Introduction à l'économétrie

Une approche moderne

2e édition

Jeffrey M. **Wooldridge**

Traduction de la 6e édition américaine
par P. André, M. Beine, S. Béreau, M. de la Rupelle, A. Durré, J.-Y. Gnabo, C. Heuchenne, M. Leturcq et M. Petitjean

LES +

- Nouvelle édition revue et augmentée
- Nouveaux exercices et problèmes
- Questions de réflexion et réponses
- Glossaire exhaustif

OUVERTURES ÉCONOMIQUES

deboeck SUPÉRIEUR

Introduction à l'économétrie

Introduction à l'économétrie

2e édition

Une approche moderne

Jeffrey M. **Wooldridge**

Traduction de la 6e édition américaine
par P. André, M. Beine, S. Béreau, M. de la Rupelle, A. Durré, J.-Y. Gnabo, C. Heuchenne, M. Leturcq et M. Petitjean

deboeck SUPÉRIEUR

Ouvrage original :
Introductory Econometrics. A Modern Approach, 6th edition by Jeffrey M. Wooldridge

Pour toute information sur notre fonds et les nouveautés dans votre domaine de spécialisation, consultez notre site web : **www.deboecksuperieur.com**

Rue du Bosquet, 7 – B-1348 Louvain-la-Neuve
Pour la traduction en langue française.

2e édition

Dépôt légal :
Bibliothèque nationale, Paris : juillet 2018
Bibliothèque royale de Belgique, Bruxelles : 2018/13647/109

ISSN : 2030-501X
ISBN : 978-2-8073-0683-7

SOMMAIRE

Partie 2
Analyse économétrique des séries temporelles

Partie 3
Thèmes avancés

AVANT-PROPOS

En rédigeant cet ouvrage, j'ai voulu combler le fossé qui existait entre la façon dont l'économétrie était enseignée dans le premier cycle universitaire et la manière dont les chercheurs pensaient et appliquaient les méthodes économétriques dans leurs travaux empiriques. J'ai en effet acquis la conviction au fil des ans qu'enseigner un cours d'introduction à l'économétrie en adoptant le point de vue d'un utilisateur professionnel permettait de simplifier la présentation de cette discipline, tout en la rendant plus attrayante.

Si j'en crois les réactions positives que les éditions précédentes de ce livre ont suscitées, il me semble avoir eu là une bonne intuition. Des enseignants aux parcours et aux intérêts divers, confrontés à des publics dont les niveaux de préparation étaient très inégaux, ont adopté l'approche moderne de l'économétrie que j'introduis dans cet ouvrage. L'application de l'économétrie à des problèmes concrets revêt une importance encore plus grande dans cette nouvelle édition. Le choix d'une méthode économétrique est toujours motivé par des problématiques auxquelles sont confrontés les chercheurs qui utilisent des données non expérimentales. L'objectif de ce livre est de comprendre et d'interpréter les hypothèses d'un modèle à la lumière d'applications empiriques concrètes. Le niveau requis en mathématiques est celui du premier cycle universitaire, que ce soit pour l'algèbre, les statistiques descriptives ou le calcul des probabilités.

UN LIVRE CONÇU POUR L'ENSEIGNANT D'AUJOURD'HUI EN ÉCONOMÉTRIE

Cette cinquième édition conserve l'organisation globale de la quatrième. La principale caractéristique de ce livre est que les thèmes identifiés le sont en fonction du type de données analysées. De ce point de vue, il s'écarte clairement de l'approche traditionnelle qui présente le modèle, en énumère toutes les hypothèses de travail, et puis s'attache à en défendre les résultats sans les relier clairement aux hypothèses de travail. L'approche que j'adopte dans la première partie est de traiter l'analyse de régression multiple à l'aide de données en coupe transversale en recourant à l'hypothèse d'échantillonnage aléatoire. Cette approche devrait convenir aux étudiants qui ont découvert l'échantillonnage aléatoire dans leur cours d'introduction à la statistique. Cela permet également aux étudiants d'opérer une distinction entre les hypothèses propres au modèle issu de la population, auxquelles nous pouvons donner une signification économique ou comportementale, et les hypothèses relatives à l'échantillonnage des données. Une fois que les étudiants ont acquis une bonne compréhension du modèle de régression basé sur l'échantillonnage aléatoire, il est alors envisageable de discuter de manière intuitive des conséquences liées à l'utilisation d'un échantillon non aléatoire.

Une autre caractéristique importante de l'approche que j'adopte dans ce livre, réside dans le fait qu'une variable est considérée comme résultant d'un processus stochastique, que cette variable soit dépendante ou explicative. Dans le cadre des sciences sociales, l'hypothèse de variables aléatoires est plus réaliste

que l'hypothèse traditionnelle de variables non aléatoires. Cette approche permet également de réduire le nombre d'hypothèses que les étudiants doivent assimiler. En réalité, l'approche traditionnelle de l'analyse de régression, qui considère les variables explicatives comme fixes d'un échantillon à l'autre, s'applique à des données collectées dans un cadre expérimental. Or, cette approche est encore omniprésente dans les ouvrages d'introduction à l'économétrie et les contorsions cérébrales nécessaires à la compréhension de ces hypothèses déroutent souvent les étudiants.

Dans le modèle issu de la population, je souligne que les hypothèses fondamentales qui sous-tendent l'analyse de régression (comme l'hypothèse d'espérance nulle de l'erreur) sont en réalité conditionnelles aux variables explicatives. Cela permet une meilleure compréhension des problèmes économétriques qui peuvent invalider les procédures classiques d'inférence statistique, telle que l'hétéroscédasticité (qui se traduit par une variance non constante de l'erreur). En me concentrant sur la population, je parviens à écarter plusieurs idées fausses que l'on rencontre dans certains ouvrages d'économétrie. Par exemple, j'explique la raison pour laquelle la mesure classique du R carré reste une mesure valide de la qualité d'ajustement d'un modèle en présence d'hétéroscédasticité (chapitre VIII) ou d'autocorrélation dans les écarts-types estimés (chapitre 12). Je montre que les tests sur la forme fonctionnelle ne devraient pas être considérés comme des tests généraux d'omission de variables (chapitre 9). J'identifie également la raison pour laquelle il est toujours intéressant d'inclure, dans un modèle de régression, des variables de contrôle supplémentaires qui ne sont pas corrélées à la variable explicative d'intérêt (chapitre 6).

Comme les hypothèses relatives à l'analyse en coupe transversale sont à la fois relativement simples et réalistes, les étudiants peuvent assez rapidement se frotter aux applications empiriques, sans devoir se préoccuper de problèmes plus épineux qui sont omniprésents dans les modèles de régression sur séries chronologiques (comme les problèmes de tendance temporelle, saisonnalité, autocorrélation, forte persistance et régression fallacieuse). En procédant de la sorte, j'espérais que mon analyse de la régression en coupe transversale, qui précède celle sur les séries chronologiques, allait être particulièrement appréciée par les enseignants dont les intérêts de recherche se situent dans le domaine de la microéconomie appliquée ; et il semble que ce soit effectivement le cas. Les personnes dont l'intérêt porte avant tout sur les séries chronologiques ont également été enthousiasmées par la structure de cet ouvrage. En retardant le traitement économétrique des séries chronologiques, je peux analyser plus sérieusement les pièges potentiels qui leur sont spécifiques. L'économétrie des séries chronologiques reçoit enfin le traitement qu'elle méritait dans un ouvrage d'introduction.

Comme dans les éditions précédentes, j'ai soigneusement sélectionné les thèmes en fonction de leur lien avec la littérature scientifique et la recherche empirique de base. Pour chaque thème, j'ai délibérément omis de nombreux tests et procédures d'estimation qui n'ont pas résisté à l'épreuve du temps, même s'ils sont encore inclus dans d'autres manuels d'économétrie. De la même façon, j'ai mis en évidence des thèmes plus récents qui ont clairement démontré leur utilité, comme le calcul de statistiques de tests robustes à l'hétéroscédasticité (ou à l'autocorrélation) de forme inconnue, l'utilisation de données portant sur plusieurs années pour l'analyse de politiques, dites discrétionnaires, ou encore l'utilisation de variables instrumentales pour faire face au problème de variable omise. Mes choix semblent avoir été judicieux car je n'ai reçu que quelques suggestions d'ajout ou de suppression.

J'adopte une approche systématique dans ce manuel : chaque thème est logiquement introduit à partir des éléments vus au préalable, et les hypothèses ne sont introduites qu'au fur et à mesure des besoins. Par exemple, les chercheurs qui utilisent l'économétrie comme outil empirique savent bien que toutes les hypothèses de Gauss-Markov ne sont pas nécessaires pour démontrer que les moindres carrés ordinaires (MCO) ne sont pas biaisés. La majorité des manuels d'économétrie présentent pourtant l'ensemble de ces hypothèses avant de prouver l'absence de biais des MCO. Il arrive même que l'hypothèse de normalité soit incluse parmi les hypothèses nécessaires à la démonstration du théorème de Gauss-Markov, alors que la normalité ne joue aucun rôle pour démontrer que les estimateurs des MCO sont les meilleurs estimateurs linéaires sans biais. L'approche systématique que j'adopte dans ce manuel est illustrée par l'ordre des hypothèses que j'utilise

pour introduire la régression multiple dans la première partie. Cet ordre suit une progression naturelle, qui nous donne l'occasion de résumer brièvement l'objectif de chaque hypothèse.

- RLM.1. Introduire le modèle issu de la population et en interpréter les paramètres (que nous espérons estimer correctement par la suite).
- RLM.2. Introduire l'échantillonnage aléatoire obtenu à partir de la population et décrire les données utilisées pour estimer les paramètres de la population.
- RLM.3. Ajouter l'hypothèse portant sur les variables explicatives, qui rend possible le calcul des estimations à l'aide de notre échantillon ; il s'agit de l'hypothèse d'absence de colinéarité parfaite.
- RLM.4. Supposer que la moyenne de l'erreur du modèle de la population, que nous ne pouvons pas observer, ne dépend pas des valeurs prises par les variables explicatives ; il s'agit de l'hypothèse d'« indépendance de la moyenne » de l'erreur, qui se résume souvent par une espérance nulle de l'erreur dans la population. Sans elle, l'absence de biais des MCO est impossible.

En utilisant les hypothèses RLM.1 à RLM.3, il est possible d'examiner les propriétés algébriques des MCO, c'est-à-dire les propriétés des MCO qui s'appliquent à n'importe quel jeu particulier de données. Si l'hypothèse RLM.4 est ajoutée aux trois premières, les MCO sont sans biais (et convergents). L'hypothèse RLM.5 d'homoscédasticité est utile pour dériver le théorème de Gauss-Markov et rendre valides les habituelles formules de variance des MCO. Sous les cinq premières hypothèses, les estimateurs des MCO sont les meilleurs estimateurs linéaires sans biais. L'hypothèse RLM.6 de normalité est la dernière des six hypothèses sur lesquelles repose le modèle linéaire classique. Ces six hypothèses sont requises pour obtenir des tests exacts d'inférence statistique et des estimateurs des MCO dont la variance est la plus petite parmi tous les estimateurs sans biais, qu'ils soient linéaires ou pas.

Dans la seconde partie, je me lance dans l'étude des propriétés des MCO en grand échantillon et l'analyse de régression sur séries chronologiques. Une présentation et une discussion minutieuses des hypothèses de travail permettent une transition plus facile vers la troisième partie. Dans cette troisième et dernière partie, j'aborde des sujets plus pointus, tels que l'utilisation de données empilées, l'exploitation de bases de données en panel, et l'application de variables instrumentales. En règle générale, je me suis efforcé de donner une vision unifiée de l'économétrie selon laquelle tous les estimateurs et les statistiques de tests sont obtenus en se reposant sur quelques principes à la fois logiques sur le plan intuitif et rigoureusement justifiés sur le plan formel. Par exemple, les étudiants comprennent d'autant plus facilement les tests d'hétéroscédasticité et d'autocorrélation qu'ils ont acquis une maîtrise de la régression. Cette manière de procéder peut être mise en contraste avec le traitement décousu de recettes qui s'appliquent souvent à des procédures de tests dépassées.

Dans ce manuel, j'insiste particulièrement sur les relations *ceteris paribus*. C'est la raison pour laquelle je passe directement de l'analyse de régression simple à l'analyse de régression multiple, l'objectif étant que les étudiants puissent analyser le plus rapidement possible des sujets empiriques intéressants. J'accorde de l'importance à l'analyse de politiques publiques en utilisant des données diverses et variées. Par exemple, j'ai tenu à introduire le plus simplement possible deux exemples de sujets importants sur le plan pratique : l'utilisation de variables de substitution dans le but d'obtenir des effets *ceteris paribus* et l'interprétation des effets partiels dans les modèles à termes d'interaction.

QUOI DE NEUF DANS CETTE ÉDITION ?

J'ai ajouté de nouveaux exercices dans presque tous les chapitres, y compris les annexes. La plupart des nouveaux exercices sur ordinateur reposent sur de nouveaux jeux de données. Par exemple, un nouveau jeu de données porte sur la performance dans le premier cycle universitaire d'étudiants préalablement inscrits dans une école secondaire du réseau catholique. Il y a également des séries chronologiques sur la cote de

popularité des présidents américains et sur le prix de l'essence. J'ai également ajouté quelques problèmes plus compliqués, exigeant le recours à des démonstrations.

J'ai apporté quelques changements au texte qu'il me semble important de souligner. Le chapitre 2 inclut une discussion plus approfondie de la relation entre le coefficient de régression et le coefficient de corrélation. Le chapitre 3 clarifie la question de la comparaison des R-carrés dans des modèles où des données manquent pour certaines variables, ce qui réduit automatiquement la taille de l'échantillon pour les régressions disposant d'un plus grand nombre de variables.

Le chapitre 6 introduit la notion d'effet marginal moyen (EMM) dans les modèles à la fois linéaires par rapport aux paramètres et non linéaires par rapport aux variables. Cette analyse s'effectue en recourant à deux fonctions non linéaires très fréquentes par rapport aux variables : les fonctions quadratiques et l'ajout de termes d'interaction. La notion d'EMM, qui était implicite dans les éditions précédentes, est devenue un concept important dans les études empiriques ; calculer et interpréter les EMM dans le contexte des estimations des MCO est une compétence très utile. Dans le cadre de cours plus avancés, l'introduction du chapitre 6 facilite l'étude des EMM dans les modèles non linéaires par rapport aux paramètres étudiés au chapitre 17. Le chapitre 17 inclut désormais une plus large discussion des EMM ainsi que des tableaux supplémentaires dans lesquels sont directement reprises les estimations des coefficients calculés par la technique des EMM.

Dans le chapitre 8, j'ai affiné la discussion concernant le problème d'hétéroscédasticité en incluant une analyse plus poussée des tests de Chow ainsi qu'une description plus précise des moindres carrés pondérés lorsque les poids doivent être estimés. Le chapitre 9, qui aborde des sujets facultatifs, légèrement plus pointus, inclut des définitions de termes que l'on retrouve fréquemment dans la littérature très vaste portant sur les données manquantes. Une pratique courante dans les travaux empiriques consiste, pour chaque variable présentant des données manquantes, à inclure une variable indicatrice de données manquantes dans une régression linéaire multiple. Le chapitre 9 discute de la mise en œuvre de cette méthode et la possibilité d'obtenir des estimateurs sans biais et convergents.

Dans le chapitre 14, le traitement des effets inobservés dans les modèles de panel a été approfondi. L'objectif a été d'inclure une discussion sur les jeux de données en panel non cylindré et sur la possibilité d'utiliser, dans un tel cas de figure, des effets fixes, aléatoires non, ou aléatoires corrélés. Ce chapitre contient également une analyse plus détaillée de l'utilisation des effets fixes ou aléatoires lorsqu'il s'agit de traiter des échantillons par grappes. J'ai également abordé quelques problèmes subtils qui peuvent survenir lorsque des écarts-types groupés sont utilisés sur des données issues d'un processus d'échantillonnage aléatoire.

Le chapitre 15 contient désormais une discussion plus fouillée du problème lié à l'utilisation de variables instrumentales faibles. De cette manière, les étudiants pourront en comprendre l'essentiel sans devoir consulter des documents plus pointus.

UN OUVRAGE CONÇU POUR LES ÉTUDIANTS UNIVERSITAIRES DU PREMIER CYCLE, MAIS ÉGALEMENT ADAPTABLE AUX ÉTUDIANTS DU SECOND CYCLE

Ce livre est conçu pour des étudiants universitaires du premier cycle (licence ou baccalauréat universitaire), inscrits en économie ou en gestion. Ces étudiants ont généralement suivi des cours d'algèbre, de statistique et d'introduction au calcul de probabilités. Si tel ne devait pas en être le cas, les annexes A, B et C contiennent toutes les références contextuelles nécessaires. Un cours d'économétrie organisé sur un seul trimestre (ou semestre) ne peut pas aborder les thèmes plus avancés de la troisième partie. Un cours classique d'introduction

à l'économétrie couvre les chapitres 1 à 8, qui abordent les bases des régressions simple et multiple pour les données en coupe transversale. Ces chapitres doivent être accessibles à l'écrasante majorité des étudiants de premier cycle, à condition que l'accent soit mis sur l'intuition et l'interprétation d'exemples empiriques. La plupart des enseignants désireront également traiter, au moins en partie et à des degrés divers, les chapitres portant sur l'utilisation de séries chronologiques dans l'analyse de régression (chapitres 10, 11 et 12). Dans mon cours organisé sur un semestre à l'université d'État du Michigan, j'étudie le chapitre 10 en détail ; je donne un aperçu du chapitre 11 ; et je ne fais qu'évoquer l'autocorrélation du chapitre 12. Il me semble que ce cours d'un semestre donne aux étudiants une assise suffisante pour leur permettre de réaliser des travaux empiriques de qualité par la suite. Le chapitre 9 contient des sujets assez spécifiques à l'utilisation de données en coupe transversale, tels que la présence d'observations isolées ou d'échantillons non aléatoires. Dans le cadre d'un cours organisé sur un semestre, ce chapitre peut être laissé de côté sans mettre en péril la cohérence de l'ensemble.

La structure du manuel convient également à un cours consacré exclusivement à l'analyse de régression sur données en coupe transversale, dont l'intérêt peut porter sur l'analyse de politiques publiques par exemple. Les chapitres relatifs aux séries chronologiques (chapitre 10, 11 et 12) peuvent être laissés de côté et être remplacés par des thèmes abordés dans les chapitres 9, 13, 14 et 15. Le chapitre 13 est « avancé » dans le sens où il traite de données dont la structure est originale ; il s'agit de données en coupe transversale empilées et de données de panel sur deux périodes uniquement. Ce type de données est particulièrement utile pour l'analyse de politiques discrétionnaires (que le pouvoir politique ou le conseil d'administration d'une entreprise peut instaurer, par exemple). La compréhension de ce chapitre ne posera aucun problème aux étudiants ayant bien assimilé les chapitres 1 à 8. En revanche, le chapitre 14 aborde des méthodes plus avancées en économétrie des données de panel ; il devrait plutôt faire l'objet d'un second cours. Pour conclure en beauté un cours sur l'analyse en coupe transversale, je conseille d'introduire les bases de l'estimation par variables instrumentales, présentées au chapitre 15.

Pour un séminaire consacré à la réalisation de travaux de recherche plus pointus, je me suis servi de plusieurs thèmes abordés dans la troisième partie de ce livre, en particulier dans les chapitres 13, 14, 15 et 17. Lorsque les étudiants ont suivi un cours d'introduction à l'économétrie et qu'ils ont été sensibilisés à l'utilisation des données de panel, des variables instrumentales, et des modèles à variable dépendante limitée, ils sont capables de comprendre une très grande partie de la littérature empirique consacrée à l'étude des sciences sociales. Le chapitre 17 propose d'ailleurs une introduction aux modèles à variable dépendante limitée les plus répandus.

Ce texte convient également à un cours d'introduction à l'économétrie organisé durant le second cycle universitaire, en reconnaissant que l'accent doit être mis davantage sur les applications empiriques que sur les démonstrations réalisées à l'aide de l'algèbre matricielle. Plusieurs enseignants ont utilisé ce manuel au niveau du « master » dans le cadre de l'analyse de politiques discrétionnaires. Pour les enseignants qui désirent présenter l'économétrie sous forme matricielle, les annexes D et E contiennent un rappel des notions d'algèbre matricielle et une introduction au modèle de régression multiple sous forme matricielle.

Les doctorants de l'Université d'État du Michigan, dont les thèses portent sur des problématiques très diverses (en comptabilité, économie de l'agriculture, économie du développement, économie de l'éducation, finance, économie internationale, économie du travail, macroéconomie, science politique ou finances publiques), ont également apprécié ce manuel en raison du pont qu'il permet de jeter entre la théorie économétrique et la nature empirique de leurs travaux.

CARACTÉRISTIQUES DE L'OUVRAGE

De nombreuses questions de réflexion, assez brèves, sont insérées dans le corps même des chapitres de ce manuel. Ces questions, dont les réponses sont reprises dans l'annexe F, permettent aux étudiants de vérifier rapidement si

les notions qu'ils viennent de découvrir ont été correctement assimilées. Chaque chapitre contient également de nombreux exemples numérotés, véritables études de cas en miniature, qui sont inspirés d'articles publiés dans la littérature scientifique. Je me suis toutefois permis d'en simplifier l'analyse, en veillant à ne pas en trahir l'esprit.

Les exercices disponibles à la fin des chapitres, ainsi que les exercices sur ordinateur, sont axés sur l'analyse empirique plutôt que sur les démonstrations théoriques. Les étudiants doivent apprendre à développer un raisonnement précis, en se basant sur ce qu'ils ont appris dans le chapitre de référence. Les exercices informatiques permettent souvent d'approfondir les exemples qui ont été analysés dans le chapitre. De nombreux exercices requièrent l'utilisation de données tirées ou inspirées d'articles publiés dans la littérature scientifique et dont les étudiants peuvent disposer gratuitement.

Une des particularités de ce manuel est le glossaire relativement exhaustif, dont les définitions brèves rafraîchiront la mémoire des étudiants qui doivent plancher sur leurs examens, lire la littérature en économétrie ou réaliser des travaux empiriques. Cette cinquième édition contient plusieurs nouvelles entrées.

BASES DE DONNÉES DISPONIBLES EN SIX FORMATS[1]

Cette nouvelle édition permet dorénavant d'importer directement les bases de données disponibles dans les logiciels R et Minitab®. L'enseignant a l'embarras du choix : plus d'une centaine de bases de données sont disponibles ; chacune d'entre elles peut être directement importée dans les logiciels Stata®, EViews®, Minitab®, Microsoft® Excel, R et TeX. Comme la plupart de ces bases de données sont tirées d'articles publiés dans la littérature scientifique, la taille de certaines d'entre elles est importante. Ces bases de données ne sont naturellement pas reproduites intégralement dans le corps du texte, même s'il s'est parfois révélé utile d'en tirer quelques extraits pour en illustrer la diversité. Comme je l'ai déjà précisé, ce manuel donne une place de prédilection aux analyses empiriques que les exercices sur ordinateur permettent de réaliser.

UN MANUEL DE DESCRIPTION DES BASES DE DONNÉES (en anglais)

Les bases de données utilisées dans cet ouvrage sont décrites dans un manuel disponible en ligne. Ce manuel, unique en son genre et créé par l'auteur lui-même, identifie les sources de ces données et propose plusieurs pistes que l'enseignant peut suivre s'il désire construire ses propres exercices et travaux. Sont également indiquées les pages de l'ouvrage principal, auxquelles chaque base de données est mentionnée, ce qui permet à l'étudiant et à l'enseignant de saisir rapidement l'utilisation qui en a été faite. Les étudiants désireront sans doute lire en priorité la description des séries disponibles, alors que les enseignants chercheront plutôt à créer de nouveaux exercices et problèmes. C'est la raison pour laquelle le manuel mentionne également des bases de données qui ne sont pas utilisées dans le texte principal. Il contient même des suggestions que l'enseignant peut suivre s'il souhaite améliorer ces bases de données. Ce manuel est disponible sur le site compagnon du livre à l'adresse suivante : http://login.cengage.com. Les étudiants peuvent y accéder gratuitement à l'adresse www.cengagebrain.com.

L'ORGANISATION PLUS DÉTAILLÉE DE VOTRE COURS

J'ai donné précédemment plusieurs indications quant à la structure générale d'un cours d'économétrie du premier ou second cycle universitaire. J'ai également commenté le contenu de plusieurs chapitres. Je donne ici un aperçu plus spécifique des sections qu'un enseignant peut décider d'inclure ou non dans son cours.

1 Ces bases de données sont disponibles uniquement en anglais. Plus d'informations sur www.deboecksuperieur.com.

Le chapitre 9 contient plusieurs exemples intéressants, comme le cas de la régression du salaire qui inclut le quotient intellectuel comme variable explicative. Il est possible de présenter ces exemples aux étudiants sans devoir passer par une discussion formelle des variables de substitution. En règle générale, je parle plus en détail des variables de substitution après avoir couvert l'analyse de la régression en coupe transversale. Dans le cadre d'un cours organisé sur un semestre, je laisse tomber l'inférence robuste à l'autocorrélation et les modèles dynamiques d'hétéroscédasticité, qui sont introduits dans le chapitre 12.

Même dans le cadre d'un second cours, je consacre peu de temps au chapitre 16, qui porte sur les équations simultanées. Les opinions des enseignants diffèrent lorsqu'il s'agit de statuer sur l'utilité d'enseigner les modèles à équations simultanées aux étudiants du premier cycle universitaire. Mon sentiment est que l'utilisation des modèles à équations simultanées est souvent abusive (voir le chapitre 16 pour une discussion plus approfondie). Dans bien des cas, lorsque la problématique empirique est analysée avec soin, l'estimation par variables instrumentales se justifie davantage par l'omission d'une variable ou la présence d'une erreur de mesure, que par une détermination simultanée des variables. C'est la raison pour laquelle, dans le chapitre 15, j'ai recouru prioritairement au problème d'omission de variables pour justifier l'estimation par variables instrumentales. Bien entendu, les modèles à équations simultanées sont indispensables pour estimer les fonctions d'offre et de demande ; ils s'appliquent également à d'autres cas importants.

Le seul chapitre qui porte sur les modèles intrinsèquement non linéaires dans leurs paramètres est le chapitre 17, dont la compréhension requiert un effort supplémentaire de la part des étudiants. Ce chapitre débute par l'analyse des modèles probit et logit, dont la variable de réponse est binaire dans les deux cas. Ce chapitre couvre également le modèle Tobit et la régression censurée, ce qui peut être considéré comme inhabituel dans un manuel d'introduction à l'économétrie. J'indique clairement que le modèle Tobit est intéressant, dans le contexte d'un échantillonnage aléatoire, lorsque la variable de réponse donne lieu à de nombreuses solutions en coin. Quant au modèle de régression censurée, il est approprié lorsque le processus aléatoire de collecte de données conduit à n'observer la variable dépendante qu'en dessous (ou qu'au-dessus) d'un seuil connu, souvent fixé de manière arbitraire.

Le chapitre 18 porte sur des thèmes plus avancés de l'économétrie des séries chronologiques, notamment les tests de racine unitaire et la cointégration. Je n'aborde ces sujets que dans le cadre d'un second cours d'économétrie, que ce cours soit organisé au niveau du premier cycle ou au niveau du « master ». Le chapitre 18 inclut également une introduction détaillée à la prévision.

Le chapitre 19 devrait être inclus dans un cours au terme duquel la rédaction d'un travail empirique est exigée. Plus approfondi que dans d'autres ouvrages d'économétrie, ce chapitre opère une synthèse des méthodes qui permettent un traitement approprié des structures de données et problèmes auxquels les étudiants sont le plus souvent confrontés ; j'identifie les pièges méthodologiques à éviter ; j'explique en détail la marche à suivre lors de la rédaction d'un travail empirique ; et je conclus en proposant quelques idées de recherche empirique.

REMERCIEMENTS

Je remercie les personnes qui ont relu le texte de la cinquième édition, en n'oubliant pas celles qui ont commenté la quatrième.

Erica Johnson,
Gonzaga University

Mary Ellen Benedict,
Bowling Green State University

Yan Li,
Temple University

Melissa Tartari,
Yale University

Michael Allgrunn,
University of South Dakota

Gregory Colman,
Pace University

Yoo-Mi Chin,
Missouri University of Science and Technology

Arsen Melkumian,
Western Illinois University

Kevin J. Murphy,
Oakland University

Kristine Grimsrud,
University of New Mexico

Will Melick,
Kenyon College

Philip H. Brown,
Colby College

Argun Saatcioglu,
University of Kansas

Ken Brown,
University of Northern Iowa

Michael R. Jonas,
University of San Francisco

Melissa Yeoh,
Berry College

Nikolaos Papanikolaou,
SUNY at New Paltz

Konstantin Golyaev,
University of Minnesota

Soren Hauge,
Ripon College

Kevin Williams,
University of Minnesota

Hailong Qian,
Saint Louis University

Rod Hissong,
University of Texas at Arlington

Steven Cuellar,
Sonoma State University

Yanan Di,
Wagner College

John Fitzgerald,
Bowdoin College

Philip N. Jefferson,
Swarthmore College

Yongsheng Wang,
Washington and Jefferson College

Sheng-Kai Chang,
National Taiwan University

Damayanti Ghosh,
Binghamton University

Susan Averett,
Lafayette College

Kevin J. Mumford,
Purdue University

Nicolai V. Kuminoff,
Arizona State University

Subarna K. Samanta,
The College of New Jersey

Jing Li,
South Dakota State University

Gary Wagner,
University of Arkansas – Little Rock

Kelly Cobourn,
Boise State University

Timothy Dittmer,
Central Washington University

Daniel Fischmar,
Westminster College

Subha Mani,
Fordham University

John Maluccio,
Middlebury College

James Warner,
College of Wooster

Christopher Magee,
Bucknell University

Andrew Ewing,
Eckerd College

Debra Israel,
Indiana State University

Jay Goodliffe,
Brigham Young University

Stanley R. Thompson,
The Ohio State University

Michael Robinson,
Mount Holyoke College

Ivan Jeliazkov,
University of California, Irvine

Heather O'Neill,
Ursinus College

Leslie Papke,
Michigan State University

Timothy Vogelsang,
Michigan State University

Stephen Woodbury,
Michigan State University

Plusieurs changements que j'ai évoqués précédemment ont été introduits dans cette édition à la suite des commentaires que ces collègues ont eu la gentillesse de me transmettre. Je poursuis d'ailleurs la réflexion sur les modifications à apporter dans les éditions ultérieures.

De nombreux étudiants et assistants, trop nombreux pour que je puisse les nommer ici, ont repéré des coquilles qui subsistaient dans les éditions précédentes. Ils m'ont également suggéré de reformuler certains paragraphes. Je leur en suis reconnaissant.

J'ai pris une nouvelle fois beaucoup de plaisir à collaborer avec l'équipe de South-Western/Cengage Learning. Mike Worls, responsable des acquisitions, que je connais depuis longtemps, a appris à me guider, avec délicatesse et fermeté. Julie Warwick est rapidement parvenue à relever le défi que constitue l'édition d'un manuel technique et dense. Sa lecture attentive du manuscrit et son sens aiguisé du détail ont considérablement amélioré la qualité de cette cinquième édition.

Jean Buttrom a brillamment rempli son rôle de directeur de production et Karunakaran Gunasekaran, de PreMediaGlobal, a supervisé la réalisation du projet et la composition du manuscrit avec beaucoup d'efficacité et de professionnalisme.

Je remercie tout particulièrement Martin Biewen, de l'université de Tübingen, qui a créé les diapositives PowerPoint qui illustrent les chapitres de cet ouvrage. Mes remerciements vont aussi à Francis Smart qui a aidé à la création des séries de données en R.

Ce livre est dédié à mon épouse, Leslie Papke, qui a directement contribué à cette édition en rédigeant les versions initiales des diapositives en *Word scientifique* pour la troisième partie. Elle a également utilisé ces diapositives dans le cours de politique publique qu'elle enseigne à l'université. Enfin, la contribution de nos enfants doit être soulignée : Edmund m'a aidé à mettre à jour le manuel de données et Gwenyth nous a agréablement divertis grâce à ses talents artistiques.

Jeffrey M. Wooldridge

REMERCIEMENTS DES TRADUCTEURS

Nous remercions Jean-Charles Wijnandts, Thomas Renault et Aude de la Rupelle d'avoir accepté de relire certains chapitres et annexes de cet ouvrage. Si des erreurs se sont glissées dans cette nouvelle édition, nous invitons chaleureusement les lecteurs à nous les communiquer.

À PROPOS DE L'AUTEUR

Jeffrey M. Wooldridge est professeur d'économie à l'Université d'État du Michigan (MSU) où il enseigne depuis 1991. De 1986 à 1991, il a été professeur d'économie au Massachusetts Institute of Technology (MIT). Il a obtenu sa licence en économie et informatique à l'Université de Californie à Berkeley en 1982, et sa thèse de doctorat en économie à l'Université de Californie à San Diego en 1986. Le professeur Wooldridge a publié de nombreux articles dans des revues de renommée internationale, ainsi que plusieurs chapitres de livres. Il est également l'auteur d'*Econometric Analysis of Cross Section and Panel Data*. Il a reçu de nombreuses récompenses : une bourse de recherche de la Fondation Alfred P. Sloan, le prix Plura Scripsit de la revue *Econometric Theory*, le prix Sir Richard Stone du *Journal of Applied Econometrics*, et le titre d'enseignant de l'année du second cycle au MIT, à trois reprises. Il est membre de l'*Econometric Society* et du *Journal of Econometrics* ; il est le coéditeur du *Journal of Econometric Methods*. Dans le passé, il a été l'éditeur du *Journal of Business and Economic Statistics* et le coéditeur en économétrie de la revue *Economics Letters*. Il a été membre du comité de rédaction d'*Econometric Theory*, *Journal of Economic Literature*, *Journal of Economics*, *Review of Economics and Statistics* et *Stata Journal*. Il a également été consultant occasionnel pour Arthur Andersen, Charles River Associates, le Washington State Institute for Public Policy et Stratus Consulting.

CHAPITRE 1

LA NATURE DE L'ÉCONOMÉTRIE ET LA STRUCTURE DES DONNÉES ÉCONOMIQUES

Traduction de Marion Leturcq

Le chapitre 1 définit le champ d'application de l'économétrie ; il soulève également des questions d'ordre général, qui se posent lors de l'analyse de données en économétrie. La section 1.1 discute brièvement de l'objectif et de la portée de l'économétrie. Son intégration dans l'analyse économique est également abordée dans cette section. La section 1.2 présente des exemples montrant que la théorie économique sert à construire des modèles dont l'estimation requiert l'utilisation de données. La section 1.3 examine les types de bases de données qui sont utilisées en gestion et en économie. La section 1.4 explique de manière intuitive les difficultés auxquelles il faut faire face lorsqu'il s'agit de déduire des liens de causalité dans le domaine des sciences sociales.

1.1 QU'EST-CE QUE L'ÉCONOMÉTRIE ?

Imaginez que vous soyez recruté par les pouvoirs publics pour évaluer l'efficacité d'un programme de formation professionnelle financé par des fonds publics. Ce programme enseigne aux employés les différentes utilisations possibles de l'ordinateur dans le cycle de production de l'entreprise. Il s'étale sur vingt semaines et offre des cours en dehors des heures de travail. Tous les salariés sont libres de participer à l'ensemble du programme ou à une partie seulement. Le cas échéant, vous devez évaluer l'effet du programme de formation professionnelle sur le salaire horaire de chaque employé.

Imaginez maintenant que vous travailliez pour une banque d'investissement. Vous devez étudier les rendements de plusieurs stratégies qui consistent à investir dans des bons du trésor américains de différentes maturités, l'objectif étant de vérifier si ces stratégies sont conformes aux théories économiques sous-jacentes.

À première vue, répondre à ces questions peut sembler insurmontable. À ce stade, vous n'avez qu'une vague idée des données auxquelles il faudrait recourir. À la fin de cet ouvrage, vous devriez être capable d'utiliser les méthodes économétriques les plus appropriées pour évaluer en bonne et due forme un programme de formation professionnelle ou tester une théorie économique simple.

L'économétrie est fondée sur le développement de méthodes statistiques dont le but est d'estimer des relations économiques, tester des théories économiques, évaluer et mettre en œuvre la politique du gouvernement et des entreprises. Une des utilisations les plus courantes de l'économétrie consiste à prédire l'évolution de variables macroéconomiques importantes, comme les taux d'intérêt, les taux d'inflation ou le produit intérieur brut. Les prévisions d'indicateurs économiques sont très visibles et largement diffusées, mais les méthodes économétriques peuvent aussi être utilisées dans des domaines de l'économie qui n'ont aucun rapport avec la prévision macroéconomique. Nous étudierons par exemple les effets des dépenses de campagne électorale sur les résultats des élections. Dans le domaine de l'éducation, nous analyserons l'effet de subsides octroyés aux écoles sur la performance des étudiants. Nous apprendrons aussi à utiliser les méthodes économétriques pour générer des prévisions à partir de séries chronologiques.

L'économétrie s'est progressivement développée comme une discipline distincte de la statistique mathématique au fur et à mesure qu'elle s'est intéressée aux problèmes inhérents à la collecte et à l'analyse de données économiques non-expérimentales. Les **données non-expérimentales** ne proviennent pas d'expérimentations contrôlées sur les individus, les entreprises ou certains segments de l'économie. (Les données non-expérimentales sont parfois appelées **données observationnelles** ou **données rétrospectives**, afin de mettre en valeur le fait que le chercheur recueille les données de manière passive.) Les données expérimentales sont souvent issues d'expérimentations, réalisées au sein de laboratoires en sciences naturelles ; elles sont bien plus difficiles à obtenir en sciences sociales. Même s'il est parfois possible de concevoir des expérimentations sociales pour répondre à des questions économiques, leur réalisation est souvent impossible, hors de prix ou moralement inacceptable. Nous verrons quelques exemples précis de différences entre des données expérimentales et des données non-expérimentales dans la section 1.4.

Bien sûr, les économètres se sont inspirés des statisticiens dès que cela leur était possible. La méthode des régressions multiples est au cœur de ces deux disciplines, mais son champ d'analyse et son interprétation peuvent différer de manière notable. Les économistes ont également mis au point des techniques nouvelles pour tenir compte de la complexité des données économiques et pour tester les prédictions des théories économiques.

1.2 LES ÉTAPES DE L'ANALYSE ÉCONOMIQUE EMPIRIQUE

Les méthodes économétriques sont utilisées dans quasiment toutes les branches de l'économie appliquée. Elles entrent en jeu dès que nous avons une théorie économique à tester ou qu'il existe un lien logique entre plusieurs variables. La nature de ce lien peut d'ailleurs revêtir une importance toute particulière lorsqu'il s'agit de prendre une décision commerciale ou de recommander une politique économique. Une **analyse empirique** utilise des données pour tester une théorie ou estimer le lien entre plusieurs variables.

Comment doit-on structurer une analyse économique empirique ? Même si cela peut paraître évident, il faut d'abord insister sur l'importance que revêt, dans toute analyse empirique, la formulation de la question d'intérêt. La question peut consister à tester un aspect particulier d'une théorie économique ; il peut s'agir également de tester les effets d'une politique menée par un gouvernement. En principe, les méthodes économétriques peuvent être utilisées pour répondre à un large éventail de questions.

Dans certains cas, en particulier lorsqu'il s'agit de tester une théorie économique, l'élaboration d'un **modèle économique** formel est requise. Un modèle économique est composé d'équations mathématiques qui décrivent des liens divers entre variables. Les économistes sont connus pour leur capacité à modéliser une large palette de comportements. Par exemple, en microéconomie, les décisions individuelles de consommation, sous contrainte de budget, sont décrites par des modèles mathématiques. Le postulat de base sous-jacent à ces modèles est la *maximisation de l'utilité*. L'hypothèse selon laquelle les individus, soumis à des contraintes de ressources, font des choix dans le but de maximiser leur bien-être, offre un cadre d'analyse puissant qui permet la mise en place de modèles économiques dont les solutions sont analytiques et les prédictions sont claires. Dans le contexte des décisions de consommation, la maximisation de l'utilité conduit à un ensemble d'équations de demande. Dans une équation de demande, la quantité de chaque produit dépend de son prix, du prix des biens compléments et substituts, du revenu du consommateur et des caractéristiques individuelles qui affectent les goûts. Ces équations peuvent former la base d'une analyse économétrique de la demande du consommateur.

Les économistes ont utilisé des outils économiques de base, comme le cadre d'analyse de la maximisation de l'utilité, pour expliquer des comportements qui ne sont pas, à première vue, de nature économique. Un exemple classique est le modèle économique de Becker (1968) pour expliquer la criminalité.

EXEMPLE 1.1
Un modèle économique de la criminalité

Dans un article précurseur, le prix Nobel Gary Becker a proposé de décrire la participation d'un individu à des activités criminelles au moyen d'un cadre d'analyse de maximisation de l'utilité. Si certaines activités criminelles conduisent à une récompense économique claire, la plupart des comportements criminels sont aussi coûteux. Ce type de comportements empêche le criminel de participer à d'autres activités comme l'emploi légal, ce qui constitue son coût d'opportunité. Il y a également des coûts associés à la possibilité d'être arrêté et, si on est reconnu coupable, des coûts liés à l'incarcération. Dans la perspective de Becker, la décision d'entreprendre une activité illégale est une décision d'allocation de ressources, qui prend en compte les coûts et avantages des activités en lice.

Sous des hypothèses très générales, il est possible de déduire une équation du temps consacré à une activité criminelle en fonction de plusieurs facteurs. On pourrait représenter cette fonction par :

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \qquad [1.1]$$

où
y = nombre d'heures passées à des activités criminelles,
x_1 = « salaire » pour une heure passée à une activité criminelle,
x_2 = salaire horaire pour un emploi légal,
x_3 = revenu de sources différentes de l'emploi ou du crime,
x_4 = probabilité d'être arrêté,
x_5 = probabilité d'être reconnu coupable si arrêté,
x_6 = sentence si reconnu coupable,
x_7 = âge.

La décision d'une personne de participer à une activité criminelle peut être affectée par d'autres facteurs, mais la liste ci-dessus est représentative de ce qui pourrait être issu d'une analyse économique formelle. Comme de coutume en économie, nous n'avons pas spécifié la fonction $f(.)$ en (1.1). Elle dépend d'une fonction d'utilité sous-jacente, qui est rarement connue. Néanmoins, on peut utiliser la théorie économique (ou l'introspection) pour prédire l'effet que chaque variable pourrait avoir sur l'activité criminelle. Tels sont les éléments de base d'une analyse économétrique de la criminalité individuelle.

La modélisation économique formelle est parfois le point de départ de l'analyse empirique, mais il est courant d'utiliser la théorie économique de manière moins systématique, voire de se reposer entièrement sur son intuition. Vous conviendrez que les déterminants de la criminalité qui apparaissent dans l'équation (1.1) sont de l'ordre du bon sens et que nous pourrions aboutir directement à cette équation sans partir d'un principe de maximisation de l'utilité. C'est en effet un point de vue acceptable, même si la modélisation apporte, dans certaines circonstances, un éclairage fort utile, que l'intuition seule est incapable d'apporter. L'exemple suivant propose une équation qui découle d'un raisonnement moins formel.

EXEMPLE 1.2
Formation professionnelle et productivité du salarié

Considérons le problème qui a été introduit au début de la section 1.1. Un économiste du travail veut étudier les effets de la formation professionnelle sur la productivité des employés. Dans ce cas, le recours à une théorie économique formelle n'est pas absolument nécessaire. Il suffit de comprendre les bases de l'économie pour se rendre compte que des facteurs tels que l'éducation, l'expérience et la formation professionnelle auront un effet sur la productivité de l'employé. D'ailleurs, les économistes savent très bien que les salariés sont payés en fonction de leur productivité. Ce raisonnement simple aboutit au modèle suivant :

$$wage = f(educ, exper, training) \qquad [1.2]$$

où
wage = salaire horaire,
educ = nombre d'années d'études,
exper = nombre d'années d'expérience professionnelle,
training = nombre de semaines de formation professionnelle.

Naturellement, d'autres facteurs affectent le taux de salaire, mais l'équation (1.2) capture l'essentiel du problème.

Après avoir spécifié le modèle économique, il est nécessaire de le transformer en ce qu'on appelle un **modèle économétrique**. Il est important de savoir comment nous passons de l'un à l'autre, puisque cet ouvrage est précisément consacré à l'étude des modèles économétriques. Partons de l'équation (1.1). Il faut d'abord spécifier la forme de la fonction $f(.)$ avant d'entreprendre une analyse économétrique. L'équation (1.1) présente un autre problème : que faisons-nous des variables qui, dans les faits, ne peuvent pas être observées ? Pensons par exemple au salaire qu'une personne peut tirer de l'exercice d'activités criminelles. En principe, cette quantité est définie, mais il serait difficile, voire impossible, de mesurer ce salaire pour un individu donné. Même des variables, comme la probabilité d'être arrêté, ne peuvent pas être obtenues pour chaque individu ; on peut néanmoins observer des statistiques pertinentes sur le nombre d'arrestations et en déduire des variables qui donnent une approximation de la probabilité d'être arrêté pour un individu. Il y a tellement d'autres facteurs qui peuvent avoir un effet sur le comportement criminel qu'on ne peut pas en établir la liste, encore moins les observer. Il faudra pourtant en tenir compte, d'une manière ou d'une autre.

Les ambiguïtés intrinsèques du modèle économique de la criminalité sont résolues en spécifiant le modèle économétrique suivant :

$$crime = \beta_0 + \beta_1 wage + \beta_2 othinc + \beta_3 freqarr + \beta_4 freqconv + \beta_5 avgsen + \beta_6 age + u \qquad [1.3]$$

où
crime = une mesure de la fréquence de l'activité criminelle,
wage = le salaire qui peut être touché dans l'emploi légal,
othinc = autres sources de revenu (biens financiers, héritage, etc.),
freqarr = la fréquence des arrestations lors d'infractions antérieures (dans l'espoir de se rapprocher de la probabilité d'arrestation pour chaque individu),
freqconv = la fréquence de condamnation,
avgsen = la durée moyenne de la sentence en cas de condamnation.

Le choix de ces variables est déterminé par la théorie économique mais aussi par des considérations liées aux données. Le terme u contient des facteurs inobservés, comme le salaire provenant d'activités criminelles, les valeurs morales, le contexte familial ; il contient également les erreurs incluses dans la mesure des variables, comme pour la fréquence de l'activité criminelle et la probabilité d'être arrêté. Même si nous pouvons ajouter des variables concernant le contexte familial (comme le nombre de frères et sœurs, l'éducation des parents, etc.), il est impossible d'éliminer u entièrement. En réalité, la prise en compte de ce *terme d'erreur* ou *terme de perturbation* est sans doute la composante la plus importante de l'analyse économétrique.

Les constantes $\beta_0, \beta_1, \ldots, \beta_6$ sont les *paramètres* du modèle économétrique. Elles décrivent dans quelles directions et dans quelle mesure la variable *crime* est reliée aux facteurs utilisés dans le modèle pour l'expliquer.

Le modèle économétrique qui correspond à l'exemple 1.2 pourrait s'écrire de la manière suivante :

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u \qquad [1.4]$$

où le terme u contient des facteurs comme les aptitudes innées, la qualité des études, le contexte familial, et une myriade d'autres facteurs qui peuvent influencer le salaire d'une personne. Si nous sommes intéressés par l'effet de la formation professionnelle, alors β_3 est le paramètre d'intérêt.

En règle générale, l'analyse économétrique débute par la spécification d'un modèle économétrique qui ne requiert pas la prise en compte des détails techniques liés à la dérivation du modèle théorique sous-jacent. Dans cet ouvrage, nous allons adopter cette approche, car l'élaboration complète d'un modèle économique, comme celui portant sur la criminalité, requiert beaucoup de temps et nous conduirait à aborder des aspects techniques, souvent compliqués, de la théorie économique. Dans les exemples que nous allons rencontrer,

le raisonnement économique joue un rôle important et nous tiendrons compte des implications de la théorie économique sous-jacente dans la spécification du modèle économétrique. Dans le cas du modèle économique de la criminalité, nous partirons du modèle économétrique décrit dans l'équation (1.3) et nous utiliserons le raisonnement économique, ainsi que notre bon sens, pour nous guider dans le choix des variables. Bien que cette approche ne permette pas de rendre pleinement compte de la finesse de la théorie économique, elle est dans les faits couramment utilisée par des chercheurs dont la rigueur analytique n'est plus à démontrer.

Après la spécification d'un modèle économétrique, comme celui de l'équation (1.3) ou (1.4), il est possible de formuler des *hypothèses* portant sur les paramètres inconnus du modèle. Par exemple, dans l'équation (1.3), nous pourrions faire l'hypothèse que *wage*, le salaire qui peut être touché dans l'emploi légal, n'a pas d'effet sur l'activité criminelle. Dans le contexte de ce modèle économétrique précis, l'hypothèse se traduit par $\beta_1 = 0$.

Par définition, une analyse empirique fait appel à des données. Après avoir collecté les données concernant les variables pertinentes du modèle, plusieurs méthodes économétriques peuvent être utilisées pour estimer les paramètres du modèle et tester formellement les hypothèses qui nous intéressent. Dans certains cas, le modèle économétrique est utilisé pour générer des prévisions auxquelles une théorie ou une politique pourrait conduire.

En raison de l'importance que revêt la collecte de données, la section 1.3 décrit les types de données qui sont fréquemment utilisées dans les travaux empiriques.

1.3 LA STRUCTURE DES DONNÉES ÉCONOMIQUES

Il existe différents types de bases de données économiques. Alors que certaines méthodes économétriques peuvent être directement appliquées à de nombreux types de bases de données, certaines méthodes présentent des particularités dont il faut tenir compte si nous désirons en exploiter le plein potentiel.

Données en coupe transversale

Une base de **données en coupe transversale**[1] est composée d'un échantillon d'individus, ménages, entreprises, villes, États, pays, ou autres unités, observés à un certain moment dans le temps. Il arrive que les données n'aient pas été recueillies exactement au même moment pour l'ensemble des unités d'observation. Par exemple, lors d'une enquête, plusieurs familles peuvent être interrogées au cours de différentes semaines d'une même année. Dans une analyse en coupe transversale pure, on a tendance à ignorer ces petits décalages temporels qui interviennent au moment de la collecte de données. Autrement dit, même si toutes les familles ne sont pas interrogées au cours de la même semaine, la base de données sera malgré tout considérée comme une base de données en coupe transversale.

Une caractéristique importante des données en coupe transversale est la possibilité d'obtenir un échantillonnage aléatoire à partir de la population sous-jacente. Par exemple, si nous pouvons obtenir des informations sur le salaire, le niveau d'études et l'expérience en tirant aléatoirement 500 personnes de la population active, alors nous disposons d'un échantillon aléatoire de cette population. Cette stratégie d'échantillonnage est la plus couramment abordée dans un cours d'introduction à la statistique et son utilisation simplifie l'analyse des données en coupe transversale. L'annexe C propose une révision de l'échantillonnage aléatoire.

Dans certaines circonstances, il n'est pas approprié d'analyser des données en coupe transversale en se reposant sur l'hypothèse d'échantillonnage aléatoire. Par exemple, si nous désirons étudier les facteurs qui déterminent l'accumulation de richesse dans une famille, nous pouvons mener une enquête auprès d'un

1 Les termes « données de coupe transversale » et « données en coupe instantanée » sont équivalentes (note de la traduction.)

échantillon aléatoire mais certaines familles refuseront de divulguer leur patrimoine. Or, si la probabilité de refus est plus élevée pour les familles les plus riches, cet échantillon ne correspondra pas un échantillon aléatoire et ne sera pas représentatif de la population. Ce point illustre le problème de sélection de l'échantillon que nous aborderons plus en détails dans le chapitre 17.

L'hypothèse d'échantillonnage aléatoire est également violée lorsque le nombre d'unités dans l'échantillon est proche de la taille de la population ; c'est souvent le cas pour les unités géographiques. Dans ces cas-là, le problème potentiel est que la taille de la population n'est pas suffisamment grande pour respecter l'hypothèse selon laquelle les observations sont tirées de manière indépendante. Par exemple, si l'on cherche à étudier le développement de nouvelles activités commerciales dans différentes régions en fonction des niveaux de salaire, prix des énergies, impôts sur les entreprises, impôts fonciers, disponibilité des services, qualité de la main d'œuvre, et autres caractéristiques de la région, il est très improbable que les activités commerciales qui se développent dans deux régions voisines soient indépendantes.

Les méthodes économétriques que nous abordons dans cet ouvrage fonctionnent toujours dans ce genre de situations, mais elles doivent parfois être raffinées. Dans la plupart des cas, nous allons ignorer ces complexités et nous analyserons ces situations dans le cadre de l'échantillonnage aléatoire, même s'il n'est pas techniquement rigoureux de procéder ainsi.

Les données en coupe transversale sont largement utilisées en sciences sociales. En économie, l'analyse de données en coupe transversale s'inscrit fortement dans le champ de la microéconomie appliquée, comme en économie du travail, finance publique, économie industrielle, économie urbaine, démographie et économie de la santé. Les données sur les individus, les ménages, les entreprises et les villes, que l'on récolte à un moment donné dans le temps, sont importantes pour tester les hypothèses microéconomiques et pour évaluer des politiques diverses et variées.

Les données en coupe transversale que nous utilisons dans cet ouvrage sont disponibles sous format informatique, ce qui permet de les consulter et de les stocker sur un ordinateur. Le tableau 1.1 contient un extrait d'une base de données composée de 526 personnes en emploi au cours de l'année 1976. (Il s'agit d'un extrait de la base de données WAGE1). Les variables sont *wage* (en dollars par heure), *educ* (nombre d'années d'études), *exper* (nombre d'années d'expérience potentielle sur le marché du travail), *female* (pour indiquer si la personne est une femme), et *married* (statut marital). Ces deux dernières variables sont par nature binaires (zéro-un) et servent à indiquer les caractéristiques qualitatives des individus (la personne est une femme ou non, la personne est mariée ou non). Nous discuterons longuement des variables binaires à partir du chapitre 7.

La variable *obsno* dans le tableau 1.1 représente le numéro d'observation attribué à chaque personne de l'échantillon. Contrairement aux autres variables, il ne s'agit pas d'une caractéristique de l'individu. Tous les logiciels économétriques attribuent un numéro à chaque unité d'observation. En vous fiant à votre intuition, vous devez comprendre que, pour des données comme celles du tableau 1.1, peu importe de savoir quelle personne est étiquetée observation 1, quelle personne est étiquetée observation 2, et ainsi de suite. Le fait que l'ordre des données n'a pas d'importance pour l'analyse économétrique est une caractéristique fondamentale des bases de données obtenues par échantillonnage aléatoire.

Tableau 1.1 *Base de données en coupe transversale indiquant les salaires et d'autres caractéristiques individuelles*

obsno	wage	educ	exper	female	married
1	3,10	11	2	1	0
2	3,24	12	22	1	1

obsno	wage	educ	exper	female	married
3	3,00	11	2	0	0
4	6,00	8	44	0	1
5	5,30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11,56	16	5	0	1
526	3,50	14	5	1	0

Dans les bases de données en coupe transversale, il arrive que certaines variables ne correspondent pas exactement aux mêmes périodes de temps. Par exemple, afin de déterminer les effets de la politique du gouvernement sur la croissance économique de long terme, les économistes ont étudié le lien entre la croissance réelle, mesurée par le produit intérieur brut (PIB) par habitant au cours d'une certaine période (par exemple, de 1960 à 1985), et un ensemble de variables déterminées en partie par la politique du gouvernement (ici, les dépenses publiques en 1960 exprimées en pourcentage du PIB et la proportion d'adultes diplômés du secondaire en 1960). Ce type de base de données se présente sous une forme similaire au tableau 1.2, qui est en fait une partie de la base de données utilisée pour l'étude comparative des taux de croissance entre pays par De Long et Summers (1991).

La variable *gpcrgdp* représente la croissance réelle moyenne du PIB par habitant au cours de la période allant de 1960 à 1985. Le fait que les variables *govcons60* (dépenses publiques exprimées en pourcentage du PIB) et *second60* (pourcentage de la population adulte diplômée du secondaire) correspondent à l'année 1960, alors que *gpcrgdp* est la croissance moyenne au cours de la période 1960-1985, ne pose aucun problème particulier ; nous pouvons les traiter comme des données en coupe transversale. Les observations sont ici ordonnées par pays de manière alphabétique, mais ce rangement n'affecte en rien les analyses qui en découlent.

Tableau 1.2 *Une base de données sur les taux de croissance économique et les caractéristiques du pays.*

obsno	country	gpcrgdp	govcons60	second60
1	Argentina	0,89	9	32
2	Austria	3,32	16	50
3	Belgium	2,56	13	69
4	Bolivia	1,24	18	12
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
61	Zimbabwe	2,30	17	6

Séries chronologiques

Une base de séries chronologiques[2] est composée d'une ou de plusieurs variables observées au cours du temps à plusieurs reprises. Comme exemples de séries chronologiques, on peut citer les prix des actions, l'offre de monnaie, l'indice des prix à la consommation, le produit intérieur brut, le taux d'homicides par an, et le chiffre d'affaire de l'industrie automobile. En sciences sociales, le temps représente une dimension importante du fait que les évènements passés peuvent influencer les évènements à venir et que les comportements ne se modifient pas instantanément.

Une caractéristique fondamentale des séries chronologiques, qui les rendent plus difficiles à analyser que les données en coupe transversale, est que les observations économiques ne sont (presque) jamais indépendantes au cours du temps. Dans la plupart des cas, ces séries chronologiques sont fortement dépendantes de leur passé récent. Par exemple, l'estimation du produit intérieur brut au cours du trimestre précédent nous renseignera plutôt bien sur l'ordre de grandeur du PIB pour le trimestre en cours ; le PIB a en effet tendance à rester relativement stable d'un trimestre à l'autre.

La plupart des procédures économétriques peuvent être utilisées tant sur données en coupe transversale que sur séries chronologiques. Pour justifier l'utilisation des méthodes économétriques standards, il est néanmoins nécessaire de préciser davantage les conditions sous lesquelles les modèles économétriques sur séries chronologiques sont valides. Ces méthodes économétriques ont d'ailleurs fait l'objet de modifications et d'améliorations visant, par exemple, à mieux tenir compte de la dépendance naturelle ou de la tendance temporelle présente dans les séries chronologiques en économie.

Une autre caractéristique des séries chronologiques nécessite une attention particulière : il s'agit de la **fréquence** à laquelle les données sont collectées. En économie, les fréquences les plus courantes sont la journée, la semaine, le mois, le trimestre et l'année. Les prix des actions sont souvent enregistrés à une journée d'intervalle (en excluant les jours fériés, les samedis et les dimanches). L'offre de monnaie de l'économie américaine est enregistrée toutes les semaines. De nombreuses séries macroéconomiques sont annoncées une fois par mois, notamment l'inflation et le taux de chômage. D'autres séries macroéconomiques sont enregistrées de façon moins fréquente, par exemple chaque trimestre. Le produit intérieur brut est un exemple bien connu de série trimestrielle. D'autres séries chronologiques, comme le taux de mortalité infantile par État aux États-Unis, ne sont disponibles que sur base annuelle.

De nombreuses séries chronologiques, observées sur base hebdomadaire, mensuelle ou trimestrielle, font état d'une forte saisonnalité dont il faut tenir compte dans l'analyse de séries chronologiques. Par exemple, les variations mensuelles observées pour les constructions de logement s'explique d'abord par les changements de conditions météorologiques. Nous apprendrons à régler le problème de la saisonnalité dans le chapitre 10.

Le tableau 1.3 présente une base de séries chronologiques que Castillo-Freeman et Freeman (1992) utilisent pour analyser les effets du salaire minimum au Puerto Rico. Dans cette base de données, la première observation correspond à l'année disponible la plus ancienne ; la dernière observation correspond à l'année disponible la plus récente. Quand les méthodes économétriques sont utilisées pour analyser des séries chronologiques, il est conseillé de conserver les données dans l'ordre chronologique.

2 Les termes « données chronologiques », « données temporelles », « séries temporelles » sont interchangeables (note de la traduction.)

Tableau 1.3 *Salaire minimum, chômage et données associées pour le Puerto Rico*

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0,20	20,1	15,4	878,7
2	1951	0,21	20,7	16,0	925,0
3	1952	0,23	22,6	14,8	1 015,9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3,35	58,1	18,9	4 281,6
38	1987	3,35	58,2	16,8	4 496,7

La variable *avgmin* fait référence au salaire minimum moyen annuel ; *avgcov* est le taux de couverture moyen (c'est-à-dire le pourcentage de salariés couverts par la loi sur le salaire minimum) ; *prunemp* est le taux de chômage ; et *prgnp* est le produit intérieur brut, en millions de dollars (exprimé en dollars de 1954). Dans les chapitres consacrés à l'étude des séries chronologiques, nous analyserons ces données plus en détails afin de mesurer l'effet du salaire minimum sur l'emploi.

Données empilées

Certaines bases de données ont à la fois des caractéristiques propres aux coupes transversales et aux séries chronologiques. Supposons par exemple que l'on mène aux États-Unis deux enquêtes sur les ménages, l'une en 1985 et l'autre en 1990. En 1985, nous tirons aléatoirement un échantillon de ménages à partir desquels nous obtenons des informations sur le revenu, l'épargne, la taille de la famille, etc. En 1990, un *nouvel* échantillon de ménages est tiré aléatoirement ; l'enquête est similaire et permet de récolter le même type de données. Afin d'accroître la taille de notre échantillon, on peut combiner les deux années pour construire des **données empilées**[3].

Empiler des coupes transversales pour différentes années est souvent efficace lorsqu'il s'agit d'analyser les effets d'une nouvelle politique menée par les pouvoirs publics. Le principe de base consiste à recueillir des données au cours des années qui précèdent et suivent un changement de politique majeur. Considérons par exemple une base de données sur les prix de biens immobiliers observés en 1993 et en 1995, juste avant et après la décision de diminuer les impôts fonciers en 1994. Supposons que nous ayons des informations sur 250 maisons en 1993 et sur 270 maisons en 1995. Le tableau 1.4 présente une façon de construire ce type de base de données.

Les observations numérotées 1 à 250 correspondent aux maisons vendues en 1993 ; les observations numérotées 251 à 520 correspondent aux 270 maisons vendues en 1995. Même si l'ordre dans lequel les données sont conservées ne s'avère pas crucial, indiquer l'année d'observation est en général très important. C'est précisément la raison pour laquelle la variable *year* est incluse dans la base de données.

3 On rencontre parfois les termes « coupes transversales empilées », « coupes transversales regroupées » ou « coupes transversales agrégées » (note de la traduction).

Tableau 1.4 *Données empilées : les prix de l'immobilier pour deux années*

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85 500	42	1 600	3	2,0
2	1993	67 300	36	1 440	3	2,5
3	1993	134 000	38	2 000	4	2,5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243 600	41	2 600	4	3,0
251	1995	65 000	16	1 250	2	1,0
252	1995	182 400	20	2 200	4	2,0
253	1995	97 500	15	1 540	3	2,0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57 200	16	1 100	2	1,5

Les données empilées sont plus ou moins analysées de la même façon que les données en coupe transversale classiques, à cette différence près que l'évolution des variables au cours du temps est un objectif explicite de l'analyse sur données empilées. L'utilisation de données empilées permet d'augmenter la taille de l'échantillon et surtout d'étudier l'évolution de la relation d'intérêt au cours du temps.

Données de panel

Une base de **données de panel** (ou *données longitudinales*) contient des séries chronologiques pour *chacune des unités* reprises dans la coupe transversale. Par exemple, une telle base de données vous permet d'observer le salaire, le niveau d'étude et l'expérience professionnelle d'un ensemble d'individus que l'on suit au cours du temps, sur une période de dix ans. Il est également possible de recueillir des informations sur la structure financière et les investissements pour un même groupe d'entreprises pendant 5 ans. Les données en panel peuvent aussi concerner des unités géographiques. Par exemple, considérant un ensemble fixe de comtés aux États-Unis, nous pouvons obtenir, pour les années 1980, 1985 et 1990, des données sur les flux d'immigration, les taux d'imposition, les taux de salaire, les dépenses publiques, etc.

La caractéristique fondamentale des données de panel, qui les distingue de simples données empilées, est que les unités que nous suivons au cours du temps restent *les mêmes*. Dans les exemples précédents, cela signifie que les différentes coupes transversales contiennent les mêmes individus, entreprises ou comtés. Les données du tableau 1.4 ne sont pas considérées comme des données de panel parce que les maisons vendues en 1993 ne sont pas forcément les mêmes que celles vendues en 1995 ; si certaines maisons peuvent apparaître à deux reprises, cela relève plus de l'exception que de la règle et le nombre de cas est souvent

négligeable. En revanche, dans le tableau 1.5, nous avons des données de panel concernant un échantillon fixe de 150 villes aux États-Unis, dont on observe notamment le taux de la criminalité à deux moments dans le temps, en 1986 et 1990.

Le tableau 1.5 présente quelques caractéristiques intéressantes. Tout d'abord, un numéro a été attribué à chaque ville, ce numéro allant de 1 à 150. Il n'est pas nécessaire de savoir quelle ville correspond à ville 1, ville 2, etc. Dans une base de données de panel, l'ordre au sein de la coupe transversale n'a aucune importance, comme c'est également le cas au sein d'une coupe transversale pure. On pourrait éventuellement utiliser le nom de la ville au lieu du numéro ; en réalité, il est souvent utile d'avoir les deux.

Tableau 1.5 *Une base de données de panel sur la criminalité urbaine*

obsno	city	year	murders	population	unem	police
1	1	1986	5	350 000	8,7	440
2	1	1990	8	359 200	7,2	471
3	2	1986	2	64 300	5,4	75
4	2	1990	1	65 100	5,5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260 700	9,6	286
298	149	1990	6	245 000	9,8	334
299	150	1986	25	543 000	4,3	520
300	150	1990	32	546 200	5,2	493

Une autre caractéristique est que les deux années pour la ville 1 occupent les deux premières lignes. Les observations 3 et 4 correspondent à la ville 2, et ainsi de suite. Pour chacune des 150 villes, on observe deux lignes de données ; tous les logiciels économétriques identifieront 300 observations. Le traitement informatique des données de panel n'est d'ailleurs pas différent de celui des données empilées, en gardant naturellement à l'esprit que les mêmes villes apparaissent chaque année dans un panel. Comme nous le verrons aux chapitres 13 et 14, la structure d'un panel permet d'étudier des problématiques que nous ne pouvons pas aborder en utilisant de simples données empilées.

Pour ranger les observations du tableau 1.5, nous plaçons, pour chaque ville, les deux années de données l'une à côté de l'autre, avec la première année placée avant la seconde. C'est la façon la plus pratique d'ordonner les données de panel. Comparez cet agencement à celui du tableau 1.4 pour les données empilées. Pour faire bref, les données en panel sont rangées de cette manière pour faciliter la transformation des données qui intervient par la suite.

Les bases de données en panel sont plus difficiles à obtenir que les bases de données empilées, en particulier lorsqu'elles concernent des individus, des ménages ou des entreprises. La constitution d'un panel exige en effet le suivi des mêmes unités au cours du temps. Évidemment, observer les mêmes unités au

cours du temps présente des avantages que les données en coupe transversale ou les données empilées n'ont pas. L'avantage de disposer de plusieurs observations pour les mêmes unités est de pouvoir tenir compte de l'influence de certaines caractéristiques non observées des individus, des entreprises, etc. Nous aurons l'occasion de le souligner à plusieurs reprises dans le reste de l'ouvrage. À ce stade, il suffit de préciser qu'il est très difficile d'inférer une relation de causalité entre les variables sans cet avantage, lorsqu'une seule coupe transversale est disponible par exemple. L'autre avantage d'un panel est qu'il nous permet d'étudier l'importance des décalages de comportements dans le temps et de mieux évaluer le résultat d'un processus décisionnel. Ce genre d'information est important car l'impact de nombreuses politiques économiques ne se fait sentir qu'après un laps de temps.

Au niveau de la licence universitaire[4], la plupart des livres n'abordent pas les méthodes économétriques pour données de panel. Les économistes reconnaissent néanmoins qu'il est aujourd'hui difficile, voire impossible, de répondre de manière satisfaisante à certaines questions sans recourir à de telles données. Vous constaterez par la suite qu'une simple analyse en panel permet de réaliser de grandes avancées, sans qu'il soit nécessaire de recourir à des méthodes plus compliquées que celles utilisées pour des données en coupe transversale.

Remarque sur la structure des données

La partie 1 de ce livre est consacrée à l'analyse des données en coupe transversale, dont les difficultés conceptuelles et techniques sont les moins nombreuses. Dans cette partie, nous aurons l'occasion d'illustrer tous les thèmes fondamentaux de l'analyse économétrique. Les méthodes et les enseignements de l'analyse en coupe transversale nous servirons dans les autres parties de l'ouvrage.

Bien que les analyses économétriques des coupes transversales et des séries chronologiques partagent de nombreux outils, le traitement des séries chronologiques en économie est plus complexe, en raison de la tendance temporelle et de la forte persistance qu'elles affichent souvent. Il est d'ailleurs admis aujourd'hui que de nombreux exemples qui ont servi à illustrer l'application des méthodes économétriques aux séries chronologiques, présentent de sérieuses lacunes. Cela n'aurait aucun sens de recourir à de tels exemples, en particulier au début de l'ouvrage ; cela ne servirait qu'à renforcer des pratiques économétriques douteuses. L'analyse économétrique des séries chronologiques fera donc l'objet de la deuxième partie de l'ouvrage ; nous y aborderons des questions importantes, comme celles liées à la tendance temporelle, la persistance, la dynamique ou la saisonnalité.

Dans la partie 3, nous traiterons explicitement des données empilées et des données de panel. L'analyse des données empilées de manière indépendante et des données de panel simple sont des extensions assez naturelles de l'analyse en coupe transversale pure. Il faudra attendre le chapitre 13 avant de nous consacrer à l'étude de ces sujets.

1.4 LA CAUSALITÉ ET LA SIGNIFICATION DE *CETERIS PARIBUS* DANS L'ANALYSE ÉCONOMÉTRIQUE

Lorsqu'il s'agit de tester des théories économiques ou d'évaluer des politiques publiques, l'objectif ultime de l'économiste est de déterminer si une variable, comme le niveau d'études, a un **effet causal** sur une autre variable, comme la productivité du salarié. L'identification d'un lien de dépendance entre ces variables peut

4 En France, la licence correspond au « Bachelor's degree » américain : elle couvre les trois premières années d'études à l'université. En Belgique, le terme de « licence » a fait place à celui de « baccalauréat universitaire » dont la durée est de trois ans également.

donner une indication de leur lien causal, mais cette indication n'en reste pas moins très vague et rarement convaincante (à moins que la causalité puisse être établie par ailleurs).

La notion de ***ceteris paribus*** – qui signifie « toutes choses (pertinentes étant) égales par ailleurs » – joue un rôle important dans l'analyse causale. Sans l'avoir explicitement indiqué jusqu'ici, cette notion était implicitement admise dans les explications que nous avons apportées précédemment, notamment pour les exemples 1.1 et 1.2.

Dans les cours d'introduction à l'économie, la plupart des questions économiques reposent sur le raisonnement *ceteris paribus*. Par exemple, lorsque nous analysons la demande du consommateur, nous cherchons à évaluer l'effet d'une variation du prix d'un bien sur la quantité demandée, tout en gardant tous les autres facteurs constants (comme le revenu, les prix des autres biens, et les préférences individuelles). Si les autres facteurs ne sont pas gardés constants[5], on ne peut pas connaître l'effet causal d'une variation du prix sur la quantité demandée.

Garder les autres facteurs constants est également essentiel pour l'analyse de politiques diverses et variées. Dans l'exemple de la formation professionnelle (exemple 1.2), il pourrait être intéressant de connaître l'effet d'une semaine de formation professionnelle supplémentaire sur les salaires, en gardant toutes les autres composantes inchangées (en particulier le niveau d'études et l'expérience). Si nous arrivons à garder égaux tous les autres facteurs pertinents et que nous trouvons un lien entre la formation professionnelle et les salaires, nous pouvons conclure que cette formation professionnelle a un effet causal sur la productivité du travailleur. Bien que cela puisse paraître simple à réaliser, il est important de souligner, même à ce niveau peu avancé de l'analyse, que nous ne parviendrons pas à garder littéralement *toutes* les choses égales par ailleurs, sauf dans certains cas bien particuliers. Dans la plupart des études empiriques, la question centrale sera plutôt la suivante : avons-nous tenu compte de l'influence de *suffisamment* de facteurs pour bien mesurer le lien de causalité entre les deux variables qui nous intéressent plus particulièrement ? Rares sont les études économétriques dont la qualité ne dépend pas de la réponse apportée à cette question centrale.

Dans la plupart des applications intéressantes en économie, le nombre de facteurs qui peuvent influencer la variable d'intérêt, comme l'activité criminelle ou le salaire, est si élevé que la tentative d'isoler l'effet d'une variable en particulier semble vouée à l'échec. Pourtant, nous verrons que les méthodes économétriques permettent de simuler une expérimentation *ceteris paribus* lorsqu'elles sont appliquées avec soin.

À ce stade de l'analyse, nous ne disposons pas encore de suffisamment d'outils pour aborder les méthodes économétriques dont nous avons besoin pour estimer les effets *ceteris paribus* ; nous allons néanmoins considérer quelques problèmes classiques qui se posent en économie dans le domaine de l'analyse causale. Nous n'allons pas utiliser d'équation dans cette discussion. Pour chaque exemple, nous allons montrer que le problème de l'inférence causale est résolu à chaque fois que nous pouvons mener à bien l'expérimentation qui convient. Il sera utile d'en décrire la marche à suivre, tout en admettant que la récolte de données expérimentales est impossible dans la plupart des cas. Il sera également important de comprendre pourquoi les données expérimentales jouissent de caractéristiques désirables que n'ont pas les données disponibles dans la réalité.

Nous comptons maintenant sur votre compréhension intuitive de plusieurs termes (comme *aléatoire*, *indépendance* et *corrélation*) ; ces termes vous sont familiers si vous avez suivi un cours de statistique ou d'introduction aux probabilités. (Ces concepts sont présentés dans l'annexe B.) Nous commençons par un exemple qui illustre la discussion que nous venons de tenir.

5 Les expressions « demeurer constant », « garder constant », « tenir compte de l'influence de », « corriger l'influence de » seront utilisées de manière interchangeable (note de la traduction).

EXEMPLE 1.3
Rendements des terres agricoles et engrais

Il faut savoir que l'effet de l'utilisation de nouveaux engrais sur les rendements agricoles a fait l'objet d'études économétriques pionnières [par exemple, Griliches (1957)]. Dans cet exemple, nous allons nous intéresser à la culture du soja. Notons d'abord que la quantité d'engrais utilisée n'est pas le seul facteur qui détermine le rendement d'une culture ; d'autres facteurs jouent, parmi lesquels les précipitations, la qualité de la terre, ou la présence de parasites. Il est donc impératif de recourir à une analyse *ceteris paribus*. Une façon de déterminer l'effet causal de l'utilisation d'engrais sur le rendement des cultures de soja est de mener une expérimentation qui pourrait inclure les étapes suivantes : choisir plusieurs terrains d'un demi-hectare ; appliquer différentes quantités d'engrais sur chaque terrain ; et mesurer les rendements. Ces trois étapes aboutissent à la constitution d'une base de données en coupe transversale. La dernière étape consiste à utiliser les méthodes statistiques, qui seront introduites dans le chapitre 2, pour mesurer le lien entre les rendements et les quantités d'engrais.

Comme nous l'avons expliqué précédemment, cette expérimentation ne suit pas une démarche très rigoureuse : nous n'avons pas précisé qu'il fallait choisir des terrains identiques en tous points, à l'exception de la quantité d'engrais qui y est répandue. En pratique, choisir des terrains en *tous* points identiques est impossible : certaines caractéristiques, comme la qualité de la terre, ne sont d'ailleurs pas parfaitement observables. Dans ce cas, comment peut-on savoir si les résultats de cette expérimentation peuvent être utilisés pour mesurer l'effet *ceteris paribus* de l'utilisation d'engrais ? La réponse à cette question dépend de la façon précise dont les quantités d'engrais ont été choisies. Si les niveaux d'engrais répandus sur les terrains ont été déterminés *indépendamment* des autres caractéristiques du terrain qui affectent le rendement (ce qui implique, par exemple, que la qualité de la terre n'a pas été prise en compte au moment de déterminer les quantités d'engrais), alors le tour est joué.

L'exemple suivant est sans doute plus représentatif des difficultés auxquelles nous sommes confrontés lorsqu'il s'agit d'inférer un lien de causalité en économie appliquée.

EXEMPLE 1.4
Rendement de l'éducation

Les économistes du travail et les responsables politiques se sont intéressés depuis longtemps à la question du « rendement de l'éducation » (provenant de l'expression anglophone « return to education »). De façon quelque peu informelle, la question peut se formuler comme suit : si on choisit un individu dans la population et qu'on lui attribue une année d'étude supplémentaire, dans quelle mesure son salaire va-t-il augmenter ? Comme pour les exemples précédents, il s'agit d'une analyse *ceteris paribus*, qui implique que tous les autres facteurs doivent être maintenus constants au moment où l'individu bénéficie d'une année d'étude supplémentaire.

De la même manière qu'un chercheur en agronomie peut mettre au point une expérimentation pour estimer l'effet de l'utilisation d'engrais sur les cultures agricoles, nous pouvons imaginer un planificateur social désireux de mettre au point une expérimentation visant à mesurer le rendement de l'éducation. Supposons pour le moment que le responsable politique ait la possibilité d'assigner n'importe quel niveau d'étude à n'importe quelle personne. Comment ce planificateur peut-il réussir son expérimentation aussi bien que dans l'exemple 1.3 ? Le planificateur devra choisir un groupe de personnes et assigner aléatoirement un niveau d'étude à chaque personne du groupe ; on attribuerait un niveau de fin de collège (secondaire inférieur) à certains, un niveau de fin de lycée (secondaire supérieur) à d'autres, un niveau de licence à d'autres encore, et ainsi de suite. Le planificateur devra ensuite mesurer les salaires pour tous ces individus (en supposant qu'ils aient un emploi). Les individus sont en quelque sorte comparables à des terrains agricoles ; le niveau d'étude joue le rôle de l'engrais et le taux de salaire celui du rendement en soja. Comme dans l'exemple 1.3, si les niveaux d'éducation sont assignés de manière indépendante des autres facteurs qui affectent la productivité (comme l'expérience et les aptitudes innées), alors cette analyse, qui ignore toutes les autres caractéristiques des individus, produira des résultats utiles. À ce stade, cela exige de notre part une certaine ouverture d'esprit. Nous prendrons soin de justifier davantage cette assertion dans le chapitre 2.

Contrairement à l'exemple sur l'utilisation d'engrais, l'expérimentation décrite dans l'exemple 1.4 est impossible à réaliser. Sur le plan éthique, des questions évidentes se posent quant à la manière d'attribuer de façon aléatoire un niveau d'études à un groupe d'individus, sans parler des coûts économiques qu'une telle expérimentation peut générer. Enfin, comment pourrait-on attribuer un niveau de fin de primaire à quelqu'un qui possède déjà un diplôme universitaire ?

Même s'il est impossible d'obtenir des données expérimentales pour mesurer le rendement de l'éducation, nous pouvons toujours recueillir des données non expérimentales sur le niveau d'études et le salaire. Il suffit de constituer un échantillon aléatoire suffisamment représentatif de la population des personnes en emploi. De telles données sont d'ailleurs disponibles dans les enquêtes utilisées en économie du travail. Ces bases de données présentent néanmoins une caractéristique qui rend plus difficile l'estimation de l'effet *ceteris paribus* du niveau d'études sur le salaire. Les gens *choisissent* leur niveau d'études. Par conséquent, le niveau d'études n'est certainement pas déterminé indépendamment des autres facteurs qui affectent le salaire. Ce problème est une caractéristique propre à la plupart des bases de données non expérimentales.

L'expérience professionnelle est un facteur qui peut avoir un effet sur le salaire. Or, ceux qui ont plus d'éducation ont souvent moins d'expérience : en général, un individu doit remettre à plus tard son entrée sur le marché du travail s'il désire poursuivre ses études. Dans une base de données non expérimentales, il est donc probable que le niveau d'études soit négativement corrélé à une autre variable clé qui affecte également le salaire. Par ailleurs, il est communément admis que les personnes ayant de meilleures aptitudes innées choisissent un niveau d'études plus élevé. Comme de meilleures aptitudes innées sont liées à des salaires plus élevés, nous avons un autre exemple de corrélation entre le niveau d'études et un facteur crucial qui affecte le salaire.

Les facteurs omis dans l'exemple sur les salaires, comme l'expérience et les aptitudes innées, ont leurs analogues dans l'exemple sur les engrais. L'expérience est en général facile à mesurer ; elle peut être assimilée à une variable comme les précipitations. Par contre, les aptitudes innées correspondent à un concept vague et difficile à quantifier ; elles sont similaires à la qualité de la terre dans l'exemple sur les engrais. Nous verrons tout au long de l'ouvrage que la prise en compte d'autres facteurs observés, comme l'expérience, ne pose pas de problème majeur lorsqu'il s'agit de déterminer l'effet *ceteris paribus* d'une variable comme le niveau d'étude. Par contre, nous constaterons qu'il est beaucoup plus difficile de prendre en compte l'effet de facteurs non observables, comme les aptitudes innées. De nombreux travaux récents en économétrie ont été motivés par le défi que représentent les facteurs non observés dans les modèles économétriques.

Nous pouvons établir un dernier parallèle entre les exemples 1.3 et 1.4. Supposons que les quantités d'engrais ne soient pas définies de manière complètement aléatoire. Par exemple, un chercheur en agronomie peut penser qu'il est préférable de répandre plus d'engrais sur les terrains de meilleure qualité. (Ce chercheur doit avoir une certaine idée des terrains dont la qualité est supérieure, même s'il n'est pas capable d'en quantifier parfaitement les différences.) Cette situation est parfaitement analogue au lien qui existe entre niveau d'études et aptitudes innées dans l'exemple 1.4. Puisque les meilleures terres ont de meilleurs rendements et qu'une plus grande quantité d'engrais est répandue sur les meilleures terres, le lien observé entre rendement et engrais peut être fallacieux.

L'utilisation de données plus « agrégées » (relatives à des villes plutôt qu'à des individus, par exemple) rend plus difficile l'inférence du lien de causalité entre les variables. Nous l'expliquons dans l'exemple 1.5.

EXEMPLE 1.5
Maintien de l'ordre et criminalité urbaine

Il existe, et existera sans doute toujours, un vif débat autour des stratégies qu'il faudrait mettre en place pour lutter efficacement contre la criminalité urbaine. À cet égard, la question suivante revêt une importance particulière : la présence d'un plus grand nombre d'agents de police dans les rues diminue-t-elle la criminalité urbaine ?

Dans le cadre de l'analyse *ceteris paribus*, la question n'est pas difficile à reformuler : si une ville est choisie de manière aléatoire et qu'un plus grand nombre de policiers lui est attribué (disons dix agents de police supplémentaires), à quelle diminution du taux de criminalité peut-on s'attendre ? Une autre façon de formuler la question est : si deux villes sont identiques à tous égards, à la seule différence près que la ville A dispose de dix agents de police de plus que la ville B, quelle sera la différence entre les taux de criminalité enregistrés dans les deux villes ?

Dans la réalité, il est quasiment impossible de trouver deux villes identiques à tous égards, hormis au niveau de leurs effectifs de police. L'analyse économétrique n'a heureusement pas besoin de cela. Il faut néanmoins déterminer s'il est possible de recueillir des données expérimentales sur la criminalité urbaine et sur les effectifs de police. On peut imaginer une véritable expérimentation à laquelle participerait un grand nombre de villes auxquelles on attribuerait à l'avance, de manière aléatoire, un nombre d'agents de police qui seraient mis à leur disposition l'année suivante.

On peut cependant difficilement imaginer qu'une ville accepte de se voir imposer des effectifs de police qui ne lui conviennent pas. Par ailleurs, si le nombre d'agents de police est déterminé par le pouvoir politique local en fonction d'autres facteurs qui déterminent le taux de criminalité (comme le taux de pauvreté ou le niveau d'études), alors les données doivent être considérées comme non expérimentales. Une autre façon de voir ce problème est de reconnaître que les effectifs de police et le taux de criminalité sont *déterminés simultanément*. On tiendra explicitement compte de cette simultanéité dans le chapitre 16.

EXEMPLE 1.6
Salaire minimum et chômage

L'effet du salaire minimum sur le taux de chômage constitue une question politique importante et fait souvent l'objet de controverse. Différents types de données (coupes transversales, séries chronologiques, panel, etc.) peuvent servir à estimer cette relation. Les séries chronologiques sont souvent utilisées pour en étudier les effets agrégés. Le tableau 1.3 en donne un bel exemple.

Si le salaire minimum se fixe à un niveau plus élevé que le salaire d'équilibre du marché, l'analyse standard en termes d'offre et de demande implique que l'emploi total diminue (car l'offre de travail excède la demande de travail). Pour quantifier cet effet, nous pouvons étudier le lien entre l'emploi et le salaire minimum au cours du temps. Déterminer la causalité entre ces deux variables n'est pas facile et cela ne provient pas uniquement des difficultés rencontrées lors de l'utilisation de séries chronologiques. En réalité, le salaire minimum n'est pas déterminé dans un vide expérimental. Il dépend des forces économiques et politiques en vigueur au moment de sa détermination. (Une fois déterminé, le salaire minimum est en place pour plusieurs années, sauf s'il est indexé sur l'inflation.) Il est donc probable que le salaire minimum soit relié à d'autres facteurs qui influencent également le niveau d'emploi.

Imaginons que le gouvernement désire conduire une expérimentation pour déterminer les effets du salaire minimum sur l'emploi (plutôt que de se préoccuper du bien-être des travailleurs à bas salaire). Le gouvernement pourrait décider de fixer aléatoirement le salaire minimum chaque année, les valeurs de différents indicateurs sur l'emploi pourraient être consignés et les séries chronologiques expérimentales qui en résulteraient pourraient faire l'objet d'une analyse économétrique assez simple. Ce scénario n'a pas grand-chose à voir avec la façon dont le salaire minimum est déterminé en réalité.

Si nous parvenons à tenir compte de l'influence de suffisamment de facteurs sur l'emploi, nous pouvons encore espérer estimer l'effet *ceteris paribus* du salaire minimum sur l'emploi. En ce sens, le problème est très similaire aux exemples précédents en coupe transversale.

Les exemples 1.3, 1.4 et 1.5 auxquels nous venons de recourir se basent sur des données en coupe transversale, à différents niveaux d'agrégation (de l'individu à la ville, par exemple). Nous rencontrons les mêmes obstacles lorsqu'il s'agit d'inférer des liens de causalité à partir de séries chronologiques, comme l'illustre l'exemple 1.6.

Même quand les théories économiques ne nous renseignent pas sur la causalité attendue, elles offrent souvent des prédictions qui peuvent être testées sur le plan économétrique. L'exemple suivant illustre cette approche.

EXEMPLE 1.7
Théorie des anticipations

Selon la théorie des anticipations en économie financière, les taux d'intérêt *espérés* de deux obligations de maturités différentes doivent être identiques, étant donné toute l'information disponible au moment d'investir. Considérons les deux stratégies d'investissement suivantes : (1) acheter un bon du trésor à trois mois, ayant un prix actuel inférieur à 10 000 €, pour recevoir la valeur faciale de 10 000 € dans trois mois ; (2) acheter un bon du trésor à six mois, ayant un prix inférieur à 10 000 €, pour le revendre, dans trois mois, au prix du bon du trésor dont la maturité est égale à trois mois. Chaque stratégie nécessite à peu près le même montant de capital au départ, mais il y a une différence importante. Pour la première stratégie, le rendement exact est connu au moment de l'achat, puisque le prix initial et la valeur faciale du bon du trésor à trois mois sont connus. Ce n'est pas vrai pour la deuxième stratégie : même si on connaît le prix des bons du trésor à six mois au moment de l'achat, on ignore le prix auquel on pourra le revendre dans trois mois. Par conséquent, le second investissement est incertain pour quelqu'un dont l'horizon d'investissement est plus court que six mois.

La théorie des anticipations prédit pourtant que le rendement espéré de ces deux investissements sera identique car les investisseurs les percevront comme des substituts parfaits. En réalité, les rendements effectifs de ces deux investissements seront différents la plupart du temps. Cette théorie se révèle assez facile à tester, comme nous le verrons dans le chapitre 11.

RÉSUMÉ

Dans ce chapitre introductif, nous avons défini le but et le cadre de l'analyse économétrique. L'économétrie est utilisée dans tous les champs de l'économie appliquée. Elle sert à tester des théories économiques. Elle permet d'informer les gouvernements et les organismes privés qui désirent mettre en place ou évaluer des politiques. Elle est aussi utilisée pour produire des prévisions économiques. Parfois, le modèle économétrique est dérivé d'un modèle économique formel ; dans d'autres cas, les modèles économétriques se basent sur l'intuition ou sur des raisonnements économiques plus informels. Toute analyse économétrique conduit à l'estimation des paramètres du modèle et à la réalisation de tests d'hypothèses sur ces paramètres. La validité d'une théorie économique ou l'effet d'une politique est déterminé en fonction de la valeur et du signe que les paramètres du modèle prennent.

Les types de données les plus couramment utilisés en économétrie appliquée sont les données en coupe transversale, les séries chronologiques, les données empilées et les données de panel. Les bases de données qui incluent une dimension temporelle, comme les séries chronologiques et les données de panel, demandent un traitement particulier en raison de la corrélation qui existe au cours du temps entre la quasi-totalité des observations en économie. D'autres problèmes, comme la saisonnalité ou la présence d'une tendance, sont propres aux séries chronologiques et n'affectent pas les données en coupe transversale.

Dans la section 1.4, nous avons expliqué la signification de *ceteris paribus* et souligné la difficulté de mener une analyse de causalité. Dans la plupart des cas, les hypothèses en sciences sociales reposent, par essence, sur la notion de *ceteris paribus* : tous les autres facteurs doivent être gardés constants lorsqu'il s'agit d'étudier le lien entre deux variables. Comme les données en sciences sociales ne sont généralement pas issues d'expérimentation, la mise à jour de liens de causalité représente souvent un véritable défi.

MOTS-CLÉS

Analyse empirique p. 23
Ceteris paribus p. 34
Données empilées (ou coupes transversales empilées) p. 30
Données en coupe transversale p. 26
Données expérimentales p. 22
Données non expérimentales p. 22
Données observationnelles (ou données non expérimentales) p. 22
Données de panel (ou données longitudinales) p. 31
Échantillonnage aléatoire p. 26
Effet causal p. 33
Fréquence des données p. 29
Modèle économétrique p. 25
Modèle économique p. 25
Séries chronologiques (ou série temporelle) p. 29

EXERCICES

1. On vous demande de mener une étude dont l'objectif est de déterminer si la réduction de la taille des classes dans les écoles élémentaires permet d'améliorer la performance des élèves de quatrième année[6].

i. Si vous en aviez la liberté et la possibilité, quelle expérimentation souhaiteriez-vous mener ? Soyez précis.

ii. Supposez plus raisonnablement que vous recueilliez des données observées sur plusieurs milliers d'élèves de quatrième année dans les écoles élémentaires d'une région donnée. Vous pouvez obtenir le nombre d'élèves dans leur classe de quatrième année d'école élémentaire ainsi que les résultats obtenus à un examen standardisé passé à la fin de la même année. Pourquoi devriez-vous vous attendre à trouver une corrélation négative entre le nombre d'élèves par classe et le résultat à l'examen ?

iii. Une corrélation négative montre-t-elle qu'un nombre réduit d'élèves par classe conduit à de meilleures performances ?

2. On justifie souvent les programmes de formation professionnelle par le fait qu'ils améliorent la productivité des salariés. On vous demande précisément de tester cette relation. Cependant, au lieu d'avoir des données pour chaque salarié, vous avez accès à des données pour des entreprises localisées dans l'Ohio. Pour chaque entreprise, vous disposez d'informations sur le nombre d'heures de formation professionnelle par salarié (*training*) et le nombre de produits non défectueux par salarié et par heure (*output*).

i. Décrivez de manière minutieuse l'expérimentation *ceteris paribus* imaginaire que vous auriez à mener.

ii. Pensez-vous que la décision d'une entreprise de former ses employés est indépendante de leurs caractéristiques ? Quelles pourraient être ces caractéristiques, mesurables et non mesurables ?

iii. Citez un facteur qui n'est pas une caractéristique du salarié mais qui peut en affecter la productivité.

iv. Si vous trouvez une corrélation positive entre *output* et *training*, aurez-vous montré de façon convaincante que la formation professionnelle rend les travailleurs plus productifs ? Expliquez.

6 Équivalent, en France, du CM1. (note de la traduction)

3. On vous demande de déterminer quelle relation existe, dans votre université, entre le nombre d'heures par semaine consacrées à l'étude des cours (*study*) et le nombre d'heures par semaine passées à travailler pour un employeur (*work*). Est-il sensé de chercher à déterminer le sens de la causalité entre *study* et *work* ? Expliquez.

4. Les États (et les provinces) qui ont le contrôle de leur fiscalité peuvent parfois réduire les impôts pour tenter de stimuler la croissance économique. Supposez que vous êtes embauché par un État pour évaluer l'effet des taux d'imposition des sociétés sur, par exemple, la croissance du produit intérieur brut (PIB) par habitant de cet État.

i. De quel type de données auriez-vous besoin pour effectuer une analyse statistique ?

ii. Est-il possible de faire une expérience contrôlée ? Que faudrait-il faire ?

iii. Une analyse de corrélation entre la croissance du PIB de cet État et les taux d'imposition est-elle susceptible d'être convaincante ? Expliquez.

EXERCICES SUR ORDINATEUR

C1. Utilisez la base de données WAGE1 pour cet exercice.

i. Calculez le niveau d'étude moyen dans l'échantillon. Quel est le minimum et le maximum du nombre d'années d'études dans cet échantillon ?

ii. Calculez le salaire horaire moyen dans l'échantillon. Vous paraît-il faible ou élevé ?

iii. Le salaire est exprimé en dollars de 1976. Sur base du dernier « Rapport Économique du Président » (*Economic Report of the President*), relevez l'indice des prix à la consommation (*Consumer Price Index, CPI*), pour la période de 1976 à 2013.

iv. Utilisez les valeurs du *CPI* de la question (iii) pour calculer le salaire horaire moyen en dollars de 2013. Le salaire horaire moyen vous semble-t-il plus raisonnable ?

v. Combien de femmes sont observées dans l'échantillon ? Combien d'hommes ?

C2. Utilisez la base de données BWGHT pour répondre à ces questions.

i. Combien de femmes sont observées dans l'échantillon ? Combien de femmes déclarent fumer au cours de la grossesse ?

ii. Quel est le nombre moyen de cigarettes fumées par jour ? S'agit-il d'une bonne mesure du comportement de la femme « typique » dans l'échantillon ? Expliquez.

iii. Parmi les femmes qui fument au cours de la grossesse, quel est le nombre moyen de cigarettes fumées par jour ? Ce nombre est-il comparable à votre réponse à la question (ii) ? Pourquoi ?

iv. Calculez la moyenne de *fatheduc* (nombre d'années d'études du père) dans l'échantillon. Pourquoi la moyenne est-elle calculée sur 1 192 observations seulement ?

v. Calculez le revenu de la famille moyen ainsi que son écart-type.

C3. Les données contenues dans MEAP01 concernent l'état du Michigan en 2001. Utilisez ces données pour répondre aux questions suivantes.

i. Trouvez les plus grandes et les plus petites valeurs de la variable *math4* (résultat à l'examen de mathématique en fin de quatrième année d'école élémentaire aux États-Unis). L'amplitude a-t-elle du sens ? Expliquez.

ii. Combien d'écoles ont un taux de réussite de 100 % à l'examen de mathématiques ? Quel pourcentage de l'échantillon global cela représente-t-il ?

iii. Combien d'écoles ont un taux de réussite à l'examen de mathématiques égal à 50 % exactement ?

iv. Comparez les taux de réussite moyen pour les examens de mathématiques et de lecture. Lequel des deux examens est le plus dur à réussir ?

v. Trouvez la corrélation entre *math4* et *read4* (résultat à l'examen de mathématique et de lecture respectivement, en fin de quatrième année d'école élémentaire aux États-Unis). Quelle conclusion en tirez-vous ?

vi. La variable *exppp* représente les dépenses effectuées par élève. Indiquez la moyenne de *exppp* ainsi que son écart-type. Diriez-vous que la variation des dépenses par élève est importante ?

vii. Soient deux écoles A et B ; l'école A dépense 6 000 US$ par élèves et l'école B dépense 5 500 US$ par élève. De quel pourcentage les dépenses de l'école A dépassent les dépenses de l'école B ? Comparez votre résultat à $100 \times [\log(6\,000) - \log(5\,500)]$, qui représente l'approximation d'une différence en pourcentage (calculée à partir d'une différence en logarithmes naturels). (Voir la section A.4 de l'annexe A.)

C4. Les données de la base JTRAIN2 proviennent d'un programme de formation professionnelle conduite en 1976-1977 auprès d'hommes touchant un bas salaire. (Voir Lalonde (1986).)

i. Utilisez la variable indicatrice *train* pour déterminer la proportion d'hommes qui ont suivi la formation professionnelle.

ii. La variable *re78* indique le salaire perçu en 1978 (mais mesuré en dollars de 1982). Trouvez la moyenne de *re78* pour l'échantillon d'hommes qui ont bénéficié de la formation professionnelle et comparez-la à celle obtenue pour l'échantillon d'hommes qui ne l'ont pas suivie. Cette différence est-elle substantielle d'un point de vue économique ?

iii. La variable *unem78* indique si un homme est au chômage en 1978 (ou pas). Quelle est la proportion d'hommes qui ont suivi la formation professionnelle et qui sont au chômage ? Qu'en est-il pour les hommes qui n'ont pas bénéficié de cette formation ? Commentez la différence.

iv. À partir de vos réponses aux questions (ii) et (iii), que pensez-vous de l'efficacité du programme de formation ? Que peut-on faire pour rendre ces conclusions plus convaincantes ?

C5. Les données de la base FERTIL2 ont été collectées auprès de femmes vivant au Botswana en 1988. La variable *children* fait référence au nombre d'enfants (vivants). La variable *electric* est une variable binaire qui vaut 1 si le lieu de résidence est raccordé à l'électricité, et 0 sinon.

i. Trouvez les plus grandes et les plus petites valeurs de la variable *children* dans l'échantillon. Quelle est la moyenne de la variable *children* ?

ii. Quel est le pourcentage de femmes qui ont l'électricité à la maison ?

iii. Calculez la moyenne de la variable *children* pour celles qui n'ont pas l'électricité et faites de même pour celles qui en ont. Commentez vos résultats.

iv. À partir de la question (iii), pouvez-vous conclure que les femmes ont moins d'enfants lorsque leur maison est raccordée à l'électricité ? Expliquez.

C6. Utilisez les données de la base COUNTYMURDERS pour répondre à cette question. Utilisez seulement l'année 1996. La variable *murders* donne le nombre de meurtres signalés dans le comté. La variable *execs* correspond au nombre d'exécutions de personnes condamnées à mort qui ont eu lieu dans le comté concerné. La plupart des États américains appliquent la peine de mort, mais pas tous.

i. Combien de comtés y a-t-il dans la base de données ? Combien d'entre eux n'ont enregistré aucun meurtre ? Quel est le pourcentage de comtés où il n'y a pas eu d'exécutions ? (Attention, utilisez uniquement les données de 1996.)

ii. Quel est le plus grand nombre de meurtres enregistrés dans un comté ? Quel est le plus grand nombre d'exécutions réalisées dans un comté ? Pourquoi le nombre moyen d'exécutions par comté est-il si faible ?

iii. Calculez le coefficient de corrélation entre les meurtres et les exécutions et décrivez ce que vous trouvez.

iv. Vous devriez avoir calculé une corrélation positive dans la partie (iii). Pensez-vous que plus d'exécutions provoquent plus de meurtres ? Comment expliqueriez-vous cette corrélation positive ?

C7. La base de données ALCOHOL fournit des informations sur un échantillon d'hommes aux États-Unis. Elle comporte deux variables principales, qui sont la situation d'emploi autodéclarée et l'abus d'alcool, ainsi que de nombreuses autres variables. Les variables *employ* et *abuse* sont toutes deux des variables binaires (ou indicatrices) : elles ne prennent que les valeurs zéro et un.

i. Quel est le pourcentage d'hommes de l'échantillon qui déclarent avoir abusé de l'alcool ? Quel est le taux d'emploi ?

ii. Considérez le groupe d'hommes qui ont déclaré avoir abusé de l'alcool. Quel est le taux d'emploi pour cette population ?

iii. Quel est le taux d'emploi du groupe d'hommes qui déclare ne pas avoir abusé de l'alcool ?

iv. Discutez de la différence entre vos réponses aux parties (ii) et (iii). Cela vous permet-il de conclure que l'abus d'alcool est une cause de chômage ?

PARTIE 1

L'ANALYSE DE RÉGRESSION SUR DONNÉES EN COUPE TRANSVERSALE

La première partie de ce livre couvre l'analyse de régression en coupe transversale. Elle s'appuie sur une base solide d'algèbre, de probabilités et de statistiques. Les annexes A, B et C en offrent une révision complète.

Le chapitre 2 est consacré à la régression linéaire simple dans laquelle une variable n'est expliquée que par une seule autre. Bien que la régression simple ne soit pas couramment utilisée en économétrie appliquée, elle constitue un point de départ naturel, l'algèbre requise et les interprétations du modèle restant relativement élémentaires.

Les chapitres 3 et 4 portent sur les fondements de la régression multiple dans laquelle plusieurs variables peuvent affecter celle que l'on cherche à expliquer. La régression multiple est encore aujourd'hui le modèle empirique le plus couramment utilisé. Ces chapitres méritent donc une attention toute particulière. Le chapitre 3 traite de l'algèbre utilisée dans le cadre de la méthode des moindres carrés ordinaires (MCO), tout en identifiant les conditions sous lesquelles l'estimateur des MCO peut être sans biais et constituer le meilleur estimateur linéaire sans biais. Le chapitre 4 couvre le sujet primordial de l'inférence statistique.

Le chapitre 5 traite des propriétés asymptotiques des estimateurs des MCO, c'est-à-dire des propriétés qui ne concernent que les grands échantillons (théoriquement infinis). Ce chapitre explique les raisons pour lesquelles l'utilisation des procédures d'inférence décrites dans le chapitre 4 peut être justifiée même lorsque les erreurs d'un modèle de régression ne sont pas distribuées selon une loi normale. Le chapitre 6 est consacré à des problématiques plus spécifiques de l'analyse de régression, comme celles liées à la forme fonctionnelle, à l'échelle de mesure des données, aux prévisions, et à la qualité d'ajustement. Le chapitre 7 explique de quelle manière des informations qualitatives peuvent être incorporées dans les modèles de régression multiple.

Le chapitre 8 porte sur les tests et les méthodes de correction liés à la présence d'hétéroscédasticité, c'est-à-dire la présence d'une variance non constante dans le terme d'erreur. Nous montrons que les tests basés sur les MCO peuvent être ajustés et nous présentons également une extension de la méthode des MCO, appelée méthode des *moindres carrés pondérés*, qui tient explicitement compte du problème lié à une variance non constante dans le terme d'erreur. Le chapitre 9 approfondit l'étude de l'important problème lié à la corrélation entre le terme d'erreur et une ou plusieurs variables explicatives. Nous démontrons que la disponibilité d'une variable de substitution peut résoudre le problème lié à l'omission d'une variable importante dans le modèle. En outre, nous prouvons que les estimateurs des MCO sont biaisés et non convergents en présence de certaines formes d'erreur de mesure dans les variables du modèle. Plusieurs problèmes liés plus spécifiquement aux données sont également abordés, notamment le problème lié à l'existence d'observations isolées, voire aberrantes.

CHAPITRE 2

LE MODÈLE DE RÉGRESSION LINÉAIRE SIMPLE

Traduction de Mikael Petitjean

Le modèle de régression linéaire simple (RLS) est utilisé pour étudier la relation entre deux variables. Comme nous le verrons plus tard, l'utilisation de la régression simple, en tant qu'outil d'analyse empirique, est limitée. Elle reste néanmoins appropriée dans certaines circonstances bien spécifiques. Par ailleurs, apprendre à interpréter la régression simple reste une pratique recommandée avant de se lancer dans l'étude de la régression linéaire multiple, ce que nous ferons dans les chapitres suivants.

2.1 LA DÉFINITION DU MODÈLE DE RÉGRESSION LINÉAIRE SIMPLE

Une grande partie de l'analyse économétrique appliquée débute par l'énoncé des éléments de bases suivants : y et x sont deux variables et l'objectif est d'« expliquer y en fonction de x » ou encore d'« étudier comment y varie en fonction de x ». Le chapitre 1 contient plusieurs exemples de ce type. Par exemple, y est le rendement des cultures de soja et x est la quantité d'engrais ; y est le salaire horaire et x représente les années d'études ; ou y est le taux de criminalité au sein d'une communauté donnée et x est le nombre de policiers.

En élaborant un modèle qui cherche à « expliquer y en fonction de x », nous faisons face à trois problèmes. Tout d'abord, comment peut-on tenir compte de l'influence que d'autres facteurs peuvent avoir sur y, sachant qu'il est impossible de caractériser la relation exacte qui existe entre deux variables ? En second lieu, quelle relation fonctionnelle entre y et x doit-on privilégier ? En dernier lieu, comment peut-on s'assurer que l'effet *ceteris paribus* de x sur y soit bien mesuré (si tel est l'objectif souhaité) ?

Nous pouvons répondre à ces défis en écrivant une équation qui relie y à x. Une équation simple est :

$$y = \beta_0 + \beta_1 x + u. \qquad \textbf{[2.1]}$$

L'équation (2.1), que l'on suppose être vérifiée dans la population, définit le **modèle de régression linéaire simple**. Il est également appelé *modèle de régression linéaire à deux variables* ou *modèle de régression linéaire bivariée* car il relie tout simplement deux variables entre elles, x et y. Donnons maintenant une signification à chacun des éléments présents dans l'équation (2.1). [Cela dit en passant, nous n'expliquons pas l'origine du terme de « régression » car cela n'a pas d'implication particulière sur l'utilisation actuelle de la régression en économétrie. Voir Stigler (1986) pour un récit historique et divertissant de l'analyse de régression.]

Dans le cadre de l'équation (2.1), les variables y et x sont dénommées de plusieurs manières : y est appelée **variable dépendante**, **variable expliquée**, **variable prédite**, **variable de réponse**, **variable endogène**, **variable résultat** ou **variable contrôlée** ; x est appelée **variable explicative**, **variable indépendante**, **variable prédictive**, **régresseur**, **variable stimulus**, **variable exogène** ou **variable de contrôle**. (Le terme « covariable » est parfois utilisé pour x). Les termes « variable indépendante » et « variable dépendante » sont sans doute les appellations les plus fréquemment utilisées en économétrie. Gardons également à l'esprit que le qualificatif « indépendant » ne se rapporte pas à la notion statistique d'indépendance entre variables aléatoires (voir l'annexe B).

Les qualificatifs « expliquée » et « explicative » sont probablement les plus explicites. « Réponse » et « contrôle » sont utilisés surtout en sciences expérimentales, lorsque la variable x est sous le contrôle de l'expérimentateur. Bien que les termes de « variable prédite » et de « variable prédictive » apparaissent parfois dans les analyses de prévision pure, ils ne seront pas utilisés dans cet ouvrage centré avant tout sur les relations de cause à effet. La terminologie de base, utilisée dans le cadre de la régression simple, est résumée au tableau 2.1.

Tableau 2.1 *Terminologie de base dans le modèle de régression simple*

y	*x*
Variable dépendante	Variable indépendante
Variable expliquée	Variable explicative
Variable de réponse	Variable de contrôle
Variable prédite	Variable prédictive
Variable endogène	Variable exogène

La variable u représente le **terme d'erreur** ; ce terme traduit les **perturbations** qui affectent y et proviennent d'autres facteurs que x. La régression simple considère en effet que tous les facteurs affectant y, et différents de x, sont inobservables. On peut considérer u comme représentant l'ensemble des variables « non observées ».

L'équation (2.1) dispose également d'une relation fonctionnelle entre y et x bien particulière. Si les autres facteurs compris dans u sont maintenus constants, de telle sorte que la variation de u soit nulle, $\Delta u = 0$, alors x a un effet linéaire sur y :

$$\Delta y = \beta_1 \Delta x \text{ si } \Delta u = 0. \qquad [2.2]$$

La variation de y est donc tout simplement égale au produit de β_1 par la variation de x. Cela signifie que β_1 est le coefficient de **la pente** dans la relation entre y et x, tous les autres facteurs dans u étant maintenus constants ; ce coefficient revêt une importance toute particulière en économie appliquée. Le coefficient β_0 représente **la constante** ; il est parfois dénommé *ordonnée à l'origine*. Bien qu'il se trouve rarement au cœur de l'analyse, le coefficient β_0 est également utile.

EXEMPLE 2.1
Rendement des cultures de soja et engrais

Imaginez que le rendement des cultures de soja soit déterminé par le modèle suivant :

$$yield = \beta_0 + \beta_1 fertilizer + u. \qquad [2.3]$$

Le rendement (*yield*) est symbolisé par y et la quantité d'engrais (*fertilizer*) est représenté par x. En économie agricole, ce modèle de base peut servir à étudier l'effet des engrais sur le rendement des cultures de soja, toutes choses étant égales par ailleurs (*ceteris paribus*). Cet effet est donné par β_1. Le terme d'erreur u contient des facteurs tels que la qualité de la terre, les précipitations, etc. Le coefficient β_1 mesure l'effet des engrais sur le rendement, les autres facteurs étant maintenus constants : $\Delta yield = \beta_1 \Delta fertilizer$.

EXEMPLE 2.2
Salaire horaire et niveau d'instruction

En tenant compte des facteurs non observés, le modèle le plus élémentaire qui puisse expliquer le salaire horaire d'un individu (*wage*) par son niveau d'études (*educ*) peut s'écrire de la manière suivante :

$$wage = \beta_0 + \beta_1 educ + u \qquad [2.4]$$

Si *wage* est calculé en dollars par heure et *educ* indique les années d'études, alors β_1 mesure l'effet sur le salaire horaire d'une année supplémentaire d'instruction, toutes choses étant égales par ailleurs. Les facteurs non observés incluent le niveau d'expérience de l'individu sur le marché du travail, ses facultés innées, son ancienneté auprès de l'employeur actuel, son éthique au travail, et bien d'autres choses.

Dans l'équation (2.1), l'effet d'une variation d'une unité de x sur y est identique quelle que soit la valeur initiale de x. Dans bon nombre d'applications économiques, cette caractéristique n'est pas réaliste. Par exemple, dans le cas (2.2) basé sur la relation entre salaire et années d'études, il pourrait être judicieux d'autoriser la présence de rendements d'échelle *croissants*. Dans un tel cas de figure, chaque année d'études supplémentaire a un effet croissant sur les salaires. Nous apprendrons à modéliser un tel effet dans la section 2.4.

La question la plus difficile à traiter est de savoir si le modèle (2.1) nous permet vraiment de tirer des conclusions valides quant à l'effet *ceteris paribus* de x sur y. Dans l'équation (2.2), β_1 mesure l'effet de x sur y en supposant que *tous les autres facteurs (inclus dans le terme u) soient fixes*. Peut-on dès lors conclure que la question du lien de causalité est résolue ? Malheureusement, non. Nous devons encore déterminer s'il est possible de bien appréhender l'effet *ceteris paribus* de x sur y en supposant fixes tous les autres facteurs que nous ignorons dans le modèle (2.1) par ailleurs.

Dans la section 2.5, nous montrerons qu'il est possible d'obtenir, à partir d'un échantillon aléatoire de données, des estimateurs fiables de β_0 et β_1 à condition de poser une hypothèse portant sur la manière dont le terme non observé u est relié à la variable explicative x. Sans cette hypothèse, il est impossible d'estimer l'effet *ceteris paribus*, β_1. Étant donné que u et x sont des variables aléatoires, nous avons besoin d'un concept fondé sur la probabilité.

Avant d'énoncer cette hypothèse fondamentale sur lien entre x et u, nous pouvons poser une hypothèse sur u. Pour autant que le coefficient β_0 soit inclus dans l'équation, nous pouvons sans problème poser que la valeur moyenne de u dans la population est égale à zéro. Sur le plan mathématique, nous supposons que son espérance est nulle :

$$\mathrm{E}(u) = 0. \qquad \textbf{[2.5]}$$

L'hypothèse (2.5) ne donne aucune information sur la nature de la relation entre u et x. Elle porte uniquement sur la distribution des facteurs non observés dans la population. Sur base des exemples précédents, nous pouvons voir que l'hypothèse (2.5) n'est pas très restrictive. Dans l'exemple 2.1, nous ne perdons rien en supposant que les facteurs non observés qui affectent le rendement du soja, tels que la qualité de la terre, ont une moyenne égale à zéro dans la population de toutes les parcelles cultivées. La même chose est vraie des facteurs non observés dans l'exemple 2.2. Sans perte de généralité, nous pouvons en effet supposer que la capacité de toutes les personnes dans la population est nulle *en moyenne*. Si vous n'êtes toujours pas convaincu, vous devriez résoudre l'exercice 2 pour constater que nous pouvons toujours redéfinir le coefficient d'ordonnée à l'origine de l'équation (2.1) afin de respecter l'hypothèse (2.5).

Nous passons maintenant à l'hypothèse cruciale portant sur la manière dont le terme u et la variable x sont liés. Une mesure naturelle de l'association entre les deux variables aléatoires est le coefficient de corrélation (voir l'annexe B pour la définition et les propriétés du coefficient de corrélation). Si u et x ne sont pas corrélés, alors ils ne sont pas liés sur le plan linéaire. Cette absence de corrélation linéaire traduit la notion d'indépendance entre u et x dans l'équation (2.1) mais elle ne le fait qu'en partie car la corrélation ne mesure que le degré de dépendance linéaire entre u et x. En ce sens, la corrélation est problématique puisqu'il est possible que u ne soit pas corrélé avec x tout en l'étant avec des fonctions non linéaires de x, comme x^2 (voir la section B.4 pour poursuivre la discussion). Cette situation n'est pas acceptable dans la plupart des cas car elle fausse l'interprétation du modèle et rend la dérivation des propriétés statistiques problématique. Une meilleure façon d'énoncer l'hypothèse impliquerait donc *la valeur espérée de u étant donné x.*

Nous pouvons définir la distribution conditionnelle de u étant donné x puisque ces deux variables sont aléatoires. En particulier, il est possible d'obtenir la valeur espérée (ou la moyenne) de u pour chaque

tranche de la population décrite par la valeur de x, quelle que soit cette dernière. Le point crucial est que la valeur moyenne de u ne dépend pas de la valeur de x. Nous pouvons écrire cette hypothèse comme

$$E(u|x) = E(u). \qquad [2.6]$$

L'équation (2.6) indique que non seulement la valeur moyenne des variables non observées est la même pour toutes les tranches de la population, tranches déterminées par la valeur de x, mais aussi que la moyenne commune à ces tranches est nécessairement égale à la moyenne de u sur l'ensemble de la population. Lorsque l'hypothèse (2.6) est vérifiée, on dit que **l'espérance de u est indépendante de x**. (Bien évidemment, cette indépendance de l'espérance résulte de l'indépendance totale entre u et x, une hypothèse souvent utilisée en probabilités et statistiques.) Lorsque nous combinons cette indépendance de l'espérance à l'hypothèse (2.5), nous obtenons l'hypothèse que **l'espérance conditionnelle est égale à zéro**, $E(u|x) = 0$. Il est essentiel de se rappeler que l'équation (2.6) est l'hypothèse qui définit l'effet *ceteris paribus*. Quant à l'hypothèse (2.5), elle définit essentiellement la constante, β_0.

Voyons ce que l'équation (2.6) implique dans l'exemple portant sur le salaire. Pour simplifier la discussion, supposons que u représente l'aptitude innée d'une personne, facteur qui ne peut pas être directement observé. L'hypothèse (2.6) exige que le niveau moyen de l'aptitude innée soit le même, quel que soit le nombre d'années d'études. Par exemple, si $E(apt|8)$ désigne l'aptitude innée moyenne du groupe de personnes qui ont suivi un enseignement pendant huit ans, et que $E(apt|16)$ indique l'aptitude innée moyenne des personnes qui ont suivi un enseignement pendant seize ans, alors (2.6) implique que les deux moyennes doivent être les mêmes. En réalité, le niveau moyen d'aptitude innée doit être le même pour *tous* les niveaux d'enseignement. Si, par exemple, nous pensons que l'aptitude innée en moyenne augmente avec les années d'études, alors (2.6) est violée. (Ce sera le cas si, en moyenne, les personnes ayant une plus grande aptitude innée choisissent de s'instruire davantage.) Comme nous ne pouvons pas observer l'aptitude innée d'une personne, nous n'avons aucun moyen de savoir si elle est en moyenne effectivement la même pour tous les niveaux d'enseignement. C'est une question qu'il faut néanmoins se poser avant de recourir à une analyse de régression simple.

Dans l'exemple portant sur les engrais, (2.6) est vérifiée si les montants d'engrais sont choisis indépendamment des autres caractéristiques de la parcelle de terrain. Si tel est le cas, la qualité moyenne des terres ne dépendra pas de la quantité d'engrais. Toutefois, si plus (ou moins) d'engrais est répandu sur les parcelles dont la qualité de la terre est meilleure, alors la valeur attendue de u change avec le niveau d'engrais et (2.6) n'est pas respectée.

Pour aller plus loin 2.1

Imaginons que la note finale reçue à un examen, *score*, dépende à la fois de la proportion de cours qui ont été suivis par les étudiants (*attend*) et de plusieurs facteurs non observés qui influencent la performance des étudiants à l'examen (comme l'aptitude innée de l'étudiant). Dans ce cas,

$$score = \beta_0 + \beta_1\, attend + u. \qquad [2.7]$$

Dans quelles circonstances l'équation (2.6) est vérifiée ?

L'hypothèse de nullité de l'espérance conditionnelle offre une autre interprétation à β_1, qui se révèle souvent utile. En considérant la valeur attendue de (2.1), conditionnelle à x, et en utilisant $E(u|x) = 0$, on obtient

$$E(y|x) = \beta_0 + \beta_1 x. \qquad [2.8]$$

L'équation (2.8) correspond à la **fonction de régression de la population (FRP)**, $E(y|x)$, qui est une fonction linéaire de x. La linéarité implique qu'une augmentation d'une unité de x induit une variation de *la valeur attendue* de y égale à β_1. Pour chaque valeur de x, la distribution de y est centrée sur $E(y|x)$, comme représenté sur la figure 2.1.

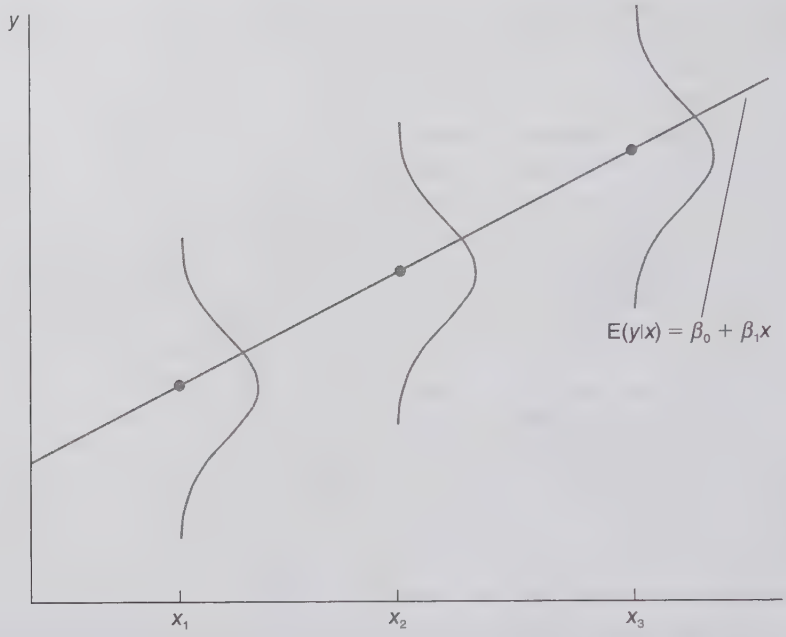


Figure 2.1 *E(y|x) en tant que fonction linéaire de x.*

Il est important de comprendre que l'équation (2.8) porte sur la valeur *moyenne* de y et sur sa variation en fonction de x ; elle ne dit pas que y est égale à $\beta_0 + \beta_1 x$ pour tous les éléments de la population. Par exemple, supposons que x soit la moyenne générale obtenue aux examens à la sortie du lycée (soit à la fin du secondaire supérieur). Aux États-Unis, cette moyenne correspond au « high school Grade Point Average » (*hsGPA*). Supposons également que y représente la moyenne générale obtenue aux examens de la licence universitaire (soit à la fin des trois premières années d'étude à l'université). Aux États-Unis, cette moyenne correspond au « college Grade Point Average » (*colGPA*). Imaginons que nous connaissons la FRP, soit $E(colGPA|hsGPA) = 1{,}5 + 0{,}5\ hsGPA$. [Bien sûr, dans la réalité, nous ne connaissons jamais la constante et la pente de la FRP, mais il est utile pour le moment de prétendre que nous le pouvons afin de mieux comprendre la nature de l'équation (2.8).] Cette FRP nous donne la note générale que les étudiants peuvent espérer *en moyenne* à la sortie de leur licence, étant donné leur note générale à la sortie du secondaire supérieur. Supposons que $hsGPA = 3{,}6$. Dans ce cas, la *moyenne* de *colGPA* pour tous les étudiants qui sont sortis du lycée avec une note de 3,6, sera égale à $1{,}5 + 0{,}5(3{,}6) = 3{,}3$. Nous n'affirmons certainement pas que *chaque* étudiant pour lequel $hsGPA = 3{,}6$ aura une note égale à 3,3 à la fin de la licence, ce qui serait évidemment faux. La FRP nous donne une relation entre le niveau moyen de y pour différents niveaux de x. Certains étudiants qui ont obtenu $hsGPA = 3{,}6$ auront $colGPA > 3{,}3$ alors que d'autres auront $colGPA < 3{,}3$. Le fait que la note obtenue à la fin de leur licence par ces étudiants soit supérieure ou inférieure à 3,3 va dépendre de facteurs non observés, compris dans u, qui varient au sein de cette tranche donnée de la population d'étudiants pour lesquels $hsGPA = 3{,}6$.

Étant donné l'hypothèse selon laquelle l'espérance conditionnelle du terme d'erreur est égale à zéro, soit $E(u|x) = 0$, il se révèle instructif de décomposer l'équation (2.1) en deux parties. La première partie, $\beta_0 + \beta_1 x$, représente $E(y|x)$ et caractérise la *partie systématique* de y, c'est-à-dire la partie de y qui est expliquée par x. La seconde partie, u, représente *la partie spécifique*, c'est-à-dire la partie de y qui n'est pas expliquée par x. Au chapitre 3, lorsque nous introduirons plusieurs variables explicatives, nous serons amenés à évaluer l'importance relative des parties systématique et spécifique.

Dans la section suivante, nous allons utiliser les hypothèses (2.5) et (2.6) pour justifier l'utilisation des estimateurs β_0 et β_1 étant donné un échantillon aléatoire de données. L'hypothèse selon laquelle l'erreur conditionnelle est nulle en moyenne joue un rôle crucial dans l'analyse statistique de la section 2.6.

2.2 LA DÉRIVATION DES ESTIMATEURS DES MOINDRES CARRÉS ORDINAIRES

Après avoir introduit les éléments fondamentaux du modèle de régression linéaire simple, nous allons chercher à estimer les paramètres β_0 et β_1 de l'équation (2.1). Pour cela, nous avons besoin d'un échantillon issu de la population. Soit un échantillon aléatoire de taille n issu de la population, tel que $\{(x_i, y_i) : i = 1, \ldots, n\}$. Sur base de l'équation (2.1), nous pouvons écrire que

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad [2.9]$$

pour chaque i. Dans ce cas, u_i est le terme d'erreur correspondant à l'observation i puisqu'il contient tous les facteurs non observés qui affectent y_i.

Par exemple, y_i et x_i pourraient respectivement correspondre à l'épargne et au revenu de la famille i pour une année donnée. Si nous rassemblons ces informations pour 15 familles, $n = 15$. Sur la figure 2.2, sont représentés le graphique en nuage de points et la FRP (nécessairement imaginaire) de l'épargne sur le revenu qui lui correspond.

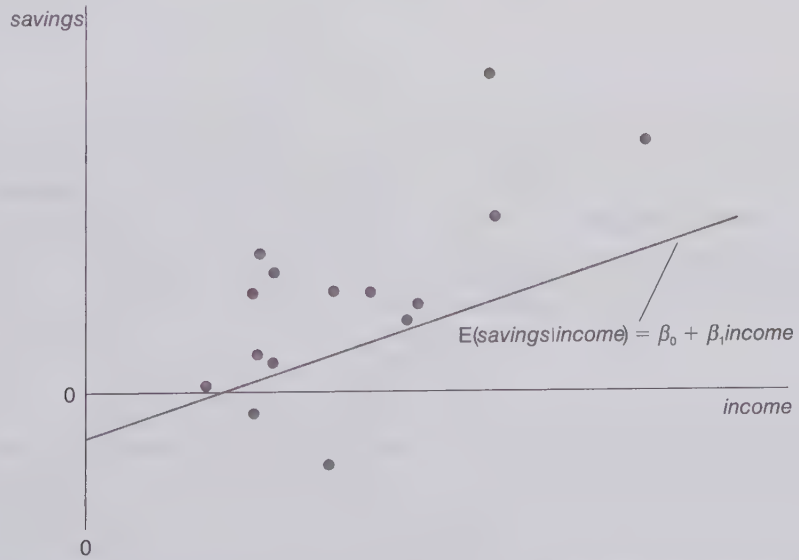


Figure 2.2 *Diagramme de dispersion de l'épargne (savings) et du revenu (income) pour 15 familles, et fonction de régression de la population* $E(\text{saving}|\text{income}) = \beta_0 + \beta_1 \text{income}$.

Nous devons maintenant décider de l'utilisation que nous allons faire de ces données pour obtenir des estimations de la constante et de la pente de cette FRP.

Il existe plusieurs manières de justifier le recours à la procédure d'estimation qui suit. Nous allons utiliser l'hypothèse (2.5) et une implication importante de l'hypothèse (2.6) : le terme d'erreur u n'est pas

corrélé avec x dans la population. Par conséquent, nous constatons que la valeur espérée de u est égale à zéro et que la *covariance* entre x et u est aussi égale à zéro :

$$\mathrm{E}(u) = 0 \qquad \textbf{[2.10]}$$

et

$$\mathrm{Cov}(x,u) = \mathrm{E}(x\,u) = 0. \qquad \textbf{[2.11]}$$

où la première égalité dans (2.11) est déduite de (2.10). (Voir la section B.4 pour la définition et les propriétés de la covariance.) Sur base des variables observables x et y et des paramètres inconnus β_0 et β_1, les équations (2.10) et (2.11) s'écrivent

$$\mathrm{E}(y - \beta_0 - \beta_1 x) = 0 \qquad \textbf{[2.12]}$$

et

$$\mathrm{E}[x(y - \beta_0 - \beta_1 x)] = 0, \qquad \textbf{[2.13]}$$

respectivement. Les équations (2.12) et (2.13) impliquent donc deux restrictions concernant la distribution de probabilité jointe de (x, y) dans la population. Comme il y a deux paramètres inconnus à estimer, nous espérons que les équations (2.12) et (2.13) peuvent être utilisées pour obtenir des estimateurs fiables de β_0 et β_1. En réalité, c'est le cas. Pour y parvenir, nous devons trouver les estimations, $\hat{\beta}_0$ et $\hat{\beta}_1$, qui permettent de résoudre les équations (2.12) et (2.13), ce qui requiert l'utilisation d'*un échantillon de données*. Sur base de cet échantillon, on obtient

$$n^{-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \textbf{[2.14]}$$

et

$$n^{-1}\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \textbf{[2.15]}$$

Il s'agit d'une application de la *méthode des moments* dans le cadre de l'estimation de paramètres. (Voir la section C.4 pour une discussion portant sur les différentes approches d'estimation.) Ces deux équations peuvent être résolues par rapport à $\hat{\beta}_0$ et $\hat{\beta}_1$.

En utilisant les propriétés de base de l'opérateur de sommation décrites dans l'annexe A, l'équation (2.14) devient

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \qquad \textbf{[2.16]}$$

où $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ est la moyenne de l'échantillon de y_i et de manière équivalente pour $\bar{x}$. Cette équation nous permet d'écrire $\hat{\beta}_0$ en fonction de $\hat{\beta}_1$, $\bar{y}$ et $\bar{x}$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \textbf{[2.17]}$$

Par conséquent, dès que nous obtenons l'estimation de la pente $\hat{\beta}_1$, le calcul de $\hat{\beta}_0$ est direct, étant donné $\bar{x}$ et $\bar{y}$.

En laissant tomber n^{-1} dans (2.15) (puisque cela ne change rien à la solution) et en insérant (2.17) dans (2.15), on obtient

$$\sum_{i=1}^{n} x_i\,[y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i] = 0,$$

qui, après réarrangement des termes, donne

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}).$$

En utilisant les propriétés élémentaires de l'opérateur de sommation [voir (A.7) et (A.8)],

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{et} \quad \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Par conséquent, à condition que

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0, \qquad \textbf{[2.18]}$$

l'estimation de la pente nous est donnée par

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \textbf{[2.19]}$$

L'équation (2.19) indique que l'estimation de la pente est égale à la covariance entre x_i et y_i divisée par la variance de x_i, toutes deux calculées sur base de l'échantillon. En utilisant un peu d'algèbre, nous pouvons également écrire que :

$$\hat{\beta}_1 = \hat{\rho}_{xy} \cdot \left(\frac{\hat{\sigma}_y}{\hat{\sigma}_x} \right),$$

où $\hat{\rho}_{xy}$ est le coefficient de corrélation entre x_i et y_i au sein de l'échantillon et $\hat{\sigma}_x$, $\hat{\sigma}_y$ désignent les écarts-types de l'échantillon. (Voir l'annexe C pour les définitions de la corrélation et de l'écart-type. Le fait de diviser le numérateur et le dénominateur par $n - 1$ ne change rien.). Une implication immédiate de (2.19) est que si x_i et y_i sont positivement corrélées dans l'échantillon, $\hat{\beta}_1$ est positif ; si x_i et y_i sont négativement corrélées, $\hat{\beta}_1$ est négatif.

Sans surprise, la formule de $\hat{\beta}_1$ est l'équivalent d'échantillonnage de la relation qui existe dans la population :

$$\beta_1 = \rho_{xy} \cdot \left(\frac{\sigma_y}{\sigma_x} \right),$$

dans laquelle toutes les quantités sont définies pour la population entière. Le fait que β_1 est une simple mesure de ρ_{xy} affectée d'un facteur d'échelle souligne une limite importante de la régression linéaire simple lorsqu'il est impossible d'obtenir des données expérimentales : une régression simple est, en réalité, une analyse de corrélation entre deux variables et il faut être prudent lorsqu'il s'agit d'en déduire un lien de causalité.

Bien que la méthode utilisée pour obtenir (2.17) et (2.19) repose sur (2.6), la seule contrainte pour obtenir les estimations de β_0 et β_1 à partir d'un échantillon donné est (2.18). Cette contrainte n'en est pas vraiment une : (2.18) est vérifiée si les observations x_i dans l'échantillon ne sont pas toutes égales à la même valeur. Si (2.18) est violée, soit nous avons été particulièrement malchanceux lors de la constitution de l'échantillon, soit nous avons choisi d'aborder un problème dénué de tout intérêt (puisque la variation de x serait nulle au sein de la population). Par exemple, si $y = salary$ et $x = educ$, (2.18) n'est violée que dans le cas où chaque personne reçoit le même niveau d'instruction (par exemple, dans le cas où chaque personne réussit sa douzième année d'études et ne les poursuit pas ; voir la figure 2.3). Il suffit qu'une seule personne

n'ait pas le même nombre d'années d'instruction pour que (2.18) soit vérifiée et que l'on puisse obtenir les estimations de β_0 et β_1.

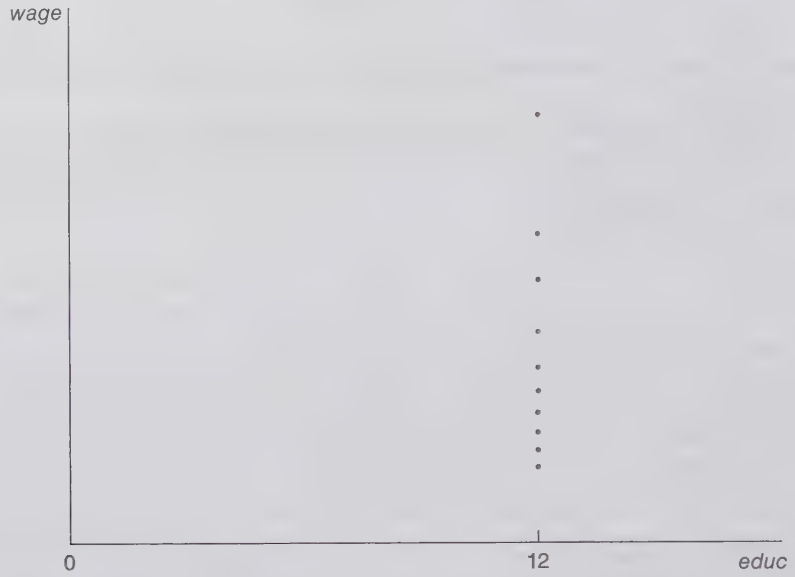


Figure 2.3 *Diagramme de dispersion du salaire et du niveau d'instruction lorsque* $\text{educ}_i = 12$ *pour tout* i.

Les estimations obtenues à partir de (2.17) et (2.19) sont appelées les estimations des **moindres carrés ordinaires** (**MCO**) de β_0 et β_1. Pour mieux comprendre cette appellation, définissons, pour tout $\hat{\beta}_0$ et $\hat{\beta}_1$, une **valeur ajustée** de y lorsque $x = x_i$. Nous obtenons

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \qquad \textbf{[2.20]}$$

Il s'agit de notre estimation de y lorsque $x = x_i$ pour des estimations données de la constante et de la pente. Il existe une valeur ajustée pour chaque observation dans l'échantillon. Le **résidu** pour cette observation i est égal à la différence entre la valeur observée de y_i dans l'échantillon et sa valeur ajustée :

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \qquad \textbf{[2.21]}$$

De nouveau, il existe n résidus, un résidu pour chaque observation. [Les résidus *ne* sont *pas* identiques aux erreurs de l'équation (2.9), une différence sur laquelle nous reviendrons dans la section 2.5.] Les valeurs ajustées et les résidus sont indiqués sur la figure 2.4.

Supposons maintenant que nous devons choisir $\hat{\beta}_0$ et $\hat{\beta}_1$ de manière à minimiser la **somme des carrés des résidus (SCR)**,

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \qquad \textbf{[2.22]}$$

Dans l'annexe 2A, il est démontré que les conditions nécessaires à la minimisation de (2.22) par rapport à $(\hat{\beta}_0, \hat{\beta}_1)$ sont exactement données par les équations (2.14) et (2.15), sans n^{-1}. Les équations (2.14) et (2.15) sont souvent appelées les **conditions de premier ordre** relatives aux estimations des MCO, un terme qui provient des techniques d'optimisation et du calcul différentiel (voir l'annexe A). Sur base de nos

calculs précédents, nous savons que les solutions aux conditions de premier ordre des MCO sont données par (2.17) et (2.19). Le terme de « moindres carrés ordinaires » vient du fait que ces estimations de β_0 et β_1 minimisent la somme des carrés des résidus.

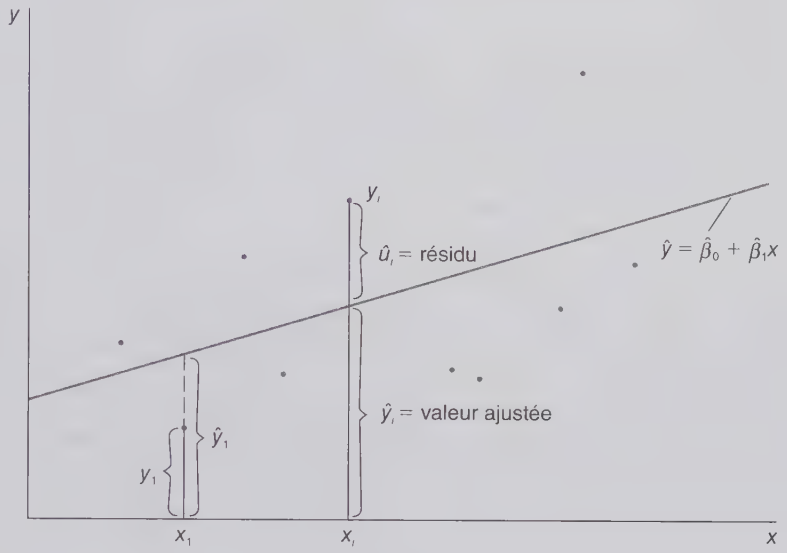


Figure 2.4 *Valeurs ajustées et résidus.*

À ce stade, il est naturel de se demander si une autre fonction que celle de la somme des carrés des résidus n'aurait pas pu être utilisée comme, par exemple, celle de la somme des valeurs absolues des résidus. En réalité, minimiser la somme des valeurs absolues des résidus est parfois très utile, comme nous le verrons dans la section 9.4. Cette fonction a néanmoins quelques inconvénients. Tout d'abord, il est impossible d'obtenir la formule des estimateurs ; l'alternative consiste à estimer les paramètres sur base de l'échantillon en utilisant des procédures d'optimisation numérique. Il en résulte que la théorie statistique des estimateurs qui minimisent la somme des résidus en valeur absolue est très compliquée. Minimiser d'autres fonctions des résidus, comme la somme des résidus élevés à la puissance quatre, par exemple, rencontre les mêmes inconvénients. (Nous ne devrions pas non plus chercher à minimiser la somme des résidus tels quels, étant donné que les résidus de grande ampleur mais de signes opposés pourraient se compenser.) Par contre, la méthode des MCO nous permet de dériver assez facilement les propriétés d'absence de biais et de convergence, parmi d'autres. En outre, comme le montre la section 2.5 et comme le sous-tendent les équations (2.13) et (2.14), la méthode des MCO est particulièrement adaptée pour estimer les paramètres de la FRP (2.8).

Dès que les estimations de la constante et de la pente sont calculées, nous obtenons **la droite de régression des MCO** :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \qquad \textbf{[2.23]}$$

où $\hat{\beta}_0$ et $\hat{\beta}_1$ sont calculés à l'aide des équations (2.17) et (2.19). La notation $\hat{y}$, se lisant « y chapeau », indique que les valeurs obtenues à partir de l'équation (2.23) sont des estimations. La constante, $\hat{\beta}_0$, est la valeur estimée de y lorsque $x = 0$. Dans certains cas, fixer $x = 0$ n'a aucun sens et $\hat{\beta}_0$ n'est alors, en soi, pas très intéressant. Si (2.23) est utilisée pour calculer les valeurs estimées de y pour différentes valeurs de x, nous devons néanmoins tenir compte de la constante. L'équation (2.23) désigne également la **fonction de**

régression de l'échantillon (FRE) car il s'agit de la version estimée de la FRP, soit $E(y|x) = \beta_0 + \beta_1 x$. Il est important de se souvenir que la FRP est unique mais inconnue. Vu que la FRE est obtenue à partir d'un échantillon particulier de données, un autre échantillon générera une pente et une constante différentes dans l'équation (2.23).

Dans la plupart des cas, l'estimation de la pente, qui est égale à

$$\hat{\beta}_1 = \Delta\hat{y} / \Delta x, \qquad \textbf{[2.24]}$$

est le point d'intérêt central de la droite de régression des MCO. Elle nous informe sur la variation de $\hat{y}$ suite à une variation d'une unité de x. De manière équivalente,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x, \qquad \textbf{[2.25]}$$

si bien que nous pouvons calculer la variation de y pour n'importe quelle variation de x, positive ou négative.

Nous allons maintenant introduire plusieurs exemples de régression simple, basés sur des données réelles. En d'autres termes, nous allons calculer la constante et la pente de la droite de régression à l'aide des équations (2.17) et (2.19). Comme ces exemples reposent sur l'utilisation de nombreuses données, les calculs ont été réalisés à l'aide d'un logiciel économétrique. À ce stade, gardez à l'esprit que vous ne devez pas tirer de conclusions trop hâtives de ces régressions simples concernant la relation causale qui existerait entre y et x. Nous n'avons encore rien dit des propriétés statistiques des MCO. Nous le ferons dans la section 2.5 après avoir posé plusieurs hypothèses sur le modèle de régression linéaire simple décrit par l'équation (2.1).

EXEMPLE 2.3
Salaire des PDG et rendement des capitaux propres

Soit y le salaire annuel (*salary*) en milliers de dollars américains (USD) pour une population de Présidents-Directeurs Généraux (PDG). Par conséquent, $y = 856{,}3$ représente un salaire annuel de 856 300 USD et $y = 1\,452{,}6$ représente un salaire annuel de 1 452 600. Soit x, le rendement moyen des capitaux propres (*roe*) réalisé par l'entreprise du PDG sur les trois dernières années. (Le rendement des capitaux propres est égal au bénéfice net divisé par les fonds propres ordinaires ; il est exprimé en pourcentage.) Par exemple, si *roe* = 10, le rendement moyen des capitaux propres = 10 %.

Dans le but d'étudier la relation entre la performance d'une entreprise et la rémunération de son PDG, nous proposons le modèle élémentaire suivant :

$$salary = \beta_0 + \beta_1\, roe + u.$$

Le paramètre de la pente β_1 mesure la variation du salaire annuel (en milliers de USD) lorsque le rendement sur fonds propres augmente d'un point de pourcentage ($\Delta roe = 1$). (Notez que si le rendement sur fonds propres passe de 4 % à 5 %, la variation est à la fois égale à un *point* de pourcentage et à 25 *pourcents*.) Étant donné qu'un *roe* plus élevé est considéré comme étant bénéfique pour l'entreprise, on s'attend à $\beta_1 > 0$.

La base de données CEOSAL1 contient ces informations pour 209 PDG au cours de l'année 1990 ; ces données proviennent du magazine *Business Week* publié le 6 mai 1991. Dans cet échantillon, le salaire annuel moyen est de 1 281 120 USD ; le moins élevé est égal à 223 000 USD et le plus élevé à 14 822 000 USD. Entre 1988 et 1990, le rendement moyen des capitaux propres est 17,18 % ; le moins élevé est égal à 0,5 % et le plus élevé à 56,3 %.

Sur base de cet échantillon de données, la droite de régression de *salary* sur le *roe*, estimée par les MCO, est

$$\widehat{salary} = 963{,}191 + 18{,}501\, roe$$
$$n = 209, \qquad \textbf{[2.26]}$$

dans laquelle les estimations de la constante et de la pente sont arrondies à trois chiffres après la virgule ; nous utilisons le « *salary* chapeau » pour indiquer qu'il s'agit d'une équation obtenue après estimation. Quelle interprétation pouvons-nous donner à cette équation ? Tout d'abord, si le rendement moyen des capitaux propres est nul, soit $roe = 0$, le salaire estimé est égal à la valeur de la constante, soit 963 191 USD (puisque *salary* est mesuré en milliers de dollars américains). Ensuite, nous pouvons exprimer la variation estimée de *salary* en fonction de la variation de *roe* : $\Delta\widehat{salary} = 18{,}501\ (\Delta roe)$. Si le rendement moyen des capitaux propres augmente d'un point de pourcentage, soit $\Delta roe = 1$, nous estimons une variation de *salary* égale à 18,501, soit 18 501 USD. Étant donné que (2.26) correspond à une équation linéaire, 18 501 USD représente la variation estimée du salaire quel que soit son niveau initial.

En nous servant de (2.26), nous pouvons aisément comparer les salaires estimés pour différentes valeurs du *roe*. Si nous supposons que $roe = 30$, $\widehat{salary} = 963{,}191 + 18{,}501\ (30) = 1\ 518{,}221$, soit juste au-dessus de 1,5 million de dollars américains. Cela *ne* signifie *pas* pour autant qu'un PDG en particulier, dont l'entreprise affiche un $roe = 30$, gagnera exactement 1 518 221 USD. Bien d'autres facteurs influencent le salaire. Il s'agit juste de notre estimation basée sur la droite de régression linéaire simple des MCO donnée par (2.26). Cette droite, qui correspond à la FRE, est indiquée sur la figure 2.5, à côté de la FRP, soit $E(salary|roe)$. Notez bien que nous ne sommes jamais capables d'observer la FRP dans la réalité ; la FRP est indiquée sur la figure dans le seul but pédagogique de rappeler que la FRE ne lui correspond pas nécessairement (et presque jamais, d'ailleurs). Nous ne sommes de toute manière pas capables de mesurer la distance qui sépare ces deux fonctions. Un autre échantillon de données donnera également une autre droite de régression, donc une autre FRE, qui pourra être plus ou moins proche de la FRP.

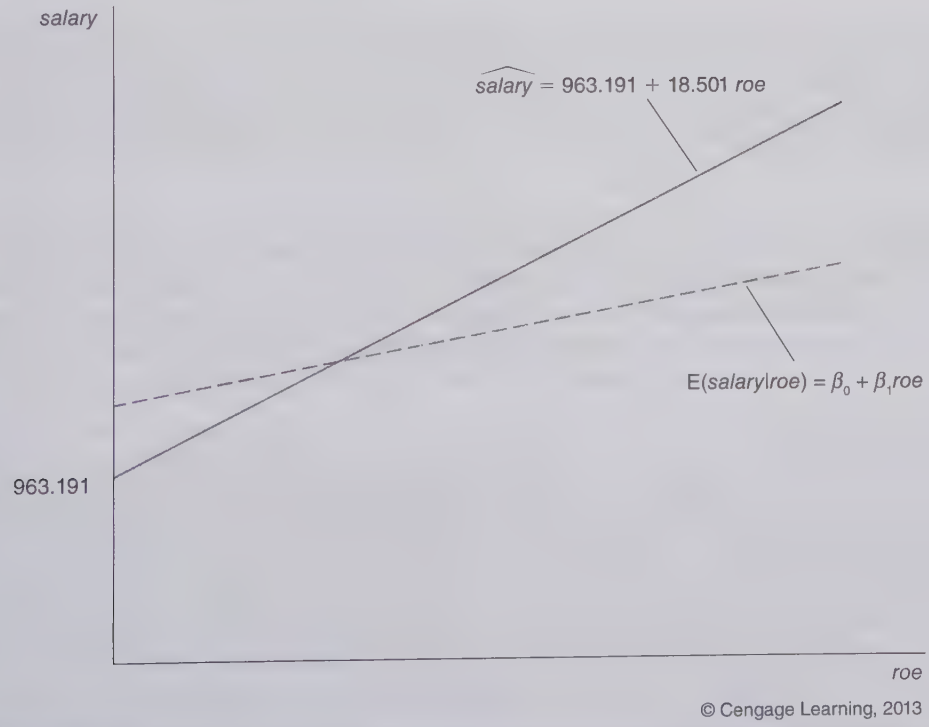


Figure 2.5 *Droite de régression des MCO,* $\widehat{salary} = 963{,}191 + 18{,}501\ roe$, *et fonction (inconnue) de régression de la population.*

EXEMPLE 2.4
Salaire horaire et niveau d'instruction

Soit $y = wage$, correspondant au salaire horaire, mesuré en USD, pour une population de personnes actives en 1976. Si $wage = 6{,}75$, cela signifie qu'une personne gagne un salaire horaire de 6,75 USD. Soit $x = educ$, correspondant au nombre d'années d'études. Par exemple, $educ = 12$ indique que la personne a terminé ses études secondaires (et n'est pas allée plus loin). Le salaire horaire moyen dans l'échantillon est égal à 5,90 USD. Cela équivaut à 19,06 USD en 2003, en corrigeant pour l'inflation grâce à l'Indice des Prix à la Consommation.

La base de données WAGE1 contient des informations pour 526 personnes ($n = 526$). La droite de régression linéaire simple (ou FRE) du salaire sur les années d'études, obtenue par les MCO, est :

$$\widehat{wage} = -0{,}90 + 0{,}54\ educ$$
$$n = 526. \qquad [2.27]$$

Nous devons interpréter ces résultats avec prudence. Une valeur de – 0,90 pour la constante signifie littéralement que le salaire horaire estimé d'une personne n'ayant jamais été scolarisée est de – 90 centimes de dollars américains par heure de travail. Bien entendu, cela n'a pas de sens. Il s'avère que seules 18 personnes sur 526 ont été scolarisées pendant moins de huit années dans l'échantillon. Par conséquent, il n'est pas étonnant qu'à ce faible niveau d'instruction, la droite de régression donne des estimations médiocres. Pour une personne dont les années d'études sont égales à huit, le salaire estimé est $\widehat{wage} = -0{,}90 + 0{,}54\ (8) = 3{,}42$, soit 3,42 USD par heure de travail (en 1976).

Pour aller plus loin 2.2

Lorsque $educ = 8$, le salaire horaire estimé à partir de (2.27) est égal à 3,42 USD en 1976. Que vaudrait ce salaire horaire en 2003 ? (*Astuce* : il y a suffisamment d'information disponible dans l'exemple 2.4 pour répondre à cette question).

L'estimation de la pente dans (2.27) implique qu'une année d'études supplémentaire permet d'augmenter le salaire de 54 centimes par heure de travail. Par conséquent, quatre années d'études supplémentaires conduisent à une augmentation estimée du salaire horaire de 4 (0,54) = 2,16, soit 2,16 USD. Il s'agit de variations relativement importantes. En raison de la linéarité de (2.27), toute année d'études supplémentaire augmente le salaire horaire du même montant, quel que soit le niveau initial d'instruction. Dans la section 2.4, nous introduirons plusieurs méthodes qui permettent de tenir compte d'éventuels effets marginaux non constants de x sur y.

EXEMPLE 2.5
Résultats du scrutin et dépenses électorales

Le fichier VOTE1 contient des données sur les dépenses de campagne électorale et les résultats du scrutin pour 173 joutes électorales lors des élections à la Chambre des représentants des États-Unis en 1988. Il y a deux candidats dans la course, A et B. Soit *voteA*, le pourcentage des votes que le candidat A reçoit, et *shareA*, le pourcentage du total des dépenses électorales dont le candidat A est responsable. D'autres facteurs que *shareA* influencent le résultat électoral (dont la qualité des candidats et éventuellement le *montant total* en USD dépensés par les deux candidats). En gardant cela à l'esprit, nous pouvons néanmoins estimer une droite de régression linéaire des MCO pour déterminer si une dépense électorale plus importante que le concurrent implique un pourcentage de votes plus élevé.

Basée sur ces 173 observations, la FRE est

$$\widehat{voteA} = 26{,}81 + 0{,}464\ shareA$$
$$n = 173. \qquad \textbf{[2.28]}$$

Si la part du candidat A dans les dépenses électorales totales augmente d'un point de pourcentage, le candidat A captera un peu moins d'un demi-point de pourcentage en plus du total des votes (soit 0,464 %). Bien qu'il soit difficile de savoir s'il s'agit réellement d'un effet causal, cette estimation ne semble pas farfelue. Si *shareA* = 50, *voteA* est estimé à environ 50 %, soit la moitié des votes.

Dans certains cas, l'analyse de régression n'est pas utilisée pour déterminer la causalité entre deux variables mais simplement pour étudier la nature positive ou négative de leur relation, comme pourrait le dévoiler une analyse de corrélation classique. Une illustration en est donnée dans l'exercice sur ordinateur C3. Cet exercice est basé sur les données de Biddle et Hamermesh (1990) qui analysent les heures de sommeil et de travail dans le but de savoir s'il existe un effet de substitution entre les deux.

Pour aller plus loin 2.3

Dans l'exemple 2.5, quelle est l'estimation du pourcentage de vote que le candidate A capte si *shareA* = 60 (ce qui signifie 60 % des dépenses électorales) ? La réponse est-elle cohérente ?

Remarque sur la terminologie

Dans la plupart des cas, nous reporterons les estimations obtenues par les MCO en indiquant les FRE telles que (2.26), (2.27) ou (2.28). Par souci de concision, il est parfois utile de ne pas reporter les résultats de la droite de régression des MCO. Dans ce cas, nous préciserons que l'équation (2.23) est estimée en écrivant que *nous effectuons une régression de*

$$y \text{ sur } x, \qquad \textbf{[2.29]}$$

ou, tout simplement, que nous *régressons* y sur x. L'ordre respectif de y et x dans (2.29) indique que la première variable est la variable dépendante et que la seconde correspond à la variable indépendante. Nous devons naturellement toujours régresser la variable dépendante sur la variable indépendante. Lorsqu'il s'agit d'analyser des relations entre des variables bien spécifiques, nous remplaçons y et x par leur nom respectif. Par exemple, nous régressons *salary* sur *roe* pour obtenir (2.26) et nous régressons *voteA* sur *shareA* pour obtenir (2.28).

Lorsque cette terminologie est utilisée, l'objectif est d'estimer à la fois la constante $\hat{\beta}_0$ et la pente $\hat{\beta}_1$. Ce sera le cas dans l'écrasante majorité des applications dans ce livre. Dans quelques circonstances bien spécifiques, nous chercherons à estimer la relation entre y et x *en supposant* que la constante est égale à zéro (de telle sorte que si $x = 0$, $\hat{y} = 0$) ; nous aborderons brièvement ce cas spécifique dans la section 2.6. Sans indication contraire, nous cherchons toujours à estimer la constante et la pente de la droite de régression.

2.3 LES PROPRIÉTÉS DES MCO EN ÉCHANTILLON

Dans la section précédente, nous avons utilisé quelques notions d'algèbre pour dériver les formules de la constante et de la pente, à partir desquelles nous obtenons les estimations. Dans cette section, nous étudions d'autres propriétés algébriques de la FRE. Pour le moment, la meilleure chose à faire est de considérer que ces propriétés s'appliquent, par construction, à *n'importe quel* échantillon particulier de données. Le travail

plus ardu, qui consistera à étudier les propriétés statistiques des MCO (en se basant sur l'ensemble de tous les échantillons aléatoires de données), sera réalisé à la section 2.5.

Plusieurs propriétés algébriques que nous allons dériver sembleront triviales. Une bonne compréhension de ces propriétés nous aidera néanmoins à comprendre ce qu'il advient des estimations des MCO et des tests statistiques lorsque les données sont modifiées, à la suite d'un changement des unités de mesure des variables x et y par exemple.

Valeurs ajustées et résidus

Supposons que les estimations de la constante et de la pente, $\hat{\beta}_0$ et $\hat{\beta}_1$, soient obtenues à partir d'un échantillon de données. Étant donné $\hat{\beta}_0$ et $\hat{\beta}_1$, nous pouvons calculer la valeur ajustée $\hat{y}_i$ pour chaque observation i. [Voir l'équation (2.20).] Par définition, chaque valeur ajustée $\hat{y}_i$ se trouve sur la droite de régression des MCO. Le résidu associé à l'observation i, soit $\hat{u}_i$, mesure la différence entre l'observation y_i et sa valeur ajustée $\hat{y}_i$, comme indiqué à l'équation (2.21). Si $\hat{u}_i$ est positif, la droite des MCO « sous-estime » y_i ; si $\hat{u}_i$ est négatif, la droite des MCO surestime y_i. Le cas idéal pour l'observation i est lorsque $\hat{u}_i = 0$. Néanmoins, dans la plupart des cas, *tous* les résidus ne sont pas nuls. Il n'est d'ailleurs pas nécessaire que toutes les observations correspondent à leurs valeurs ajustées et se situent sur la droite des MCO.

EXEMPLE 2.6
Salaire des PDG et rendement des capitaux propres

Dans le tableau 2.2, sont affichées les valeurs des variables indépendante (*roe*) et dépendante (*salary*) pour les 15 premiers PDG de la base de données. Sont également reprises les valeurs ajustées de *salaire* (*salarychap*) et les résidus qui leur correspondent (*uchap*).

Les quatre premiers PDG perçoivent des salaires inférieurs à ceux qui sont estimés par la droite de régression des MCO (2.26). En d'autres termes, étant donné le *roe de leur entreprise*, ces PDG gagnent moins que ce que nous pourrions justifier sur base de la droite des MCO. Comme la valeur positive de *uchap l'indique*, le cinquième PDG de l'échantillon gagne plus que ne le prédit la droite de régression.

Tableau 2.2 *Valeurs ajustées et résidus pour les 15 premiers PDG*

obs	roe	salary	salarychap	uchap
1	14,1	1 095	1 224,058	– 129,0581
2	10,9	1 001	1 164,854	– 163,8542
3	23,5	1 122	1 397,969	– 275,9692
4	5,9	578	1 072,348	– 494,3484
5	13,8	1 368	1 218,508	149,4923
6	20,0	1 145	1 333,215	– 188,2151
7	16,4	1 078	1 266,611	– 188,6108
8	16,3	1 094	1 264,761	– 170,7606
9	10,5	1 237	1 157,454	79,54626
10	26,3	833	1 449,773	– 616,7726

obs	roe	salary	salarychap	uchap
11	25,9	567	1 442,372	– 875,3721
12	26,8	933	1 459,023	– 526,0231
13	14,8	1 339	1 237,009	101,9911
14	22,3	937	1 375,768	– 438,7678
15	56,3	2 011	2 004,808	6,191895

Propriétés algébriques des statistiques dérivées de la méthode des MCO

Il existe plusieurs propriétés algébriques dont disposent les estimations et autres statistiques dérivées de la méthode des MCO. Nous allons identifier les trois propriétés les plus importantes.

(1) La somme des résidus est égale à zéro. Il en va, par conséquent, de même pour la moyenne des résidus. Sur le plan mathématique,

$$\sum_{i=1}^{n} \hat{u}_i = 0. \qquad \textbf{[2.30]}$$

Cette propriété ne requiert aucune démonstration. Elle découle directement de la condition de premier ordre (2.14) des MCO, en notant que $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Autrement dit, les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ sont *déterminées* de telle sorte que la somme des résidus est égale à zéro (quel que soit l'échantillon). Cela ne dit naturellement rien quant à la valeur résiduelle correspondant à chaque observation i.

(2) La covariance entre les valeurs explicatives et les résidus des MCO est nulle. Cette propriété découle de la condition de premier ordre (2.15), dans laquelle on peut faire apparaître les résidus de la manière suivante :

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0. \qquad \textbf{[2.31]}$$

Comme la moyenne des résidus est nulle, la partie gauche de (2.31) est proportionnelle à la covariance entre x_i et $\hat{u}_i$.

(3) Le point $(\bar{x}, \bar{y})$ est toujours situé sur la droite de régression des MCO. En d'autres termes, si nous considérons l'équation (2.23) et que nous remplaçons x par $\bar{x}$, la valeur ajustée de y sera égale à $\bar{y}$, ce que l'équation (2.16) nous démontrait précisément.

EXEMPLE 2.7
Salaire horaire et niveau d'instruction

Dans la base de données WAGE1, le salaire horaire moyen est égal à 5,90, arrondi à deux chiffres après la virgule, et la moyenne du niveau d'instruction est égale à 12,56. Si nous utilisons *educ* = 12,56 dans la FRE (2.27), nous obtenons $\widehat{wage} = -0{,}90 + 0{,}54(12{,}56) = 5{,}8824$, ce qui équivaut à 5,9 lorsque le résultat est arrondi à un chiffre après la virgule. Les valeurs ne sont pas exactement identiques car nous avons arrondi les moyennes du salaire et du niveau d'instruction, ce que nous avons également fait pour les estimations de la constante et de la pente. Si nous avions utilisé les valeurs exactes, nous aurions obtenu des réponses beaucoup plus proches, sans que cela ne modifie en rien nos conclusions.

Une autre manière d'interpréter une régression des MCO est de partir de la constatation que la valeur observée y_i est égale à la somme de sa valeur ajustée et de son résidu. Pour chaque i, on obtient

$$y_i = \hat{y}_i + \hat{u}_i. \qquad \textbf{[2.32]}$$

De la propriété (1), nous savons que la moyenne des résidus est égale à zéro ; dès lors, la moyenne des valeurs observées, y_i, est égale à la moyenne des valeurs ajustées, $\hat{y}_i$, soit $\bar{y} = \bar{\hat{y}}$. En outre, les propriétés (1) et (2) peuvent servir à démontrer que la covariance entre $\hat{y}_i$ et $\hat{u}_i$ est égale à zéro. En résumé, nous pouvons considérer la régression des MCO comme une manière de diviser chaque observation y_i en deux composantes, une valeur ajustée et une valeur résiduelle, ces valeurs n'étant pas corrélées dans l'échantillon.

Grâce à cette décomposition, nous pouvons définir la **somme des carrés totaux (SCT)**, la **somme des carrés expliqués (SCE)** et **la somme des carrés des résidus (SCR)** de la manière suivante :

$$\text{SCT} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad \textbf{[2.33]}$$

$$\text{SCE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \qquad \textbf{[2.34]}$$

$$\text{SCR} = \sum_{i=1}^{n}\hat{u}_i^2. \qquad \textbf{[2.35]}$$

SCT est une mesure de la variation totale entre les y_i de l'échantillon ; autrement dit, SCT mesure le degré de dispersion des y_i dans l'échantillon. Si nous divisons SCT par $n - 1$, nous obtenons la variance de y dans l'échantillon, comme indiqué dans l'annexe C. De manière équivalente, SCE mesure la variation au sein des $\hat{y}_i$, en notant que $\bar{\hat{y}} = \bar{y}$. Enfin, SCR, que l'on appelle également somme des résidus au carré, mesure la variation observée entre les $\hat{u}_i$. La variation totale de y (SCT) peut donc être exprimée comme la somme de la variation expliquée (SCE) et de la variation résiduelle (SCR), soit

$$\textbf{SCT = SCE + SCR.} \qquad \textbf{[2.36]}$$

Il n'est pas difficile de démontrer (2.36). Cela requiert néanmoins l'utilisation de toutes les propriétés de l'opérateur de sommation, qui sont reprises dans l'annexe A. Par un simple artifice mathématique, nous pouvons écrire que :

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=0}^{n}[\hat{u}_i + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \text{SCR} + 2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) + \text{SCE}.\end{aligned}$$

Par conséquent, (2.36) est vérifiée si nous pouvons montrer que

$$\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) = 0. \qquad \textbf{[2.37]}$$

Or, nous avons indiqué que la covariance entre les résidus et les valeurs ajustées est nulle et que cette covariance est précisément égale à (2.37), divisée par $n - 1$. L'équation (2.36) est donc validée.

Il n'y a malheureusement pas de consensus quant à la terminologie à employer concernant les trois statistiques SCT, SCE, et SCR. Concernant la première, il n'y a pas trop de problème. La somme des carrés totaux (SCT) est également appelée « somme des carrés totale ». Pour les deux suivantes, le risque de confusion est plus grand. La somme des carrés expliqués est parfois dénommée « somme des carrés de la régression », dont l'abréviation se confond alors avec celle de la somme des carrés des résidus (SCR). Dans certains logiciels économétriques, notons également que la somme des carrés expliqués est appelée « somme des carrés du modèle ».

Pour rendre les choses encore plus compliquées, la somme des carrés des résidus est souvent confondue, à tort, avec la somme des carrés des erreurs (ou de l'erreur). Comme nous le verrons à la section 2.5, les erreurs et les résidus sont des quantités différentes. Pour éviter de confondre les deux, nous utiliserons systématiquement les termes de « somme des carrés des résidus », « somme des carrés résiduelle » ou « somme des résidus au carré » pour caractériser l'équation (2.35). En anglais, pour désigner SCR, il existe deux sigles, SSR (« Sum of Squared Residuals ») et RSS (« Residual Sum of Squares »), le premier étant plus fréquemment utilisé dans les logiciels économétriques que le second.

Qualité d'ajustement

Jusqu'ici, nous n'avons vu aucun outil qui nous aide à déterminer si la variable explicative, x, explique correctement la variable dépendante, y. Il est souvent utile de recourir à une statistique qui mesure précisément la qualité d'ajustement de la droite de régression des MCO aux données. Dans la discussion qui suit, gardez en mémoire que la pente de la droite est estimée en même temps que la constante.

En considérant que la somme des carrés totaux (SCT) n'est pas nulle (ce qui est vrai, sauf dans le cas dénué d'intérêt où tous les y_i sont égaux à la même valeur), nous pouvons diviser (2.36) par SCT et obtenir : 1 = SCE/SCT + SCR/SCT. De cette égalité, on obtient le « ***R* carré** » de la régression, également appelé le **coefficient de détermination**, soit

$$R^2 = \text{SCE} / \text{SCT} = 1 - (\text{SCR} / \text{SCT}). \qquad \textbf{[2.38]}$$

Le R^2 est calculé en divisant la variation expliquée par la variation totale ; il représente la *fraction de la variation de y qui est expliquée par x au sein de l'échantillon*. La seconde égalité dans (2.38) constitue une autre manière de calculer le R^2.

Grâce à (2.36), nous savons que la valeur du R^2 sera toujours comprise entre zéro et un, étant donné que SCE ne peut pas être plus élevé que SCT. Lorsqu'il s'agit d'interpréter le R^2, on le multiple par 100 pour obtenir un pourcentage : 100 R^2 est *le pourcentage de la variation de y présente dans l'échantillon, qui est expliquée par x.*

Si tous les points correspondant aux données observées se situent sur la même droite, la méthode des MCO offre un ajustement parfait de la droite de régression aux données observées. Dans ce cas, $R^2 = 1$. Une valeur du R^2 proche de zéro indique que l'ajustement est de piètre qualité : la variation entre les $\hat{y}_i$ (qui se trouvent tous sur la droite de régression des MCO) ne capture quasiment rien de la variation observée entre les y_i. En réalité, on peut démontrer que le R^2 est égal au *carré* du coefficient de corrélation entre y_i et $\hat{y}_i$ au sein de l'échantillon. Le terme « R carré » en découle. La lettre R a traditionnellement été utilisée pour symboliser l'estimation du coefficient de corrélation de la population, soit *rho* (ρ) en grec. Cet usage a perduré dans le domaine de la régression linéaire.

Dans le domaine des sciences sociales, obtenir des valeurs faibles pour le R^2 n'est pas inhabituel, particulièrement dans le cas d'une analyse transversale. Nous traiterons de cette problématique de manière

plus systématique dans le cadre de l'analyse de régression linéaire multiple. Il est néanmoins important de souligner qu'un faible R^2 ne signifie pas nécessairement que la régression des MCO ne sert à rien. Dans l'exemple 2.8, il est en effet possible que (2.39) soit une estimation fiable de la relation qui existe entre *salary* et *roe*, toutes choses étant égales par ailleurs ; la faible valeur du R^2 ne nous renseigne pas sur la fiabilité de cette estimation. Les étudiants qui découvrent l'économétrie ont tendance à donner trop de poids à la valeur du R^2 lorsqu'ils évaluent les résultats d'une régression linéaire. À ce stade, gardez à l'esprit que l'utilisation du R^2 comme outil principal d'évaluation d'une analyse économétrique peut poser problème.

EXEMPLE 2.8
Salaire des PDG et rendement des capitaux propres

Dans la régression du salaire des PDG,

$$\widehat{Salary} = 963{,}191 + 18{,}501\ roe$$
$$n = 209,\ R^2 = 0{,}0132. \qquad \textbf{[2.39]}$$

Nous reprenons les résultats de la droite de régression des MCO ainsi que le nombre d'observations. Nous y ajoutons le R^2, arrondi à quatre décimales, pour évaluer le pourcentage de la variation de *salary* qui est réellement expliquée par le rendement des capitaux propres. Il s'agit d'un très faible pourcentage. Le rendement des capitaux propres d'une grande entreprise cotée en bourse explique à peine 1,3 % de la variation totale des salaires que l'on observe pour les 209 PDG inclus dans l'échantillon. Autrement dit, 98,7 % de la somme du carré des écarts de chaque salaire par rapport à la moyenne restent inexpliqués. Ce faible pouvoir explicatif n'est pas nécessairement une surprise puisque de nombreuses caractéristiques propres à l'entreprise et au PDG sont susceptibles d'influencer le salaire. Dans une régression simple telle que (2.39), ces facteurs sont tout simplement inclus dans les erreurs.

Il arrive aussi que la variable x parvienne à expliquer une part conséquente de la variation totale de y dans l'échantillon.

EXEMPLE 2.9
Résultats du scrutin et dépenses électorales

Dans l'équation (2.28) portant sur les résultats du scrutin, $R^2 = 0{,}856$. Par conséquent, les dépenses électorales expliquent 86 % de la variation totale observée pour les résultats des élections au sein de l'échantillon. Cela représente un pourcentage substantiel.

2.4 LES UNITÉS DE MESURE ET LA FORME FONCTIONNELLE

En économie appliquée, il est important de : (1) comprendre l'impact qu'un changement des unités de mesure des variables présentes dans le modèle peut avoir sur les estimations des MCO ; (2) parvenir à utiliser, dans le cadre d'une régression linéaire, les formes fonctionnelles que l'on rencontre le plus fréquemment en économie. Le développement mathématique nécessaire à une compréhension approfondie du sujet portant sur les formes fonctionnelles est disponible dans l'annexe A.

Effets du changement des unités de mesure sur les statistiques des MCO

Dans l'exemple 2.3, nous avons choisi de mesurer le salaire annuel en milliers de dollars américains, et d'exprimer le rendement sur capitaux propres en pourcentage (plutôt que sous la forme de décimales). Il est impératif de le savoir avant de donner une interprétation aux estimations de l'équation (2.39).

Il est également important de noter que les estimations obtenues par les MCO changent d'une manière totalement prévisible lorsque les unités de mesure des variables dépendante et indépendante sont modifiées. Supposons que nous mesurions le salaire en dollars, plutôt qu'en milliers de dollars. Soit *salardol*, le salaire en dollars (*salardol* = 845 761 implique un salaire de 845 761 USD). La relation entre *salardol* et le salaire mesuré en milliers de dollars est simple : *salardol* = 1 000 *salary*. Nous n'avons donc pas besoin d'estimer la régression de *salardol* sur *roe* pour savoir que la FRE sera :

$$\widehat{salardol} = 963\ 191 + 18\ 501\ roe. \qquad \textbf{[2.40]}$$

Nous obtenons les estimations de la constante et de la pente de (2.40) en multipliant les estimations de (2.39) par 1 000. L'interprétation des équations (2.39) et (2.40) est identique. Dans (2.40), si *roe* = 0, alors $\widehat{salardol}$ = 963 191, soit un salaire estimé à 963 191 USD. Cette valeur est identique à celle que nous avons obtenue à partir de l'équation (2.39). En outre, si *roe* augmente de 1 (point de pourcentage), l'augmentation du salaire sera estimée à 18 501 USD, encore une fois identique aux conclusions que nous avons tirées de notre analyse de l'équation (2.39).

En règle générale, il est aisé de déterminer les estimations de la constante et de la pente lorsque la variable dépendante change d'unités de mesure. Si la variable dépendante est multipliée par le facteur d'échelle c, ce qui signifie que chaque valeur est multipliée par c, alors les estimations de la constante et de la pente sont également multipliées par c. (On suppose naturellement que l'unité de mesure de la variable indépendante ne change pas). Dans l'exemple sur le salaire des PDG, c = 1 000 lorsque nous passons de *salary* à *salardol*.

Nous pouvons également utiliser l'exemple sur le salaire des PDG pour examiner l'impact d'un changement dans les unités de mesure d'une variable indépendante. Soit *roedec* = *roe*/100, qui est l'équivalent de *roe* sous la forme décimale ; *roedec* = 0,23 signifie donc que le rendement sur capitaux propres est égal à 23 %. Pour analyser l'effet propre du changement de l'unité de mesure de la variable indépendante, nous retournons à notre variable dépendante de départ, *salary*, qui est mesurée en milliers de dollars. Lorsque nous régressons *salary* sur *roedec*, nous obtenons

$$\widehat{salary} = 963{,}191 + 1\ 850{,}1\ roedec \qquad \textbf{[2.41]}$$

Le coefficient de *roedec* est égal à 100 fois le coefficient de *roe* dans (2.39), conformément aux attentes. Comme Δroe = 1 équivaut à $\Delta roedec$ = 0,01, nous obtenons dans (2.41) que $\Delta\widehat{salary}$ = 1 850,1 (0,01) = 18,501. Le résultat est identique à celui obtenu sur base de (2.39). Comme la variable indépendante est divisée par 100 en passant de (2.39) à (2.41), l'estimation de la pente des MCO doit être multipliée par 100. L'égalité de l'équation est préservée et son interprétation est inchangée. En règle générale, si la variable indépendante est divisée ou multipliée par un facteur non nul, c, alors le coefficient de la pente des MCO doit être respectivement multiplié ou divisé par c.

La constante dans (2.41) n'a pas changé, étant donné que *roedec* = 0 est identique à *roe* = 0. Sur un plan plus général, la modification des unités de mesure d'une variable indépendante ne donne lieu à aucun changement de la constante.

Pour aller plus loin 2.4

Soit *salarhun*, le salaire des PDG mesuré en centaines de dollars, plutôt qu'en milliers de dollars. Quelles seront les estimations des MCO pour la constante et la pente de la droite de régression de *salarhun* sur *roe* ?

Dans la section précédente, nous avons défini le R^2 comme une mesure de la qualité d'ajustement de la droite de régression des MCO aux données. Nous pouvons également nous demander ce qu'il advient du R^2 lorsque l'unité de mesure de la variable indépendante ou dépendante change. Sans avoir besoin d'algèbre, nous pouvons en deviner la réponse : la qualité d'ajustement du modèle ne dépend pas des unités de mesure des variables. Par exemple, l'ampleur de la variation des salaires de PDG qui est expliquée par le rendement sur capitaux propres, ne doit naturellement pas dépendre du fait que le salaire est mesuré en dollars ou en milliers de dollars ; ou que le rendement sur fonds propres est donné en pourcentage ou sous la forme décimale. Une preuve mathématique de cette intuition existe : en utilisant la définition du R^2, on peut montrer que le R^2 est insensible aux changements d'unités de y ou de x.

Tenir compte de la non-linéarité dans une régression simple

Jusqu'ici, nous nous sommes focalisés sur des relations *linéaires* entre la variable dépendante et la variable indépendante. Comme nous l'avons mentionné au chapitre 1, les relations linéaires ne sont pas généralisables à l'ensemble des applications économiques. Il est néanmoins relativement aisé d'incorporer différentes formes de non-linéarité dans un modèle de régression simple en exprimant les variables dépendante et indépendante de manière appropriée. À ce stade, nous allons envisager deux possibilités que l'on retrouve fréquemment dans les travaux empiriques.

Dans la littérature empirique consacrée aux sciences sociales, vous aurez souvent l'occasion d'analyser des modèles de régression dont la variable dépendante est exprimée sous la forme logarithmique. Comment peut-on justifier ce choix ? Rappelez-vous le modèle « salaire-éducation » dans lequel nous avions régressé le salaire horaire sur le nombre d'années d'études. Nous avions obtenu une estimation de la pente égale à 0,54 [voir l'équation (2.27)], ce qui signifiait que chaque année d'instruction supplémentaire conduisait à une augmentation estimée du salaire horaire de 54 cents. Autrement dit, étant donné la nature linéaire de l'équation (2.27), 54 cents représente l'augmentation du salaire horaire quel que soit le niveau d'instruction initial, qu'il s'agisse de la première année d'études ou, par exemple, de la vingtième. Prendre le raccourci de la fonction linéaire peut donc nous amener à mal interpréter la véritable relation qui existe entre le salaire horaire et le niveau d'instruction.

Il est possible d'envisager une spécification alternative de la manière dont évolue le salaire horaire en fonction des années d'études : chaque année supplémentaire d'instruction peut augmenter le salaire d'un *pourcentage* constant. Par exemple, le niveau d'études peut passer de 5 à 6 ans et induire une augmentation du salaire de 8 % ; dans ce cas, une augmentation du niveau d'instruction de 11 à 12 ans conduira également à une augmentation de 8 %. Lorsque l'effet de la variable indépendante sur la variable dépendante est (approximativement) constant en pourcentage, le modèle peut s'écrire de la manière suivante :

$$\log(wage) = \beta_0 + \beta_1\, educ + u, \qquad \textbf{[2.42]}$$

où log(.) représente le logarithme naturel (ou népérien), en base e. (Voir l'annexe A pour une revue de la fonction logarithmique.) En particulier, si $\Delta u = 0$, alors

$$\%\Delta wage = (100\ \beta_1)\ \Delta educ \qquad \textbf{[2.43]}$$

Pour obtenir la variation en pourcentage de *wage* étant donné la variation en années de *educ*, nous multiplions β_1 par 100. Puisque chaque année supplémentaire d'instruction conduit à la même variation en pourcentage de *wage*, la variation de *wage* en unités monétaires (USD) *augmente* lorsque le niveau d'instruction s'élève. En d'autres termes, (2.42) implique que le « rendement » d'une année d'instruction supplémentaire est *croissant* s'il est mesuré en USD et *constant* s'il est mesuré en pourcentage. En utilisant la transformation exponentielle, (2.42) s'écrit : $wage = \exp(\beta_0 + \beta_1 educ + u)$. Cette équation est représentée sur la figure 2.6, avec $u = 0$.

L'utilisation d'une régression pour estimer un modèle comme celui de (2.42) ne pose aucun problème. Il suffit de définir la variable dépendante, y, telle que $y = \log(wage)$. La variable indépendante est représentée par $x = educ$. La méthode des MCO fonctionne de la même manière : la constante et la pente de la droite sont estimées à l'aide des formules (2.17) et (2.19). En d'autres termes, nous pouvons obtenir $\hat{\beta}_0$ et $\hat{\beta}_1$ par les MCO en régressant $\log(wage)$ sur *educ*, sans aucune difficulté supplémentaire.

Figure 2.6 $wage = exp(\beta_0 + \beta_1 educ)$, *avec* $\beta_1 > 0$.

EXEMPLE 2.10
Une équation du salaire en log

Si nous utilisons les données de l'exemple 2.4 avec une fonction logarithmique pour la variable dépendante, nous obtenons la relation suivante :

$$\widehat{\log(wage)} = 0{,}581 + 0{,}083\ educ$$
$$n = 526,\ R^2 = 0{,}186. \qquad \textbf{[2.44]}$$

Lorsque le coefficient de *educ* est multiplié par 100, il s'interprète en pourcentage. $\widehat{wage}$ augmente de 8,3 % pour chaque année supplémentaire d'instruction. Selon les économistes, ce pourcentage mesure « le rendement de l'éducation », c'est-à-dire l'augmentation en pourcentage du salaire provenant d'une année d'instruction supplémentaire.

Il est important de se rappeler que la principale raison motivant l'utilisation du log de *wage* dans (2.42) est d'imposer un effet constant en pourcentage (*et donc variable en unité monétaire*) du niveau d'instruction sur *wage*. Après estimation du modèle, le log de *wage* n'a pas d'intérêt particulier au niveau de l'interprétation des résultats. Par exemple, il est faux de conclure qu'une année d'études en plus augmente log(*wage*) de 8,3 %. C'est le salaire horaire qui augmente de 8,3 %.

La constante dans (2.44) n'est pas utile car elle donne la valeur estimée de log(*wage*), et non de *wage*, lorsque *educ* = 0. Le R^2 indique que *educ* explique environ 18,6 % de la variation de log(*wage*), et non de *wage*. Enfin, l'équation (2.44) ne tient pas nécessairement compte de tous les aspects non linéaires qui peuvent exister dans la relation entre salaire et niveau d'études. S'il existe un « effet diplôme », alors la douzième année d'instruction, correspondant à la sortie du secondaire supérieur, pourrait conduire à une plus grande augmentation du salaire que la onzième, par exemple. Vous apprendrez à tenir compte de cette forme de non-linéarité au chapitre 7.

Il est également fréquent de recourir au logarithme naturel pour estimer un **modèle à élasticité constante**.

EXEMPLE 2.11
Salaire des PDG et chiffre d'affaires

Estimons un modèle à élasticité constante pour expliquer le salaire des PDG (*salary*) par le chiffre d'affaires de l'entreprise (*sales*) qu'ils dirigent. La base de données est identique à celle que nous avons utilisée dans l'exemple 2.3, à la seule différence que nous expliquons le salaire (*salary*) par le chiffre d'affaires (*sales*). Soit *sales*, le chiffre d'affaires de l'entreprise, égal au total des ventes mesurées en millions de dollars américains. Le modèle à élasticité constante est

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u \qquad [2.45]$$

où β_1 mesure l'élasticité de *salary* par rapport à *sales*. Il s'agit bien d'un modèle de régression simple dans lequel la variable dépendante est $y = \log(salary)$ et la variable indépendante est $x = \log(sales)$. L'estimation du modèle par les MCO donne

$$\widehat{\log(salary)} = 4{,}822 + 0{,}257 \log(sales)$$
$$n = 209,\ R^2 = 0{,}211. \qquad [2.46]$$

Le coefficient de log(*sales*) représente l'estimation de l'élasticité de *salary* par rapport à *sales*. Une augmentation d'1 % du chiffre d'affaires conduit à une augmentation du salaire du PDG d'environ 0,257 %, ce qui correspond à l'interprétation habituelle d'une élasticité.

Les deux formes fonctionnelles que nous venons d'utiliser dans cette section vont réapparaître souvent dans les autres chapitres du livre. Nous avons sélectionné les modèles basés sur l'emploi du log naturel car ils sont fréquemment utilisés dans les travaux d'analyse empirique. Leur interprétation ne sera pas vraiment différente lorsque nous passerons au cas de la régression multiple.

Il est également instructif d'étudier les conséquences sur la constante et la pente de la droite d'un changement d'unité de mesure de la variable dépendante quand celle-ci est exprimée sous la forme logarithmique. Étant donné que le passage à une forme en log se traduit par une interprétation en pourcentage, ce changement d'unité de mesure ne doit avoir *aucune* conséquence sur la valeur de la pente. Nous pouvons le vérifier en utilisant un facteur d'échelle égal à c_1 pour chaque observation i de y_i. Si l'équation de départ est $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$ et que nous ajoutons $\log(c_1)$ à la gauche et à la droite du signe d'égalité, nous obtenons $\log(c_1) + \log(y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$, soit $\log(c_1 y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$. (Rappelez-vous

que la somme de deux logarithmes naturels est égale au log de leur produit, comme indiqué à l'annexe A.) La pente de la droite reste bien égale à β_1. Par contre, la constante est désormais égale à $\log(c_1) + \beta_0$. De manière équivalente, si la variable indépendante x est en log et que nous modifions son unité de mesure (avant de recourir à la forme logarithmique), la pente restera la même mais la constante sera modifiée. Dans le problème 9 à la fin du chapitre, vous devrez le démontrer.

Tableau 2.3 *Synthèse des formes fonctionnelles ayant recours au log naturel*

Modèle	Variable dépendante	Variable indépendante	Interprétation de β_1
Niveau-niveau	y	x	$\Delta y = \beta_1 \Delta x$
Niveau-log	y	$\log(x)$	$\Delta y = (\beta_1 / 100) \%\Delta x$
Log-niveau	$\log(y)$	x	$\%\Delta y = (100\ \beta_1) \Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Nous pouvons identifier quatre combinaisons de formes fonctionnelles, selon que nous décidons de conserver une variable sous sa forme initiale (en niveau) ou de l'utiliser sous sa forme logarithmique (en log). Dans le tableau 2.3, x et y représentent les variables dans leur forme initiale. Le modèle dont les variables dépendante et indépendante sont respectivement y et x, est désigné sous le terme de modèle « niveau-niveau », car chaque variable est présente sous sa forme initiale ou en niveau, sans qu'aucune transformation n'ait eu lieu. Le modèle dont la variable dépendante est $\log(y)$ et la variable indépendante est x, est le modèle dit « log-niveau ». Nous n'entamons pas de discussion plus poussée du modèle « niveau-log » à ce stade pour la simple raison qu'il est moins utilisé dans la pratique. De toute manière, nous en verrons plusieurs exemples dans les chapitres qui suivent.

La dernière colonne du tableau 2.3 est consacrée à l'interprétation de β_1. Dans le modèle log-niveau, on considère parfois que $100\ \beta_1$ représente la **semi-élasticité** de y par rapport x. Comme nous l'avons mentionné dans l'exemple 2.11, dans le modèle log-log, β_1 est l'**élasticité** de y par rapport à x. Le tableau 2.3 mérite toute votre attention ; nous nous y référerons souvent dans les autres chapitres.

La signification du qualificatif « linéaire »

Le modèle de régression simple que nous avons étudié dans ce chapitre est également désigné sous le terme de modèle de régression *linéaire* simple. Pourtant, comme nous venons de le voir, ce modèle permet de tester plusieurs formes de non-linéarité. Dès lors, qu'entend-on exactement par « linéaire » dans un tel modèle ? Nous pouvons répondre à cette question en nous concentrant sur l'équation de départ (2.1) : $y = \beta_0 + \beta_1 x + u$. L'élément clé est la linéarité de l'équation dans ses *paramètres* β_0 et β_1. Il n'existe aucune restriction sur la forme que doit prendre y ou x. Comme nous l'avons vu aux exemples 2.10 et 2.11, y et x peuvent être les logarithmes naturels de variables quelconques ; c'est relativement fréquent dans les travaux empiriques. Il n'y a d'ailleurs aucune raison de nous borner à l'utilisation de la fonction logarithmique. Par exemple, rien ne nous empêche d'utiliser le modèle de régression linéaire simple pour estimer un modèle tel que $cons = \beta_0 + \beta_1 \sqrt{inc} + u$, où *cons* est la consommation annuelle et *inc* est le revenu annuel.

Bien que la méthode des MCO ne soit pas affectée par la manière dont les variables y et x sont définies, l'interprétation des coefficients en dépend. Or, le succès d'une étude empirique repose davantage sur l'aptitude à bien interpréter les coefficients que sur celle qui consiste à trouver l'équation (2.19) nécessaire à l'estimation de la pente. Nous aurons l'occasion de nous entraîner à cet art lorsqu'il s'agira d'interpréter les résultats de régressions multiples estimées par les MCO.

De nombreux modèles *ne* peuvent *pas* être conceptualisés sous la forme d'un modèle de régression linéaire, car ils ne sont pas linéaires dans leurs paramètres. Un exemple est : $cons = 1 / (\beta_0 + \beta_1 inc) + u$. L'estimation de ce type de modèles nous conduirait à explorer l'univers des *modèles de régressions non linéaires*, ce qui est au-delà de la portée de cet ouvrage. Dans la plupart des applications, il suffit généralement de déterminer un modèle qui rentre dans le cadre de la régression linéaire.

2.5 ESPÉRANCES ET VARIANCES DES ESTIMATEURS DES MCO

Dans la section 2.1, nous avons défini le modèle issu de la population, $y = \beta_0 + \beta_1 x + u$, dont l'hypothèse fondamentale précise que l'espérance de u est nulle, quelle que soit la valeur de x. Dans les sections 2.2, 2.3, et 2.4, nous avons dérivé les propriétés algébriques des MCO. Nous y revenons pour en étudier les propriétés *statistiques*. En d'autres termes, nous considérons désormais $\hat{\beta}_0$ et $\hat{\beta}_1$ comme les *estimateurs* des paramètres β_0 et β_1. Cela implique que nous allons étudier les propriétés des distributions de $\hat{\beta}_0$ et $\hat{\beta}_1$ sur base d'échantillons aléatoires tirés au sein de la population. (L'annexe C inclut une définition des estimateurs et une revue de leurs propriétés fondamentales.)

Absence de biais des estimateurs des MCO

Nous partons de la propriété d'absence de biais des MCO que nous déterminons sur base d'un ensemble restreint d'hypothèses. Pour y faire appel plus facilement par la suite, il est utile de numéroter ces hypothèses en les précédant du préfixe « régression simple » pour régression linéaire simple. La première hypothèse définit le modèle issu de la population.

Lorsque nous avons élaboré le modèle (2.47), nous avons considéré que y, x, et u étaient toutes des variables aléatoires. Nous avons longuement discuté de son interprétation dans la section 2.1 en recourant à plusieurs exemples. Dans la section précédente, nous avons également découvert que (2.47) n'était pas aussi restrictif que nous pouvions le penser au départ ; en déterminant y et x de manière appropriée, il est possible de tester l'existence de plusieurs types de relations non linéaires (dans un modèle à élasticité constante, par exemple).

Hypothèse RLS.1
Linéarité dans les paramètres

Dans le modèle issu de la population, la variable dépendante, y, est liée à la variable indépendante, x, et au terme d'erreur (ou de perturbation), u, comme suit :

$$y = \beta_0 + \beta_1 x + u, \qquad [2.47]$$

où β_0 et β_1 sont respectivement les paramètres de la constante et de la pente au sein de la population.

Notre intérêt porte maintenant sur l'utilisation de données concernant y et x dans le but d'estimer les paramètres β_0 et, plus spécialement, β_1. Nous supposons que nos données sont tirées d'un échantillon aléatoire. (Voir l'annexe C pour une révision des principes de l'échantillonnage aléatoire.)

Hypothèse RLS.2
Échantillonnage aléatoire

Nous disposons d'un échantillon aléatoire de taille n, $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$, tiré de la population sur laquelle repose le modèle (2.47).

Dans les chapitres portant sur les séries chronologiques et les problèmes d'échantillonnage, nous aurons à affronter la difficulté de ne pas pouvoir compter sur cette hypothèse. Même si tous les échantillons constitués à partir de données transversales ne sont pas aléatoires, l'hypothèse est vérifiée pour beaucoup d'entre eux.

Nous pouvons maintenant écrire (2.47) sur base d'un échantillon aléatoire :

$$y_i = \beta_0 + \beta_1 x_i + u_i, \; i = 1, 2, \ldots, n, \qquad \textbf{[2.48]}$$

où u_i représente l'erreur ou la perturbation, propre à l'observation i : par exemple, la personne i, l'entreprise i, la ville i, etc. Par conséquent, u_i représente les facteurs non observés, spécifiques à l'observation i, qui affectent y_i. L'erreur u_i ne doit pas être assimilée au résidu, $\hat{u}_i$, que nous avons défini dans la section 2.3. Par la suite, nous explorerons la relation entre les erreurs et les résidus. Quand il s'agit d'interpréter β_0 et β_1 dans le cadre d'une application, (2.47) est plus instructif ; (2.48) reste néanmoins nécessaire dans le cadre de certaines dérivations statistiques.

La relation (2.48) peut être représentée sur base d'observations propres à un échantillon, comme sur la figure 2.7.

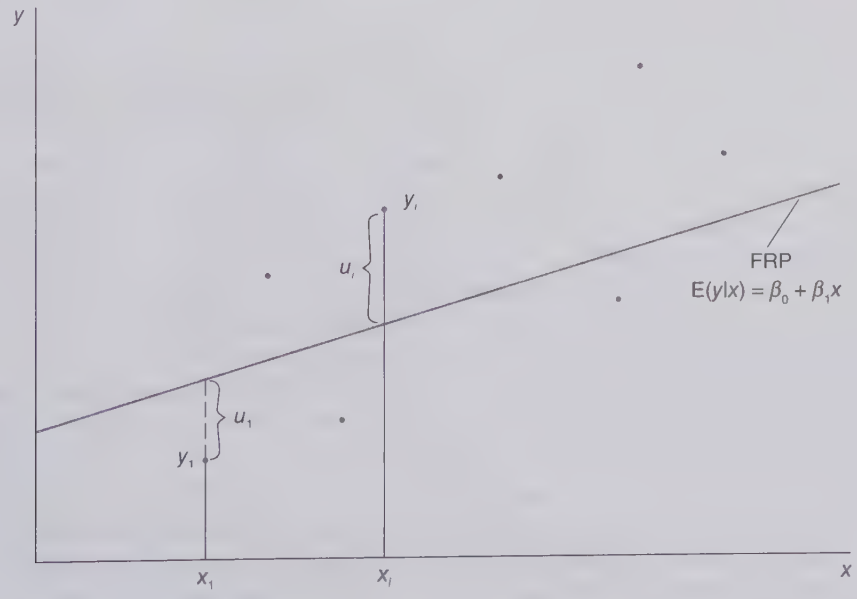


Figure 2.7 *Graphique de* $y = \beta_0 + \beta_1 x_i + u_i$.

Comme nous l'avons vu dans la section 2.2, il est indispensable que la variable indépendante affiche une variation non nulle au sein de l'échantillon pour qu'il soit possible d'estimer la pente et la constante des MCO. Il convient donc d'ajouter une hypothèse concernant la variation de x_i dans notre liste.

Hypothèse RLS.3
Variation de la variable explicative au sein de l'échantillon

Les éléments de x au sein de l'échantillon, à savoir $\{x_i, i = 1, \ldots, n\}$, n'ont pas tous la même valeur.

Il s'agit d'une hypothèse indispensable mais peu contraignante, sur laquelle il est inutile d'insister. Si x varie au sein de la population, il est plus que probable qu'un échantillon aléatoire de x aura également une variation non nulle, à moins que la variation au sein de la population soit minime ou que l'échantillon soit de très petite taille. Un simple examen des statistiques de base sur les x_i permet de le vérifier et de savoir si l'hypothèse RLS.3 est violée : si l'écart-type estimé de x_i est égal à zéro, ce sera le cas ; sinon, l'hypothèse est validée.

Hypothèse RLS.4
Espérance conditionnelle de l'erreur égale à zéro

Le terme d'erreur u affiche une espérance égale à zéro, quelle que soit la valeur de x. Autrement dit,

$$E(u|x) = 0.$$

Enfin, pour obtenir des estimateurs de β_0 et β_1 sans biais, il est impératif d'ajouter l'hypothèse de nullité de l'espérance conditionnelle qui a déjà fait l'objet d'une discussion poussée dans la section 2.1.

Dans le cas d'un échantillon aléatoire, cette hypothèse implique que $E(u_i|x_i) = 0$, pour tout $i = 1, 2, \ldots, n$.

Au-delà de la restriction qu'elle impose sur la relation entre u et x au sein de la population, l'hypothèse d'espérance conditionnelle nulle, combinée à l'hypothèse RLS.2 sur l'échantillonnage aléatoire, permet de recourir à un raccourci technique très commode. Il nous permet de dériver les propriétés statistiques des estimateurs des MCO, *étant donné* les valeurs de x_i dans notre échantillon. Sur un plan plus technique, cette possibilité nous permet de dériver les propriétés statistiques des estimateurs en considérant que les x_i sont *fixes en échantillons répétés* : d'un échantillon à un autre, on considère que les valeurs prises pour chaque x_i restent inchangées. Nous pouvons l'expliquer de la manière suivante. Nous devons obtenir, dans un premier temps, un échantillon de n valeurs pour les x_i, une valeur pour chaque $x_1, x_2, \ldots, x_n$. (Cette procédure peut d'ailleurs être répétée autant de fois qu'on le désire.) *Étant donné* ces valeurs x_i et après voir constitué un échantillon aléatoire de n valeurs pour les u_i, nous pouvons obtenir un échantillon pour la variable y, constitué lui-même de n valeurs, allant de $y_1, y_2, \ldots,$ à y_n. Ensuite, un autre échantillon de y peut être constitué sur la base des *mêmes* valeurs de $x_1, x_2, \ldots, x_n$ (mais différentes valeurs de $u_1, u_2, \ldots, u_n$). L'étape précédente peut à nouveau être répétée en utilisant les mêmes $x_1, x_2, \ldots, x_n$ (mais, à nouveau, différentes valeurs pour $u_1, u_2, \ldots, u_n$) et ainsi de suite.

Ce scénario, qui suppose que les valeurs de x sont fixes en échantillonnage répété, n'est pas très réaliste dans le contexte non expérimental des sciences sociales. Par exemple, dans le cadre de la relation entre salaire et années d'études, il semble absurde de déterminer les années d'éducation à l'avance pour ensuite constituer un échantillon d'individus en fonction de ces différents niveaux d'instruction. La plupart des jeux de données en sciences sociales sont basés sur le principe de l'échantillonnage aléatoire selon lequel les individus sont sélectionnés au hasard avant que ne soient enregistrées leurs caractéristiques propres, c'est-à-dire le salaire et le niveau d'instruction dans le cas qui nous intéresse. Si nous disposons d'un échantillon aléatoire et que nous *faisons l'hypothèse* que $E(u_i|x_i) = 0$, il est vrai que rien ne change sur le plan technique des dérivations statistiques en considérant x_i comme étant non stochastique, c'est-à-dire fixe d'un échantillon à l'autre. Le danger de cette hypothèse de « fixité » est de considérer qu'elle implique

que u_i et x_i sont *à coup sûr* indépendants. En réalité, *l'hypothèse* RLS.4 ne se vérifie pas automatiquement. Elle est d'ailleurs déterminante lorsqu'il s'agit de savoir si la méthode des MCO permet d'obtenir des estimateurs sans biais.

Nous pouvons à présent démontrer que, sous ces hypothèses, les estimateurs sont sans biais. Étant donné que $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$ (voir l'annexe A), l'estimateur de la pente de l'équation (2.19) peut s'écrire sous la forme suivante :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad [2.49]$$

Comme nous étudions le comportement de $\hat{\beta}_1$ dans tous les échantillons possibles, $\hat{\beta}_1$ doit être considéré, à juste titre, comme une variable aléatoire.

Nous pouvons également écrire $\hat{\beta}_1$ en fonction des coefficients de la population et du terme d'erreur en utilisant (2.48) dans (2.49). Nous obtenons

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\text{SCT}_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\text{SCT}_x}, \qquad [2.50]$$

où la variation totale de x_i est égale à $\text{SCT}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$. (Ce n'est pas exactement égal à la variance des x_i au sein de l'échantillon puisque nous n'avons pas divisé par $n - 1$.) En utilisant les propriétés de base de l'opérateur de sommation, le numérateur de $\hat{\beta}_1$ devient

$$\begin{aligned}&\sum_{i=1}^{n}(x_i - \bar{x})\beta_0 + \sum_{i=i}^{n}(x_i - \bar{x})\beta_1 x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i \\ &= \beta_0\sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1\sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i \qquad [2.51]\end{aligned}$$

Comme démontré dans l'annexe A, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ et $\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \text{SCT}_x$. Par conséquent, nous pouvons considérer le numérateur de $\hat{\beta}_1$ comme étant égal à $\beta_1\text{SCT}_x + \sum_{i=1}^{n}(x_i - \bar{x})u_i$. Divisé par le dénominateur, cela donne

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\text{SCT}_x} = \beta_1 + (1/\text{SCT}_x)\sum_{i=1}^{n}d_i u_i, \qquad [2.52]$$

où $d_i = (x_i - \bar{x})$. Nous pouvons voir que l'estimateur $\hat{\beta}_1$ est égal à la somme du coefficient de population pour la pente, β_1, et d'un terme qui correspond à une combinaison linéaire des erreurs $\{u_1, u_2, \ldots, u_n\}$. Étant donné les valeurs de x_i, le caractère aléatoire de $\hat{\beta}_1$ provient uniquement des erreurs. Le fait que ces erreurs ne sont généralement pas égales à zéro explique la différence entre $\hat{\beta}_1$ et β_1.

En se basant sur (2.52), nous pouvons démontrer la première propriété statistique importante des MCO, soit le théorème 2.1.

Théorème 2.1
Absence de biais des MCO

En utilisant les hypothèses RLS.1 à RLS.4,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0, \text{ et } \mathrm{E}(\hat{\beta}_1) = \beta_1, \qquad \textbf{[2.53]}$$

quelle que soit la valeur de β_0 et β_1. En d'autres termes, $\hat{\beta}_0$ est un estimateur sans biais de β_0, et $\hat{\beta}_1$ est un estimateur sans biais de β_1.

PREUVE : Dans cette démonstration, les espérances sont conditionnelles aux valeurs observées pour la variable indépendante au sein de l'échantillon. Autrement dit, les éléments x_i sont donnés ex ante ou connus à l'avance. Puisque SCT_x et d_i sont des fonctions des x_i (et d'eux seuls), ces fonctions ne seront pas stochastiques mais déterministes. Par conséquent, en partant de (2.52) et étant donné $\{x_1, x_2, \ldots, x_n\}$, nous obtenons :

$$\mathrm{E}(\hat{\beta}_1) = \beta_1 + \mathrm{E}[(1/\mathrm{SCT}_x)\sum_{i=1}^{n} d_i u_i] = \beta_1 + (1/\mathrm{SCT}_x)\sum_{i=1}^{n}\mathrm{E}(d_i u_i)$$

$$= \beta_1 + (1/\mathrm{SCT}_x)\sum_{i=1}^{n} d_i\ \mathrm{E}(u_i) = \beta_1 + (1/\mathrm{SCT}_x)\sum_{i=1}^{n} d_i\ 0 = \beta_1.$$

Grâce aux hypothèses RLS.2 et RLS.4, nous avons pu indiquer que la valeur attendue de chaque u_i est nulle, étant donné $\{x_1, x_2, \ldots, x_n\}$. Notez bien que la propriété d'absence de biais est vérifiée quelles que soient les valeurs $\{x_1, x_2, \ldots, x_n\}$. Par conséquent, cette propriété est vérifiée même si nous ne conditionnons pas les espérances par rapport aux $\{x_1, x_2, \ldots, x_n\}$.

La démonstration pour $\hat{\beta}_0$ est évidente. Il suffit de calculer la moyenne de (2.48) par rapport à i pour obtenir $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$ et l'utiliser dans la formule de $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \beta_0 + \beta_1\bar{x} + \bar{u} - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}.$$

En conditionnant le calcul aux valeurs x_i, on obtient

$$\mathrm{E}(\hat{\beta}_0) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)\bar{x}] + \mathrm{E}(\bar{u}) = \beta_0 + \mathrm{E}(\beta_1 - \hat{\beta}_1)\bar{x},$$

vu que $\mathrm{E}(\bar{u}) = 0$ sous les hypothèses RLS.2 et RLS.4. Nous avons également démontré que $\mathrm{E}(\hat{\beta}_1) = \beta_1$, ce qui est équivalent à $\mathrm{E}(\hat{\beta}_1 - \beta_1) = 0$. Dès lors, $\mathrm{E}(\hat{\beta}_0) = \beta_0$. Ces développements sont valides pour n'importe quelle valeur de β_0 ou de β_1 ; nous avons donc réussi à démontrer l'absence de biais pour les estimateurs des MCO.

Gardez bien à l'esprit que la propriété d'absence de biais est une caractéristique des distributions d'échantillonnage de $\hat{\beta}_1$ et $\hat{\beta}_0$, ce qui ne nous dit rien sur *l'estimation* que nous pouvons obtenir à partir d'un échantillon donné. Si cet échantillon est représentatif de la population, nous pouvons espérer que l'estimation, soit $\hat{\beta}_1$ (ou $\hat{\beta}_0$), sera proche de la valeur dans la population, soit β_1 (ou β_0). Comme il nous est impossible d'observer cette valeur « vraie », nous ne sommes *jamais* certains que l'estimation s'y trouve à proximité. Il existe toujours un risque que nous obtenions un échantillon atypique et, par conséquent, une estimation éloignée de la « vraie » valeur. Si vous désirez explorer davantage le sujet des estimateurs sans biais, lisez l'annexe C, en particulier l'exercice de simulation du tableau C.1 qui illustre le concept d'absence de biais.

En règle générale, la propriété d'absence de biais est violée dès que l'une des quatre hypothèses ne tient pas. Il est donc important d'évaluer le bien-fondé de ces hypothèses à chaque fois que les MCO sont utilisés en pratique. L'hypothèse RLS.1 exige que la relation entre y et x soit linéaire (dans les paramètres), en tenant compte d'un terme d'erreur additif. Il est clair que cette hypothèse peut être violée. Nous savons néanmoins qu'il est possible de tester des relations non linéaires instructives en exprimant y et x sous une forme adéquate. L'estimation de relations non linéaires plus complexes exige l'utilisation de méthodes plus sophistiquées qui sortent du cadre d'analyse de cet ouvrage.

Lorsque nous étudierons les séries chronologiques, nous serons contraints d'assouplir l'hypothèse RLS.2, celle concernant l'échantillonnage aléatoire. Il arrive également qu'un échantillon constitué à partir de données transversales ne soit pas représentatif de la population sous-jacente. Il existe aussi des bases de données dans lesquelles certaines catégories de la population sont délibérément surpondérées. Nous aborderons le sujet de l'échantillonnage dirigé aux chapitres 9 et 17.

Comme nous en avons déjà discuté, l'hypothèse RLS.3 est presque toujours vérifiée. Sans elle, il serait impossible de calculer les estimateurs des MCO.

L'hypothèse qui mérite une plus grande attention est RLS.4. Si RLS.4 est vérifiée, les estimateurs des MCO sont sans biais. Inversement, si RLS.4 est violée, les estimateurs seront généralement biaisés. Il est d'ailleurs possible de déterminer le signe et l'ampleur du biais, comme nous le verrons au chapitre 3.

Le risque que x soit corrélé avec u représente presque toujours un sujet de préoccupation dans les régressions simples basées sur des données non expérimentales, telles que celles utilisées en sciences sociales. Nous l'avions déjà souligné dans la section 2.1 en recourant à plusieurs exemples. Si le terme d'erreur d'une régression simple contient des facteurs qui influencent y, tout en étant corrélés avec x, alors le résultat de cette régression sera biaisé en raison de l'existence d'une *corrélation fallacieuse* entre y et x. Alors que la relation entre y et x nous semble valide et significative, elle s'explique par la relation qui existe entre y et les autres facteurs non observés inclus dans u, qui sont également et malencontreusement corrélés avec x.

EXEMPLE 2.12
Performance des étudiants en maths et distribution de repas scolaires subventionnés par l'État

La variable *math*10 correspond au pourcentage des élèves âgés d'une quinzaine d'années qui ont réussi leur examen de mathématiques. (Ces étudiants ont atteint le « grade 10 », aux États-Unis.) Imaginez que vous désirez estimer l'effet sur le taux de réussite à cet examen d'un programme de distribution à l'école de repas subventionnés. On pourrait s'attendre à ce que le programme ait un effet positif sur la performance de ces élèves, toutes choses étant égales par ailleurs : si l'effet de tous les autres facteurs influençant le taux de réussite est neutralisé, un élève, trop pauvre pour pouvoir manger régulièrement sur le temps de midi, a plus de chance de réussir son examen suite à la distribution de repas subventionnés. En considérant que la variable *lnchprg* correspond au pourcentage d'élèves qui ont accès au programme de distribution, le modèle de régression simple peut s'écrire

$$math10 = \beta_0 + \beta_1 lnchprg + u \qquad [2.54]$$

où u incorpore toutes les autres caractéristiques propres aux élèves et aux établissements scolaires, qui peuvent influencer le taux de réussite scolaire. Sur base du jeu de données MEAP93 portant sur 408 écoles secondaires du Michigan au cours de l'année scolaire 1992–1993, nous obtenons

$$\widehat{math10} = 32{,}14 - 0{,}319\ lnchprg$$
$$n = 408,\ R^2 = 0{,}171.$$

Cette équation indique que le pourcentage d'élèves ayant réussi l'examen de math *diminue* de 3,2 points de pourcentage lorsque le pourcentage d'élèves ayant accès au programme de distribution de repas subventionnés augmente de 10 points de pourcentage. Faut-il en conclure qu'un taux de participation plus élevé à ce programme *conduit* à une moins bonne performance ? La réponse est non, très vraisemblablement. Une meilleure explication de ce résultat surprenant est que le terme d'erreur u de l'équation (2.54) est corrélé avec la variable *lnchprg*. En fait, u comprend des facteurs qui peuvent être fortement (et positivement) corrélés avec la variable explicative *lnchprg*, comme le pourcentage des élèves de l'établissement qui vivent sous le seuil de pauvreté. Il y a aussi la qualité de l'enseignement et celle des ressources matérielles que l'établissement offre à ses élèves. Ces variables sont également comprises dans u et sont susceptibles d'être (négativement) corrélées avec *lnchprg*. Certes, comme l'estimation – 0,319 n'est propre qu'à cet échantillon, sa nature atypique pourrait expliquer ce résultat surprenant. Le signe et l'ampleur de cette estimation nous amènent néanmoins à penser que u et x sont corrélés, biaisant ainsi les résultats de cette régression simple.

L'omission d'une variable n'est pas la seule raison pour laquelle x est corrélée avec u dans le modèle de régression simple. Comme cette problématique se présente également dans le cadre des modèles de régression multiple, nous en traiterons plus systématiquement par la suite.

Variances des estimateurs des MCO

Nous avons vu que la distribution d'échantillonnage de $\hat{\beta}_1$ est centrée sur β_1 ($\hat{\beta}_1$ est sans biais). Il importe maintenant de savoir dans quelle mesure $\hat{\beta}_1$ sera éloigné de β_1 en moyenne. Cela nous permettra, entre autres, de sélectionner le meilleur estimateur parmi tous les estimateurs sans biais ou, à tout le moins, parmi un large éventail d'estimateurs. La mesure de dispersion la plus fréquente pour une distribution telle que celle de $\hat{\beta}_1$ (et $\hat{\beta}_0$) est la variance ou sa racine carrée, l'écart-type. (Voir l'annexe C pour une discussion plus détaillée.)

Il s'avère que la variance des estimateurs des MCO peut être calculée sous les hypothèses RLS.1 à RLS.4 mais leur formulation reste compliquée. Nous allons plutôt ajouter une hypothèse qui s'applique traditionnellement à l'analyse en coupe transversale. Sous cette hypothèse, la variance de l'erreur u, conditionnelle à x, est constante. Il s'agit de l'hypothèse d'**homoscédasticité** ou de « variance constante ».

Hypothèse RLS.5
Homoscédasticité

La variance de l'erreur u est constante, quelle que soit la valeur de x. En d'autres termes,

$$\text{Var}(u|x) = \sigma^2.$$

Il est important de souligner que l'hypothèse d'homoscédasticité est clairement différente de l'hypothèse selon laquelle l'espérance conditionnelle est nulle, $\text{E}(u|x) = 0$. L'hypothèse RLS.4 concerne la *valeur attendue* de u, alors que l'hypothèse RLS.5 concerne la *variance* de u (toutes deux conditionnelles à x). Notez bien que nous avons démontré la propriété d'absence de biais sans recourir à l'hypothèse RLS.5 : l'hypothèse d'homoscédasticité ne joue aucun rôle lorsqu'il s'agit de prouver que $\hat{\beta}_0$ et $\hat{\beta}_1$ sont sans biais. Nous ajoutons l'hypothèse RLS.5 parce qu'elle simplifie le calcul de la *variance* de $\hat{\beta}_0$ et $\hat{\beta}_1$ et parce qu'elle permet aux estimateurs des MCO d'afficher certaines propriétés d'efficacité désirables, comme nous le verrons au chapitre 3. Si nous avions directement supposé que u et x étaient *indépendants*, la distribution de u, étant donné x, n'aurait évidemment pas dépendu de x ; nous aurions obtenu à la fois que $\text{E}(y|x) = \text{E}(u) = 0$ et

que $\text{Var}(u|x) = \sigma^2$. Malheureusement, l'indépendance, qui ne se limite pas à l'absence de corrélation linéaire entre deux variables, est une hypothèse trop restrictive dans certains cas.

Étant donné que $\text{Var}(u|x) = \text{E}(u^2|x) - [\text{E}(u|x)]^2$ et que $\text{E}(u|x) = 0$, $\sigma^2 = \text{E}(u^2|x)$. Cela implique que σ^2 est aussi égale à l'espérance *inconditionnelle* de u^2 (puisque σ^2 est une constante). Par conséquent, $\sigma^2 = \text{E}(u^2) = \text{Var}(u)$, puisque $\text{E}(u) = 0$. En d'autres termes, σ^2 est la variance *inconditionnelle* de u ; σ^2 est souvent dénommée **la variance de l'erreur** ou la variance des perturbations. La racine carrée de σ^2, σ, représente l'écart-type du terme d'erreur. Lorsque σ^2 prend une valeur élevée, la distribution des facteurs non observés qui affectent y est moins resserrée autour de la moyenne : elle affiche une dispersion plus grande.

Il est souvent utile d'exprimer les hypothèses RLS.4 et RLS.5 en fonction de l'espérance et de la variance conditionnelles de y :

$$\text{E}(y|x) = \beta_0 + \beta_1 x. \qquad \textbf{[2.55]}$$

$$\text{Var}(y|x) = \sigma^2. \qquad \textbf{[2.56]}$$

L'espérance conditionnelle de y, étant donné x, est linéaire en x ; par contre, la variance de y, étant donné x, est constante. Ces deux résultats sont représentés graphiquement sur la figure 2.8 en supposant que $\beta_0 > 0$ et $\beta_1 > 0$.

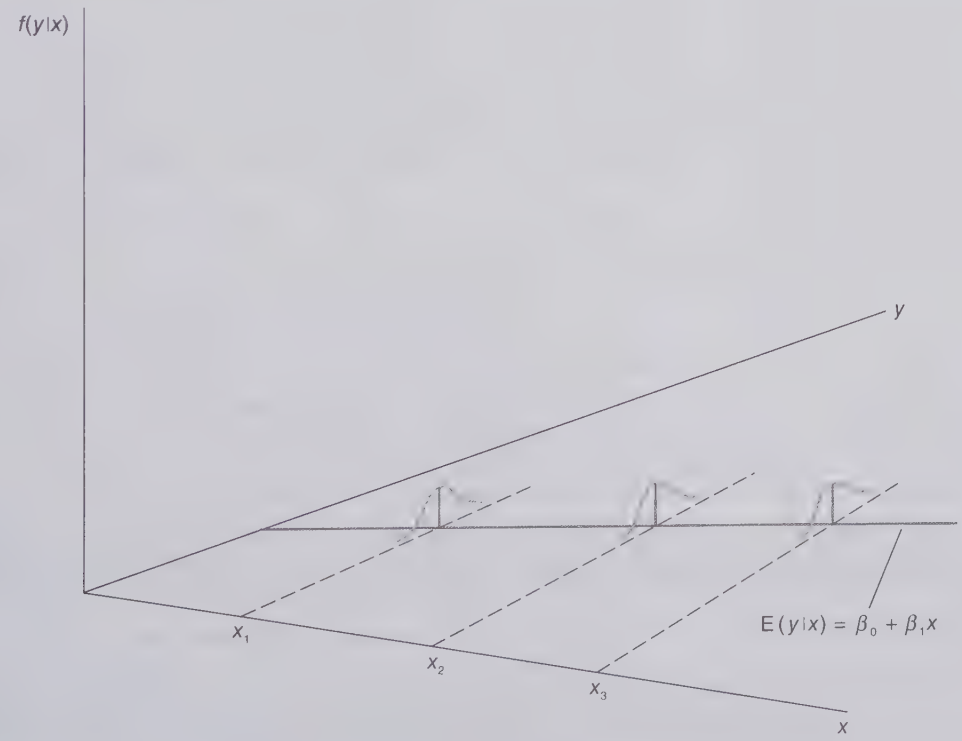


Figure 2.8 *Le modèle de régression simple sous l'hypothèse d'homoscédasticité.*

Lorsque $\text{Var}(u|x)$ dépend de x, le terme d'erreur souffre d'hétéroscédasticité (ou d'une variance qui n'est pas constante). Puisque $\text{Var}(u|x) = \text{Var}(y|x)$, l'hétéroscédasticité est présente chaque fois que $\text{Var}(y|x)$ est une fonction de x.

EXEMPLE 2.13
Hétéroscédasticité dans l'équation sur le salaire

Si nous voulons obtenir un estimateur sans biais de l'effet *ceteris paribus* de *educ* sur *wage*, nous devons poser l'hypothèse que $E(u|educ) = 0$, ce qui conduit à $E(wage|educ) = \beta_0 + \beta_1$. Si nous faisons appel à l'hypothèse d'homoscédasticité, alors $Var(u|educ) = \sigma^2$ ne dépend pas du niveau d'éducation, ce qui revient à écrire que $Var(wage|educ) = \sigma^2$. Sous ces deux hypothèses, le salaire moyen peut naturellement augmenter en fonction du niveau d'instruction – c'est précisément ce taux de croissance que nous cherchons à estimer – mais les écarts de salaire autour du salaire moyen doivent rester inchangés, quel que soit le niveau d'instruction. Est-ce vraiment réaliste ? Les personnes dont le niveau d'instruction est élevé ont généralement des opportunités d'emploi et des centres d'intérêt plus variés, ce qui se traduit par une plus grande variabilité du salaire. À l'opposé, les personnes dont le niveau d'instruction est rudimentaire décrochent des emplois plus standardisés et gagnent souvent le salaire minimum ; les écarts de salaire sont beaucoup plus faibles à ce niveau d'instruction. Cet état de fait est représenté à la figure 2.9. Que cela se traduise par une violation de l'hypothèse RLS.5 est, en fin de compte, une question d'ordre empirique. Au chapitre 8, nous étudierons les tests qui nous permettront de répondre à cette question.

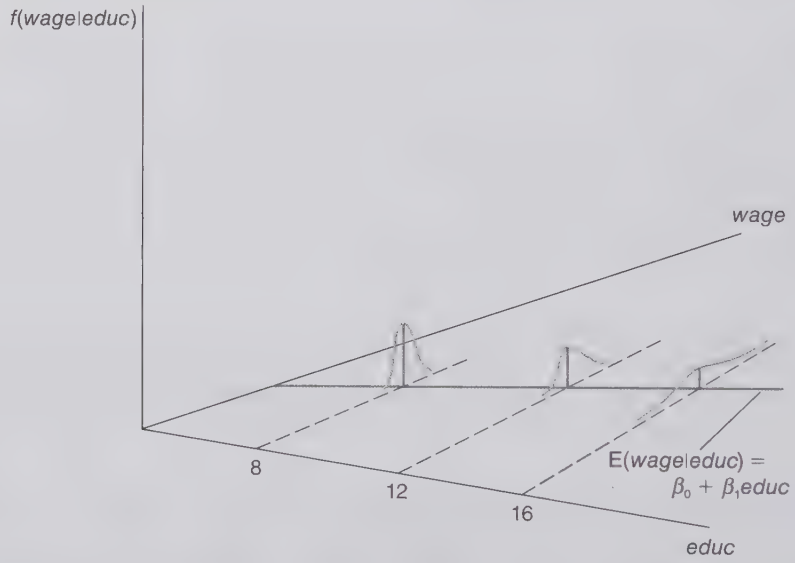


Figure 2.9 *Var(wage|educ) est une fonction croissante de* educ.

Après avoir défini l'hypothèse d'homoscédasticité, nous pouvons la démontrer.

Théorème 2.2
Variances d'échantillonnage des estimateurs

Sous les hypothèses RLS.1 à RLS.5,

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2/\text{SCT}_x \qquad [2.57]$$

et

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2 \, n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad [2.58]$$

en soulignant que les variances sont conditionnelles aux valeurs $\{x_1, \ldots, x_n\}$ observées dans l'échantillon.

PREUVE : Nous allons nous contenter de dériver la formule pour $\text{Var}(\hat{\beta}_1)$; l'autre dérivation est abordée au problème 10. Le point de départ est l'équation (2.52) : $\hat{\beta}_1 = \beta_1 + (1/\text{SCT}_x)\sum_{i=1}^{n} d_i u_i$. Notons tout d'abord que β_1 est une constante. Vu que notre analyse est conditionnelle aux x_i, SCT_x et $d_i = x_i - \bar{x}$ sont non aléatoires. Par ailleurs, étant donné que les u_i sont des variables aléatoires indépendantes en i (grâce à l'échantillonnage aléatoire), la variance de la somme est égale à la somme des variances. Sur cette base, nous obtenons :

$$\text{Var}(\hat{\beta}_1) = (1/\text{SCT}_x)^2 \, \text{Var}\left(\sum_{i=1}^{n} d_i u_i\right) = (1/\text{SCT}_x)^2 \left(\sum_{i=1}^{n} d_i^2 \, \text{Var}(u_i)\right)$$

$$= (1/\text{SCT}_x)^2 \left(\sum_{i=1}^{n} d_i^2 \sigma^2\right) [\text{car } \text{Var}(u_i) = \sigma^2 \text{ pour tout } i]$$

$$= \sigma^2 \, (1/\text{SCT}_x)^2 \left(\sum_{i=1}^{n} d_i^2\right) = \sigma^2 \, (1/\text{SCT}_x)^2 \, \text{SCT}_x = \sigma^2/\text{SCT}_x,$$

ce qui correspond à ce que nous voulions démontrer.

Les équations (2.57) et (2.58) sont les deux formules classiques auxquelles l'analyse de la régression simple recourt le plus souvent. Ces formules ne sont pas valides en présence d'hétéroscédasticité. Cela aura une importance particulière lorsque nous étudierons les intervalles de confiance et les tests d'hypothèse dans le cadre de la régression multiple.

Dans la plupart des cas, notre attention se porte sur $\text{Var}(\hat{\beta}_1)$. Il est facile d'expliquer la manière dont la variance de cet estimateur dépend de la variance de l'erreur, σ^2, et de la variation totale au sein de $\{x_1, x_2, \ldots, x_n\}$, SCT_x. Plus la variance de l'erreur est élevée, plus $\text{Var}(\hat{\beta}_1)$ l'est également. Ce résultat est logique : une plus grande variation dans les facteurs non observés rend l'estimation de β_1 moins précise. Par contre, une plus grande variabilité dans la variable explicative est désirable : plus les variations entre les x_i sont grandes, plus la variance de $\hat{\beta}_1$ diminue. Ce résultat correspond également à notre intuition : plus l'échantillon de la variable explicative contient un large éventail de valeurs différentes pour les x_i, plus il est facile de caractériser la relation entre $E(y|x)$ et x et, par conséquent, d'estimer β_1. Inversement, si l'échantillon de x ne contient que des valeurs proches, la variation dans x est faible et il est difficile de déterminer la manière dont y varie en fonction de x. Enfin, il existe une relation positive entre la taille de l'échantillon et la variation *totale* de x. Plus l'échantillon contient d'observations, plus la somme des écarts au carré entre chaque x_i et la moyenne est grande. Par conséquent, l'utilisation d'un plus grand échantillon permet de diminuer la variance de $\hat{\beta}_1$.

Cette analyse montre que nous devrions choisir une série de x_i la plus dispersée possible, en supposant que nous en ayons la possibilité. C'est parfois le cas lorsque nous travaillons avec des données expérimentales. En sciences sociales, c'est un luxe : nous sommes plutôt contraints d'accepter les x_i que l'échantillonnage aléatoire a généré. Dans certains cas, il est possible d'agrandir la taille de l'échantillon aléatoire, à condition que cela ne soit pas trop coûteux.

Pour aller plus loin 2.5

Lorsqu'il s'agit d'estimer β_0, montrez que l'idéal est d'avoir $\bar{x} = 0$. Que devient $\text{Var}(\hat{\beta}_0)$ dans un tel cas de figure ? [*Astuce* : Quelles que soient les valeurs des x_i dans l'échantillon, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, l'égalité ne valant que lorsque $\bar{x} = 0$.]

Lorsqu'il s'agira de construire des intervalles de confiance et de calculer les statistiques liées aux tests d'hypothèse, nous aurons besoin des écarts-types de $\hat{\beta}_1$ et $\hat{\beta}_0$, soit $\sigma(\hat{\beta}_1)$ et $\sigma(\hat{\beta}_0)$. En anglais, on les dénomme "standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_0$", soit $\text{sd}(\hat{\beta}_1)$ et $\text{sd}(\hat{\beta}_0)$. Ils sont égaux à la racine carrée des variances telles que décrites en (2.57) et (2.58). En particulier, $\sigma(\hat{\beta}_1) = \sigma / \sqrt{\text{SCT}_x}$, où σ est la racine carrée de σ^2 et $\sqrt{\text{SCT}_x}$ est la racine carrée de SCT_x.

L'estimation de la variance de l'erreur

Les formules (2.57) et (2.58) nous permettent d'identifier les facteurs qui influencent $\text{Var}(\hat{\beta}_1)$ et $\text{Var}(\hat{\beta}_0)$. L'inconvénient est qu'elles contiennent des inconnues, sauf dans le cas extrêmement rare où σ^2 est observable. Nous pouvons néanmoins évaluer σ^2 en utilisant des données, le but ultime étant d'estimer $\text{Var}(\hat{\beta}_1)$ et $\text{Var}(\hat{\beta}_0)$.

Le moment est venu de souligner la différence entre les *erreurs* (ou perturbations) et les *résidus*. Cette distinction est capitale lorsqu'il s'agit de déterminer un estimateur de σ^2. L'équation (2.48) nous montre comment il convient d'écrire le modèle issu de la population en fonction d'observations échantillonnées aléatoirement, soit $y_i = \beta_0 + \beta_1 x_i + u_i$ où u_i est l'erreur relative à l'observation i. Nous pouvons également exprimer y_i en fonction de sa valeur ajustée et de son résidu. En suivant (2.32), on obtient : $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$. La comparaison de ces deux équations nous montre que l'erreur apparaît dans l'équation relative à la *population*, celle qui inclut les paramètres de la population, β_0 et β_1. Quant aux résidus, ils font partie de l'équation *estimée*, celle qui incorpore $\hat{\beta}_0$ et $\hat{\beta}_1$. Les erreurs ne peuvent jamais être observées alors que les résidus sont calculés à partir d'une base de données.

Nous pouvons utiliser l'équation (2.32) et (2.48) pour exprimer les résidus en fonction des erreurs :

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

ou encore

$$\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i. \qquad \textbf{[2.59]}$$

On constate que $\hat{u}_i$ n'est pas égale à u_i. C'est la différence *attendue* entre ces deux termes qui est égale à zéro, comme c'est le cas entre $\hat{\beta}_0$ et β_0, d'une part, et $\hat{\beta}_1$ et β_1, d'autre part.

Maintenant que nous comprenons la différence entre les erreurs et les résidus, nous pouvons estimer σ^2. Comme $\sigma^2 = \text{E}(u^2)$, on pourrait penser que $n^{-1} \sum_{i=1}^{n} u_i^2$ est un estimateur sans biais de σ^2. Ce n'est malheureusement pas le cas pour la simple raison qu'il est impossible d'observer les erreurs u_i. La bonne nouvelle est que nous disposons d'estimations pour les u_i, à savoir les résidus des MCO, $\hat{u}_i$. Si nous remplaçons les erreurs par les résidus, nous obtenons $n^{-1} \sum_{i=1}^{n} \hat{u}_i^2 = \text{SCR} / n$. Il s'agit bien d'un « vrai estimateur » car il offre

une règle de calcul qui s'applique à n'importe quel échantillon de données. L'inconvénient de cet estimateur est qu'il est biaisé, bien que ce biais soit négligeable lorsque n est grand. Comme le calcul de l'estimateur sans biais n'est pas compliqué, nous allons y recourir.

L'estimateur SCR / n est biaisé pour la principale raison qu'il ne tient pas compte de deux contraintes que les résidus des MCO doivent respecter. Ces contraintes sont données par les deux conditions de premier ordre des MCO :

$$\sum_{i=1}^{n}\hat{u}_i = 0, \quad \sum_{i=1}^{n}x_i\hat{u}_i = 0. \qquad \textbf{[2.60]}$$

Une manière d'interpréter ces deux conditions est de considérer que nous perdons deux **degrés de liberté** pour pouvoir les remplir. Si nous connaissons la valeur des $n - 2$ résidus dans notre échantillon, nous sommes contraints de choisir les deux derniers résidus de sorte que les conditions de premier ordre soient satisfaites (2.60). C'est la raison pour laquelle il n'y a que $n - 2$ degrés de liberté dans les résidus, contrairement aux n degrés de liberté dans les erreurs. Si nous décidions de remplacer $\hat{u}_i$ par u_i dans (2.60), ces deux conditions ne seraient plus remplies.

L'estimateur sans biais de σ^2 que nous allons utiliser incorpore l'ajustement relatifs aux degrés de liberté :

$$\hat{\sigma}^2 = \frac{1}{(n-2)}\sum_{i=1}^{n}\hat{u}_i^2 = \text{SCR}/(n-2) \qquad \textbf{[2.61]}$$

(On parle aussi d'erreur quadratique moyenne ou du carré moyen des erreurs, correspondant au MSE, « mean squared error », en anglais. L'estimateur est parfois représenté par s^2, mais nous allons continuer à utiliser la convention qui consiste à placer des « chapeaux » sur les estimateurs.)

Théorème 2.3
Estimation sans biais de σ^2

Sous les hypothèses RLS.1 à RLS.5,

$$\text{E}(\hat{\sigma}^2) = \sigma^2$$

PREUVE : Si nous calculons la moyenne de l'équation (2.59) en fonction des i et que nous tenons compte du fait que la moyenne des résidus des MCO est égale à zéro, nous obtenons : $0 = \bar{u} - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)\bar{x}$. Si cette égalité est soustraite de (2.59), cela donne $\hat{u}_i = (u_i - \bar{u}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$. Par conséquent, $\hat{u}_i^2 = (u_i - \bar{u})^2 + (\hat{\beta}_1 - \beta_1)^2(x_i - \bar{x})^2 - 2(u_i - \bar{u})(\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$. En sommant par rapport à i,

$\sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i=1}^{n}(u_i - \bar{u})^2 + (\hat{\beta}_1 - \beta_1)^2\sum_{i=1}^{n}(x_i - \bar{x})^2 - 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}u_i(x_i - \bar{x})$. Calculons maintenant l'espérance de cette égalité.

La valeur espérée du premier terme de droite est $(n - 1)\sigma^2$, ce que nous démontrons dans l'annexe C. La valeur attendue du deuxième terme est tout simplement égale à σ^2 étant donné que $\text{E}[(\hat{\beta}_1 - \beta_1)^2] = \text{Var}(\hat{\beta}_1) = \sigma^2/\text{SCT}_x$. Enfin, on peut démontrer que le troisième terme s'écrit $2(\hat{\beta}_1 - \beta_1)^2\text{SCT}_x$; son espérance donne $2\sigma^2$. En rassemblant les trois termes, nous obtenons : $E\left(\sum_{i=1}^{n}\hat{u}_i^2\right) = (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2$, si bien que $\text{E}[\text{SCR}/(n-2)] = \sigma^2$

Si nous insérons $\hat{\sigma}^2$ dans les formules (2.57) et (2.58), nous obtenons des estimateurs sans biais de $\text{Var}(\hat{\beta}_1)$ et $\text{Var}(\hat{\beta}_0)$. Nous aurons également besoin d'estimateurs pour les écarts-types de $\hat{\beta}_1$ et $\hat{\beta}_0$. Comme ceux-ci reposent sur l'estimation de σ, il nous faut, tout d'abord, trouver un estimateur pour σ. Le plus naturel est **l'écart-type de la régression (ETR)** que nous définissons comme suit :

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}. \qquad \textbf{[2.62]}$$

Il est parfois défini comme $\hat{\sigma}_u$. On désigne également cet estimateur sous les deux sigles anglais suivants : RMSE (« root mean squared error ») et SER (« standard error of the regression »). Bien que $\hat{\sigma}$ ne soit pas un estimateur sans biais de σ, il en constitue un estimateur *convergent* dont les propriétés se révéleront très utiles malgré tout (voir l'annexe C).

L'estimation $\hat{\sigma}$ est intéressante car elle mesure l'écart-type des facteurs non observés. En effet, elle évalue l'écart-type qui subsiste dans y après intégration de l'effet de x ; autrement dit, elle mesure l'écart-type de y qui n'a pas pu être expliqué par x. Les logiciels de régression linéaire affichent très fréquemment l'estimation de $\hat{\sigma}$, à côté de celles du R carré, de la constante, de la pente, etc. Pour l'instant, nous cherchons à utiliser $\hat{\sigma}$ pour estimer les écarts-types de $\hat{\beta}_0$ et $\hat{\beta}_1$. Étant donné que $\sigma(\hat{\beta}_1) = \sigma/\sqrt{\text{SCT}_x}$, l'estimateur naturel de $\sigma(\hat{\beta}_1)$ est

$$\hat{\sigma}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{SCT}_x} = \hat{\sigma}/\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2}.$$

Cet estimateur est dénommé **l'écart-type estimé de $\hat{\beta}_1$**. En anglais, on parle de « standard error of $\hat{\beta}_1$ » dont le symbole est se($\hat{\beta}_1$). Notez bien que $\hat{\sigma}(\hat{\beta}_1)$ doit être considéré comme une variable aléatoire puisque $\hat{\sigma}$ varie à chaque fois que nous utilisons un échantillon différent et que nous régressons y sur x. Pour un échantillon donné, $\hat{\sigma}(\hat{\beta}_1)$ représente juste une valeur de la distribution sous-jacente, à l'instar de $\hat{\beta}_1$.

De même, $\hat{\sigma}(\hat{\beta}_0)$ provient de $\sigma(\hat{\beta}_0)$ à la seule différence que σ est remplacé par $\hat{\sigma}$. L'écart-type estimé mesure l'incertitude avec laquelle cette estimation a pu être calculée sur base de l'estimateur. Les écarts-types estimés jouent un rôle fondamental dans les chapitres suivants ; nous en aurons besoin pour construire les statistiques des tests d'hypothèse ainsi que les intervalles de confiance, notamment au chapitre 4.

2.6 RÉGRESSION PASSANT PAR L'ORIGINE ET RÉGRESSION SUR CONSTANTE

Dans de rares circonstances, il est désirable d'obtenir une valeur attendue de y égale à zéro lorsque $x = 0$, ce qui requiert une régression sans constante. Par exemple, si le revenu (x) est égal à zéro, l'impôt sur le revenu (y) ne peut être que nul. Il existe également des modèles dont la constante n'est pas égale à zéro mais qui peuvent être transformés en modèles sans constante.

Sur un plan plus formel, nous choisissons un estimateur de la pente, $\tilde{\beta}_1$, et une droite de régression

$$\tilde{y} = \tilde{\beta}_1 x, \qquad \textbf{[2.63]}$$

dans laquelle le tilde est placé au-dessus de β_1 et de y pour le distinguer du cas où la pente *et* la constante sont incluses dans le modèle. L'équation (2.63) représente une **régression passant par l'origine** car elle passe par la coordonnée $x = 0$, $\tilde{y} = 0$. L'estimation de la pente dans (2.63) s'effectue également à l'aide des moindres carrés ordinaires dont l'objectif est de minimiser la somme des carrés des résidus suivante :

$$\sum_{i=1}^{n}(y_i - \tilde{\beta}_1 x_i)^2. \qquad \textbf{[2.64]}$$

En utilisant la dérivée première par rapport à x_i, nous obtenons la condition de premier ordre :

$$\sum_{i=1}^{n} x_i(y_i - \tilde{\beta}_1 x_i) = 0. \qquad \textbf{[2.65]}$$

La solution de (2.65) par rapport à $\tilde{\beta}_1$ est :

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \qquad [2.66]$$

à condition que tous les éléments x_i ne soient pas égaux à zéro, une éventualité que nous pouvons raisonnablement exclure.

Comparons $\tilde{\beta}_1$ à l'estimateur de la pente d'une droite de régression comportant une constante [voir l'équation (2.49) pour $\hat{\beta}_1$]. Ces deux estimateurs seront identiques si, et seulement si, $\bar{x} = 0$. Dans la littérature empirique, on ne recourt pas très souvent à une régression passant par l'origine pour estimer β_1 ; la raison en est simple : si la valeur de β_0 dans la population est différente de zéro ($\beta_0 \neq 0$), $\tilde{\beta}_1$ est un estimateur biaisé de β_1. Vous aurez à le démontrer pour résoudre l'exercice 8.

Dans le cas où la régression passant par l'origine est appropriée, l'interprétation du *R* carré peut néanmoins poser problème. Comme indiqué dans l'équation (2.33), le dénominateur du *R* carré tient explicitement compte de $\bar{y}$, la moyenne de $\{y_i : i = 1, \ldots, n\}$ au sein de l'échantillon. Or, dans certains logiciels économétriques, le *R* carré d'une régression est calculé en considérant que $\bar{y}$ vaut zéro. Dans un tel cas,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \tilde{\beta}_1 x_i)^2}{\sum_{i=1}^{n} y_i^2} \qquad [2.67]$$

Le numérateur est logique car il correspond à la somme des carrés des résidus (SCR). Quant au dénominateur, il suppose que la valeur moyenne de *y* dans la population est connue et égale à zéro. Notez aussi que le *R* carré sera toujours positif dans un tel cas de figure, puisque le dénominateur sera toujours plus grand que le numérateur. En réalité, dans le cas de la régression passant par l'origine, il est préférable d'utiliser la définition traditionnelle de la somme des carrés totaux (SCT), celle de l'équation (2.33). Dans ce cas,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \tilde{\beta}_1 x_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \qquad [2.68]$$

Contrairement à (2.33) et (2.67), le *R* carré peut être négatif dans (2.68) [car le numérateur de (2.68) ne correspond pas à celui de (2.33)]. Il sera négatif lorsque l'utilisation de $\bar{y}$ permet de mieux expliquer la variation des y_i que ne le fait la régression passant par l'origine, basée sur les x_i. C'est la raison pour laquelle (2.68) est plus intéressant que (2.67). Dans (2.68), si le *R* carré est négatif, la variable *x* ne sert à rien et peut être ignorée.

Cette discussion sur l'utilité d'une régression passant par l'origine et sur les différentes mesures de qualité d'ajustement nous amène à nous poser une autre question : qu'en est-il d'**une régression de *y* sur une constante** uniquement ? Autrement dit, que se passe-t-il si nous fixons la valeur de la pente égale à zéro et que nous n'estimons que la constante ? La réponse est simple : la constante sera égale à $\bar{y}$. Comme il s'agit à nouveau de trouver la plus petite somme des écarts aux carrés pour *y*, la solution des MCO pour la constante sera égale à la moyenne de *y* dans l'échantillon, ce que nous pourrions également démontrer à l'aide de statistiques élémentaires. Sous cet angle de vue, l'équation (2.68) permet de comparer la qualité d'ajustement d'une régression sur *x* en passant par l'origine à celle d'une régression sur la constante. Un *R* carré négatif nous indique qu'une régression sur la constante est préférable.

RÉSUMÉ

Dans ce chapitre, nous avons étudié le modèle de régression linéaire simple et nous en avons défini les propriétés de base. À l'aide d'un échantillon aléatoire, la méthode des moindres carrés ordinaires permet d'estimer les paramètres de la pente et de la constante de la population, qu'il nous est impossible d'observer. Nous avons présenté les fondements mathématiques et statistiques de la droite de régression des MCO. Nous avons appris à en calculer les valeurs ajustées et les résidus. Nous avons également appris à interpréter les estimations de la pente pour déterminer l'effet d'une variation de x sur y. Dans la section 2.4, nous avons abordé deux sujets importants d'un point de vue pratique : (1) l'impact que peut avoir un changement des unités de mesure des variables x et y sur les estimations des MCO ; (2) le recours au logarithme naturel pour construire des modèles à élasticité ou semi-élasticité constante.

Dans la section 2.5, nous avons montré que, sous les hypothèses RLS.1 à RLS.4, les estimateurs des MCO étaient sans biais. L'hypothèse RLS.4 revêt une importance toute particulière. Sous cette hypothèse, l'espérance du terme d'erreur u est nulle, quelle que soit la valeur de x. Il y a néanmoins plusieurs raisons de penser que cette hypothèse est violée dans bon nombre d'applications en sciences sociales : les facteurs non observés et compris dans u sont souvent corrélés avec x. Grâce à l'hypothèse RLS.5 sous laquelle la variance de l'erreur, étant donné x, est constante, nous pouvons obtenir des formules simples pour les variances d'échantillonnage relatives aux estimateurs des MCO. Comme nous l'avons vu, la variance de l'estimateur de la pente $\hat{\beta}_1$ augmente lorsque la variance de l'erreur augmente ; elle diminue quand il y a une plus grande variation de la variable indépendante au sein de l'échantillon. Nous avons également dérivé un estimateur sans biais pour $\sigma^2 = Var(u)$.

Dans la section 2.6, nous avons brièvement discuté de la régression passant par l'origine dans laquelle l'estimateur de la pente est dérivé sous l'hypothèse que la constante est égale à zéro. Ce type de régression est utile dans certains cas spécifiques mais son utilisation dans les travaux empiriques reste relativement rare.

Il nous reste encore beaucoup de travail à accomplir. Par exemple, nous ne savons toujours pas effectuer de test d'hypothèse sur les paramètres de la population, β_0 et β_1. Certes, nous savons que l'absence de biais pour les estimateurs des MCO est vérifiée sous les hypothèses RLS.1 à RLS.4, mais nous n'avons toujours aucun outil à notre disposition pour induire les caractéristiques inconnues des paramètres de la population à partir d'un échantillon issu de cette population. D'autres sujets n'ont pas été abordés, tels que l'efficacité des MCO par rapport à des méthodes d'estimation alternatives.

La construction des intervalles de confiance, la réalisation de tests d'hypothèse, et l'efficacité des estimateurs sont des thèmes tout aussi importants dans le cadre de la régression multiple. Étant donné que la régression simple est un cas particulier de la régression multiple et que les méthodologies requises sont très similaires, nous utiliserons mieux notre temps en abordant ces sujets dans le cadre de la régression multiple dont l'utilisation est beaucoup plus fréquente dans les travaux empiriques. L'objectif du chapitre 2 était de vous familiariser aux concepts économétriques les plus fondamentaux dans un environnement technique le plus simple possible.

LES HYPOTHÈSES DE GAUSS-MARKOV DANS LA RÉGRESSION SIMPLE

À des fins pratiques, nous résumons les **hypothèses de Gauss-Markov** que nous avons utilisées dans ce chapitre. Rappelez-vous que seules les hypothèses RLS.1 à RLS.4 sont requises pour démontrer que $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais. Nous avons ajouté l'hypothèse d'homoscédasticité, RLS.5, dans le but d'obtenir les deux formules traditionnelles de la variance des estimateurs, (2.57) et (2.58).

Hypothèse RLS.1 (Linéarité dans les paramètres)

Dans le modèle issu de la population, la variable dépendante, y, est liée à la variable indépendante, x, et au terme d'erreur, u, comme suit :

$$y = \beta_0 + \beta_1 x + u, \qquad [2.47]$$

où β_0 et β_1 sont respectivement les paramètres de la constante et de la pente au sein de la population.

Hypothèse RLS.2 (Échantillonnage aléatoire)

Nous disposons d'un échantillon aléatoire de taille n, $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$, tiré du modèle issu de la population décrit sous l'hypothèse RLS.1.

Hypothèse RLS.3 (Variation de la variable explicative au sein de l'échantillon)

Les éléments de x au sein l'échantillon, à savoir $\{x_i, i = 1, \ldots, n\}$, n'ont pas tous la même valeur.

Hypothèse RLS.4 (Espérance conditionnelle de l'erreur égale à zéro)

Le terme d'erreur u affiche une espérance égale à zéro, quelle que soit la valeur de x. Autrement dit,

$$E(u|x) = 0.$$

Hypothèse RLS.5 (Homoscédasticité)

La variance de l'erreur u est constante, quelle que soit la valeur de x. En d'autres termes,

$$Var(u|x) = \sigma^2.$$

MOTS-CLÉS

Coefficient de détermination p. 63
Coefficient de la constante p. 47
Coefficient de la pente p. 47
Conditions de premier ordre p. 54
Covariable p. 47
Degrés de liberté p. 81
Droite de régression des MCO p. 55
Écart-type de la régression (ETR) p. 81
Écart-type estimé de $\hat{\beta}_1$ p. 82
Élasticité p. 69
Fonction de régression de l'échantillon (FRE) p. 55
Fonction de régression de la population (FRP) p. 49
Hétéroscédasticité p. 77
Homoscédasticité p. 76
Hypothèse d'espérance conditionnelle de l'erreur égale à zéro p. 49
Hypothèses de Gauss-Markov p. 84
Modèle à élasticité constante p. 68
Modèle de régression linéaire simple p. 46
Moindres carrés ordinaires (MCO) p. 54
R carré p. 63
Régresseur p. 46

Régression passant par l'origine p. 82
Résidu p. 54
Semi-élasticité p. 69
Somme des carrés des résidus (SCR) p. 54, 62
Somme des carrés expliqués (SCE) p. 62
Somme des carrés totaux (SCT) p. 62
Terme d'erreur (perturbation) p. 47
Valeur ajustée p. 54
Variable de contrôle p. 46
Variable dépendante p. 46
Variance de l'erreur p. 77
Variable de réponse p. 46
Variable explicative p. 46
Variable expliquée p. 46
Variable indépendante p. 46
Variable prédictive p. 46
Variable prédite p. 46

EXERCICES

1. La variable *kids* inclut le nombre d'enfants par femme et *educ* correspond à leur niveau d'instruction (en nombre d'années d'études). Un modèle de régression simple de la fécondité peut consister à régresser la fécondité sur le niveau d'instruction :

$$kids = \beta_0 + \beta_1 \, educ + u$$

où u est le terme d'erreur (non observé).

i. Quels types de facteurs sont inclus dans u ? Sont-ils susceptibles d'être corrélés avec le niveau d'instruction ?

ii. Cette analyse de régression simple permet-elle d'identifier l'effet *ceteris paribus* du niveau d'instruction sur la fécondité ? Expliquez.

2. Dans le modèle de régression linéaire simple $y = \beta_0 + \beta_1 x + u$, imaginez que $E(u) \neq 0$. En posant que $a_0 = E(u)$, montrez que le modèle conserve la même pente mais qu'il incorpore une nouvelle constante et un nouveau terme d'erreur dont l'espérance est égale zéro.

3. Le tableau suivant contient le résultat obtenu par huit étudiants au test « ACT » (American College Testing) ; ce test est basé sur des QCM et sa note maximale est 36. Le tableau reprend également la moyenne « GPA » (Grade Point Average) obtenue aux examens à la sortie du lycée (soit à la fin du secondaire supérieur), dont la note maximale est 4. Ces deux évaluations sont notamment utilisées aux États-Unis pour accéder aux universités.

Étudiant	GPA	ACT
1	2,8	21
2	3,4	24
3	3,0	26
4	3,5	27
5	3,6	29

Étudiant	GPA	ACT
6	3,0	25
7	2,7	25
8	3,7	30

i. Estimez la relation entre *GPA* et *ACT* en utilisant la méthode des moindres carrés ordinaires. Autrement dit, calculez les estimations de la constante de la pente de l'équation

$$\widehat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 \text{ ACT}.$$

Décrivez la nature de cette relation. L'interprétation de la constante est-elle utile ? Expliquez. Quelle est la variation estimée de *GPA* si la note obtenue à l'*ACT* augmente de 5 points ?

ii. Calculez les valeurs ajustées et les résidus pour chaque observation. Vérifiez que la somme des résidus est (approximativement) égale à zéro.

iii. Quelle est la valeur estimée (ou « valeur prédite ») de *GPA* lorsque *ACT* = 20 ?

iv. Quel pourcentage de *GPA* est expliqué par *ACT* ? Expliquez.

4. La base de données BWGHT contient des informations sur les naissances aux États-Unis. Les deux variables qui nous intéressent ici sont le poids du nouveau-né (*bwght*), en onces (1 once = 28,35 grammes), et le nombre de cigarettes fumées en moyenne chaque jour par la mère durant la grossesse (*cigs*). La régression simple, estimée sur n = 1 388 naissances, donne les résultats suivants :

$$\widehat{bwght} = 119{,}77 - 0{,}514\ cigs.$$

i. Quel est le poids du nouveau-né estimé par le modèle lorsque *cigs* = 0 ? Qu'en est-il lorsque *cigs* = 20 (un paquet par jour) ? Commentez.

ii. Pensez-vous que cette régression simple capture la relation causale qui existe entre le poids du nouveau-né et la consommation de tabac de la mère ? Expliquez.

iii. Quelle est la valeur de *cigs* si l'estimation du poids du nouveau-né est égale à 125 onces, soit 3,5 kg environ ? Commentez.

iv. Dans l'échantillon, la proportion de femmes qui ne fument pas durant leur grossesse est égale à 0,85. Cela vous aide-t-il à mieux expliquer le résultat obtenu au point (iii) ?

5. La fonction linéaire de consommation suivante est estimée sur base d'un échantillon de 100 familles dont la consommation annuelle (cons) et le revenu annuel (*inc*) sont mesurés en dollars :

$$\widehat{cons} = \hat{\beta}_0 + \hat{\beta}_1\ inc,$$

où la *propension marginale à consommer* (PmC) est estimée par la pente, $\hat{\beta}_1$, alors que la *propension moyenne à consommer* (PMC) est égale à $\widehat{cons}/inc = \hat{\beta}_0/inc + \hat{\beta}_1$. L'estimation de l'équation par les MCO donne :

$$\widehat{cons} = -124{,}84 + 0{,}853\ inc,$$
$$n = 100,\ R^2 = 0{,}692.$$

i. Interprétez l'estimation de la constante de cette équation. Quel est votre commentaire sur son signe et son ampleur ?

ii. Quelle est la consommation à laquelle on doit s'attendre lorsque le revenu annuel de la famille est égal à 30 000 dollars ?

iii. En utilisant *inc* sur l'axe des *x*, construisez un graphique pour représenter la PMC et la PmC que vous avez estimées précédemment.

6. Sur base des données de Kiel et McClain (1995) portant sur 135 transactions immobilières effectuées en 1988 à Andover au Massachusetts, on obtient l'équation suivante, qui explique le prix de vente des biens immobiliers (*price*) par la distance qui les sépare d'un incinérateur de déchets (*dist*) :

$$\widehat{\log(price)} = 9{,}40 + 0{,}312 \log(dist),$$
$$n = 135,\ R^2 = 0{,}162.$$

i. Interprétez le coefficient de log(*dist*). S'agit-il du signe auquel vous vous attendiez ?

ii. Pensez-vous que la régression simple permette d'obtenir un estimateur sans biais de l'élasticité de *price* par rapport à *dist*, toutes choses étant égales par ailleurs (*ceteris paribus*) ? (Pensez à la diversité des quartiers dans une ville et à la décision de l'autorité politique portant sur le lieu d'implantation de l'incinérateur.)

iii. Quels sont les autres facteurs d'un bien immobilier qui peuvent en influencer le prix ? Pourraient-ils être corrélés avec la distance qui sépare l'incinérateur de ce bien ?

7. Considérez la fonction d'épargne

$$sav = \beta_0 + \beta_1\, inc + u,\ u = \sqrt{inc} \cdot e,$$

où *inc* est le revenu, *sav* est l'épargne, et *e* est une variable aléatoire pour laquelle $E(e) = 0$ et $Var(e) = \sigma_e^2$. Supposez que *e* est indépendant de *inc*.

i. Montrez que $E(u|inc) = 0$, c'est-à-dire que l'hypothèse RLS.4 est satisfaite. [*Astuce :* Si *e* est indépendant de *inc*, alors $E(e|inc) = E(e)$.]

ii. Montrez que $Var(u|inc) = \sigma_e^2 inc$, c'est-à-dire que l'hypothèse RLS.5 est violée. En particulier, montrez que la variance de *sav* augmente avec *inc*. [*Astuce* : $Var(e|inc) = Var(e)$, si *e* et *inc* sont indépendants.]

iii. Identifiez les arguments en faveur de l'idée selon laquelle la variance de l'épargne augmente en fonction du revenu.

8. Considérez le modèle de régression simple $y = \beta_0 + \beta_1 x + u$. Sous les hypothèses de Gauss-Markov (RLS.1 à RLS.5), les estimateurs traditionnels des MCO, $\hat{\beta}_0$ et $\hat{\beta}_1$, sont des estimateurs sans biais des paramètres respectifs de la population. Soit $\tilde{\beta}_1$, l'estimateur de β_1 lorsque la constante est égale à zéro (voir la section 2.6).

i. Exprimez $E(\tilde{\beta}_1)$ en fonction des x_i, β_0, et β_1. Vérifiez que $\tilde{\beta}_1$ est un estimateur sans biais de β_1 lorsque la valeur de la constante au sein de la population (β_0) est égale à zéro. Existe-t-il d'autres cas pour lesquels $\tilde{\beta}_1$ est sans biais ?

ii. Calculez la variance de $\tilde{\beta}_1$. (*Astuce* : la variance ne dépend pas de β_0.)

iii. Montrez que $Var(\tilde{\beta}_1) \leq Var(\hat{\beta}_1)$. [*Astuce* : pour tout échantillon, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$ avec une inégalité stricte, sauf lorsque $\bar{x} = 0$.]

iv. Expliquez le compromis qui existe entre le biais et la variance lorsqu'il s'agit de faire un choix entre $\hat{\beta}_1$ et $\tilde{\beta}_1$.

9. i. Soit $\hat{\beta}_0$ et $\hat{\beta}_1$, l'ordonnée à l'origine et la pente de la droite de régression de y_i sur x_i, avec n observations. Soit c_1 et c_2, deux constantes (ou facteurs d'échelle), avec $c_2 \neq 0$. Soit $\tilde{\beta}_0$ et $\tilde{\beta}_1$, l'ordonnée à l'origine et la pente de la droite de régression de $c_1 y_i$ sur $c_2 x_i$. Montrez que $\tilde{\beta}_1 = (c_1/c_2)\hat{\beta}_1$ et $\tilde{\beta}_0 = c_1\hat{\beta}_0$, vérifiant de cette manière les affirmations de la section 2.4 concernant les unités de mesure. [*Astuce* : pour obtenir $\tilde{\beta}_1$, utilisez les versions de x et y affectés des facteurs d'échelle dans (2.19). Ensuite, utilisez (2.17) pour obtenir $\tilde{\beta}_0$, en veillant à insérer les x et y avec les facteurs d'échelle *ainsi que* la « bonne » pente.]

ii. Supposons maintenant que $\tilde{\beta}_0$ et $\tilde{\beta}_1$ soient issus de la régression de $(c_1 + y_i)$ sur $(c_2 + x_i)$, sans aucune restriction sur c_1 et c_2. Montrez que $\tilde{\beta}_1 = \hat{\beta}_1$ et $\tilde{\beta}_0 = \hat{\beta}_0 + c_1 - c_2\hat{\beta}_1$.

iii. Supposons maintenant que $\hat{\beta}_0$ et $\hat{\beta}_1$ correspondent aux estimations des MCO provenant de la régression de $\log(y_i)$ sur x_i, où nous supposons que $y_i > 0$ pour tout i. Soit $\tilde{\beta}_0$ et $\tilde{\beta}_1$, l'ordonnée à l'origine et la pente de la droite de régression de $\log(c_1 y_i)$ sur x_i, avec $c_1 > 0$. Montrez que $\tilde{\beta}_1 = \hat{\beta}_1$ et $\tilde{\beta}_0 = \log(c_1) + \hat{\beta}_0$.

iv. Soit $\tilde{\beta}_0$ et $\tilde{\beta}_1$, l'ordonnée à l'origine et la pente de la droite de régression de y_i sur $\log(c_2 x_i)$, avec $x_i > 0$ pour tout i. Comparez $\tilde{\beta}_0$ et $\tilde{\beta}_1$ respectivement à l'ordonnée à l'origine et à la pente de la droite de régression de y_i sur $\log(x_i)$.

10. Soit $\hat{\beta}_0$ et $\hat{\beta}_1$, les estimateurs des MCO de l'ordonnée à l'origine et de la pente. Soit $\bar{u}$, la moyenne des erreurs (et non des résidus !).

i. Montrez que $\hat{\beta}_1$ peut s'écrire sous la forme $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$, où $w_i = d_i/\text{SCT}_x$ et $d_i = x_i - \bar{x}$.

ii. En partant du point (i) et sachant que $\sum_{i=1}^{n} w_i = 0$, montrez que $\hat{\beta}_1$ et $\bar{u}$ ne sont pas corrélés. [*Astuce* : On vous demande de montrer que $\text{E}[(\hat{\beta}_1 - \beta_1) \cdot \bar{u}] = 0$].

iii. Montrez que $\hat{\beta}_0$ peut s'écrire sous la forme $\hat{\beta}_0 = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x}$.

iv. Utilisez les points (ii) et (iii) pour montrer que $\text{Var}(\hat{\beta}_0) = \sigma^2/n + \sigma^2(\bar{x})^2/\text{SCT}_x$.

v. Simplifiez l'expression du point (iv) pour aboutir à l'équation (2.58). [*Astuce* : $\text{SCT}_x/n = n^{-1}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$].

11. Vous cherchez à quantifier la relation entre, d'une part, le nombre d'heures d'études consacrées par semaine à un cours de remise à niveau (*study*) et, d'autre part, la moyenne des résultats obtenus par les étudiants universitaires de première année (*gpa*).

i. En quoi consisterait une étude contrôlée (ou expérimentation) dans un tel contexte ? Est-il possible de constituer un groupe de contrôle ?

ii. Envisagez un cas plus réaliste : les étudiants choisissent le temps qu'ils désirent consacrer au cours de remise à niveau. Vous tirez ensuite un échantillon de *gpa* et *study* au sein de la population. Le modèle issu de la population est

$$gpa = \beta_0 + \beta_1\, study + u$$

où nous supposons que $\text{E}(u) = 0$. Identifiez au moins deux facteurs qui sont inclus dans u. Sont-ils susceptibles d'être corrélés avec *study* ?

iii. Dans l'équation du point (ii), quel devrait être le signe de β_1 si le cours de remise à niveau est utile ?

iv. Dans l'équation du point (ii), quelle interprétation donnez-vous à β_0 ?

12. Reconsidérez le problème décrit à la fin de la section 2.6, celui de la régression sur constante.

i. Étant donné un échantillon $\{y_i : i = 1, 2, \ldots, n\}$, soit $\tilde{\beta}_0$ la solution au problème de minimisation suivant :

$$\min_{b_0} \sum_{i=1}^{n} (y_i - b_0)^2.$$

Montrez que $\tilde{\beta}_0 = \bar{y}$, autrement dit, que la moyenne de l'échantillon minimise la somme des carrés des résidus. (*Astuce* : vous pouvez directement le démontrer en ajoutant et en soustrayant $\bar{y}$ à l'intérieur des résidus au carré ; un calcul élémentaire suffit par la suite.)

ii. Soit les résidus $\tilde{u}_i = y_i - \bar{y}$. Montrez que ces résidus sont toujours égaux à zéro.

EXERCICES SUR ORDINATEUR

C1. Les données 401K proviennent d'une étude réalisée par Papke (1995). Cette étude cherche à expliquer la participation des travailleurs au plan d'épargne-pension « 401(k) », qui est un système par capitalisation très largement utilisé aux États-Unis. La variable dépendante, *prate*, représente le taux de participation des travailleurs, égal au pourcentage de travailleurs éligibles qui ont effectivement ouvert un compte d'épargne-pension. La variable explicative, *mrate*, caractérise la « générosité » du plan de pension, dans le sens où elle mesure la proportion de la contribution venant de l'employeur relativement à celle du travailleur. Par exemple, si *mrate* = 0,50, une contribution de 1 dollar de la part du travailleur est accompagnée d'une contribution de 0,5 dollar venant de l'employeur.

i. Calculez le taux de participation moyen ainsi que la contribution relative moyenne de l'employeur dans l'échantillon.

ii. Estimez le modèle de régression simple pour obtenir

$$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate$$

et indiquez les résultats de l'estimation en n'oubliant pas de préciser la taille de l'échantillon et le R carré.

iii. Interprétez la constante et le coefficient de *mrate*.

iv. Calculez la valeur « prédite » de *prate* lorsque *mrate* = 3,5. S'agit-il d'une estimation raisonnable ? Expliquez.

v. Quel pourcentage de la variation de *prate* est expliqué par *mrate* ? S'agit-il d'une valeur élevée selon vous ?

C2. La base de données CEOSAL2 contient des informations relatives aux PDG de grandes entreprises américaines. La variable *salary* représente la rémunération annuelle, en milliers de dollars ; *ceoten* est égal au nombre d'années d'expérience au poste de PDG.

i. Calculez le salaire moyen et l'ancienneté moyenne des PDG dans l'échantillon.

ii. Combien de PDG occupent cette fonction depuis moins d'un an (autrement dit, *ceoten* = 0) ? Quelle est la plus grande ancienneté à ce poste dans l'échantillon ?

iii. Estimez le modèle de régression simple

$$\log(salary) = \beta_0 + \beta_1 ceoten + u,$$

et affichez les résultats de la manière habituelle. Quelle est votre estimation de l'augmentation de salaire lorsqu'un PDG acquiert une année d'expérience en plus à ce poste ?

C3. Les données comprises dans SLEEP75 sont celles de Biddle et Hamermesh (1990). Ils étudient le compromis entre heures de sommeil (*sleep*, en minutes) et heures de travail rémunéré (*totwrk*, en minutes). Remarquez que ces deux variables peuvent être utilisées comme variable dépendante. Pour simplifier, estimez le modèle

$$sleep = \beta_0 + \beta_1 totwrk + u.$$

i. Indiquez les résultats de l'estimation en précisant le nombre d'observations et le R^2. Quelles interprétations donnez-vous à la constante et à la pente de la droite de régression ?

ii. Si *totwrk* augmente de 2 *heures*, quelle sera la réduction estimée de *sleep* ? S'agit-il d'un impact conséquent ?

C4. Utilisez les données WAGE2 pour estimer une régression simple du salaire mensuel (*wage*) sur le résultat obtenu à un test de QI (*IQ*).

i. Calculez la moyenne du salaire et celle du test de QI dans l'échantillon. Quel est l'écart-type de *IQ* dans l'échantillon ? (Les résultats au test de QI sont standardisés de telle sorte que dans *la population*, la moyenne est égale à 100 et l'écart-type à 15.)

ii. Estimez un modèle de régression simple dans lequel une augmentation d'un point au test de QI implique une variation du salaire d'un montant constant en dollars. Utilisez ce modèle pour estimer l'augmentation attendue du salaire lorsque l'augmentation de *IQ* est égale à 15 points. Pensez-vous que *IQ* explique l'essentiel de la variation de *wage* ?

iii. Estimez maintenant un modèle dans lequel une augmentation d'un point au test de QI implique une variation de *wage* en *pourcentage*. Si *IQ* augmente de 15 points, quelle est votre estimation (approximative) de l'augmentation en pourcentage de *wage* ?

C5. Soit *rd*, les dépenses annuelles de recherche et développement, et *sales*, les ventes annuelles, toutes deux en millions de dollars. L'échantillon comprend des entreprises de l'industrie chimique.

i. Écrivez un modèle (et non une équation estimée) qui implique une élasticité constante entre *rd* et *sales*. Quel paramètre représente l'élasticité ?

ii. Estimez le modèle à l'aide du jeu de données RDCHEM. Affichez les résultats de l'équation estimée sous le format habituel. Quelle est l'estimation de l'élasticité de *rd* par rapport à *sales* ? Expliquez ce que l'élasticité signifie.

C6. Comme dans l'exemple 2.12, nous utilisons la base de données MEAP93. Nous explorons ici la relation entre *le taux* de réussite à l'examen de math (*math*10, en pourcentage) et les dépenses de l'établissement par étudiant (*expend*, en dollars).

i. Pensez-vous que chaque dollar supplémentaire que l'établissement dépense par étudiant aura un effet constant sur le taux de réussite ? Un effet marginalement décroissant n'est-il pas réaliste ? Expliquez.

ii. Dans le modèle issu de la population

$$math10 = \beta_0 + \beta_1 \log(expend) + u,$$

montrez que $\beta_1/10$ correspond à la variation en point de pourcentage de *math*10 étant donné une augmentation de 10 % de *expend*.

iii. Utilisez MEAP93 pour estimer le modèle décrit au point (ii). Affichez les résultats de l'équation estimée, en incluant la taille de l'échantillon et le *R* carré.

iv. Quelle importance faut-il donner à l'influence des dépenses sur le taux de réussite ? En particulier, si les dépenses augmentent de 10 %, quelle est l'augmentation attendue de *math*10 *en point de pourcentage* ?

v. Il serait inquiétant que l'estimation de ce modèle nous donne des valeurs ajustées de *math*10 supérieures à 100. Pourquoi ne faut-il pas trop s'en inquiéter dans le cas particulier de cette base de données ?

C7. Pour répondre aux questions suivantes, utilisez la base de données CHARITY de Franses et Paap (2001).

i. Sur base de cet échantillon comprenant 4 268 personnes, calculez la moyenne des dons qui ont été faits (*gift*, en florin néerlandais). Quel est le pourcentage de personnes qui n'ont fait aucun don ?

ii. Quel est le nombre moyen de sollicitations envoyées par la poste sur une année (*mailsyear*) ? Quelles en sont les valeurs minimale et maximale ?

iii. Estimez le modèle

$$gift = \beta_0 + \beta_1 \, mailsyear + u$$

par les MCO et affichez les résultats, en précisant la taille de l'échantillon et le *R* carré.

iv. Donnez une interprétation au coefficient de la pente. Si chaque sollicitation envoyée par la poste coûte un florin, est-il possible pour l'organisme de bienfaisance de réaliser un profit net à la suite de l'envoi de sollicitations ? Cela signifie-t-il qu'un profit est réalisé sur chaque envoi ? Expliquez.

v. Quelle est la valeur du plus petit don que vous pouvez observer dans l'échantillon ? Sur base des résultats de la régression, pouvez-vous être amené à prédire qu'une personne ne fera aucun don ?

C8. Pour réaliser cet exercice, vous devez utiliser un logiciel qui génère des données à partir de la distribution uniforme et de la distribution normale.

i. Utilisez la distribution uniforme pour générer 500 observations, x_i, comprises entre [0,10]. (La plupart des logiciels ont une ligne de code pour la distribution uniforme [0,1], de moyenne nulle et de variance unitaire ; utilisez-la et multipliez ensuite les observations générées par 10.) Quels sont la moyenne et l'écart-type des x_i dans cet échantillon ?

ii. Générez de manière aléatoire 500 erreurs, u_i, à partir d'une distribution normale [0,36], de moyenne nulle et de variance égale à 36. (Si vous générez des observations sur base de la normale [0,1], multipliez ensuite les valeurs générées par six.) Obtenez-vous une estimation de la moyenne des u_i exactement égale à zéro ? Pourquoi ? Quelle est la valeur de l'écart-type pour cette série de u_i ?

iii. Vous pouvez maintenant générer les y_i comme suit :

$$y_i = 1 + 2\,x_i + u_i \equiv \beta_0 + \beta_1 \, x_i + u_i.$$

La constante et la pente dans la population sont respectivement égales à un et deux. Utilisez les données des points (i) et (ii) pour effectuer une régression de y_i sur x_i. Quelles sont les estimations de la constante et de la pente ? Sont-elles égales aux valeurs issues de la population ? Expliquez.

iv. Il est désormais possible d'obtenir les résidus des MCO, $\hat{u}_i$, et de vérifier que l'équation (2.60) tient effectivement (abstraction faite des erreurs d'arrondi).

v. Utilisez maintenant les erreurs u_i, au lieu des résidus, pour recalculer les éléments de l'équation (2.60). Quelles sont vos conclusions ?

vi. Répétez les étapes (i), (ii), et (iii) afin de générer une nouvelle base de données, en commençant par générer les x_i. Qu'obtenez-vous maintenant pour $\hat{\beta}_0$ et $\hat{\beta}_1$? Pourquoi ces valeurs sont-elles différentes de celles obtenues auparavant ?

C9. Utilisez les données COUNTYMURDERS pour répondre à cette question. Utilisez les données de l'année 1996 uniquement.

i. Dans combien de districts n'y a-t-il eu aucun meurtre en 1996 ? Dans combien de districts y a-t-il eu au moins une exécution ? Quel est le plus grand nombre d'exécutions ?

ii. Estimez l'équation

$$murders = \beta_0 + \beta_1 execs + u$$

par MCO et indiquez les résultats de l'estimation comme de coutume, en incluant la taille de l'échantillon et le R-carré.

iii. Interprétez le coefficient de la pente estimé au point (ii). La peine capitale a-t-elle un effet dissuasif ?

iv. Quel est le plus petit nombre de meurtres que peut prédire ce modèle ? Quel est le résidu pour un district dont le nombre d'exécution et le nombre de meurtre sont nuls ?

v. Expliquez la raison pour laquelle une analyse de régression simple n'est pas appropriée lorsqu'il s'agit de déterminer si la peine capitale a un effet dissuasif sur le nombre de meurtres.

C10. Le jeu de données CATHOLIC contient des informations sur les résultats obtenus à des tests par plus de 7 000 étudiants qui étaient inscrits en dernière année du secondaire inférieur aux États-Unis en 1988. Plus précisément, les variables *math*12 et *read*12 correspondent à des résultats obtenus à des tests standardisés de mathématiques et de lecture, s'échelonnant de 0 à 12.

i. Combien d'étudiants contient l'échantillon ? Calculez les moyennes et les écarts-types des variables *math*12 et *read*12.

ii. Effectuez une régression simple de *math*12 sur *read*12 afin d'obtenir les estimations par MCO de la constante et de la pente. Indiquez les résultats de la manière suivante, en veillant à indiquer les valeurs de $\hat{\beta}_0$ et $\hat{\beta}_1$ et à remplacer les points d'interrogation :

$$\widehat{math12} = \hat{\beta}_0 + \hat{\beta}_1 read12$$
$$n = ?\ ,\ R^2 = ?$$

iii. Est-il possible de donner une interprétation sensée à la constante estimée au point (ii) ? Expliquez.

iv. La valeur de $\hat{\beta}_1$ est-elle surprenante ? Qu'en est-il du R-carré ?

v. Supposons que vous présentiez ces résultats à la directrice générale d'un district scolaire et qu'elle vous dise : « vos résultats montrent que pour améliorer les résultats au test de math, il nous faut juste améliorer les résultats au test de lecture ; nous devrions donc engager un plus grand nombre tuteurs de lecture. » Que répondriez-vous à ce commentaire ? (Astuce : Si vous effectuez la régression de *read*12 sur *math*12, quel résultat pensez-vous obtenir ?)

ANNEXE 2A

Minimisation de la somme des carrés des résidus

Nous démontrons que les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ obtenues par les MCO permettent effectivement de minimiser la somme des carrés des résidus, comme nous l'affirmons dans la section 2.2. Sur un plan formel, il s'agit de résoudre le problème de minimisation suivant :

$$\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

dont b_0 et b_1 en sont les arguments ; pour plus de simplicité, appelons cette fonction $Q(b_0, b_1)$. Grâce à un résultat fondamental de l'analyse multivariée (voir annexe A), nous savons qu'une condition nécessaire pour que $\hat{\beta}_0$ et $\hat{\beta}_1$ résolvent ce problème de minimisation est que les dérivées partielles de $Q(b_0, b_1)$ par rapport à β_0 et β_1 soient égales à zéro lorsqu'elles sont évaluées en $\hat{\beta}_0$, $\hat{\beta}_1$. Autrement dit, $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_0 = 0$ et $\partial Q(\hat{\beta}_0, \hat{\beta}_1)/\partial b_1 = 0$. En utilisant le théorème de dérivation des fonctions composées (appelé « règle de la chaîne », en anglais), ces deux équations peuvent s'écrire :

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

$$-2\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

Ces deux équations sont égales à (2.14) et (2.15), multipliées par – 2n ; par conséquent, leurs solutions correspondent également à $\hat{\beta}_0$ et $\hat{\beta}_1$.

Sommes-nous certains d'avoir effectivement minimisé la somme des carrés des résidus ? Certes, les conditions de premier ordre sont nécessaires, mais elles ne sont pas suffisantes. Une manière de vérifier que nous avons effectivement minimisé la somme des carrés des résidus est d'écrire, pour tout b_0 et b_1,

$$\begin{aligned} Q(b_0, b_1) &= \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1) x_i]^2 \\ &= \sum_{i=1}^{n} [\hat{u}_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1) x_i]^2 \\ &= \sum_{i=1}^{n} \hat{u}_i^2 + n(\hat{\beta}_0 - b_0)^2 + (\hat{\beta}_1 - b_1)^2 \sum_{i=1}^{n} x_i^2 + 2(\hat{\beta}_0 - b_0)(\hat{\beta}_1 - b_1) \sum_{i=1}^{n} x_i, \end{aligned}$$

où nous avons utilisé les équations (2.30) et (2.31). Le premier terme ne dépend ni de b_0 ni de b_1 ; la somme des trois derniers termes est égale à

$$\sum_{i=1}^{n} [(\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1) x_i]^2,$$

ce que nous pouvons vérifier à l'aide d'un peu d'algèbre. Comme il s'agit d'une somme de termes au carré, son minimum est égal à zéro et est atteint lorsque $b_0 = \hat{\beta}_0$ et $b_1 = \hat{\beta}_1$.

CHAPITRE 3

LE MODÈLE DE RÉGRESSION LINÉAIRE MULTIPLE

Traduction de Maëlys de la Rupelle

Dans le chapitre 2, nous avons vu que la régression linéaire simple (RLS) pouvait être utilisée pour expliquer la variable dépendante, y, en fonction d'une seule variable indépendante, x. Le principal problème lié à l'utilisation de la régression simple est qu'il est très difficile de tirer des conclusions empiriques valides concernant l'impact de x sur y, *toutes choses égales par ailleurs*. En règle générale, l'hypothèse RLS.4 ne tient pas car la variable x est souvent corrélée à un autre facteur qui influence y.

Le modèle de régression linéaire multiple (RLM), appelé également modèle de régression multiple, convient davantage à un raisonnement *ceteris paribus*, car il nous permet de prendre en compte de manière explicite de nombreux facteurs qui affectent simultanément la variable dépendante. C'est important aussi bien pour tester des théories économiques que pour évaluer des politiques publiques à partir de données non expérimentales. Comme les modèles de régression multiple peuvent inclure de nombreuses variables explicatives, éventuellement corrélées entre elles, il est possible de mener une analyse causale dans des situations où l'utilisation de la régression simple serait inadéquate.

Si nous ajoutons des facteurs utiles pour expliquer y dans notre modèle, nous parviendrons naturellement à expliquer une plus grande partie de la variation de y. L'utilisation de la régression multiple peut donc conduire à une meilleure prédiction de la variable dépendante.

Un autre avantage de la régression multiple est qu'elle permet de recourir à des formes fonctionnelles diverses et variées. Ce n'est pas le cas de la régression simple dans laquelle n'apparaît qu'une seule fonction de la variable explicative, x. Comme nous le verrons, la régression multiple offre beaucoup plus de flexibilité.

La section 3.1 introduit le modèle de régression multiple et en présente les avantages par rapport au modèle de régression simple. La section 3.2 est consacrée à la méthode des moindres carrés ordinaires (MCO) sur laquelle repose l'estimation des paramètres du modèle de régression multiple. Les sections 3.3, 3.4 et 3.5 décrivent les différentes propriétés statistiques de l'estimateur des MCO, notamment son absence de biais et son efficacité.

Le modèle de régression multiple est encore de nos jours l'outil le plus utilisé pour analyser des données, aussi bien en économie que dans les autres sciences sociales. Quant à la méthode des moindres carrés ordinaires, elle reste la méthode d'estimation du modèle de régression multiple la plus populaire.

3.1 LES AVANTAGES DU MODÈLE DE RÉGRESSION LINÉAIRE MULTIPLE

Le modèle à deux variables indépendantes

Commençons par quelques exemples simples. L'objectif est de montrer que la régression multiple permet de résoudre des problèmes que la régression simple est incapable de surmonter.

Le premier exemple est une simple variante de l'équation du salaire, introduite au chapitre 2, qui vise à obtenir l'effet du niveau d'instruction (mesuré en années d'études) sur le salaire horaire :

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u, \qquad \textbf{[3.1]}$$

où *exper* est l'expérience professionnelle, en nombre d'années. La variable *wage* est expliquée par deux variables explicatives : le niveau d'instruction et l'expérience. Les autres facteurs non observés sont inclus dans u. Nous sommes particulièrement intéressés par l'effet *ceteris paribus* d'*educ* sur *wage*, mesuré par β_1 ; il faut donc que les autres facteurs affectant *wage* demeurent constants.

Si nous comparons l'équation (3.1) à l'équation de la régression simple qui explique *wage* par *educ* uniquement, nous constatons que le terme d'erreur de l'équation (3.1) ne contient plus *exper* puisque cette

variable est présente dans l'équation. Quant à β_2, le coefficient associé à *exper*, il mesure l'effet *ceteris paribus* d'*exper* sur *wage*.

Comme dans le modèle de régression simple, nous allons devoir poser des hypothèses sur la manière dont *u*, dans l'équation (3.1), est relié aux variables indépendantes, *educ* et *exper*. Néanmoins, comme nous le verrons dans la section 3.2, nous pouvons être certains d'une chose : comme (3.1) tient compte de l'expérience, nous serons capables de mesurer l'effet du niveau d'instruction sur le salaire, *pour un niveau d'expérience donné*. Dans un modèle de régression simple, comme le terme d'erreur contient l'expérience, nous devons *supposer* que l'expérience n'est pas corrélée avec les années d'études. Or, cette hypothèse est contestable.

Considérons à présent un second exemple. Nous cherchons à estimer l'impact de la dépense publique moyenne par élève (*expend*) sur les résultats scolaires obtenus dans le secondaire supérieur, au lycée. Supposons que les résultats scolaires d'un élève, mesurés par la moyenne obtenue à un examen national (*avgscore*), dépendent également du revenu moyen de sa famille (*avginc*), ainsi que d'autres facteurs non observés. Le modèle est :

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u. \qquad [3.2]$$

L'intérêt du décideur politique portera vraisemblablement sur l'effet *ceteris paribus* de la variable *expend* sur *avgscore*, mesuré par le coefficient β_1. En incluant *avginc* comme variable explicative dans le modèle, nous tenons explicitement compte de son effet sur *avgscore*. Il est important de le faire pour deux raisons. Le revenu familial a un impact sur les résultats scolaires *et* il est corrélé avec les dépenses publiques d'éducation. En effet, le niveau de dépense des collectivités territoriales dépend en partie de l'impôt local et de l'impôt foncier, qui eux-mêmes dépendent directement ou indirectement des revenus des ménages. Or, dans un modèle de régression simple, *avginc* est absorbé par le terme d'erreur. Ce dernier est donc corrélé avec *expend*, ce qui introduit un biais dans l'estimateur des MCO de β_1.

Dans les deux exemples précédents, nous avons montré que l'ajout de facteurs observables dans un modèle de régression multiple a permis d'aboutir à une analyse plus fine de la relation entre la variable dépendante et la variable d'intérêt, c'est-à-dire *educ* dans l'équation (3.1) et *expend* dans l'équation (3.2). De manière plus générale, nous pouvons écrire un modèle à deux variables indépendantes de la manière suivante :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \qquad [3.3]$$

où
β_0 est l'ordonnée à l'origine, ou la constante.
β_1 mesure la variation de *y suite à une variation de* x_1, les autres facteurs étant fixés.
β_2 mesure la variation de *y suite à une variation de* x_2, les autres facteurs étant fixés.

La régression multiple permet également d'affiner la modélisation de la relation fonctionnelle qui peut exister entre plusieurs variables. Par exemple, supposons que la consommation familiale (*cons*) soit une fonction quadratique du revenu familial (*inc*) :

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u \qquad [3.4]$$

où *u* comprend les autres facteurs affectant la consommation. Dans ce modèle, la consommation dépend d'un seul facteur observé, le revenu. À première vue, on pourrait penser que l'utilisation d'une régression simple pourrait convenir. Ce n'est en fait pas le cas, car ce modèle comprend deux fonctions du revenu, *inc* et inc^2 (et donc trois paramètres, β_0, β_1, et β_2). Par contre, la fonction de consommation peut être estimée facilement par un modèle de régression à deux variables indépendantes, en posant $x_1 = inc$ et $x_2 = inc^2$.

Sur un plan technique, la méthode d'estimation des paramètres des équations (3.1) et (3.4) est identique ; elle sera introduite dans la section 3.2. Chacune de ces deux équations peut s'écrire sous la forme

de (3.3). Il y a, cependant, une différence de taille dans la manière dont on *interprète* les paramètres. Dans l'équation (3.1), β_1 est l'effet *ceteris paribus* d'*educ* sur *wage*. Dans (3.4), il n'est pas possible de mesurer l'effet d'*inc* sur *cons* en gardant inc^2 constant : si *inc* varie, alors inc^2 varie aussi. Pour connaître la variation de la consommation entraînée par une variation du revenu, c'est-à-dire la propension marginale à consommer, on doit utiliser l'approximation suivante :

$$\frac{\Delta cons}{\Delta inc} \approx \beta_1 + 2\beta_2 inc$$

Si nécessaire, consultez l'annexe A pour comprendre comment cette équation s'obtient. L'effet marginal du revenu sur la consommation dépend donc de β_2, de β_1 et du niveau du revenu. Cet exemple montre que la définition des variables indépendantes est cruciale, quelle que soit l'application empirique. Nous le préciserons encore davantage dans le chapitre 6. Il est pour l'instant inutile d'aller plus loin.

Dans le modèle avec deux variables indépendantes, l'hypothèse principale concernant la relation entre u, x_1 et x_2 est

$$E(u|x_1, x_2) = 0. \qquad \textbf{[3.5]}$$

L'interprétation de la condition (3.5) est semblable à l'interprétation de l'hypothèse RLS.4 faite au cours de l'analyse de la régression simple. Elle signifie que, pour n'importe quelle valeur de x_1 et de x_2 dans la population, l'espérance des facteurs non observés est égale à zéro. Comme pour la régression simple, le point important de cette hypothèse est que la valeur espérée de u est la même *pour toutes les combinaisons possibles de* x_1 *et* x_2 ; le fait que cette valeur soit égale à zéro n'est pas contraignant tant que l'ordonnée à l'origine β_0 est incluse dans le modèle (voir la section 2.1).

Dans les exemples précédents, comment pouvons-nous interpréter l'hypothèse selon laquelle l'erreur conditionnelle est nulle en moyenne ? Dans l'équation (3.1), l'hypothèse est E(*u*|*educ*, *exper*) = 0. Cela signifie que les autres facteurs influant sur *wage* ne sont en moyenne pas reliés à *educ* et *exper*. Par conséquent, si nous pensons que les capacités innées font partie de *u*, nous devons faire l'hypothèse que les capacités innées sont en moyenne identiques pour toutes les combinaisons de niveau d'études et d'expérience dans la population active. Cela peut être vrai ou faux. Comme nous le verrons dans la section 3.3, nous devons absolument nous poser cette question pour déterminer si la méthode des moindres carrés produira des estimateurs non biaisés.

L'équation (3.2) portant sur les résultats scolaires est similaire à celle sur le salaire. L'hypothèse de l'espérance conditionnelle nulle est E(*u*|*expend*, *avginc*) = 0, ce qui implique que les autres facteurs affectant les résultats scolaires (comme les caractéristiques de l'école ou de l'élève) sont, en moyenne, sans lien avec les dépenses moyennes par élève ou le revenu familial moyen.

Pour aller plus loin 3.1

Nous voulons expliquer le taux de criminalité urbain (*murdrate*) en fonction de la probabilité d'être condamné (*prbconv*) et de la durée moyenne des peines (*avgsen*). Soit le modèle de régression multiple suivant :

$$murdrate = \beta_0 + \beta_1\, prbconv + \beta_2 avgsen + u.$$

Quels sont les facteurs contenus dans *u* ? Pensez-vous que l'hypothèse clé (3.5) soit vérifiée ?

Quand l'hypothèse de l'espérance conditionnelle nulle de l'erreur est appliquée à la fonction quadratique de consommation décrite en (3.4), elle a une interprétation légèrement différente. Écrite sous sa forme littérale, l'égalité (3.5) implique E(*u*|*inc*, inc^2) = 0. Comme inc^2 est connu lorsque *inc* l'est, inclure inc^2 dans le terme d'espérance est redondant : E(*u*|*inc*, inc^2) = 0 est identique à E(*u*|*inc*) = 0. Lorsque nous posons l'hypothèse (3.5), il n'est pas faux d'ajouter inc^2 à *inc* mais E(*u*|*inc*) = 0 est plus concis.

Le modèle avec k variables indépendantes

Dans un modèle de régression multiple, il n'y a pas de raison de se limiter à deux variables indépendantes. Dans la régression multiple, de nombreux facteurs observés peuvent affecter y. Dans l'exemple du salaire, nous pourrions ajouter le temps de formation, le nombre d'années d'expérience acquises auprès de l'employeur actuel, une mesure des capacités innées, et même des variables démographiques, comme le nombre de frères et sœurs ou le nombre d'années d'études de la mère. Dans l'exemple des résultats scolaires, nous pourrions ajouter, comme variables supplémentaires, la qualité des enseignants, la taille de l'école, la taille de la classe, etc.

Le **modèle de régression linéaire multiple** dans la population peut prendre la forme générale suivante :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u \qquad [3.6]$$

avec
β_0, l'ordonnée à l'origine, ou la constante
β_1, le paramètre associé à x_1,
β_2, le paramètre associé à x_2, et ainsi de suite.

Comme il y a k variables indépendantes et une ordonnée à l'origine, l'équation (3.6) contient $k + 1$ paramètres de population (inconnus). Par souci de simplification, les paramètres associés aux variables explicatives seront souvent appelés **paramètres de la pente**, bien que ce terme n'indique pas toujours ce qu'ils représentent exactement. [Par exemple, dans l'équation (3.4), ni β_1 ni β_2 ne constituent une pente en tant que tels ; ils déterminent ensemble la pente du rapport entre la consommation et le revenu].

La terminologie de la régression multiple est semblable à celle de la régression simple ; elle est donnée dans le tableau 3.1. Comme dans la régression simple, la variable u est le **terme d'erreur**, ou la **perturbation**. Il contient les autres facteurs, différents de x_1, x_2, …, x_k, qui affectent y. Quel que soit le nombre de variables explicatives dans notre modèle, il y aura toujours des facteurs que nous ne pourrons pas inclure ; ils seront tous compris dans u.

Pour pouvoir utiliser le modèle de régression multiple, nous devons savoir comment interpréter les paramètres. Avant de nous pencher sur cette question, il est utile de rappeler certains éléments que nous avons déjà abordés auparavant. Supposons que le salaire d'un PDG (*salary*) dépende du chiffre d'affaire de son entreprise (*sales*) et de son ancienneté à ce poste au sein l'entreprise (*ceoten*), selon la relation suivante :

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u. \qquad [3.7]$$

Tableau 3.1 *Terminologie de base dans le modèle de régression multiple*

Y	$x_1, x_2, \ldots, x_k$
Variable dépendante	Variable indépendante
Variable expliquée	Variable explicative
Variable de réponse	Variable de contrôle
Variable prédite	Variable prédictive
Variable endogène	Variable exogène

Cela correspond bien à un modèle de régression multiple (avec $k = 3$) pour lequel on aurait $y = \log(salary)$, $x_1 = \log(sales)$, $x_2 = ceoten$, et $x_3 = ceoten^2$. Comme nous l'avons vu dans le chapitre 2, le coefficient β_1 est *l'élasticité* (*ceteris paribus*) de *salary* par rapport à *sales*. Par ailleurs, si $\beta_3 = 0$, alors $100\beta_2$ donne approximativement

le pourcentage d'augmentation de *salary* quand *ceoten* augmente d'une unité – ici une année, toutes choses égales par ailleurs. Quand $\beta_3 \neq 0$, l'effet de *ceoten* sur *salary* est plus compliqué. Les modèles qui incorporent un terme quadratique feront l'objet d'une étude plus approfondie dans le reste de l'ouvrage, en particulier au chapitre 6.

L'équation (3.7) nous permet de souligner une dimension importante de l'analyse par régression multiple. Le terme « linéaire », dans un modèle de régression multiple, signifie que l'équation (3.6) est *linéaire par rapport à ses paramètres*, les β_j. L'équation (3.7) est un exemple de régression multiple où le modèle est linéaire par rapport aux β_j mais où la relation entre *salary* et les variables *explicatives* n'est pas linéaire. De nombreuses applications de la régression multiple impliquent des relations non linéaires entre la variable expliquée et les variables explicatives.

L'hypothèse fondamentale du modèle de régression multiple est simple à exprimer :

$$E(u|x_1, x_2, \ldots, x_k) = 0. \qquad \textbf{[3.8]}$$

L'équation (3.8) requiert qu'il n'existe aucune sorte de lien entre les facteurs inclus dans le terme d'erreur et les variables explicatives présentes dans le modèle. Elle exige également que nous ayons correctement appréhendé les relations fonctionnelles entre la variable expliquée et les variables explicatives. Tout problème qui entrainerait une corrélation entre u et une des variables indépendantes invaliderait (3.8). Dans la section 3.3, nous montrerons que l'hypothèse (3.8) implique que les MCO (moindres carrés ordinaires) ne sont pas biaisés. Nous calculerons également le biais qui résulte de l'omission d'une variable importante. Dans les chapitres 15 et 16, nous étudierons les autres raisons qui peuvent invalider (3.8) et nous expliquerons ce qui peut être fait pour y remédier.

3.2 UNE INTERPRÉTATION DE LA RÉGRESSION MULTIPLE EN TERMES D'EFFET PARTIEL

Nous allons à présent résumer les caractéristiques statistiques et algébriques essentielles de la méthode des moindres carrés ordinaires lorsqu'elle est appliquée à un ensemble particulier de données. Nous allons également discuter de la manière dont il convient d'interpréter les résultats de l'équation estimée.

Le calcul des estimateurs des MCO

Pour commencer, nous allons estimer le modèle avec deux variables indépendantes. Comme pour la régression simple, nous pouvons écrire l'équation estimée par les MCO de la manière suivante :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \qquad \textbf{[3.9]}$$

avec

$\hat{\beta}_0$ = l'estimation de β_0.
$\hat{\beta}_1$ = l'estimation de β_1.
$\hat{\beta}_2$ = l'estimation de β_2.

Comment pouvons-nous obtenir $\hat{\beta}_0$, $\hat{\beta}_1$, et $\hat{\beta}_2$? La méthode des **moindres carrés ordinaires** sélectionne les estimations qui minimisent la somme des carrés des résidus. Autrement dit, pour n observations de y, x_1, et x_2, $\{(x_{i1}, x_{i2}, y_i) : i = 1, 2, \ldots, n\}$, les estimations $\hat{\beta}_0$, $\hat{\beta}_1$, et $\hat{\beta}_2$ sont simultanément sélectionnées de manière à ce que

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \qquad \textbf{[3.10]}$$

soit le plus petit possible.

Pour comprendre ce que les MCO impliquent, il est important d'insister sur la signification des indices dans (3.10). Les variables indépendantes ont deux éléments en indice : i suivi par 1 ou 2. L'indice i indique le numéro de l'observation. Dans (3.10), la somme est donc calculée sur toutes les observations, i allant de 1 à n. Le deuxième indice, c'est-à-dire le chiffre suivant i, permet de distinguer les différentes variables indépendantes. Dans l'exemple liant *wage* à *educ* et à *exper*, $x_{i1} = educ_i$ indique le niveau d'études de l'individu i dans l'échantillon, et $x_{i2} = exper_i$ est l'expérience de cet individu i. La somme des carrés des résidus dans l'équation (3.10) est $\sum_{i=1}^{n}(wage_i - \hat{\beta}_0 - \hat{\beta}_1 educ_i - \hat{\beta}_2 exper_i)^2$. Par la suite, nous utiliserons systématiquement l'indice i pour indiquer le numéro de l'observation. En écrivant « x_{ij} », nous nous référons à la $i^{ème}$ observation de la $j^{ème}$ variable indépendante. (Certains auteurs préfèrent inverser cet ordre ; dans ce cas, x_{ij} serait l'observation i de la première variable. C'est juste une question de notation.)

Dans le cas général avec k variables indépendantes, nous cherchons les estimations $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k$ de l'équation suivante

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \qquad \textbf{[3.11]}$$

Les estimations des MCO, au nombre de $k + 1$, sont choisies afin de minimiser la somme des carrés des résidus :

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2 \qquad \textbf{[3.12]}$$

Ce problème de minimisation peut être résolu en utilisant le calcul multivarié (voir l'annexe 3A). Cela conduit à $k + 1$ équations linéaires à $k + 1$ inconnues $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k$:

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \\
&\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \\
&\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \\
&\vdots \\
&\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0
\end{aligned} \qquad \textbf{[3.13]}$$

Elles sont souvent appelées les **conditions du premier ordre** des MCO. Comme nous l'avons expliqué dans la section 2.2 pour la régression simple, les conditions du premier ordre des MCO peuvent être obtenues par la méthode des moments : sous l'hypothèse (3.8), $E(u) = 0$ et $E(x_j) = 0$, avec $j = 1, 2, \ldots, k$. Les équations figurant dans (3.13) sont les équivalents d'échantillonnage des moments de la population, bien que nous n'ayons pas divisé par la taille de l'échantillon, n.

Même pour des valeurs peu élevées de n et de k, résoudre à la main les équations (3.13) est particulièrement fastidieux. Heureusement, les logiciels d'économétrie permettent de résoudre rapidement ces équations, même pour de très grandes valeurs de n et k.

Il y a néanmoins un point auquel nous devons prêter attention : nous devons supposer que les équations (3.13) ont une solution unique pour les $\hat{\beta}_j$. Cette hypothèse est souvent vérifiée dans les modèles

qui ont été bien spécifiés. C'est dans la section 3.3 que nous précisons l'hypothèse permettant d'obtenir des estimations des MCO uniques (voir l'hypothèse RLM.3).

Comme dans la régression simple, l'équation (3.11) est appelée la **droite de régression des MCO** ou la **fonction de régression de l'échantillon (FRE)**. $\hat{\beta}_0$ est **l'estimation de l'ordonnée à l'origine par les MCO** (ou estimation de la constante par les MCO) ; $\hat{\beta}_1, \ldots, \hat{\beta}_k$ représentent **les estimations de la pente par les MCO** correspondant aux variables indépendantes $x_1, x_2, \ldots, x_k$.

Pour indiquer que nous avons procédé à une régression estimée par les MCO, nous dirons que « nous avons réalisé une régression des MCO de y *sur* $x_1, x_2, \ldots, x_k$ » ou bien que « nous avons régressé y sur $x_1, x_2, \ldots, x_k$ », en veillant à remplacer les variables y et les x_j par leur nom respectif (comme par exemple *wage*, *educ*, *exper*). Nous indiquons ainsi que nous avons utilisé la méthode des moindres carrés ordinaires afin d'obtenir les estimations de l'équation (3.11). À défaut d'indication contraire, nous inclurons toujours une ordonnée à l'origine dans la régression.

Interprétation de l'équation de régression des MCO

L'interprétation de l'équation estimée est plus importante que les détails qui sous-tendent le calcul des β_j. Nous commençons par le cas où deux variables explicatives sont présentes dans l'équation :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \qquad \textbf{[3.14]}$$

$\hat{\beta}_0$ représente la valeur estimée (ou prédite) de y quand $x_1 = 0$ et $x_2 = 0$. Dans certains cas, l'estimation de l'ordonnée à l'origine donne des informations intéressantes ; dans d'autres, elle n'a pas de sens. Notons cependant que cette ordonnée à l'origine est indispensable si nous désirons obtenir une estimation de y à partir de la droite de régression des MCO lorsque $x_1 = 0$ et $x_2 = 0$. Sans cette ordonnée à l'origine, la valeur de y sera toujours nulle lorsque $x_1 = 0$ et $x_2 = 0$.

Les estimations $\hat{\beta}_1$ et $\hat{\beta}_2$ s'interprètent comme des **effets marginaux** ou des **effets *ceteris paribus***. De l'équation (3.14), nous déduisons que

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

de manière à obtenir la variation estimée (ou prédite) de y suite aux variations de x_1 et x_2. (Notons d'ailleurs que l'ordonnée à l'origine n'a rien à voir avec les variations de y.) Par exemple, quand x_2 est maintenue constante, de telle sorte que $\Delta x_2 = 0$, alors

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1$$

pour x_2 fixé. Le point crucial est que l'inclusion de x_2 dans notre modèle permet de donner une interprétation *ceteris paribus* au coefficient de x_1. La régression multiple est donc plus utile que la régression simple à cet égard. De même,

$$\Delta\hat{y} = \hat{\beta}_2 \Delta x_2$$

avec x_1 fixé.

EXEMPLE 3.1
Les déterminants des résultats obtenus à l'université

La base de données GPA1 contient des informations sur un échantillon de 141 étudiants inscrits dans une grande université américaine. Ces informations incluent la moyenne générale des notes obtenus à l'université (*colGPA*, soit *college Grade Point Average*), la moyenne générale des notes au lycée (*hsGPA*, soit *high school GPA*), et la note obtenue à un test d'évaluation utilisé pour accéder aux études supérieures (*ACT*, soit *American College Testing*). Les moyennes générales obtenues au lycée et à l'université sont calculées sur une échelle qui va de 1 à 4 ; le test *ACT* est noté sur un total de 36 points. En cherchant à expliquer la moyenne générale à l'université à partir de la moyenne générale au lycée et de la note obtenue au test *ACT*, nous obtenons la droite de régression des MCO suivante :

$$\widehat{colGPA} = 1{,}29 + 0{,}453\,hsGPA + 0{,}0094\,ACT \qquad [3.15]$$
$$n = 141.$$

Comment interpréter cette équation ? D'abord, l'ordonnée à l'origine, égale à 1,29, correspond à l'estimation de la moyenne universitaire lorsque les variables *hsGPA* et *ACT* sont toutes deux égales à zéro. Comme aucun étudiant ne peut accéder à l'université en ayant obtenu zéro à l'une de ces deux évaluations, cette ordonnée à l'origine n'a pas de sens en elle-même.

Les estimations des coefficients de *hsGPA* et de *ACT* sont plus intéressantes. Comme on pouvait s'y attendre, l'effet *ceteris paribus* de *hsGPA* sur *colGPA* est positif : si nous maintenons *ACT* inchangé, un point de plus de moyenne générale au lycée est associé à une augmentation de *colGPA* égale à un demi-point environ (soit 0,453 point). Autrement dit, si nous choisissons deux étudiants, A et B, dont les résultats au test de niveau *ACT* sont identiques, nous pouvons prédire que l'étudiant A aura une moyenne à l'université plus élevée de 0,453 que celle de l'étudiant B *si l'étudiant A est parvenu à obtenir un point de plus que l'étudiant B à* hsGPA. Notez que cela ne nous renseigne pas nécessairement sur deux étudiants particuliers, qui seraient repris dans l'échantillon ; il s'agit simplement d'une prédiction basée sur toute l'information disponible dans notre échantillon et obtenue à l'aide d'un modèle de régression multiple.

Le signe de *ACT* implique que si nous maintenons *hsGPA* inchangé, un changement de 10 points dans la note obtenue au test *ACT* affectera *colGPA* d'un dixième de point environ. Un changement de 10 points est une amélioration considérable de la note au test *ACT*. En effet, la moyenne de l'échantillon est égale à 24 environ ; l'écart-type est inférieur à 3, pour une note maximale égale à 36. Par contre, un effet positif d'un dixième de point est minime. Par conséquent, à partir du moment où la moyenne générale au lycée (*hsGPA*) est prise en compte, la réussite au test *ACT* n'est pas un facteur déterminant lorsqu'il s'agit de prédire la réussite d'un étudiant à l'université. (Naturellement, de nombreux autres facteurs contribuent à la moyenne ; ici nous tenons uniquement compte de statistiques de réussite disponibles pour les élèves du lycée.) Dans le chapitre 4, nous montrerons que le coefficient d'*ACT* n'est pas différent de zéro sur le plan statistique.

Si nous effectuons une régression simple de *colGPA* sur *ACT*, nous obtenons

$$\widehat{colGPA} = 2{,}40 + 0{,}0271\,ACT$$
$$n = 141.$$

Dans ce cas, le coefficient d'*ACT* est presque trois fois plus grand que l'estimation que nous avons obtenue dans (3.15). Notez bien que le modèle de régression simple ne nous *permet pas* de comparer deux personnes qui ont la même moyenne générale au lycée. Elle correspond à une expérience différente. Nous approfondirons les différences entre les régressions simple et multiple ultérieurement.

Le raisonnement est similaire lorsque nous incluons plus de deux variables indépendantes dans le modèle de régression multiple. La droite de régression des MCO est :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \qquad [3.16]$$

ou, écrite en termes de changements :

$$\Delta\hat{y} = \hat{\beta}_1\Delta x_1 + \hat{\beta}_2\Delta x_2 + \ldots + \hat{\beta}_k\Delta x_k \qquad \textbf{[3.17]}$$

Toutes choses étant égales par ailleurs, le coefficient de x_1 mesure le changement de $\hat{y}$ dû à une augmentation de x_1 d'une unité. Soit :

$$\Delta\hat{y} = \hat{\beta}_1\Delta x_1 \qquad \textbf{[3.18]}$$

en gardant x_2, x_3, ..., x_k constants. De cette manière, nous *neutralisons* l'effet des variables x_2, x_3, ..., x_k quand nous estimons l'effet de x_1 sur y. Les autres coefficients ont une interprétation similaire.

L'encadré 3.2 offre un exemple basé sur trois variables indépendantes.

EXEMPLE 3.2
L'équation du salaire horaire

En utilisant les 526 observations de WAGE1, nous cherchons à expliquer le salaire horaire d'employés américains sur base de plusieurs variables : *educ* (les années d'études), *exper* (les années d'expérience sur le marché du travail) et *tenure* (les années d'ancienneté avec l'employeur actuel). L'équation estimée est

$$\widehat{\log(wage)} = 0{,}284 + 0{,}092\,educ + 0{,}0041\,exper + 0{,}022\,tenure \qquad \textbf{[3.19]}$$

$$n = 526.$$

Comme dans le cas de la régression simple, l'interprétation des coefficients doit se faire en pourcentage [puisque y est en *log* et que les x_j sont en niveau]. Par contre, dans la régression multiple, ces coefficients ont une interprétation *ceteris paribus*. Le coefficient 0,092 signifie que, si nous maintenons *exper* et *tenure* constants, une année d'études en plus doit augmenter log(*wage*) de 0,092, soit une augmentation d'environ 9,2 % de *wage*. Autrement dit, si nous considérons deux individus dont les niveaux d'expérience et d'ancienneté sont les mêmes, le coefficient d'*educ* donne une estimation de la variation de salaire imputable à une année d'études supplémentaire, exprimée en pourcentage du salaire de départ. Cette estimation des rendements de l'éducation s'obtient en maintenant inchangés deux facteurs de productivité importants. Afin de déterminer s'il s'agit d'une bonne mesure du rendement d'une année d'études supplémentaire, nous devons analyser les propriétés statistiques des MCO (voir section 3.3).

Sur la signification de ceteris paribus *dans la régression multiple*

L'interprétation des coefficients de la pente comme des effets marginaux est souvent source de confusion ; c'est la raison pour laquelle nous y revenons de nouveau.

Dans l'exemple 3.1, nous observons que le coefficient d'*ACT* mesure la différence prédite de *colGPA* pour un niveau donné de *hsGPA*. L'intérêt de la régression linéaire multiple est qu'elle fournit une interprétation *ceteris paribus*, même si les données n'ont *pas* été collectées dans des conditions expérimentales qui permettent d'avoir tous les facteurs externes sous contrôle. Pour pouvoir donner une interprétation *ceteris paribus* au coefficient d'*ACT*, on pourrait penser qu'il faut nécessairement interroger des individus dont la moyenne générale au lycée était la même mais dont les résultats au test *ACT* étaient différents. Ce n'est pas le cas. Les données proviennent d'un échantillon aléatoire au sein d'une grande université américaine : au moment de la collecte des données, il n'y avait aucune restriction quant aux valeurs que *hsGPA* ou *ACT* pouvaient prendre dans l'échantillon. En sciences sociales, nous avons rarement le luxe de maintenir certaines variables fixes dans notre échantillon. *Si* nous avions pu collecter un échantillon d'individus ayant la même moyenne générale au lycée, nous pourrions effectuer une analyse de régression simple reliant *colGPA* à *ACT*.

La régression multiple nous permet de simuler cette situation, sans pour autant restreindre les valeurs que peuvent prendre les autres variables explicatives.

La force de la régression linéaire multiple est qu'elle nous permet d'agir, dans un environnement non expérimental, comme si nous étions dans un laboratoire de sciences naturelles, dans des conditions d'expérimentation contrôlées qui nous permettent de maintenir les autres facteurs fixés.

Faire varier plusieurs variables indépendantes en même temps

Dans la régression multiple, nous pouvons aisément faire varier plusieurs variables explicatives en même temps pour obtenir leur effet global sur la variable dépendante, comme l'indique l'équation (3.17). Par exemple, dans l'équation (3.19), nous pouvons obtenir l'effet estimé sur *wage* si un employé décide de rester une année de plus dans la même entreprise. Dans un tel cas de figure, *exper* (l'expérience professionnelle) et *tenure* (l'ancienneté) augmentent chacune d'une année. En gardant *educ* fixée (*ceteris paribus*), l'effet total est

$$\Delta\widehat{\log(wage)} = 0{,}0041\Delta exper + 0{,}022\Delta tenure = 0{,}0041 + 0{,}022 = 0{,}0261$$

soit environ 2,6 %. Comme *exper* et *tenure* augmentent chacune d'une année, nous ajoutons les coefficients associés à *exper* et *tenure* et nous les multiplions par 100 pour calculer l'effet en pourcentage.

Valeurs ajustées et résidus des MCO

Après avoir obtenu la droite de régression (3.11), nous pouvons obtenir une valeur ajustée, ou valeur prédite, pour chaque observation. Pour l'observation i, la valeur ajustée est simplement :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} \qquad \textbf{[3.20]}$$

qui est la valeur prédite obtenue en utilisant les valeurs des variables indépendantes pour l'observation i dans l'équation (3.11). Nous ne devons pas oublier l'ordonnée à l'origine lorsque nous calculons la valeur ajustée ; sinon, nous courrons le risque de commettre des erreurs importantes de calcul. Par exemple, dans (3.15), si $hsGPA_i = 3{,}5$ et $ACT_i = 24$, alors $colGPA = 1{,}29 + 0{,}453(3{,}5) + 0{,}0094(24) = 3{,}101$ (en arrondissant à la troisième décimale).

Normalement, pour n'importe quelle observation i, la vraie valeur y_i, celle que nous observons dans l'échantillon, ne sera pas égale à sa valeur prédite : les MCO minimisent l'erreur *moyenne* de prévision au carré, ce qui ne nous renseigne pas sur l'erreur de prévision pour *une* observation donnée. Le **résidu** de l'observation i est défini de la même manière que pour la régression simple :

$$\hat{u}_i = y_i - \hat{y}_i \qquad \textbf{[3.21]}$$

Il y a un résidu pour chaque observation. Si $\hat{u}_i > 0$, alors $\hat{y}_i$ est inférieur à y_i, ce qui signifie que, pour cette observation, y_i est sous-estimé. Si $\hat{u}_i < 0$, alors $y_i < \hat{y}_i$, et y_i est surestimé.

Les valeurs ajustées et les résidus des MCO ont des propriétés importantes qui sont des extensions immédiates du cas à une seule variable :

1. Comme la moyenne des résidus est égale à zéro, $\bar{y} = \bar{\hat{y}}$.

2. Comme la covariance d'échantillon entre chaque variable indépendante et les résidus des MCO est égale à zéro, la covariance d'échantillon entre les valeurs ajustées des MCO et les résidus des MCO est aussi égale à zéro.

3. Le point $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$ est toujours situé sur la droite de régression.

Pour aller plus loin 3.2

Dans l'exemple 3.1, nous avons expliqué comment calculer la moyenne générale obtenue à l'université (*colGPA*) à l'aide de deux variables : la moyenne générale obtenue au lycée (*hsGPA*) et la note obtenue au test *ACT*. La droite de régression des MCO est :

$$\widehat{colGPA} = 1{,}29 + 0{,}453\,hsGPA + 0{,}0094\,ACT.$$

Si la moyenne générale au lycée est 3,4 et que la note moyenne au test ACT est 24,2, quelle sera la moyenne de *colGPA* dans l'échantillon ?

Les deux premières propriétés sont des conséquences immédiates de l'ensemble des équations utilisées pour obtenir les estimations des MCO. La première équation de (3.13) dit que la somme des résidus est égale à zéro. Les autres équations sont de la forme $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, ce qui implique que chaque variable indépendante a une covariance d'échantillon nulle avec $\hat{u}_i$. La propriété (3) découle directement de la propriété (1).

Une interprétation de la régression linéaire multiple en termes d'effet net

Quand nous utilisons les MCO dans la pratique, nous n'avons pas besoin de connaître les formules explicites des $\hat{\beta}_j$ qui résolvent le système d'équation (3.13). Cependant, ces formules sont bien utiles pour certains calculs. Elles permettent également de mieux comprendre la méthode des MCO.

Considérons à nouveau le cas basé sur $k = 2$ variables indépendantes, soit $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Pour plus de facilité, concentrons-nous sur $\hat{\beta}_1$. Une manière d'exprimer $\hat{\beta}_1$ est

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} \hat{r}_{i1} y_i\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right) \qquad \textbf{[3.22]}$$

où les $\hat{r}_{i1}$ correspondent aux résidus d'une régression simple de x_1 sur x_2 en utilisant le même échantillon. Nous régressons notre première variable indépendante, x_1, sur notre seconde variable indépendante, x_2 ; nous obtenons ensuite les résidus, y ne jouant aucun rôle ici. Comme l'équation (3.22) le montre, il suffit ensuite de régresser y sur $\hat{r}_1$ pour obtenir $\hat{\beta}_1$. (Notons que les résidus $\hat{r}_1$ ont une moyenne d'échantillon nulle ; $\hat{\beta}_1$ est donc l'estimation de la pente habituellement obtenue par une régression simple.)

L'équation (3.22) offre une interprétation alternative de $\hat{\beta}_1$ en termes d'effet partiel. Les résidus $\hat{r}_{i1}$ sont la partie de x_{i1} qui n'est pas corrélée avec x_{i2}. Pour le dire autrement, $\hat{r}_{i1}$ n'est rien d'autre que x_{i1} dont nous avons *retiré* ou *déduit* les effets de x_{i2} ; il s'agit de x_{i1} *nettoyé* de l'influence de x_{i2}. Par conséquent, $\hat{\beta}_1$ mesure la relation entre y et x_1 dans l'échantillon, *après avoir purgé* x_1 *de l'effet de* x_2. C'est en ce sens que $\hat{\beta}_1$ mesure l'effet *marginal* de x_1.

Dans une régression simple, il n'existe pas d'interprétation en termes d'effet *net de l'influence d'autres variables*, puisque la régression simple n'inclut aucune variable de contrôle. Dans l'exercice sur ordinateur C5, qui repose sur les mêmes données de salaire que celles utilisées dans l'exemple 3.2, nous aurons à purger une variable explicative de l'effet d'une autre variable afin d'en identifier l'effet *net* sur y ; il faudra donc se débarrasser de l'influence des autres variables explicatives avant de pouvoir calculer cet effet. Dans l'équation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, l'élément important sur un plan pratique est que $\hat{\beta}_1$ mesure bien la variation de y suite à l'augmentation de x_1 d'une unité, x_2 étant égal par ailleurs.

L'équation (3.22) permet également de calculer $\hat{\beta}_1$ dans le modèle général avec k variables explicatives, mais les résidus $\hat{r}_{i1}$ proviennent de la régression de x_1 sur x_2, ..., x_k. Encore une fois, $\hat{\beta}_1$ mesure l'effet de x_1 sur y, *net* de l'influence de x_2, ..., x_k. En économétrie, le résultat général d'interprétation en termes d'effet net est connu sous le nom de théorème de Frish-Waugh. Il a de nombreuses utilisations en économétrie théorique et appliquée. Nous verrons comment il s'applique aux régressions de séries temporelles dans le chapitre 10.

Comparaison des estimations par régressions simple et multiple

Il existe deux cas spécifiques où la régression simple de y sur x_1 produira les mêmes résultats pour x_1 que la régression multiple de y sur x_1 *et* x_2. Pour plus de clarté et de précision, écrivons la régression simple de y sur x_1 comme suit : $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$; écrivons la régression multiple comme suit : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Nous savons que le coefficient de la régression simple, $\tilde{\beta}_1$, n'est généralement pas égal au coefficient de la régression multiple, $\hat{\beta}_1$. En réalité, il existe une relation simple entre $\hat{\beta}_1$ et $\tilde{\beta}_1$, qui permet de mieux comparer la régression simple et la régression multiple :

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1 \qquad [3.23]$$

où $\tilde{\delta}_1$ est le coefficient de la pente de la régression simple de x_{i2} sur x_{i1}, $i = 1, \ldots, n$. Cette équation nous montre comment $\tilde{\beta}_1$ diffère de l'effet *ceteris paribus* de x_1 sur $\hat{y}$. Le second terme de l'équation (3.23) correspond à l'effet *ceteris paribus* de x_2 sur $\hat{y}$, multiplié par la pente de la régression simple de x_2 sur x_1. (Consultez la section 3A.4 dans les annexes pour une démonstration plus générale.)

La relation entre $\hat{\beta}_1$ et $\tilde{\beta}_1$ montre qu'il existe deux cas particuliers dans lesquels $\hat{\beta}_1$ et $\tilde{\beta}_1$ sont égaux :

1. L'effet marginal de x_2 sur $\hat{y}$ est nul dans l'échantillon. Soit $\hat{\beta}_2 = 0$.
2. Les variables x_1 et x_2 ne sont pas corrélées dans l'échantillon. Soit $\tilde{\delta}_1 = 0$.

Même si les estimations obtenues par la régression simple ne sont presque jamais égales à celles de la régression multiple, nous pouvons utiliser la formule précédente pour identifier les raisons pour lesquelles elles peuvent différer. Par exemple, si la valeur de $\hat{\beta}_2$ est faible, nous pouvons nous attendre à ce que les estimations de β_1 ne soient pas influencées par le choix de la régression simple ou de la régression multiple. Dans l'exemple 3.1, la corrélation d'échantillon entre *hsGPA* (x_1) et *ACT* (x_2) est d'environ 0,346 ; cette corrélation, correspondant à $\tilde{\delta}_1$, est loin d'être négligeable. Par contre, le coefficient d'*ACT* ($\hat{\beta}_2$) est relativement faible. Il n'est donc pas surprenant que la régression simple de *colGPA* sur *hsGPA* produise une estimation de la pente proche de l'estimation obtenue en (3.15) par la régression multiple (0,482 *versus* 0,453).

Dans le cas où nous incluons k variables indépendantes, la régression simple de y sur x_1 et la régression multiple de y sur x_1, x_2, ..., x_k ne donneront des estimations *identiques* de l'effet de x_1 sur y que si : (1) les coefficients des MCO de x_2, ..., x_k sur x_1 sont tous égaux à zéro ; *ou* (2) la variable x_1 n'est corrélée avec *aucune* des variables x_2, ..., x_k. Ni (1) ni (2) ne sont très réalistes. Par contre, si les estimations des coefficients de x_2 à x_k sur x_1 sont faibles, *ou* si la corrélation d'échantillon entre x_1 et les autres variables indépendantes est faible, alors la régression simple et la régression multiple donneront des estimations *similaires* de l'effet de x_1 sur y.

EXEMPLE 3.3
Participation des employés au plan d'épargne-retraite *401(k)*

Nous utilisons la base de données 401K portant sur la participation des travailleurs au plan d'épargne-retraite *401(k)* ; il s'agit d'un système par capitalisation très largement utilisé aux États-Unis. Notre objectif est d'estimer l'effet du taux de contribution des entreprises (*mrate*) sur le taux de participation des travailleurs (*prate*) à ce plan. Le taux de contribution des entreprises est le montant que l'entreprise investit dans le fond de pension pour chaque dollar que le travailleur verse (jusqu'à une certaine limite).

Ainsi, *mrate* = 0,75 signifie que l'entreprise verse 75 cents pour chaque dollar versé par le travailleur. Le taux de participation est le pourcentage de travailleurs qui ont un compte *401(k)* parmi les travailleurs éligibles. Il y a 1 534 fonds de pension dans le jeu de données ; le *prate* moyen est 87,36 (%) ; le *mrate* moyen est 0,732, et l'ancienneté moyenne des plans d'épargne-pension (*age*) est égale à 13,2 ans.

Si on régresse *prate* sur *mrate* et *age*, on obtient :

$$\widehat{prate} = 80{,}12 + 5{,}52\,mrate + 0{,}243\,age$$
$$n = 1\,534.$$

Nous pouvons constater que *mrate* et *age* influencent *prate* dans le sens attendu, étant donné le signe positif des coefficients estimés. Que se passe-t-il si nous jugeons qu'il est inutile de tenir compte de l'effet de la variable *age* dans cette régression ? L'effet estimé de la variable *age* est loin d'être négligeable. Nous pouvons donc nous attendre à un changement important de l'effet estimé de *mrate* si nous décidons de retirer *age* de la régression. En réalité, la régression simple de *prate* sur *mrate* nous donne $\widehat{prate} = 83{,}08 + 5{,}86\,mrate$. Certes, l'estimation de l'effet de *mrate* sur *prate* dans la régression simple est différente de l'estimation de la régression multiple, mais cette différence n'est pas très grande. En effet, l'estimation de la régression simple est supérieure de seulement 6,2 % à celle de la régression multiple. Cette faible différence peut s'expliquer par le fait que la corrélation entre *mrate* et *age* est elle-même faible : elle n'est que de 0,12.

Qualité de l'ajustement

Comme pour la régression simple, nous pouvons définir la **somme des carrés totaux (SCT),** la **somme des carrés expliqués (SCE)** et la **somme des carrés des résidus (SCR)** de la manière suivante :

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \textbf{[3.24]}$$

$$SCE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \textbf{[3.25]}$$

$$SCR = \sum_{i=1}^{n} \hat{u}_i^2 \qquad \textbf{[3.26]}$$

En utilisant le même raisonnement que dans le cas de la régression simple, nous pouvons montrer que

$$\text{SCT} = \text{SCE} + \text{SCR}. \qquad \textbf{[3.27]}$$

Autrement dit, la variation totale des $\{y_i\}$ est la somme des variations totales des $\{\hat{y}_i\}$ et des $\{\hat{u}_1\}$.

Si nous faisons l'hypothèse que la variation totale de y n'est pas nulle, ce qui est le cas si y_i n'est pas constant dans l'échantillon, nous pouvons diviser (3.27) par SCT pour obtenir :

$$\text{SCR/SCT} + \text{SCE/SCT} = 1.$$

Comme dans le cas de la régression simple, le R carré est défini de la manière suivante :

$$R^2 = \text{SCE/SCT} = 1 - \text{SCR/SCT} \qquad \textbf{[3.28]}$$

et il est interprété comme la proportion de la variation de y_i dans l'échantillon qui est expliquée par la régression des MCO. Par définition, le R^2 est un nombre entre zéro et un.

On peut aussi montrer que le R^2 est égal au carré du coefficient de corrélation entre les valeurs observées y_i et leurs valeurs prédites $\hat{y}_i$. Soit :

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)} \qquad [3.29]$$

[Nous avons utilisé la moyenne de $\hat{y}$ dans (3.29) pour être cohérent avec la formule du coefficient de corrélation ; nous savons que cette moyenne est égale à $\bar{y}$ car $y_i = \hat{y}_i + \hat{u}_i$ et la moyenne d'échantillon des résidus est nulle.]

Une caractéristique importante du R^2 est qu'il ne diminue jamais ; il augmente même presque systématiquement quand on ajoute une variable explicative dans la régression. En raison de sa définition algébrique, la somme du carré des résidus ne peut jamais augmenter lorsqu'un nouveau régresseur est ajouté au modèle. Par exemple, le dernier chiffre du numéro de sécurité sociale d'un travailleur n'a rien à voir avec son salaire horaire ; pourtant, si on l'ajoute à une équation de salaire, le R^2 augmentera – ne fut-ce que très légèrement, en raison du hasard.

Ce qui vient d'être dit sur le R carré repose sur l'hypothèse que les différentes variables omises n'ont pas de valeurs manquantes. Si deux régressions mobilisent des ensembles d'observations différents, nous ne pouvons pas comparer leurs R carrés, même si une des deux régressions utilise un sous-ensemble des régresseurs de l'autre. Si on a l'ensemble des données pour les variables y, x_1 et x_2, mais que pour certaines observations, les données sur x_3 manquent, alors nous ne pouvons pas dire que le R carré de la régression de y sur x_1 et x_2 sera inférieur au R carré de la régression de y sur x_1, x_2 et x_3. Il peut être ou plus grand, ou plus petit. Les valeurs manquantes peuvent engendrer des problèmes pratiques, et nous y reviendrons dans le chapitre 9.

Comme le R^2 ne diminue jamais après l'ajout d'une variable *quelle qu'elle soit*, ce n'est pas un critère très utile pour décider si une ou plusieurs variables doivent être ajoutées au modèle. Une variable explicative ne doit être incluse dans un modèle qu'à la condition que son effet *ceteris paribus* sur y ne soit pas nul dans la *population*. Nous montrerons comment tester cette hypothèse dans le chapitre 4 consacré à l'inférence statistique. Nous verrons aussi que le R^2, utilisé correctement, permet de *tester* si un groupe de variables joue un rôle important dans l'explication de y. Pour l'instant, nous l'utiliserons comme une mesure de la qualité d'ajustement d'un modèle donné.

EXEMPLE 3.4
Analyse des résultats obtenus à l'université

Reprenons la régression cherchant à expliquer la moyenne générale des notes obtenues à l'université (*colGPA*). Nous indiquons cette fois le R^2 de la régression.

$$\widehat{colGPA} = 1{,}29 + 0{,}453\ hsGPA + 0{,}0094\ ACT$$
$$n = 141,\ R^2 = 0{,}176.$$

Le R carré signifie que *hsGPA* et *ACT* expliquent ensemble environ 17,6 % de la variation de *colGPA* dans cet échantillon d'étudiants. Ce pourcentage peut sembler faible, mais il ne faut pas oublier qu'il n'y a que deux variables explicatives dans cette régression. Or, de nombreux autres facteurs contribuent à la performance d'un étudiant, comme l'environnement familial, la personnalité, la qualité de la formation reçue au lycée, l'intérêt porté aux études universitaires. Si *hsGPA* et *ACT* expliquaient la quasi-totalité de la variation de *colGPA*, cela signifierait que la réussite à l'université serait prédéterminée par la réussite au lycée !

EXEMPLE 3.5
Modélisation du nombre d'arrestations

Le fichier CRIME1 contient une base de données sur les arrestations réalisées au cours de l'année 1986, sur un échantillon de 2 725 hommes nés en Californie en 1960-1961. Chaque homme présent dans l'échantillon a été arrêté au moins une fois avant l'année 1986. La variable *narr86* est le nombre de fois où l'homme a été arrêté en 1986. Elle vaut zéro pour la plupart des hommes de l'échantillon (72,69 %), et elle varie entre 0 et 12. Le pourcentage d'hommes arrêtés une seule fois en 1986 est égal à 20,51 %. La variable *pcnv* est la proportion (et non le pourcentage) des arrestations antérieures (celles ayant eu lieu avant 1986) qui ont été suivies d'une condamnation. La variable *avgsen* est la durée moyenne des peines de prisons purgées lors des condamnations antérieures ; elle est égale à zéro pour la plupart des hommes. La variable *ptime86* indique les mois passés en prison en 1986 (de zéro à douze) et *qemp86* est le nombre de trimestres durant lesquels l'individu a travaillé (de zéro à quatre).

On cherche à expliquer le nombre d'arrestations, et pour cela, on considère le modèle de régression multiple suivant :

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 ptime86 + \beta_4 qemp86 + u$$

pcnv tente de capturer la probabilité *attendue* d'être reconnu coupable d'un délit ou d'un crime ; *avgsen* donne une idée de la sévérité *attendue* de la peine en cas de condamnation. La variable *ptime86* prend en compte l'impact mécanique de l'incarcération sur le crime : si un individu est en prison, il ne peut pas être arrêté pour un crime commis à l'extérieur de la prison. *Qemp86* est un indicateur élémentaire des opportunités qui existent sur le marché du travail.

Dans un premier temps, nous estimons le modèle sans la variable *avgsen*. Nous obtenons :

$$\widehat{narr86} = 0,712 - 0,150\, pcnv - 0,34\, ptime86 - 0,104\, qemp86$$
$$n = 2\,725,\ R^2 = 0,0413.$$

Cette équation nous indique que les trois variables, *pcnv*, *ptime86* et *qemp86*, expliquent ensemble 4,1 % environ de la variation de *narr86*.

Tous les coefficients de la pente ont le signe attendu. Une augmentation de la proportion des arrestations suivies d'une condamnation diminue le nombre prédit d'arrestations. Si nous augmentons *pcnv* de 0,5 (soit 50 points de pourcentage, ce qui est une très forte augmentation de la probabilité d'être reconnu coupable), alors $\widehat{\Delta narr86} = -0,150(0,50) = -0,075$, en maintenant les autres facteurs inchangés. Cela peut sembler surprenant à première vue, car le nombre d'arrestations est un nombre entier et ne peut donc pas varier d'une fraction. En réalité, on peut utiliser cette valeur pour obtenir la variation prédite du nombre d'arrestations attendues pour un groupe d'hommes suffisamment important. Par exemple, pour 100 hommes qui ont déjà fait l'objet d'une arrestation dans le passé (et pour lesquels *ptime86* et *qemp86* sont identiques), on prédit une diminution du nombre d'arrestations de 7,5 si *pcnv* augmente de 0,5.

Le coefficient de *ptime86* a également le signe attendu. Un temps plus long d'incarcération en 1986 implique un nombre plus faible d'arrestations. Si *ptime86* passe de 0 à 6, les arrestations prédites pour un homme en particulier chuteront de 0,034(6) = 0,204, *ceteris paribus*. La possibilité de travailler pendant un trimestre (supplémentaire) implique une diminution des arrestations de 0,104, soit 10,4 arrestations pour 100 hommes.
Si nous ajoutons *avgsen* au modèle, l'équation estimée est :

$$\widehat{narr86} = 0,707 - 0,151\, pcnv + 0,0074\, avgsen - 0,037\, ptime86 - 0,103\, qemp86$$
$$n = 2\,725,\ R^2 = 0,0422$$

Quand on ajoute la variable mesurant la durée moyenne des peines, le R^2 augmente de 0,0413 à 0,0422, ce qui est un effet plutôt faible. Le signe de la variable *avgsen* n'est pas non plus celui que nous attendions : plus la durée moyenne des peines est longue, plus le nombre d'arrestations est élevé.

L'exemple 3.5 nous invite à analyser les résultats avec précaution. Le fait que les quatre variables explicatives de la deuxième régression expliquent environ 4,2 % de la variation de *narr86* ne signifie pas nécessairement que la régression multiple est inutile. Même si ces variables ne parviennent pas à bien expliquer la variation des arrestations, il est tout à fait possible que ces estimations soient des estimations fiables de l'effet *ceteris paribus* de chacune des variables indépendantes sur *narr86*. Comme nous le verrons plus tard, cela ne dépend pas directement de la valeur du R^2. De manière générale, un R^2 faible implique qu'il est difficile de prédire avec précision les résultats de y pour des cas particuliers. Nous y reviendrons plus en détail au chapitre 6. Dans l'exemple des arrestations, la faible valeur du R^2 nous rappelle ainsi une question classique en sciences sociales : prédire le comportement d'un individu s'avère souvent difficile.

Régression passant par l'origine

Abordons à présent brièvement le cas où la théorie ou l'intuition économique indique que β_0 doit être égal à zéro. Nous cherchons alors à estimer une équation de la forme suivante :

$$\tilde{y} = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \ldots + \tilde{\beta}_k x_k \qquad \textbf{[3.30]}$$

Le symbole « ~ », placé au-dessus des coefficients estimés, est utilisé pour différencier ces estimations de celles que nous obtenons par les MCO lorsque la régression contient une ordonnée à l'origine [comme dans (3.11)]. Dans (3.30), quand $x_1 = 0$, $x_2 = 0$, ..., $x_k = 0$, la valeur prédite est zéro. Dans ce cas, la droite de régression passe par l'origine ; $\tilde{\beta}_1, \ldots, \tilde{\beta}_k$ sont les coefficients des MCO de y sur x_1, x_2, ..., x_k, *en l'absence d'ordonnée à l'origine dans la régression.*

Comme d'habitude, les estimations des MCO de (3.30) minimisent la somme des carrés des résidus, mais il y a une contrainte en plus : l'ordonnée à l'origine doit être égale à zéro. Notez bien que les propriétés des MCO que nous avons dérivées auparavant ne sont plus valables dans le cas d'une régression sans ordonnée à l'origine. En effet, la moyenne des résidus des MCO au sein de l'échantillon n'est plus égale à zéro. Par ailleurs, si on écrit le R^2 sous la forme 1 – SCR/SCT, avec SCT donné dans (3.24), alors le R^2 peut même être négatif, car la SCR est à présent égale à $\sum_{i=1}^{n} (y_i - \tilde{\beta}_1 x_{i1} - \ldots - \tilde{\beta}_k x_{ik})^2$. Cela signifie que la simple moyenne, $\bar{y}$, « explique » davantage la variation des y_i que l'ensemble des variables explicatives incluses dans la régression. Nous devons alors soit inclure une ordonnée à l'origine dans la régression, soit conclure que les variables explicatives ne parviennent pas à expliquer la variation totale de y. En réalité, il n'y a pas de règle établie pour calculer le R carré d'une régression sans ordonnée à l'origine. Pour éviter d'avoir un R carré négatif dans une régression sans ordonnée à l'origine, certains économistes préfèrent calculer le R^2 comme le carré du coefficient de corrélation entre les observées et les valeurs prédites de y, comme en (3.29). Notez bien que la valeur ajustée moyenne, $\bar{\hat{y}}$, n'est alors plus égale à $\bar{y}$.

Un des principaux défauts de la régression sans ordonnée à l'origine est que les estimateurs des MCO pour les paramètres de la pente seront biaisés si l'ordonnée à l'origine β_0 dans le modèle de population est différente de zéro. Dans certains cas, le biais peut être très important. Par contre, si le modèle inclut une ordonnée à l'origine alors que sa vraie valeur est nulle, les estimateurs des coefficients de la pente ne seront pas biaisés. Seule leur variance sera supérieure à celle que nous aurions obtenue sans ordonnée à l'origine.

3.3 L'ESPÉRANCE DES ESTIMATEURS DES MCO

Nous abordons à présent les propriétés statistiques des MCO qui s'avèrent cruciales pour estimer les paramètres d'un modèle pour une population sous-jacente. Dans cette section, nous calculons la valeur attendue des estimateurs des MCO. Nous définissons et analysons quatre hypothèses selon lesquelles les estimateurs des MCO sont des estimateurs sans biais des paramètres de population. Ces hypothèses sont des extensions

directes des hypothèses du modèle de régression simple. Nous caractérisons également le biais des MCO qui provient de l'omission d'une variable importante dans la régression.

Rappelez-vous que les propriétés statistiques ne sont pas déduites d'un échantillon spécifique ; elles sont liées aux propriétés des estimateurs que l'on observe lors d'un échantillonnage aléatoire répété. C'est la raison pour laquelle les sections 3.3, 3.4, et 3.5 sont quelque peu abstraites. Même s'il nous arrivera de parler d'estimateur biaisé pour des modèles estimés sur base d'un seul échantillon, il n'est pertinent de définir les propriétés statistiques d'un ensemble d'estimations que si elles proviennent d'un échantillonnage aléatoire répété.

La première hypothèse que nous posons définit simplement le modèle de régression multiple.

Hypothèse RLM.1
Linéarité dans les paramètres

Le modèle de la population peut être écrit comme

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \qquad [3.31]$$

où $\beta_0, \beta_1, \ldots, \beta_k$ sont les paramètres inconnus qui nous intéressent et u est le terme d'erreur aléatoire non observé, ou perturbation.

L'équation (3.31) représente le **modèle de la population**, aussi appelé le **vrai modèle**, dans le sens où il est possible que nous estimions un modèle différent de (3.31) sur base d'un échantillon. L'élément crucial à noter est que le modèle est linéaire par rapport aux paramètres $\beta_0, \beta_1, \ldots, \beta_k$. Le modèle (3.31) est assez flexible car il permet de modéliser les variables d'intérêt sous-jacentes en recourant à des formes fonctionnelles diverses et variées, comme la fonction logarithmique ou la fonction carré [voir par exemple l'équation (3.7)].

Hypothèse RLM.2
Échantillonnage aléatoire

Nous avons un échantillon aléatoire de n observations $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$ portant sur les variables du modèle de la population décrit ci-dessus.

Nous pouvons également écrire le modèle pour une observation particulière i provenant de l'échantillon aléatoire. Dans ce cas,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i \qquad [3.32]$$

Rappelez-vous que i se réfère à l'observation et que le second indice associé à x correspond au numéro de la variable. Par exemple, nous pouvons écrire l'équation du salaire des PDG pour un PDG en particulier de la manière suivante :

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ceoten_i + \beta_3 ceoten_i^2 + u_i \qquad [3.33]$$

Le terme u_i comprend les caractéristiques non observées du PDG i qui affectent son salaire. Dans les applications empiriques, il est souvent plus rapide d'écrire le modèle en reprenant le modèle de la population, comme celui de (3.31). Le modèle de la population est plus dépouillé et rappelle que nous sommes avant tout intéressés par la relation qui existe au sein de la population.

À la lumière du modèle (3.31), les estimateurs des MCO $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$, qui proviennent de la régression de y sur $x_1, \ldots, x_k$, sont à présent considérés comme les estimateurs de $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Dans la section 3.2, nous avons vu que les MCO déterminent les paramètres du modèle sur base d'un échantillon particulier de telle sorte que la moyenne des résidus soit égale à zéro et que la corrélation d'échantillon

entre chaque variable indépendante et les résidus vaille également zéro. Nous allons maintenant définir une condition sans laquelle il est impossible de bien définir les estimations des MCO pour un échantillon donné.

Hypothèse RLM.3
Pas de colinéarité parfaite

Dans l'échantillon (et par conséquent dans la population), aucune variable indépendante ne correspond à une constante et il n'existe pas de combinaison linéaire exacte entre les variables indépendantes.

L'hypothèse RLM.3 est plus compliquée que sa contrepartie pour la régression simple, car nous devons à présent considérer les relations entre toutes les variables indépendantes. Si une variable indépendante de (3.31) est une combinaison linéaire exacte des autres variables indépendantes, alors nous disons que le modèle souffre de **colinéarité parfaite** ; il ne peut pas être estimé par les MCO.

Il est important de noter que, sous l'hypothèse RLM.3, les variables indépendantes *peuvent* être corrélées, sans pouvoir l'être *parfaitement*. Si aucune corrélation entre les variables indépendantes n'était autorisée, alors l'intérêt de la régression multiple serait très limité. Par exemple, dans le modèle reliant les résultats scolaires (*avgscore*) aux dépenses d'éducation (*expend*) et au revenu familial moyen (*avginc*),

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u$$

il est logique que les variables *expend* et *avginc* soient corrélées : les districts scolaires où les revenus familiaux sont élevés ont tendance à dépenser davantage par étudiant. C'est d'ailleurs la raison principale qui nous a conduits à inclure *avginc* dans l'équation, en plus de la variable *expend*. Cela nous permet de mesurer l'effet *ceteris paribus* de la variable *avginc*. L'hypothèse RLM.3 exclut une corrélation *parfaite* entre *expend* et *avginc* au sein de l'échantillon. Nous serions particulièrement malchanceux s'il s'avérait que les dépenses par étudiant devaient être parfaitement corrélées avec le revenu familial moyen au sein de l'échantillon. Par contre, une corrélation non nulle, même élevée, sera possible, voire même attendue.

Deux variables sont parfaitement corrélées lorsqu'une variable est le multiple d'une autre, tout simplement. Cela peut arriver si un chercheur introduit par inadvertance dans sa régression deux fois la même variable exprimée dans des unités de mesure différentes. Par exemple, dans la relation entre la consommation et le revenu, cela n'a pas de sens d'inclure à la fois le revenu mesuré en dollars et le revenu mesuré en milliers de dollars. Une de ces deux variables indépendantes est redondante. Comment pourrait-on imaginer faire varier le revenu mesuré en dollars tout en maintenant constant le revenu mesuré en milliers de dollars ?

L'hypothèse RLM.3 n'interdit pas l'utilisation de fonctions non linéaires d'une même variable comme régresseurs. Par exemple, le modèle $cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$ respecte l'hypothèse RLM.3. Même si $x_2 = inc^2$ est une fonction exacte de $x_1 = inc$, inc^2 n'est pas une fonction *linéaire* exacte de *inc*. Inclure *inc*, à la fois en niveau et au carré, est une manière de modéliser une forme quadratique. Il ne s'agit pas de mesurer *inc* dans des *unités* de mesure différentes.

Dans certaines situations plus subtiles, l'inclusion d'une variable indépendante multiple d'une autre est désirable, bien qu'elle reste strictement impossible. Supposons que nous voulions estimer une extension de la fonction de consommation à élasticité constante. Pour introduire une élasticité variable en fonction du revenu, nous pourrions naïvement proposer un modèle de ce type :

$$\log(cons) = \beta_0 + \beta_1 \log(inc) + \beta_2 \log(inc^2) + u \qquad \textbf{[3.34]}$$

où $x_1 = \log(inc)$ et $x_2 = \log(inc^2)$. Néanmoins, en utilisant les propriétés de base du logarithme népérien (voir l'annexe A), nous obtenons $\log(inc^2) = 2\log(inc)$. Autrement dit, $x_2 = 2x_1$, ce qui viole l'hypothèse RLM.3. À côté de $\log(inc)$, nous devrions plutôt inclure $[\log(inc)]^2$, et non $\log(inc^2)$. Cette extension du modèle à

élasticité constante peut être estimée par les MCO. Le chapitre 6 sera, entre autres, consacré à l'interprétation de ce type de modèles.

Des cas de colinéarité parfaite se rencontrent également lorsqu'une variable indépendante correspond à une fonction linéaire exacte d'*au moins deux* autres variables indépendantes. Par exemple, supposons que nous voulions estimer les effets des dépenses de campagne électorale sur les résultats du scrutin. Pour simplifier, admettons que deux candidats se présentent à chaque élection. Soit *voteA*, le pourcentage de votes récoltés par le candidat A ; *expendA*, les dépenses électorales du candidat A ; *expendB*, les dépenses de campagnes du candidat B ; et *totexpend*, le total des dépenses électorales. Les trois dernières variables sont mesurées en dollars. Si nous désirons séparer les effets des dépenses effectuées par chaque candidat de ceux liés aux dépenses totales, il peut sembler approprié de tester le modèle suivant :

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 totexpend + u \qquad \textbf{[3.35]}$$

Pourtant, ce modèle viole l'hypothèse RLM.3 et ne pourra pas être estimé par les MCO, car $x_3 = x_1 + x_2$ par définition. Le paramètre β_1 dans l'équation (3.35) devrait mesurer l'effet *ceteris paribus* d'une augmentation, égale à un dollar, des dépenses du candidat A sur le pourcentage de votes récoltés par le candidat A. Cela exige que les dépenses du candidat B *et* que les dépenses totales fixées restent inchangées, ce qui est absurde. Si *expendB* et *totexpend* doivent rester constants, *expendA* ne peut pas augmenter.

Le problème de colinéarité parfaite dans l'équation (3.35) disparaît si nous retirons une des trois variables du modèle, par exemple *totexpend*. Dans ce cas, le coefficient d'*expendA* mesure l'effet d'une augmentation des dépenses électorales du candidat A sur le pourcentage de votes reçus par A, en maintenant les dépenses de B inchangées.

Comme le montrent les exemples précédents, le respect de l'hypothèse RLM.3 exige que nous portions une attention particulière à la spécification du modèle. Par ailleurs, l'hypothèse RLM.3 ne tient pas non plus si la taille de l'échantillon, n, est trop petite par rapport au nombre de paramètres estimés. Dans le modèle de régression général de l'équation (3.31) où il y a $k + 1$ paramètres, il est impossible de vérifier RLM.3 si $n < k + 1$. Il suffit de se fier à son intuition : pour estimer $k + 1$ paramètres, nous avons besoin de $k + 1$ observations *au minimum*. Il est d'autant plus facile de respecter RLM.3 que le nombre d'observations est élevé. Nous reviendrons sur l'importance de disposer d'un échantillon de grande taille lorsque nous calculerons la variance dans la section 3.4.

Pour aller plus loin 3.3

Dans l'exemple (3.35), vous désirez remplacer la variable *totexpend* par *shareA*, sachant que *shareA* = 100 (*expendA/totexpend*). La variable *shareA* mesure donc la proportion en pourcentage des dépenses électorales faites par le candidat A. L'hypothèse RLM.3 est-elle violée ?

Dans de rares cas, il arrive que l'hypothèse RLM.3 ne soit pas vérifiée même lorsque la spécification du modèle est judicieuse et que $n \geq k + 1$. Il faut néanmoins jouer de malchance. Par exemple, dans l'équation du salaire, il est possible de tomber sur un échantillon aléatoire dans lequel, par le plus grand des hasards, le nombre d'années d'études de *chaque* individu soit précisément égal au double des années d'expérience. Il est impossible de remplir les conditions de l'hypothèse RLM.3 dans un tel scénario. Fort heureusement, à moins que l'échantillon soit de très petite taille, ce scénario reste très improbable.

L'absence de biais de l'estimateur des MCO repose sur une dernière hypothèse, la plus importante, qui correspond à une extension directe de l'hypothèse RLS.4.

Hypothèse RLM.4
Espérance conditionnelle de l'erreur égale à zéro

La valeur attendue du terme d'erreur u est égale à zéro, quelles que soient les valeurs prises par les variables indépendantes. En d'autres termes,

$$E(u|x_1, x_2, \ldots, x_k) = 0. \qquad [3.36]$$

L'hypothèse RLM.4 est fausse si la relation fonctionnelle entre la variable expliquée et les variables explicatives est mal spécifiée dans l'équation (3.31). Tel sera le cas si, par exemple, nous oublions d'introduire le terme quadratique inc^2 avant d'estimer la fonction de consommation $cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$. La forme fonctionnelle sera également mal spécifiée si nous utilisons une variable en niveau alors que le modèle de la population requiert une variable en log, et *vice versa*. Par exemple, les estimateurs seront biaisés si nous utilisons *wage* comme variable dépendante dans notre analyse de régression alors que le modèle de la population doit avoir log(*wage*) comme variable dépendante. Bien que le choix d'une bonne spécification repose sur l'intuition, nous étudierons dans le chapitre 9 des procédures qui permettent de détecter des erreurs de spécification dans la forme fonctionnelle.

L'omission d'un facteur important, par ailleurs corrélé avec une des variables $x_1, x_2, \ldots, x_k$, invalide également RLM.4. Etant donné que la régression multiple permet d'inclure de nombreuses variables explicatives, ce risque d'omission est beaucoup plus grand dans le cadre de l'analyse par régression simple. Néanmoins, dans toute analyse empirique, il peut exister des facteurs auxquels nous ne pensons pas ou que nous ne pouvons tout simplement pas observer par manque de données. S'il s'agit de facteurs importants, qui sont corrélés avec une ou plusieurs variables indépendantes, alors l'hypothèse RLM.4 est invalidée. Nous allons bientôt être amenés à calculer le biais imputable à l'omission de tels facteurs.

L'omission de variables est une des raisons qui explique l'existence d'une corrélation entre une variable explicative et le terme d'erreur u ; ce n'est pas la seule. Dans les chapitres 9 et 15, nous discuterons du problème de l'erreur de mesure d'une variable explicative. Dans le chapitre 16, nous aborderons le problème, plus compliqué à conceptualiser, de la détermination simultanée de différentes variables. Nous y faisons face lorsque nous voulons modéliser les prix en fonction des quantités, et *vice versa* ; ces deux variables sont déterminées conjointement par l'intersection entre l'offre et la demande. Avant de pouvoir examiner ces questions, nous devons identifier l'ensemble des hypothèses dont dépend le bon fonctionnement de l'analyse par régression multiple.

Quand l'hypothèse RLM.4 est vérifiée, il est fréquent d'affirmer que les **variables explicatives** sont **exogènes**. Si x_j est corrélée avec u, quelle que soit la raison, elle est souvent considérée comme une **variable explicative endogène.** Les qualificatifs « endogène » et « exogène » proviennent des systèmes d'équations simultanées (voir chapitre 16). De nos jours, le terme « variable endogène » désigne toute situation dans laquelle une variable explicative est corrélée avec le terme d'erreur.

Avant de démontrer que l'estimateur des MCO est sans biais sous les hypothèses RLM.1 à RLM.4, il est important de ne pas confondre les hypothèses RLM.3 et RLM.4, comme le font souvent les étudiants qui découvrent l'économétrie. L'hypothèse RLM.3 exclut certaines relations entre les variables du modèle, mais elle n'a *rien à voir* avec le terme d'erreur u. Si la condition RLM.3 n'est pas remplie, le modèle ne pourra pas être estimé par la méthode des MCO. Quant à l'hypothèse RLM.4, la plus importante des deux, elle restreint la relation entre les variables explicatives et les facteurs non observés, présents dans u. Malheureusement, il est impossible de savoir avec certitude si l'espérance conditionnelle des facteurs non observés (donc, de l'erreur) est nulle ou si elle dépend effectivement des valeurs prises par les variables explicatives. C'est précisément la raison pour laquelle il s'agit d'une hypothèse cruciale.

Vérifions à présent que l'estimateur des MCO n'est pas biaisé lorsque les quatre premières hypothèses de la régression multiple sont vérifiées. Comme dans le cas de la régression simple, les espérances sont conditionnelles aux valeurs des variables explicatives dans l'échantillon. Nous le démontrons explicitement dans l'annexe 3A.

Théorème 3.1
Absence de biais des MCO

Sous les hypothèses RML.1 à RML.4,

$$E(\hat{\beta}_j) = \beta_j, \; j = 0, 1, \ldots, k \qquad [3.37]$$

quelle que soit la valeur du paramètre de la population, β_j. En d'autres termes, les estimateurs des MCO sont des estimateurs sans biais des paramètres de population.

Dans les exemples empiriques que nous avons vus, l'hypothèse RLM.3 était satisfaite puisque nous avons été capables de calculer les estimateurs des MCO. De plus, pour la majorité d'entre eux, les échantillons sont choisis de manière aléatoire au sein d'une population bien définie. Si nous pensons que les modèles sont correctement spécifiés de telle sorte que l'hypothèse RLM.4 soit valide, alors nous pouvons conclure que, dans ces exemples, les MCO offrent des estimateurs sans biais.

Maintenant que nous allons pouvoir utiliser la régression multiple dans le but de réaliser des études empiriques plus élaborées, il est important d'insister sur la signification de l'absence de biais. Dans des exemples similaires à celui de l'équation du salaire en (3.19), il est tentant d'écrire que « 9,2 % est une estimation sans biais du rendement de l'éducation ». Comme nous le savons, considérer qu'une estimation est sans biais n'a pas de sens : une estimation est un nombre donné, obtenu à partir d'un échantillon particulier. Cette estimation est rarement égale au paramètre de la population, et nous ne pouvons d'ailleurs pas le savoir avec certitude. Quand nous disons que les MCO sont sans biais sous les hypothèses RLM.1 à RLM.4, nous voulons signifier que la *procédure*, par laquelle sont obtenues les estimations des MCO, est sans biais, en imaginant que cette procédure soit appliquée à tous les échantillons aléatoires possibles. Notre espoir est d'avoir obtenu un échantillon qui nous donne une estimation proche de la valeur de la population ; malheureusement, nous n'en avons aucune garantie. Par contre, sous les hypothèses RLM.1 à RLM.4, nous avons la certitude que l'estimation n'a aucune chance d'être *a priori* supérieure ou inférieure à sa vraie valeur. Tel serait le cas si l'estimateur des MCO était biaisé.

Inclusion de variables non pertinentes dans une régression

Le problème lié à **l'inclusion d'une variable non pertinente** dans une régression est équivalent à un problème de **surspécification du modèle**. Y remédier n'est pas compliqué. Une variable non pertinente (appelée également variable superflue ou variable redondante) est une variable indépendante, présente dans le modèle, dont l'effet *ceteris paribus* sur y est nul dans la population. Autrement dit, la vraie valeur du coefficient de cette variable est égale à zéro.

Pour illustrer ce problème, spécifions le modèle suivant, qui respecte les hypothèses RLM.1 à RLM.4 :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u. \qquad [3.38]$$

Si la variable x_3 n'a aucun effet sur y, après avoir pris en compte x_1 et x_2, alors elle ne « sert à rien » et $\beta_3 = 0$. Notez que le caractère superflu de x_3 n'est pas déterminé par son degré de corrélation avec les deux autres variables ; x_3 est une variable superflue si son effet sur y, purgé de l'influence de x_1 et x_2, est nul dans la population. En termes d'espérance conditionnelle, $E(y|x_1, x_2, x_3) = E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Dans la réalité, nous ne pouvons pas observer β_3 et savoir s'il est effectivement égal à zéro. Nous sommes contraints d'estimer l'équation en incluant x_3 :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 \qquad \textbf{[3.39]}$$

Nous avons inclus la variable non pertinente dans notre régression. Quelles sont les conséquences liées à l'inclusion de x_3 dans (3.39), sachant que son coefficient dans le modèle de la population (3.38) est égal à zéro ? Cela n'a *aucun effet* sur l'absence de biais de x_1 et x_2. Cette conclusion ne requiert pas de calcul particulier ; elle découle directement du théorème 3.1. Souvenez-vous : l'absence de biais signifie que $E(\hat{\beta}_j) = \beta_j$ pour *toute* valeur de β_j, donc y compris $\beta_j = 0$. Il en ressort que $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_2) = \beta_2$, et $E(\hat{\beta}_3) = 0$ (pour toute valeur de β_0, β_1, et β_2). Même si $\hat{\beta}_3$ ne sera jamais exactement égal à zéro, sa valeur moyenne dans l'ensemble des échantillons aléatoires le sera.

La conclusion que nous venons d'établir à partir de cet exemple a une portée générale : la surspécification d'un modèle, c'est-à-dire l'inclusion d'une ou de plusieurs variables non pertinentes dans une régression multiple, n'affecte pas les propriétés d'absence de biais des estimateurs des MCO. Faut-il en conclure que l'inclusion d'une variable redondante n'a aucun effet indésirable ? Absolument pas. Comme nous le verrons dans la section 3.4, inclure une variable superflue peut avoir des conséquences fâcheuses sur la *variance* des estimateurs des MCO.

Biais de variable omise : un cas simple

Au lieu d'inclure une variable non pertinente, supposons que nous omettions une variable qui appartient effectivement au véritable modèle, c'est-à-dire au modèle de la population. Il s'agit du problème lié à l'**exclusion d'une variable pertinente**, équivalent à un problème de **sous-spécification du modèle**. À plusieurs reprises déjà, nous avons indiqué que ce problème introduit un biais dans les estimateurs des MCO. Il est à présent temps de calculer le sens et l'ampleur attendus de ce biais.

La présence d'un biais causé par l'omission d'une variable omise provient d'**une erreur de spécification**. Nous allons commencer par définir le vrai modèle de population, qui respecte les hypothèses RLM.1 à RLM.4 par définition. Supposons qu'il contienne deux variables explicatives et un terme d'erreur :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u. \qquad \textbf{[3.40]}$$

Supposons que notre intérêt premier porte sur β_1, soit l'effet *ceteris paribus* de x_1 sur y. Soit y, le salaire horaire (*wage*) ; x_1, le niveau d'instruction (*educ*) ; et x_2, une mesure des aptitudes innées (*abil*). Afin d'obtenir un estimateur sans biais de β_1, nous *aimerions* idéalement effectuer une régression de y sur les deux variables, x_1 *et* x_2 (ce qui nous permettrait d'obtenir des estimateurs sans biais de β_0, β_1, et β_2). Cependant, en raison de notre ignorance ou du manque de données (concernant x_2), nous sommes contraints d'estimer le modèle *sans* x_2. En d'autres termes, nous réalisons une régression simple de y sur x_1 seulement. Nous obtenons l'équation :

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \qquad \textbf{[3.41]}$$

Nous utilisons le symbole "~" plutôt que "^" pour souligner le fait que $\tilde{\beta}_1$ vient d'un modèle sous-spécifié. Il est important d'opérer une distinction claire entre le vrai modèle sous-jacent, (3.40) dans notre cas, et le modèle que nous estimons effectivement par la régression décrite en (3.41). Il peut sembler illogique d'omettre la variable x_2 si elle appartient au vrai modèle ; rappelez-vous néanmoins que nous n'observons pas le vrai modèle (3.40) et, même lorsque nous parvenons à le deviner, nous n'avons parfois pas le choix, les données n'étant pas disponibles par exemple. Supposons que *wage* soit déterminé par

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u. \qquad \textbf{[3.42]}$$

Comme les aptitudes innées (*abil*) ne sont pas (directement) observables, nous devons souvent nous rabattre sur le modèle

$$wage = \beta_0 + \beta_1 educ + v$$

où $v = \beta_2 abil + u$. Nous notons $\tilde{\beta}_1$ l'estimateur de β_1 obtenu par la régression simple de *wage* sur *educ*.

Pour identifier le biais de variable omise dont souffre $\tilde{\beta}_1$, nous allons dériver la valeur espérée de $\tilde{\beta}_1$ conditionnellement aux valeurs d'échantillon de x_1 et x_2. Cette tâche n'est pas difficile car $\tilde{\beta}_1$ n'est autre que l'estimateur de la pente des MCO d'une régression simple ; nous avons déjà abondamment étudié cet estimateur dans le chapitre 2. La seule différence est que nous en analysons les propriétés lorsque le modèle de régression simple a été mal spécifié, en raison de l'omission d'une variable.

Le travail de dérivation du biais de l'estimateur $\tilde{\beta}_1$ de la régression simple est presque terminé. En utilisant l'équation (3.23), nous avons la relation algébrique $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$. Les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ sont les estimateurs de la pente que nous aurions pu obtenir en estimant le modèle de régression multiple (3.40), soit

$$y_i \text{ sur } x_{i1}, x_{i2}, i = 1, \ldots, n. \qquad \textbf{[3.43]}$$

L'estimateur $\tilde{\delta}_1$ est la pente obtenue par la régression simple de

$$x_{i2} \text{ sur } x_{i1}, i = 1, \ldots, n. \qquad \textbf{[3.44]}$$

Dans le calcul de $E(\tilde{\beta}_1)$, nous considérons $\tilde{\delta}_1$ comme donné (non aléatoire) puisqu'il ne dépend que des variables indépendantes dans l'échantillon. Par ailleurs, comme le modèle (3.40) respecte les hypothèses RLM.1 à RLM.4, nous savons que $\hat{\beta}_1$ et $\hat{\beta}_2$ sont des estimateurs sans biais de β_1 et β_2. Par conséquent,

$$\begin{aligned} E(\tilde{\beta}_1) &= E(\hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1) = E(\hat{\beta}_1) + E(\hat{\beta}_2)\tilde{\delta}_1 \\ &= \beta_1 + \beta_2\tilde{\delta}_1 \end{aligned} \qquad \textbf{[3.45]}$$

Par conséquent,

$$\text{Biais}(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1 = \beta_2\tilde{\delta}_1 \qquad \textbf{[3.46]}$$

Comme le biais provient de l'omission de la variable explicative x_2, le membre de droite de l'équation (3.46) est souvent appelé le **biais de variable omise**, ou parfois le **biais d'omission**.

L'équation (3.46) indique qu'il existe deux cas où $\tilde{\beta}_1$ est sans biais. Le premier est plutôt évident : si x_2 n'apparaît pas dans le vrai modèle (3.40), soit $\beta_2 = 0$, alors $\tilde{\beta}_1$ est sans biais. Rien de nouveau : nous en étions parfaitement conscients depuis l'analyse de régression simple du chapitre 2. Le deuxième cas est plus intéressant : si $\tilde{\delta}_1 = 0$, alors $\tilde{\beta}_1$ est un estimateur sans biais de β_1, même si $\beta_2 \neq 0$.

Comme $\tilde{\delta}_1$ mesure la covariance d'échantillonnage entre x_1 et x_2 divisée par la variance d'échantillon de x_1, $\tilde{\delta}_1 = 0$ si et seulement si x_1 et x_2 ne sont pas corrélées dans l'échantillon. Nous aboutissons à une conclusion importante : si x_1 et x_2 ne sont pas corrélées dans l'échantillon, alors $\tilde{\beta}_1$ est sans biais. Cela n'est pas surprenant. Dans la section 3.2, nous avons montré que l'estimateur $\tilde{\beta}_1$ de la régression simple est égal à l'estimateur $\hat{\beta}_1$ de la régression multiple lorsque x_1 et x_2 ne sont pas corrélées dans l'échantillon. [Nous pouvons également montrer que, si $E(x_2|x_1) = E(x_2)$, $\tilde{\beta}_1$ est sans biais ; même si on ne tient pas compte des x_{i2}, l'estimation de β_1 est sans biais. Autrement dit, le terme d'erreur peut inclure x_2 sans violer l'hypothèse de moyenne conditionnelle nulle des erreurs, puisque l'ordonnée à l'origine peut s'ajuster].

Quand x_1 et x_2 sont corrélées, $\tilde{\delta}_1$ a le même signe que la corrélation entre x_1 et x_2 : $\tilde{\delta}_1 > 0$ lorsque x_1 et x_2 sont corrélées positivement ; $\tilde{\delta}_1 < 0$ si x_1 et x_2 sont corrélées négativement. Le signe du biais de $\tilde{\beta}_1$ dépend des signes de β_2 et $\tilde{\delta}_1$. Les quatre combinaisons possibles sont indiquées dans le tableau 3.2. Le tableau 3.2 mérite toute votre attention. Par exemple, le biais de $\tilde{\beta}_1$ sera positif si $\beta_2 > 0$ (x_2 a un effet positif sur y) et si x_1 et x_2 sont corrélées positivement ; le biais sera négatif si $\beta_2 > 0$ et si x_1 et x_2 sont corrélées négativement, etc.

Le tableau 3.2 donne des indications quant à la direction du biais, mais que pouvons-nous dire de son ampleur ? La taille du biais est déterminée par les valeurs de β_2 et $\tilde{\delta}_1$. Qu'il soit positif ou négatif, un biais de faible ampleur ne doit pas être source d'inquiétude. Par exemple, si le rendement de l'éducation dans la population est 8,6 % alors que le biais de l'estimateur des MCO est égal à 0,1 %, l'existence du biais n'a pas d'implication réelle sur le plan pratique.

Comme β_2 est un paramètre inconnu de la population, nous n'avons pas de certitude quant à la vraie valeur de l'effet marginal de x_2 sur y. Nous pouvons néanmoins nous faire une idée relativement bonne de son signe. Quant à la corrélation entre x_1 et x_2 ($\tilde{\delta}_1$), il nous est impossible de la mesurer si x_2 n'est pas observé ; pourtant, dans la plupart des cas, nous pouvons en déterminer le signe *a priori*.

Retournons à l'équation du salaire (3.41). De plus grandes aptitudes innées conduisent nécessairement à une productivité plus forte et des salaires plus élevés, si bien que $\beta_2 > 0$. Par ailleurs, nous avons des raisons de croire qu'*educ* et *abil* sont corrélés positivement : les individus dont les capacités innées sont plus grandes visent, en moyenne, des niveaux d'études plus élevés. Par conséquent, les estimations de la régression simple [$wage = \beta_0 + \beta_1 educ + v$] par les MCO seront *en moyenne* trop élevés. Cela ne signifie pas que les estimations obtenues *à partir de notre échantillon* soient nécessairement trop grandes. Nous pouvons seulement conclure que si nous constituons un grand nombre d'échantillons aléatoires et que nous calculons les estimations de la régression simple pour chacun d'entre deux, alors la moyenne de ces estimations sera supérieure à β_2.

Tableau 3.2 *Résumé du biais de $\tilde{\beta}_1$ quand x^2 n'est pas inclus dans l'équation estimée (3.40)*

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Biais positif	Biais négatif
$\beta_2 < 0$	Biais négatif	Biais positif

EXEMPLE 3.6
Équation de salaire horaire

Supposons que le modèle $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + u$ respecte les hypothèses RLM.1 à RLM.4. Le jeu de données WAGE1 ne contient pas d'information sur les capacités innées. Nous estimons donc β_1 à partir de la régression simple

$$\widehat{\log(wage)} = 0{,}584 + 0{,}083 educ \qquad [3.47]$$
$$n = 526,\ R^2 = 0{,}186.$$

Ce résultat provient d'un seul échantillon ; nous ne pouvons donc pas affirmer que β_1 est égal à 0,083. Le vrai rendement de l'éducation pourrait être inférieur ou supérieur à 8,3 % ; d'ailleurs, nous ne pourrons jamais le savoir avec certitude. Par contre, en raison du biais de variable omise, nous pouvons conclure que la moyenne des estimations de β_1 obtenues à partir de tous les échantillons aléatoires possibles sera trop élevée.

Imaginons un second exemple dans lequel la moyenne obtenue à un examen national par les élèves d'un lycée est déterminée par le niveau des dépenses publiques par élève (*expend*) et le taux de pauvreté des familles (*povrate*) dans le district, soit :

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 povrate + u. \qquad [3.48]$$

Les données disponibles par district scolaire incluent le pourcentage d'élèves qui ont obtenu une note supérieure à la moyenne et les dépenses publiques par élève. Par contre, nous n'avons pas d'informations sur

le taux de pauvreté. Nous sommes donc contraints d'estimer β_1 à partir d'une régression simple d'*avgscore* sur *expend*.

Nous pouvons néanmoins caractériser le biais probable de $\tilde{\beta}_1$. Notons tout d'abord que β_2 est vraisemblablement négatif : en effet, de nombreuses études montrent que les enfants issus de familles pauvres ont des résultats scolaires en moyenne moins bons que les autres enfants. Ensuite, les dépenses publiques par élève sont probablement corrélées négativement avec le taux de pauvreté : plus le taux de pauvreté est élevé, plus il est difficile de financer la dépense publique (l'impôt local ne rapportant pas grand-chose) ; donc, $\text{Corr}(x_1, x_2) < 0$. En conclusion, sur base du tableau 3.2, $\tilde{\beta}_1$ a un biais positif. Cette observation a une implication importante. Il est probable que l'estimation par régression simple de β_1 nous donne une valeur positive en raison du biais de variable omise, alors que le véritable effet *ceteris paribus* des dépenses publiques est nul, à savoir $\beta_1 = 0$. Le biais de variable omise pourrait nous faire croire que les dépenses publiques sont importantes alors qu'elles ne le sont pas en réalité.

Dans les travaux empiriques en économie, il est important de maîtriser la terminologie associée aux estimateurs biaisés. Lorsque la variable x_2 est omise dans le modèle (3.40) et que $\text{E}(\tilde{\beta}_1) > \beta_1$, alors nous disons que $\tilde{\beta}_1$ souffre d'**un biais vers le haut**. Si $\text{E}(\tilde{\beta}_1) < \beta_1$, alors $\tilde{\beta}_1$ est affecté d'un **biais vers le bas.** Que β_1 soit positif ou négatif, les définitions restent les mêmes. Dans le cas où $\text{E}(\tilde{\beta}_1)$ est plus proche de zéro que ne l'est β_1, $\tilde{\beta}_1$ affiche également un **biais vers zéro.** Par exemple, si $\beta_1 > 0$ et que $\tilde{\beta}_1$ est biaisé vers le bas, on dira que $\tilde{\beta}_1$ est également biaisé vers zéro. De même, si $\beta_1 < 0$ et que $\tilde{\beta}_1$ est biaisé vers le haut, alors $\tilde{\beta}_1$ sera également biaisé vers zéro.

Biais de variable omise : le cas général

Il est plus difficile d'identifier le signe d'un biais de variable omise lorsqu'il y a de multiples régresseurs dans le modèle estimé. Rappelez-vous qu'il suffit généralement que la corrélation entre une seule variable explicative et l'erreur soit non nulle pour que *tous* les estimateurs des MCO soient biaisés.

Par exemple, supposons que le modèle de population suivant respecte les hypothèses RLM.1 à RLM.4 :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \qquad \textbf{[3.49]}$$

Imaginons que nous estimions à la place le même modèle en omettant x_3 :

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 \qquad \textbf{[3.50]}$$

À présent, supposons que les variables x_2 et x_3 ne soient pas corrélées mais que x_1 soit corrélée avec x_3. En d'autres termes, x_1 est corrélée avec la variable omise, tandis que x_2 ne l'est pas. Il est tentant de penser que, même si $\tilde{\beta}_1$ est probablement biaisé (voir les dérivations effectuées dans la section précédente), $\tilde{\beta}_2$ ne le sera pas puisque x_2 et x_3 ne sont pas corrélées. Malheureusement, ce n'est vrai que si x_1 et x_2 ne sont pas corrélées. En général, $\tilde{\beta}_1$ et $\tilde{\beta}_2$ seront donc tous les deux biaisés.

Même dans le modèle plutôt simple décrit ci-dessous, il peut se révéler difficile d'obtenir la direction du biais de $\tilde{\beta}_1$ et $\tilde{\beta}_2$ en raison de la corrélation qui existe entre x_1, x_2, et x_3. Nous pouvons néanmoins recourir à une approximation pour nous faciliter la tâche. Si nous supposons que x_1 et x_2 ne sont pas corrélées, alors il nous est possible d'étudier le biais de $\tilde{\beta}_1$ comme si x_2 était absent à la fois du modèle de la population et du modèle estimé. En fait, quand x_1 et x_2 ne sont pas corrélées, nous pouvons montrer que :

$$E(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i3} - \bar{x}_3)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} = \beta_1 + \beta_3 \tilde{\delta}_1$$

Cette équation est similaire à l'équation (3.45), si ce n'est que β_3 remplace β_2 et que $\tilde{\delta}_1$ mesure la pente de la régression de x_3 sur x_1 dans (3.44), et non de x_2 sur x_1. Par conséquent, le biais de $\tilde{\beta}_1$ est obtenu en remplaçant β_2 par β_3 et x_2 par x_3 dans le tableau 3.2. Si $\beta_3 > 0$ et Corr $(x_1, x_3) > 0$, le biais de $\tilde{\beta}_1$ est positif, etc.

Reprenons le modèle du salaire en y ajoutant la variable *exper*, soit

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u.$$

Si nous ne tenons pas compte de l'aptitude innée des travailleurs (*abil*) en omettant cette variable du modèle, les estimateurs de β_1 et β_2 sont biaisés, même si nous supposons que la variable *exper* est non corrélée avec la variable *abil*. (Pour que β_2 ne soit pas biaisé, il faudrait également que la corrélation entre *exper* et *educ* soit nulle ; en réalité, on observe souvent une corrélation négative entre *educ* et *exper*.) Comme l'intérêt de cette régression porte sur la mesure du rendement de l'éducation, il est utile de déterminer si $\tilde{\beta}_1$ souffre d'un biais vers le haut ou vers le bas, en raison de l'omission de la variable *abil*. Il est souvent impossible de le déterminer sans poser une ou plusieurs hypothèses supplémentaires. Comme *approximation*, supposons que la variable *exper* ne soit corrélée ni avec *abil*, ni avec *educ*. (Autrement dit, faisons comme si *exper* était une variable superflue, ne jouant aucun rôle dans la régression.) Dans ce cas, vu que $\beta_3 > 0$ et que les variables *educ* et *abil* sont corrélées positivement, $\tilde{\beta}_1$ devrait avoir un biais vers le haut, ce qui correspond à la conclusion à laquelle nous avions abouti lorsque la variable *exper* n'était pas dans le modèle.

Ce type de raisonnement est souvent tenu lorsqu'il s'agit d'anticiper le biais de variable omise attendu dans des modèles plus complexes. En règle générale, nous nous concentrons sur une variable explicative bien spécifique, disons x_1, et sur un facteur omis en particulier, disons x_5. Gardez bien à l'esprit que cette pratique, qui consiste à ignorer toutes les autres variables explicatives (de x_2 à x_4, par exemple), n'est valide que si aucune d'entre elles n'est corrélée avec x_1. Ce ne sera jamais le cas dans la réalité mais ce genre d'analyse reste tout de même un point de départ utile. L'annexe 3A présente une analyse plus rigoureuse du biais provenant de l'omission d'une variable pertinente dans une régression multiple.

3.4 LA VARIANCE DES ESTIMATEURS DES MCO

Nous allons à présent calculer la variance des estimateurs des MCO. Après avoir étudié le mode de la distribution d'échantillonnage des $\hat{\beta}_j$ à l'aide de la moyenne, nous cherchons maintenant à obtenir une mesure de la dispersion de cette distribution à l'aide de la variance. Pour ce faire, nous devons ajouter l'hypothèse d'homoscédasticité, comme nous l'avons fait dans le chapitre 2. Deux raisons motivent le recours à cette hypothèse. En premier lieu, l'hypothèse de variance constante des erreurs permet de simplifier le calcul des formules. En second lieu, sous l'hypothèse d'homoscédasticité, les estimateurs des MCO jouissent de la propriété importante d'efficacité, comme nous le verrons dans la section 3.5.

Dans le cadre de la régression multiple, l'homoscédasticité est définie de la manière suivante.

Hypothèse RLM.5
Homoscédasticité

La variance de l'erreur u est la même quelle que soit la valeur des variables explicatives. En d'autres termes, $\text{Var}(u|x_1, \ldots, x_k) = \sigma^2$

L'hypothèse RLM.5 signifie que, conditionnellement aux variables explicatives, la variance du terme d'erreur, u, est *la même* pour toutes les combinaisons de résultats des variables explicatives. Si cette hypothèse n'est pas respectée, alors le modèle souffre d'hétéroscédasticité. Dans l'équation suivante :

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

l'homoscédasticité nécessite que la variance du terme d'erreur non observé u ne dépende pas du niveau des études, de l'expérience, ou du type de contrat. En d'autres termes,

$$\text{Var}(u|educ, exper, tenure) = \sigma^2$$

Si cette variance change lorsqu'une de ces trois variables explicatives change, alors il y a de l'hétéroscédasticité.

L'ensemble des hypothèses RLM.1 à RLM.5 représente les **hypothèses de Gauss-Markov**. Jusqu'ici, les hypothèses que nous avons posées ne s'appliquent qu'à l'analyse en coupe transversale avec un échantillonnage aléatoire. Nous verrons que les hypothèses de Gauss-Markov pour les séries temporelles ou les données de panel sont plus compliquées à formuler, bien qu'il existe de nombreuses similarités.

Dans la discussion qui suit, nous utilisons le symbole **x** pour représenter l'ensemble des variables indépendantes, $(x_1, \ldots, x_k)$. Par exemple, dans la régression du salaire sur *educ*, *exper*, et *tenure*, **x** = (*educ*, *exper*, *tenure*). Nous pouvons donc écrire les hypothèses RLM.1 et RLM.4 de la manière suivante :

$$E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Quant à l'hypothèse RLM.5, elle devient $\text{Var}(y|\mathbf{x}) = \sigma^2$. En écrivant les hypothèses de cette manière, nous pouvons clairement remarquer que l'hypothèse RLM.5 diffère de l'hypothèse RLM.4. D'un côté, sous l'hypothèse RLM.4, la valeur espérée de y, étant donné **x**, est linéaire par rapport aux paramètres et dépend très certainement de $x_1, x_2, \ldots, x_k$. De l'autre, sous l'hypothèse RLM.5, la variance de y, étant donné **x**, *ne* dépend *pas* des valeurs prises par les variables indépendantes.

À présent, nous pouvons obtenir les variances des $\hat{\beta}_j$ en conditionnant nos calculs aux valeurs prises par les variables indépendantes au sein de l'échantillon. La démonstration figure dans l'annexe de ce chapitre.

Théorème 3.2
Variance d'échantillonnage des estimateurs de la pente des MCO

Sous les hypothèses RLM.1 à RLM.5, étant donné les valeurs prises par les variables indépendantes dans l'échantillon,

$$\text{Var}(\hat{\beta}_j) = \sigma^2/[\text{SCT}_j(1 - R_j^2)] \qquad \textbf{[3.51]}$$

pour $j = 1, 2, \ldots, k$. $\text{SCT}_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ est la variation totale de x_j dans l'échantillon. R_j^2 est le R carré de la régression de x_j sur toutes les autres variables indépendantes (en incluant une ordonnée à l'origine).

Le lecteur attentif peut se demander s'il existe une formule simple de la variance des $\hat{\beta}_j$ qui n'exige pas un calcul conditionnel. Malheureusement, il n'en existe aucune qui soit réellement utile. En effet, la formule (3.51) est une fonction particulièrement non linéaire des x_{ij}, ce qui rend virtuellement impossible le calcul d'une moyenne à partir de la distribution de population des variables explicatives. Heureusement, l'équation (3.51) répond à toutes nos questions d'ordre pratique. Même lorsque nous aborderons les propriétés asymptotiques des MCO dans le chapitre 5, l'équation (3.51) nous permettra d'obtenir la valeur dont nous avons besoin en grands échantillons, à condition que les hypothèses RLM.1 à RLM.5 tiennent.

Avant d'étudier l'équation (3.51) plus en détail, il est important de noter que toutes les hypothèses de Gauss-Markov sont requises pour parvenir à cette formule. L'hypothèse d'homoscédasticité dépend donc de l'hypothèse d'absence de biais. L'inverse n'est pas vrai : l'hypothèse d'homoscédasticité n'est pas requise lorsqu'il s'agit d'obtenir des estimateurs des MCO sans biais.

D'un point de vue pratique, la valeur que prend $\text{Var}(\hat{\beta}_j)$ est très importante. Une variance plus élevée signifie que l'estimateur est moins précis ; cela se traduit par des intervalles de confiance plus larges et des

tests d'hypothèses moins exacts (comme nous le verrons dans le chapitre 4). Dans la section suivante, nous discutons des différents éléments qui composent l'équation (3.51).

Les composants de la variance des MCO et la multicolinéarité

Dans l'équation (3.51), la variance des $\hat{\beta}_j$ dépend de trois facteurs : σ^2, SCT_j, et R_j^2. Souvenez-vous que l'indice j se réfère simplement à l'une des variables indépendantes (comme le niveau d'études ou le taux de pauvreté). Commençons par analyser la relation entre la variance de l'erreur et $Var(\hat{\beta}_j)$.

La variance de l'erreur, σ^2

L'équation (3.51) indique que la variance de l'estimateur des MCO sera plus grande au fur et à mesure que σ^2 s'élève. Cela ne doit pas nous surprendre : davantage de « bruit » dans un modèle (soit un σ^2 plus élevé) rend plus difficile l'estimation de l'effet *ceteris paribus* d'une variable indépendante, quelle qu'elle soit. Cela se traduit par des variances plus élevées pour les estimateurs de la pente des MCO. Comme σ^2 est propre à la population, il ne dépend pas de la taille de l'échantillon. Il s'agit d'ailleurs du seul élément de (3.51) qui est inconnu. Nous verrons plus tard comment obtenir un estimateur sans biais de σ^2.

Pour une variable dépendante donnée y, la seule manière de réduire la variance des erreurs consiste à ajouter des variables explicatives dans l'équation (ce qui revient à retirer des facteurs explicatifs du terme d'erreur). Dans certains cas, il n'est malheureusement pas possible de trouver d'autres facteurs clés qui aient un effet sur y et qui soient observables.

La variation totale des x_j dans l'échantillon, SCT_j

Selon l'équation (3.51), plus la variation totale des x_j est élevée, plus $Var(\hat{\beta}_j)$ est faible. Par conséquent, toutes choses étant égales par ailleurs, il est préférable que x_j affiche la variation la plus grande possible dans l'échantillon. Nous l'avions déjà constaté dans le cas de la régression simple, au chapitre 2. Bien qu'il soit rarement possible d'augmenter cette variation en *choisissant* des valeurs éloignées les unes des autres pour x_j, il est possible d'y arriver en augmentant la taille de l'échantillon. En réalité, lorsqu'un échantillon aléatoire est prélevé à partir d'une population, SCT_j augmente sans limite en fonction de la taille de l'échantillon. Contrairement à σ^2, SCT_j est un élément de la variance qui dépend systématiquement de la taille de l'échantillon.

Lorsque SCT_j est faible, la valeur de $Var(\hat{\beta}_j)$ peut être très élevée. Malgré tout, ce type de scénario ne viole pas l'hypothèse RLM.3. Il est vrai que $Var(\hat{\beta}_j)$ tendra vers l'infini si SCT_j tend vers zéro. Néanmoins, sur le plan technique, seul le cas extrême d'absence de variation des x_j, soit $SCT_j = 0$, est exclu par l'hypothèse RLM.3.

La force de la relation linéaire *entre* les variables indépendantes, R_j^2

Des trois éléments constitutifs de la variance décrite en (3.51), le terme R_j^2 est le plus compliqué à comprendre. Ce terme n'apparaît pas dans l'analyse par régression simple puisqu'il n'y a qu'une seule variable explicative dans une régression simple. Notez bien que ce R_j^2 est différent du R carré de la régression de y sur $x_1, x_2, \ldots, x_k$. Le R_j^2 s'obtient à partir d'une régression n'impliquant que les variables indépendantes du modèle original dans lequel x_j joue le rôle de variable dépendante.

Considérons d'abord le cas où $k = 2$, soit $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. Dans ce cas, $Var(\hat{\beta}_1) = \sigma^2/[SCT_1(1 - R_1^2)]$, où R_1^2 est le R carré de la régression simple de x_1 sur x_2 (en ajoutant une ordonnée à l'origine, comme

d'habitude). Comme le R carré mesure la qualité d'ajustement de la régression, une valeur de R_1^2 proche de 1 indique que x_2 explique la majorité de la variation de x_1 dans l'échantillon. Cela signifie que x_1 et x_2 sont fortement corrélées.

$\text{Var}(\hat{\beta}_j)$ s'élève au fur et à mesure que R_1^2 s'approche de 1. Par conséquent, une relation linéaire solide entre x_1 et x_2 peut conduire à des variances élevées pour les estimateurs de la pente des MCO. Le raisonnement est similaire pour $\text{Var}(\hat{\beta}_2)$. La figure 3.1 indique clairement la relation qui existe entre $\text{Var}(\hat{\beta}_1)$ et le R carré de la régression de x_1 sur x_2.

Considérons maintenant le cas général. R_j^2 mesure la proportion de la variation totale des x_j qui peut être expliquée par les *autres* variables indépendantes incluses dans l'équation. Pour une SCT_j et un σ^2 donnés, $\text{Var}(\hat{\beta}_j)$ est minimisée lorsque $R_j^2 = 0$, ce qui n'a lieu que si, et seulement si, la corrélation entre x_j et *chacune des autres* variables indépendantes est nulle. C'est là le scénario qui nous permet d'estimer β_j avec la plus grande précision.

L'autre scénario extrême, $R_j^2 = 1$, est exclu par l'hypothèse RLM.3. Si $R_j^2 = 1$, x_j est une combinaison linéaire *parfaite* d'une ou plusieurs autres variables indépendantes dans l'échantillon. Considérons le cas plus pertinent dans lequel R_j^2 « s'approche » de la valeur 1. Si nous examinons l'équation (3.51) et la figure 3.1, nous constatons que $\text{Var}(\hat{\beta}_j)$ sera très élevée. En fait, $\text{Var}(\hat{\beta}_j) \to \infty$ lorsque $R_j^2 \to 1$. Bien qu'imparfaite, cette forme de corrélation élevée entre deux variables indépendantes est appelée **multicolinéarité.**

Avant d'affiner notre compréhension de la multicolinéarité, il est important de souligner à nouveau qu'une situation où R_j^2 est proche de 1 *n'est pas* une violation de l'hypothèse RLM.3.

Comme la multicolinéarité ne va à l'encontre d'aucune des hypothèses de Gauss-Markov, il est important d'expliquer les circonstances dans lesquelles la multicolinéarité peut représenter un « problème ». En général, nous disons que la multicolinéarité intervient lorsque R_j^2 est « proche » de 1 ; l'utilisation des guillemets indique bien qu'il n'existe pas de nombre absolu nous permettant de conclure que la multicolinéarité représente un problème. Par exemple, $R_j^2 = 0{,}9$ signifie que 90 % de la variation de x_j dans l'échantillon peut être expliquée par les autres variables indépendantes du modèle de régression. Cela signifie sans aucun doute que la relation linéaire entre x_j et les autres variables indépendantes est forte. Néanmoins, $\text{Var}(\hat{\beta}_j)$ ne sera pas *nécessairement* élevée pour autant, car elle ne dépend pas uniquement du degré de multicolinéarité. Les deux autres facteurs, σ^2 et SCT_j, sont également déterminants. (Si le facteur SCT_j est beaucoup plus grand que le facteur σ^2, la présence de multicolinéarité n'est pas nécessairement un problème). Comme nous le verrons dans le chapitre 4, la qualité de l'inférence statistique repose avant tout sur la valeur absolue de $\hat{\beta}_j$ par rapport à son écart-type estimé.

De la même manière qu'un R_j^2 élevé *peut* conduire à une $\text{Var}(\hat{\beta}_j)$ élevée, une faible valeur de SCT_j conduira à une plus grande imprécision dans le calcul des estimateurs. Par conséquent, un échantillon de plus petite taille peut également conduire à une variance d'échantillonnage plus grande. S'inquiéter de la présence d'une corrélation élevée entre les variables indépendantes revient à s'inquiéter de l'utilisation d'un échantillon de petite taille ; ces deux situations contribuent à augmenter $\text{Var}(\hat{\beta}_j)$. Arthur Goldberger, le célèbre économétricien de l'Université du Wisconsin, a réagi à l'obsession des économétriciens vis-à-vis de la multicolinéarité en créant, non sans une ironie certaine, le terme « **micronumérosité** ». Il le définit comme « le problème lié à l'utilisation d'un échantillon de petite taille ». [Pour une discussion passionnante sur la multicolinéarité et sur la micronumérosité, voir Goldberger (1991).]

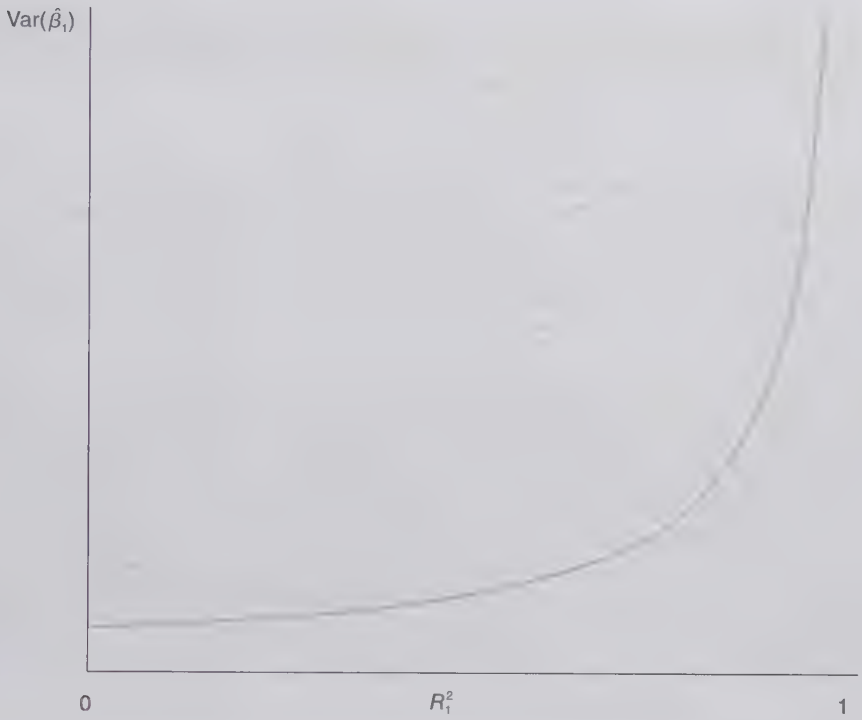


Figure 3.1 *Évolution de* $\text{Var}(\hat{\beta}_1)$ *en fonction de* R_1^2.

Même si la définition du problème de multicolinéarité n'est pas très précise, une chose est sure : lorsqu'il s'agit d'estimer β_j, il est préférable d'avoir le moins de corrélation possible entre x_j et les autres variables indépendantes, toutes choses égales par ailleurs. De nombreuses discussions portent d'ailleurs sur la meilleure manière de « résoudre » le problème de la multicolinéarité. Comme les données sont généralement obtenues de manière « passive » en sciences sociales, la réduction de la variance d'un estimateur non biaisé passe par la collecte du plus grand nombre d'observations possible. Pour une base de données déjà constituée, nous pouvons retirer du modèle les variables indépendantes dont la corrélation avec la variable d'intérêt est élevée. Malheureusement, si nous supprimons une variable qui appartient au modèle de population, cela peut conduire à l'apparition d'un biais de variable omise, ce que nous avons étudié dans la section 3.3.

À ce stade, il est sans doute utile de recourir à un exemple pour révéler davantage de subtilités de la multicolinéarité. Supposons que nous souhaitions estimer l'effet de différents types de dépenses scolaires sur les résultats des élèves, comme le paiement des salaires des enseignants, l'achat de matériels éducatifs, l'achat d'équipements sportifs, etc. Il est probable que la corrélation entre ces dépenses soit élevée. Les écoles les plus riches ont tendance à dépenser davantage dans tous les domaines alors que c'est l'inverse pour les écoles les plus pauvres. Sans surprise, il sera plutôt difficile d'estimer l'effet *ceteris paribus* d'une catégorie particulière de dépenses sur les résultats des élèves. En effet, la variation observée dans une catégorie sera largement expliquée par les variations enregistrées dans les autres catégories de dépenses, ce qui implique un R_j^2 élevé pour chacune des variables de dépenses. Ce type de multicolinéarité peut être atténué par une collecte de données supplémentaires. Ceci dit, en posant des questions auxquelles il est difficile d'apporter des réponses précises, nous sommes en partie responsable du problème que nous rencontrons. Nous pouvons certainement aboutir à de meilleurs résultats en modifiant l'objectif de l'analyse, par exemple en regroupant toutes les catégories de dépenses et en renonçant à estimer l'effet propre de chaque catégorie.

Un autre point important mérite d'être souligné. Une forte corrélation entre certaines variables explicatives peut n'avoir aucun impact sur la qualité de l'estimation des autres paramètres du modèle. Par exemple, considérons un modèle incluant trois variables indépendantes :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

Si x_2 et x_3 sont fortement corrélées, alors $\text{Var}(\hat{\beta}_2)$ et $\text{Var}(\hat{\beta}_3)$ peuvent être très élevées. Néanmoins, l'ampleur de la corrélation entre x_2 et x_3 n'a aucun effet direct sur $\text{Var}(\hat{\beta}_1)$. En effet, si x_1 n'est pas corrélée avec x_2 et x_3, alors $R_1^2 = 0$ et $\text{Var}(\hat{\beta}_1) = \sigma^2/\text{SCT}_1$, indépendamment de l'ampleur de la corrélation entre x_2 et x_3. Si β_1 est le paramètre qui nous intéresse, nous n'avons pas vraiment à nous préoccuper du degré de corrélation entre x_2 et x_3.

Pour aller plus loin 3.4

Supposons que vous proposiez un modèle expliquant les résultats universitaires obtenus à un examen de fin d'année en fonction de la présence en cours. La variable dépendante est la note obtenue à l'examen et la variable explicative clé est le nombre de cours auxquels l'étudiant a assisté. Pour prendre en compte les capacités innées des étudiants et leur travail réalisé en dehors des heures de cours, vous incluez la moyenne générale des notes obtenues à l'université (*colGPA*), le résultat au test *SAT* utilisé pour l'admission à l'université (aux États-Unis), et la moyenne générale des notes obtenues au lycée (*hsGPA*). Quelqu'un vous dit : « Vous ne pourrez tirer aucun enseignement de cet exercice car ces trois variables de contrôle sont probablement très corrélées. » Quelle est votre réponse ?

L'observation précédente est importante car les économistes incluent souvent un grand nombre de variables de contrôle pour isoler l'effet *ceteris paribus* d'une variable en particulier. Par exemple, pour identifier l'existence d'une discrimination (mesurée par le pourcentage de personnes issues de minorités dans les environs) dans l'accès aux prêts bancaires (mesuré par le taux d'approbation des prêts), nous devons inclure plusieurs variables de contrôle comme le revenu moyen, la valeur moyenne des habitations, différentes mesures de solvabilité, etc. Ces facteurs doivent être pris en compte pour pouvoir tirer des conclusions valables quant à l'existence d'une discrimination. Le revenu, le prix des logements et la solvabilité sont en général des variables fortement corrélées. Néanmoins, lorsqu'il s'agit d'estimer l'existence d'une discrimination dans l'accès aux prêts bancaires, une forte corrélation *entre ces variables* n'a pas beaucoup d'importance. (Ce sont les corrélations entre la variable de discrimination et les trois variables de contrôle qui jouent un rôle important).

Dans le but de mesurer l'ampleur de la multicolinéarité dans une régression, certains chercheurs jugent utile de calculer certaines statistiques qui peuvent malheureusement être utilisées à mauvais escient. Comme nous l'avons déjà dit, nous ne pouvons pas déterminer précisément le niveau à partir duquel la corrélation entre des variables explicatives doit être considérée comme « trop élevée ». Par exemple, certains tests de multicolinéarité sont basés sur des statistiques « globales » dans le sens où elles détectent une relation linéaire forte entre des sous-ensembles de variables explicatives. Leur utilité est limitée car elles peuvent tout simplement révéler un « problème » de multicolinéarité entre deux variables de contrôle, dont les coefficients ne nous intéressent pas du tout. [La statistique globale de multicolinéarité la plus fréquente est le *nombre de conditionnement*, qui dépend de la matrice complète des données ; cela dépasse la portée de cet ouvrage. Voir Belsley, Kuh, et Welsh (1980), par exemple.]

Certaines statistiques associées aux coefficients individuels sont un peu plus utiles, bien qu'elles puissent également faire l'objet d'un usage inapproprié. Parmi ces statistiques, le **facteur d'inflation de la variance** noté **VIF**, (pour « variance inflation factor » en anglais) est le plus populaire. Ce facteur est directement calculé à partir de l'équation (3.51). Le VIF du coefficient de la pente j est égal à $\text{VIF}_j = 1/(1 - R_j^2)$. Il correspond à

l'élément de $\text{Var}(\hat{\beta})$, qui est précisément déterminé par la corrélation entre x_j et les autres variables explicatives. Nous pouvons réécrire l'équation (3.51) comme suit :

$$\text{Var}(\hat{\beta}_j) = \left(\sigma^2 / SCT_j\right) VIF_j$$

Cette formulation montre bien qu'une augmentation de VIF_j, provoquée par une plus grande corrélation entre x_j et les autres variables explicatives, conduit à une augmentation de $\text{Var}(\hat{\beta}_j)$. Comme VIF_j est une fonction croissante de R_j^2 (l'ordonnée de la figure 3.1 permet de caractériser VIF_1), notre discussion précédente peut être reformulée en fonction du VIF. Si nous avions le choix, nous préférerions que VIF_j soit le plus petit possible, toutes choses égales par ailleurs. Nous avons cependant rarement le choix. Par exemple, si nous pensons que certaines variables explicatives de contrôle doivent être incluses dans la régression pour bien mesurer l'effet *ceteris paribus* de x_j, alors nous hésiterons à les supprimer, même si nous pensons que le VIF_j est « trop élevé », car cela pourrait introduire un biais de variable omise dans $\hat{\beta}_j$. Par contre, nous pouvons ignorer entièrement les VIF des *autres* coefficients si notre principal intérêt réside dans la mesure de l'effet causal de x_1 sur y. Enfin, il est arbitraire et peu utile de définir un seuil au-delà duquel le VIF démontrerait que la multicolinéarité est un « problème ». Une valeur égale à 10 est parfois choisie : si VIF_j est plus grand que 10 (ou, de manière équivalente, si R_j^2 est supérieur à 0,9), alors il faudrait conclure que la multicolinéarité pose « problème » pour estimer β_j. Pourtant, un VIF_j supérieur à 10 ne signifie pas que l'écart-type de $\hat{\beta}_j$ sera trop élevé puisque deux autres facteurs interviennent, σ^2 et SCT_j, la valeur du dernier pouvant également augmenter en fonction de la taille de l'échantillon. Par conséquent, comme c'était déjà le cas pour R_j^2, il est peu utile de se focaliser sur une valeur absolue du VIF_j, bien que certains le fassent.

Variance de l'estimateur dans un modèle mal spécifié

Le choix d'inclure ou non une variable dans un modèle de régression dépend en fait d'un compromis entre le biais et la variance de l'estimateur. Dans la section 3.3, nous avons calculé le biais provenant de l'omission d'une variable pertinente lorsque le vrai modèle contient deux variables explicatives. Poursuivons l'analyse de ce modèle en comparant cette fois les variances des estimateurs des MCO.

Nous écrivons le vrai modèle de la population, qui respecte les hypothèses de Gauss-Markov, de la manière suivante :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

Nous considérons deux estimateurs de β_1. L'estimateur $\hat{\beta}_1$ vient de la régression multiple :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \qquad \textbf{[3.52]}$$

En d'autres termes, nous incluons x_2 aux côtés de x_1 dans le modèle de régression. L'estimateur $\tilde{\beta}_1$ est obtenu en effectuant une régression simple de y sur x_1, après omission de la variable x_2 :

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \qquad \textbf{[3.53]}$$

Si $\beta_2 \neq 0$, l'équation (3.53) exclut une variable pertinente du modèle. Comme nous l'avons vu dans la section 3.3, cela implique que $\tilde{\beta}_1$ souffre d'un biais de variable omise, sauf si x_1 et x_2 ne sont pas corrélées. Par contre, $\hat{\beta}_1$ est un estimateur sans biais de β_1 pour n'importe quelle valeur de β_2, y compris $\beta_2 = 0$. Si l'absence de biais est notre seul critère de sélection, $\hat{\beta}_1$ est préférable à $\tilde{\beta}_1$.

Si maintenant nous évaluons la variance, nous ne pouvons plus affirmer que $\hat{\beta}_1$ est préférable à $\tilde{\beta}_1$ dans tous les cas. Étant donné les valeurs de x_1 et de x_2 dans l'échantillon, nous obtenons, à partir de (3.51), la relation suivante :

$$\text{Var}(\hat{\beta}_1) = \sigma^2 / [\text{SCT}_1 (1 - R_1^2)] \qquad \textbf{[3.54]}$$

où SCT_1 est la variation totale de x_1 et R_1^2 est le R carré de la régression de x_1 sur x_2. Par ailleurs, en modifiant quelque peu la démonstration du chapitre 2 concernant la régression simple, il est facile de démontrer que

$$\text{Var}(\tilde{\beta}_1) = \sigma^2 / SCT_1 \qquad [3.55]$$

En comparant (3.55) avec (3.54), nous constatons que $\text{Var}(\tilde{\beta}_1)$ est toujours *plus petite* que $\text{Var}(\hat{\beta}_1)$ sauf lorsque x_1 et x_2 ne sont pas corrélées dans l'échantillon, auquel cas les deux estimateurs $\tilde{\beta}_1$ et $\hat{\beta}_1$ sont identiques. Lorsque x_1 et x_2 ne sont pas corrélées, nous pouvons établir les conclusions suivantes.

1. Quand $\beta_2 \neq 0$, $\tilde{\beta}_1$ est biaisé, $\hat{\beta}_1$ est sans biais, et $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$.

2. Quand $\beta_2 = 0$, $\tilde{\beta}_1$ et $\hat{\beta}_1$ sont sans biais, et $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$.

La deuxième conclusion indique clairement que $\tilde{\beta}_1$ est l'estimateur le plus intéressant dans le cas où $\beta_2 = 0$. Intuitivement, si x_2 n'a aucun effet marginal sur y, son inclusion dans le modèle ne peut qu'exacerber le problème de multicolinéarité, ce qui conduit à un estimateur moins efficace de β_1. Le fait d'inclure une variable non pertinente dans le modèle a un coût, à savoir une plus grande variance de l'estimateur de β_1. Le cas où $\beta_2 \neq 0$ est plus ambigu. L'omission de x_2 conduit à un estimateur biaisé de β_1, dont la variance est néanmoins moins grande. Pour décider de l'omission de x_2, les économétriciens ont traditionnellement suggéré de comparer la taille probable du biais de x_1 avec la diminution de la variance mesurée par la taille de R_1^2. (L'omission de x_2 se justifie d'autant plus que le biais de variable omise attendu est ténu et que le R_1^2 est faible.) Deux raisons peuvent néanmoins nous pousser à inclure x_2 dans la régression. La première raison est liée à la taille de l'échantillon. D'une part, le biais de variable omise de $\tilde{\beta}_1$ ne diminue pas lorsque la taille de l'échantillon augmente. Comme l'évolution du biais ne suit pas de scénario prévisible, nous pouvons, sur un plan pratique, considérer que la taille de l'échantillon n'affecte pas le biais de variable omise de $\tilde{\beta}_1$. D'autre part, la multicolinéarité due à l'ajout de x_2 est de moins en moins importante lorsque la taille de l'échantillon augmente. En effet, $\text{Var}(\tilde{\beta}_1)$ et $\text{Var}(\hat{\beta}_1)$ tendent tous deux vers zéro quand n devient grand. Dans le cas où l'échantillon est grand, nous préférerons donc $\tilde{\beta}_1$.

Une autre raison, plus subtile, nous incite à inclure x_2 dans la régression et à préférer $\hat{\beta}_1$. Dans la formule (3.55), la variance de $\tilde{\beta}_1$ est conditionnelle aux valeurs de x_{i1} et x_{i2} dans l'échantillon, ce qui correspond au scénario idéal. En réalité, dans le cas où $\beta_2 \neq 0$, la variance de $\tilde{\beta}_1$ n'est conditionnelle que par rapport à x_1 ; elle est donc plus grande que ne le fait penser (3.55). Intuitivement, lorsque $\beta_2 \neq 0$ et que x_2 est exclu du modèle, la variance de *l'erreur*, σ^2, augmente car l'erreur contient effectivement une partie de x_2. Or, l'équation (3.55) ne prend pas en compte l'augmentation de la variance de l'erreur car elle traite les *deux* régresseurs comme non aléatoires. Une discussion approfondie sur le choix des variables à prendre en compte dans le calcul conditionnel dépasse l'objectif de cet ouvrage. Il suffit ici de préciser que (3.55) surestime la précision de $\tilde{\beta}_1$. Heureusement, les logiciels d'analyse statistique utilisent le bon estimateur de la variance, il n'est donc pas nécessaire de connaître les subtilités des formules théoriques. Une fois que vous aurez lu la prochaine sous-partie, vous pourrez étudier les problèmes 14 et 15 pour approfondir la question.

Estimation de σ^2 et écarts-types estimés des MCO

Pour obtenir un estimateur sans biais de $\text{Var}(\tilde{\beta}_j)$, il est impératif de choisir un estimateur sans biais de σ^2.

Comme $\sigma^2 = E(u)^2$, l'idéal serait d'utiliser la moyenne des erreurs au carré dans l'échantillon, $n^{-1}\sum_{i=1}^{n} u_i^2$. C'est malheureusement impossible puisque nous ne pouvons pas observer les u_i. Rappelez-vous l'égalité suivante : $u_i = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i1} - \ldots - \beta_k x_{ik}$. La raison pour laquelle nous ne pouvons pas observer les u_i est que les β_j ne sont pas connus. Si nous remplaçons les β_j par leurs estimations des MCO, nous obtenons les résidus des MCO :

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}.$$

Il est donc naturel de vouloir estimer σ^2 en remplaçant u_i par $\hat{u}_i$. Le travail n'est pourtant pas fini car nous avons vu que cela conduisait à un estimateur biaisé dans le cas de la régression simple. Pour obtenir l'estimateur sans biais de σ^2 dans le cas général de la régression multiple, nous devons écrire :

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} \hat{u}_i^2 \right) / (n - k - 1) = \text{SCR}/(n - k - 1) \qquad \textbf{[3.56]}$$

Nous avons déjà rencontré cet estimateur dans le cas spécifique de la régression simple où $k = 1$. Le terme $n - k - 1$ dans (3.56) représente **les degrés de liberté (*ddl*)** dans le cas général des MCO où il y a n observations et k variables indépendantes. Comme il y a $k + 1$ paramètres dans un modèle de régression, soit k variables indépendantes et une ordonnée à l'origine, nous pouvons écrire :

$$ddl = n - (k + 1) \qquad \textbf{[3.57]}$$

$$= \text{(nombre d'observations)} - \text{(nombre de paramètres estimés)}.$$

Il s'agit de la manière la plus simple de calculer les degrés de liberté dans une application particulière. Il suffit de compter le nombre de paramètres, en n'oubliant pas l'ordonnée à l'origine, puis de soustraire cette valeur du nombre total d'observations. (Le nombre de paramètres est égal à k dans les rares cas où le modèle ne contient aucune ordonnée à l'origine).

D'un point de vue mathématique, la division par $n - k - 1$ dans (3.56) se justifie par le fait que la valeur attendue de la somme des carrés des résidus est précisément : $\text{E(SCR)} = (n - k - 1)\sigma^2$. D'un point de vue intuitif, l'importance que représentent les degrés de liberté peut se comprendre en revenant aux conditions de premier ordre des estimateurs des MCO, soit $\sum_{i=1}^{n} \hat{u}_i = 0$ et $\sum_{i=1}^{n} x_{ij} \hat{u}_i = 0$, avec $j = 1, 2, \ldots, k$. Par conséquent, nous imposons $k + 1$ restrictions sur les résidus des MCO lorsque nous cherchons à obtenir les estimateurs des MCO. Cela signifie que, si nous considérons $n - (k + 1)$ résidus, les $(k + 1)$ restants sont connus. Il n'y a que $n - (k + 1)$ degrés de liberté dans les résidus (à la différence des *erreurs* u_i, qui disposent de n degrés de liberté dans l'échantillon).

Le théorème 3.3 nous permet de synthétiser la discussion que nous venons de tenir. Nous l'avons démontré au chapitre 2 dans le cadre de la régression simple (voir théorème 2.3). (L'annexe E contient la démonstration générale qui requiert l'utilisation de l'algèbre matricielle.)

Théorème 3.3
Estimation sans biais de σ^2

Sous les hypothèses de Gauss-Markov RLM.1 à RLM.5, $\text{E}(\hat{\sigma}^2) = \sigma^2$.

La racine carrée de $\hat{\sigma}^2$, notée $\hat{\sigma}$, est appelée **l'écart-type de la régression (ETR)**. Contrairement à l'écart-type de l'erreur (σ), l'ETR correspond à un écart-type estimé puisqu'il est calculé à partir des résidus de la régression. Il correspond à un estimateur de l'écart-type du terme d'erreur. Cet estimateur est également désigné sous les deux sigles anglais suivants : RMSE (« root mean squared error ») et SER (« standard error of the regression). Dans les logiciels économétriques, il est souvent calculé automatiquement, même s'il peut apparaître sous différentes appellations.

Pour un échantillon donné, il est important de noter que $\hat{\sigma}$ peut diminuer *ou* augmenter quand une variable explicative est ajoutée à la régression. Dans un tel cas de figure, les degrés de liberté diminuent d'une unité. La SCR sera également moins élevée mais nous ne pouvons pas dire *a priori* quelle sera l'ampleur de cette baisse. Comme la SCR est au numérateur et que les *ddl* sont au dénominateur, l'effet global sur $\hat{\sigma}$ est incertain.

Pour construire les intervalles de confiance et effectuer les tests que nous décrirons dans le chapitre 4, nous avons besoin d'estimer l'**écart-type de** $\hat{\beta}_j$, qui correspond tout simplement à la racine carrée de la variance, soit :

$$\sigma(\hat{\beta}_j) = \sigma/[\text{SCT}_j(1 - R_j^2)]^{1/2}.$$

Notez que $\sigma(\hat{\beta}_j)$ est parfois écrit « sd($\hat{\beta}_j$) », ce qui correspond à l'abréviation de « standard deviation » en anglais. Comme σ est inconnu, nous le remplaçons par son estimateur, $\hat{\sigma}$. Cela nous donne l'**écart-type estimé de** $\hat{\beta}_j$, soit :

$$\hat{\sigma}(\hat{\beta}_j) = \hat{\sigma} / [SCT_j(1 - R_j^2)]^{1/2}. \qquad \textbf{[3.58]}$$

Notez également que $\hat{\sigma}(\hat{\beta}_j)$ est parfois écrit « se($\hat{\beta}_j$) », ce qui correspond à l'abréviation de « standard error » en anglais. Que ce soient les estimations des β_j ou les écarts-types estimés des $\hat{\beta}_j$, nous pouvons les obtenir par les MCO à partir de n'importe quel échantillon donné. Vu que $\hat{\sigma}(\hat{\beta}_j)$ dépend de $\hat{\sigma}$, l'écart-type estimé dispose d'une distribution d'échantillonnage qui jouera un rôle primordial dans le chapitre 4.

Nous devons également souligner que les écarts-types estimés ne sont valides que si l'hypothèse d'homoscédasticité est respectée. Comme (3.58) est obtenue directement à partir de la formule de la variance (3.51), et que (3.51) repose sur l'hypothèse d'homoscédasticité RLM.5, la formule de l'écart-type estimé en (3.58) n'est *pas* un estimateur valide de $\sigma(\hat{\beta}_j)$ dans le cas où les erreurs sont hétéroscédastiques. Si la présence d'hétéroscédasticité n'entraîne pas de biais dans les $\hat{\beta}_j$, elle conduit à un biais dans la formule de Var($\hat{\beta}_j$), ce qui invalide le calcul des écarts-types estimés. C'est un élément important car les logiciels économétriques calculent par défaut l'écart-type estimé pour chaque coefficient à partir de (3.58), avec une représentation quelque peu différente pour la constante. Dans le chapitre 8, nous étudierons les méthodes qui permettent d'identifier l'hétéroscédasticité et d'en tenir compte dans le calcul des écarts-types estimés.

Dans certains cas, il est utile d'écrire :

$$\hat{\sigma}(\hat{\beta}_j) = \hat{\sigma} / [\sqrt{n}\sigma(x_j)\sqrt{1 - R_j^2}\,] \qquad \textbf{[3.59]}$$

où $\sigma(x_j) = \sqrt{n^{-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$. $\sigma(x_j)$ représente l'écart-type de population, dont la somme des carrés totaux est divisée par n plutôt que par $n - 1$. L'équation (3.59) indique que la taille de l'échantillon, n, affecte directement les écarts-types estimés. Les trois autres termes dans la formule, soit $\hat{\sigma}$, $\sigma(x_j)$ et R_j^2, varient en fonction de l'échantillon choisi mais ils ne dépendent pas de n. Grâce à l'équation (3.59), nous pouvons constater que les écarts-types estimés tendent vers zéro à la vitesse $1/\sqrt{n}$. Cette formule montre qu'il est intéressant de constituer des échantillons de grande taille : la précision des $\hat{\beta}_j$ augmente lorsque n augmente. Par exemple, la précision des $\hat{\beta}_j$ sera deux fois plus grande si le nombre de données est multiplié par 4. Dans le chapitre 5, nous étudierons en détail les propriétés des MCO lorsque la taille de l'échantillon est très grande. (Souvenez-vous, néanmoins, que la propriété d'absence de biais ne dépend pas de n ; le caractère non biaisé d'un estimateur vaut pour n'importe quelle taille de l'échantillon, à condition naturellement qu'il soit possible de le calculer.)

3.5 EFFICACITÉ DES MCO : LE THÉORÈME DE GAUSS-MARKOV

Dans cette section, nous définissons et discutons le théorème de Gauss-Markov qui montre que l'estimateur des MCO est préférable à d'autres estimateurs alternatifs. Nous avons déjà justifié l'utilisation des MCO sur base des hypothèses RLM.1 à RLM.4 : sous ces hypothèses, les estimateurs des paramètres, calculés à l'aide

de la méthode des MCO, sont sans biais. Il existe néanmoins de *nombreux* estimateurs qui jouissent de la propriété d'absence de biais lorsque ces hypothèses sont respectées (voir le problème 13, par exemple). La question est maintenant de savoir si la *variance* de l'estimateur sans biais des MCO est plus faible que celle de tous les estimateurs sans biais disponibles.

Si nous définissons la classe des estimateurs alternatifs de manière appropriée, nous pouvons effectivement montrer que la méthode des MCO permet d'obtenir le meilleur estimateur. Plus précisément, nous pouvons montrer que, sous les hypothèses RLM.1 à RLM.5, l'estimateur des MCO de β_j, soit $\hat{\beta}_j$, est le **meilleur estimateur linéaire sans biais**. En anglais, il s'agit du « **best linear unbiased estimator** » (**BLUE**). Avant de formuler le théorème de Gauss-Markov, il est important de bien comprendre ce que nous entendons exactement par « BLUE ». Nous savons tout d'abord qu'un estimateur représente une formule qui permet d'obtenir une estimation pour n'importe quel échantillon de données. Nous savons ensuite qu'un estimateur de β_j, soit $\tilde{\beta}_j$, sera sans biais si $E(\tilde{\beta}_j) = \beta_j$, pour tout $\beta_0, \beta_1, \ldots, \beta_k$.

Quant à l'adjectif « linéaire », il signifie que l'estimateur de β_j, soit $\tilde{\beta}_j$, est linéaire si et seulement si cet estimateur peut être exprimé comme une fonction linéaire des observations de la variable dépendante :

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i \qquad \textbf{[3.60]}$$

où chaque w_{ij} est une fonction des valeurs prises par les variables indépendantes au sein de l'échantillon. Les estimateurs des MCO sont linéaires, comme nous pouvons le constater dans l'équation (3.22).

Enfin, quelle définition pouvons-nous donner à l'adjectif « meilleur » ? Dans le théorème de Gauss-Markov, il désigne l'estimateur *qui a la plus petite variance.* Étant donné deux estimateurs sans biais, il est logique de préférer celui qui a la variance la plus petite (voir l'annexe C).

Supposons à présent que $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ représentent les estimateurs du modèle (3.31) que nous obtenons par la méthode des MCO sous les hypothèses RLM.1 à RLM.5. Selon le théorème de Gauss-Markov, $\text{Var}(\hat{\beta}_j) \leq \text{Var}(\tilde{\beta}_j)$ pour tout estimateur $\tilde{\beta}_j$ qui est *linéaire* et *sans biais*, l'inégalité étant généralement stricte. En d'autres termes, dans la classe des estimateurs linéaires sans biais, les MCO permettent d'obtenir l'estimateur dont la variance est la plus faible (sous les cinq hypothèses de Gauss-Markov). Le théorème en dit même davantage. Si nous voulons estimer une fonction linéaire quelconque des β_j, alors la combinaison linéaire des estimateurs des MCO qui en résulte affichera également la plus petite variance parmi tous les estimateurs linéaires sans biais. Le théorème de Gauss-Markov est démontré dans l'annexe 3A.

Théorème 3.4
Théorème de Gauss-Markov

Sous les hypothèses RLM.1 à RLM.5, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ sont les meilleurs estimateurs linéaires sans biais (BLUE) de $\beta_0, \beta_1, \ldots, \beta_k$ respectivement.

C'est en raison de ce théorème que les hypothèses RLM.1 à RLM.5 sont également appelées les hypothèses de Gauss-Markov.

Le théorème de Gauss-Markov revêt une importance particulière : lorsque les hypothèses classiques tiennent, il est inutile de chercher des estimateurs sans biais alternatifs, tels que décrits en (3.60), car aucun d'entre eux ne sera meilleur que celui des MCO. De manière équivalente, si nous rencontrons un estimateur qui est à la fois linéaire et sans biais, alors nous savons que la variance de cet estimateur est au moins aussi grande que la variance de l'estimateur des MCO. Aucun calcul supplémentaire n'est requis.

Dans le contexte de la régression multiple, le théorème 3.4 justifie l'utilisation des MCO comme méthode d'estimation. Si l'une des hypothèses de Gauss-Markov est violée, alors le théorème de Gauss-Markov ne tient plus. Nous savons déjà que si l'hypothèse d'espérance conditionnelle nulle (hypothèse RLM.4) n'est pas respectée, l'estimateur des MCO sera biaisé et le théorème 3.4 ne s'appliquera pas. Nous savons aussi que l'hétéroscédasticité (qui invalide l'hypothèse RLM.5) n'introduit pas de biais dans l'estimateur des MCO. Par contre, en présence d'hétéroscédasticité, l'estimateur des MCO ne dispose plus de la plus petite variance parmi les estimateurs linéaires sans biais. Dans le chapitre 8, nous verrons qu'il est possible de calculer un estimateur qui, en présence d'hétéroscédasticité, est supérieur à celui des MCO.

3.6 QUELQUES COMMENTAIRES SUR LA TERMINOLOGIE

Il arrive fréquemment que les débutants, et parfois même les chercheurs expérimentés, affirment qu'ils ont « estimé un modèle des MCO ». Même si nous comprenons sans difficulté ce qu'ils veulent dire, cette affirmation est fausse, tant sur le fond que sur la forme. Elle trahit une mauvaise compréhension de l'analyse par régression multiple.

Il faut tout d'abord garder à l'esprit que les moindres carrés ordinaires (MCO) ne représentent qu'une méthode d'estimation. Les MCO ne correspondent pas à un modèle. Un modèle décrit une population sous-jacente et dépend de paramètres inconnus. Dans la population, le *modèle linéaire* que nous avons étudié dans ce chapitre peut s'écrire sous la forme

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \qquad \textbf{[3.61]}$$

les paramètres étant représentés par les β_j. Notez que nous pouvons interpréter les β_j sans utiliser de données. Même s'il est vrai que nous n'apprendrons pas grand-chose au sujet des β_j sans les estimer, l'interprétation des β_j dépend du modèle linéaire (3.61) et non des données.

Une fois que nous disposons d'un échantillon représentatif de données, nous pouvons estimer les paramètres du modèle. Ces estimations peuvent s'obtenir en appliquant la méthode des MCO que nous avons privilégiée jusqu'ici. En réalité, il existe un grand nombre de techniques d'estimation sur lesquelles nous pouvons nous appuyer. Nous nous sommes concentrés sur les MCO en raison de leur popularité, justifiée par les considérations statistiques que nous avons abordées précédemment dans ce chapitre. La méthode des MCO repose néanmoins sur des hypothèses bien précises (RLM.1 à RLM.5). Comme nous le verrons ultérieurement dans différents chapitres, si une ou plusieurs de ces hypothèses ne sont pas respectées, il existe d'autres méthodes d'estimation préférables aux MCO. Parmi les méthodes d'estimation alternatives, nous pouvons citer les moindres carrés pondérés (MCP), les moindres déviations absolues (MDA) et les variables instrumentales que nous aborderons dans les chapitres 8, 9 et 15. Quelle que soit la méthode d'estimation choisie, notre modèle reste celui caractérisé par l'équation (3.61).

Certains lecteurs pourraient considérer que cette discussion est tatillonne et que l'expression « estimer un modèle *des* MCO » est équivalente à l'expression « estimer un modèle *par les* MCO ». Nous devons néanmoins nous souvenir que le modèle (3.61) « n'appartient pas » à la méthode des MCO. Nous avons étudié les propriétés des estimateurs des MCO sous différentes hypothèses. Par exemple, nous savons que l'estimateur des MCO est sans biais si les quatre premières hypothèses de Gauss-Markov sont respectées mais qu'il n'est pas efficace sans l'hypothèse RLM.5. Nous avons également vu que l'omission d'une variable importante pouvait conduire à la violation de l'hypothèse RLM.4 et à un estimateur des MCO biaisé. L'utilisation d'un vocabulaire imprécis s'accompagne souvent d'une analyse superficielle des hypothèses sur lesquelles repose le modèle sous-jacent. Or, le choix d'un estimateur repose sur la détermination des hypothèses auxquelles nous pouvons raisonnablement recourir.

Comme application du modèle (3.61), considérons l'équation suivante dont l'objectif est d'expliquer les résultats obtenus à un examen de mathématiques en quatrième année d'études :

$$maths4 = \beta_0 + \beta_1 classize4 + \beta_2 maths3 + log(income) + \beta_4 motheduc + \beta_5 fatheduc + u \quad \textbf{[3.62]}$$

Nous pouvons tout d'abord déterminer s'il est raisonnable de maintenir l'hypothèse RLM.4 en réfléchissant aux facteurs qui sont laissés dans le terme d'erreur *u*. Nous pouvons également vérifier s'il est nécessaire d'introduire des relations fonctionnelles plus complexes, un sujet que nous étudierons en détail dans le chapitre 6. Nous pouvons ensuite décrire le jeu de données, qui est idéalement obtenu par échantillonnage aléatoire, et commenter les estimations obtenues par les MCO à partir de l'échantillon. Une bonne manière de lancer la discussion sur les estimations est de dire : « j'ai estimé l'équation (3.62) par les moindres carrés ordinaires. Dans le cadre d'un échantillonnage aléatoire et sous l'hypothèse qu'aucune variable importante n'a été omise, l'estimateur des MCO de l'effet *ceteris paribus* de la taille de la classe (*classize*), soit β_1, n'est pas biaisé. Si le terme d'erreur *u* a une variance constante, l'estimateur des MCO est également le meilleur estimateur linéaire sans biais ». Comme nous le verrons aux chapitres 4 et 5, nous en dirons bien davantage sur les MCO. Nous pouvons évidemment aussi souligner que le modèle ne tient pas compte de tous les facteurs qui différencient les élèves : outre les résultats de maths en troisième année (*maths*3), le revenu familial (*income*) et le niveau d'études des parents (*fatheduc* et *motheduc*), d'autres facteurs peuvent jouer – par exemple, *u* inclut la motivation de l'élève ou l'implication des parents – auquel cas l'estimateur des MCO peut être biaisé.

Il existe une autre raison, plus subtile, pour laquelle il convient de ne pas assimiler le modèle de la population à la méthode d'estimation. Une méthode d'estimation, comme celle des MCO, peut être utilisée dans le cadre d'un simple exercice de prévision, sans que nous ne devions pour autant nous soucier du modèle sous-jacent et des propriétés statistiques usuelles d'absence de biais et d'efficacité. Par exemple, nous pouvons utiliser les MCO pour estimer une droite dans le seul but de prédire la moyenne générale à l'université qu'obtiendront un ensemble d'élèves de lycée aux caractéristiques données.

RÉSUMÉ

1. Le modèle de régression linéaire multiple nous permet d'étudier l'effet d'une variable indépendante sur la variable dépendante tout en tenant compte de l'effet des autres variables indépendantes présentes dans le modèle. Le modèle peut ainsi tenir explicitement compte de la corrélation qui peut exister entre les variables indépendantes.

2. Pour autant que le modèle soit linéaire dans ses *paramètres*, il peut servir à estimer par les MCO des relations non linéaires, à condition d'introduire les variables dépendante et indépendantes sous une forme pertinente.

3. Il est facile d'estimer le modèle de régression linéaire multiple par les moindres carrés ordinaires. L'estimation du coefficient d'une variable indépendante mesure son effet *ceteris paribus* sur la variable dépendante.

4. Le *R* carré mesure la proportion de la variation de *y* au sein de l'échantillon qui est expliquée par les x_j. Il mesure la qualité d'ajustement d'un modèle. Il ne faut pas lui donner trop d'importance, sans le négliger pour autant.

5. Sous les quatre premières hypothèses de Gauss-Markov (RLM.1 à RLM.4), les estimateurs des MCO sont sans biais. Cela implique que l'ajout d'une variable supplémentaire dans le modèle n'apporte rien : cette variable est superflue. Son inclusion n'aura aucun effet sur l'absence de biais dans les estimateurs de

l'ordonnée à l'origine et de la pente. D'un autre côté, l'oubli d'une variable pertinente implique que les MCO sont biaisés. Dans de nombreux cas, il est possible de déterminer la direction de ce biais.

6. Sous les cinq hypothèses de Gauss-Markov, la variance d'un estimateur des MCO pour la pente est donnée par $\text{Var}(\hat{\beta}_j) = \sigma^2 / [\text{SCT}_j(1 - R_j^2)]$. Plus la variance σ^2 augmente, plus $\text{Var}(\hat{\beta}_j)$ augmente. Quand la variation des x_j dans l'échantillon, SCT_j, augmente, $\text{Var}(\hat{\beta}_j)$ diminue. Le terme R_j^2 mesure l'ampleur de la colinéarité entre x_j et les autres variables explicatives. Quand R_j^2 tend vers un, $\text{Var}(\hat{\beta}_j)$ tend vers l'infini.

7. L'ajout d'une variable non pertinente introduit généralement de la multicolinéarité et augmente les variances des estimateurs des MCO.

8. Sous les hypothèses de Gauss-Markov (RLM.1 à RLM.5), les estimateurs des MCO sont les meilleurs estimateurs linéaires sans biais (BLUE).

9. À partir du chapitre 4, nous utiliserons les écarts types estimés des coefficients des MCO pour calculer les intervalles de confiance des paramètres de population. Nous les utiliserons également pour calculer des statistiques de test permettant de tester des hypothèses sur les paramètres de population. Ainsi, lorsque nous présenterons les résultats de régression, nous inclurons à présent les écarts types estimés. Dans les équations, les écarts types estimés sont habituellement reportés entre parenthèses en dessous des coefficients estimés. La même convention est souvent utilisée dans les tableaux reportant les résultats des MCO.

LES HYPOTHÈSES DE GAUSS-MARKOV

Nous résumons les cinq **hypothèses de Gauss-Markov** que nous avons utilisées dans ce chapitre. Souvenez-vous que seules les quatre premières hypothèses, RLM.1 à RLM.4, sont requises pour démontrer l'absence de biais de l'estimateur des MCO. La cinquième hypothèse, RLM.5, permet d'obtenir les formules traditionnelles de la variance. Sous ces cinq hypothèses, l'estimateur des MCO est le meilleur estimateur linéaire sans biais. Autrement dit, l'estimateur des MCO est « BLUE ».

Hypothèse RLM.1 (Linéarité dans les paramètres)

Le modèle dans la population peut être écrit comme

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

où $\beta_0, \beta_1, \ldots, \beta_k$ sont les paramètres d'intérêt inconnus et u est une erreur ou perturbation aléatoire.

Hypothèse RLM.2 (Échantillonnage aléatoire)

Nous disposons d'un échantillon aléatoire de n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$, tiré du modèle de population décrit sous l'hypothèse RLM.1.

Hypothèse RLM.3 (Absence de colinéarité parfaite)

Dans l'échantillon (et par conséquent dans la population), aucune des variables indépendantes n'est constante et il n'y a aucune relation *linéaire exacte* entre les variables indépendantes.

Hypothèse RLM.4 (Espérance conditionnelle de l'erreur égale à zéro)

Le terme d'erreur u affiche une espérance égale à zéro, quelle que soit la valeur prise par les variables indépendantes. En d'autres termes,

$$\text{E}(u|x_1, x_2, \ldots, x_k) = 0.$$

Hypothèse RLM.5 (Homoscédasticité)

La variance de l'erreur u est constante, quelle que soit la valeur prise par les variables explicatives. En d'autres termes,

$$\text{Var}(u|x_1, \ldots, x_k) = \sigma^2.$$

MOTS-CLÉS

Biais de variable omise ou biais d'omission p. 118
Biais vers le bas p. 120
Biais vers le haut p. 120
Biais vers zéro p. 120
Colinéarité parfaite p. 113
Conditions du premier ordre p. 101
Degré(s) de liberté (*ddl*) p. 129
Droite de régression des MCO p. 102
Écart-type de $\hat{\beta}_j$, soit $\sigma(\hat{\beta}_j)$ ou $\text{sd}(\hat{\beta}_j)$ p. 130
Écart-type estimé de $\hat{\beta}_j$, soit $\hat{\sigma}(\hat{\beta}_j)$ ou $\text{se}(\hat{\beta}_j)$ p. 130
Écart-type (estimé) de la régression (ETR ou SER), soit $\hat{\sigma}$ p. 129
Écart-type de l'erreur, soit σ p. 130
Effet *ceteris paribus* p. 102
Effet marginal p. 102
Erreur de spécification p. 117
Estimation de la pente par MCO p. 102
Estimation de l'ordonnée à l'origine par MCO p. 102
Exclusion d'une variable pertinente p. 117
Facteur d'inflation de la variance (VIF) p. 126
Fonction de régression de l'échantillon (FRE) p. 102
Hypothèses de Gauss-Markov p. 122
Inclusion d'une variable non pertinente (ou superflue) p. 116
Meilleur estimateur linéaire sans biais (BLUE, en anglais) p. 131
Micronumérosité p. 124
Modèle de régression linéaire multiple (RLM) p. 96, 99
Modèle vrai p. 112
Moindres carrés ordinaires (MCO) p. 100
Multicolinéarité p. 124
Ordonnée à l'origine p. 99
Paramètre de la pente p. 99
Perturbation p. 99
Résidu p. 105
Somme des carrés des résidus (SCR) p. 108
Somme des carrés expliqués (SCE) p. 108
Somme des carrés totaux (SCT) p. 108
Sous-spécifier un modèle p. 117
Surspécifier un modèle p. 116
Terme d'erreur p. 99
Théorème de Frish-Waugh p. 107
Théorème de Gauss-Markov p. 130

Variable explicative endogène p. 115
Variable explicative exogène p. 115

EXERCICES

1. En utilisant la base de données GPA2 portant sur 4 137 étudiants inscrits à l'université aux États-Unis, l'équation suivante a été estimée par les MCO :

$$\widehat{colgpa} = 1{,}392 - 0{,}0135\,hsperc + 0{,}00148\,sat$$
$$n = 4\,137,\ R^2 = 0{,}273$$

où *colgpa* mesure, sur une échelle de quatre points, la moyenne des résultats obtenus à l'université ; *hsperc* est le percentile auquel se situe l'étudiant au sein de sa promotion (défini de telle sorte que *hsperc* = 5 désigne les 5 % supérieurs de la promotion) ; *sat* représente la combinaison de résultats obtenus en mathématiques et en vocabulaire à un test (le « SAT » ou « Scholastic Assessment Test ») que les étudiants aux États-Unis passent avant l'entrée à l'université.

i. Pourquoi est-il logique que le coefficient de *hsperc* soit négatif ?

ii. Quelle est la moyenne générale prédite à l'université (*colgpa*) quand *hsperc* = 20 et *sat* = 1 050 ?

iii. Supposons que deux lycéens, A et B, se situent au même percentile dans leurs promotions respectives mais que la note obtenue au « SAT » par l'étudiant A soit de 140 points supérieure à celle obtenue par l'étudiant B (ce qui correspond à l'écart-type observé dans l'échantillon pour la variable *sat*). Quelle sera l'estimation de la différence entre la moyenne obtenue à l'université (*colgpa*) par A et celle obtenue par B ? Cette différence est-elle importante ?

iv. Si la variable *hsperc* demeure constante (*ceteris paribus*), quelle est la variation dans la note obtenue au « SAT » qui conduit à une augmentation d'un demi-point dans la moyenne générale obtenue à l'université (soit une différence de *colgpa* égale à 0,50) ? Commentez votre réponse.

2. Les données du fichier WAGE2 sur la population employée masculine ont été utilisées pour estimer l'équation suivante :

$$\widehat{educ} = 10{,}36 - 0{,}094\,sibs + 0{,}131\,meduc + 9{,}210\,feduc$$
$$n = 722,\ R^2 = 0{,}214$$

où *educ* est le nombre d'années d'études, *sibs* est le nombre de frères et sœurs, *meduc* est le nombre d'années d'études de la mère, et *feduc* est celui du père. La variable *sibs* a-t-elle l'effet attendu ? Expliquez. Les autres variables (*meduc* et *feduc*) demeurant constantes (*ceteris paribus*), quelle est l'augmentation de *sibs* qui conduit à une diminution d'une année d'études ? (La réponse ne correspond pas forcément à un nombre entier).

i. Donnez une interprétation au coefficient de *meduc*.

ii. Supposez que l'individu A est fils unique et que ses parents ont chacun fait 12 années d'études. L'individu B est également fils unique mais ses parents ont fait chacun 16 ans d'études. Quelle sera la différence d'instruction prédite entre A et B, en nombre d'années d'études ?

3. Le modèle suivant est une version simplifiée du modèle de régression multiple utilisé par Biddle et Hamermesh (1990) pour étudier la relation entre sommeil et travail, en tenant compte d'autres facteurs qui affectent la relation. Soit

$$sleep = \beta_0 + \beta_1\,totwrk + \beta_2\,educ + \beta_3\,age + u$$

où le sommeil (*sleep*) et le travail (*totwrk*, « total work » en anglais) sont mesurés en minutes par semaine. Les variables *educ* et *age* sont mesurées en années. (Voir aussi l'exercice sur ordinateur C3 dans le chapitre 2.)

i. Si les adultes dorment moins pour travailler, quel est le signe attendu de β_1 ?

ii. Quels sont, à votre avis, les signes attendus de β_2 et β_3 ?

iii. En utilisant les données de SLEEP75, l'équation estimée est :

$$\widehat{sleep} = 3638{,}25 - 0{,}148\,towrk - 11{,}13\,edu + 2{,}20\,age$$
$$n = 706,\ R^2 = 0{,}113$$

Quel est l'effet *ceteris paribus* sur le sommeil (estimé en minutes) d'une augmentation de cinq heures de travail par semaine ? Est-ce important ?

iv. Que pensez-vous du signe et de la taille du coefficient estimé pour *educ* ?

v. Estimez-vous que *totwrk*, *educ*, et *age* expliquent une grande part de la variation de *sleep* ? Quels autres facteurs pourraient affecter le sommeil ? Sont-ils susceptibles d'être corrélés avec *totwrk* ?

4. Le modèle suivant vise à expliquer le salaire médian (*salary*) que les nouveaux diplômés en droit aux États-Unis perçoivent à l'embauche :

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + u$$

L'échantillon regroupe des écoles de droit aux États-Unis. *LSAT* est la note médiane obtenue par la promotion à un test d'entrée (le « Law School Admission Test ») ; *GPA* est la médiane de la moyenne générale obtenue à l'université par la promotion ; *libvol* est le nombre d'ouvrages disponibles dans la bibliothèque de l'école ; *cost* est le coût d'inscription annuel à l'école, et *rank* est la position de l'école dans un classement national (*rank* = 1 désigne la meilleure école).

i. Pourquoi doit-on s'attendre à $\beta_5 \leq 0$?

ii. Quels sont les signes attendus pour les autres paramètres ? Justifiez vos réponses.

iii. En utilisant les données LAWSCH85, nous pouvons obtenir les estimations suivantes :

$$\widehat{\log(salary)} = 8{,}34 + 0{,}0047\,LSAT + 0{,}248\,GPA + 0{,}095\log(libvol) + 0{,}038\log(cost) - 0{,}0033\,rank$$
$$n = 136,\ R^2 = 0{,}842$$

Toutes choses étant égales par ailleurs (*ceteris paribus*), quelle est l'estimation de la différence de salaire entre des écoles dont la médiane de la moyenne générale (*GPA*) diffèrent d'un point ? Indiquez votre réponse en pourcentage.

iv. Interprétez le coefficient de la variable log(*libvol*).

v. Est-il préférable d'être dans une école de droit mieux classée ? Que « rapporte », *ceteris paribus*, une amélioration de 20 places dans le classement (en termes de salaire estimé à l'embauche) ?

5. Vous distribuez un questionnaire à plusieurs étudiants dans le but d'estimer le lien entre la moyenne générale obtenue à l'université (*GPA*) et le temps consacré à différentes activités. Il est notamment demandé aux étudiants d'évaluer les heures consacrées chaque semaine à leurs diverses activités en les classant obligatoirement dans les quatre catégories suivantes : les études, le sommeil, le travail rémunéré, et le divertissement. Pour chaque étudiant, la somme des heures consacrées aux quatre activités doit donc être égale à 168 heures, soit 7 journées de 24 heures.

i. Dans le modèle

$$GPA = \beta_0 + \beta_1\, study + \beta_2\, sleep + \beta_3\, work + \beta_4\, leisure + u$$

est-il sensé de faire varier *study* en cherchant à maintenir *sleep*, *work*, et *leisure* constants ?

ii. Pourquoi ce modèle viole-t-il l'hypothèse RLM.3 ? Expliquez.

iii. Comment pourriez-vous reformuler ce modèle en respectant l'hypothèse RLM.3 et en lui permettant d'avoir une interprétation utile ?

6. Considérons le modèle de régression multiple suivant, qui contient trois variables indépendantes et respecte les hypothèses RLM.1 à RLM.4 :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Vous souhaitez estimer la somme des paramètres associés à x_1 et x_2 ; soit $\theta_1 = \beta_1 + \beta_2$.

i. Montrez que $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ est un estimateur sans biais de θ_1.

ii. Exprimez $\text{Var}(\hat{\theta}_1)$ en fonction de $\text{Var}(\hat{\beta}_1)$, $\text{Var}(\hat{\beta}_2)$ et $\text{Corr}(\hat{\beta}_1, \hat{\beta}_2)$.

7. Lequel de ces éléments peut conduire l'estimateur des MCO à être biaisé ?

i. L'hétéroscédasticité.

ii. L'oubli d'une variable importante.

iii. Une corrélation d'échantillon de 0,95 entre deux variables indépendantes incluses toutes deux dans le modèle.

8. Supposons que la productivité moyenne des employés dans l'industrie manufacturière (*avgprod*) dépende de deux facteurs : le nombre moyen d'heures de formation (*avgtrain*) et les capacités moyennes des employés :

$$avgprod = \beta_0 + \beta_1 avgtrain + \beta_2 avgabil + u$$

Supposons que cette équation respecte les hypothèses de Gauss-Markov. Si des subventions sont accordées aux entreprises dont les employés ont des capacités moyennes plus faibles (de telle sorte que les variables *avgtrain* et *avgabil* sont corrélées négativement), quel sera le biais probable de $\tilde{\beta}_1$, sachant que $\tilde{\beta}_1$ correspond à l'estimation de la pente d'une régression simple de *avgprod* sur *avgtrain* ?

9. L'équation suivante décrit le prix médian de l'immobilier dans un quartier en fonction de deux variables : le degré de pollution dans l'atmosphère, mesuré par le protoxyde d'azote (*nox* pour « nitrous oxide » en anglais), et le nombre moyen de pièces dans les logements du quartier (*rooms*). Soit

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u$$

i. Quels sont les signes probables de β_1 et β_2 ? Quelle est l'interprétation de β_1 ? Expliquez.

ii. Pourquoi *nox* [ou plus précisément log(*nox*)] et *rooms* pourraient être négativement corrélés ? Si tel est le cas, la régression simple de log(*price*) sur log(*nox*) produirait-elle un estimateur de β_1 biaisé vers le haut ou vers le bas ?

iii. En utilisant le jeu de données HPRICE2, nous pouvons estimer les équations suivantes par les MCO :

$$\widehat{\log(price)} = 11{,}71 - 1{,}043 \log(nox),\ n = 506,\ R^2 = 0{,}264$$

$$\widehat{\log(price)} = 9{,}23 - 0{,}718 \log(nox) + 0{,}306 rooms.\ n = 506,\ R^2 = 0{,}514$$

Nous obtenons deux estimations de l'élasticité du prix (*price*) par rapport à la pollution (*nox*) ; la première provient d'une régression simple alors que la seconde résulte d'une régression multiple. Étant donné votre réponse à la question (ii), est-ce la différence que vous anticipiez ? Cela signifie-t-il que l'estimation $-0{,}718$ est plus proche de la vraie élasticité que l'estimation $-1{,}043$?

10. Vous désirez estimer la relation *ceteris paribus* entre y et x_1. Pour ce faire, vous obtenez des données sur deux variables de contrôle, x_2 et x_3. Pour être plus concret, on peut imaginer que y est la note finale obtenue à un examen ; x_1 est la présence en classe ; x_2 est la moyenne générale (GPA) obtenue au cours du semestre précédent ; et x_3 est la note obtenue à un examen standardisé au niveau national (comme le SAT ou l'ACT aux États-Unis). Soit $\tilde{\beta}_1$, l'estimation du coefficient de x_1 provenant de la régression simple de y sur x_1 ; soit $\hat{\beta}_1$, l'estimation obtenue à partir de la régression multiple de y sur x_1, x_2, x_3.

i. Si x_1 est fortement corrélée avec x_2 et x_3 au sein de l'échantillon et que x_2 et x_3 ont des effets *ceteris paribus* de grande ampleur sur y, quelle différence anticipez-vous entre $\tilde{\beta}_1$ et $\hat{\beta}_1$? Expliquez.

ii. Si x_1 n'est quasiment pas corrélée avec x_2 et x_3 mais que x_2 et x_3 sont extrêmement corrélées, les deux estimations de $\tilde{\beta}_1$ et $\hat{\beta}_1$ auront-elles tendance à être proches ou éloignées ? Expliquez.

iii. Si x_1 est fortement corrélée avec x_2 et x_3 et que les effets marginaux de x_2 et x_3 sur y sont faibles, pensez-vous que $\hat{\sigma}(\hat{\beta}_1)$ sera plus petit ou plus grands que $\hat{\sigma}(\tilde{\beta}_1)$? Expliquez.

iv. Si x_1 n'est quasiment pas corrélée avec x_2 et x_3, que x_2 et x_3 ont de grands effets marginaux sur y, et que x_2 et x_3 sont fortement corrélées, vous attendez-vous à ce que $\hat{\sigma}(\tilde{\beta}_1)$ soit plus petit que $\hat{\sigma}(\hat{\beta}_1)$, ou à l'inverse ? Expliquez.

11. Supposons que le modèle issu de la population soit

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

et que ce modèle respecte les hypothèses RLM.1 à RLM.4. En réalité, nous estimons le même modèle en omettant la variable x_3. Soit $\tilde{\beta}_0$, $\tilde{\beta}_1$, et $\tilde{\beta}_2$, les estimateurs des MCO de la régression de y sur x_1 et x_2. Étant donné les valeurs des variables indépendantes dans l'échantillon, montrez que la valeur attendue de $\tilde{\beta}_1$ est :

$$E(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r}_{i1} x_{i3}}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$$

où les $\hat{r}_{i1}$ sont les résidus de la régression par MCO de x_1 sur x_2. [*Astuce* : la formule des $\tilde{\beta}_1$ vient de l'équation (3.23). Utilisez $y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \beta_3 x_{i3} + u_i$ dans l'équation. Après un peu de calcul, prenez l'espérance en traitant x_{i3} et $\hat{r}_{i1}$ comme non aléatoires.]

12. L'équation suivante représente les effets sur l'emploi de la composition des recettes fiscales pour la population des districts aux États-Unis :

$$growth = \beta_0 + \beta_1 share_P + \beta_2 share_I + \beta_3 share_S + other\ factors$$

La variable *growth* désigne la variation en pourcentage de l'emploi entre 1980 et 1990. Les variables « *share...* » désignent la contribution de différents types de taxes aux recettes fiscales dans le district : $share_P$ est la proportion des recettes fiscales provenant des taxes foncières ; $share_I$ est la proportion des recettes fiscales imputable à l'impôt sur le revenu ; et $share_S$ est la proportion des recettes fiscales dérivant de la taxe sur la valeur ajoutée. Toutes ces variables sont mesurées en 1980. Il y a une catégorie de taxe qui est omise, $share_F$, qui contient les autres frais et taxes résiduelles. Par définition, la somme des quatre catégories est égale à un. La variable *other factors* inclut les dépenses du district liées à l'éducation, aux infrastructures, etc. (toutes mesurées en 1980).

i. Pourquoi doit-on omettre une des quatre catégories de recettes fiscales dans l'équation ?

ii. Interprétez le coefficient β_1 avec attention.

13. i. Considérons le modèle de régression simple, $y = \beta_0 + \beta_1 x + u$, qui respecte les quatre premières hypothèses de Gauss-Markov. Soit $g(x)$, une fonction de x. Par exemple, $g(x) = x^2$ ou $g(x) = \log(1 + x^2)$. Ecrivons $z_i = g(x_i)$ et définissons un estimateur de la pente comme suit :

$$\tilde{\beta}_1 = \frac{\left(\sum_{i=1}^{n}(z_i - \bar{z})y_i\right)}{\left(\sum_{i=1}^{n}(z_i - \bar{z})x_i\right)}$$

Montrez que $\tilde{\beta}_1$ est linéaire et sans biais. Comme $E(u|x) = 0$, vous pouvez traiter x_i et z_i comme étant non aléatoires dans vos calculs.

ii. Ajoutez l'hypothèse d'homoscédasticité. Montrez que :

$$var(\tilde{\beta}_1) = \frac{\sigma^2\left(\sum_{i=1}^{n}(z_i - \bar{z})^2\right)}{\left(\sum_{i=1}^{n}(z_i - \bar{z})x_i\right)}$$

iii. Sous les hypothèses de Gauss-Markov, montrez que $\text{Var}(\hat{\beta}_1) \leq \text{Var}(\tilde{\beta}_1)$, où $\hat{\beta}_1$ est l'estimateur des MCO. [*Astuce* : l'inégalité de Cauchy-Schwartz dans l'annexe B implique que :

$$\left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})\right)^2 \leq \left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^2\right)\left(n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

Remarquez que nous pouvons supprimer $\bar{x}$ de la covariance d'échantillon.]

14. Supposez que vous avez un échantillon de taille n de trois variables, y, x_1, et x_2, et que vous êtes avant tout intéressé par l'effet de x_1 sur y. Soit $\tilde{\beta}_1$ le coefficient de x_1 dans la régression simple et $\hat{\beta}_1$ le coefficient de x_1 dans la régression de y sur x_1 et x_2. Les écarts types estimés par n'importe quel logiciel d'économétrie sont :

$$\text{se}(\tilde{\beta}_1) = \frac{\tilde{\sigma}}{\sqrt{\text{SST}_1}}$$

$$\text{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\text{SST}_1}} \cdot \sqrt{\text{VIF}_1},$$

où $\tilde{\sigma}$ est l'écart type estimé de la régression multiple, $\text{VIF}_1 = 1/(1 - R_1^2)$, et R_1^2 est le R carré de la régression de x_1 sur x_2. Expliquez pourquoi $\text{se}(\hat{\beta}_1)$ peut être plus petit ou plus grand que $\text{se}(\tilde{\beta}_1)$.

15. Les équations suivantes utilisent les données du fichier MLB1, qui contiennent des informations sur les salaires des joueurs de la principale compétition de baseball américaine, la "major league". La variable dépendante, *lsalary*, est le log du salaire. Les deux variables explicatives sont le nombre d'années où les joueurs ont joué dans les major leagues (*years*) et les points produits (*rbisyr*, abréviation pour "runs batted in per year") :

$$\widehat{lsalary} = 12.373 + .1770years$$
$$(.098)\ (.0132)$$

$$n = 353,\ \text{SCR} = 326.196,\ \text{ETR} = .964,\ R^2 = .337$$

$$\widehat{lsalary} = 11.861 + .0904years + .0302rbisyr$$
$$(.084)\ (.0118)\qquad (.0020)$$

$$n = 353,\ \text{SCR} = 198.475,\ \text{ETR} = .753,\ R^2 = .597$$

i. Combien de degrés de liberté y a-t-il dans chaque régression ? Comment cela se fait-il que l'ETR de la deuxième régression est plus petit que celui de la première ?

ii. Le coefficient de corrélation d'échantillon entre *years* et *rbisyr* est d'environ 0.487. Cela vous semble-t-il intuitif ? Quel est le facteur d'inflation de la variance (il n'y en a qu'un) des coefficients de pente dans la régression multiple ? Diriez-vous de la colinéarité entre *years* et *rbisyr* qu'elle est faible, modérée, ou forte ?

iii. Comment expliquez-vous que l'écart type estimé du coefficient de *years* dans la régression multiple est plus faible que dans la régression simple ?

16. Les équations suivantes ont été estimées en utilisant les données du fichier LAWSCH85 :

$$\widehat{lsalary} = 9.90 - .0041rank + .294GPA$$
$$(.24)\ (.0003)\qquad (.069)$$
$$n = 142,\ R^2 = .8238$$

$$\widehat{lsalary} = 9.86 - .0038rank + .295GPA + .00017age$$
$$(.29)\ (.0004)\qquad (.083)\qquad (.00036)$$
$$n = 99,\ R^2 = .8036$$

Comment cela se fait-il que le R carré soit plus petit quand on ajoute la variable *âge* à l'équation ?

EXERCICES SUR ORDINATEUR

C1. Dans le domaine de la santé publique, un problème important pour les décideurs politiques est de déterminer les effets de la consommation de tabac par la mère durant la grossesse (*cigs*) sur la santé de son enfant. Le poids à la naissance (*bwght*) est une mesure de santé infantile. Un poids à la naissance trop faible peut augmenter le risque de contracter différentes maladies. Il existe naturellement d'autres facteurs qui affectent le poids du bébé à la naissance. Comme ces facteurs sont susceptibles d'être corrélés avec la consommation de cigarettes, nous devons les prendre en compte. Par exemple, un revenu familial (*faminc*) plus élevé facilite l'accès aux soins avant la naissance et assure également une meilleure alimentation à la mère. Considérons l'équation suivante, qui tient compte du revenu comme variable de contrôle :

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u$$

i. Quel est le signe le plus probable pour β_2 ?

ii. Pensez-vous que *cigs* et *faminc* sont susceptibles d'être corrélées ? Expliquez pourquoi la corrélation pourrait être positive *ou* négative.

iii. En recourant au jeu de données du fichier BWGHT, estimez l'équation par les MCO avec et sans *faminc*. Affichez les résultats sous la forme d'une équation, en précisant la taille de l'échantillon et le R carré. Analysez les résultats. L'ajout de *faminc* change-t-il de manière substantielle l'effet estimé de *cigs* sur *bwght* ?

C2. Utilisez les données du fichier HPRICE1 pour estimer le modèle

$$price = \beta_{0+}\beta_1 sqrft + \beta_2 bdrms + u$$

où *price* est le prix d'une maison (en milliers de dollars) ; *sqrft* mesure sa superficie (en pieds carrés ; un pied carré = 0,3 m^2) ; et *bdrms* représente le nombre de chambres.

i. Affichez les résultats sous la forme d'une équation.

ii. En maintenant la superficie constante (*ceteris paribus*), quelle est l'augmentation estimée du prix d'une maison disposant d'une pièce supplémentaire ?

iii. Quelle est l'augmentation estimée du prix d'une maison si vous décidez d'augmenter sa superficie de 140 pieds carrés en lui ajoutant une pièce supplémentaire ? Comparez votre réponse à celle obtenue en (ii).

iv. Quel pourcentage de la variation du prix est expliqué par la superficie et par le nombre de chambres ?

v. La première maison de l'échantillon dispose de 4 chambres (*bdrms* = 4) et d'une superficie de 2 438 pieds carrés (*sqrft* = 2 438). Estimez le prix de vente de cette maison à partir de la droite de régression des MCO.

vi. En réalité, cette maison s'est vendue à $300 000, soit son prix de vente effectif ou observé dans l'échantillon (*price* = 300). Calculez le résidu pour cette maison. Si votre modèle est fiable, cela suggère-t-il que l'acheteur a plutôt sous-payé ou surpayé la maison ?

C3. Le fichier CEOSAL2 contient un jeu de données portant sur 177 présidents-directeurs généraux (PDG) aux États-Unis. Ces données peuvent être utilisées pour examiner les effets de la performance des entreprises sur le salaire des PDG.

i. Estimez un modèle expliquant le salaire annuel du PDG par le chiffre d'affaires (*sales*) et la valeur de marché (*mktval*) de son entreprise. Faites en sorte que ce modèle soit un modèle à élasticité constante pour les deux variables indépendantes. Affichez les résultats sous la forme d'une équation.

ii. Ajoutez la variable *profits* au modèle utilisé en (i). Pourquoi ne doit-on pas inclure cette variable sous forme logarithmique ? Considérez-vous que ces variables de performance des entreprises expliquent l'essentiel de la variation des salaires des PDG ?

iii. Ajoutez l'expérience du PDG (*ceoten*) comme variable explicative au modèle utilisé en (ii). Toutes choses étant égales par ailleurs, quel est le « rendement » estimé d'une année supplémentaire d'expérience ? (Exprimez ce rendement en pourcentage).

iv. Calculez le coefficient de corrélation entre les variables log(*mktval*) et *profits* dans l'échantillon. Ces variables sont-elles fortement corrélées ? Qu'est-ce-que cela implique pour les estimateurs des MCO ?

C4. Utilisez les données du fichier ATTEND pour cet exercice.

i. Quel est le minimum, le maximum et la valeur moyenne des variables *atndrte* (taux de présence aux cours), *priGPA* (moyenne des examens à l'université), et *ACT* (test d'aptitude à l'entrée de l'université) ?

ii. Estimez le modèle

$$atndrte = \beta_{0+}\beta_1 priGPA + \beta_2 ACT + u$$

et affichez les résultats sous forme d'équation. Interprétez l'ordonnée à l'origine. A-t-elle un sens utile ?

iii. Interprétez les coefficients estimés de la pente. Est-ce surprenant ?

iv. Quelle est l'estimation de *atndrte* lorsque *priGPA* = 3,65 et *ACT* = 20 ? Comment interprétez-vous ce résultat ? Existe-t-il des étudiants dans l'échantillon, dont les résultats correspondent à ces valeurs ?

v. Si un étudiant (A) obtient un *priGPA* = 3,1 et un *ACT* = 21 alors qu'un autre (B) obtient un *priGPA* = 2,1 et un *ACT* = 26, quelle sera la différence prédite dans leur taux de présence en classe ?

C5. Dans le cadre de l'exemple 3.2, confirmez l'interprétation en termes d'effet *net* (ou *purgé*) que nous pouvons donner aux estimations des MCO. Pour ce faire, vous devez d'abord régresser *educ* sur *exper* et *tenure* pour sauvegarder les résidus, $\hat{r}_1$. Ensuite, vous devez régresser log(*wage*) sur $\hat{r}_1$. Enfin, il faut comparer les coefficients de $\hat{r}_1$ avec le coefficient de la variable *educ* dans la régression de log(*wage*) sur *educ*, *exper*, et *tenure*.

C6. Utilisez la base de données du fichier WAGE2 pour cet exercice. Comme d'habitude, vérifiez que toutes les régressions suivantes contiennent une ordonnée à l'origine.

i. Effectuez une régression simple de *IQ* sur *educ* pour obtenir le coefficient de la pente, soit $\tilde{\delta}_1$.

ii. Effectuez une régression simple de log(*wage*) sur *educ*, et obtenez le coefficient de la pente, soit $\tilde{\beta}_1$.

iii. Effectuez la régression multiple de log(*wage*) sur *educ* et *IQ*, et obtenez les coefficients de la pente, soit $\hat{\beta}_1$ et $\hat{\beta}_2$, respectivement.

iv. Vérifiez que $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$

C7. Utilisez les données du fichier MEAP93 pour répondre à cette question.

i. Estimez le modèle

$$math10 = \beta_0 + \beta_1 \log(expend) + \beta_2 lnchprg + u$$

et reportez les résultats de la manière habituelle, en précisant la taille de l'échantillon et le *R* carré. Les signes des coefficients de pente sont-ils ceux que vous anticipiez ? Expliquez.

ii. Que faites-vous de l'ordonnée à l'origine que vous avez estimée au point (i) ? Est-il raisonnable de choisir une valeur nulle pour ces deux variables explicatives ? [*Astuce* : souvenez-vous que log(1)=0.]

iii. Effectuez à présent la régression de *math10* sur log(*expend*) et comparez le coefficient de la pente avec l'estimation obtenue au point (i). L'estimation de l'effet des dépenses est-t-elle désormais plus grande ou plus petite que celle obtenue au point (i) ?

iv. Calculez la corrélation entre *lexpend* = log(*expend*) et *lnchprg*. Pouvez-vous justifier son signe ?

v. Utilisez votre réponse obtenue au point (iv) pour expliquer vos résultats du point (iii).

C8. Le fichier DISCRIM contient des données de prix pour différents produits vendus dans des enseignes de restauration rapide. Deux États américains sont concernés, le New Jersey et la Pennsylvanie, et les informations sont disponibles par « codes ZIP », ce qui correspond aux codes postaux en Europe. On y trouve aussi des informations sur les caractéristiques de la population vivant dans chacune des zones « ZIP ». La question est de savoir si les enseignes de restauration rapide pratiquent des prix plus élevés dans les endroits où il y a une plus grande proportion de personnes d'origine afro-américaine (*prpblck*), tout en prenant en compte le niveau des revenus (*income*).

i. Quelles sont les valeurs moyennes et les écarts-types des variables *prpblck* et *income* dans l'échantillon ? Quelles sont les unités de mesure de *prpblck* et *income* ?

ii. Considérons un modèle dans lequel le prix du soda, *psoda*, est une fonction de la proportion de la population d'origine afro-américaine et du revenu médian :

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + u$$

Estimez ce modèle par les MCO et reportez les résultats sous la forme d'une équation, en incluant la taille de l'échantillon et le *R* carré. (N'utilisez pas de notation scientifique en reportant les estimations). Interprétez le coefficient de *prpblck*. Est-il important sur le plan « économique » ?

iii. Comparez l'estimation obtenue au point (ii) avec celle de *psoda* sur *prpblck*. Lorsque vous tenez compte du revenu, l'effet « discrimination » est-il atténué ou accentué ?

iv. Il est possible qu'un modèle à élasticité constante du prix par rapport au revenu soit plus approprié. Reportez les estimations du modèle suivant :

$$\log(psoda) = \beta_0 + \beta_1 prpblack + \beta_2 \log(income) + u$$

Si *prpblck* augmente de 0,20 (20 points de pourcentage), quel sera le changement estimé de *psoda* (exprimé en pourcentage) ? (*Astuce* : la réponse est égale à 2,*xx*. À vous de trouver la valeur des « *xx*. »)

v. Ajoutez la proportion de personnes vivant sous le seuil de pauvreté (*prppov*) dans la régression utilisée au point précédent. Comment évolue le coefficient estimé de *prpblack* ?

vi. Calculez la corrélation entre log(*income*) et *prppov*. S'agit-il grosso modo d'une estimation que vous anticipiez ?

vii. Évaluez l'affirmation suivante : « log(*income*) et *prppov* sont des variables tellement corrélées que cela n'a pas de sens de les mettre dans la même équation ».

C9. Utilisez les données du fichier CHARITY pour répondre aux questions suivantes.

i. Estimez l'équation suivante par les MCO :

$$gift = \beta_0 + \beta_1 mailsyear + \beta_2 giftlast + \beta_3 propresp + u$$

Reportez les résultats de la manière habituelle, en incluant la taille de l'échantillon et le *R* carré. Quelle est la différence entre ce *R* carré et celui d'une régression simple excluant *giftlast* (le montant moyen du dernier don) *et propresp* (taux de réponse aux envois postaux) ? Pour rappel (voir l'exercice C7 du chapitre 2), *gift* correspond à la moyenne des dons réalisés sur l'année (en florin néerlandais) et *mailsyear* donne le nombre moyen de sollicitations envoyées par la poste sur l'année.

ii. Interprétez le coefficient de *mailsyear*. Est-il plus grand ou plus petit que le coefficient estimé à partir de la régression simple ?

iii. Interprétez le coefficient de *propresp*. Faites bien attention à l'unité de mesure de cette variable.

iv. Ajoutez à présent la variable *avggift* (montant moyen des dons effectués dans les années précédentes). Que devient l'effet estimé de *mailsyear* ?

v. Dans l'équation utilisée au point (iv), qu'est-il arrivé au coefficient de *giftlast* ? À votre avis, que se passe-t-il ?

C10. Utilisez les données du fichier HTV pour répondre aux questions suivantes. Cette base de données porte sur un échantillon de 1 230 employés masculins en 1991 ; elle inclut des informations sur leur salaire, leur niveau d'études, le niveau d'études de leurs parents, etc.

i. Quel est l'éventail des valeurs prises par *educ* dans l'échantillon ? Quel est le pourcentage d'employés qui ont terminé leurs études à la fin du lycée (ce qui correspond au « 12^{th} grade » aux États-Unis) ? En moyenne, les employés repris dans l'échantillon ont-ils un niveau d'études plus élevé que leurs parents ?

ii. Estimez le modèle de régression

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + u$$

par les MCO et reportez les résultats de la manière habituelle. Quelle variation de *educ* dans l'échantillon est expliquée par l'éducation des parents ? Interprétez le coefficient de *motheduc*.

iii. Ajoutez la variable *abil* (une mesure des capacités cognitives) dans la régression estimée en (ii) et reportez les résultats de la manière habituelle. La variable *abil* nous aide-t-elle à mieux expliquer les variations observées dans le nombre d'années d'études, même après avoir pris en compte l'éducation des parents ? Expliquez.

iv. Estimez une équation où *abil* apparaît sous une forme quadratique :

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u$$

En utilisant β_3 et β_4, utilisez un peu d'algèbre pour trouver la valeur de *abil*, soit *abil**, pour laquelle *educ* est minimisée (en considérant que le nombre d'années d'études des parents ne varie pas). Remarquez que la variable *abil* est mesurée de telle sorte que des valeurs négatives sont possibles. Vous pouvez également vérifier que la dérivée seconde est positive, démontrant l'existence d'un minimum.

v. Démontrez que seule une petite fraction d'employés dans l'échantillon ont une capacité cognitive inférieure à *abil**, correspondant à la valeur minimale de *abil*, calculée au point (iv). Pourquoi cela est-il important ?

vi. Utilisez les estimations obtenues en (iv) pour représenter sur un graphique la relation entre le nombre d'années d'études prédit et *abil*, en considérant que les valeurs de *motheduc* et *fatheduc* sont égales à leur moyenne respective dans l'échantillon, soit 12,18 et 12,45.

C11. Utilisez les données du fichier MEAPSINGLE pour étudier les résultats scolaires en maths des enfants de familles monoparentales. Ces données proviennent d'un sous-ensemble des écoles du Sud-Est du Michigan en 2000. Les variables socio-économiques de la base sont définies au niveau du code postal de l'école.

i. Faites la régression simple de *math4* sur *pctsgle* et reportez les résultats sous la forme habituelle. L'effet de la monoparentalité vous semble-t-il petit ou grand ?

ii. Ajoutez les variables *lmedinc* et *free* à l'équation. Que devient le coefficient de *pctsgle* ? Expliquez ce qu'il se passe.

iii. Calculez la corrélation d'échantillon entre *lmdinc* et *free*. A-t-elle le signe que vous attendiez ?

iv. L'ampleur de la corrélation entre *lmedinc* et *free* signifie-t-elle que vous devez supprimer une des deux variables pour mieux estimer l'effet causal de la monoparentalité sur les résultats des élèves ? Expliquez.

v. Trouvez les facteurs d'inflation de la variance (VIF) de chacune des variables explicatives apparaissant dans la régression du (iii). Pour quelle variable le VIF est-il le plus fort ? Cette information change-t-elle le modèle que vous voudriez utiliser pour analyser l'effet causal de la monoparentalité sur les résultats scolaires en mathématiques ?

C12. Les données du fichier ECONMATH contiennent les résultats obtenus par les étudiants d'une grande université publique américaine à différents examens : moyennes générales (GPA, pour grade point averages), résultat à un examen standardisé au niveau national (appelé ACT), note obtenue à un cours introductif à l'économie. La variable que nous voulons expliquer est *score*, la note finale obtenue à ce cours d'économie, exprimée en pourcentage.

i. Combien d'étudiants ont reçu la note maximale pour ce cours ? Quelle était la note moyenne ? Donnez les moyennes et les écarts types de *actmth* et *acteng*, et comparez-les.

ii. Estimez une équation linéaire expliquant *score* par *colgpa*, *actmth* et *acteng*, où *colgpa* est mesurée au début du semestre. Reportez les résultats sous la forme habituelle.

iii. Parmi les résultats obtenus à l'examen standardisé ACT, diriez-vous que ce sont les notes de maths ou les notes d'anglais qui prédisent le mieux les résultats obtenus dans le cours d'économie ? Interprétez.

iv. Discutez la taille du R carré dans la régression.

ANNEXE 3A

3A.1 Dérivation des conditions du premier ordre dans l'équation (3.13)

L'analyse est très similaire au cas de la régression simple. Nous devons caractériser les solutions au problème :

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_k x_{ik})^2$$

En prenant les dérivées partielles par rapport à chacun des b_j (voir annexe A), en les annulant, on obtient :

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$-2\sum_{i=1}^{n} x_{ij} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0, \text{ pour tout } j = 1, \ldots k$$

En laissant tomber le -2, on obtient les conditions du premier ordre données en (3.13).

3A.2 Dérivation de l'équation (3.22)

Pour obtenir (3.22), écrivez x_{i1} en fonction de ses valeurs ajustées et de ses résidus, tels qu'obtenus par la régression de x_1 sur $x_2, \ldots, x_k$: $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$, pour tout $i = 1, \ldots, n$. À présent, utilisez cette expression à l'intérieur de la seconde équation de (3.13) pour obtenir :

$$\sum_{i=1}^{n} (\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \qquad \textbf{[3.63]}$$

Étant donné la définition du résidu des MCO $\hat{u}_i$, comme $\hat{x}_{i1}$ est simplement une fonction linéaire des variables explicatives $x_{i2}, \ldots, x_{ik}$, il s'ensuit que $\sum_{i=1}^{n} \hat{x}_{i1} \hat{u}_i = 0$. Par conséquent, l'équation (3.63) peut être écrite comme suit :

$$\sum_{i=1}^{n} \hat{r}_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \qquad \textbf{[3.64]}$$

Comme les $\hat{r}_{i1}$ sont les résidus de la régression de x_1 sur $x_2, \ldots, x_k$, $\sum_{i=1}^{n} x_{ij} \hat{r}_{i1} = 0$, pour tout $j = 2, \ldots, k$. Par conséquent, (3.64) est équivalent à $\sum_{i=1}^{n} \hat{r}_{i1} (y_i - \hat{\beta}_1 x_{i1}) = 0$. Enfin, nous utilisons le fait que $\sum_{i=1}^{n} \hat{x}_{i1} \hat{r}_{i1} = 0$, ce qui signifie que $\hat{\beta}_1$ est tel que :

$$\sum_{i=1}^{n} \hat{r}_{i1} (y_i - \hat{\beta}_1 \hat{r}_{i1}) = 0$$

Avec un peu de calcul, on obtient (3.22), à condition, naturellement, que $\sum_{i=1}^{n} \hat{r}_{i1}^2 > 0$; ce que l'hypothèse RLM.3 garantit.

3A.3 Démonstration du théorème 3.1

Nous allons démontrer le théorème 3.1 pour $\hat{\beta}_1$; la démonstration pour les autres paramètres de la pente est quasi identique (voir l'annexe E pour une démonstration plus succincte, reposant sur la notation matricielle). Sous l'hypothèse RLM.3, les estimateurs des MCO existent et nous pouvons écrire $\hat{\beta}_1$ comme en (3.22). Sous l'hypothèse RLM.1, nous pouvons écrire y_i comme en (3.32) et substituer cette expression à y_i dans (3.22). Ensuite, en utilisant $\sum_{i=1}^{n} \hat{r}_{i1} = 0$, $\sum_{i=1}^{n} x_{ij} \hat{r}_{i1} = 0$ pour tout $j = 2, \ldots, k$, et $\sum_{i=1}^{n} x_{i1} \hat{r}_{i1} = \sum_{i=1}^{n} \hat{r}_{i1}^2$, nous obtenons

$$\hat{\beta}_1 = \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1} u_1 \right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right) \qquad \textbf{[3.65]}$$

À présent, sous les hypothèses RLM.2 et RLM.4, la valeur attendue de chaque u_i est nulle, étant donné toutes les variables indépendantes de l'échantillon. Comme les $\hat{r}_{i1}$ sont simplement des fonctions des variables indépendantes de l'échantillon, il s'ensuit que :

$$\begin{aligned} \mathrm{E}(\hat{\beta}_1|\mathbf{X}) &= \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1}\mathrm{E}(u_i|\mathbf{X})\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right) \\ &= \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1} \cdot 0\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right) = \beta_1 \end{aligned}$$

où X désigne les observations de toutes les variables indépendantes, $\mathrm{E}(\hat{\beta}_1|\mathbf{X})$ est la valeur de $\hat{\beta}_1$ étant donné x_{i1}, …, x_{ik} pour tout $i = 1, \ldots, n$. Cela clôt notre démonstration.

3A.4 Biais de variable omise – le cas général

Nous pouvons calculer le biais lié à l'omission d'une variable dans le modèle général de l'équation (3.31), sous les quatre premières hypothèses de Gauss-Markov. Soit les $\hat{\beta}_j$ ($j = 0, 1, \ldots, k$), caractérisant les estimateurs de la régression basée sur l'ensemble des variables explicatives. Soit les $\tilde{\beta}_j$ ($j = 0, 1, \ldots, k-1$), représentant les estimateurs des MCO de la régression dans laquelle la variable x_k a été omise. Soit $\tilde{\delta}_j$, $j = 1, \ldots, k-1$, le coefficient de la pente associé à x_j suite à la régression de x_{ik} sur x_{i1}, x_{i2}, … $x_{i,k-1}$, $i = 1, \ldots, n$. Une expression utile est donnée par :

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k\tilde{\delta}_j \qquad \textbf{[3.66]}$$

Quand nous ne prenons pas en compte x_k dans la régression, (3.66) montre explicitement que l'effet marginal estimé de x_j est égal à la somme de deux éléments : l'effet marginal de xj lorsque la variable xk est incluse ; le produit entre l'effet marginal de x_k sur $\hat{y}$ et celui de x_j sur x_k (pour $j \neq k$). Conditionnellement à l'ensemble de toutes les variables explicatives, $\mathbf{X}$, nous savons que tous les $\hat{\beta}_j$ seront des estimateurs sans biais des β_j correspondants, pour $j = 1, \ldots, k$. De plus, comme $\tilde{\delta}_j$ est simplement une fonction de $\mathbf{X}$, nous avons :

$$\begin{aligned} \mathrm{E}(\tilde{\beta}_j|\mathbf{X}) &= \mathrm{E}(\hat{\beta}_j|\mathbf{X}) + \mathrm{E}(\hat{\beta}_k|\mathbf{X})\tilde{\delta}_j \\ &= \beta_j + \beta_k\tilde{\delta}_j \end{aligned} \qquad \textbf{[3.67]}$$

L'équation (3.67) montre que $\tilde{\beta}_j$ est biaisé par rapport à β_j, à moins que β_k ne soit égal à zéro, auquel cas x_k n'a pas d'effet marginal dans la population, ou bien que $\tilde{\delta}_j$ soit égal à zéro, ce qui signifie que la corrélation partielle entre x_{ik} et x_{ij} dans l'échantillon est nulle. Pour obtenir l'équation (3.67), l'équation clé est (3.66). Pour arriver à l'équation (3.66), nous pouvons utiliser plusieurs fois l'équation (3.22). Pour plus de simplicité, nous nous concentrons sur le cas $j = 1$, si bien que $\tilde{\beta}_1$ est le coefficient de la pente dans la régression simple de y_i sur $\tilde{r}_{i1}$, $i = 1, \ldots, n$, où les $\tilde{r}_{i1}$ sont les résidus des MCO de la régression de x_{i1} sur x_{i2}, x_{i3}, …, $x_{i,k-1}$. Considérons le numérateur de l'expression de $\tilde{\beta}_1$: $\sum_{i=1}^{n} \tilde{r}_{i1}y_i$. Pour chaque i, nous pouvons écrire $y_i = \hat{\beta}_0 + \hat{\beta}_1x_{i1} + \ldots + \hat{\beta}_kx_{ik} + \hat{u}_i$; nous pouvons donc remplacer y_i par cette expression. Grâce aux propriétés des résidus obtenus par les MCO, les $\tilde{r}_{i1}$ ont une moyenne d'échantillon nulle et ne sont pas corrélés avec x_{i2}, x_{i3}, …, $x_{i,k-1}$ dans l'échantillon. De même, les $\hat{u}_i$ ont une moyenne d'échantillon nulle et ne sont pas corrélés avec x_{i1}, x_{i2}, …, x_{ik}. Il s'ensuit que $\tilde{r}_{i1}$ et $\hat{u}_i$ ne sont pas corrélés dans l'échantillon (puisque les $\tilde{r}_{i1}$ correspondent à des combinaisons linéaires de x_{i1}, x_{i2}, …, $x_{i,k-1}$). Ainsi,

$$\sum_{i=1}^{n} \tilde{r}_{i1}y_i = \hat{\beta}_1\left(\sum_{i=1}^{n} \tilde{r}_{i1}x_{i1}\right) + \hat{\beta}_k\left(\sum_{i=1}^{n} \tilde{r}_{i1}x_{ik}\right) \qquad \textbf{[3.68]}$$

À présent, $\sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} = \sum_{i=1}^{n} \tilde{r}_{i1}^2$; cette expression caractérise également le dénominateur de $\tilde{\beta}_1$. Par conséquent, nous avons montré que :

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_k \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} \right) \Big/ \left(\sum_{i=1}^{n} \tilde{r}_{i1}^2 \right)$$
$$= \hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1$$

Il s'agit bien de la relation que nous voulions démontrer.

3A.5 Démonstration du théorème 3.2

A nouveau, nous allons faire la démonstration pour $j = 1$. Nous écrivons $\hat{\beta}_1$ comme dans l'équation (3.65). À présent, sous RLM.5, $\text{Var}(u_i|\mathbf{X}) = \sigma^2$, pour tout $i = 1, \ldots, n$. En présence d'un échantillonnage aléatoire, les u_i sont indépendants, même conditionnellement à $\mathbf{X}$, et les $\hat{r}_{i1}$ ne sont pas aléatoires conditionnellement à $\mathbf{X}$. Par conséquent,

$$\text{Var}(\hat{\beta}_1|\mathbf{X}) = \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \, \text{Var}(u_i|\mathbf{X}) \right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)^2$$
$$= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \sigma^2 \right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)^2 = \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)$$

À présent, comme $\sum_{i=1}^{n} \hat{r}_{i1}^2$ représente la somme des carrés des résidus de la régression de x_1 sur $x_2, \ldots, x_k$, alors $\sum_{i=1}^{n} \hat{r}_{i1}^2 = SCT_1(1 - R_1^2)$, ce qui conclut notre démonstration.

3A.6 Démonstration du théorème 3.4

Nous allons montrer que, pour n'importe quel autre estimateur linéaire sans biais de β_1, soit $\tilde{\beta}_1$, $\text{Var}(\tilde{\beta}_1) \geq \text{Var}(\hat{\beta}_1)$, où $\hat{\beta}_1$ est l'estimateur des MCO. Nous posons ici $j = 1$ sans pour autant perdre en généralité.
Pour $\tilde{\beta}_1$ tel que décrit dans l'équation (3.60), nous pouvons réécrire y_i afin d'obtenir

$$\tilde{\beta}_1 = \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1} x_{i2} + \ldots + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik} + \sum_{i=1}^{n} w_{i1} x_i$$

À présent, comme les w_{i1} sont fonction des x_{ij}, on a :

$$\text{E}(\tilde{\beta}_1|\mathbf{X}) = \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1} x_{i2} + \ldots + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik} + \sum_{i=1}^{n} w_{i1} \text{E}(u_i|\mathbf{X})$$
$$= \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1} x_{i2} + \ldots + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik}$$

car $\text{E}(u_i|\mathbf{X}) = 0$, pour tout $i = 1, \ldots, n$, sous les hypothèses MLR.2 et MLR.4. Par conséquent, pour que $\text{E}(\tilde{\beta}_1|\mathbf{X})$ soit égal à β_1 pour toute valeur des paramètres, nous devons avoir :

$$\sum_{i=1}^{n} w_{i1} = 0, \sum_{i=1}^{n} w_{i1} x_{i1} = 1, \sum_{i=1}^{n} w_{i1} x_{ij} = 0, \; j = 2, \ldots, k. \qquad \textbf{[3.69]}$$

À présent, notons $\hat{r}_{i1}$, les résidus de la régression de x_{i1} sur $x_{i2}, \ldots, x_{ik}$. De (3.69), il découle que :

$$\sum_{i=1}^{n} w_{i1} r_{i1} = 1 \qquad \textbf{[3.70]}$$

car $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$ et $\sum_{i=1}^{n} w_{i1}\hat{x}_{i1} = 0$. À présent, considérons la différence entre $\text{Var}(\tilde{\beta}_1|X)$ et $\text{Var}(\hat{\beta}_1|X)$ sous RLM.1 à RLM.5 :

$$\sigma^2 \sum_{i=1}^{n} w_{i1}^2 - \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right) \qquad \textbf{[3.71]}$$

En ne prenant pas en compte σ^2 dans l'expression, nous pouvons utiliser (3.70) pour obtenir :

$$\sum_{i=1}^{n} w_{i1}^2 - \left(\sum_{i=1}^{n} w_{i1}\hat{r}_{i1} \right)^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right) \qquad \textbf{[3.72]}$$

Remarquons que (3.72) se simplifie en :

$$\sum_{i=1}^{n} (w_{i1} - \hat{\gamma}_1 \hat{r}_{i1})^2 \qquad \textbf{[3.73]}$$

où $\hat{\gamma}_1 = \left(\sum_{i=1}^{n} w_{i1}\hat{r}_{i1} \right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)$

On peut le vérifier en élevant les termes de (3.73) au carré, en les sommant, et en simplifiant. Comme (3.74) correspond tout simplement à la somme des carrés des résidus de la régression simple de w_{i1} sur $\hat{r}_{i1}$ (souvenez-vous que la moyenne d'échantillon de $\hat{r}_{i1}$ est zéro), l'expression (3.73) est forcément non négative, ce qui clôt la démonstration.

CHAPITRE 4

RÉGRESSION MULTIPLE : INFÉRENCE

Traduction de Jean-Yves Gnabo

Dans ce chapitre, nous poursuivons notre analyse des modèles de régression multiple. Nous nous intéressons maintenant au problème des tests d'hypothèses sur les paramètres de la population des modèles de régression. Nous commençons par nous intéresser à l'identification de la distribution des estimateurs par la méthode des MCO sous l'hypothèse additionnelle de normalité des erreurs dans la population. Les sections 4.2 et 4.3 traitent ensuite des tests d'hypothèses sur les paramètres individuels, alors que la section 4.4 aborde la question des tests d'hypothèses uniques relatives à plus d'un paramètre. La section 4.5 examine enfin plus avant la question des restrictions multiples en mettant l'accent sur le cas où l'on doit déterminer si un groupe de variables indépendantes peut être exclu du modèle.

4.1 DISTRIBUTIONS D'ÉCHANTILLONNAGE DES ESTIMATEURS DES MCO

Jusqu'à présent, nous avons formulé un ensemble d'hypothèses suivant lesquelles l'estimateur des MCO est sans biais, puis nous avons dérivé et commenté le biais généré par l'omission de variables explicatives. Dans la section 3.4, nous avons calculé la variance de l'estimateur des MCO sous les hypothèses de Gauss-Markov. Enfin, nous avons montré dans la section 3.5, que cette variance est la plus petite parmi la classe des estimateurs linéaires sans biais.

Connaître l'espérance et la variance de l'estimateur des MCO s'avère utile pour décrire sa précision. En revanche, la seule information sur les deux premiers moments des estimateurs $\hat{\beta}_j$ reste insuffisante pour procéder à de l'inférence statistique. Pour ce faire, nous avons besoin de connaître l'intégralité de la distribution d'échantillonnage des $\hat{\beta}_j$. Or, même sous les hypothèses de Gauss-Markov, la distribution des $\hat{\beta}_j$ peut potentiellement prendre n'importe quelle forme.

Lorsque nous travaillons de manière conditionnelle aux valeurs prises par les variables indépendantes dans notre échantillon, il apparaît clairement que les distributions d'échantillonnage des estimateurs des MCO des paramètres du modèle, dépendent de la distribution sous-jacente des erreurs. De façon à rendre facilement manipulable les distributions d'échantillonnage des $\hat{\beta}_j$, nous faisons maintenant l'hypothèse que l'erreur non observée de la population est *normalement distribuée*. Nous appelons cette hypothèse, l'**hypothèse de normalité des erreurs**.

HYPOTHÈSE RLM.6
(Normalité des erreurs)

L'erreur u dans la population, est *indépendante* des variables explicatives $x_1, x_2, \ldots, x_k$ et suit une distribution normale de moyenne nulle et de variance σ^2, soit : $u \sim \text{Normale}(0,\sigma^2)$

L'hypothèse RLM.6 est bien plus forte que l'ensemble des hypothèses que nous avons retenues jusqu'à présent. En effet, dans la mesure où u est indépendante des x_j sous RLM.6, il suit que : $E(u|x_1, \ldots, x_k) = E(u) = 0$ et $\text{Var}(u) = \sigma^2$. De ce fait, faire l'hypothèse RLM.6, implique nécessairement de supposer valides les hypothèses RLM.4 et RLM.5. Pour souligner que nous allons dans ce qui suit nous appuyer sur un ensemble d'hypothèses plus fortes que précédemment, nous nous référerons dorénavant aux hypothèses RLM.1 à RLM.6.

Pour les applications du modèle de régressions aux données en coupe transversale, les hypothèses allant de RLM.1 à RLM.6 sont appelées **hypothèses du modèle linéaire classique** (**MLC**). Dès lors, le modèle s'inscrivant dans le cadre défini par ces six hypothèses est dénommé **modèle linéaire classique**. Il est préférable de considérer les hypothèses MLC comme un ensemble d'hypothèses comprenant outre les hypothèses de Gauss-Markov, celle de normalité des erreurs.

Sous les hypothèses MLC, les estimateurs par les MCO des paramètres $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ ont de meilleurs propriétés que sous les seules hypothèses de Gauss-Markov. Ainsi, il est possible de montrer que les estimateurs des MCO sont dans ce cas les **estimateurs sans biais à variance minimale**, ce qui signifie qu'ils possèdent la plus petite variance parmi la classe d'estimateurs sans biais et non plus seulement des estimateurs linéaires sans biais. Cette propriété des estimateurs des MCO sous les hypothèses MLC est discutée plus en détails dans l'annexe E.

Il est possible de synthétiser les hypothèses MLC sur la population comme suit :

$$y|\mathbf{x} \sim \text{Normale}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \sigma^2).$$

où $\mathbf{x}$ désigne le vecteur $(x_1, \ldots, x_k)$. Ainsi, conditionnellement à $\mathbf{x}$, y suit une distribution normale avec pour moyenne une combinaison linéaire des $x_1, \ldots, x_k$ et une variance constante. Dans le cas d'une unique variable explicative x, la situation est illustrée par la figure 4.1.

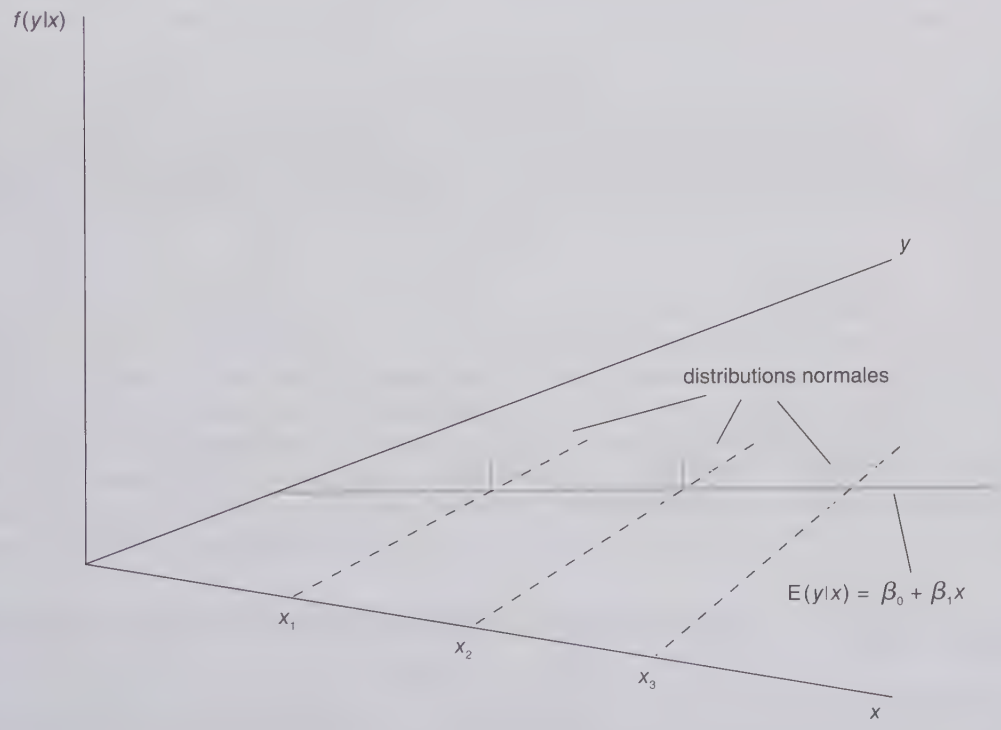


Figure 4.1 *La distribution normale homoscédastique avec une seule variable explicative.*

En règle générale, on justifie l'hypothèse de normalité des erreurs à l'aide de l'argument suivant : puisque l'erreur, u, est la somme de plusieurs facteurs non-observés qui affectent y, on peut recourir au théorème central limite (voir l'annexe C) pour conclure que u suit approximativement une distribution normale. Cet argument présente certains avantages. Il souffre cependant également de quelques faiblesses. Tout d'abord, les différents facteurs regroupés dans le terme d'erreur, u peuvent avoir des distributions très différentes dans la population (par exemple, les facteurs d'aptitude et de compétence scolaires se retrouvant dans le terme d'erreur d'une équation de salaire). Bien que cela n'invalide pas le théorème central limite (TCL), l'approximation de la distribution par une distribution normale peut être relativement mauvaise selon le nombre de facteurs inclus dans u et de l'importance des différences dans leurs distributions.

Un problème plus sérieux avec cet argument tient au fait qu'il repose sur l'hypothèse que l'ensemble des facteurs inobservés composant le terme d'erreur, affectent y de façon séparée et exclusivement additive. Or, rien ne le garantit. Si u est une combinaison complexe de ces différents facteurs inobservés, l'argument du TCL ne tient plus.

Dans toute application, le fait de savoir s'il convient de faire l'hypothèse de normalité de u demeure une question empirique. Par exemple, il n'existe aucun théorème démontrant que la variable *wage* conditionnellement aux valeurs prises par les variables *educ*, *exper*, et *tenure* soit normalement distribuée. D'ailleurs, on aurait plutôt tendance à penser le contraire puisque les salaires n'étant jamais négatifs, on ne peut pas s'attendre à ce que la variable suive *stricto sensu* une distribution normale. En outre, une fraction importante de la population des pays industrialisés gagne exactement le salaire minimum en raison des lois en vigueur, ce qui est également en contradiction avec l'hypothèse de normalité. Pour autant, en pratique, la question qui se pose est surtout de savoir si la distribution conditionnelle des salaires est relativement « proche » d'une distribution normale. Dans le cas d'espèce, les études empiriques passées nous enseignent que la normalité n'est *pas* une hypothèse pertinente pour les salaires.

Souvent, il est possible de mieux se conformer à l'hypothèse de normalité en utilisant des transformations des variables initiales et en particulier, la transformation logarithmique. Par exemple, une transformation du type, log(*price*), permet généralement d'obtenir des séries dont la distribution se rapproche plus d'une distribution normale que celle de la variable brute *price*. De nouveau, il s'agit ici d'une question empirique. Nous discuterons des conséquences de la non normalité des termes d'erreur pour l'inférence statistique dans le chapitre 5.

Dans un certain nombre de cas, l'hypothèse RLM.6 doit clairement être réfutée. Lorsque la variable dépendante, y, par exemple prend seulement un nombre limité de valeurs, sa distribution ne peut se rapprocher de celle d'une loi normale. La variable dépendante de l'exemple 3.5 en offre une bonne illustration. La variable, *narr86* récence le nombre de fois qu'un homme jeune a fait l'objet d'une arrestation en 1986. Elle prend naturellement un nombre limité de valeurs entières, et est égale à zéro pour la plupart des hommes de l'échantillon. Que convient-il de faire dans ce cas ? Comme nous le verrons dans le chapitre 5 – et ceci est particulièrement important – les conséquences de la non normalité des erreurs doivent être relativisées en présence de grand échantillon. Pour l'instant, nous supposerons simplement que l'hypothèse de normalité des erreurs est vérifiée.

La normalité des erreurs implique également la normalité de la distribution d'échantillonnage de l'estimateur des MCO :

THÉORÈME 4.1
Distributions d'échantillonnage normales

Sous les hypothèses MLC définies par les hypothèses RLM.1 à RLM.6, conditionnellement aux valeurs prises dans l'échantillon par les variables indépendantes du modèle, on a :

$$\hat{\beta}_j \sim \text{Normale}\,[\beta_j\,,\ \text{Var}\,(\hat{\beta}_j)], \qquad \textbf{[4.1]}$$

avec $\text{Var}(\hat{\beta}_j)$ défini dans le chapitre 3 [équation (3.51)]. Dès lors,

$$(\hat{\beta}_j - \beta_j)/\sigma(\hat{\beta}_j) \sim \text{Normale}(0,1).$$

La démonstration de (4.1) n'est pas très difficile et s'appuie principalement sur les propriétés des variables aléatoires normalement distribuées rappelées en annexe B. Chaque $\hat{\beta}_j$ peut s'écrire comme $\hat{\beta}_j = \beta_j + \sum_{i=1}^{n} w_{ij}u_i$, avec $w_{ij} = \hat{r}_{ij}\, SCR_j$, $\hat{r}_{ij}$ le $i^{ème}$ résidu issu de la régression des x_j sur toutes les autres variables explicatives, et SCR_j la somme des carrés des résidus de cette régression [voir l'équation (3.62)]. Puisque les w_{ij} ne dépendent que des variables indépendantes, ils peuvent être assimilés à des éléments non aléatoires. Dès lors, $\hat{\beta}_j$ se comprend comme une combinaison linéaire des erreurs dans l'échantillon, $\{u_i : i = 1, 2, \ldots, n\}$. Sous l'hypothèse RLM.6 (et celle d'échantillonnage aléatoire décrite dans RLM.2), les erreurs sont indépendantes et identiquement distribuées selon une loi Normale $(0, \sigma^2)$. Une propriété importante des variables indépendantes normalement distribuées tient au fait qu'une combinaison linéaire d'entre elles est elle aussi normalement distribuée (voir annexe B). Cette propriété nous permet de clore la démonstration. Dans la section 3.3 du chapitre 3, nous avions montré que $E(\hat{\beta}_j) = \beta_j$ et dérivé l'expression de la $Var(\hat{\beta})$ dans la section 3.4 ; il n'est donc pas utile de revenir ici sur ces résultats.

Pour aller plus loin 4.1

Supposons que u soit indépendant des variables explicatives, et prenne les valeurs $-2, -1, 0, 1$, avec pour chacune d'entre elles une probabilité 1/5. Les hypothèses de Gauss-Markov sont-elles violées ? Qu'en est-il des hypothèses MLC ?

La seconde partie du théorème découle directement du fait que la standardisation d'une variable aléatoire en retranchant sa moyenne et en la divisant par son écart-type, permet d'obtenir une variable aléatoire normale centrée réduite.

Les conclusions du théorème 4.1 peuvent être étendues. En complément de (4.1), toute combinaison linéaire de $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ est également normalement distribuée, et tout sous-ensemble de $\hat{\beta}_j$ est caractérisé par une distribution normale *jointe*. Ces éléments sont au cœur des résultats relatifs aux procédures de tests présentés dans l'introduction de ce chapitre. Par la suite, dans le chapitre 5, nous montrons que la normalité des estimateurs des MCO est toujours une *approximation* correcte en présence de grands échantillons même en l'absence de normalité des erreurs.

4.2 TESTS D'HYPOTHÈSES SUR UN UNIQUE PARAMÈTRE DE LA POPULATION : LE TEST DE STUDENT

Cette section traite de la question primordiale en économétrie des tests d'hypothèses sur un unique paramètre de la population des modèles de régression. Le modèle de la population peut être écrit comme suit :

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \qquad \textbf{[4.2]}$$

On suppose que ce modèle satisfait les hypothèses MLC. Nous savons que la procédure des MCO permet d'obtenir des estimateurs sans biais des β_j. Dans cette section, nous étudions comment procéder à des tests d'hypothèses relativement à certains β_j spécifiques. Afin de bien comprendre la mise en œuvre des tests d'hypothèses, il est nécessaire de garder à l'esprit que les β_j sont des caractéristiques inconnues de la population, nous n'aurons donc jamais la possibilité de les connaître avec certitude. Pour autant, nous pouvons émettre des *hypothèses* sur la valeur des β_j puis utiliser les méthodes de l'inférence statistique dans le but de vérifier la vraisemblance ces hypothèses.

Pour construire les tests d'hypothèses, nous avons besoin de recourir au résultat suivant :

THÉORÈME 4.2
Distribution de Student pour les estimateurs standardisés

Sous les hypothèses MLC allant de RLM.1 à RLM.6, on a :

$$(\hat{\beta}_j - \beta_j)/\hat{\sigma}(\hat{\beta}_j) \sim t_{n-k-1} = t_{ddl} \qquad \textbf{[4.3]}$$

avec $k + 1$ le nombre de paramètres inconnus dans le modèle de population $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$ (k paramètres de pente et une constante β_0) et $n - k - 1$ le nombre de degrés de liberté (*ddl*).

Ce résultat diffère du théorème 4.1 par un certain nombre d'aspects fondamentaux. Le théorème 4.1 que nous avons démontré précédemment, stipulait que sous les hypothèses MLC, $(\hat{\beta}_j - \beta_j)/\sigma(\hat{\beta}_j) \sim$ Normale (0,1). La distribution de Student dans (4.3) vient du fait que le paramètre σ dans l'expression de $\sigma(\hat{\beta}_j)$ a été remplacé par la variable aléatoire $\hat{\sigma}$. La preuve mathématique que cette quantité suit une distribution de Student à $n - k - 1$ degrés de liberté est difficile et ne revêt qu'un intérêt limité. Elle repose pour l'essentiel, sur le fait que (4.3) peut être exprimée sous la forme d'un ratio de variables aléatoires normales centrées réduites $(\hat{\beta}_j - \beta_j)/\sigma(\hat{\beta}_j)$ sur la racine carrée de $\hat{\sigma}^2/\sigma^2$. Il est possible de montrer que ces variables aléatoires sont indépendantes et que $(n - k - l)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}$. Le résultat final découle alors de la définition d'une variable aléatoire de Student (voir section B.5).

Le théorème 4.2 est un résultat important puisque il nous permet de réaliser des tests d'hypothèses impliquant le paramètre β_j. Dans la plupart des applications, notre intérêt premier réside dans la possibilité de tester l'**hypothèse nulle** :

$$H_0 : \beta_j = 0 \qquad \textbf{[4.4]}$$

où j correspond à l'une des k variables indépendantes. Il est essentiel de bien comprendre la signification de l'équation (4.4) et de pouvoir décrire cette hypothèse avec des mots simples dans le cadre d'une application donnée. Puisque β_j mesure l'effet marginal de x_j sur (la valeur attendue de) y, après avoir pris en compte l'influence des autres variables indépendantes, (4.4) signifie que, une fois que l'influence des x_1, x_2, …, x_{j-1}, x_{j+1}, …, x_k a été prise en compte, x_j n'a *pas d'effet* sur la valeur espérée de y. Nous pouvons alors exprimer l'hypothèse nulle comme suit « x_j n'a pas d'effet marginal sur y » puisque cela est vrai pour toute valeur de β_j autre que zéro. Les procédures des tests classiques sont indiquées pour tester des *hypothèses simples* comme (4.4).

À titre d'exemple, considérons l'équation de salaire suivante :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

L'hypothèse nulle $H_0 : \beta_2 = 0$ signifie que, une fois l'impact du niveau d'éducation et la titularisation pris en compte, le nombre d'années d'expérience (*exper*) n'a plus aucun effet sur le salaire horaire. C'est une hypothèse pertinente du point de vue économique. Si elle est vraie, elle implique que les années d'expérience antérieures à l'occupation professionnelle courante d'un individu n'ont aucun impact sur son salaire. Si en revanche, $\beta_2 > 0$ alors, l'expérience passée contribue à la productivité et par conséquent, au salaire.

Vous vous souvenez certainement des rudiments de vos cours de statistiques relatifs aux tests d'hypothèses appliqués à la moyenne d'une population. (Voir l'annexe C pour un rappel.) Les mécanismes à l'œuvre en (4.4) dans le contexte de la régression multiple sont très similaires. La principale difficulté réside dans l'obtention des coefficients estimés, des écarts-types estimés ainsi que des valeurs critiques, ces informations

étant en général fournies par les logiciels économétriques. Notre tâche ici est de comprendre comment les sorties d'estimation peuvent être utilisées pour tester les hypothèses d'intérêt.

La statistique que nous utilisons pour tester (4.4) (contre n'importe quelle hypothèse alternative) est appelée « la » **statistique de Student** ou « la » **statistique** *t* de $\hat{\beta}_j$ et se définit comme suit :

$$t_{\hat{\beta}_j} = \hat{\beta}_j / \hat{\sigma}(\hat{\beta}_j) \qquad \textbf{[4.5]}$$

Nous avons mis l'article « la » entre guillemets car, comme nous le verrons par la suite, il est nécessaire d'utiliser une forme plus générale de la statistique *t* pour tester d'autres hypothèses relatives à β_j. Pour le moment, gardons à l'esprit que l'expression (4.5) convient uniquement pour tester l'hypothèse (4.4). En pratique, il est souvent utile d'indicer la statistique *t* en y associant le nom de la variable indépendante ; par exemple, t_{educ} serait la statistique *t* de $\hat{\beta}_{educ}$.

La statistique *t* de $\hat{\beta}_j$ se calcule simplement à partir de la valeur estimée $\hat{\beta}_j$ et de son écart-type estimé. En pratique, cette statistique est reportée par défaut dans la plupart des sorties de logiciels d'économétrie en parallèle des paramètres estimés et des écarts-types estimés.

Avant de présenter plus avant comment utiliser l'expression (4.5) pour tester formellement $H_0 : \beta_j = 0$ il est utile de mentionner les propriétés qui font de $t_{\hat{\beta}_j}$ une statistique pertinente pour détecter si $\beta_j \neq 0$. Tout d'abord, puisque $\hat{\sigma}(\hat{\beta}_j)$ est toujours positif, $t_{\hat{\beta}_j}$ possède toujours le même signe que $\hat{\beta}_j$: si $\hat{\beta}_j$ est positif, alors $t_{\hat{\beta}_j}$ l'est également et inversement. De plus, pour une valeur donnée de $\hat{\sigma}(\hat{\beta}_j)$, plus la valeur de $\hat{\beta}_j$ est grande, plus la valeur de la statistique $t_{\hat{\beta}_j}$ l'est également. De même si $\hat{\beta}_j$ devient plus négative, il en va de même pour $t_{\hat{\beta}_j}$.

Puisque nous cherchons à tester $H_0 : \beta_j = 0$, il semble naturel d'étudier le comportement de notre estimateur sans biais de β_j, soit $\hat{\beta}_j$. En pratique, l'estimation ponctuelle de $\hat{\beta}_j$ ne sera *jamais* exactement zéro, qu'H_0 soit vraie ou non. La question qui se pose est donc de savoir si $\hat{\beta}_j$ est suffisamment éloigné de zéro pour rejeter cette hypothèse au sens statistique du terme. Une valeur estimée de $\hat{\beta}_j$ très éloignée de zéro sera plutôt en faveur du rejet de l'hypothèse $H_0 : \beta_j = 0$. Toutefois, il est également possible que notre estimation $\hat{\beta}_j$ soit sujette à une erreur d'échantillonnage, il convient donc d'en tenir compte en pondérant $\hat{\beta}_j$ par cette dernière. Puisque la statistique calculée $t_{\hat{\beta}_j}$ reporte au numérateur la valeur du paramètre estimé et au dénominateur, celle de l'écart-type estimé associé, elle mesure à combien d'écart-type estimé $\hat{\beta}_j$ se trouve éloigné de zéro. La démarche est exactement similaire à celle que nous adoptons en statistiques lorsque nous nous intéressons à la moyenne d'une population et testons son égalité à zéro à partir de statistiques *t* standards. Les valeurs de $t_{\hat{\beta}_j}$ qui sont suffisamment éloignées de zéro entraîneront quant à elles le rejet de l'hypothèse H_0. La règle de décision précise dépend de la formulation de l'hypothèse alternative et du seuil de significativité choisi pour le test.

Déterminer une règle de décision pour l'hypothèse (4.4) au seuil de significativité donné – c'est-à-dire la probabilité de rejeter H_0 lorsqu'elle est vraie – requiert la connaissance de la distribution d'échantillonnage de $t_{\hat{\beta}_j}$ sous l'hypothèse que H_0 est vraie. D'après le théorème 4.2, nous savons que cette statistique suit une distribution t_{n-k-1}. C'est là le résultat clé nécessaire à la mise en œuvre de nos procédures de tests. (4.4).

Avant d'aller plus loin, il est important de se souvenir que nous testons des hypothèses relatives aux paramètres de la *population*. Nous ne testons *pas* des hypothèses relatives à des estimations particulières pour un échantillon donné. Dès lors, cela n'a aucun sens de formuler une hypothèse telle que « $H_0 : \hat{\beta}_1 = 0$ » ou pire encore, comme « $H_0 : 0{,}237 = 0$ » dans le cas où le paramètre estimé à partir des données de l'échantillon est de 0,237. Ce que nous testons, c'est bien la possibilité que la valeur inconnue du paramètre au niveau de la population soit égale à zéro.

Il arrive que des logiciels renvoient en guise de statistique t la *valeur absolue* de (4.5), de sorte que la statistique soit toujours positive. Cette transformation présente le désavantage de rendre plus ardue la mise en œuvre d'un test d'hypothèse unilatéral. Dans cet ouvrage, la statistique t a toujours le même signe que celui du paramètre associé estimé par les MCO.

Test d'hypothèse unilatéral

Afin de déterminer une règle de décision pour H_0, nous devons identifier au préalable l'**hypothèse alternative**. Nous considérons tout d'abord une **alternative unilatérale** de la forme :

$$H_1 : \beta_j > 0 \qquad \textbf{[4.6]}$$

Lorsque nous posons l'hypothèse alternative comme dans l'équation (4.6), nous considérons que l'hypothèse nulle est donnée par : $H_0 : \beta_j \leq 0$. Par exemple, si β_j est le coefficient associé à la variable *educ*, capturant le nombre d'années d'études, dans l'équation de salaire, nous cherchons à détecter si β_j est différent de zéro lorsque la valeur de β_j est positive. Comme vous le savez sans doute depuis vos cours de statistiques, l'hypothèse la plus difficile à rejeter en faveur de celle décrite en (4.6) est l'hypothèse $\beta_j = 0$. En d'autres termes, si nous sommes en mesure de rejeter l'hypothèse nulle $\beta_j = 0$ alors nous rejetterons aussi automatiquement $\beta_j < 0$. Dès lors, il suffit de procéder comme si nous testions $H_0 : \beta_j = 0$ contre $H_1 : \beta_j > 0$ en ignorant $\beta_j < 0$. C'est l'approche qui sera adoptée dans cet ouvrage.

Comment déterminer la règle de décision associée à ce test ? Nous devons d'abord nous décider sur le **seuil de significativité** (ou tout simplement « seuil » du test, pour faire court) [qui correspond au risque de première espèce], soit la probabilité de rejeter H_0 alors qu'elle est vraie. Concrètement, à supposer que l'on ait décidé de fixer un seuil de significativité de 5 %, comme c'est le cas dans la plupart des applications, cela signifie que l'on ne souhaite pas rejeter à tort l'hypothèse H_0 lorsqu'elle est vraie plus de 5 % du temps. Alors que $t_{\hat{\beta}_j}$ suit une distribution de Student sous H_0 – et est donc de moyenne nulle – sous l'alternative, $\beta_j > 0$, la valeur attendue de $t_{\hat{\beta}_j}$ est positive. Nous sommes donc à la recherche de valeurs positives de $t_{\hat{\beta}_j}$ qui soient « suffisamment grandes » de façon à rejeter $H_0 : \beta_j = 0$ en faveur de $H_1 : \beta_j > 0$. Des valeurs négatives de $t_{\hat{\beta}_j}$ n'apportent aucune preuve en faveur de l'hypothèse H_1.

L'expression « suffisamment grande » pour un seuil de significativité de 5 % doit se comprendre comme le 95e centile associé à la distribution de Student t à $n - k - 1$ degrés de liberté ; que l'on nomme ici c. En d'autres termes, la **règle de décision** est telle que H_0 est rejetée en faveur de H_1 au seuil de significativité de 5 % si :

$$t_{\hat{\beta}_j} > c \qquad \textbf{[4.7]}$$

Du fait de notre choix de la **valeur critique** c, le rejet de H_0 sera effectif pour 5 % de notre échantillon aléatoire lorsque H_0 est vraie.

La règle de décision stipulée en (4.7) est un exemple de **test unilatéral**. Pour obtenir c, nous n'avons besoin que du niveau de significativité et du nombre de degrés de liberté. Par exemple, pour un test au seuil de 5 % avec $n - k - 1 = 28$ degrés de liberté, la valeur critique est donnée par $c = 1{,}701$. Si $t_{\hat{\beta}_j} \leq 1{,}701$, nous ne pouvons rejeter l'hypothèse H_0 en faveur de (4.6) au seuil de 5 %. À noter qu'une valeur négative pour $t_{\hat{\beta}_j}$ quelque fût son amplitude en valeur absolue, eût mené au rejet de l'hypothèse H_0 en faveur de 4.6. (Voir figure 4.2.)

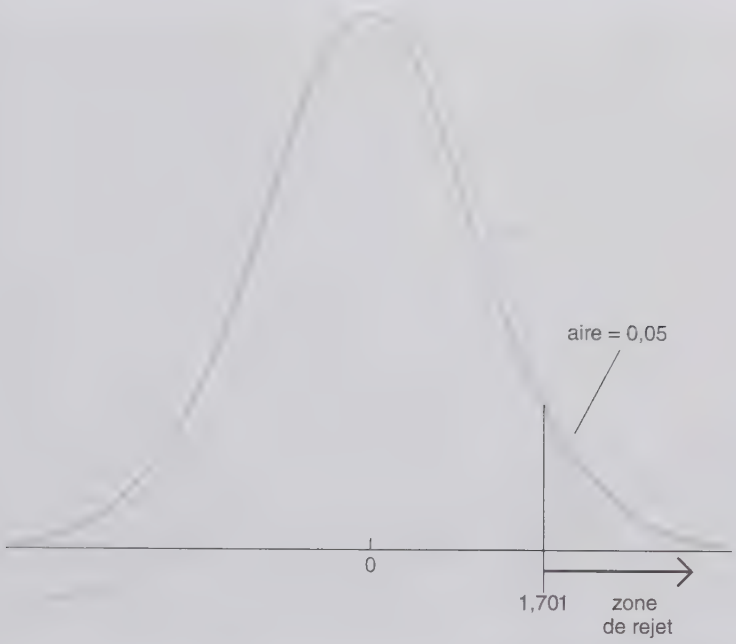


Figure 4.2 *Règle de décision au seuil de 5 % contre une hypothèse alternative $H_1 : \beta_j > 0$ avec 28 ddl.*

Une procédure similaire peut être mise en œuvre pour d'autres seuils de significativité. Au seuil de 10 % et lorsque *ddl* = 21, la valeur critique $c = 1{,}323$. Au seuil de 1 % pour le même nombre de degrés de liberté, $c = 2{,}518$. Toutes ces valeurs critiques sont obtenues par lecture directe de la table statistique G.2. Nous pouvons observer une tendance dans les valeurs prises par la valeur critique : à mesure que le seuil de significativité diminue, la valeur critique augmente, nécessitant alors une valeur d'autant plus élevée de t_{β_j} pour rejeter l'hypothèse H_0. De ce fait, si H_0 est rejetée au seuil de 5 %, elle le serait automatiquement au seuil de 10 %. En effet, cela n'a pas de sens de rejeter l'hypothèse nulle à 5 % et de devoir pratiquer à nouveau le test pour déterminer l'issue d'un test au seuil de 10 %.

À mesure que les degrés de liberté de la distribution de Student deviennent grands, celle-ci converge vers une distribution normale centrée réduite. Par exemple, lorsque $n - k - 1 = 120$, la valeur critique au seuil de 5 % pour l'alternative unilatérale mentionnée en (4.7) est de 1,658, à comparer avec la valeur critique au même seuil d'une normale centrée réduite, de 1,645. Ces valeurs sont suffisamment proches pour pouvoir considérer en pratique qu'au delà de 120 degrés de liberté, il est acceptable d'utiliser les valeurs critiques d'une loi normale.

EXEMPLE 4.1
Équation du salaire horaire

L'estimation d'une équation de salaire horaire sur les données contenues dans WAGE1 donne les résultats d'estimation suivants :

$$\widehat{\log(wage)} = 0{,}284 + 0{,}092\,educ + 0{,}0041\,exper + 0{,}022\,tenure$$

$$(0{,}104)\quad (0{,}007)\qquad (0{,}0017)\qquad (0{,}003)$$

$$n = 526,\ R^2 = 0{,}316,$$

où les écarts-types estimés sont mentionnés entre parenthèses, sous les coefficients estimés. Nous suivrons cette convention d'écriture tout au long de l'ouvrage. Cette équation peut être utilisée pour tester si les rendements de *exper*, en tenant compte de l'influence d'*educ* et *tenure*, sont de valeur nulle pour la population, contre l'alternative qu'ils soient positifs. En posant $H_0 : \beta_{exper} = 0$ contre $H_1 : \beta_{exper} > 0$ (En pratique indicer le paramètre par le nom de la variable à laquelle il se rapporte est un moyen commode et pratique de labelliser les paramètres, le recours aux indices numériques pouvant prêter à confusion.) Rappelez vous que β_{exper} fait référence à la vraie valeur du paramètre pour la population. Il serait donc absurde d'écrire « $H_0 : 0{,}0041 = 0$ » ou « $H_0 : \hat{\beta}_{exper} = 0$ ».

Le nombre de degrés de liberté étant de 522, nous pouvons utiliser les valeurs critiques tabulées pour une distribution normale. Au seuil de 5 % celle-ci s'élève à 1,645, à celui de 1 % à 2,326. La statistique t pour $\hat{\beta}_{exper}$ est donnée par :

$$t_{exper} = 0{,}0041/0{,}0017 \approx 2{,}41,$$

de ce fait $\hat{\beta}_{exper}$, ou *exper*, apparaît statistiquement significatif, même au seuil de 1 %. Nous pouvons également conclure que « $\hat{\beta}_{exper}$ est statistiquement plus grand que zéro au seuil de 1 %. »

Les rendements estimés d'une année additionnelle d'expérience professionnelle, toutes choses égales par ailleurs, n'apparaissent pas particulièrement élevés. Par exemple, ajouter trois années d'expérience augmente le log des salaires, log(*wage*), de 3(0,0041) = 0,0123, impliquant une hausse de salaire de seulement 1,2 %. Pour autant, nous avons montré de façon convaincante que l'effet marginal de l'expérience *est* positif à l'échelle de la population.

Il arrive qu'à l'inverse du cas précédent, l'on soit amené à considérer l'hypothèse alternative que le paramètre est inférieur à zéro, ce qui s'écrit :

$$H_1 : \beta_j < 0 \qquad \textbf{[4.8]}$$

La règle de décision compte-tenu de l'alternative décrite en (4.8) est simplement le miroir du cas précédent. Ainsi, la valeur critique sera localisée à l'extrémité gauche de la distribution de la statistique t. En pratique, il est plus aisé d'appréhender la règle de décision comme suit :

$$t_{\hat{\beta}_j} < -c \qquad \textbf{[4.9]}$$

avec c la valeur critique pour l'hypothèse alternative $H_1 : \beta_j > 0$. Par soucis de simplicité, nous ferons toujours l'hypothèse que la valeur seuil c est positive, puisque c'est la manière dont elles sont reportées dans les tables de Student, les valeurs critiques négatives étant données par $-c$.

Par exemple, si l'on considère un seuil de significativité de 5 % et 18 degrés de liberté, alors $c = 1{,}734$, il suit que $H_0 : \beta_j = 0$ est rejetée en faveur de $H_1 : \beta_j < 0$ au seuil de 5 % si $t_{\hat{\beta}_j} < -1{,}734$. Il est important de garder à l'esprit que pour rejeter l'hypothèse H_0 contre l'alternative négative stipulée en (4.8), la valeur de la statistique t doit être négative. Une valeur positive quelque fût son amplitude, ne produit en aucune façon une preuve en faveur de (4.8). La règle de décision est illustrée à la figure 4.3.

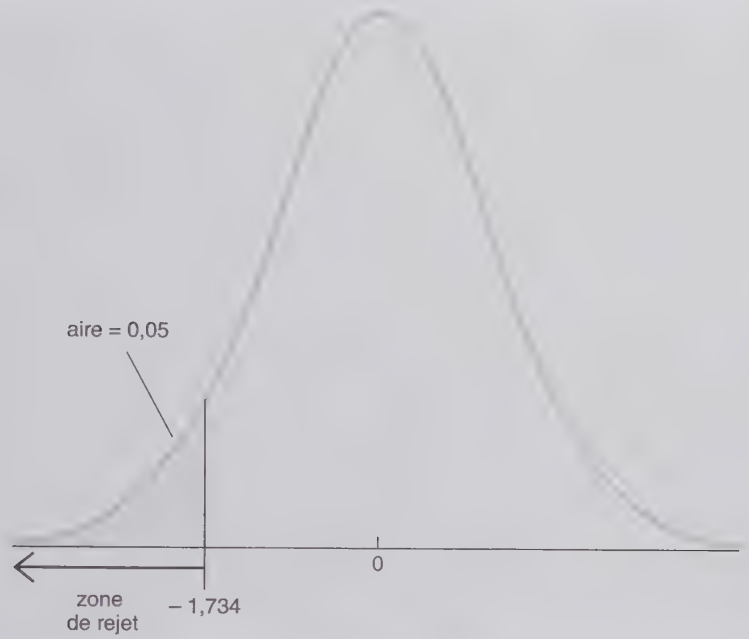


Figure 4.3 *Règle de décision au seuil de 5 % contre une hypothèse alternative avec* $H_1 : \beta_j < 0$ *18 ddl.*

Pour aller plus loin 4.2

Supposons que le taux d'approbation de prêts communautaires soit déterminé par l'équation suivante :

$$apprate = \beta_0 + \beta_1\, percmin + \beta_2\, avginc + \beta_3\, avgwlth + \beta_4\, avagdebt + u$$

avec *percmin* le pourcentage de minorités au sein de la communauté, *avginc* le revenu moyen, *avgwlth* la richesse moyenne, et *avgdebt* une mesure d'endettement moyen. Comment établiriez-vous l'hypothèse nulle qu'il n'existe aucune différence dans le taux d'octroi des prêts entre communautés ethniques, une fois pris en compte le revenu moyen, la richesse moyenne et le niveau d'endettement moyen ? Comment définiriez-vous l'hypothèse alternative qu'il existe de la discrimination à l'encontre des minorités dans le taux d'octroi des prêts ?

EXEMPLE 4.2
Réussite scolaire et taille des classes

Il existe depuis longtemps un intérêt marqué pour l'étude de l'impact de la taille des classes sur la réussite scolaire. (Voir par exemple, *The New York Times Magazine*, daté du 28/05/95.) Certains soutiennent que toutes choses égales par ailleurs, les élèves scolarisés dans des établissements de petite taille obtiennent de meilleurs résultats que ceux scolarisés dans des institutions de grande taille. Cette hypothèse est supposée valide même après avoir pris en compte les différences entre les tailles de classe des établissements.

Le fichier MEAP93.RA contient une base de données rassemblant des informations sur 408 écoles du secondaire dans l'état du Michigan pour l'année 1993. Nous pouvons dès lors utiliser ces données pour tester l'hypothèse nulle selon laquelle la taille de l'école n'a pas d'effet sur les résultats scolaires standardisés contre l'hypothèse alternative d'un effet négatif. La réussite est mesurée par le pourcentage d'étudiants ayant

obtenu une note leur permettant de valider le test de mathématiques du "Michigan Educational Assessment Program" (MEAP), soit une note standardisée à dix (*math*10). La taille de l'école est quant à elle mesurée par le taux d'inscription des étudiants (*enroll*). L'hypothèse nulle est donnée par $H_0 : \beta_{enroll} = 0$, et l'alternative $H_1 : \beta_{enroll} < 0$. Nous allons dans un premier temps ajouter deux variables de contrôle additionnelles, à savoir, la rémunération annuelle moyenne des enseignants (*totcomp*) et le nombre d'employé par millier d'élèves dans l'établissement (*staff*). L'introduction de la variable relative au salaire moyen des enseignants vise à capturer la qualité de l'enseignement, alors que le nombre d'employé est une mesure plus ou moins précise de l'attention portée aux élèves.

L'équation estimée, avec les écarts-types estimés reportés entre parenthèses, est donnée par :

$$\widehat{math10} = 2{,}274 + 0{,}00046\,totcomp + 0{,}048\,staff - 0{,}00020\,enroll$$
$$(6{,}113)\quad (0{,}00010)\qquad (0{,}040)\qquad (0{,}00022)$$
$$n = 408,\ R^2 = 0{,}0541.$$

Le coefficient de la variable *enroll*, $-0{,}00020$, confirme l'intuition de départ d'un impact négatif de la taille de l'école sur la réussite scolaire. En effet, plus le nombre d'inscrits est important plus le taux de réussite à l'examen de mathématiques est faible. (À noter que les coefficients de *totcomp* et *staff* présentent eux aussi les signes attendus.) Pour autant, le fait que le coefficient de la variable *enroll* soit différent de zéro pourrait être dû à l'erreur d'échantillonnage ; pour être certain de l'effet mesuré, nous devons conduire un test de Student.

Puisque $n - k - 1 = 408 - 4 = 404$, nous avons recours ici aux valeurs critiques d'une distribution normale. Au seuil de 5 %, la valeur critique est donnée par $-1{,}65$; la statistique t du coefficient de *enroll* doit être *inférieure* à $-1{,}65$ pour rejeter l'hypothèse H_0 au seuil de 5 %.

La statistique t relative à *enroll* est de $-0{,}0020/0{,}00022 \approx -0{,}91$ soit une valeur supérieure à $-1{,}65$: nous ne sommes *pas* en mesure de rejeter l'hypothèse H_0 en faveur de H_1 au seuil de 5 %. De la même manière, au seuil de 15 %, la valeur critique est de $-1{,}04$, et puisque $-0{,}91 > -1{,}04$, nous ne pouvons rejeter H_0, même au seuil de 15 %. Nous déduisons de ces résultats que le coefficient de *enroll* n'est pas statistiquement significatif au seuil de 15 %.

La variable *totcomp* est quant à elle statistiquement significative au seuil de 1 % puisque sa statistique t est de 4,6. À l'inverse, la statistique t de *staff* est de 1,2, et nous ne pouvons dans ces conditions rejeter $H_0 : \beta_{staff} = 0$ contre $H_1 : \beta_{staff} > 0$ même au seuil de 10 %. (La valeur critique obtenue à partir d'une distribution normale est de $c = 1{,}28$.)

Afin d'illustrer l'incidence de la spécification de notre modèle sur les conclusions de l'étude économétrique, nous estimons un autre modèle dans lequel l'ensemble des variables sont introduites en logarithmes. Cette transformation permet par exemple de faire en sorte que l'effet de la taille de l'établissement sur les résultats scolaires diminue à mesure que la taille augmente. L'équation estimée est donnée par :

$$\widehat{math10} = -207{,}66 + 21{,}16\log(totcomp) + 3{,}98\log(staff) - 1{,}29\log(enroll)$$
$$(48{,}70)\quad (4{,}06)\qquad (4{,}19)\qquad (0{,}69)$$
$$n = 408,\ R^2 = 0{,}0654.$$

La statistique t de log(*enroll*) est d'environ $-1{,}87$; dans la mesure où cette valeur est inférieure à la valeur critique au seuil de 5 %, $-1{,}65$, nous rejetons $H_0 : \beta_{\log(enroll)} = 0$ en faveur de $H_1 : \beta_{\log(enroll)} < 0$ au seuil de 5 %.

Dans le chapitre 2, nous avions étudié le cas d'un modèle dans lequel la variable dépendante était introduite sans transformation préalable (on parle alors de variable *en niveau*), alors que la variable indépendante était prise en logarithme (on parle alors de modèle *niveau-log*). L'interprétation d'un tel modèle dans le contexte

multivarié est le même qu'auparavant si ce n'est évidemment, que nous donnons une interprétation des coefficients estimés toutes choses égales par ailleurs. À supposer *totcomp* et *staff* fixés, nous obtenons $\Delta\widehat{math10} = -1{,}29[\Delta\log(enroll)]$ de sorte que : $\widehat{math10} \approx -(1{,}29/100)(\%\Delta enroll) \approx -0{,}013(\%\Delta enroll)$.

À nouveau, nous utilisons ici le fait que le taux de variation de log(*enroll*) multiplié par 100, correspond approximativement au pourcentage de variation de *enroll*. Ainsi, si le nombre d'inscrits dans un établissement augmente de 10 %, alors le modèle prédit que le résultat au test de mathématiques sera 0,013(10) = 0,13 points de pourcentage plus faible.

Quel modèle est-il alors préférable d'utiliser ? Le modèle introduisant le niveau ou la transformation logarithmique d'*enroll* ? Dans notre exemple, le modèle en niveau-niveau ne permet pas de conclure à l'incidence de la taille de l'établissement sur les résultats scolaires à l'inverse du modèle niveau-log qui identifie un effet négatif. Cette différence se traduit par une valeur du R-carré du modèle niveau-log plus élevée, ce qui signifie que les variations de *math*10 sont mieux expliquées lorsque nous utilisons un modèle niveau-log (6,5 % contre 5,4 %). Le modèle niveau-log est par conséquent préféré au modèle niveau-niveau car il permet de mieux capturer le lien entre les résultats scolaires et la taille des établissements. Nous détaillerons plus avant dans le chapitre 6 le calcul ainsi que l'utilisation qui peut être faite de l'indicateur du R-carré pour la sélection de modèle.

Alternatives bilatérales

Dans la plupart des applications, il est courant de tester l'hypothèse nulle $H_0 : \beta_j = 0$ contre une **hypothèse alternative bilatérale**, c'est-à-dire :

$$H_1 : \beta_j \neq 0. \qquad \textbf{[4.10]}$$

Sous cette hypothèse alternative, x_j a un effet *ceteris paribus* sur y sans spécifier si cet effet est positif ou négatif. Il s'agit là de l'hypothèse alternative la plus appropriée lorsque ni la théorie économique ni le bon sens ne permettent de définir précisément le signe de β_j. Même lorsque l'on connaît le signe de β_j sous l'hypothèse alternative, il est prudent de pratiquer un test bilatéral. En effet, le recours à un tel test, permet de nous éviter d'étudier l'équation estimée de façon à définir l'hypothèse alternative selon le signe de $\hat{\beta}$. L'inférence classique requiert que nous établissions les hypothèses nulle et alternative relatives à la population avant d'étudier les données et non de les adapter en fonction des résultats estimés. Par exemple, il ne serait pas correct de commencer par estimer la relation entre les performances scolaires et le nombre d'inscrits puis après avoir constaté que l'effet estimé est négatif, décider de formuler l'hypothèse alternative comme suit : $H_1 : \beta_{enroll} < 0$.

Lorsque l'hypothèse alternative est bilatérale, nous nous intéressons à la *valeur absolue* de la statistique *t*. La règle de décision pour l'hypothèse nulle $H_0 : \beta_j = 0$ contre (4.10) est alors donnée par :

$$|t_{\hat{\beta}_j}| > c, \qquad \textbf{[4.11]}$$

avec |.| la fonction valeur absolue et c la valeur critique adéquate. Pour trouver c, nous devons à nouveau spécifier le seuil de significativité, mettons 5 %. Dans le cadre d'un **test bilatéral**, c est choisi de façon à ce que l'aire sous chacune des queues de la distribution de Student soit égale à 2,5 %. En d'autres termes, c correspond au 97,5^e centile de la distribution de Student *à* $n - k - 1$ degrés de liberté. Lorsque $n - k - 1 = 25$, la valeur critique à 5 % du test bilatéral est donnée par $c = 2{,}060$. La figure 4.4 propose une illustration de cette distribution.

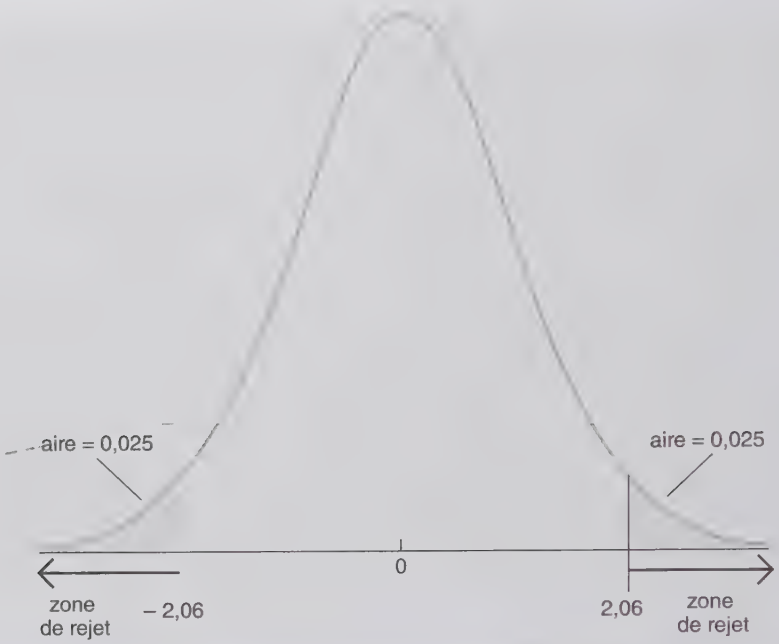


Figure 4.4 *Règle de décision au seuil de 5 % contre une hypothèse alternative $H_1 : \beta_j \neq 0$ avec 25 ddl.*

Lorsque l'hypothèse alternative n'est pas spécifiée, il est d'usage d'utiliser celle du test bilatéral. Par la suite, nous considérerons par défaut l'alternative bilatérale au seuil de significativité de 5 %. En pratique toutefois, lors de la réalisation d'une étude économétrique, il est toujours bon de mentionner explicitement l'hypothèse alternative et le seuil de significativité. Si H_0 est rejetée en faveur de (4.10) au seuil de 5 % nous dirons généralement que « x_j est **statistiquement significative**, ou statistiquement différente de zéro au seuil de 5 %. » Si H_0 n'est pas rejetée, nous dirons à l'inverse que « x_j est **statistiquement non significative** au seuil de 5 %. »

EXEMPLE 4.3
Déterminants de la réussite universitaire

Nous avons recours aux données contenues dans le fichier GPA1 pour estimer le modèle expliquant la moyenne obtenue par les étudiants en premier cycle universitaire (*colGPA*). En complément des variables traditionnelles, nous ajoutons dans notre équation de régression la variable *skipped* qui mesure le nombre de cours manqués par semaine. Le modèle estimé est donné par :

$$\widehat{colGPA} = 1{,}39 + 0{,}412\,hsGPA + 0{,}015\,ACT - 0{,}083\,skipped$$

$$(0{,}33)\ (0{,}094)\qquad (0{,}011)\qquad (0{,}026)$$

$$n = 141,\ R^2 = 0{,}234$$

Nous pouvons aisément calculer les statistiques t afin d'identifier les variables statistiquement significatives, en utilisant à chaque fois un test bilatéral. La valeur critique au seuil de 5 % est d'environ 1,96, puisque les degrés de liberté (141 – 4 = 137) sont suffisamment grands pour recourir à l'approximation normale. La valeur critique à 1 % est quant à elle d'environ 2,58.

La statistique t de *hsGPA* est de 4,38, ce qui apparaît très significatif même pour des valeurs très faibles du seuil de significativité. Ainsi, nous dirons généralement que « *hsGPA* est statistiquement significative aux seuils de significativité traditionnels ». La statistique t associée à *ACT* est de 1,36. Nous ne pouvons donc pas rejeter l'hypothèse nulle dans ce cas au seuil de 10 %. Il est également intéressant de noter la faible valeur du coefficient estimé. Ainsi, une augmentation de 10 points de la variable *ACT*, ce qui est important dans la réalité, devrait selon les prédictions du modèle, augmenter la moyenne obtenue à l'université de 0,15 points uniquement. Dès lors, la variable *ACT* est en pratique tout comme au sens statistique du terme, non significative.

Le coefficient attaché à la variable capturant l'absentéisme, *skipped*, a une statistique $t - 0{,}083/0{,}026 = -3{,}19$. Nous pouvons donc conclure que *skipped* est statistiquement significative au seuil de 1 % ($3{,}19 > 2{,}58$). Le coefficient associé à cette variable implique une diminution attendue de 0,083 de la moyenne générale (*colGPA*) par cours supplémentaire qui serait manqué par semaine. Autrement dit, toutes choses égales par ailleurs, la différence de moyenne attendue entre un étudiant n'ayant pas manqué de cours et un étudiant ayant manqué cinq cours sera de 0,42. Rappelez vous que ces résultats ne sont en rien reliés à la performance d'un ou plusieurs étudiants en particulier, la valeur de 0,42 devant s'interpréter comme l'effet moyen sur l'ensemble d'étudiants constitutifs de notre échantillon.

Dans cet exemple, nous pourrions utiliser un test unilatéral pour tester la significativité de chacune des variables. Dans le cas des variable *hsGPA* et *skipped*, celles-ci s'avèrent toutes deux très significatives dans le cadre d'un test bilatéral avec des signes conformes aux attentes ; il n'y a donc pas lieu de procéder à un test unilatéral. En revanche, les conclusions pour la variable *ACT* seraient sensiblement différentes si nous avions opté pour un test unilatéral puisque la variable ne serait alors significative qu'au seuil de 10 % mais pas à celui de 5 %. Le coefficient associé dans tous les cas demeure relativement faible.

Tester d'autres hypothèses relatives à β_j

Bien que $H_0 : \beta_j = 0$ soit l'hypothèse la plus courante, il arrive que nous soyons amenés à tester d'autres hypothèses relatives à β_j et notamment la possibilité que le paramètre prenne d'autres valeurs constantes. Deux cas standards illustrent cette possibilité et sont donnés par $\beta_j = 1$ ou $\beta_j = -1$. En général, l'hypothèse nulle est posée comme suit :

$$H_0 : \beta_j = a_j \qquad \textbf{[4.12]}$$

avec a_j la valeur hypothétique de β_j que nous stipulons. Dans ce cas, la statistique t s'écrit :

$$t = (\hat{\beta}_j - a_j)/\hat{\sigma}(\hat{\beta}_j).$$

Comme précédemment, la statistique t mesure simplement le nombre d'écarts-types qui sépare $\hat{\beta}_j$ de sa valeur hypothétique β_j. La formulation générale de la statistique t est donnée par :

$$t = \frac{(\textit{estimateur} - \textit{valeur hypothétique})}{\textit{écart-type estimé}} \qquad \textbf{[4.13]}$$

Sous l'hypothèse nulle (4.12), cette statistique t suit, d'après le théorème 4.2, une distribution de Student t_{n-k-1}. La statistique t usuelle est obtenue lorsque $a_j = 0$.

La formulation générale de la statistique t peut être utilisée pour réaliser des tests unilatéraux et bilatéraux. Par exemple, si l'hypothèse nulle et l'hypothèse alternative sont les suivantes, $H_0 : \beta_j = 1$ et $H_1 : \beta_j > 1$, nous identifions de la même façon que précédemment la valeur critique pour un test unilatéral, c : la seule différence se situe dans le calcul de la statistique t, et non dans l'établissement des valeurs critiques. À l'instar des résultats précédents, nous rejetons H_0 en faveur de l'hypothèse alternative H_1 est $t > c$. Nous dirons dans ce cas que « $\hat{\beta}_j$ est statistiquement plus grand que un » au seuil de significativité retenu.

EXEMPLE 4.4
Des liens entre la criminalité sur les campus américains et le taux d'inscription à l'université

Considérons un modèle de régression simple dans lequel le nombre de crimes sur les campus américains par an (*crime*) est expliqué par le nombre d'inscrits (*enroll*) :

$$\log(crime) = \beta_0 + \beta_1 \log(enroll) + u.$$

Il s'agit là d'un modèle à élasticités constantes, avec β_1 l'élasticité du crime par rapport au nombre d'inscrits. Il paraît peu pertinent dans ce cadre de tester l'hypothèse nulle $H_0 : \beta_1 = 0$ puisqu'il paraît assez évident que le nombre de crimes augmente avec la taille du campus. Une question plus intéressante semble être celle de savoir si l'élasticité de la criminalité par rapport au nombre d'inscriptions est de un : $H_0 : \beta_1 = 1$. Cela signifie qu'une augmentation de 1 % des inscriptions devrait s'accompagner, en moyenne, d'un accroissement du nombre de crimes de 1 %. L'hypothèse alternative, $H_1 : \beta_1 > 1$ implique qu'une augmentation de 1 % des inscriptions sera associée à une hausse du nombre de crimes *supérieure* à 1 %. Les crimes seraient dès lors un problème plus aigu sur les grands campus. Une manière de mieux visualiser ce phénomène est de considérer l'exponentiel de notre équation :

$$crime = exp(\beta_0) enroll^{\beta_1} exp(u).$$

(Voir l'annexe A pour les propriétés des fonctions logarithme népérien et exponentielle.) Une représentation graphique de cette fonction est proposée à la figure 4.5 lorsque $\beta_0 = 0$ et $u = 0$, dans les cas où $\beta_1 < 1$, puis $\beta_1 > 1$.

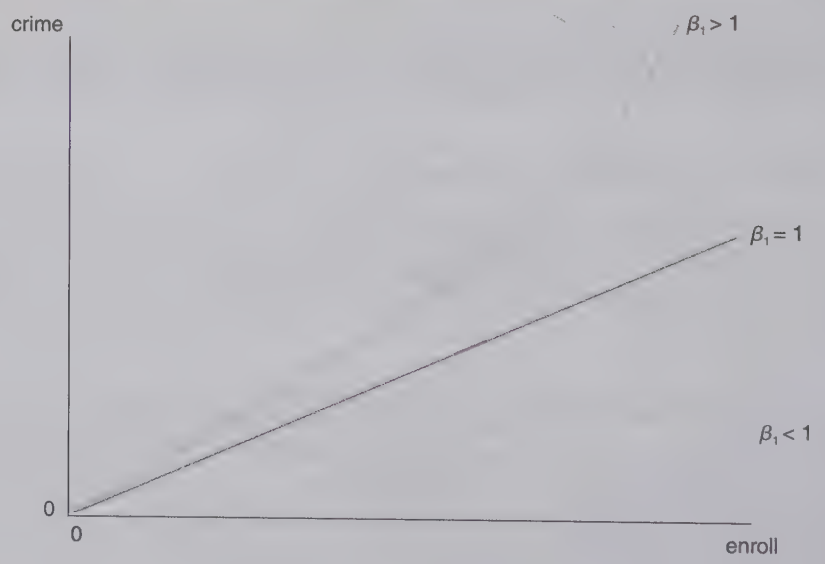


Figure 4.5 *Graphique de crime = pour et* © Cengage Learning, 2013

En résumé, nous testons l'hypothèse nulle $\beta_1 = 1$ contre l'hypothèse alternative $\beta_1 > 1$ à l'aide de données relatives à 97 universités aux États-Unis collectées en 1992. Ces données sont contenues dans le fichier CAMPUS. Elles sont issues du "Uniform Crime Reports" mis à disposition par le FBI. La moyenne des crimes par campus est de 394, alors que la moyenne du nombre d'inscriptions est quant à elle de 16076. Les résultats d'estimation sont les suivants :

$$\widehat{\log(crime)} = -6{,}63 + 1{,}27 \log(enroll) \quad [4.14]$$
$$(1{,}03)\ (0{,}11)$$
$$n = 97,\ R^2 = 0{,}585$$

L'élasticité estimée du crime en fonction du nombre d'inscrits, 1,27, présente le même signe que celle supposée pour H_1, $\beta_1 > 1$. Est-ce toutefois suffisant pour conclure que $\beta_1 > 1$? Nous devons être prudent pour procéder à ce test d'hypothèse, et ce notamment en raison du fait que les informations fournies par défaut par le logiciel sont plus riches en général que celles reportées dans l'équation (4.14). Intuitivement, nous serions tentés de construire une statistique t sur base du modèle à l'instar de ce que nous avons fait au début du chapitre, en divisant le coefficient estimé associé à la variable log(*enroll*) par son écart-type estimé associé. Cette statistique est celle généralement reportée par défaut dans les sorties des logiciels économétriques. Ce n'est toutefois pas la bonne statistique pour l'hypothèse nulle qui nous intéresse, $H_0 : \beta_1 = 1$. La statistique t correcte s'obtient à partir de (4.13) : il convient donc de soustraire la valeur hypothétique $\beta_1 = 1$, du coefficient estimé avant de diviser le tout par l'écart-type estimé tel que : $t = (1{,}27 - 1)/0{,}11 = 0{,}27/0{,}11 \approx 2{,}45$. La valeur critique de ce test au seuil de 5 % pour une distribution de Student avec $97 - 2 = 95$ degrés de liberté est d'environ 1,66 (en utilisant *ddl* = 120). Par conséquent, nous pouvons clairement rejeter l'hypothèse nulle $\beta_1 = 1$ en faveur de l'hypothèse alternative $\beta_1 > 1$ au seuil de 5 %. Par ailleurs on notera que l'hypothèse nulle peut également être rejetée au seuil de 1 %, pour laquelle la valeur critique est de 2,37.

Il est important de garder à l'esprit que nous n'avons pas tenu compte dans notre étude de l'influence de facteurs tiers. Par conséquent, la valeur de 1,27 n'est pas forcément une bonne approximation de l'effet *ceteris paribus*, entre la criminalité et le nombre d'inscrits. Il est possible par exemple que la taille des universités soit corrélée avec d'autres facteurs eux aussi susceptibles d'influencer le nombre de crimes comme le taux de criminalité des quartiers dans lesquels se situent les campus. Nous pourrions tenir compte de tels facteurs en collectant des données supplémentaires sur les taux de criminalité des villes hébergeant ces institutions par exemple.

Pour un test bilatéral par exemple, $H_0 : \beta_j = -1$, $H_1 : \beta_j \neq -1$, nous calculerons une statistique t conforme à (4.13) : $t = (\hat{\beta}_j + 1)/\hat{\sigma}(\hat{\beta}_j)$ (notez que le fait de soustraire -1 se traduit par ajouter 1 au numérateur). La règle de décision reste celle utilisée précédemment pour un test bilatéral : H_0 pourra être rejetée si $|t| > c$, c étant la valeur critique. Si H_0 est rejetée nous dirons que « $\hat{\beta}_j$ est statistiquement différent de moins un » au seuil retenu.

EXEMPLE 4.5
De l'impact de la pollution de l'air sur le prix des maisons

L'échantillon pour l'estimation du modèle est composé de 506 quartiers dans la région de Boston. Le prix médian d'une maison (*price*) dans un quartier est supposé dépendre d'un ensemble de caractéristiques du quartier comme, *nox*, la concentration du dioxyde d'azote dans l'air, en partie par million, *dist*, la distance pondérée entre le quartier et les cinq plus proches bassins d'emploi, *rooms*, le nombre moyen de chambres par maison dans le quartier, et enfin, *stratio*, le nombre moyen d'élèves par enseignant dans les écoles du quartier. Le modèle économétrique s'écrit comme suit :

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 stratio + u.$$

β_1 représente l'élasticité du prix par rapport à nox. Nous souhaitons ici tester l'hypothèse nulle $H_1 : \beta_1 = -1$ contre l'hypothèse alternative $H_1 : \beta_1 \neq -1$. La statistique t pour réaliser ce test est donnée par : $t = (\hat{\beta}_1 + 1)/\hat{\sigma}(\hat{\beta}_1)$.

Les données pour estimer le modèle sont contenues dans le fichier HPRICE2. L'équation reportant les paramètres ainsi que les écarts-types estimés (entre parenthèses) est donnée comme suit :

$$\widehat{\log(price)} = 11{,}8 - 0{,}954 \log(nox) - 0{,}134 \log(dist) + 0{,}255 rooms - 0{,}052 stratio$$

$$(0{,}32)\ (0{,}117) \qquad (0{,}043) \qquad (0{,}19) \qquad (0{,}006)$$

$$n = 506,\ R^2 = 0{,}581.$$

L'ensemble des coefficients estimés présentent le signe attendu. Par ailleurs, ils sont tous statstiquement significatifs au seuil de 5 %, log(*nox*) y compris. Toutefois, le résultat qui nous intéresse au premier chef n'est pas celui-ci, i.e. $\beta_1 = 0$. L'hypothèse nulle que nous souhaitons tester est la suivante $H_0 : \beta_1 = -1$, la statistique de test correspondante est $(-0{,}954 + 1)/0{,}117 = 0{,}393$, ceci nous amène à ne pas rejeter l'hypothèse nulle : l'élasticité du prix des maisons par rapport au taux de dioxyde d'azote n'est pas statistiquement différente de -1.

Calcul des p-valeurs pour les tests de Student

Jusqu'à maintenant, nous avons réalisé les tests d'hypothèse en utilisant la procédure classique : après avoir posé les hypothèses nulle et alternative, un seuil de significativité est choisi, permettant de déterminer une valeur critique. Cette valeur critique, une fois identifiée, est comparée à la valeur de la statistique t. L'hypothèse est enfin rejetée ou non pour un certain niveau de confiance.

Même après avoir décidé quelle était l'hypothèse alternative la plus appropriée, il reste une part d'arbitraire dans cette procédure puisque la règle de décision sera toujours conditionnelle au seuil choisi *a priori*. Or, il n'existe pas de seuil de significativité « correct » dans l'absolu. Le choix de ce dernier dépend souvent de la sensibilité du chercheur ainsi que de la question abordée dans le cadre du test d'hypothèse.

L'application d'une règle de rejet en fonction d'un seuil de significativité fixé à l'avance peut dissimuler une information importante sur les résultats du test. Par exemple, supposons que l'on souhaite tester l'hypothèse nulle d'absence de significativité d'une variable contre l'hypothèse alternative d'un effet significatif de cette même variable. La valeur obtenue pour la statistique t est de 1,85. La comparaison de cette statistique à la valeur critique au seuil de 5 % pour une distribution de Student à 40 degrés de liberté, $c = 2{,}02$ amène à ne pas rejeter l'hypothèse nulle. Fort de ce résultat, nous pourrions simplement conclure à l'absence d'effet de la variable d'intérêt au seuil de 5 %. Le test peut également être renouvelé avec un autre seuil, par exemple 10 %, pour évaluer la solidité de notre conclusion. Or, dans le cas d'espèce, il s'avère que l'hypothèse nulle peut être rejetée puisque la valeur critique passe à $c = 1{,}684$.

Pratiquer des tests d'hypothèses pour plusieurs seuils de significativité peut vite devenir fastidieux. C'est pourquoi, il est plus informatif de se poser la question suivante : étant donné la statistique t, quel est le plus petit seuil de significativité auquel l'hypothèse nulle serait rejetée ? Cette valeur, s'appelle la ***p*-valeur** (voir l'annexe C). Dans l'exemple précédent, nous savons que la p-valeur est supérieure à 0,05, puisque l'hypothèse nulle ne peut être rejetée à ce seuil. En suivant la même logique, nous la savons également inférieure à 10 % puisque l'hypothèse nulle peut être rejetée à 10 %. Le calcul exact de la p-valeur s'obtient en calculant la probabilité d'observer une valeur supérieure ou égale à 1,85 en valeur absolue pour une distribution de Student à 40 degrés de liberté. Autrement dit, la p-valeur correspond au seuil de significativité du test lorsque la valeur de la statistique t, 1,85 dans notre exemple, est utilisée comme valeur critique pour le test. La p-valeur est présentée à la figure 4.6.

Puisque la p-valeur est une probabilité, sa valeur sera toujours comprise entre 0 et 1. Son calcul exact à partir des tables statistiques n'est pas chose aisée puisque ces tables présentent une information discontinue. En pratique, ce calcul sera effectué par l'intermédiaire des logiciels économétriques qui offrent pour la plupart, des commandes permettant d'obtenir les paramètres estimés par la méthode des MCO ainsi que leurs p-valeurs associées. Par défaut, la p-valeur correspond habituellement à l'hypothèse nulle $H_0 : \beta_j = 0$ contre l'hypothèse alternative $H_1 : \beta_j \neq 0$. La p-valeur dans ce cas s'écrit :

$$P(|T| > |t|), \qquad \textbf{[4.15]}$$

où dans un souci de clarté, T désigne une variable aléatoire suivant une distribution de Student avec $n - k - 1$ degrés de liberté alors que t désigne la valeur numérique prise par la statistique de test.

La *p*-valeur permet de résumer de manière pratique les forces et les faiblesses des preuves empiriques contre l'hypothèse nulle. L'interprétation la plus parlante est peut-être la suivante : la *p*-valeur est la probabilité d'observer une statistique *t* aussi grande sous l'hypothèse qu'H_0 est vraie. Plus cette *p*-valeur est faible, plus on sera confiant dans le fait de ne pas se tromper en rejetant H_0. Par exemple, si la *p*-valeur = 0,50 (celle-ci sera toujours reportée avec une décimale, plutôt qu'un pourcentage), nous observerions une valeur de la statistique *t* de cette grandeur pour 50 % des échantillons aléatoires pour lesquels l'hypothèse se vérifie. Il s'agit ici d'une très faible preuve à l'encontre de l'hypothèse nulle.

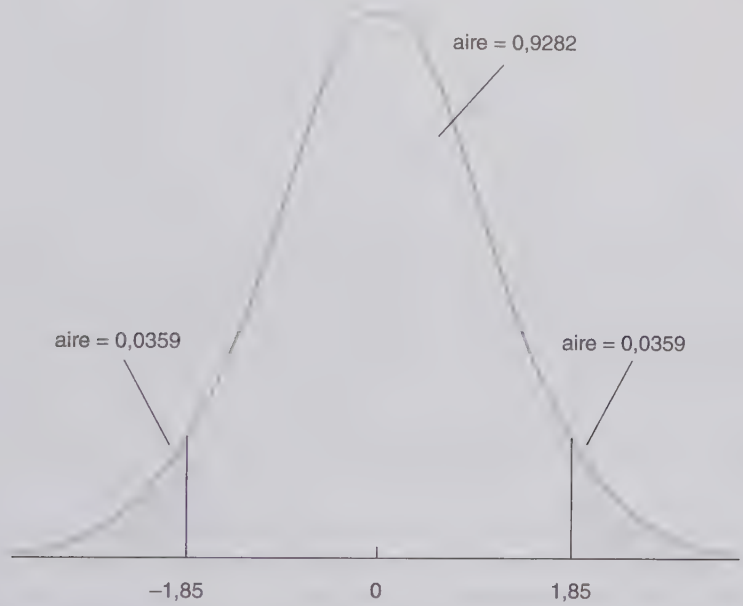


Figure 4.6 *Obtention de la* p*-valeur contre une alternative bilatérale, lorsque t = 1,85 et ddl = 40.*

Dans notre exemple avec *ddl* = 40 et *t* = 1,85, la *p*-valeur est calculée comme suit :

$$p\text{-valeur} = P(|T| > 1{,}85) = 2P(T > 1{,}85) = 2(0{,}0359) = 0{,}0718,$$

où $P(T > 1{,}85)$ est l'aire sous la courbe de la distribution de Student à 40 *ddl*, à droite de la valeur 1,85. (Cette valeur a été calculée à l'aide du logiciel économétrique Stata ; elle n'est pas reportée dans la table statistique G.2.) Ce résultat signifie que si l'hypothèse nulle est vraie, nous devrions observer une valeur absolue de la statistique *t* supérieure ou égale à 1,85 environ 7,2 pourcents du temps. La valeur tend à supporter le rejet de l'hypothèse nulle, même si nous ne pourrions le faire au seuil de significativité de 5 %.

Les exemples précédents ont permis d'illustrer la possibilité de réaliser des tests d'hypothèses à n'importe quel seuil de significativité dès que la *p*-valeur a été calculée. La règle de décision consiste simplement, pour α, le seuil de significativité du test, à décider de rejeter l'hypothèse H_0 si la *p*-valeur $< \alpha$; et de ne pas rejeter H_0 sinon, au seuil de α %.

Le calcul de la *p*-valeur pour un test unilatéral ne pose pas de problème particulier. Supposons, par exemple, que nous testions l'hypothèse nulle $H_0 : \beta_j = 0$ contre l'hypothèse alternative $H_1 : \beta_j > 0$. Si $\hat{\beta}_j < 0$, le calcul de la *p*-valeur n'a pas de sens puisque l'on sait que cette *p*-valeur est supérieure à 0,50, ce qui ne permettra jamais de rejeter l'hypothèse H_0 en faveur de l'hypothèse alternative H_1. Si $\hat{\beta}_j > 0$, en revanche, la statistique sera supérieure à zéro, $t > 0$, et la *p*-valeur reflètera la probabilité d'observer une valeur supérieure à *t* chez une

variable aléatoire suivant une distribution de Student avec le nombre de degrés de libertés adéquat. La plupart des logiciels économétriques reportent la *p*-valeur du test bilatéral. L'application au test unilatéral toutefois se fait très facilement en divisant cette *p*-valeur par deux.

Si l'hypothèse alternative est $H_1 : \beta_j < 0$, le calcul de la *p*-valeur n'a de sens que si $\hat{\beta}_j < 0$ (et ainsi $t < 0$) : *p*-valeur = $P(T < t) = P(T > |t|)$ car la distribution de Student est symétrique par rapport à 0. De nouveau, cette *p*-valeur peut être obtenue en divisant simplement par deux la *p*-valeur du test bilatéral. L'application de ces procédures de tests permet de se familiariser rapidement avec les ordres de grandeurs permettant de rejeter l'hypothèse nulle aux seuils habituels, notamment lorsque les valeurs de la statistique sont très élevées. Le report de la *p*-valeur peut dès lors sembler accessoire. Toutefois, celle-ci permet d'avoir une information plus fine sur le fait de savoir si l'on se trouve loin ou non du seuil de significativité, ce qui peut être utile au lecteur. Enfin, lorsque nous discuterons la mise en œuvre des tests de Fisher dans la section 4.5, nous verrons qu'il est important de calculer la *p*-valeur car les valeurs critiques pour le test *F* ne sont pas faciles à mémoriser. –

Pour aller plus loin 4.3

Supposez que vous estimiez un modèle de régression vous permettant d'obtenir une valeur estimée du paramètre inconnu : $\hat{\beta}_1 = 0{,}56$. La *p*-valeur est de 0,086 pour un test bilatéral avec $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$. Quelle est la *p*-valeur du test $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 > 0$?

Rappel du jargon des tests d'hypothèses classiques

Dans le jargon économétrique, lorsque H_0 n'est pas rejetée, on aura tendance à utiliser l'expression « on ne peut rejeter H_0 au seuil de *x* % », plutôt que « H_0 est acceptée au seuil de *x* % ». Cette manière de formuler les résultats du test peut être illustrée par l'exemple 4.5. Dans cet exemple, l'élasticité prix estimée par rapport à *nox* est de −0,954, et la statistique *t* associée pour tester $H_0 : \beta_{nox} = -1$ est $t = 0{,}393$. Par conséquent, nous ne pouvons rejeter H_0. Ce résultat ne veut pas dire pour autant que l'on connait de façon certaine la valeur de paramètre de la population car il y a d'autres valeurs de β_{nox} pour lesquelles l'hypothèse nulle ne peut également être rejetée. Par exemple, la statistique *t* pour $\beta_{nox} = -0{,}9$ est $(-0{,}954 + 0{,}9)/0{,}117 = -0{,}462$. Cette hypothèse ne peut elle non plus être rejetée. Naturellement, les hypothèses $\beta_{nox} = -1$ et $\beta_{nox} = -0{,}9$ ne peuvent être toutes les deux vraies. Dire que l'on « accepte » l'une ou l'autre hypothèse n'a dès lors pas de sens. Tout ce que nous pouvons dire est que les données ne nous permettent de rejeter aucune de ces deux hypothèses au seuil de significativité de 5 %.

Significativité statistique et significativité économique ou pratique

Après avoir mis en évidence l'importance de l'inférence statistique tout au long de cette section, nous allons maintenant aller au delà de la grandeur de la statistique *t* pour nous intéresser à la taille du paramètre estimé. Contrairement à la significativité statistique d'une variable déterminée par la taille de $t_{\hat{\beta}_j}$, la signification économique d'une variable dépend de la taille (et du signe) de $\hat{\beta}_j$. Il est important de se rappeler que la significativité statistique d'un paramètre dépend de deux éléments : la taille du paramètre estimé et l'écart-type estimé pour ce paramètre estimé, $t_{\hat{\beta}_j} = \hat{\beta}_j / \hat{\sigma}(\hat{\beta}_j)$. Par conséquent, on peut avoir un paramètre qui est très significatif parce que son écart-type calculé est très faible, $\hat{\sigma}(\hat{\beta}_j)$ ou bien parce que le coefficient estimé, $\hat{\beta}_j$, est très « grand » (en valeur absolue). En revanche, si le coefficient estimé, $\hat{\beta}_j$, est faible (en valeur absolue), la variable associée à ce coefficient n'aura quand même pas beaucoup de poids dans l'explication de la variable dépendante.

EXEMPLE 4.6
Participation au plan d'épargne retraite 401(k)

Dans l'exemple 3.3, nous utilisions les données sur les plans d'épargne retraite américains, 401(k)[1] afin d'estimer un modèle mesurant l'impact de l'ancienneté du plan sur le taux de participation des salariés. En complément des variables explicatives *mrate* et *age*, nous ajoutons maintenant une mesure de la taille de l'entreprise, à savoir le nombre total d'employés (*totemp*).

L'équation reportant les paramètres ainsi que l'écart-type (entre parenthèses) estimés est la suivante :

$$\widehat{\text{prate}} = 80,29 + 5,44\,mrate + 0,269\,age - 0,000013\,totemp$$
$$(0,78)\ \ (0,52)\qquad (0,045)\qquad (0,00004)$$
$$n = 1534,\ R^2 = 0,100$$

La variable ayant la plus petite statistique t en valeur absolue est *totemp* : $t = -0,00013/0,00004 = -3,25$. Par conséquent, nous pouvons rejeter l'hypothèse nulle selon laquelle le paramètre de la population est égal à zéro, pour l'ensemble des variables explicatives (la p-valeur d'un test bilatéral pour cette statistique t est d'environ 0,001).

Maintenant que l'on a déterminé la significativité statistique de la variable *totemp*, nous allons analyser son importance économique. Notre modèle prédit, une baisse de la participation des employés au plan d'épargne de 10000(0,00013) = 1,3 points de pourcentage, *ceteris paribus*, si les effectifs augmentent de 10 000 personnes, et que les variables *mrate* et *age* restent inchangées. Cet exemple illustre la faible ampleur du lien entre les deux variables, malgré la significativité statistique, puisqu'un très grand changement du nombre d'employés n'occasionne qu'une variation très faible du taux de participation au plan d'épargne.

L'exemple précédent montre l'importance de toujours bien interpréter la taille des coefficients en sus de la statistique t notamment en présence de grands échantillons. Pour rappel, une grande taille d'échantillon aura tendance à améliorer la précision des estimateurs et donc à diminuer les écarts-types. Dès lors, un coefficient même de faible ampleur pourra être associé à une statistique t supérieure à la valeur critique calculée en fonction des seuils de significativité traditionnels.

EXEMPLE 4.7
Effet de la formation professionelle sur les taux de mise au rebut

Le taux de mise au rebut de production pour une entreprise est le nombre d'articles défectueux – un produit qui ne peut être vendu – sur 100 articles produits. Une diminution de ce taux reflète une amélioration de la productivité des travailleurs.

Dans l'application qui suit, nous utilisons le taux de rebut pour mesurer l'impact de la formation sur la productivité des travailleurs. Les données utilisées sont contenues dans le fichier JTRAIN. Elles rassemblent un ensemble d'observations sur 29 entreprises ne disposant pas en leur sein de structure syndicale pour l'année 1987. L'équation estimée est donnée par :

$$\widehat{\log(scrap)} = 12,46 - 0,029\,hrsemp - 0,962\log(sales) + 0,761\log(employ)$$
$$(5,69)\ \ (0,023)\qquad (0,453)\qquad (0,407)$$
$$n = 29,\ R^2 = 0,262$$

1 Le plan 401(k), ou 401(k), est un système d'épargne retraite par capitalisation utilisé aux États-Unis. L'employé est libre de participer ou non au plan 401(k) offert par son employeur, et peut également décider de limiter les sommes épargnées.

avec *hrsemp*, le nombre d'heures de formation suivies par les employés annuellement, *sales*, le montant des ventes annuelles pour l'entreprise (en dollars), et *employ*, le nombre d'employés de l'entreprise. Pour l'année 1987, le taux de mise au rebut moyen dans l'échantillon se situe aux alentours de 4,6, la moyenne d'heures de formation est elle d'environ 8,9.

Les résultats relatifs à notre principale variable d'intérêt nous montrent qu'une heure de plus de formation par employé réduit le log*(scrap)* de 0,029, ce qui correspond à une baisse d'environ 2,9 % du taux de mise au rebut. Si *hrsemp* augmente de 5 – nombre d'heures supplémentaires par employé sur l'année – le taux de rebut tombe à 5(2,9) = 14,5 %. Les gains de productivité liés à la mise en œuvre d'heures de formation semblent par conséquent substantiels. Ceci ne donne pas d'indication définitive toutefois sur l'opportunité pour une entreprise de développer ces programmes de formation. Une telle décision nécessitera d'aller plus avant dans l'analyse en s'appuyant notamment sur une étude coût bénéfice mettant en regard les coûts associés aux programmes de formation et les gains attendus d'une baisse de la mise au rebut. Les informations en notre possession ne nous permettent pas en l'état de réaliser cette étude.

Si l'analyse précédente a pu mettre en évidence la significativité économique de la formation sur la productivité, nous nous posons maintenant la question de sa significativité statistique. Étonnamment, la valeur de la statistique t : $-0{,}029/0{,}023 = -1{,}26$ se situe bien en dessous des valeurs critiques traditionnelles, comme celle associée au seuil de 5 % par exemple. L'effet n'est donc pas assez fort pour pouvoir conclure que la variable *hrsemp* est statistiquement significative au seuil de 5 %. Ce résultat demeure même si l'on considère un test unilatéral à $29 - 4 = 25$ *ddl*, puisque sous l'hypothèse alternative $H_1 : \beta_{hrsemp} < 0$, la valeur critique, $-1{,}71$ reste supérieure à la statistique.

Comment comprendre ce résultat ? La petite taille de l'échantillon utilisé explique sans doute en grande partie la difficulté à rejeter l'hypothèse nulle. Nous devrions par conséquent assouplir notre règle de décision en considérant un seuil moins strict. La valeur critique d'un test unilatéral au seuil de 10 %, $-1{,}32$, permet quasiment de conclure à la significativité statistique de la variable. Ceci se voit nettement avec la p-valeur associée, $P(T_{25} < -1{,}26) = 0{,}110$. À ce stade, certains pourront conclure que les preuves empiriques sont suffisantes pour rejeter l'hypothèse nulle, d'autres trouveront néanmoins la p-valeur encore trop importante pour arriver à une conclusion similaire. La sensibilité du chercheur ou bien les conventions dans les domaines d'application demeurent souvent pour ces ordres de grandeur, les seuls guides pour trancher.

Afin de limiter l'influence de la taille de l'échantillon sur les résultats des tests, certains chercheurs ajustent le seuil de significativité en considérant des seuils plus petits pour les grandes tailles d'échantillons. Par exemple, imaginons le cas d'une règle de rejet conditionnée au seuil de 5 % pour un échantillon de taille n est composé d'une centaine d'observations, nous devrions adopter un seuil de significativité plus dur à 1 % lorsque n contient plusieurs milliers d'observations. En ajustant le seuil de significativité à la taille de l'échantillon, on espère mieux faire coïncider les notions de significativité statistique et économique. Il n'y a toutefois pas d'assurance que ceci fonctionne correctement. L'exemple précédent illustre bien ce point, puisque la variable *totemp*, marginale d'un point de vue économique, reste significative au seuil de 1 %.

La plupart des chercheurs appliquent également le raisonnement inverse, en utilisant des seuils moins restrictifs pour les échantillons de petite taille. Là encore, la pertinence de tels ajustements dépendra du problème posé et de l'objectif de l'étude.

Rappelons que la multicolinéarité (soit la forte corrélation entre les variables indépendantes du modèle) peut également contribuer au gonflement des écarts-types, en dépit de tailles d'échantillon raisonnables. Comme déjà évoqué dans la section 3.4 du chapitre 3, peu de choses peuvent être mises en place pour résoudre le problème de multicolinéarité et son impact sur la précision des estimateurs, si ce n'est l'augmentation de la taille de l'échantillon ou le repositionnement de l'étude en supprimant ou combinant des variables indépendantes fortement liées. Comme dans le cas des petits échantillons, la mesure des effets marginaux reste incertaine et difficile à obtenir lorsque les variables exhibent une corrélation élevée (voir la section 4.5 pour des exemples supplémentaires).

Nous terminerons cette section par quelques recommandations pour discuter les questions de significativité économique et statistique dans les modèles de régression multiple :

1. Vérifiez tout d'abord la significativité statistique de vos variables. Si la variable est significative continuez l'analyse en discutant la grandeur du coefficient afin de tirer vos premières conclusions sur l'ampleur des effets réels ou économiques. Cette dernière étape nécessite de prendre certaines précautions en raison notamment de la forme que prennent les variables dépendante et indépendantes de l'équation (en particulier, on s'interrogera sur les unités de mesures choisies, mais aussi sur l'utilisation de la transformation logarithmique.)

2. Si la variable n'est pas significative aux seuils habituels (10 %, 5 %, 1 %), il convient tout de même de se demander si le signe de l'effet sur la variable d'intérêt est conforme aux attentes et si l'ampleur de l'effet prédit par le modèle est importante. Si c'est le cas, la statistique t et la p-valeur peuvent être calculées. Dans le cas d'un petit échantillon, la règle de rejet peut être assouplie afin de rejeter l'hypothèse nulle pour des p-valeurs allant jusque 0,20 (il est important de noter cependant qu'il n'existe pas de règle précise qui fasse consensus en la matière). Les conclusions tirées des variables associées à des p-valeurs importantes, c'est-à-dire des statistiques t faibles, sont toujours très délicates car les valeurs élevées des paramètres estimés peuvent être simplement dues à une erreur d'échantillonnage : le triage d'un échantillon aléatoire autre pourrait aboutir à des résultats sensiblement différents.

3. Il arrive par ailleurs de trouver des variables avec un mauvais signe, c'est-à-dire un signe contraire aux attentes initiales, et une faible valeur de la statistique t. Ces variables sont en pratique ignorées : on conclut à l'absence de significativité statistique de la variable. Un cas plus ennuyeux est celui où le signe est de nouveau contre intuitif mais l'effet sur la variable d'intérêt important. On aura tendance en général dans ce cas à questionner la spécification du modèle ainsi que la nature des données. Souvent, les variables statistiquement significatives avec un signe contre-intuitif sont le résultat d'un problème de variable omise ou d'autres problèmes plus importants qui seront abordés dans les chapitres 9 et 15.

4.3 INTERVALLES DE CONFIANCE

Sous les hypothèses MLC, nous pouvons également facilement construire un **intervalle de confiance (IC)** pour un paramètre de la population β_j. Cet intervalle de confiance fournit l'ensemble des valeurs possibles du paramètre de la population, et pas simplement une estimation ponctuelle de cette valeur.

En utilisant le fait que $(\hat{\beta}_j - \beta_j)/\hat{\sigma}(\hat{\beta}_j)$ suit une distribution de Student à $n - k - 1$ degrés de liberté [voir l'équation (4.3)], un simple réarrangement de l'expression permet de calculer l'intervalle de confiance ayant 95 % de chance de contenir le paramètre inconnu β_j :

$$\hat{\beta}_j \pm c \cdot \hat{\sigma}(\hat{\beta}_j), \qquad \textbf{[4.16]}$$

où la constante c correspond à la valeur du 95^e centile obtenue à partir des tables de la distribution de Student à $n - k - 1$ degrés de liberté. Plus précisément, les bornes inférieure et supérieure de l'intervalle de confiance sont données par l'expression suivante :

$$\underline{\beta}_j \equiv \hat{\beta}_j - c \cdot \hat{\sigma}(\hat{\beta}_j)$$

et

$$\overline{\beta}_j \equiv \hat{\beta}_j + c \cdot \hat{\sigma}(\hat{\beta}_j)$$

À ce stade, il est utile de rappeler la signification d'un intervalle de confiance. Imaginons que l'on procède à un grand nombre de tirages successifs d'échantillons aléatoires au sein de la même population en calculant à chaque fois $\underline{\beta}_j$ et $\overline{\beta}_j$. Dans ce cas, le paramètre (inconnu) de la population, β_j, devrait se trouver

dans 95 % des cas entre ces deux bornes. Lors de notre estimation, nous ne pouvons malheureusement pas savoir si le paramètre inconnu fait réellement parti de l'intervalle de confiance que nous avons construit sur un unique échantillon. Nous ne pouvons dès lors qu'espérer être dans 95 % des intervalles de confiance contenant β_j. Nous n'avons toutefois aucune garantie que ce soit le cas.

En pratique, la mise en œuvre du calcul des intervalles de confiance se fait aisément grâce aux logiciels économétriques. Seules trois grandeurs sont nécessaires pour y parvenir : $\hat{\beta}_j$, $\hat{\sigma}(\hat{\beta}_j)$, et c. Les deux premières sont généralement reportées par défaut dans les tableaux de sortie. La dernière, c, dépend du nombre de degrés de liberté, $n - k - 1$, ainsi que du niveau de confiance choisi, 95 % dans notre exemple. Une fois ces valeurs renseignées, c s'obtient simplement à partir des tables de la distribution de Student pour $n - k - 1$ degrés de liberté.

Ainsi, l'intervalle de confiance d'un paramètre β_j, pour un niveau de confiance de 95 % et un nombre de degrés de liberté de $n - k - 1 = 25$ sera donné par l'expression suivante : $[\hat{\beta}_j - 2{,}06\hat{\sigma}(\hat{\beta}_j), \hat{\beta}_j + 2{,}06\hat{\sigma}(\hat{\beta}_j)]$.

Pour les grands échantillons, la loi normale standardisée constitue une bonne approximation de la distribution de Student. Si $n - k - 1 > 120$, nous pouvons donc utiliser le 97,5^e centile de la loi normale standardisée pour construire l'intervalle de confiance à 95 % IC : $\hat{\beta}_j \pm 1{,}96\hat{\sigma}(\hat{\beta}_j)$. En pratique, le calcul peut être d'avantage simplifié lorsque l'on travaille à un niveau de confiance de 95 %, en remplaçant la valeur critique par une approximation, $1.96 \simeq 2$. Par ce bais, lorsque $n - k - 1 > 50$, nous pouvons appliquer la règle empirique consistant à soustraire deux écarts-types au coefficient estimé, $\hat{\beta}_j$, pour calculer la borne inférieure de l'intervalle et à effectuer la même opération en ajoutant deux écarts-types pour calculer la borne supérieure. Pour des degrés de liberté inférieurs, en revanche, il conviendra de se référer aux tables de la distribution de Student.

Le calcul peut naturellement être étendu à d'autres niveaux de confiance. À 90 % par exemple, le seuil, c, correspond au 95^e centile de la distribution de Student à $n - k - 1$ *ddl*. Pour un nombre de degrés de liberté de $n - k - 1 = 25$, on obtient une valeur critique, $c = 1{,}71$. L'intervalle de confiance correspondant est donc $\hat{\beta}_j \pm 1{,}71\hat{\sigma}(\hat{\beta}_j)$. Cet intervalle de confiance est naturellement plus étroit que celui obtenu précédemment pour un niveau de confiance à 95 %. Si l'on considère maintenant un niveau de confiance à 99 %, la valeur du seuil, c, correspondra au 99,5^e centile de la distribution de Student à 25 *ddl*. L'intervalle de confiance associé s'écrira $\hat{\beta}_j \pm 2{,}79\hat{\sigma}(\hat{\beta}_j)$.

La plupart des logiciels économétriques actuels calculent par défaut l'intervalle de confiance des paramètres inconnus à un niveau de confiance de 95 %, en plus des paramètres estimés et des écarts-types estimés. Grace à cette information, on peut aisément procéder à un test bilatéral pour n'importe quelle hypothèse nulle. En effet, si $H_0 : \beta_j = a_j$, nous déciderons de rejeter H_0 en faveur de l'hypothèse alternative $H_1 : \beta_j \neq a_j$ au seuil de significativité de 5 % si et seulement si, a_j, ne se trouve pas dans l'intervalle de confiance à 95 %.

EXEMPLE 4.8
Déterminants des dépenses de R&D

Les spécialistes d'économie industrielle s'intéressent depuis longtemps à la nature des liens entre la taille des entreprises – souvent mesurée par le montant des ventes annuelles – et les dépenses en recherche et développement (R&D). Le modèle employé suppose en général une élasticité constante entre ces deux variables d'intérêt. Une autre question connexe ayant fait l'objet d'une attention particulière concerne l'effet *ceteris paribus* de la marge bénéficiaire nette, c'est-à-dire des bénéfices nets en pourcentage du chiffre d'affaire, sur les dépenses de R&D. Afin d'apporter un éclairage sur cette question à l'aide de données réelles, nous utilisons la base de données contenue dans le fichier RDCHEM. Cette base comprend des données

relatives à 32 entreprises de l'industrie chimique. Nous estimons l'équation suivante (les écarts-types estimés sont reportés entre parenthèses) :

$$\widehat{\log(rd)} = -4{,}38 + 1{,}084\log(sales) + 0{,}0217\,profmarg$$
$$(0{,}47)\ \ (0{,}060)\qquad\qquad (0{,}0128)$$
$$n = 32,\ R^2 = 0{,}918$$

L'élasticité estimée des dépenses de R&D en fonction des ventes est de 1,084. Pour un niveau de marge bénéficiaire donné, une augmentation de 1 % des ventes est associée à une hausse de 1,084 % des dépenses de R&D. (On notera que les dépenses de R&D et les ventes sont toutes les deux exprimées en millions de dollars ; l'unité de mesure n'a toutefois pas d'incidence sur l'estimation de l'élasticité). Grace aux informations reportées dans le tableau de sortie, nous pouvons calculer l'intervalle de confiance à 95 % du paramètre associé à la variable log(*sales*). La table statistique G.2 nous donne la valeur seuil correspondant au 97e centile d'une distribution de Student à $n - k - 1 = 32 - 2 - 1 = 29$ *ddl* soit $c = 2{,}045$. L'intervalle de confiance pour le paramètre inconnu, $\beta_{\log(sales)}$, à 95 % est donné par : $1{,}084 \pm 0{,}60(2{,}045)$ où environ (0,961 ; 1,21). Un des premiers constats est que zéro ne fait pas partie de l'intervalle. Ce résultat ne nous étonne guère puisque l'on s'attend à ce que les montants investis en dépenses de R&D augmentent en fonction de la taille de l'entreprise. Un résultat plus intéressant concerne l'hypothèse nulle $H_0 : \beta_{\log(sales)} = 1$ contre l'hypothèse alternative $H_1 : \beta_{\log(sales)} \neq 1$ au seuil de 5 %. La valeur 1 fait partie de l'intervalle de confiance. Ceci signifie donc que nous ne pouvons rejeter l'hypothèse nulle selon laquelle l'élasticité entre les dépenses de R&D et les ventes ne sont pas statistiquement différentes de 1 au seuil de significativité de 5 % (on remarquera également que la valeur estimée est extrêmement proche de 1).

Le coefficient estimé associé à la variable *profmarg* présente quant à lui un signe positif. L'intervalle de confiance du paramètre $\beta_{profmarg}$, à 95 % est $0{,}0217 \pm 0{,}0128(2{,}045)$, soit environ (−0,0045 ; 0,0479). Contrairement au cas précédent, la valeur zéro se trouve maintenant entre les deux bornes de l'intervalle. Nous ne pouvons donc pas rejeter l'hypothèse nulle $H_0 : \beta_{profmarg} = 0$ contre l'hypothèse alternative $H_1 : \beta_{profmarg} \neq 0$ au seuil de 5 %. Le calcul de la statistique de Student, $t = 1{,}70$, et de la *p*-valeur associée, environ 0,10, apporte une information complémentaire intéressante. Si l'hypothèse nulle ne peut être rejetée au seuil de 5 %, nous pouvons conclure au seuil de 10 % que la variable présente un impact statistiquement significatif. Il est possible de tirer une conclusion similaire si l'on considère un seuil de 5 % mais avec cette fois-ci l'hypothèse alternative suivante : $H_1 : \beta_{profmarg} > 0$. Attardons nous maintenant sur l'interprétation économique de nos résultats. Le résultat obtenu concernant la variable, *profmarg*, nous dit qu'une augmentation de la marge bénéficiaire nette d'un pourcent, toute chose égale par ailleurs, devrait en moyenne accroître les dépenses en R&D de $100(0{,}0217) \approx 2{,}2$ %.

Avant de conclure cette partie, il convient de rappeler l'importance des hypothèses du modèle linéaire classique sous-jacentes au calcul de l'intervalle de confiance et sa pertinence pour l'analyse. Imaginons par exemple que des facteurs importants omis dans la spécification soient corrélés avec les variables indépendantes. Dans ce cas, nos estimateurs des MCO ne peuvent être utilisés pour calculer les intervalles de confiance car ils souffrent d'un biais d'omission. Si maintenant les erreurs sont hétéroscédastiques – ceci serait le cas par exemple dans notre exemple précédent si la variance de la variable *log*(*rd*) dépendait des variables explicatives – les écarts-types estimés ne peuvent plus être considérés comme des estimateurs fiables de $\sigma(\hat{\beta}_j)$ (voir la discussion dans la section 3.4 du chapitre 3). Nous ne pourrons dès lors plus considérer que la vraie valeur du paramètre inconnu se trouve pour un niveau de confiance donné, dans l'intervalle de confiance obtenu à partir des écarts-types estimés. Enfin, l'hypothèse de normalité joue aussi un rôle important puisqu'elle permet d'obtenir les valeurs critiques utilisées dans le calcul des bornes. Comme nous le verrons dans le chapitre 5 néanmoins, cette hypothèse perd de son importance lorsque l'échantillon contient plusieurs centaines d'observations.

4.4 TESTS D'HYPOTHÈSES SUR UNE COMBINAISON LINÉAIRE SIMPLE DES PARAMÈTRES

Les deux sections précédentes nous ont permis de voir comment utiliser les hypothèses du modèle linéaire classique afin de procéder au calcul des intervalles de confiance et aux tests d'hypothèses usuels sur un paramètre, β_j. En pratique, il arrive souvent de devoir tester des hypothèses impliquant plus d'un paramètre de la population. Dans cette section, nous présenterons comment effectuer un test d'hypothèse faisant intervenir plus d'un paramètre. La section 4.5, traite quant à elle de la question des tests d'hypothèses multiples.

Afin d'illustrer l'approche générale des tests sur une combinaison linéaire de paramètres, nous considérons un modèle simple servant à expliquer les rendements de l'éducation et les comparons pour les cursus du supérieur américains de type court, organisés au sein des "*junior colleges*", qui s'étalent sur deux ans, et ceux de type long, sur quatre ans, organisés au sein des "*four-year colleges*". Pour simplifier notre exemple, on considèrera que les programmes de type courts sont rattachés aux "*colleges*" et ceux de type long aux universités. [Kane et Rouse (1995) proposent une analyse comparée des rendements de l'éducation des programmes de "*college*" en deux et quatre ans.] La population étudiante considérée comprend les indvidus qui travaillent ayant obtenu un diplôme du secondaire. Le modèle estimé est repris ci-dessous :

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_{f3} exper + u \qquad \textbf{[4.17]}$$

où

jc = nombre d'années d'étude dans un programme du supérieur de type court, "*two year college*"
$univ$ = nombre d'années d'étude dans un programme du supérieur de type long, "*four-year college*"
$exper$ = nombre de mois d'expérience professionnelle

On notera concernant les années d'éducation, que toute combinaison est possible pour les deux cursus d'intérêt. Il est possible notamment qu'une personne ne soit passée ni par l'un ni par l'autre cursus, c'est-à-dire que $jc = 0$ et $univ = 0$.

La question qui nous intéresse ici est de savoir si les rendements associés à une année passée au "*college*" comparativement à l'université sont les mêmes. Formellement, l'hypothèse testée s'écrit comme suit :

$$H_0 : \beta_1 = \beta_2. \qquad \textbf{[4.18]}$$

Sous l'hypothèse H_0, une année supplémentaire au "*college*" ou à l'université conduisent, toutes choses égales par ailleurs, à une augmentation de salaire équivalente. L'hypothèse alternative sera que le rendement de l'éducation d'une année supplémentaire au "*college*" sera inférieur à celui d'une année supplémentaire à l'université :

$$H_1 : \beta_1 < \beta_2. \qquad \textbf{[4.19]}$$

À la différence des cas de figure abordés jusqu'à présent, les hypothèses (4.18) et (4.19) impliquent les deux paramètres β_1 et β_2. Il n'est dès lors plus possible de se limiter aux calculs des statistiques de Student des deux paramètres pour procéder au test. Ceci ne veut pas dire qu'il n'est pas envisageable de construire une statistique de Student permettant de répondre à notre question. En effet, nous pouvons remanier l'écriture de l'hypothèse nulle ainsi que celle de l'hypothèse alternative de sorte à obtenir les expressions suivantes : $H_0 : \beta_1 - \beta_2 = 0$ et $H_1 : \beta_1 - \beta_2 < 0$, Dans ce cas, nous pouvons calculer une statistique de Student qui ne dépendra pas d'un paramètre mais de l'écart entre les deux paramètres estimés, $\hat{\beta}_1 - \hat{\beta}_2$. Pour rejeter l'hypothèse nulle (4.18) en faveur de l'hypothèse alternative (4.19) cet écart doit être suffisamment inférieur à zéro. Comme précédemment, il convient de prendre en compte l'erreur d'échantillonnage dans le calcul du test. Pour ce faire, on procède à la standardisation de la statistique en divisant $\hat{\beta}_1 - \hat{\beta}_2$ par son écart-type :

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\hat{\sigma}(\hat{\beta}_1 - \hat{\beta}_2)} \qquad \textbf{[4.20]}$$

Une fois la statistique de Student obtenue (4.20), nous pouvons procéder au test comme dans le cas d'un test sur un seul paramètre. Ainsi, nous choisissons un seuil de significativité pour le test, puis en fonction du nombre de *ddl*, nous obtenons une valeur critique qui sera comparée à la valeur de la statistique de test. Comme l'hypothèse alternative est de la forme (4.19), la règle de rejet impliquera de ne pas accepter l'hypothèse nulle si $t < -c$, c étant une valeur positive extraite de la table statistique appropriée. Une alternative peut consister à utiliser la p-valeur correspondant à la statistique de Student pour appliquer la règle de décision (voir section 4.2).

La procédure décrite est par conséquent très similaire à celle appliquée précédemment. La principale différence entre ces deux approches se situe dans le calcul de l'écart-type lorsque l'on considère plus d'un paramètre. Les logiciels économétriques fournissent en général par défaut dans leur tableau de sortie de régressions par les MCO, les écarts-types associés aux coefficients estimés. Par exemple, si l'on estime la relation entre le *log(wage)* et ses principaux déterminants sur les données utilisées par Kane et Rouse (1995), contenues dans le fichier TWOYEAR, nous obtenons les résultats suivants :

$$\widehat{\log(wage)} = 1{,}472 + 0{,}0667\,jc + 0{,}0769\,univ + 0{,}0049\,exper$$

$$(0{,}021)\ (0{,}0068) \qquad (0{,}0023) \qquad (0{,}0002)$$

$$n = 6\,763.\ R^2 = 0{,}222. \qquad \textbf{[4.21]}$$

On voit très clairement à partir de ces résultats que les variables *jc* et *univ* sont à la fois significatives d'un point de vue économique et statistique pour expliquer le salaire. Cette question est intéressante mais n'est pas celle que nous souhaitions traiter. Concernant l'importance relative des deux types de cursus, on notera une différence négative entre les deux coefficients estimés d'intérêt : $\hat{\beta}_1 - \hat{\beta}_2 = -0{,}0102$. Le rendement à attendre d'une année supplémentaire au "*college*" est donc inférieur d'un point environ à celui d'une année supplémentaire à l'université. D'un point de vue économique, cette différence semble assez faible. Pour procéder à un test statistique toutefois, nous devons aller plus avant dans notre analyse. L'écart entre les paramètres estimés nous donne la valeur du numérateur de notre statistique de Student.

Malheureusement, nous ne disposons pas de l'information suffisante pour calculer la statistique puisque le logiciel ne nous indique pas la valeur de l'écart-type estimé de $\hat{\beta}_1 - \hat{\beta}_2$. Il pourrait être tentant de considérer que $\hat{\sigma}(\hat{\beta}_1 - \hat{\beta}_2) = \hat{\sigma}(\hat{\beta}_1) - \hat{\sigma}(\hat{\beta}_2)$. Ceci n'est cependant pas correct. Pour le voir, imaginons simplement que l'on renverse l'ordre des paramètres $\hat{\beta}_1$ et $\hat{\beta}_2$. Nous aurions alors un écart-type négatif pour l'écart entre les paramètres ce qui n'est évidemment pas possible, un écart-type devant toujours être positif. Bien que l'écart-type de $\hat{\beta}_1 - \hat{\beta}_2$ dépende bien de $\sigma(\hat{\beta}_1)$ et $\sigma(\hat{\beta}_2)$, la forme de ce lien n'est pas triviale. Afin de déterminer $\hat{\sigma}(\hat{\beta}_1 - \hat{\beta}_2)$, nous procédons d'abord au calcul de la variance de la différence (voir l'annexe B pour plus de détails sur les calculs de variance). On obtient l'expression suivante :

$$\mathrm{Var}(\hat{\beta}_1 - \hat{\beta}_2) = \mathrm{Var}(\hat{\beta}_1) + \mathrm{Var}(\hat{\beta}_2) - 2\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2). \qquad \textbf{[4.22]}$$

On distingue très clairement dans (4.22) l'introduction de manière additive des variances respectives de $\hat{\beta}_1$ et $\hat{\beta}_2$ ainsi que la soustraction à deux reprises de leur covariance. L'écart-type découle directement de ce calcul en prenant la racine carrée de la variance. Pour pouvoir procéder au calcul il nous faut des estimateurs de chacune de ces grandeurs. Comme $[\hat{\sigma}(\hat{\beta}_1)]^2$ et $[\hat{\sigma}(\hat{\beta}_2)]^2$ sont des estimateurs sans biais de $\mathrm{Var}(\hat{\beta}_1)$ et $\mathrm{Var}(\hat{\beta}_2)$, nous utiliserons l'expression suivante :

$$\hat{\sigma}(\hat{\beta}_1 - \hat{\beta}_2) = \{[\hat{\sigma}(\hat{\beta}_1)]^2 + [\hat{\sigma}(\hat{\beta}_2)]^2 - 2s_{12}\}^{1/2} \qquad \textbf{[4.23]}$$

où s_{12} désigne un estimateur de la covariance $Cov(\hat{\beta}_1, \hat{\beta}_2)$. Nous n'avons pour l'instant pas montré la formule de $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)$. Certains logiciels renvoient les valeurs de covariance, s_{12}, auquel cas il est possible de procéder

au calcul de l'écart-type (4.23), puis de la statistique de Student (4.20). L'annexe E détaille comment utiliser les outils d'algèbre linéaire pour obtenir s_{12}.

Certains logiciels économétriques proposent des routines incorporant des options avancées propres à l'implémentation de tests de restrictions linéaires. Ici, nous nous contenterons de présenter l'approche qui peut être mise en œuvre avec la plupart des logiciels. Plutôt que de chercher à calculer $\hat{\sigma}(\hat{\beta}_1 - \hat{\beta}_2)$ à partir de (4.23), il est beaucoup plus simple d'estimer un modèle différent permettant de revenir au cas de figure connu du test sur un seul paramètre. Pour ce faire, nous définissons un nouveau paramètre dont la valeur correspond à l'écart entre β_1 et β_2 : $\theta_1 = \beta_1 - \beta_2$. À partir de cette réécriture, nous pouvons reformuler nos hypothèses de test comme suit :

$$H_0 : \theta_1 = 0 \text{ contre } H_1 : \theta_1 < 0 \qquad \textbf{[4.24]}$$

La statistique de Student dans (4.20) associée au paramètre $\hat{\theta}_1$ est simplement $t = \hat{\theta}_1/\hat{\sigma}(\hat{\theta}_1)$. La seule difficulté ici tient à l'estimation de la valeur de $\sigma(\hat{\theta}_1)$.

Nous pouvons l'obtenir en réécrivant le modèle de telle sorte à faire apparaitre θ_1. Comme $\theta_1 = \beta_1 - \beta_2$, nous pouvons également écrire que $\beta_1 = \theta_1 + \beta_2$. En remplaçant β_1 par $\theta_1 + \beta_2$ dans l'expression (4.22) et en réarrangeant quelque peu l'expression, on obtient l'équation suivante :

$$\begin{aligned} \log(wage) &= \beta_0 + (\theta_1 + \beta_2)jc + \beta_2 univ + \beta_3 exper + u \\ &= \beta_0 + \theta_1 jc + \beta_2(jc + univ) + \beta_3 exper + u. \end{aligned} \qquad \textbf{[4.25]}$$

Le principal résultat de cette opération est que le paramètre d'intérêt, θ_1, intervient maintenant dans l'équation en étant rattaché à la variable indépendante *jc*. La constante du modèle reste elle la même, β_0. Il n'y a pas de changement non plus concernant la variable *exper* qui est toujours associée au paramètre β_3. Le seul changement notable suite à l'apparition du paramètre θ_1 dans le modèle, concerne l'introduction d'une nouvelle variable, *jc + univ*, associée au paramètre β_2. Par conséquent, si l'on souhaite directement estimer θ_1 afin d'obtenir l'écart-type $\hat{\theta}_1$, il est nécessaire d'introduire une nouvelle variable *jc + univ* dans le modèle. Cette variable viendra se substituer à la variable *univ* du modèle original. Dans notre exemple, la nouvelle variable a une interprétation directe : c'est le nombre d'années total passées dans le supérieur. On appelle *totcoll = jc + univ*. Le nouveau modèle s'écrira alors :

$$\log(wage) = \beta_0 + \theta_1 jc + \beta_2 totcoll + \beta_3 exper + u. \qquad \textbf{[4.26]}$$

Le paramètre β_1 a disparu du modèle, alors que θ_1 y apparait explicitement. Le modèle reste le même que l'original, en étant simplement écrit différemment. La seule raison pour laquelle nous avons procédé à cette réécriture du modèle est qu'elle permet d'associer le coefficient $\hat{\theta}_1$, à la variable explicative *jc* et par celà, obtenir directement la valeur de $\hat{\sigma}(\hat{\theta}_1)$. En procédant de la sorte, la statistique de Student que nous souhaitions obtenir précédemment peut nous être fournie directement par n'importe quel logiciel économétrique.

Nous réalisons l'estimation sur les 6 763 observations utilisées précédemment. Le résultat est alors le suivant :

$$\begin{gathered} \widehat{\log(wage)} = 1{,}472 - 0{,}0102\, jc + 0{,}769\, totcoll + 0{,}0049\, exper \\ (0{,}021)\ (0{,}0069)\quad (0{,}0023)\qquad (0{,}0002) \\ n = 6\ 763,\ R^2 = 0{,}222. \end{gathered} \qquad \textbf{[4,27]}$$

Le seul chiffre de cette équation que nous ne pouvions obtenir à partir de (4.21) est l'écart-type de $\hat{\theta}_1$. Les valeurs estimées du paramètre et de l'écart-type étant de $-0{,}0102$ et $0{,}0069$ respectivement, la statistique de Student pour tester (4.18) se calcule directement, $-0{,}0102/0{,}0069 = -1{,}48$. Pour un test unilatéral avec

pour hypothèse alternative (4.19), la *p*-valeur se situe aux alentours de 0,070. Nous pouvons donc rejeter l'hypothèse nulle (4.18) contre l'hypothèse alternative (4.19) au seuil de significativité de 10 %.

Notons que ni la valeur de la constante, ni celle du coefficient estimé attaché à la variable *exper* n'ont changé par rapport à l'équation (4.21). Il en va de même pour les écarts-types estimés qui leur sont associés. Ce résultat doit être vrai pour toute procédure d'estimation similaire. Cette comparaison permet de vérifier que la transformation a bien été faite correctement. Par ailleurs, on peut également noter que les coefficient et écart-type associés à *totcoll*, sont identiques à ceux d'*univ* dans l'équation (4.21). Nous savons qu'il doit en être ainsi en comparant les expressions (4.17) et (4.25).

Le calcul des intervalles de confiance autour de θ_1, ne revêt pas de difficulté particulière. Ainsi, si l'on utilise l'approximation de la loi normale standardisée, l'intervalle de confiance pour un niveau de confiance de 95 % sera $\hat{\theta}_1 \pm 1{,}96\hat{\sigma}(\hat{\theta}_1)$ c'est-à-dire $-0{,}0102 \pm 0{,}0135$.

L'approche consistant à réécrire le modèle de départ afin de faire apparaitre le paramètre d'intérêt dans l'équation estimée, présente l'avantage de pouvoir être appliquée dans tous les cas et d'être par ailleurs très simple à mettre en œuvre (pour aller plus loin, voir les exercices sur ordinateur C1 et C3 pour d'autres exemples).

4.5 TESTER DES RESTRICTIONS LINÉAIRES MULTIPLES : LE TEST DE FISHER

La statistique de Student associée à un estimateur par la méthode des MCO peut être utilisée afin de tester si le paramètre inconnu de la population correspondant est égal à une valeur donnée (qui est en générale, mais pas toujours, zéro). Nous avons vu par ailleurs dans la section précédente comment tester une hypothèse pour une combinaison linéaire de paramètres β_j en réarrangeant l'équation de départ et en estimant le modèle avec les variables transformées. Toutefois, jusqu'à maintenant, nous nous sommes contentés de couvrir des cas impliquant une seule restriction. Il arrive fréquemment cependant, que l'on ait besoin de tester plusieurs hypothèses sur les paramètres de la population β_0, β_1, ..., β_k. Nous commençons par traiter le cas le plus courant suivant lequel on cherche à tester si l'ensemble des variables explicatives n'ont pas d'effet marginal sur la variable dépendante.

Tester les restrictions d'exclusion

Nous savons déjà comment procéder pour savoir si une variable particulière à un effet marginal ou non sur la variable dépendante : nous utilisons la statistique de Student. Dans le cas présent, nous souhaitons tester si un groupe de variables prises ensemble n'a pas d'effet sur la variable dépendante. Plus précisément, l'hypothèse nulle est qu'un ensemble de variables n'a pas d'effet sur *y*, une fois que l'on a pris en compte l'influence de facteurs tiers.

Pour illustrer cette question, considérons le modèle suivant qui explique les salaires des joueurs de baseball de la ligue professionnelle américaine :

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u, \qquad \textbf{[4.28]}$$

où *salary* est le salaire total en 1993, *years* le nombre d'années passées au sein de la ligue, *gamesur* le nombre moyen de rencontres disputées dans l'année, *bavg* la « moyenne à la batte » (ou "*batting average*" en anglais) sur l'ensemble de la carrière du joueur (par exemple, *bavg* = 250), *hrunsyr* le nombre de "*home runs*" par an, et enfin *rbisyr* le nombre de points produits par un frappeur au cours d'une année. Supposons que l'on souhaite tester l'hypothèse nulle selon laquelle, une fois avoir pris en compte l'influence du nombre d'années au sein de la ligue de baseball ainsi que le nombre de rencontres disputées par an, les statistiques

de mesure de performance usuelles, *bavg*, *hrunsyr*, et *rbisyr*, n'ont pas d'effet sur le salaire. En somme, cette hypothèse nulle nous dit que la productivité des joueurs telle que mesurée par les statistiques traditionnelles n'a pas d'effet sur le salaire.

Formellement, l'hypothèse nulle peut s'écrire comme suit :

$$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0. \qquad \textbf{[4.29]}$$

L'hypothèse nulle (4.29) implique trois **restrictions d'exclusion** : si (4.29) est vraie, alors *bavg*, *hrunsyr*, et *rbisyr* n'ont pas d'effet sur *log*(*salary*) après avoir pris en compte *years* et *gamesyr*. Cet exemple illustre la question des **restrictions multiples** car nous imposons plus d'une restriction sur la valeur des paramètres du modèle (4.28). Nous verrons plus tard des exemples plus généraux de restrictions multiples sur les paramètres. Un test imposant des restrictions multiples est appelé **test d'hypothèses multiples** ou **test d'hypothèses jointes**.

Quelle peut être l'hypothèse alternative à (4.29) ? Si la théorie économique ou notre intuition personnelle nous amène à penser que les statistiques sur les performances des joueurs jouent un rôle dans la fixation de leur salaire, même après avoir tenu compte de l'influence de l'expérience au sein de la ligue et du nombre de rencontres disputées par an, alors l'hypothèse alternative la plus appropriée est simplement donnée par :

$$H_1 : H_0 \text{ n'est pas vraie.} \qquad \textbf{[4.30]}$$

L'hypothèse alternative (4.30) tient si au moins un des paramètres β_3, β_4, ou β_5 est différent de zéro (l'hypothèse nulle sera rejetée dès qu'un seul des trois paramètres sera différent de zéro). Le test que nous étudions ici est construit de manière à détecter toute violation de l'hypothèse H_0. Il est également valide lorsque l'hypothèse alternative est telle que $H_1 : \beta_3 > 0$, ou $\beta_4 > 0$, ou encore $\beta_5 > 0$. Toutefois, d'autres tests sont plus performants sous ces hypothèses alternatives. Nous n'avons pas la place ici ni les bases suffisantes en statistiques pour aborder les tests ayant plus de puissance sous l'hypothèse alternative multiple sur une queue de distribution.

Comment doit-on s'y prendre pour procéder au test de (4.29) contre (4.30) ? Il serait tentant de tester (4.29) en utilisant la statistique de Student sur les différentes variables d'intérêt, *bavg*, *hrunsyr*, et *rbisyr* dans le but de déterminer si chaque variable prise individuellement est statistiquement significative. Cette démarche n'est toutefois pas appropriée. En effet, aucune restriction n'est imposée sur les autres paramètres lors du calcul de la statistique de Student. En procédant de la sorte par ailleurs, nous aurions trois résultats à prendre en compte pour prendre notre décision – un résultat par statistique de Student.

Quelle règle de décision doit dès lors être adoptée pour tester (4.29) au seuil de 5 % par exemple ? Les trois statistiques de Student doivent-elles être conjointement significatives à 5 % ? Ou peut-on se contenter d'une statistique de Student significative pour arrêter notre conclusion ? Ces questions sont difficiles à résoudre, mais nous verrons dans un instant que nous n'avons heureusement pas besoin d'y répondre. Par ailleurs, tester séparément la significativité des différents coefficients en utilisant la statistique de Student peut conduire à des conclusions erronées. Afin de répondre à notre question de départ, il nous faut trouver un moyen de tester conjointement les restrictions d'exclusion.

Pour illustrer cette question, nous estimons l'équation (4.28) en utilisant les données du fichier MLB1. Nous obtenons les résultats suivants :

$$\begin{aligned} \widehat{\log(salary)} &= 11{,}19 + 0{,}0689\, years + 0{,}0126\, gamesyr \\ &\qquad (0{,}29)\ \ (0{,}0121)\qquad\ (0{,}0026) \\ &\quad + 0{,}00098\, bavg + 0{,}0144\, hrunsyr + 0{,}0108\, rbisyr \\ &\qquad (0{,}00110)\qquad\ (0{,}0161)\qquad\ (0{,}0072) \\ n &= 353,\ \text{SCR} = 183{,}186,\ R^2 = 0{,}6278 \end{aligned} \qquad \textbf{[4.31]}$$

où SCR est la somme des carrés des résidus (nous utiliserons cette information plus tard). Les valeurs du SCR et du R-carré sont reportées avec plusieurs chiffres après la virgule pour faciliter les comparaisons. L'équation (4.32) révèle que si les variables *years* et *gamesyr* sont statistiquement significatives aux seuils usuels, aucune autre variable parmi *bavg*, *hrunsyr*, et *rbisyr* n'est significativement différente de zéro au seuil de 5 % (on notera toutefois que la variable *rbisyr* est proche du seuil de significativité, sa *p*-valeur pour un test bilatéral étant de 0,134). Sur base des trois statistiques de Student, il semble que nous ne puissions pas rejeter H_0.

Cette conclusion s'avère fausse cependant. Pour s'en rendre compte, nous devons dériver un test de restrictions multiples dont la distribution est connue et les valeurs tabulées. La somme des carrés des résidus nous fournit dans ce cadre une information précieuse pour procéder au calcul de la statistique de test. Nous allons également montrer que l'indice du R-carré peut être utilisé dans certains cas pour réaliser des tests de restriction.

La connaissance de la somme des carrés des résidus de (4.31) ne nous dit rien en soit sur l'hypothèse (4.29). En revanche, celle de la hausse de la SCR consécutive à l'exclusion des variables *bavg*, *hrunsyr*, et *rbisyr* sera précieuse pour notre analyse. Souvenons-nous que, les estimateurs des MCO étant choisis pour minimiser la somme des carrés des résidus, la SCR augmente toujours lorsque certaines des variables du modèle sont exclues. La question qui se pose est alors de savoir si cette augmentation est suffisante, par rapport à la SCR du modèle incluant toutes les variables, pour nous conduire à rejeter l'hypothèse nulle.

Le modèle sans les trois variables en question est donné par :

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + u. \qquad \textbf{[4.32]}$$

Dans le cadre des tests d'hypothèses, l'équation (4.32) est qualifiée de **modèle contraint** pour le test (4.29) ; le modèle (4.28) est quant à lui qualifié de **modèle non contraint**. Le modèle contraint a toujours moins de paramètres à estimer que le modèle non contraint.

Si l'on estime le modèle contraint avec les données contenues dans le fichier MLB1, on obtient les résultats suivants :

$$\widehat{\log(salary)} = 11,22 + 0,0713\, years + 0,0202\, gamesyr$$

$$(0,11)\ \ (0,0125)\qquad (0,0013)$$

$$n = 353,\ \text{SCR} = 198,311\ R^2 = 0,5971 \qquad \textbf{[4.33]}$$

En résumé, la SCR de (4.33) est supérieure à celle de (4.31). De même, la valeur du R-carré du modèle contraint est inférieure à celle du modèle non contraint. Ce qu'il nous reste à décider maintenant est de savoir si l'augmentation constatée du SCR en passant du modèle non contraint au modèle contraint (183,186 à 198,311) est suffisante pour nous conduire au rejet de (4.29). À l'instar des autres tests d'hypothèses, la réponse dépend du seuil de significativité choisi. Toutefois, il ne nous est pas possible de mettre en œuvre un test d'hypothèse pour un seuil donné, en l'absence de statistique de test pour laquelle la distribution sous l'hypothèse nulle serait connue et les valeurs tabulées. Nous devons trouver un moyen d'utiliser l'information contenue dans les deux SCR pour obtenir une statistique de test dont la distribution statistique serait connue sous l'hypothèse nulle.

Nous allons maintenant dériver le test dans le cas général. Le modèle non contraint comprenant k variables indépendantes s'écrit comme suit :

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u\ ; \qquad \textbf{[4.34]}$$

le modèle non contraint possède $k + 1$ paramètres à estimer (rappelez-vous qu'il convient d'ajouter un paramètre correspondant à la constante). Supposons maintenant que q variables soient exclues du modèle afin de tester l'hypothèse nulle selon laquelle les coefficients associés aux q variables de (4.34) sont égaux à zéro. Pour

simplifier les notations, nous supposerons que les q variables testées sont les dernières de la liste $x_{k-q+1}, \ldots, x_k$. (l'ordre des variables est évidement arbitraire et sans incidence sur le résultat du test). L'hypothèse nulle s'écrira comme suit :

$$H_0 : \beta_{k-q+1} = 0, \ldots, \beta_k = 0, \qquad \textbf{[4.35]}$$

(4.35) impose q restrictions d'exclusion sur le modèle (4.34). L'hypothèse alternative à (4.35) est simplement que H_0 soit fausse ; ce qui signifie qu'au moins un des paramètres listés dans (4.35) est différent de zéro. Lorsque nous imposons les restrictions de l'hypothèse nulle, nous obtenons le modèle suivant :

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-q} x_{k-q} + u. \qquad \textbf{[4.36]}$$

Dans cette sous section, nous faisons l'hypothèse que les modèles contraint et non contraint contiennent tous deux une constante, comme c'est en général le cas dans la plupart des applications.

Intéressons nous maintenant à la statistique de test elle même. Nous suggérions précédemment que l'étude de l'augmentation relative du SCR lors du passage du modèle non contraint au modèle contraint était un bon indicateur pour tester l'hypothèse (4.35). La statistique de Fisher ou **statistique *F***, se définit comme suit :

$$F \equiv \frac{(SCR_c - SCR_{nc})/q}{SCR_{nc}/(n-k-1)} \qquad \textbf{[4.37]}$$

où SCR_c est la somme des carrés des résidus du modèle contraint et SCR_{nc} la somme des carrés des résidus du modèle non contraint.

Pour aller plus loin 4.4

Essayons de relier la réussite aux épreuves scolaires standardisées, *score*, à un ensemble d'autres variables. Les facteurs liés à l'environnement scolaire comprennent le nombre moyen d'élèves par classe, les dépenses moyennes par élève, la rémunération moyenne des enseignants, et le nombre total d'élèves dans l'établissement scolaire. D'autres variables ont trait à des caractéristiques plus spécifiques de l'élève comme le revenu de la famille, le niveau d'éducation de la mère, le niveau d'éducation du père et le nombre de frères et soeurs. Le modèle s'écrit comme suit :

$$score = \beta_0 + \beta_1 classize + \beta_2 expend + \beta_3 tchcomp + \beta_4 enroll + \beta_5 faminc + \beta_6 motheduc + \beta_7 fatheduc + \beta_8 siblings + u.$$

Formulez l'hypothèse nulle suivant laquelle les variables spécifiques à l'élève n'ont pas d'impact sur les résultats aux épreuves standardisées une fois prise en compte l'influence des facteurs liés à l'environnement scolaire. À quoi correspondent k et q dans cet exemple ? Écrivez l'équation du modèle contraint.

Il apparaît immédiatement que puisque SCR_c ne peut pas être plus petit que SCR_{nc}, la statistique F doit toujours être non négative (et quasiment toujours strictement positive). Par conséquent, si vous obtenez une valeur négative à l'issue du calcul de la statistique F c'est que quelque chose n'est pas correct ; en général il s'agit simplement d'une inversion de l'ordre des SCR qui se trouvent au numérateur. Par ailleurs, la SCR du dénominateur est celle du modèle non contraint. Le moyen le plus simple de se souvenir de la place de chaque SCR est de voir la statistique F comme une mesure de l'accroissement relatif de la SCR lors du passage du modèle non contraint vers le modèle contraint.

La différence entre les SCR au numérateur de la statistique F est divisée par q, qui représente le nombre de restrictions imposées lors du passage du modèle non contraint au modèle contraint (q variables indépendantes sont supprimées du modèle). Par conséquent, nous pouvons écrire :

$$q = \textbf{nombre de degrés de liberté au numérateur} = ddl_c - ddl_{nc}, \qquad \textbf{[4.38]}$$

cette expression montre également que q est la différence de degrés de liberté entre les modèles contraint et non contraint (pour rappel *ddl* = nombre d'observations – nombre de paramètres estimés). Puisque le modèle contraint a moins de paramètres et que chaque modèle est estimé avec le même nombre d'observations – ddl_c est toujours supérieur à ddl_{nc}.

La SCR au dénominateur de la statistique F est divisée par le nombre de degrés de liberté du modèle non contraint :

$$n - k - 1 = \textbf{nombre de degrés de liberté au dénominateur} = ddl_{nc} \qquad \textbf{[4.39]}$$

Le dénominateur de la statistique F est simplement l'estimateur sans biais de $\sigma^2 = \text{Var}(u)$ du modèle non contraint.

En pratique, le calcul de la statistique F est plus simple que ce qu'il y paraît. Il convient dans un premier temps d'obtenir le nombre de degrés de liberté du modèle non contraint, ddl_{nc}, puis, de dénombrer les variables exclues du modèle contraint, soit q. Les SCR étant reportées dans tous tableaux de sortie d'une estimation par les MCO, il ne reste qu'à appliquer la formule pour calculer la statistique F recherchée.

Dans le modèle de régression sur les salaires des joueurs de baseball professionnels, $n = 353$, et le modèle (4.28) contient six paramètres. Ainsi, $n - k - 1 = ddl_{nc} = 353 - 6 = 347$. Le modèle contraint (4.32) contient quant à lui trois variables indépendantes de moins (4.28). Par conséquent, $q = 3$. Nous avons là tous les ingrédients pour calculer la statistique F ; nous nous gardons de le faire pour l'instant, ne sachant pas encore comment l'interpréter.

Pour pouvoir utiliser la statistique F, nous devons connaitre la distribution d'échantillonnage sous l'hypothèse nulle. Ceci nous est nécessaire pour obtenir les valeurs critiques et définir la règle de décision. Il est possible de montrer que, sous H_0 (supposant les hypothèses du modèle linéaire classique vérifiées), la statistique F est distribuée selon une loi de Fisher à $(q, n - k - 1)$ degrés de liberté. Nous pouvons l'écrire comme suit :

$$F \sim F_{q,n-k-1}.$$

Les valeurs de la distribution de $F_{q,n-k-1}$ sont tabulées et peuvent facilement s'obtenir en consultant les tables statistiques (voir la table G.3), mais également *via* les logiciels économétriques traditionnels.

Nous ne dériverons pas ici la distribution de la statistique F de manière formelle car ceci demande de nombreuses manipulations mathématiques. Nous noterons simplement qu'il peut être démontré que l'équation (4.37) est le ratio de deux variables aléatoires indépendantes suivant chacune une distribution du chi-deux, divisée par leurs degrés de liberté respectifs. La variable aléatoire au numérateur suit donc une distribution du chi-deux à q degrés de liberté et celle au dénominateur un chi-deux à $n - k - 1$ degrés de liberté. Il s'agit bien là de la définition d'une variable aléatoire suivant une distribution de Fisher (voir l'annexe B).

Il ressort relativement clairement de la définition de F que nous rejetons H_0 en faveur de H_1 lorsque la statistique F est suffisamment « grande ». Ce que l'on entend par « suffisamment grand » dépendra du seuil de significativité fixé. Supposons que nous décidions d'adopter un seuil à 5 %. Soit c le 95e centile de la distribution $F_{q,n-k-1}$. La valeur critique dépend de q (le nombre de *ddl* du numérateur) et de $n - k - 1$ (le nombre de *ddl* du dénominateur). Il est important de garder le nombre de degrés de liberté au numérateur et au dénominateur explicite dans le calcul de la statistique.

Les valeurs critiques aux seuils de 10 %, 5 % et 1 % de la distribution de Fisher sont données dans la table statistique G.3. La règle de décision appliquée reste simple. Une fois c connu, nous rejetons l'hypothèse H_0 en faveur de H_1 au seuil de significativité retenu si :

$$F > c. \qquad \textbf{[4.40]}$$

À un seuil de 5 %, $q = 3$ et $n - k - 1 = 60$, la valeur critique est $c = 2{,}79$. Nous rejetons par conséquent H_0 au seuil de 5 % si la statistique F calculée est supérieure à 2,76. La valeur critique à 5 % ainsi que la région de rejet sont représentées à la figure 4.7. Pour le même nombre de degrés de liberté, la valeur est de 4,13 pour un seuil de 1 %.

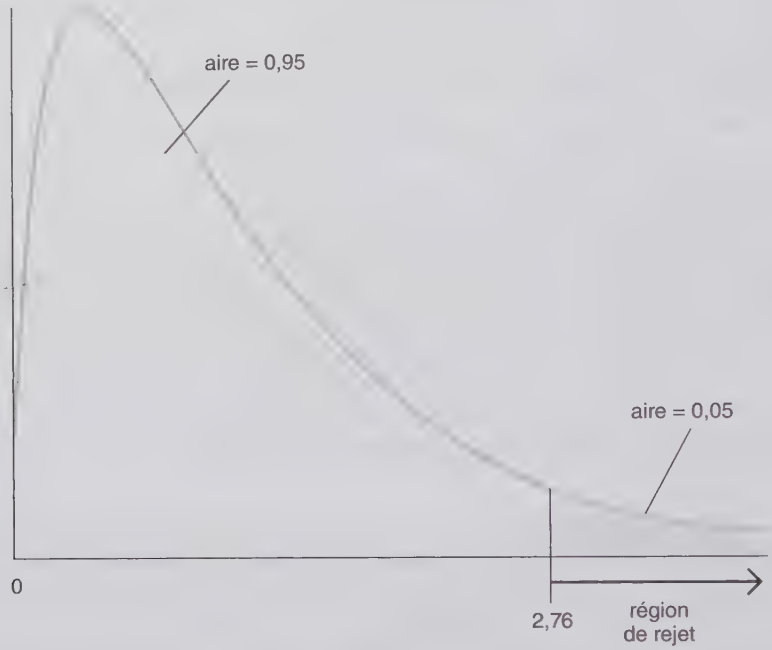


Figure 4.7 *La valeur critique au seuil de 5 % et la région de rejet d'une distribution de* $F_{3,60}$.

Dans la plupart des applications, le nombre de degrés de liberté au numérateur (q) est significativement plus petit que celui du dénominateur ($n - k - 1$). Les études empiriques dans lesquelles $n - k - 1$ est petit ont peu de chance de fournir des résultats intéressants du fait du manque de précision des estimateurs du modèle non contraint. Lorsque le nombre de degrés de liberté dépasse 120, la distribution de la statistique F n'est alors plus sensible au nombre d'observations (ce résultat est similaire à celui obtenu pour la statistique de Student qui est bien approximée par la distribution normale standardisée lorsque le nombre de degrés de liberté est important). Ainsi, il existe une colonne dans la table pour le dénominateur lorsque $ddl = \infty$. Cette colonne est utilisée en présence de grands échantillons (lorsque $n - k - 1$ est grand). Ceci est également vrai lorsque le nombre de degrés de liberté du numérateur est important, toutefois, ce cas de figure se présente plus rarement dans les applications empiriques.

Lorsque l'hypothèse nulle, H_0, peut être rejetée, nous disons que $x_{k-q+1}, \ldots, x_k$ sont **conjointement statistiquement significatives** (ou bien simplement *conjointement significatives*) au seuil de significativité adéquat. Ce test seul ne nous permet pas de dire quelle variable a un effet marginal sur y ; il se peut qu'elles affectent toutes y ou bien qu'une seule d'entre elle soit significative. Si l'hypothèse nulle est rejetée, les variables sont dites **conjointement non significatives**, impliquant souvent leur exclusion du modèle.

Dans le cas de l'exemple précédent sur les salaires des joueurs de baseball professionnels, la valeur critique pour un nombre de degrés de liberté de 3 au numérateur et de 347 au dénominateur, est de 2,60 au seuil de 5 % et de 3,78 au seuil de 1 %. La règle de décision, nous amène donc à rejeter H_0 au seuil de 1 % si la statistique F est supérieure à 3,78 ; et à faire de même au seuil de 5 % si F est supérieure à 2,60.

Nous sommes maintenant en mesure de tester l'hypothèse énoncée au début de cette section : les variables *bavg*, *hrunsyr*, et *rbisyr* n'ont-elles conjointement aucune influence sur le salaire des joueurs une fois que l'on a tenu compte de l'influence des variables *years* et *gamesyr* ? En pratique, il est plus simple de calculer $(SCR_c - SCR_{nc})/SCR_{nc}$ dans un premier temps puis de multiplier le résultat par $(n - k - 1)/q$; la raison pour laquelle la formule est exprimée comme dans (4.37) est simplement que ceci permet de garder les nombres de degrés de liberté au numérateur et au dénominateur explicites. En utilisant la SCR dans (4.31) et (4.33), nous obtenons :

$$F = \frac{(198{,}311 - 183{,}186)}{183{,}186} \cdot \frac{347}{3} \approx 9{,}55$$

La valeur obtenue est bien supérieure à la valeur critique de 3,78 au seuil de 1 % pour une distribution F à 3 et 327 degrés de liberté. Nous pouvons donc rejeter avec confiance l'hypothèse stipulant que les variables *bavg*, *hrunsyr*, et *rbisyr* n'ont pas d'effet conjoint sur le salaire.

Le résultat du test joint peut sembler surprenant à la lumière des tests individuels de significativité pour les trois variables qui nous conduisaient à ne pas rejeter l'hypothèse nulle. L'explication de ce résultat contradictoire tient au fort taux de corrélation entre les variables *hrunsyr* et *rbisyr*. Ainsi, la présence de multicolinéarité entre les variables explicatives rend difficile l'identification des effets partiels ; ceci se reflète dans la valeur des écarts-types et des statistiques de Student individuelles. La statistique F teste si ces variables sont conjointement significatives, la multicolinéarité entre *hrunsyr* et *rbisyr* n'entre dès lors plus en ligne de compte. Dans l'exercice 16, il vous est demandé de ré-estimer le modèle en excluant *rbisyr*. La variable *hrunsyr* devient alors significative. L'inverse est également vrai lorsque *hrunsyr* est exclue du modèle à la place de *rbisyr*.

La statistique F est très utile pour tester la significativité d'un groupe de variables fortement corrélées entre elles. Par exemple, supposons que nous souhaitions tester si la rémunération d'un directeur général dépend des performances de son entreprise. Il existe plusieurs indicateurs potentiels pour mesurer la performance d'une entreprise. Il est difficile en revanche de savoir *a priori* lequel est le plus pertinent pour expliquer les niveaux de rémunération. Ces indicateurs étant probablement très fortement corrélés, les tests de significativité individuelle ne permettront sans doute pas d'identifier des indicateurs pertinents. Un test F peut dans ce cas être utilisé pour déterminer si, prises ensemble, les variables capturant les performances des entreprises influencent le salaire.

Liens entre les statistiques de Fisher et de Student

Nous avons vu dans cette section comment utiliser la statistique F pour tester si un groupe de variables doit être inclus dans un modèle. La question que nous nous posons maintenant, dans la mesure où les précédents développements ne l'interdisent pas, est de savoir ce qu'il se passe lorsque la statistique F est appliquée à une seule variable explicative. Par exemple, nous pouvons poser l'hypothèse nulle suivante : $H_0 : \beta_k = 0$ et $q = 1$ (afin d'effectuer un test de restriction portant uniquement sur la variable x_k). Depuis la section 4.2, nous savons que la statistique de Student associée permet de réaliser ce test. La question est donc de savoir maintenant s'il existe une manière différente de procéder à un test sur un unique coefficient. La réponse est négative. Il est ainsi possible de montrer que la statistique F pour tester l'exclusion d'une variable est égale au carré de la statistique de Student correspondante. Puisque la distribution t^2_{n-k-1} est équivalente à une distribution $F_{1,n-k-1}$, les deux approches mènent exactement au même résultat, pour un test bilatéral. La statistique de Student néanmoins reste plus flexible pour effectuer un test sur un seul paramètre puisqu'elle permet de considérer des hypothèses alternatives bilatérale et unilatérale. Les statistiques de Student étant également plus simples à obtenir que les statistiques de Fisher, il n'y a dès lors pas de raison de favoriser l'emploi de cette dernière pour des tests sur un unique paramètre.

Nous avons vu précédemment dans le cas de l'exemple sur les déterminants du salaire des joueurs de baseball professionnels que deux (ou plusieurs) variables jugées non significatives sur base du test de Student pouvaient être conjointement très significatives. Il est également possible de conclure qu'un groupe de variables ne sont pas significatives alors qu'une des variables de ce groupe est associée à une statistique de Student impliquant sa significativité. Quelle conclusion tirer de ces résultats contrastés ? Pour être plus précis, supposons que dans un modèle à plusieurs variables nous ne puissions pas rejeter l'hypothèse nulle suivant laquelle β_1, β_2, β_3, β_4, et β_5 sont égaux à zéro au seuil de 5 %, mais que la statistique de Student de $\hat{\beta}_1$ indique elle que ce paramètre est significatif à 5 %. Logiquement nous ne devrions pas avoir à la fois que $\beta_1 \neq 0$ et que β_1, β_2, β_3, β_4 et β_5 sont tous nuls ! En statistiques cependant, il est possible d'associer un ensemble de variables non significatives avec une variable individuellement significative et de conclure que l'ensemble des variables sont conjointement non significatives (ce nouvel exemple de contradiction entre les conclusions du test de Student et du test de Fisher renforce l'idée suivant laquelle nous ne pouvons « accepter » l'hypothèse nulle ; mais simplement échouer à la rejeter). La statistique de Fisher vise à détecter si un ensemble de coefficients est différent de zéro, elle n'est toutefois pas la plus adaptée pour déterminer si un paramètre unique est différent de zéro. Le test de Student offre dans ce cas plus de flexibilité (d'un point de vue statistique, on dira qu'une statistique F pour tester $\beta_1 = 0$ a moins de puissance pour détecter $\beta_1 \neq 0$ qu'un test de Student usuel. Voir la section C.6 dans l'annexe C pour un complément d'information sur la puissance d'un test).

Malheureusement, le fait de pouvoir cacher la présence d'une variable statistiquement significative au sein d'un ensemble de variables non significatives peut mener à des abus si les résultats des régressions ne sont pas soigneusement reportés. Par exemple, supposons que, dans une étude sur les déterminants du taux d'octroi des prêts à l'échelle de la ville, x_1 est la part des ménages africain-américains de la ville. Supposons que les variables x_2, x_3, x_4, et x_5 reprennent la part des ménages dirigés par des personnes de différentes catégories d'âge. Afin de compléter le modèle pour expliquer le taux d'octroi des prêts, nous pouvons inclure des variables relatives au revenu du ménage, à sa richesse, à son score de risque bancaire, etc. Même si l'origine ethnique a un effet marginal, il se peut que les variables capturant l'origine ethnique et l'âge du chef de famille soient conjointement non significatives. Une personne souhaitant démontrer que l'origine ethnique n'influence pas le taux d'octroi des prêts aura tendance à reporter les résultats comme suit : « L'origine ethnique et l'âge du chef de famille ont été inclus dans l'équation, mais elles ne sont pas conjointement significatives au seuil de 5 % ». Heureusement, l'évaluation par les paires des études scientifiques évite ces types de conclusions trompeuses, il est important toutefois d'être conscient que ce type de manipulation est possible.

Souvent, des variables très significatives sont également testées conjointement avec d'autres variables du modèle, et le test conclut à la significativité jointe des variables. Dans ce cas, il n'est pas dérangeant de rejeter les deux hypothèses nulles.

La formulation R-carré de la statistique de Fisher

Pour tester les restrictions d'exclusion, il est souvent plus simple d'utiliser la forme de la statistique F qui dépend des coefficients de détermination R-carré des modèles contraint et non contraint. La raison principale tient au fait que le R-carré est borné entre zéro et un, alors que les SCR peuvent prendre des valeurs très grandes suivant l'échelle utilisée pour mesurer la variable dépendante, y. Faire les calculs avec les SCR est donc simplement plus fastidieux. L'expression de la statistique F (4.37) peut se réécrire en utilisant $SCR_c = SCT(1 - R_c^2)$ et $SCR_{nc} = SCT(1 - R_{nc}^2)$:

$$F = \frac{(R_{nc}^2 - R_c^2)/q}{(1 - R_{nc}^2)/(n - k - 1)} = \frac{(R_{nc}^2 - R_c^2)/q}{(1 - R_{nc}^2)/ddl_{nc}} \qquad \textbf{[4.41]}$$

(notons que cette réécriture permet de simplifier l'expression en supprimant les termes SCT). Cette expression est appelée la **forme R-carré de la statistique *F***. [À ce stade, il convient de garder à l'esprit que bien que l'équation (4.41) soit très pratique pour les tests de restrictions d'exclusion, elle ne peut pas être appliquée pour

tester toutes les restrictions linéaires. Comme nous le verrons au moment de discuter des tests de restrictions linéaires générales, la statistique F écrite en fonction de la somme des carrés des résidus reste parfois nécessaire].

Le coefficient de détermination, R-carré, étant reporté dans la plupart des régressions (à la différence de la SCR), le test d'exclusion d'une variable peut être facilement effectué à partir du R-carré des modèles contraint et non contraint. Lors du calcul de la statistique, une attention particulière doit être portée à l'ordre des R-carrés au numérateur : le R-carré du modèle non contraint arrive en premier [à la différence de la SCR dans (4.37)]. Comme $R^2_{nc} > R^2_c$, nous pouvons de nouveau constater que la statistique F sera toujours positive.

Lors du calcul de la statistique de test, il est important de ne pas mettre deux fois le coefficient de détermination au carré. En effet tous les logiciels reportent le R^2, cette valeur peut donc être directement introduite dans la formule (4.41). Dans l'exemple précédent, sur les salaires des joueurs de baseball, nous pouvons utiliser (4.41) afin d'obtenir la statistique F comme suit :

$$F = \frac{(0,6278 - 0,5971)}{(1 - 0,6278)} \cdot \frac{347}{3} \approx 9,54$$

On notera que cette valeur reste très proche de celle obtenue précédemment (la différence est simplement due à l'erreur d'arrondi).

EXEMPLE 4.9
Impact de l'éducation des parents sur le poids des nouveaux-nés

Pour illustrer à nouveau le calcul de la statistique F au moyen de cette méthode, considérons le modèle expliquant le poids des nouveaux-nés par un ensemble de facteurs :

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + u, \quad \textbf{[4.42]}$$

où

bwght = poids du nouveau-né, en livres
cigs = nombre de cigarettes fumées quotidiennement par la mère durant la grossesse
parity = ordre de naissance de l'enfant
faminc = revenu du foyer annuel
motheduc = nombre d'années d'étude de la mère
fatheduc = nombre d'années d'étude du père

Commençons par tester l'hypothèse nulle selon laquelle l'éducation n'a pas d'effet sur le poids des nouveaux-nés, après avoir pris en compte l'influence de *cigs*, *parity*, et *faminc*. L'hypothèse nulle s'écrit : $H_0 : \beta_4 = 0, \beta_5 = 0$. Il y a $q = 2$ restrictions d'exclusion à tester. Le modèle non contraint (4.42) possède $k + 1 = 6$ paramètres ; le nombre de degrés de liberté associé à ce modèle est donc égal à $n - 6$, avec n le nombre d'observations dans l'échantillon.

Nous allons tester cette hypothèse en utilisant les données contenues dans le fichier BWGHT. Cette base de données contient des informations sur 1 388 naissances, mais nous devons être prudents dans le décompte des observations utilisées pour tester l'hypothèse nulle. Il s'avère que des informations sur au moins une des variables *motheduc* et *fatheduc* manquent pour 197 naissances de l'échantillon ; ces observations ne peuvent par conséquent pas être incluses dans l'estimation du modèle non contraint. Ainsi, nous considérons finalement $n = 1\,191$ observations, menant à $1\,191 - 6 = 1\,185$ *ddl* dans le modèle non contraint. Nous devons nous assurer d'utiliser ces mêmes 1 191 observations pour l'estimation du modèle contraint (et non l'ensemble des 1 388 observations disponibles). En règle générale, lors de l'estimation du modèle contraint pour calculer la statistique F, nous devons utiliser les mêmes observations que dans le cadre de l'estimation du modèle non contraint ; sans cela, le test n'est pas valide. En l'absence de données manquantes, cette condition ne pose pas de problème.

Le *ddl* du numérateur est 2, celui du dénominateur est 1185 ; à partir de la table statistique G.3 nous pouvons déterminer que la valeur critique au seuil de 5 % est $c = 3,0$. Plutôt que de rendre compte de l'ensemble des résultats nous ne présentons, par souci de concision, que les R-carrés. Le R-carré pour le modèle complet atteint $R^2_{nc} = 0,0387$. Lorsque *motheduc* et *fatheduc* sont retirées de la régression, le R-carré chute à $R^2_c = 0,0364$. La statistique F prend pour valeur $F = [(0,0387 - 0,0364) / (1 - 0,0387)] \times (1185/2) = 1,42$. Cette valeur étant bien inférieure à la valeur critique de 5 %, nous ne parvenons pas à rejeter H_0. En d'autres termes, *motheduc* et *fatheduc* sont conjointement non significatifs dans l'équation de poids des nourrissons.

Calcul des p-valeurs pour le test de Fisher

Pour rendre compte des résultats des tests F, les p-valeurs sont particulièrement utiles. Puisque la distribution F dépend au numérateur et au dénominateur du nombre de degrés de liberté, il est difficile de se faire une idée des chances de rejet de l'hypothèse nulle par une simple lecture de la valeur de la statistique F et de l'une ou l'autre des valeurs critiques possibles. Dans le contexte des tests de Fisher, la p-valeur est définie comme

$$p\text{-valeur} = P(\mathcal{F} > F), \qquad \textbf{[4.43]}$$

où, pour simplifier la lecture, nous notons $\mathcal{F}$ une variable aléatoire suivant une distribution de Fisher à $(q, n - k - 1)$ degrés de liberté, et F la valeur de la statistique de test. La p-valeur a toujours la même interprétation que dans le cadre des statistiques de Student : il s'agit de la probabilité d'observer une valeur de F au moins aussi grande que celle obtenue, étant donné que l'hypothèse nulle est vraie. Une faible p-valeur suggère de rejeter H_0. Par exemple, p-valeur = 0,016 signifie que la chance d'observer une valeur de F aussi grande que celle calculée sous l'hypothèse nulle n'est que de 1,6 % ; nous rejetons H_0 généralement dans de tels cas. Si la p-valeur = 0,314, alors la chance d'observer une valeur de la statistique F aussi grande que celle calculée sous l'hypothèse nulle est de 31,4 %. La plupart des chercheurs considèrent cela comme une preuve trop faible pour rejeter H_0.

Pour aller plus loin 4.5

Les données contenues dans le fichier ATTEND ont été utilisées pour estimer les deux équations suivantes :

$$\widehat{atndrte} = 47,13 + 13,37\, priGPA$$
$$(2,87) \quad (1,09)$$
$$n = 680,\ R^2 = 0,183$$

et

$$\widehat{atndrte} = 75,70 + 17,26\, priGPA - 1,72\, ACT$$
$$(3,88) \quad (1,08) \qquad (?)$$
$$n = 680,\ R^2 = 0,291$$

où, comme toujours, les écarts-types estimés sont donnés entre parenthèses ; l'écart-type estimé pour *ACT* est manquant dans la seconde équation. Quelle est la valeur de la statistique de Student pour le coefficient de *ACT* ? (*Indice* : Calculez d'abord la statistique F pour la significativité de *ACT*.)

À l'instar des tests de Student, une fois la p-valeur calculée, le test de Fisher peut être effectué à n'importe quel seuil de significativité. Par exemple, si la p-valeur = 0,024, nous rejetons H_0 au seuil de significativité de 5 % mais pas au seuil de 1 %. La p-valeur pour le test F dans l'exemple 4.9 est de 0,238. Par

conséquent l'hypothèse nulle selon laquelle les paramètres associés à *motheduc* et *fatheduc* sont tous deux égaux à zéro n'est pas rejetée, même au seuil de significativité de 20 %.

Plusieurs logiciels économétriques disposent d'une fonction intégrée permettant de tester des restrictions d'exclusion multiples. Ces logiciels ont plusieurs avantages par rapport aux calculs de statistiques réalisés à la main : ils permettent de faire moins d'erreurs, les *p*-valeurs sont calculées automatiquement, et le problème des données manquantes, comme dans l'exemple 4.9, est géré sans aucun travail supplémentaire de notre part.

De l'usage de la statistique de Fisher pour tester la significativité globale d'un modèle de régression

Un ensemble particulier de restrictions d'exclusion est régulièrement testé par défaut par la plupart des logiciels économétriques. Ces restrictions ont la même interprétation, quel que soit le modèle. Dans le modèle avec k variables indépendantes, nous pouvons écrire l'hypothèse nulle suivante :

$$H_0 : x_1, x_2, \ldots, x_k \text{ ne permettent pas d'expliquer } y$$

Cette hypothèse nulle est, en quelque sorte, très pessimiste. Elle stipule qu'aucune des variables explicatives n'a d'effet sur y. Exprimée en termes de paramètres, l'hypothèse nulle revient à considérer que l'ensemble des paramètres de pente du modèle sont nuls :

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0, \qquad \textbf{[4.44]}$$

l'hypothèse alternative étant qu'au moins l'un des paramètres, β_j est différent de zéro. Une autre manière d'énoncer l'hypothèse nulle est que $H_0 : E(y|x_1, x_2, \ldots, x_k) = E(y)$, de telle sorte que la connaissance des valeurs de $x_1, x_2, \ldots, x_k$ n'affecte pas la valeur attendue de y.

On dénombre k restrictions dans (4.44), et lorsque nous les imposons, nous obtenons le modèle contraint suivant :

$$y = \beta_0 + u ; \qquad \textbf{[4.45]}$$

où *toutes* les variables indépendantes ont été retirées de l'équation. Il suit que, le R-carré issu de l'estimation de l'équation (4.45) est égal à zéro puisqu'aucune variation de y ne peut être expliquée par le modèle en l'absence de variables explicatives. Par conséquent, la statistique de test F pour tester (4.44) s'écrira :

$$\frac{R^2/k}{(1 - R^2)/(n - k - 1)}, \qquad \textbf{[4.46]}$$

où R^2 est tout simplement le R-carré habituel obtenu à partir de la régression de y sur $x_1, x_2, \ldots, x_k$.

La plupart des logiciels économétriques reportent automatiquement la statistique F dans (4.46). Il est dès lors tentant d'utiliser cette statistique pour tester des restrictions d'exclusion générales. Il est toutefois préférable d'éviter de le faire. La statistique F dans (4.41) est utilisée pour les restrictions d'exclusion générales ; elle dépend des R-carrés des modèles contraint et non contraint. La forme particulière de (4.46) n'est valable que pour tester l'exclusion conjointe de toutes les variables indépendantes. On parle parfois de test de **significativité globale du modèle de régression**.

Si nous ne parvenons pas à rejeter (4.44), alors il n'existe aucune preuve que l'une ou l'autre des variables indépendantes contribue à expliquer y. Cela signifie généralement que nous devons chercher d'autres variables pour expliquer y. Dans l'exemple 4.9, la statistique F pour le test (4.44) est d'environ 9,55 avec $k = 5$ et $n - k - 1 = 1185$ *ddl*. La *p*-valeur est de zéro à 10^{-4} près, de sorte que (4.44) est rejetée très fortement. Ainsi, nous concluons que les variables de l'équation visant à modéliser *bwght* expliquent effectivement certaines des variations de *bwght*. Le part de *bwght* expliquée est relativement modeste puisque notre modèle

n'explique que 3,87 % de la variance totale. Pour autant, cette faible valeur du R-carré permet tout de même d'obtenir une statistique F très significative. C'est pourquoi nous devons calculer cette statistique pour tester la signification globale du modèle et non uniquement nous fier à la seule taille du R-carré.

Il arrive parfois que le recours à la statistique F pour tester la pertinence d'ensemble du modèle, et donc la significativité jointe de l'ensemble des paramètres du modèle, constitue le cœur de l'étude empirique. À titre d'exemple, l'exercice 10 vous demande d'utiliser les données disponibles sur les rendements de cours de bourse pour vérifier si les rendements des actions sur un horizon de quatre ans sont prévisibles compte tenu des informations connues seulement en début de période. Sous l'hypothèse d'efficience des marchés, les rendements ne devraient pas être prévisibles ; l'hypothèse nulle est donc précisément (4.44).

Tester des restrictions linéaires générales

Les tests de restrictions d'exclusion sont de loin l'utilisation la plus courante de la statistique de Fisher. Il arrive cependant, que les restrictions qu'implique une théorie soient plus complexes que la simple exclusion de certaines variables indépendantes. La statistique de Fisher reste toutefois là aussi simple à utiliser.

À titre d'exemple, considérons l'équation suivante :

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u, \quad \textbf{[4.47]}$$

où :

$price$ = prix de la maison ;
$assess$ = la valeur attendue du logement (avant la vente de la maison) ;
$lotsize$ = surface de la maison, en pieds ;
$sqrft$ = superficie ;
$bdrms$ = nombre de chambres à coucher.

À supposer que l'on souhaite tester si le prix des logements est évalué de manière rationnelle, une variation de 1 % de l'évaluation de la maison devrait être associée à une variation de 1 % de son prix ; c'est à dire $\beta_1 = 1$. En outre, *lotsize*, *sqrft* et *bdrms* ne devraient pas contribuer à expliquer log(*price*), une fois prise en compte la valeur attendue. Ensemble, ces hypothèses peuvent être formulées comme suit :

$$H_0 : \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0. \quad \textbf{[4.48]}$$

Quatre restrictions doivent être testées ; trois sont des restrictions d'exclusion, ce qui n'est pas le cas de $\beta_1 = 1$. Comment pouvons-nous vérifier cette hypothèse en utilisant la statistique F ?

Comme dans le cas de restrictions d'exclusion, nous estimons le modèle sans restriction, c'est-à-dire selon l'équation (4.47) dans ce cas, puis nous imposons les restrictions dans (4.48) afin d'obtenir le modèle contraint. La seconde étape peut être plus délicate bien qu'il s'agisse ici simplement d'incorporer les restrictions directement dans l'écriture du modèle. Si nous écrivons le modèle (4.47) comme suit :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u, \quad \textbf{[4.49]}$$

le modèle contraint s'écrira lui comme : $y = \beta_0 + x_1 + u$. Pour imposer notre contrainte sur le coefficient de x_1, nous devrons estimer le modèle suivant :

$$y - x_1 = \beta_0 + u. \quad \textbf{[4.50]}$$

Le modèle final se limite donc à une simple constante. La variable dépendante en revanche, diffère de celle de (4.49). La procédure de calcul de la statistique F reste la même : (i) on estime d'abord (4.50), dans le but d'obtenir la SCR (SCR_c), puis on utilise cette information avec la SCR du modèle sans contrainte

(4.49) pour calculer la statistique F (4,37). Nous testons alors $q = 4$ contraintes, pour $n - k - 1 = ddl$ dans le modèle non contraint. La statistique F s'écrit alors $[(SCR_c - SCR_{nc}) / SCR_{nc}]\ [(n - 5) / 4]$.

Avant d'illustrer ce test sur données réelles, nous insistons sur un point important : nous ne pouvons pas utiliser la formulation R-carré de la statistique F dans cet exemple car la variable dépendante dans (4.50) n'est pas la même que dans (4.49). Ceci signifie que les sommes des carrés totaux des deux régressions sont différentes, et donc que l'équation (4.41) n'est plus équivalente à (4.37). Gardons dès lors comme principe de toujours employer la forme SCR de la statistique F lorsqu'une variable dépendante différente intervient dans la régression du modèle contraint.

L'estimation du modèle non contraint sur les données du fichier HPRICE1, nous permet d'obtenir les résultats suivants :

$$\widehat{\log(price)} = 0,264 + 1,043\log(assess) + 0,0074\log(lotsize)$$
$$(0,570)\ \ (0,151)\qquad\qquad (0,0386)$$
$$-0,1032\ \log(sqrft) + 0,0338\ bdrms$$
$$(0,1384)\qquad\qquad (0,0221)$$
$$n = 88,\ SCR = 1,822,\ R^2 = 0,773.$$

Si nous utilisons des statistiques de Student individuelles pour tester chaque hypothèse de (4.48), aucune ne peut être rejetée. La logique sous jacente de test cependant est celle d'un test joint sur un ensemble de paramètres. Nous devrions donc tester ces restrictions conjointement. La SCR du modèle contraint est de $SCR_c = 1,880$, et la statistique F de $[(1,880 - 1,822)/(1,822)] \times (83/4) = 0,661$. La valeur critique d'une distribution F avec (4,83) *ddl* au seuil de 5 % se situe autour de 2,50. L'hypothèse nulle peut par conséquent être rejetée.

4.6 REPORTER LES RÉSULTATS D'ESTIMATION DES MODÈLES DE RÉGRESSION

Nous terminons ce chapitre en fournissant quelques conseils sur la façon de rendre compte des résultats de régressions multiples pour des projets empiriques relativement compliqués. Ces éléments devraient vous aider à lire des travaux publiés (monographies ou articles scientifiques) dans le domaine des sciences sociales appliquées, tout en vous préparant à rédiger vos propres études empiriques. Nous développerons ce sujet plus avant en présentant et discutant les tableaux de résultats de plusieurs exemples. La plupart des points clés néanmoins peuvent être abordés dès à présent.

Il va de soi que les coefficients estimés par les MCO doivent toujours être reportés. S'agissant des variables clés de l'analyse, vous devez *interpréter* les coefficients estimés (ce qui nécessite souvent une connaissance précise des unités de mesure des variables). Par exemple, vous pouvez vous interroger sur le fait que le paramètre estimé représente ou non une élasticité, ou bien si, plus généralement, l'interprétation du paramètre nécessite une explication particulière. L'importance économique ou pratique des estimations des paramètres clés doit être discutée.

Les écarts-types doivent par ailleurs toujours être inclus en parallèle des coefficients estimés. Certains auteurs préfèrent reporter les statistiques de Student plutôt que les écarts-types (et parfois simplement la valeur absolue des statistiques de Student). Bien qu'il n'y ait rien de mal à cela, il est préférable, par souci de transparence, de fournir l'information sur les écarts-types estimés des coefficients estimés. Tout d'abord, cela vous oblige à réfléchir à l'hypothèse nulle testée ; l'hypothèse nulle la plus pertinente n'étant pas toujours que le paramètre de la population est égale à zéro. D'autre part, la présence des écarts-types facilite le calcul des intervalles de confiance.

Le R-carré de la régression doit également toujours apparaître. Nous avons vu qu'en plus de fournir une mesure de la qualité d'ajustement du modèle estimé, il peut servir au calcul des statistiques F pour les

tests de restrictions d'exclusion simples. Reporter à la fois la somme des carrés des résidus et l'écart-type de la régression est souvent pertinent, mais pas essentiel. Le nombre d'observations utilisées pour réaliser la régression doit également apparaître à proximité de l'équation estimée.

Si un nombre modéré de modèles différents est estimé, les résultats peuvent être résumés sous forme d'équation, comme nous l'avons fait jusqu'à présent. Cependant, dans la plupart des articles scientifiques, de nombreuses équations sont estimées avec différents ensembles de variables indépendantes. Il est possible en effet, de vouloir estimer la même équation pour différents groupes de personnes, voire même, estimer des modèles expliquant différentes variables dépendantes. Dans de tels cas, il est préférable de synthétiser les résultats dans un ou plusieurs tableaux. La variable dépendante doit être clairement indiquée dans le tableau, et les variables indépendantes listées dans la première colonne. Les écarts-types estimés (ou statistiques de Student) peuvent être mis entre parenthèses sous les paramètres estimés.

EXEMPLE 4.10
Analyse de l'arbitrage entre le montant des salaires et l'épargne retraite des enseignants

Soit *totcomp*, la variable mesurant la rémunération totale annuelle moyenne d'un enseignant, comprenant le salaire ainsi que les avantages sociaux (retraite, assurance maladie, etc.). Si l'on étend l'équation de salaire standard, la rémunération totale doit être fonction de la productivité et sans doute d'autres caractéristiques. Comme il est d'usage pour modéliser les salaires, nous utilisons la forme logarithmique suivante

$$\log(totcomp) = f(productivity\ characteristics, other\ factors),$$

où $f(\cdot)$ désigne une fonction quelconque pour le moment. Nous posons maintenant :

$$totcomp = salary + benefits = salary\left(1 + \frac{benefits}{salary}\right).$$

Cette équation montre que la rémunération totale est le produit de deux facteurs : *salary*, le salaire, et $1 + b/s$, où b/s désigne le « ratio allocations sur salaire ». En prenant le log de cette équation, nous obtenons $log(totcomp) = log\ (salary) + log\ (1 + b/s)$. Pour les « petites » valeurs de b/s, nous avons recours à l'approximation suivante : $\log\ (1 + b/s) \approx b/s$. Ceci conduit au modèle économétrique suivant :

$$log\ (salary) = \beta_0 + \beta_1\ (b/s) + other\ factors$$

Tester l'arbitrage entre allocations et salaire revient à tester $H_0 : \beta_1 = -1$ contre $H_1 : \beta_1 \neq -1$.

Pour ce faire, nous utilisons les données du fichier MEAP93. Cette base de données reporte un ensemble de valeurs moyennes par institution scolaire. Nombre de facteurs autres que ceux relatifs au lieu de travail sont également susceptibles d'expliquer la rémunération totale. Par manque d'information cependant, ces variables ne seront pas intégrées au modèle. Nous pouvons ainsi tenir compte de l'influence de la taille de l'école (*enroll*), des ressources disponibles (*staff*), et des mesures telles que le taux de décrochage scolaire et d'obtention de diplôme. La moyenne de b/s sur l'échantillon est d'environ 0,205, et la plus grande valeur est 0,450.

Les équations estimées sont explicitées dans le tableau 4.1, où les écarts-types estimés sont indiqués entre parenthèses juste en dessous des coefficients estimés. La variable clé est b/s, le rapport allocations-salaire.

À partir de la première colonne du tableau 4.1, nous voyons qu'en l'absence de variables de contrôle, l'estimateur des MCO de b/s est $-0,825$. La statistique de Student pour tester l'hypothèse nulle est $t = (2,825 + 1)/0,200 = 0,875$. Par conséquent, l'approche par régression linéaire simple ne permet pas de rejeter H_0. Après l'ajout de variables de contrôle telles que la taille de l'école et le taux d'encadrement (qui mesure à peu près le nombre d'élèves par enseignant), l'estimation du coefficient de b/s devient $-0,605$. Dans ce cas, le test de $H_0 : \beta_1 = -1$ donne une statistique de Student d'environ 2,39 ; permettant ainsi le rejet de H_0 au seuil de 5 % contre l'hypothèse alternative bilatérale. On note par ailleurs que les variables log (*enroll*) et log (*staff*) sont statistiquement très significatives.

Pour aller plus loin 4.6

Dans quelle mesure l'ajout des variables *droprate* et *gradrate* influencent-elles l'estimation du lien entre salaire et allocations ? Sont-elles significatives au seuil de 5 % ? Qu'en est-il au seuil de 10 % ?

Tableau 4.4 *Tester l'arbitrage salaires-avantages sociaux*

	Variable dépendante : log(*salary*)		
Variables indépendantes	**(1)**	**(2)**	**(3)**
b/s	– 0,825 (0,200)	– 0,605 (0,165)	– 0,589 (0,165)
log(*enroll*)	—	0,0874 (0,0073)	0,0881 (0,0073)
log(*staff*)	—	– 0,222 (0,050)	– 0,218 (0,050)
droprate	—	—	– 0,00028 (0,00161)
gradrate	—	—	0,00097 (0,00066)
constante	10,523 (0,042)	10,884 (0,252)	10,738 (0,258)
Observations R-carré	408 0,040	408 0,353	408 0,361

RÉSUMÉ

Dans ce chapitre nous nous sommes intéressés à un sujet central en statistique et en économétrie : l'inférence statistique. L'inférence statistique regroupe un ensemble de méthodes permettant de tirer des conclusions sur la population à partir d'un nombre d'observations limité issues d'un échantillon aléatoire. Nous résumons les principaux résultats discutés dans le corps du chapitre ci-dessous :

1. Sous les hypothèses du modèle linéaire classique RLM.1 à RLM.6, les estimateurs des paramètres du modèle de régression linéaire par la méthode des MCO sont normalement distribués.

2. Sous les hypothèses MLC, les statistiques t suivent une distribution de Student sous l'hypothèse nulle.

3. Nous avons recours à des statistiques t pour réaliser des tests d'hypothèse sur un unique paramètre du modèle, l'hypothèse alternative pouvant être unilatérale ou bilatérale. Selon les cas, le test portera alors sur l'une ou bien les deux queues de la distribution de Student respectivement. L'hypothèse nulle la plus couramment utilisée dans les applications est donnée par : $H_0 : \beta_j = 0$, mais il arrive parfois que l'on souhaite tester d'autres valeurs pour β_j sous H_0.

4. Dans le cadre des tests d'hypothèses traditionnels, nous fixons d'abord un seuil de significativité, qui en fonction du nombre de degrés de liberté (*ddl*) et de la formulation de l'hypothèse alternative, permettra de déterminer la valeur critique à laquelle sera comparée la valeur calculée de la statistique de Student. Il

est souvent plus pratique de calculer la p-valeur d'une statistique t – soit le plus petit seuil de significativité pour lequel l'hypothèse nulle peut être rejetée – pour tester une hypothèse au seuil de significativité désiré.

5. Sous les hypothèses MLC, les intervalles de confiance peuvent être construits pour chacun des paramètres β_j. Ces IC peuvent être utilisés pour tester n'importe quelle hypothèse nulle relative à β_j contre l'alternative bilatérale.

6. Les tests d'hypothèses simples concernant plus d'un paramètre peuvent toujours être réalisés en réécrivant le modèle de telle sorte à faire apparaître le paramètre d'intérêt dans l'équation. Ensuite, une statistique de Student standard peut être utilisée pour réaliser le test.

7. La statistique F est utilisée pour tester des restrictions d'exclusion multiples. Deux formes équivalentes de ce test peuvent être employées. L'une est construite sur les SCR des modèles contraint et non contraint. Une forme alternative, plus simple à manipuler, s'obtient à partir des R-carrés des deux modèles.

8. Lors du calcul d'une statistique F, le nombre de degrés de liberté au numérateur représente le nombre de contraintes sur le modèle, tandis que le nombre de degrés de liberté au dénominateur correspond au nombre de degrés de liberté du modèle non contraint.

9. L'hypothèse alternative pour le test F est bilatérale. Dans l'approche traditionnelle, nous spécifions un seuil qui, en fonction du nombre de degrés de liberté au numérateur et au dénominateur, permet de déterminer la valeur critique. L'hypothèse nulle est rejetée lorsque la statistique F, dépasse la valeur critique, c. Une alternative est de calculer la p-valeur afin d'avoir une vision d'ensemble des preuves à l'encontre de H_0 pour les différents seuils de significativités potentiellement envisagés.

10. Des restrictions linéaires générales peuvent être testées en utilisant la statistique F écrite en fonction de la somme des carrés des résidus.

11. La statistique F permet de tester la significativité globale du modèle sous l'hypothèse nulle que tous les paramètres sont égaux à zéro, à l'exception de la constante. Autrement dit, on teste si les paramètres de pente sont tous nuls. Sous l'hypothèse H_0, les variables explicatives n'ont pas d'impact sur les valeurs attendues de y.

12. Lorsque des données sont manquantes pour une ou plusieurs variables explicatives du modèle, il convient de rester prudent lors du calcul « manuel » de la statistique F, c'est-à-dire lorsque cette statistique est calculée à partir de la somme des carrés des résidus ou des R carrés issus des deux régressions. Lorsque cela est possible, il vaut mieux recourir aux calculs réalisés de façon automatique par des logiciels statistiques qui disposent de commandes pré-programmées demeurant valides en présence de données manquantes.

LES HYPOTHÈSES DU MODÈLE LINÉAIRE CLASSIQUE

Il semble maintenant opportun de revoir l'ensemble des hypothèses du modèle linéaire classique (MLC) pour des régressions en coupe transversale. À la suite de chaque hypothèse vous trouverez un commentaire explicitant son rôle dans l'analyse des régressions multiples.

Hypothèse RLM.1 (Linéarité des paramètres)

Le modèle de la population peut être écrit comme ceci :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

où β_0, β_1, …, β_k sont les paramètres inconnus (invariants) d'intérêt et u est un terme d'erreur ou perturbation aléatoire non observé(e).

L'hypothèse RLM.1 décrit la relation au sein de la population que nous espérons pouvoir estimer. Elle définit également explicitement les paramètres d'intérêt : β_j – les effets *ceteris paribus* de x_j sur y au sein de la population.

Hypothèse RLM.2 (Échantillonnage aléatoire)

Nous considérons un échantillon aléatoire de n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, \ldots, n\}$, suivant le modèle prévalant au sein de la population défini sous l'hypothèse RLM.1.

L'hypothèse d'échantillonnage aléatoire signifie que nous avons des données qui peuvent être utilisées pour estimer les β_j, et que les données ont été choisies pour être représentatives de la population définie par l'hypothèse RLM.1.

Hypothèse RLM.3 (Absence de colinéarité parfaite)

Dans l'échantillon (et par extension au sein de la population), aucune des variables indépendantes n'est supposée constante. De plus, aucune variable dépendante ne peut être exprimée comme une combinaison linéaire parfaite des autres.

Une fois que nous disposons d'un échantillon de données observées, nous avons besoin de savoir que nous pouvons les utiliser pour estimer les paramètres du modèle par la méthode des MCO, $\hat{\beta}_j$. C'est justement le rôle de l'hypothèse RLM.3 : si nous observons des variations pour chacune des variables indépendantes de l'échantillon et qu'il n'existe aucune combinaison linéaire entre les variables indépendantes, nous pouvons estimer les valeurs de $\hat{\beta}_j$.

Hypothèse RLM.4 (Moyenne conditionnelle égale à zéro)

L'erreur u a une valeur attendue de zéro, conditionnellement aux valeurs prises par les variables explicatives. En d'autres termes,

$$E[u|x_1, x_2, \ldots, x_k] = \sigma^2.$$

Comme nous l'avons vu précédemment, supposer que les facteurs non observés sont, en moyenne, non corrélés avec les variables explicatives, permet de dériver la première propriété statistique de chacun des estimateurs des MCO : soit le fait qu'il s'agit d'estimateurs sans biais des paramètres de la population. Bien évidemment, toutes les hypothèses précédentes sont utilisées pour démontrer l'absence de biais.

Hypothèse RLM.5 (Homoscédasticité)

L'erreur u a la même variance quelque soit les valeurs des variables explicatives. En d'autres termes,

$$Var(u|x_1, x_2, \ldots, x_k) = \sigma^2.$$

Au regard de l'hypothèse RLM.4, l'hypothèse d'homoscédasticité est d'une importance secondaire ; en particulier, l'hypothèse RLM.5 n'a aucune incidence sur le biais du paramètre $\hat{\beta}_j$.

L'hypothèse d'homoscédasticité a malgré tout deux conséquences importantes : (1) Nous pouvons dériver des formules pour les variances d'échantillonnage dont les composantes sont faciles à caractériser ; (2) Nous pouvons conclure, sous les hypothèses de Gauss-Markov RLM.1 à RLM.5, que les estimateurs des MCO ont la plus petite variance parmi tous les estimateurs linéaires sans biais.

Hypothèse RLM.6 (Normalité)

Le terme d'erreur de la population u est indépendant des variables explicatives $x_1, x_2, \ldots, x_k$, et est distribué normalement avec une moyenne nulle et une variance σ^2.

Dans ce chapitre, nous avons ajouté l'hypothèse RLM.6 afin d'obtenir les distributions d'échantillonnage exactes des statistiques de Student et de Fisher. De cette façon, nous pouvons mener des tests d'hypothèses exacts. Dans le chapitre suivant, nous verrons que l'hypothèse MLC.6 peut être relâchée à condition de disposer d'un échantillon de taille suffisante. L'hypothèse RLM.6 implique une propriété d'efficience plus forte de l'estimateur des MCO : celui-ci présente la plus petite variance parmi tous les estimateurs sans biais ; le groupe de comparaison ne se limitant alors plus à la classe des estimateurs linéaires dans $\{y_i : i = 1, 2, \ldots, n\}$.

MOTS-CLÉS

Conjointement non significatif p. 184
Conjointement statistiquement significatif p. 184
Contraintes multiples p. 180
Degrés de liberté du dénominateur p. 183
Degrés de liberté du numérateur p. 182
Estimateur non biaisé de variance minimale p. 153
Forme R-carré de la statistique *F* p. 186
Hypothèse alternative p. 158
Hypothèse alternative bilatérale p. 163
Hypothèse alternative unilatérale p. 158
Hypothèse de normalité p. 152
Hypothèses du modèle linéaire classique (MLC) p. 152
Hypothèse nulle p. 156
Importance économique p. 171, 191
Intervalle de confiance (IC) p. 173
Modèle contraint p. 181
Modèle non contraint p. 181
Modèle linéaire classique p. 152
Niveau de significativité p. 158
p-valeur p. 168
Règle de décision p. 158
Restrictions d'exclusion p. 180
Significativité globale de la régression p. 189
Significativité pratique p. 170
Statistique de Fisher ou statistique *F* p. 182
Statistique de Student ou statistique *t* p. 157
Statistiquement non significatif p. 164
Statistiquement significatif p. 164
Test bilatéral p. 163
Test d'hypothèses jointes p. 180
Test d'hypothèses multiples p. 180
Test unilatéral p. 158
Valeur critique p. 158

EXERCICES

1. Parmi les cas explicités ci-après, lequel d'entre eux peut invalider les statistiques de Student des estimateurs des MCO (c'est-à-dire faire en sorte que ces statistiques ne suivent pas de distribution de Student sous l'hypothèse nulle) ?

i. l'hétéroscédasticité du terme d'erreur ;

ii. un coefficient de corrélation empirique de 0,95 observé pour deux des variables indépendantes du modèle ;

iii. l'omission d'une variable explicative importante dans le modèle.

2. Considérons une équation de salaire cherchant à expliquer les revenus des PDG en fonction des ventes annuelles de l'entreprise, du retour sur fonds propres (*roe* en pourcentage), et du rendement des actions de la société (*ros*, en pourcentage) :

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 ros + u.$$

i. On pose comme hypothèse nulle que *ros* n'a aucun effet sur le salaire des PDG après avoir pris en compte l'influence du niveau des ventes et du ROE. Écrire l'hypothèse nulle. L'hypothèse alternative est que de meilleures performances boursières donnent lieu à une augmentation de salaire des PDG. Écrire l'hypothèse alternative.

ii. À partir des données contenues dans le fichier CEOSAL1, un modèle de régression linéaire est estimé par la méthode des MCO avec pour résultats :

$$\widehat{\log(salary)} = 4{,}32 + 0{,}280 \log(sales) + 0{,}0174\, roe + 0{,}00024\, ros$$
$$(0{,}32)\ \ (0{,}035)\qquad\qquad (0{,}0041)\qquad (0{,}00054)$$
$$n = 209,\ R^2 = 0{,}283$$

Quelle est la variation de salaire attendue en pourcentage si la variable *ros* augmente de 50 points ?

Le ROS a-t-il un impact important en pratique sur le salaire ?

iii. Testez l'hypothèse nulle selon laquelle *ros* n'a pas d'effet sur le salaire contre l'hypothèse alternative que *ros* a un effet positif. Testez cette hypothèse au seuil de significativité de 10 %.

iv. Seriez-vous en faveur d'inclure la variable *ros* dans le modèle final expliquant la rémunération des PDG en fonction des performances de l'entreprise ? Justifiez.

3. La variable *rdintens* mesure les dépenses de recherche et développement (R&D) au prorata des ventes exprimées en pourcentage. Les ventes sont mesurées en millions de dollars. La variable *profmarg* mesure quant à elle le profit en pourcentage des ventes.

Le modèle suivant est estimé à partir des données de 32 entreprises de l'industrie chimique contenues dans le fichier RDCHEM :

$$\widehat{rdintens} = 0{,}472 + 0{,}321 \log(sales) + 0{,}050\, profmarg$$
$$(1{,}369)\ \ (0{,}216)\qquad\qquad (0{,}046)$$
$$n = 32,\ R^2 = 0{,}099.$$

i. Interprétez le coefficient de log(*sales*). En particulier, évaluez l'impact sur *rdintens* d'une augmentation des ventes de 10 % ? S'agit-il d'un effet économique important ?

ii. Testez l'hypothèse d'absence de lien entre les ventes et l'intensité des dépenses de R&D contre l'hypothèse alternative d'une augmentation de cette intensité en fonction des ventes. Faites le test aux seuils de 5 % puis 10 %.

iii. Interprétez le coefficient de *profmarg*. Est-il économiquement important ?

iv. La variable *profmarg* a-t-elle un effet statistiquement significatif sur *rdintens* ?

4. Le prix des loyers est-il influencé par la taille de la population estudiantine dans les villes universitaires ? Soit *rent* le loyer mensuel moyen payé pour la location d'un appartement dans une ville universitaire aux États-Unis. La variable *pop* désigne le nombre d'habitants de la ville, *avginc* le revenu moyen des habitants de la ville, et *pctstu* la population étudiante sur la population totale en pourcentage. Le modèle suivant est utilisé pour tester cette relation :

$$\log(rent) = \beta_0 + \beta_1 \log(pop) + \beta_2 \log(avginc) + \beta_3 pctstu + u.$$

i. Posez l'hypothèse nulle selon laquelle la part de la population estudiantine n'a pas d'effet *ceteris paribus* sur les loyers mensuels, contre l'hypothèse alternative de présence d'effet. Quels signes attendez-vous pour β_1 et β_2 ?

ii. L'équation suivante est estimée en utilisant les données de la base RENTAL regroupant des informations sur 64 villes universitaires en 1990.

$$\widehat{\log(rent)} = 0{,}043 + 0{,}066 \log(pop) + 0{,}507 \log(avginc) + 0{,}0056\, pctstu$$
$$(0{,}844)\ \ (0{,}039) \qquad\qquad (0{,}081) \qquad\qquad (0{,}0017)$$
$$n = 64,\ R^2 = 0{,}458$$

Pourquoi l'affirmation selon laquelle : « Une augmentation de 10 % de la population est associée à une augmentation de 6,6 % des loyers » peut-elle poser problème ?

iii. Testez l'hypothèse mentionnée dans la question (i) au seuil de 1 %.

5. Considérons l'équation estimée de l'exemple 4.3, que nous utilisons ici pour étudier les effets de l'absentéisme sur les résultats aux tests du premier cycle universitaire américain (*college GPA*) :

$$\widehat{colGPA} = 1{,}39 + 0{,}412\, hsGPA + 0{,}015\, ACT - 0{,}083\, skipped$$
$$(0{,}33)\ \ (0{,}094) \qquad\qquad (0{,}011) \qquad\qquad (0{,}26)$$
$$n = 141,\ R^2 = 0{,}234,$$

i. En utilisant l'approximation de la loi normale standardisée, trouvez l'intervalle de confiance à 95 % de β_{hsGPA}.

ii. Pouvez-vous rejeter l'hypothèse $H_0 : \beta_{hsGPA} = 0{,}4$ contre l'hypothèse alternative bilatérale au seuil de 5 % ?

iii. Pouvez-vous rejeter l'hypothèse $H_0 : \beta_{hsGPA} = 1$ contre l'hypothèse alternative bilatérale au seuil de 5 % ?

6. Dans la section 4.5, nous avons utilisé comme exemple la réalisation d'un test pour évaluer dans quelle mesure le prix des logements pouvait s'expliquer de manière rationnelle dans le cadre d'un modèle log-log pour le prix, *price*, et l'évaluation, *assess* [voir l'équation (4.47)]. Ici, nous nous proposons d'utiliser une formulation de type niveau-niveau.

i. Dans le modèle de régression simple suivant :

$$price = \beta_0 + \beta_1 assess + u,$$

l'évaluation du prix est rationnelle si $\beta_1 = 1$ et $\beta_0 = 0$. L'équation estimée est donnée par :

$$\widehat{\text{price}} = -14{,}47 + 0{,}976\,assess$$
$$(16{,}27)\quad (0{,}049)$$
$$n = 88,\ \text{SCR} = 165644{,}51,\ R^2 = 0{,}820$$

Testez tout d'abord l'hypothèse $H_0 : \beta_0 = 0$ contre l'hypothèse alternative bilatérale. Ensuite, testez $H_0 : \beta_1 = 1$ contre l'hypothèse alternative bilatérale. Que pouvez-vous conclure de ces résultats ?

ii. Pour tester l'hypothèse jointe $\beta_0 = 0$ et $\beta_1 = 1$, il nous faut calculer la SCR du modèle contraint. Cela revient à calculer $\sum_{i=1}^{n}(price_i - assess_i)^2$, où $n = 88$, puisque le résidu dans le modèle contraint est simplement donné par $price_i - assess_i$. (Aucune estimation n'est nécessaire pour le modèle contraint puisque les deux paramètres sont spécifiés sous H_0.) Nous obtenons alors SCR = 209448,99. Effectuez le test F de l'hypothèse jointe.

iii. Nous souhaitons maintenant tester les contraintes suivantes $H_0 : \beta_2 = 0$, $\beta_3 = 0$, et $\beta_4 = 0$ dans le modèle.

$$price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u.$$

L'estimation du modèle sur les mêmes 88 maisons permet d'obtenir un R-carré de 0,829.

iv. Si la variance de la variable *price* change en fonction des variables *assess*, *lotsize*, *sqrft*, or *bdrms* que pouvez vous dire du test F de la partie (iii) ?

7. Dans l'exemple 4.7, nous avons utilisé des données sur les entreprises manufacturières non syndiquées pour estimer la relation entre le taux de rebut et d'autres caractéristiques de l'entreprise. Nous allons maintenant analyser cet exemple plus en détail en utilisant pour nos estimations toutes les entreprises disponibles.

i. Le modèle théorique estimé dans l'exemple 4.7 peut être écrit comme suit :

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales) + \beta_3 \log(employ) + u.$$

L'estimation du modèle à partir de l'échantillon des 43 observations disponibles pour l'année 1987, permet d'obtenir les résultats suivants :

$$\widehat{\log(scrap)} = 11{,}74 - 0{,}042\,hrsemp - 0{,}951 \log(sales) + 0{,}992 \log(employ)$$
$$(4{,}57)\ (0{,}019)\qquad (0{,}370)\qquad (0{,}360)$$
$$n = 43,\ R^2 = 0{,}310$$

Comparez cette équation à celle estimée en utilisant seulement les 29 entreprises non syndiquées de l'échantillon.

ii. Montrez que le modèle théorique peut aussi s'écrire :

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales/employ) + \theta_3 \log(employ) + u,$$

où $\theta_3 = \beta_2 + \beta_3$. [*Indice* : Rappelons que $\log(x_2 / x_3) = \log(x_2) - \log(x_3)$.] Interprétez l'hypothèse $H_0 : \theta_3 = 0$.

iii. Lorsque l'équation de la question (ii) est estimée, on obtient les résultats suivants :

$$\widehat{\log(scrap)} = 11{,}74 - 0{,}042\,hrsemp - 0{,}951 \log(sales/employ) + 0{,}041 \log(employ)$$
$$(4{,}57)\ (0{,}019)\qquad (0{,}370)\qquad (0{,}205)$$
$$n = 43,\ R^2 = 0{,}310$$

Une fois que l'on contrôle pour la formation des travailleurs et le ratio ventes par employé, les grandes entreprises affichent-elles un taux de rebut statistiquement plus significatif ?

iv. Testez l'hypothèse selon laquelle une augmentation de 1 % de *sales/employ* est associée à une baisse de 1 % du taux de rebut.

8. Considérons le modèle de régression multiple avec trois variables indépendantes, sous les hypothèses usuelles du modèle linéaire classique, de RLM.1 à RLM.6 :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

Vous souhaitez tester l'hypothèse nulle $H_0 : \beta_1 - 3\beta_2 = 1$.

i. Soit $\hat{\beta}_1$ et $\hat{\beta}_2$ les estimateurs des MCO des paramètres β_1 et β_2. Écrivez $\text{Var}(\hat{\beta}_1 - 3\hat{\beta}_2)$ en fonction des variations $\hat{\beta}_1$, $\hat{\beta}_2$ et de leur covariance. Quelle est l'expression de l'écart-type de $\hat{\beta}_1 - 3\hat{\beta}_2$?

ii. Écrivez la statistique t permettant de tester $H_0 : \beta_1 - 3\beta_2 = 1$.

iii. On définit $\theta_1 = \beta_1 - 3\beta_2$ et $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2$. Réécrivez l'équation de départ de sorte à faire apparaître β_0, θ_1, β_2 et β_3 et ainsi directement obtenir l'expression de $\hat{\theta}_1$.

9. Dans l'exercice 3 du chapitre 3, nous avions estimé l'équation :

$$\widehat{sleep} = 3638{,}25 - 0{,}148\,totwrk - 11{,}13\,educ + 2{,}20\,age$$
$$(112{,}28)\quad (0{,}017)\qquad (5{,}88)\qquad (1{,}45)$$
$$n = 706,\ R^2 = 0{,}113.$$

pour laquelle nous reportons maintenant les paramètres estimés ainsi que les écarts-types estimés associés.

i. Les variables *educ* ou *age* sont-elles individuellement significatives au seuil de 5 % contre l'hypothèse alternative bilatérale ? Justifiez votre réponse.

ii. Exclure *educ* et *age* de l'équation donne

$$\widehat{sleep} = 3586{,}38 - 0{,}151\,totwk$$
$$(38{,}91)\quad (0{,}17)$$
$$n = 706,\ R^2 = 0{,}103$$

Les variables *educ* et *age* sont-elles conjointement significatives au seuil de 5 % ? Justifiez votre réponse.

iii. Est-ce que l'introduction ou non des variables *educ* et de *age* dans le modèle affecte substantiellement l'arbitrage estimé entre les temps de sommeil et de travail ?

iv. Supposons que l'équation de sommeil contienne de l'hétéroscédasticité. Quelle incidence ceci peut avoir sur les tests calculés dans les questions (i) et (ii) ?

10. Les modèles de régression peuvent être utilisés pour tester si les marchés intègrent l'information disponible de manière efficiente dans les prix des actifs financiers. Pour procéder à un test formel, considérons la variable *return* qui mesure le rendement associé à la détention d'un titre financier durant une période de quatre ans, allant de fin 1990 à fin 1994. L'hypothèse d'efficience des marchés implique que ces rendements ne doivent pas être corrélés à l'information connue en 1990. Si les caractéristiques de l'entreprise dont les titres sont échangés sur les marchés, permettaient de prédire les rendements, cette information pourrait alors être utilisée pour sélectionner les actifs financiers permettant de réaliser un profit sans risque.

Considérons plusieurs caractéristiques d'entreprises dont les valeurs sont connues en 1990 : *dkr* désigne le ratio des dettes sur fonds propres, *eps* capture le bénéfice par action, enfin *ntinc* et *salary* mesurent respectivement le bénéfice net et le salaire total du PDG.

i. En utilisant les données du fichier RETURN, l'équation suivante a été estimée :

$$\widehat{return} = -14{,}37 + 0{,}321\,dkr + 0{,}043\,eps - 0{,}0051\,netinc + 0{,}0035\,salary$$

$$(6{,}89)\ (0{,}201) \qquad (0{,}078) \qquad (0{,}0047) \qquad (0{,}0022)$$

$$n = 142,\ R^2 = 0{,}0395$$

Testez si les variables explicatives sont conjointement significatives au seuil de 5 %. Les variables explicatives prises individuellement sont-elles statistiquement significatives ?

ii. Maintenant, estimez à nouveau le modèle en utilisant une transformation logarithmique pour les variables *netinc* et *salary* :

$$\widehat{return} = -36{,}30 + 0{,}327\,dkr + 0{,}069\,eps - 4{,}74\log(netinc) + 7{,}24\log(salary)$$

$$(39{,}37)\ (0{,}203) \qquad (0{,}080) \qquad (3{,}39) \qquad (6.31)$$

$$n = 142,\ R^2 = 0{,}0330$$

Certaines de vos conclusions de la question (i) sont-elles remises en cause ?

iii. Dans cet exemple, certaines entreprises ne possèdent pas de dette, d'autres ont enregistré des bénéfices négatifs. Devrions-nous essayer d'utiliser les transformations log(*dkr*) ainsi que log(*eps*) dans le modèle afin de voir si ceci améliore le qualité prédictive du modèle ? Justifiez.

iv. Dans l'ensemble, l'analyse des données nous fournit-elle des preuves solides sur la prédictabilité des rendements ?

11. Le tableau suivant a été généré en utilisant les données du fichier CEOSAL2. Les écarts-types estimés sont mentionnés entre parenthèses sous les coefficients estimés du modèle :

	Variable dépendante : log(*salary*)		
Variables indépendantes	**(1)**	**(2)**	**(3)**
log(*sales*)	0,224 (0,027)	0,158 (0,040)	0,188 (0,040)
log(*mktval*)	—	0,112 (0,050)	0,100 (0,049)
Profmarg	—	– 0,0023 (0,0022)	– 0,0022 (0,0021)
Ceoten	—	—	0,0171 (0,0055)
comten	—	—	– 0,0092 (0,0033)
constante	4,94 (0,20)	4,62 (0,25)	4,57 (0,25)
Observations *R*-carré	177 0,281	177 0,304	177 0,353

La variable *mktval* désigne la valeur de marché de l'entreprise, *profmarg* le profit en pourcentage des ventes, *ceoten* le nombre d'années d'expérience du PDG à son poste, *comten* désigne le nombre total d'années d'exercice du PDG au sein de l'entreprise (au poste actuel ou à un autre poste).

i. Commentez l'influence de la variable *profmarg* sur le salaire des PDG.

ii. La valeur de marché a-t-elle un effet statistiquement significatif ? Justifiez.

iii. Interprétez les coefficients de *ceoten* et *comten*. Ces variables explicatives sont-elles statistiquement significatives ?

iv. Que pensez-vous du fait que l'augmentation de l'ancienneté dans l'entreprise, toutes choses égales par ailleurs, est associée à un salaire inférieur ?

12. La régression suivante porte sur les données du fichier MEAP93. On y trouve les taux de réussite, en pourcentage, à un test de mathématiques (soit la note normalisée à 10 de ce test).

i. La variable *expend* mesure les dépenses par élève, en dollars, et *math*10 désigne le taux de réussite à l'examen. La régression simple explique les résultats au test, *math*10 par *lexpend* = log (*expend*) :

$$\widehat{math10} = -69{,}34 + 11{,}16\,lexpend$$
$$(25{,}53)\quad(3{,}17)$$
$$n = 408,\ R^2 = 0{,}0297$$

Interprétez le coefficient de *lexpend*. En particulier, calculez le changement du taux de réussite prédit par le modèle si la variable *expend* augmente de 10 %. Que pensez-vous de la valeur négative de la constante ? (La valeur minimale de *lexpend* est de 8,11 et sa valeur moyenne est de 8,37.)

ii. La faible valeur du coefficient de détermination, le R-carré, de la question (i) implique-t-elle que les dépenses sont fortement corrélées à d'autres facteurs influençant *math*10 ? Justifiez. Vous attendez-vous à obtenir un R-carré beaucoup plus élevé si les dépenses des écoles étaient décidées de manière aléatoire plutôt qu'en fonction de leur localisation ?

iii. Lorsque le log du nombre d'inscriptions ainsi que la part des élèves admissibles pour le programme de repas gratuit fédéral sont inclus, l'équation estimée devient :

$$\widehat{math10} = -23{,}14 + 7{,}75\,lexpend - 1{,}26\,lenroll - 0{,}324\,inchprg$$
$$(24{,}99)\ (3{,}04)\qquad(0{,}58)\qquad(0{,}36)$$
$$n = 408,\ R^2 = 0{,}1893$$

Commentez l'effet de cette nouvelle spécification sur le coefficient de *lexpend*. Le coefficient associé aux dépenses reste-t-il statistiquement différent de zéro ?

iv. Que pensez-vous du R-carré de l'équation estimée de la question (iii) ? Quels autres facteurs pourraient être pris en compte pour expliquer *math*10 (à l'échelle de l'école) ?

13. Les données contenues dans la base MEAPSINGLE ont été utilisées pour estimer les modèles décrits ci-dessous qui étudient les performances scolaires lors d'un test de mathématiques en fin de primaire et les caractéristiques socio-économiques des élèves de l'école. La variable *free*, mesurée à l'échelle de l'école, indique le pourcentage d'élèves éligibles au programme fédéral américain de couverture des frais de cantine[2].

2 Note des traducteurs : Il s'agit des « free lunch programs » qui permettent aux élèves d'un large panel d'établissements primaires et secondaires aux États-Unis d'accéder à un service de cantine gratuit. Pour plus de détails, on pourra se référer au descriptif de cette page du ministère américain de l'agriculture, en charge des questions alimentaires : https://www.ers.usda.gov/topics/food-nutrition-assistance/

La variable *medinc* correspond au revenu médian de la zone géographique identifiée ici par le code postal correspondant, et *pctsgle* le pourcentage d'élèves ne vivant pas avec leurs deux parents (également mesuré au niveau de la zone géographique, i.e. soit par code postal). Voir également l'exercice sur ordinateur C11 du chapitre 3.

$$\widehat{math4} = 96{,}77 - {,}844\ pctsgle$$
$$(1{,}60)\quad ({,}071)$$
$$n = 299,\ R^2 = {,}380$$

$$\widehat{math4} = 93{,}00 - {,}275\ pctsgle - {,}402\ free$$
$$(1{,}63)\quad ({,}117)\qquad (0{,}070)$$
$$n = 299,\ R^2 = {,}459$$

$$\widehat{math4} = 24{,}49 - {,}274\ pctsgle - {,}422\ free - {,}752\ lmedinc + 9{,}01\ lexppp$$
$$(59{,}24)\ ({,}161)\qquad ({,}071)\qquad (5{,}358)\qquad (4{,}04)$$
$$n = 299,\ R^2 = {,}472$$

$$\widehat{math4} = 17{,}52 - {,}259\ pctsgle - {,}420\ free + 8{,}80\ lexppp$$
$$(32{,}25)\ ({,}117)\qquad ({,}070)\qquad (3{,}76)$$
$$n = 299,\ R^2 = {,}472$$

i. Interprétez le coefficient associé à la variable *pctsgle* de la première équation. Commentez de façon détaillée ce qui arrive lorsque la variable *free* est introduite dans la régression comme variable explicative supplémentaire.

ii. Est-ce que la variable capturant les dépenses par élève, entrée sous forme logarithmique, présente un effet significatif sur la performance ? Si oui, quelle est l'amplitude de cet effet ?

iii. Si vous deviez sélectionner parmi les quatre équations proposées, celle présentant la meilleure estimation de l'impact de *pctsgle*, de façon à obtenir un intervalle de confiance à 95 % pour le coefficient associé à *bpctsgle*, quelle spécification choisiriez-vous ? Pourquoi ?

EXERCICES SUR ORDINATEUR

C1. Le modèle suivant peut être utilisé pour étudier si les dépenses de campagne influent sur les résultats des élections :

$$voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3\ prtystrA + u,$$

où *voteA* est le pourcentage du vote obtenu par le candidat A, *expendA* et *expendB* les dépenses de campagne des candidats A et B, et *prtystrA* une mesure de l'importance du parti pour le candidat A (soit le pourcentage de votes en faveur du parti du candidat A lors des élections les plus récentes).

i. Quelle est l'interprétation de β_1 ?

ii. En fonction des paramètres du modèle, spécifiez l'hypothèse nulle selon laquelle l'impact sur le vote d'une augmentation de 1 % des dépenses de A est compensée par une augmentation de 1 % des dépenses du B.

child-nutrition-programs/national-school-lunch-program.aspx

iii. Estimez le modèle en utilisant les données du fichier VOTE1 et présentez les résultats sous la forme usuelle. Les dépenses du candidat A influencent-elles le résultat ? Qu'en est-il des dépenses du candidat B ? Pouvez-vous utiliser ces résultats pour tester l'hypothèse de la question (ii) ?

iv. Estimez un modèle donnant directement la statistique de Student adéquate pour tester l'hypothèse décrite en question (ii). Que concluez-vous ? (Utilisez une hypothèse alternative bilatérale.)

C2. Utilisez les données du fichier LAWSCH85 pour cet exercice.

i. En utilisant le même modèle que dans celui de l'exercice 4 du chapitre 3, posez l'hypothèse nulle qui suppose que le classement des écoles de commerce n'a aucun effet *ceteris paribus* sur le salaire médian à la sortie des études.

ii. Les caractéristiques de la classe des étudiants venant d'entrer sur le marché du travail, *LSAT* et *GPA*, ont-elles individuellement ou conjointement un impact sur la variable *salary* ? (Assurez-vous de tenir compte des données manquantes sur *LSAT* et *GPA*.)

iii. Testez si la taille de la classe (*clsize*) ou celle du corps professoral (*faculty*) doivent être ajoutées à cette équation ; effectuez un seul test pour répondre à la question. (Prenez garde à bien tenir compte des données manquantes pour ces deux variables.)

iv. Quels sont les facteurs susceptibles d'influencer le classement des écoles de commerce absents de la régression de salaire ?

C3. Reportez-vous à l'exercice C2 du chapitre 3. Maintenant, utilisez le logarithme du prix du logement comme variable dépendante :

$$\log(price) = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u.$$

i. Vous vous intéressez à l'estimation et l'obtention d'un intervalle de confiance pour le taux de variation en pourcentage de la variable *price* lorsqu'une pièce de 150 pieds carrés est ajoutée à la configuration existante d'une maison donnée. Sous forme décimale, ceci s'écrit $\theta_1 = 150\beta_1 + \beta_2$. Utilisez les données du fichier HPRICE1 pour estimer θ_1.

ii. Écrivez β_2 en fonction de θ_1 et β_1 et introduisez ce paramètre dans l'équation expliquant le log(*price*).

iii. Utilisez les éléments de la question (ii) afin d'obtenir un écart-type estimé pour $\hat{\theta}_1$ puis calculez à partir de cet écart-type un intervalle de confiance à 95 %.

C4. Dans l'exemple 4.9, la version contrainte du modèle peut être estimée en utilisant l'ensemble des 1388 observations de l'échantillon. Calculez le R-carré de la régression de *bwght* sur les variables *cigs*, *parity* et *faminc* en utilisant toutes les observations. Comparez ce résultat au R-carré obtenu pour le modèle contraint dans l'exemple 4.9.

C5. Utilisez les données du fichier MLB1 pour cet exercice.

i. Utilisez le modèle estimé dans l'équation (4.31) en excluant la variable *rbisyr*. Qu'advient-il de la significativité statistique de *hrunsyr* ? Qu'en est-il de la taille du coefficient de *hrunsyr* ?

ii. Ajoutez les variables *runsyr* (nombre de points par an), *fldperc* (pourcentage de mise en service), et *sbasesyr* (bases « volées » par an) au modèle de régression de la question (i). Lequel de ces facteurs est individuellement statistiquement significatif ?

iii. Dans le modèle de la question (ii), testez la significativité jointe de *bavg*, *fldperc*, et *sbasesyr*.

C6. Utilisez les données du fichier WAGE2 pour cet exercice.

i. Considérons l'équation de salaire suivante

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

Posez l'hypothèse nulle qu'une année supplémentaire d'expérience professionnelle en général a le même effet sur la variable log (*salary*) qu'une année supplémentaire au sein de son entreprise actuelle.

ii. Testez l'hypothèse nulle de la question (i) contre l'hypothèse alternative bilatérale, au seuil de significativité de 5 %. Pour ce faire, construisez un intervalle de confiance à 95 %. Que pouvez-vous en conclure ?

C7. Reportez-vous à l'exemple utilisé dans la section 4.4. Vous utilisez cette fois l'ensemble des données du fichier TWOYEAR.

i. La variable *phsrank* reporte le centile dans lequel une personne se situait lors de ses études dans le secondaire. (Plus la valeur est élevée, meilleurs sont les résultats scolaires. Par exemple, 90 signifie que les résultats scolaires sont meilleurs que ceux de 90 pourcent de votre promotion.) Trouvez les valeurs minimale, maximale et moyenne de l'échantillon.

ii. Ajoutez la variable *phsrank* à l'équation (4.26) et reportez les estimations par la méthode des MCO des paramètres du modèle dans le format habituel. Peut-on en conclure que la variable *phsrank* est statistiquement significative ? Quel est le salaire attendu pour un individu se trouvant parmi les 10 meilleurs de sa promotion ?

iii. Est-ce que l'ajout de la variable *phsrank* à (4.26) modifie fortement les conclusions sur les revenus attendus suite à l'obtention d'un diplôme du supérieur pour un cursus de type court, 2 ans après les études secondaires, ou bien de type long, 4 ans après les études secondaires ? Justifiez.

iv. La base de données contient une variable appelée *id*. Expliquez pourquoi, si vous ajoutez *id* à l'une des équations (4.17) ou (4.26), vous vous attendez à ce qu'elle soit statistiquement significative. Quelle est la *p*-valeur associée au paramètre estimé de cette variable dans le cadre d'un test bilatéral ?

C8. La base de données du fichier 401KSUBS contient des informations sur le patrimoine financier net (*nettfa*), l'âge de la personne interrogée (*age*), le revenu familial annuel (*inc*), la taille de la famille (*fsize*), et la participation à certains régimes de retraite pour les personnes résidant au Royaume-Uni. Les variables de richesse et de revenu sont toutes deux enregistrées en milliers de dollars. Pour cette question, utilisez uniquement les données pour les ménages d'une seule personne (soit lorsque *fsize* = 1).

i. Combien de ménages d'une personne sont présents dans la base de données ?

ii. Utilisez la méthode des MCO pour estimer le modèle suivant :

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + u,$$

Présentez les résultats en utilisant le format habituel. Assurez-vous d'utiliser uniquement les ménages d'une personne dans l'échantillon. Interprétez les coefficients de pente. Y a-t-il des surprises dans les estimations de certains des paramètres de pente ?

iii. Est-ce que la constante de la régression du modèle présenté en question (ii) possède une signification intéressante ? Justifiez.

iv. Calculez la *p*-valeur du test $H_0 : \beta_2 = 1$ contre $H_1 : \beta_2 < 1$. Pouvez-vous rejeter H_0 au seuil de significativité de 1 % ?

v. Si vous faites une régression simple de *nettfa* sur *inc*, trouvez-vous un coefficient estimé pour la variable *inc* très différent de l'estimation réalisée en question (ii) ? Expliquez votre résultat.

C9. Utilisez des données du fichier DISCRIM afin de répondre à cette question. (Voir aussi exercice sur ordinateur C8 du chapitre 3)

i. Utilisez la méthode des MCO pour estimer le modèle suivant :

$$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + \beta_3 prppov + u,$$

et présentez les résultats sous la forme habituelle. Le paramètre estimé, $\hat{\beta}_1$, est-il statistiquement différent de zéro au seuil de 5 % contre l'hypothèse alternative bilatérale ? Qu'en est-il au seuil de 1 % ?

ii. Quelle est la corrélation entre log(*income*) et *prppov* ? Les variables apparaissent-elles chacune statistiquement significative ? Signalez les *p*-valeurs dans le cas de tests bilatéraux.

iii. Reprenez l'équation de la question (i) et ajoutez-y la variable log(*hseval*). Interprétez son coefficient et indiquez la *p*-valeur relative au test bilatéral pour $H_0 : \beta_{\log(hseval)} = 0$.

iv. Dans la régression décrite en question (iii), que se passe-t-il pour la significativité statistique individuelle des variables log(*income*) et *prppov* ? Ces variables sont-elles conjointement significatives ? (Calculez la *p*-valeur.) Qu'advient-il de vos précédentes réponses ?

v. Compte tenu des résultats des régressions précédentes, laquelle seriez-vous prêt à reporter avec le plus de confiance afin d'analyser si la composition ethnique d'une localité influence les prix des fast-foods locaux ?

C10. Utilisez les données du fichier ELEM94_95 afin de répondre à cette question. Les résultats peuvent être comparés à ceux du tableau 4.1. La variable dépendante *lavgsal* s'obtient à partir du logarithme du salaire moyen des enseignants. La variable *bs* correspond au rapport entre les prestations moyennes et le salaire moyen (par l'école).

i. Exécutez une régression simple de *lavgsal* sur *bs*. Le coefficient de pente estimé est-il statistiquement différent de zéro ? Est-il statistiquement différent de –1 ?

ii. Ajoutez les variables *lenrol* et *lstaff* à la régression de la question (i).

Qu'advient-il du coefficient de *bs* ? Comparez ces résultats à ceux du tableau 4.1 ?

iii. Comment se fait-il que l'écart-type estimé du coefficient *bs* est plus faible dans le modèle estimé en question (ii) que dans celui discuté en question (i) ? (*Astuce* : Qu'advient-il de la variance de l'erreur et du problème de multicolinéarité lorsque les variables *lenrol* et *lstaff* sont ajoutées ?)

iv. Comment se fait-il que le coefficient de *lstaff* soit négatif ? Son ordre de grandeur est-il important ?

v. Ajoutez maintenant la variable *lunch* à la régression. Si l'on conserve les autres facteurs inchangés, les enseignants sont-ils indemnisés pour enseigner aux élèves issus de milieux défavorisés ? Justifiez.

vi. Dans l'ensemble, les résultats que vous trouvez avec ELEM94_95 sont-ils conformes à ceux du tableau 4.1 ?

C11. Utilisez les données du fichier HTV afin de répondre à cette question. Voir aussi l'exercice C10 du chapitre 3.

i. Estimez le modèle de régression suivant :

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u$$

par les MCO et présentez les résultats sous la forme habituelle. Testez l'hypothèse nulle selon laquelle *educ* se trouve linéairement liée à *abil* contre l'alternative selon laquelle la relation est quadratique.

ii. En utilisant l'équation de la question (i), testez $H_0 : \beta_1 = \beta_2$ contre l'alternative bilatérale. Quelle est la *p*-valeur du test ?

iii. Ajoutez les deux variables de scolarité dans le supérieur à la régression de la question (i) et déterminez si elles sont statistiquement conjointement significatives.

iv. Quelle est la corrélation entre *tuit17* et *tuit18* ? Expliquez pourquoi l'utilisation de la moyenne des variables relatives aux frais de scolarité au cours des deux années pourrait être préférable à l'ajout de chaque variable séparément dans le modèle. Qu'advient-il lorsque vous n'utilisez que la valeur moyenne ?

v. Les résultats relatifs à la variable capturant la moyenne des frais de scolarité en question (iv) a-t-elle du sens si l'on souhaite avoir une interprétation causale des effets ? Quel pourrait être le mécanisme à l'œuvre ?

C12. Appuyez-vous sur les données de la base ECONMATH pour répondre aux questions suivantes.

i. Estimez un modèle expliquant la variable *colgpa* en fonction de *hsgpa*, *actmth*, et *acteng*. Reportez les résultats sous la forme habituelle. Les variables explicatives sont-elles toutes significatives ?

ii. Considérez une hausse de la variable *hsgpa* d'une unité d'écart-type, soit d'environ 0,343. De combien la variable *colgpa* va-t-elle augmenter toutes choses égales par ailleurs ? De combien de points d'écart-type la variable *actmth* devrait-elle augmenter pour faire varier *colgpa* de la même amplitude que précédemment ? Justifiez.

iii. Testez l'hypothèse nulle selon laquelle les variables *actmth* et *acteng* ont le même impact (au niveau de la population) contre l'alternative bilatérale. Reportez la *p*-valeur et analysez vos conclusions en détail.

iv. Supposez que la personne en charge des admissions à l'université souhaite que vous utilisiez les données relatives à la question (i) pour construire un modèle de régression expliquant au moins 50 % de la variation de *colgpa*. Que lui répondriez-vous ?

CHAPITRE 5

RÉGRESSION MULTIPLE : RÉSULTATS ASYMPTOTIQUES DES MCO

Traduction de Sophie Béreau

Dans les chapitres 3 et 5, nous avons couvert ce que l'on nomme les propriétés en *échantillon fini*, *petit échantillon* ou *propriétés exactes* des estimateurs des MCO dans le modèle issu de la population

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \qquad \textbf{[5.1]}$$

Par exemple, l'absence de biais des MCO (calculé au chapitre 3) sous les quatre premières hypothèses de Gauss-Markov est une propriété en échantillon fini car elle tient pour *tout* échantillon de taille n (sous la restriction minime que n doit être au moins aussi grand que le nombre total de paramètres du modèle de régression, soit $k + 1$). De même, le fait que l'estimateur des MCO est le meilleur estimateur de la classe des estimateurs linéaires sans biais (estimateur *BLUE*) sous l'ensemble des hypothèses de Gauss-Markov (RLM.1 à RLM.5) est une propriété en échantillon fini.

Dans le chapitre 4, nous avions ajouté la propriété classique du modèle de régression linéaire RLM.6, qui dit que le terme d'erreur u est normalement distribué et indépendant des variables explicatives. Cette condition permet de calculer la *vraie* distribution de l'estimateur des MCO (conditionnellement aux réalisations des variables explicatives de l'échantillon). Plus précisément, le théorème 4.1 montre que les estimateurs des MCO suivent des distributions normales. Ainsi, on en déduit que les statistiques t et F suivent des distributions de Student et de Fisher. Si les erreurs ne sont pas normalement distribuées, la distribution de la statistique t n'est pas exactement une Student, de même que la statistique F ne suit pas exactement une distribution de Fisher pour toute taille d'échantillon.

En sus des propriétés en échantillon fini, il est important de connaître les **propriétés asymptotiques** ou **propriétés en grand échantillon** des estimateurs et des statistiques de tests. Ces propriétés ne sont pas calculées pour une taille particulière de l'échantillon, mais sont définies à mesure que la taille de l'échantillon tend vers l'infini. Heureusement, sous les hypothèses que nous avons faites, l'estimateur des MCO a de bonnes propriétés asymptotiques. Un élément clé tient à ce que, même en l'absence de l'hypothèse de normalité des erreurs (hypothèse RLM.6), les statistiques t et F suivent *approximativement* des distributions de Student et de Fisher, lorsque les échantillons sont de grande taille. Nous discutons de ceci en détail dans la section 5.2, après avoir couvert la convergence de l'estimateur des MCO dans la section 5.1.

Du fait de la difficulté des notions abordées dans ce chapitre et de la possibilité de conduire des travaux empiriques sans une compréhension profonde de ces concepts, ce chapitre peut être ignoré en première lecture. Cependant, il est nécessaire de faire référence aux propriétés asymptotiques des estimateurs des MCO lorsque nous relâchons l'hypothèse d'homoscédasticité dans le chapitre 8 ou lorsque, nous étudions plus en détail l'estimation des modèles de séries chronologiques dans la partie 2. De plus, fondamentalement, toutes les méthodes avancées d'économétrie se justifient par des arguments asymptotiques ; les lecteurs qui poursuivent en partie 3 doivent donc être familiers des notions présentées dans ce chapitre.

5.1 CONVERGENCE

L'absence de biais des estimateurs, bien qu'étant une propriété importante, n'est pas toujours atteignable. Par exemple, comme discuté dans le chapitre 3, l'écart-type estimé de la régression, $\hat{\sigma}$, n'est pas un estimateur sans biais de σ, l'écart-type de l'erreur u dans le modèle de régression multiple. Bien que l'estimateur des MCO soit sans biais sous les hypothèses RLM.1 à RLM.4, nous verrons dans le chapitre 11 que, dans le cadre de certaines régressions en séries chronologiques, les estimateurs peuvent être biaisés. De plus, dans le partie 3 de cet ouvrage, nous étudions un certain nombre d'estimateurs biaisés et néanmoins utiles.

Alors que tous les estimateurs d'intérêt ne sont pas nécessairement sans biais, l'ensemble des économistes s'accordent sur le fait que la propriété de **convergence** est quant à elle, un minimum requis pour tout estimateur. L'économètre Clive W. J. Granger, lauréat du Prix Nobel d'Économie, a un jour fait la remarque suivante : « Si vous ne parvenez pas à avoir raison lorsque n tend vers l'infini, alors il vaut mieux changer de métier »[1]. En d'autres termes, si votre estimateur proposé pour un paramètre de population donné n'est pas convergent, alors vous perdez votre temps.

Il existe différentes approches pour décrire la convergence. Un ensemble de définitions et de résultats formels est donné en annexe C ; ici, nous nous concentrons sur la compréhension intuitive de ces concepts. De façon concrète, soit $\hat{\beta}_j$, l'estimateur des MCO de β_j pour un certain j. Pour tout n, $\hat{\beta}_j$ est caractérisé par une distribution de probabilités (représentant ses valeurs possibles dans les différents échantillons aléatoires de taille n). Du fait que $\hat{\beta}_j$ est sans biais sous les hypothèses RLM.1 à RLM.4, cette distribution est de moyenne β_j. Si l'estimateur est convergent, alors la distribution de $\hat{\beta}_j$ devient de plus en plus resserrée autour de la vraie valeur du paramètre β_j à mesure que la taille d'échantillon croît. Lorsque n tend vers l'infini, la distribution de $\hat{\beta}_j$ se réduit à un unique point β_j. Concrètement, cela signifie que si nous le voulons, nous pouvons faire en sorte que notre estimateur soit très proche de la vraie valeur du paramètre β_j en collectant suffisamment de données. La convergence est illustrée à la figure 5.1.

Naturellement, dans le cadre d'une application pratique, nous faisons face à une taille d'échantillon finie, ce qui explique qu'une propriété asymptotique comme la convergence soit difficile à saisir. La convergence nécessite de réaliser un exercice d'abstraction ou d'expérience de pensée (en anglais *thought experiment*) sur ce qui se passerait si la taille d'échantillon tendait à croître indéfiniment (alors qu'en même temps, nous pouvons considérer de nombreux échantillons aléatoires de différentes tailles finies). Si le fait d'obtenir de plus en plus de données ne nous permet pas de nous rapprocher de la vraie valeur du paramètre d'intérêt, alors nous utilisons une procédure d'estimation sans grande valeur.

Fort heureusement, le même jeu d'hypothèses implique à la fois l'absence de biais et la convergence de l'estimateur des MCO. Nous nous proposons de les résumer dans un théorème.

THÉORÈME 5.1
Convergence de l'estimateur des MCO

Sous les hypothèses RLM.1 à RLM.4, l'estimateur des MCO $\hat{\beta}_j$ est convergent pour β_j, pour tout $j = 0, 1, ..., k$.

1 *"If you can't get it right as n goes to infinity, you shouldn't be in this business"*

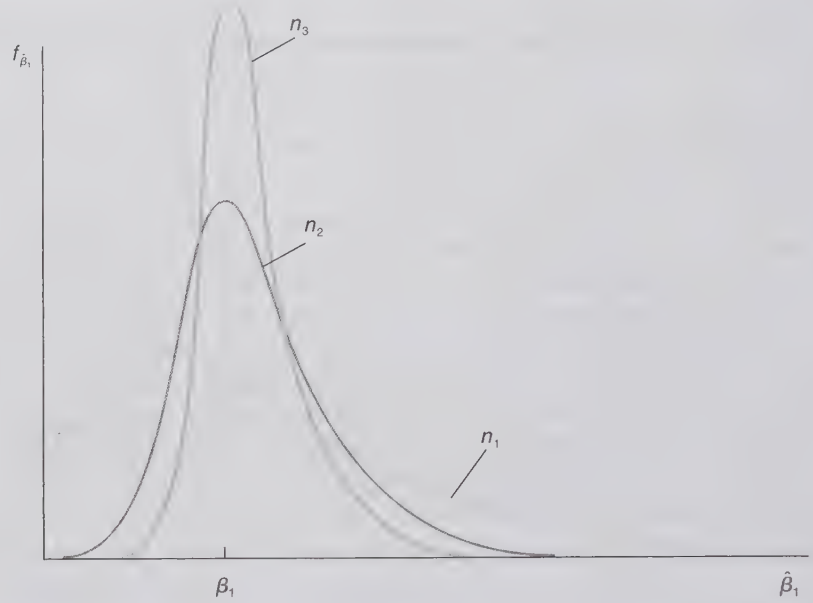


Figure 5.1 *Distributions d'échantillonnage de* β_1 *pour des tailles d'échantillons* $n_1 < n_2 < n_3$.

Une preuve générale de ce résultat est très facilement obtenue en utilisant les notations et méthodes matricielles développées dans les annexes D et E. Mais il est possible de démontrer le théorème 5.1 sans difficulté pour le modèle de régression simple. Nous nous concentrons sur l'estimateur du paramètre de pente, $\hat{\beta}_1$.

La démonstration démarre de la même manière que celle relative à l'absence de biais de l'estimateur : nous écrivons la formule pour $\hat{\beta}_1$ et l'introduisons dans $y_i = \beta_0 + \beta_1 x_{it} + u_i$:

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)\, y_i\right) \Big/ \left(\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2\right) \qquad \textbf{[5.2]}$$

$$= \beta_1 + \frac{\left(n^{-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)u_i\right)}{\left(n^{-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2\right)},$$

où nous divisons le numérateur et dénominateur par n, ce qui ne change donc rien à l'expression mais nous permet d'appliquer directement la loi des grands nombres. Lorsque nous appliquons la loi des grands nombres aux valeurs moyennes de l'équation (5.2), nous concluons que les numérateur et dénominateur convergent tous deux en probabilité vers les valeurs de la population, $\text{Cov}(x_1,u)$ et $\text{Var}(x_1)$, respectivement. Pour $\text{Var}(x_1) \neq 0$ – qui est obtenue par RLM.3 –, nous pouvons utiliser les propriétés des *limites en probabilités* (voir annexe C) comme suit :

$$\begin{aligned} \text{plim}\ \hat{\beta}_1 &= \beta_1 + \text{Cov}(x_1,u)/\text{Var}(x_1) \\ &= \beta_1 \text{ car } \text{Cov}(x_1,u) = 0 \end{aligned} \qquad \textbf{[5.3]}$$

Comme discuté dans les chapitres 2 et 3, nous avons utilisé ici la condition $E(u|x_1) = 0$ (hypothèse RLM.4) qui implique que x_1 et u sont décorrélées (c'est-à-dire qu'elles ont une covariance nulle).

D'un point de vue technique, il convient de s'assurer que les limites en probabilités existent. Nous devons donc faire l'hypothèse que $\text{Var}(x_1) < \infty$ et $\text{Var}(u) < \infty$ (ce qui implique que les distributions de probabilités ne sont pas trop étalées) mais nous ne nous inquiéterons pas ici des cas où ces hypothèses pourraient ne pas tenir. De plus, nous pourrions – et dans le cadre d'un cours avancé d'économétrie, nous le ferions très certainement – relâcher explicitement l'Hypothèse RLM.3 excluant la possibilité de parfaite colinéarité pour la seule population sous-jacente. Comme explicité, l'Hypothèse RLM.3 empêche en outre la possibilité de colinéarité parfaite entre les régresseurs dans l'échantillon étudié. Techniquement et pour être tout à fait rigoureux, il est possible de prouver la convergence des estimateurs sans imposer l'absence de colinéarité parfaite au niveau de la population, en supposant possible que nous tirions au hasard, par malchance, une base de données pour laquelle les données présenteraient de la colinéarité parfaite. D'un point de vue pratique la distinction est toutefois de peu d'importance dans la mesure où, dans les deux cas, nous sommes incapables d'estimer les paramètres par la méthode des moindres carrés ordinaires si notre échantillon ne satisfait pas l'Hypothèse RLM.3.

Les développements qui précèdent et l'équation (5.3) en particulier, montrent que l'estimateur des MCO dans le modèle de régression simple n'est convergent que si nous faisons l'hypothèse d'absence de corrélation entre x_1 et u. Ceci est également vrai dans le cas général. Nous établissons maintenant cette hypothèse.

HYPOTHÈSE RLM.4'
Espérance nulle et absence de corrélation

$E(u) = 0$ et $\text{Cov}(x_j, u) = 0$, pour $j = 1, 2, ..., k$.

L'hypothèse RLM.4' est moins forte que l'hypothèse RLM.4 dans le sens où cette dernière implique la première. Un moyen de caractériser l'hypothèse d'espérance conditionnelle nulle, $E(u|x_1, ..., x_k) = 0$, est d'imposer que *toute* fonction des variables explicatives soit décorrélée de u. L'hypothèse RLM.4' ne requiert quant à elle que le fait que chacune des variables x_j soit décorrélée de u (et que u soit de moyenne nulle dans la population). Dans le chapitre 2, nous motivons de fait l'estimateur des MCO pour le modèle de régression simple en faisant usage de l'hypothèse RLM.4', et les conditions du premier ordre pour les MCO données dans les équations (3.13) dans le cas du modèle de régression multiple, sont simplement l'équivalent des hypothèses d'absence de corrélation (et de l'hypothèse d'espérance nulle) au niveau de la population. De ce fait, l'hypothèse RLM.4' apparaît plus naturelle car elle mène directement aux estimations par les MCO. De plus, lorsque nous pensons aux éventuelles violations de l'hypothèse RLM.4, nous raisonnons en général en termes de $\text{Cov}(x_j, u) \neq 0$ pour certains j. Alors pourquoi avoir défini l'hypothèse RLM.4 de cette façon jusqu'à présent ? Il y a à cela deux raisons, toutes deux ayant été évoquées précédemment. La première tient au fait que l'estimateur des MCO est biaisé (mais convergent) sous l'hypothèse RLM.4' si $E(u|x_1, ..., x_k)$ dépend de l'un des x_j. Du fait que nous nous étions précédemment concentrés sur les propriétés statistiques exactes, c'est-à-dire en échantillon fini de l'estimateur des MCO, nous avions besoin de l'hypothèse plus forte sur la nullité de l'espérance conditionnelle.

La seconde motivation et probablement la plus importante, tient à ce que l'hypothèse de nullité de l'espérance conditionnelle signifie que nous avons correctement modélisé la fonction de régression de la population (FRP). Soit, sous l'hypothèse RLM.4 :

$$E(y|x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,$$

dès lors, nous sommes en mesure d'obtenir les effets marginaux des variables explicatives sur la valeur moyenne ou espérée de y. Si nous avions fait à la place l'hypothèse RLM.4', $\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$, ne

représenterait pas nécessairement la fonction de régression de la population et nous ferions face à la possibilité qu'une fonction non linéaire des x_j, telle que x_j^2, puisse être corrélée avec u. Une telle situation impliquerait que nous aurions négligé d'éventuelles non linéarités qui pourraient nous aider à mieux expliquer y ; si cela avait été le cas, nous aurions alors introduit de telles fonctions non linéaires. En d'autres termes, la plupart du temps, nous espérons obtenir une bonne estimation de la FRP, et l'hypothèse de nullité de l'espérance conditionnelle apparaît naturelle. Pour autant, l'hypothèse moins forte d'absence de corrélation s'avère utile pour interpréter l'estimation d'un modèle linéaire par la méthode des MCO comme étant celle proposant la meilleure approximation linéaire de la FRP. Cette hypothèse est également utilisée dans des cadres plus élaborés comme ceux abordés dans le chapitre 15, où notre intérêt n'est pas de modéliser la FRP. Pour plus de détails sur ce point subtile, voir Wooldridge (2010, chapitre 4).

Calculer la non convergence de l'estimateur des MCO

La simple violation de $E(u|x_1, \ldots, x_k) = 0$ engendre un biais de l'estimateur des MCO, la corrélation entre u et *n'importe laquelle* des variables $x_1, x_2, \ldots, x_k$ engendre en général la non convergence de *tous* les estimateurs des MCO des différents coefficients associés (c'est-à-dire la convergence vers d'autres valeurs que $\beta_0, \beta_1, \ldots, \beta_k$). Cette observation simple mais néanmoins essentielle peut être résumée en une phrase : *si l'erreur est corrélée avec n'importe laquelle des variables dépendantes, alors l'estimateur des MCO est biaisé et non convergent*. C'est évidemment très problématique puisque cela signifie que tout biais persistera à mesure que la taille d'échantillon croît.

Dans le cas de la régression simple, il est aisé de mettre en lumière la non convergence de l'estimateur des MCO à partir de l'équation (5.3), qui survient lorsque u et x_1 ne sont pas décorrélées. La **non convergence** de $\hat{\beta}_1$ (quelquefois dénommée **biais asymptotique**) est en effet donnée par :

$$\text{plim } \hat{\beta}_1 - \beta_1 = \text{Cov}(x_1,u)/\text{Var}(x_1). \qquad \textbf{[5.4]}$$

Puisque $\text{Var}(x_1) > 0$, la non convergence de $\hat{\beta}_1$ est positive si x_1 et u sont positivement corrélées et négative dans le cas contraire. Si la covariance entre x_1 et u est de faible ampleur relativement à la variance de x_1, la non convergence est négligeable ; malheureusement, nous ne sommes pas en mesure d'estimer le comportement de cette covariance car u est inobservable.

Nous pouvons utiliser la condition (5.4) pour calculer l'équivalent asymptotique du biais de variable omise (voir tableau 3.2 du chapitre 3). Supposons que le vrai modèle est donné par :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v,$$

et qu'il satisfait les quatre premières hypothèses de Gauss-Markov. Alors, l'erreur v est de moyenne nulle et décorrélée de x_1 et x_2. Si on note $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\beta}_2$, les estimateurs des MCO des paramètres de la régression de y sur x_1 et x_2, alors le théorème 5.1 implique que ces estimateurs sont convergents. Si nous omettons x_2 du modèle de régression et effectuons une régression simple de y sur x_1, alors $u = \beta_2 x_2 + v$. Soit $\tilde{\beta}_1$ l'estimateur de pente du modèle. On obtient :

$$\text{plim } \tilde{\beta}_1 = \beta_1 + \beta_2 \delta_1, \qquad \textbf{[5.5]}$$

avec :

$$\delta_1 = Cov(x_1,x_2)/Var(x_1). \qquad \textbf{[5.6]}$$

D'un point de vue pratique, nous pouvons interpréter cette non convergence comme étant de même nature qu'un biais. La différence tient à ce que cette non convergence est exprimée en termes de variance de x_1 et de covariance entre x_1 et x_2 au sein de la population, alors que le biais lui est exprimé sur base de

leurs contreparties en échantillon (puisque nous l'exprimons conditionnellement aux valeurs prises par x_1 et x_2 dans l'échantillon).

Si les variables x_1 et x_2 sont décorrélées (dans la population), alors $\delta_1 = 0$, et $\tilde{\beta}_1$ est un estimateur convergent de β_1 (bien qu'il ne soit pas nécessairement sans biais). Si x_2 a un effet marginal positif sur y, de sorte que $\beta_2 > 0$, et que x_1 et x_2 sont positivement corrélées, alors $\delta_1 > 0$, et le biais asymptotique de β_1 est alors positif. Nous pouvons obtenir le sens du biais asymptotique à partir du tableau 3.2. Si la covariance entre x_1 et x_2 est petite relativement à la variance de x_1, le biais asymptotique sera faible.

EXEMPLE 5.1
Prix de l'immobilier et proximité d'un incinérateur de déchets

Soit y, le prix d'une maison (*price*), x_1, la distance séparant la maison du nouvel incinérateur de déchets (*distance*), et x_2, la « qualité » de la maison (*quality*). La variable *quality* s'entend au sens large et peut inclure des éléments tels que la taille de la maison et du terrain, le nombre de chambres et de salles de bain ou des éléments moins tangibles tels que l'attractivité du quartier. Si la présence d'un incinérateur déprécie les prix de l'immobilier, alors β_1 doit être positif : toutes choses égales par ailleurs, une maison localisée à bonne distance de l'incinérateur doit valoir relativement plus cher. Par définition, β_2 est positif puisque plus la maison est de qualité, plus son prix de vente est accru, toutes autres choses égales par ailleurs. Si l'incinérateur a été construit plus loin en moyenne des maisons de meilleure qualité, alors *distance* et *quality* sont positivement corrélées, et $\delta_1 > 0$. Un modèle de régression simple de *price* sur *distance* [ou de log(*price*) sur log(*distance*)] aura tendance à surestimer l'effet de l'incinérateur : $\beta_1 + \beta_2\delta_1 > \beta_1$.

Pour aller plus loin 5.1

On suppose que le modèle

$$score = \beta_0 + \beta_1 skipped + \beta_2 priGPA + u$$

satisfait les quatre premières hypothèses de Gauss-Markov, avec *score*, la note obtenue à l'examen final, *skipped*, le nombre de cours manqués et *priGPA*, la valeur du GPA précédent le semestre en cours (soit la moyenne des notes obtenues aux examens à l'université). Si $\tilde{\beta}_1$ est issu d'un modèle de régression simple de *score* sur *skipped*, quelle est la direction attendue du biais asymptotique de $\tilde{\beta}_1$?

Une remarque importante relative à la non convergence de l'estimateur des MCO tient au fait que par définition, le biais ne disparaît pas lorsqu'on ajoute de nouvelles observations à l'échantillon. Au contraire, le problème devient même plus aigu puisque l'estimateur des MCO se rapproche de plus en plus de $\beta_1 + \beta_2\delta_1$ à mesure que la taille d'échantillon croît.

Déterminer le signe et l'amplitude du biais asymptotique dans le cas général de k régresseurs est plus complexe et se heurte aux mêmes difficultés que celles liées à l'établissement du biais en échantillon fini. Nous devons garder en mémoire que, si l'on considère le modèle de l'équation (5.1) où, par exemple, x_1 est corrélée avec u, les autres variables explicatives étant décorrélées avec u, *aucun* des estimateurs des MCO des paramètres du modèle n'est en général convergent. Par exemple, pour $k = 2$,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

supposons que la variable x_2 et l'erreur u sont décorrélées mais que x_1 et u sont corrélées. Alors les estimateurs des MCO $\hat{\beta}_1$ et $\hat{\beta}_2$ seront en général tous deux asymptotiquement biaisés, de même que la constante du modèle. La non convergence de $\hat{\beta}_2$ survient lorsque x_1 et x_2 sont corrélées, comme c'est le cas habituellement. Si x_1 et x_2 sont décorrélées, alors *aucune* corrélation entre x_1 et u n'entraînera de biais asymptotique de $\hat{\beta}_2$:

plim $\hat{\beta}_2 = \beta_2$. De plus, la non convergence de $\hat{\beta}_1$ est de même nature que celle décrite en (5.4). Les mêmes remarques prévalent dans le cas général : si x_1 est corrélée avec u, mais que x_1 et u sont décorrélées avec les autres variables indépendantes du modèle, alors, seul $\hat{\beta}_1$ sera asymptotiquement biaisé et sa non convergence sera caractérisée par l'expression (5.4). Le cas général est très proche du cas du biais de variable omise détaillé dans la section 3A.4 de l'annexe 3A.

5.2 NORMALITÉ ASYMPTOTIQUE ET INFÉRENCE EN GRAND ÉCHANTILLON

La convergence d'un estimateur est une propriété importante mais elle ne peut à elle seule garantir l'inférence statistique. Le simple fait de savoir qu'un estimateur se rapproche de la vraie valeur du paramètre de la population à mesure que la taille d'échantillon croît ne nous permet pas de tester des hypothèses relativement à ces paramètres. Pour mettre en œuvre des procédures de tests, nous avons besoin de connaître la distribution d'échantillonnage des estimateurs des MCO. Sous les hypothèses classiques du modèle de régression linéaire RLM.1 à RLM.6, le théorème 4.1 établit que les distributions d'échantillonnage sont normales. Ce résultat est à la base de la dérivation des distributions de Student et Fisher que nous utilisons en économétrie appliquée.

La normalité exacte des estimateurs des MCO repose en grande partie sur la normalité du terme d'erreur, u, dans la population. Si les erreurs u_1, u_2, ..., u_n sont des tirages aléatoires d'une distribution donnée autre que la loi normale, $\hat{\beta}_1$ ne sera pas normalement distribué, ce qui implique que les statistiques t ne suivront pas de distribution de Student, et de la même façon, que les statistiques F ne suivront pas de distribution de Fisher. Cela est potentiellement problématique car nos méthodes d'inférence reposaient jusqu'à présent sur la possibilité d'obtenir des valeurs critiques ou p-valeurs issues des distributions de Student ou de Fisher.

Rappelons que l'hypothèse RLM.6 revient à stipuler que la distribution de y sachant les réalisations des variables x_1, x_2, ..., x_k est normale. Puisque y est observé et que u ne l'est pas, il est en effet plus facile dans le cadre d'applications, de raisonner en considérant que la distribution de y est normale. Nous avons déjà étudié différents exemples où y ne pouvait pas être considérée comme conditionnellement normale. Une variable aléatoire distribuée normalement est symétrique autour de sa valeur moyenne, et peut prendre n'importe quelle valeur ponctuelle positive ou négative (mais avec une probabilité nulle), plus de 95 % de l'aire sous la distribution se localisant dans un intervalle de deux écarts-types de part et d'autre de la valeur moyenne.

Dans l'exemple 3.5, nous avons estimé un modèle expliquant le nombre d'arrestations d'hommes jeunes pour une année particulière (*narr86*). Dans la population considérée, la plupart des hommes ne sont pas arrêtés durant l'année, et la grande majorité a été arrêtée au plus une fois. (Dans l'échantillon de 2 725 hommes issus de la base de données CRIME1, moins de 8 % ont été arrêtés plus d'une fois en 1986.) Du fait que *narr86* ne prend que deux valeurs pour 92 % de l'échantillon, cette variable ne peut en aucun cas être considérée comme suivant une distribution normale dans la population.

Dans l'exemple 4.6, nous avons estimé un modèle expliquant le taux de participation au plan d'épargne 401(k) (*prate*). La distribution des fréquences empiriques (également appelée *histogramme*) de la figure 5.2 montre que la distribution de *prate* est très asymétrique à droite, plutôt que d'être normalement distribuée. En effet, plus de 40 % des observations de *prate* sont de valeur 100, indiquant 100 % de participation. Ceci viole l'hypothèse de normalité, même conditionnellement aux réalisations des variables explicatives.

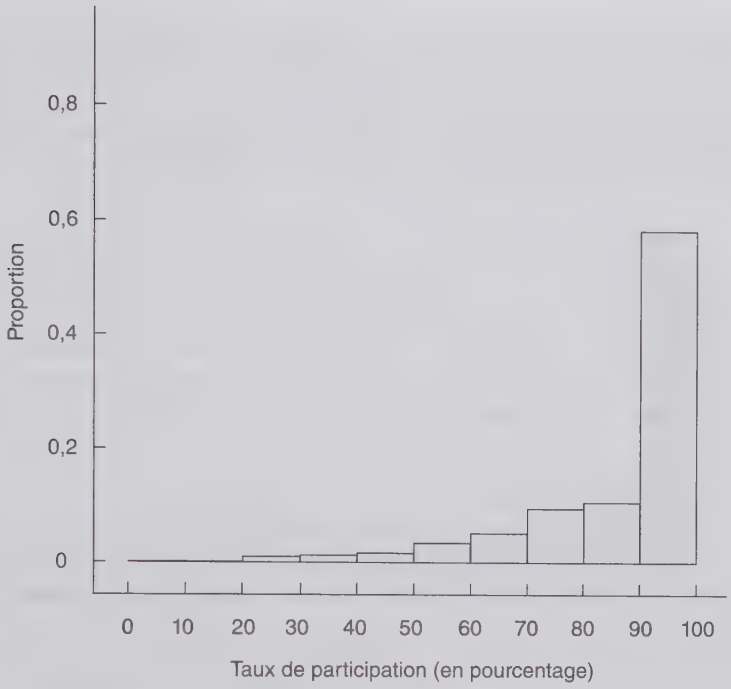


Figure 5.2 *Histogramme de* prate *utilisant les données de 401K.*

Nous savons que la normalité ne joue aucun rôle ni dans l'établissement de l'absence de biais de l'estimateur des MCO, ni dans les conclusions relatives aux caractéristiques *BLUE* de ce dernier sous les hypothèses de Gauss-Markov. Pour autant, l'inférence exacte qui repose sur les statistiques de Student et de Fisher, requiert l'hypothèse RLM.6. Cela signifie-t-il que, dans notre analyse de *prate* dans l'exemple 4.6, nous devons abandonner les statistiques *t* pour déterminer si les coefficients des variables sont statistiquement significatifs ? Heureusement, la réponse à cette question est *non*. Même si les y_i ne sont pas issus d'une distribution normale, nous pouvons recourir au théorème central limite rappelé en annexe C pour conclure que les estimations par les MCO satisfont la **normalité asymptotique**, ce qui signifie qu'ils se comportent approximativement comme des variables normalement distribuées dans des échantillons de taille suffisamment grande.

THÉORÈME 5.2
Normalité asymptotique de l'estimateur des MCO

Sous les hypothèses de Gauss-Markov RLM.1 à RLM.5, on a :

i. $\sqrt{n}(\hat{\beta}_j - \beta_j) \overset{a}{\sim}$ Normale$(0, \sigma^2 / a_j^2)$, avec $\sigma^2 / a_j^2 > 0$, la **variance asymptotique** de $\sqrt{n}(\hat{\beta}_j - \beta_j)$; pour les coefficients de pente, $a_j^2 = \text{plim}\left(n^{-1}\sum_{i=1}^{n}\hat{r}_{ij}^2\right)$, avec $\hat{r}_{ij}$ les résidus issus de la régression de x_j sur les autres variables indépendantes. On dit que $\hat{\beta}_j$ suit une *distribution normale asymptotique* (voir annexe C) ;

ii. $\hat{\sigma}^2$ est un estimateur convergent de $\sigma^2 = \text{Var}(u)$;

iii. Pour chaque j,

$$(\hat{\beta}_j - \beta_j) / \sigma(\hat{\beta}_j) \overset{a}{\sim} \text{Normale } (0,1)$$

et

$$(\hat{\beta}_j - \beta_j)/\hat{\sigma}(\hat{\beta}_j) \overset{a}{\sim} \text{Normale}(0,1), \qquad [5.7]$$

avec $\hat{\sigma}(\hat{\beta}_j)$ l'écart-type du paramètre estimé $\hat{\beta}_j$ obtenu par la méthode des MCO.

La preuve de la normalité asymptotique quelque peu complexe, est esquissée dans l'annexe pour le cas de la régression simple. L'assertion (ii) dérive de la loi des grands nombres, celle décrite en (iii) des résultats (i) et (ii) ainsi que des propriétés asymptotiques discutées en annexe C.

Le théorème 5.2 est utile en raison de l'abandon de l'hypothèse de normalité RLM.6 ; la seule restriction sur la distribution des erreurs tient à la variance finie, une hypothèse que nous conserverons systématiquement. Nous avons également supposé les hypothèses relatives à la moyenne conditionnelle (RLM.4) et l'homoscédasticité de u (RLM.5) valides.

En cherchant à comprendre le sens du théorème 5.2, il est essentiel de séparer les notions de distribution du terme d'erreur u pour la population, et les distributions d'échantillonnage des $\hat{\beta}_j$ à mesure que la taille d'échantillon croît. Il est courant de penser à tort que quelque chose va modifier la distribution de u – en particulier que celle-ci va se *« rapprocher »* de la distribution normale – à mesure que la taille d'échantillon croît. Rappelons que la distribution valant pour la population est donnée une bonne fois pour toute et est immuable, elle n'a donc rien à voir avec la taille d'échantillon. Par exemple, nous avions discuté précédemment le cas de la variable *narr86*, soit le nombre de fois qu'un homme jeune est arrêté durant l'année 1986. La nature de cette variable – elle ne prend que de petites valeurs entières limitées – est fixée dans la population. Que nous échantillonnions 10 ou 1 000 hommes issus de cette population, cela n'aura aucun impact sur la distribution valant pour la population.

Ce que nous dit le théorème 5.2, est que, indépendamment de la distribution de u valant pour la population, l'estimateur des MCO lorsqu'il est correctement standardisé, suit approximativement une distribution normale. Cette approximation vient de l'application du théorème central limite en raison de l'expression de l'estimateur des MCO à partir des moyennes empiriques – selon une combinaison complexe. En effet, la séquence de distributions des moyennes des erreurs sous-jacentes avoisine la normalité pour potentiellement toute distribution au sein de la population.

Il est à noter que quelque soit la manière dont on standardise les $\hat{\beta}_j$, en divisant la différence $\hat{\beta}_j - \beta_j$ par $\sigma(\beta_j)$ (que nous n'observons pas car il dépend de σ) ou par $\hat{\sigma}(\hat{\beta}_j)$ (que nous pouvons calculer à partir des données observées et qui dépend de $\hat{\sigma}$), ceux-ci suivent une distribution asymptotiquement normale. En d'autres termes, du point de vue asymptotique, cela ne fait aucune différence de remplacer la vraie valeur σ par $\hat{\sigma}$. Bien évidemment, ce faisant, nous modifions la distribution exacte du paramètre standardisé $\hat{\beta}_j$. Nous venons de voir au chapitre 4 que sous les hypothèses classiques du modèle linéaire, $(\hat{\beta}_j - \beta_j)/\sigma(\beta_j)$ suit exactement une distribution Normale(0,1) et $(\hat{\beta}_j - \beta_j)/\hat{\sigma}(\hat{\beta}_j)$ suit une distribution exacte de Student t_{n-k-1}.

Comment devons-nous prendre en compte le résultat de l'équation (5.7) ? Une conséquence notable est que, si nous nous apprêtons à réaliser une analyse en grand échantillon, nous devons alors faire usage de la distribution normale standard pour l'inférence plutôt que des distributions de Student. Mais d'un point de vue pratique, il est tout aussi légitime d'écrire :

$$(\hat{\beta}_j - \beta_j)/\hat{\sigma}(\hat{\beta}_j) \overset{a}{\sim} t_{n-k-1} = t_{ddl}, \qquad [5.8]$$

car t_{ddl} approxime la distribution Normale(0,1) lorsque le nombre de degrés de liberté *ddl* devient grand. Comme nous savons que par application des hypothèses classiques du modèle linéaire, la distribution de

Student t_{n-k-1} tient exactement, il est justifié de considérer que $(\hat{\beta}_j - \beta_j)/\hat{\sigma}(\hat{\beta}_j)$ suit une Student t_{n-k-1} en général, même lorsque l'hypothèse RLM.6 ne tient pas.

L'équation (5.8) nous dit que les tests de Student et la construction d'intervalles de confiance peuvent être réalisés *exactement* comme sous les hypothèses classiques du modèle linéaire. Cela signifie que notre analyse des variables dépendantes comme *prate* et *narr86* n'ont pas à être revues si les hypothèses de Gauss-Markov tiennent : dans les deux cas, nous disposons d'au moins 1 500 observations, ce qui est certainement suffisant pour justifier l'approximation issue de l'application du théorème central limite.

Si la taille d'échantillon n'est pas suffisante, alors les distributions de Student peuvent potentiellement être une mauvaise approximation de la distribution de la statistique t lorsque l'erreur u n'est pas normalement distribuée. Malheureusement, il n'y a pas de recommandation précise sur la taille optimale que doit avoir l'échantillon pour pouvoir réaliser sans risque cette approximation. Certains économètres pensent que $n = 30$ est satisfaisant, mais cela ne peut être correct pour toutes les distributions possibles de u. Selon la distribution de u, plus d'observations peuvent être nécessaires avant que le théorème central limite ne délivre une approximation valable. De plus, la qualité de l'approximation ne dépend pas simplement de n, mais des degrés de liberté *ddl*, $n - k - 1$: plus le nombre de variables indépendantes augmente dans le modèle, plus la taille d'échantillon doit être importante pour réaliser l'approximation de la distribution de la statistique t. L'étude des méthodes d'inférence pour de faibles degrés de liberté et sous l'hypothèse de non normalité des erreurs dépasse le cadre de cet ouvrage. Nous ferons simplement usage ici des statistiques t dans la mesure où nous nous trouvons toujours dans un cadre où la normalité est supposée acquise.

Il est très important de comprendre que le théorème 5.2 requiert l'hypothèse d'homoscédasticité (en sus de l'hypothèse de nullité de l'espérance conditionnelle). Si $\text{Var}(y|\mathbf{x})$ n'est pas constante, les statistiques de Student usuelles et les intervalles de confiance associés, ne sont pas valables quelque soit la taille de l'échantillon considéré ; le théorème central limite, ne nous prémunit pas contre les problèmes consécutifs à l'hétéroscédasticité. Pour cette raison, nous dédions l'entièreté du chapitre 8 à étudier ce qui peut être entrepris en présence d'hétéroscédasticité.

Une des conclusions du théorème 5.2 est que $\hat{\sigma}^2$ est un estimateur convergent de σ^2 ; nous savons déjà depuis le théorème 3.3 que $\hat{\sigma}^2$ est un estimateur sans biais de σ^2 sous les hypothèses de Gauss-Markov. La convergence implique que $\hat{\sigma}$ est un estimateur convergent de σ, condition qui s'avère importante dans l'établissement du résultat de normalité asymptotique de l'équation (5.7).

Rappelons que $\hat{\sigma}$ apparaît dans l'expression de l'écart-type estimé de chaque $\hat{\beta}_j$. En effet, la variance estimée de $\hat{\beta}_j$ est donnée par :

$$\widehat{Var(\hat{\beta}_j)} = \frac{\hat{\sigma}^2}{SCT_j(1 - R_j^2)} \qquad [5.9]$$

Pour aller plus loin 5.2

Dans un modèle de régression avec une taille d'échantillon importante, quel est l'intervalle de confiance de niveau 95 % approximatif pour β_j sous les hypothèses RLM.1 à RLM.5 ? Nous l'appellerons **intervalle de confiance asymptotique**.

avec SCT_j, la somme des carrés totaux des x_j dans l'échantillon, et R_j^2, le *R-carré* issu de la régression de x_j sur toutes les autres variables dépendantes du modèle. Dans la section 3.4, nous avons étudié chacune des composantes de l'équation (5.9) prises isolément, composantes sur lesquelles nous allons maintenant revenir dans le contexte de l'analyse asymptotique. À mesure que la taille d'échantillon croît, $\hat{\sigma}^2$ converge en

probabilité vers la valeur constante σ^2. De plus, R_j^2 s'approche d'un nombre strictement compris entre zéro et un (de telle sorte que $1 - R_j^2$ converge vers une valeur comprise entre zéro et un). La variance empirique de x_j est donnée par SCT_j/n, et de ce fait, SCT_j/n converge vers $\text{Var}(x_j)$ à mesure que la taille d'échantillon croît. Cela signifie que SCT_j croît à environ le même vitesse que la taille d'échantillon : $\text{SCT}_j \approx n\sigma_j^2$, avec σ_j^2, la variance de x_j valant pour la population. Quand nous combinons ces éléments, nous trouvons que $\text{Var}(\hat{\beta}_j)$ tend vers zéro à la vitesse de $1/n$; c'est pourquoi disposer d'échantillons plus grands est toujours préférable.

Lorsque l'erreur u n'est pas normalement distribuée, la racine carrée de l'équation (5.9) est parfois appelée **écart-type asymptotique**, et les statistiques t, les **statistiques *t* asymptotiques**. Du fait qu'il s'agit là des mêmes notions que celles abordées dans le chapitre 4, nous les appellerons unilatéralement écarts-types et statistiques t, en ayant à l'esprit que parfois ceux-ci ont des justifications asymptotiques. Un raisonnement analogue tient pour les **intervalles de confiance asymptotiques** construits à partir des écarts-types asymptotiques.

Usant des mêmes arguments que ceux évoqués précédemment pour la variance estimée, nous pouvons écrire que :

$$\hat{\sigma}(\hat{\beta}_j) \approx c_j / \sqrt{n} \qquad \textbf{[5.10]}$$

avec c_j une constante positive que ne dépend *pas* de la taille d'échantillon. En effet, il est possible de montrer que la constante c_j est définie par :

$$c_j = \frac{\sigma}{\sigma_j \sqrt{1 - \rho_j^2}}$$

avec $\sigma = \sigma(u)$, $\sigma_j = \sigma(x_j)$, et ρ_j^2, le *R-carré* sur la population issu de la régression de x_j sur les autres variables explicatives du modèle. Comme étudié dans l'équation (5.9) pour voir quelles variables affectent $\text{Var}(\hat{\beta}_j)$ sous les hypothèses de Gauss-Markov, nous devons utiliser cette expression pour c_j de façon à étudier l'influence des écarts-types de grande taille (σ), d'une plus forte variation des x_j (σ_j) ou de multicolinéarité dans la population (ρ_j^2).

L'équation (5.10) n'est qu'une approximation, mais c'est un point de repère utile : les écarts-types peuvent diminuer à une vitesse inversement proportionnelle à la *racine carrée* de la taille d'échantillon.

La normalité asymptotique des estimateurs des MCO implique en outre que les statistiques F suivent approximativement une distribution de Fisher pour les grands échantillons. De ce fait, dans le cadre de restrictions d'exclusion ou d'autres hypothèses multiples, rien ne change dans ce que nous avons étudié jusqu'à maintenant.

EXEMPLE 5.2
Écarts-types estimés et modélisation du poids des nourrissons

Nous reprenons les données contenues dans la base BWGHT pour estimer la relation entre le log du poids des nourrissons comme variable dépendante, et le nombre de cigarettes fumées par jour par la mère (*cigs*) ainsi que le log du revenu de la famille comme variables indépendantes. Le nombre total d'observations est de 1 388 points. Utilisant la moitié des observations (694), l'écart-type estimé de $\hat{\beta}_{\text{cigs}}$ est d'environ 0,0013. L'écart-type estimé sur l'ensemble des observations est quant à lui d'environ 0,00086. Le ratio des deux est donné par $0{,}00086/0{,}0013 \approx 0{,}662$, ce qui est relativement proche de $\sqrt{694/1\,388} \approx 0{,}707$, soit le ratio obtenu par l'approximation de (5.10). En d'autres termes, l'équation (5.10) suggère que les écarts-types estimés sur des échantillons de plus grande taille devraient représenter environ 70.7 % des écarts-types estimés sur des échantillons de taille plus modeste. Ce pourcentage est relativement proche des 66.2 % effectivement calculés.

Autres tests en grand échantillon : la statistique du multiplicateur de Lagrange

Maintenant que nous sommes entrés dans le monde merveilleux de l'analyse asymptotique, d'autres statistiques peuvent être utilisées pour la mise en œuvre des tests d'hypothèses. Dans la plupart des cas, il y a peu d'intérêt à aller au-delà des tests de Student et de Fisher : comme nous venons de le voir, ces statistiques, se justifient en grand échantillon en l'absence de l'hypothèse de normalité. Néanmoins, il est parfois utile de tester des restrictions d'exclusions multiples par d'autres moyens, et, à cet effet, nous abordons maintenant le cas de la **statistique du multiplicateur de Lagrange** (***LM***), qui s'est révélée très populaire dans la pratique de l'économétrie moderne.

La dénomination de la statistique du multiplicateur de Lagrange ou *Lagrange multiplier statistic* en anglais dérive de la terminologie usitée dans le cadre de programmes d'optimisation sous contraintes, un sujet qui dépasse le cadre de cet ouvrage. [voir Davidson et MacKinnon (1993).] En anglais, l'appellation de *score statistic* – qui là encore dérive de l'approche consistant à résoudre un programme d'optimisation à l'aide d'outils de calcul différentiel – peut également être utilisée en pratique. Dans le cadre du modèle de régression linéaire, il est assez simple de motiver l'usage du test du multiplicateur de Lagrange sans avoir à entrer dans ces considérations mathématiques complexes.

La forme de la statistique *LM* que nous calculons ici repose sur les hypothèses de Gauss-Markov, soient les mêmes hypothèses qui justifient le recours à la statistique de Fisher en grands échantillons. L'hypothèse de normalité n'est donc pas requise.

Pour calculer la statistique *LM*, nous considérons le modèle de régression linéaire multiple avec k variables indépendantes suivant :

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u. \qquad \textbf{[5.11]}$$

Nous souhaitons tester la nullité des q dernières variables au niveau de la population, l'hypothèse nulle est alors donnée par :

$$H_0 : \beta_{k-q+1} = 0, \ldots, \beta_k = 0, \qquad \textbf{[5.12]}$$

et s'assimile à q restrictions d'exclusion sur le modèle (5.11). À l'instar des tests de Fisher, l'hypothèse alternative à (5.12) est que l'un au moins des paramètres s'avère différent de zéro.

La statistique *LM* requiert la seule estimation du modèle *contraint*. En effet, supposons que nous ayons estimé le modèle suivant :

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \ldots + \tilde{\beta}_{k-q} x_{k-q} + \tilde{u}, \qquad \textbf{[5.13]}$$

où « ~ » indique que les estimations sont issues du modèle contraint. En particulier, $\tilde{u}$ renvoie aux résidus du modèle contraint. (Comme toujours, ceci est une manière de dire de façon abrégée que nous avons obtenu les résidus contraints pour chacune des observations de l'échantillon à l'issue de la procédure d'estimation.)

Si les variables omises x_{k-q+1} à x_k ont véritablement des coefficients associés nuls au niveau de la population, alors, $\tilde{u}$ devrait être décorrélé avec chacune de ces variables dans l'échantillon au moins approximativement. Cela suggère de régresser $\tilde{u}$ sur les variables indépendantes exclues sous H_0, ce qui est approximativement ce que le test *LM* fait. Cependant, il s'avère que pour obtenir une statistique de test utilisable, nous devons inclure *toutes* les variables indépendantes de la régression. (Nous devons inclure tous les régresseurs car, en général, les régresseurs omis dans le modèle contraint sont corrélés avec les régresseurs qui apparaissent dans le modèle.) De ce fait, nous régressons alors :

$$\tilde{u} \text{ sur } x_1, x_2, \ldots, x_k. \qquad \textbf{[5.14]}$$

Ceci est un exemple de **régression auxiliaire**, une régression utilisée pour calculer une statistique de test mais dont les coefficients estimés n'ont pas d'intérêt en tant que tels.

Comment peut-on faire usage des résultats d'estimation du modèle (5.14) pour tester (5.12) ? Si (5.12) est vraie, alors le *R-carré* de (5.14) devrait être « proche » de zéro, aux erreurs d'échantillonnage près, puisque $\tilde{u}$ devrait être approximativement décorrélée de toutes les variables indépendantes du modèle. La question, comme toujours avec les tests d'hypothèses, est de savoir comment déterminer à quels moments la valeur de la statistique est suffisamment grande pour rejeter l'hypothèse nulle à un niveau de confiance donné. Il s'avère que, sous l'hypothèse nulle, la taille d'échantillon multipliée par le *R-carré* traditionnel issu de la régression auxiliaire (5.14) est distribuée asymptotiquement comme une variable aléatoire du chi-deux à q degrés de liberté. Cette propriété mène à une procédure simple pour tester la significativité jointe d'un ensemble de q variables indépendantes.

La statistique du multiplicateur de Lagrange pour q restrictions d'exclusion :

i. Régresser y sur l'ensemble des variables indépendantes du modèle *contraint* et préserver les valeurs des résidus, $\tilde{u}$.

ii. Régresser $\tilde{u}$ sur *l'ensemble* des variables indépendantes et calculer le *R-carré*, noté, R_u^2 (pour le distinguer du *R-carré* obtenu avec y comme variable dépendante).

iii. Calculer $LM = nR_u^2$ [la taille d'échantillon fois le *R-carré* de la régression auxiliaire réalisée à l'étape (ii)].

iv. Comparer LM à la valeur critique pertinente, c, pour une distribution χ_q^2 ; si $LM > c$, l'hypothèse nulle est rejetée au niveau de confiance choisi. Encore mieux, calculer la p-valeur comme la probabilité qu'une variable aléatoire suivant une χ_q^2 excède la valeur de la statistique calculée. Si la p-valeur est inférieure au niveau de confiance désiré alors, H_0 est rejetée. Dans le cas contraire, nous échouons à rejeter H_0. La règle de décision est fondamentalement la même que dans le cadre d'un test de Fisher.

Du fait de sa forme, la statistique LM est parfois désignée comme étant la **statistique *n*-*R*-carré**. Contrairement à la statistique de Fisher, les degrés de liberté du modèle non contraint ne jouent aucun rôle dans la mise en œuvre du test. Tout ce qui compte tient au nombre de restrictions considérées (q), à la valeur du *R-carré* de la régression auxiliaire (R_u^2), et à la taille d'échantillon (n). Les degrés de liberté (*ddl*) du modèle non contraint ne jouent aucun rôle du fait de la nature asymptotique de la statistique LM. Nous devons par contre être certains de multiplier R_u^2 par la taille d'échantillon pour obtenir la valeur de la statistique LM ; une valeur en apparence faible du *R-carré* auxiliaire pouvant malgré tout mener au non-rejet de l'hypothèse de significativité jointe si n est grand.

Avant d'illustrer ceci par un exemple, il convient de mentionner quelques conseils de prudence. Si dans l'étape (i), nous régressons y sur toutes les variables indépendantes et obtenons les résidus de cette régression non contrainte utilisés ensuite dans l'étape (ii), nous n'obtenons pas une statistique pertinente puisque le *R-carré* qui en résultera aura pour valeur de zéro exactement ! Ceci vient du fait que les MCO identifient les valeurs estimées des paramètres de telle sorte que les résidus de l'échantillon soient décorrélés des variables indépendantes incluses dans le modèle [voir les équations (3.13)]. Dès lors, nous ne pouvons tester (5.12) qu'en régressant les résidus contraints sur *toutes* les variables indépendantes. (Régresser les résidus contraints sur l'ensemble restreint des variables indépendantes mènerait au même résultat, c'est-à-dire à $R^2 = 0$.)

EXEMPLE 5.3
Un modèle économique de la criminalité

Nous illustrons le test *LM* en proposant une petite extension du modèle de criminalité étudié dans l'exemple 3.5 :

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u,$$

avec :

narr86 = le nombre de fois qu'un homme a été arrêté
pcnv = la proportion d'arrestations préalables ayant mené à la condamnation
avgsen = la durée moyenne des peines réalisées dans le passé
tottime = temps total que l'homme a passé en prison avant 1986 depuis l'âge de 18 ans
ptime86 = nombre de mois passés en prison en 1986
qemp86 = nombre de trimestres en 1986 durant lesquels l'homme était employé légalement

Nous utilisons la statistique *LM* pour tester l'hypothèse nulle que *avgsen* et *tottime* n'ont pas d'effet sur *narr86* une fois l'influence des autres facteurs prise en compte.

Dans l'étape (i), nous estimons le modèle contraint en régressant *narr86* sur *pcnv*, *ptime86* et *qemp86* ; les variables *avgsen* et *tottime* étant exclues de cette régression. Nous obtenons les résidus $\tilde{u}$ pour cette régression, 2 725 au total. Puis nous régressons :

$$\tilde{u} \text{ sur } pcnv, ptime86, qemp86, avgsen, \text{ et } tottime ; \qquad \textbf{[5.15]}$$

comme toujours l'ordre des variables indépendantes importe peu. De cette seconde régression, on tire R_u^2, qui est évalué ici à 0,0015. Cela peut sembler petit mais nous devons multiplier cette valeur par la taille d'échantillon pour obtenir la valeur de la statistique *LM* : $LM = 2\,725(0{,}0015) \approx 4{,}09$. La valeur critique au seuil de 10 % d'une distribution du chi-deux à deux degrés de liberté est d'environ 4,61 (arrondie à deux chiffres après la virgule ; voir tableau G.4). Dès lors, nous échouons à rejeter l'hypothèse nulle que $\beta_{avgsen} = 0$ et $\beta_{tottime} = 0$ au seuil de 10 %. Plus généralement, la *p*-valeur étant de $P(\chi_2^2 > 4{,}09) \approx 0{,}129$, nous rejetons donc H_0 au seuil de 15 %.

À titre de comparaison, le test de Fisher pour la significativité jointe des coefficients de *avgsen* et *tottime* mène à une *p*-valeur d'environ 0,131, ce qui est assez proche du résultat obtenu avec la statistique *LM*. Il n'est donc pas surprenant qu'asymptotiquement, les deux statistiques aient la même probabilité d'erreur de première espèce. (C'est-à-dire la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie.)

Comme l'exemple précédent l'illustre, avec un échantillon de grande taille, il est rare d'observer des différences notables entre les résultats des tests *LM* et de Fisher. Nous ferons usage des statistiques de Fisher la plupart du temps car ces dernières sont en général calculées par défaut par la plupart des logiciels économétriques standards. Vous devez toutefois garder à l'esprit l'existence de la statistique *LM* qui est également utilisée dans de nombreux travaux appliqués.

Pour finir, il est à noter qu'à l'instar de la statistique de Fisher, nous devons être certains d'utiliser les mêmes observations dans les étapes (i) et (ii). Si des données sont manquantes pour l'une ou l'autre des variables indépendantes exclues sous l'hypothèse nulle, les résidus obtenus à l'étape (i) doivent alors être obtenus sur ce même sous-ensemble de données.

5.3 EFFICACITÉ ASYMPTOTIQUE DE L'ESTIMATEUR DES MCO

Nous savons que, sous les hypothèses de Gauss-Markov, l'estimateur des MCO est le meilleur estimateur de la classe des estimateurs linéaires sans biais (estimateur dit *BLUE*). L'estimateur des MCO est également, sous ces mêmes hypothèses, **asymptotiquement efficace** parmi une certaine classe d'estimateurs. Une analyse du

cas général nécessiterait des notions d'algèbre linéaire et d'analyse asymptotique avancées. Dans un premier temps, nous décrivons le résultat dans le cas de la régression simple.

Considérons le modèle suivant :

$$y = \beta_0 + \beta_1 x + u, \qquad \textbf{[5.16]}$$

où l'erreur u est supposée être de moyenne nulle sous RLM.4 : $E(u|x) = 0$. Ceci ouvre la voie à une variété d'estimateurs convergents pour β_0 et β_1 ; comme toujours, nous nous concentrons sur l'estimateur du paramètre de pente, β_1. Soit $g(x)$, une fonction quelconque de x ; par exemple, $g(x) = x^2$ ou $g(x) = 1/(1 + |x|)$. Alors, u est décorrélée de $g(x)$ (voir les propriétés CE.5 de l'annexe B). Soit $z_i = g(x_i)$ pour toute observation i. L'estimateur

$$\tilde{\beta} = \left(\sum_{i=1}^{n} (z_i - \bar{z}) y_i \right) \Big/ \left(\sum_{i=1}^{n} (z_i - \bar{z}) x_i \right) \qquad \textbf{[5.17]}$$

est convergent pour β_1, si $g(x)$ et x sont corrélées. [Rappelez-vous qu'il est possible que $g(x)$ et x soient décorrélées car la corrélation mesure la dépendance *linéaire* entre variables.] Pour comprendre cette assertion, nous pouvons introduire l'équation $y_i = \beta_0 + \beta_1 x_i + u_i$ et écrire $\tilde{\beta}_1$ comme :

$$\tilde{\beta} = \beta_1 + \left(n^{-1} \sum_{i=1}^{n} (z_i - \bar{z}) u_i \right) \Big/ \left(n^{-1} \sum_{i=1}^{n} (z_i - \bar{z}) x_i \right) \qquad \textbf{[5.18]}$$

Par application de la loi des grands nombres, les numérateur et dénominateur convergent en probabilité vers $\text{Cov}(z,u)$ et $\text{Cov}(z,x)$, respectivement. Si $\text{Cov}(z,x) \neq 0$ – de sorte que z et x sont corrélées – nous obtenons :

$$\text{plim}\ \tilde{\beta}_1 = \beta_1 + \text{Cov}(z,u)/\text{Cov}(z,x) = \beta_1,$$

puisque $\text{Cov}(z,u) = 0$ sous l'hypothèse RLM.4.

Il est plus difficile de montrer que $\tilde{\beta}_1$ est asymptotiquement normal. Néanmoins, en usant d'arguments similaires à ceux développés dans la partie annexe de ce chapitre, on peut montrer que $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$ est asymptotiquement normal de moyenne nulle et de variance asymptotique $\sigma^2 \text{Var}(z)/[\text{Cov}(z,x)]^2$. La variance asymptotique de l'estimateur des MCO est obtenue en fixant $z = x$, auquel cas, $\text{Cov}(z,x) = \text{Cov}(x,x) = \text{Var}(x)$. Dès lors, la variance asymptotique de $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$, où $\tilde{\beta}_1 = \hat{\beta}_1$ est l'estimateur du paramètre de pente par les MCO, est donnée par : $\sigma^2 \text{Var}(x)/[\text{Var}(x)]^2 = \sigma^2/\text{Var}(x)$. Par application de l'inégalité de Cauchy-Schwartz (voir l'annexe B.4) : $[\text{Cov}(z,x)]^2 \leq \text{Var}(z)\text{Var}(x)$, la variance asymptotique de $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ n'est donc pas plus grande que celle de $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$. Nous avons montré, dans le cas de la régression simple, que sous les hypothèses de Gauss-Markov, l'estimateur des MCO a une variance asymptotique plus faible que n'importe quel estimateur de la forme décrite en équation (5.17). [L'estimateur en (5.17) est un exemple d'*estimateur par la méthode des variables instrumentales*, que nous étudierons en détails dans le chapitre 15.] Si l'hypothèse d'homoscédasticité est violée, il existe des estimateurs de la forme décrite par l'équation (5.17) qui présentent alors une variance asymptotique plus faible que celle de l'estimateur des MCO. Nous discutons de cette question en détails dans le chapitre 8.

Le cas général est similaire mais beaucoup plus difficile à établir d'un point de vue formel. En considérant k régresseurs, la classe des estimateurs convergents est obtenue en généralisant les conditions du premier ordre permettant d'identifier l'estimateur des MCO :

$$\sum_{i=1}^{n} g_j(x_i)(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \ldots - \tilde{\beta}_k x_{ik}) = 0,\ j = 0, 1, \ldots, k. \qquad \textbf{[5.19]}$$

avec $g_j(\boldsymbol{x}_i)$ une fonction quelconque de l'ensemble des variables explicatives pour l'observation i. En comparant l'équation (5.19) avec les conditions du premier ordre permettant l'identification de l'estimateur des MCO

en (3.13), il apparaît clairement que l'estimateur des MCO constitue un cas particulier à savoir $g_0(\boldsymbol{x}_i) = 1$ et $g_j(\boldsymbol{x}_i) = x_{ij}$ pour tout $j = 1, 2, ..., k$. La classe d'estimateurs décrite en (5.19) est infinie puisque nous pouvons utiliser n'importe quelle fonction des x_{ij}.

THÉORÈME 5.3
Efficacité asymptotique des MCO

Sous les hypothèses de Gauss-Markov, on considère $\tilde{\beta}_j$, les estimateurs solutions des équations de la forme (5.19) et $\hat{\beta}_j$, les estimateurs par les MCO des paramètres du modèle de régression linéaire multiple. Pour $j = 0, 1, 2, ..., k$, les estimateurs des MCO sont associés aux plus petites variances asymptotiques : Avar $\sqrt{n}(\hat{\beta}_j - \beta_j) \leq$ Avar $\sqrt{n}(\tilde{\beta}_j - \beta_j)$.

Prouver la convergence des estimateurs issus de (5.19), sans parler de la normalité asymptotique, est mathématiquement ardu, voir Wooldridge (2010, chapitre 5).

RÉSUMÉ

Les éléments contenus dans ce chapitre sont pour la plupart relativement techniques mais leurs implications pratiques sont essentielles. Nous avons d'abord montré que les quatre premières hypothèses de Gauss-Markov impliquaient la convergence de l'estimateur des MCO. De plus, l'ensemble des méthodes d'inférence visant à la mise en œuvre de tests statistiques et à la construction d'intervalles de confiance abordés dans le chapitre 4 sont approximativement valides sans recourir à l'hypothèse de normalité des erreurs (de façon équivalente, il n'est pas requis que la distribution de y conditionnellement aux variables explicatives soit normale). Cela signifie que nous pouvons appliquer les MCO ainsi que les méthodes d'inférence statistique associées, pour un ensemble d'applications où la variable dépendante n'est même pas approximativement normale. Nous avons également montré que la statistique *LM* peut être mise à profit comme alternative à la statistique de Fisher pour tester les restrictions d'exclusion.

Avant de clore ce chapitre, il convient de rappeler que les exemples tels que l'exemple 5.3 met en exergue un certain nombre de problèmes qui requièrent des solutions spécifiques. Pour une variable telle que *narr86*, qui prend les valeurs zéro ou un pour la plupart des hommes de la population, un modèle linéaire n'est peut-être pas le plus adapté pour capturer la relation fonctionnelle entre *narr86* et ses variables explicatives. De plus, même si un modèle linéaire décrit effectivement la valeur espérée du nombre d'arrestations, la présence d'hétéroscédasticité dans le terme d'erreur peut être un problème. De telles violations des hypothèses de base ne peuvent être réglées avec l'accroissement de la taille d'échantillon, nous y reviendrons dans des chapitres ultérieurs.

MOTS-CLÉS

Asymptotiquement efficace p. 223
Biais asymptotique p. 214
Convergence p. 211
Écart-type asymptotique p. 220
Intervalle de confiance asymptotique p. 220
Non convergence p. 214
Normalité asymptotique p. 217
Propriétés asymptotiques p. 210
Propriétés en grand échantillon p. 210
Régression auxiliaire p. 222

EXERCICES

1. Dans le modèle de régression simple sous les hypothèses RLM.1 à RLM.4, nous avons postulé que l'estimateur de pente, $\hat{\beta}_1$, était un estimateur convergent de β_1. En utilisant $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1$, montrez que plim $\hat{\beta}_0 = \beta_0$. [Vous devez utiliser la convergence de β_1 ainsi que la loi des grands nombres, tout en notant que $\beta_0 = E(y) - \beta_1 E(x_1)$.]

2. Supposons que le modèle suivant :

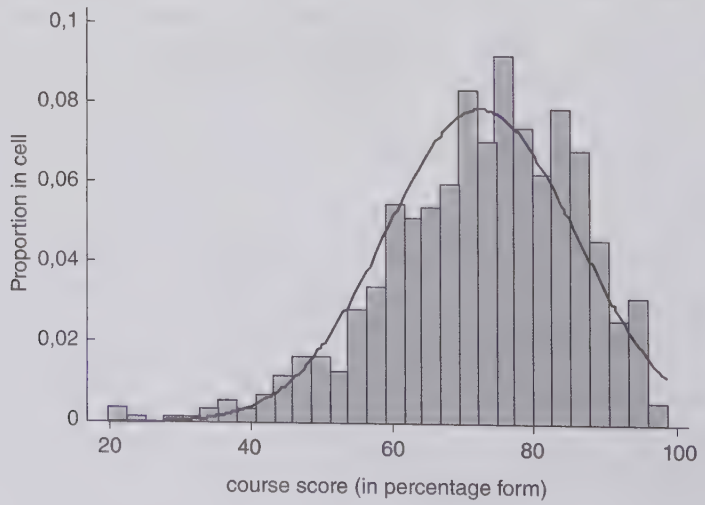
$$pctstck = \beta_0 + \beta_1 funds + \beta_2 risktol + u$$

satisfait les quatre premières hypothèses de Gauss-Markov, avec *pctstck* la part de la retraite des travailleurs investie dans des fonds de pension, *funds* le nombre de fonds mutuels entre lesquels le travailleur peut choisir d'investir son argent, et *risktol* une mesure de tolérance au risque (plus *risktol* est grand, plus une personne présente une tolérance aigue à l'égard du risque). À supposer que *funds* et *risktol* soient corrélées positivement, quel est le biais asymptotique de $\tilde{\beta}_1$, le paramètre de pente du modèle de régression de *pctstck* sur *funds* ?

3. La base de données SMOKE contient un ensemble d'informations sur le comportement à l'égard de la cigarette ainsi que d'autres variables pour un échantillon aléatoire d'adultes célibataires résidant aux États-Unis. La variable *cigs* correspond au nombre moyen de cigarettes fumées par jour. Pensez-vous que *cigs* soit distribuée normalement dans la population adulte aux États-Unis ? Justifiez.

4. Dans le modèle de régression simple présenté en (5.16), sous les quatre premières hypothèses de Gauss-Markov, nous avons montré que les estimateurs de la forme (5.17) sont convergents pour le paramètre β_1. Soit un tel estimateur, définissez un estimateur de β_0 par : $\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1\bar{x}$. Montrez que plim $\tilde{\beta}_0 = \beta_0$.

5. L'histogramme représenté ci-dessous a été généré à partir de la variable *core* issue de la base de données ECONMATH. Trente bâtons ont été utilisés pour tracer l'histogramme et la hauteur de chacune des cellules correspond à la proportion des observations associées aux différents intervalles considérés. La densité de la distribution Normale appropriée à nos données – c'est-à-dire celle dont les paramètres de moyenne et de variance ont été choisis pour correspondre à la moyenne et la variance d'échantillonnage – a été surimposée sur l'histogramme.



i. Si l'on recourt à la distribution Normale pour modéliser notre variable d'intérêt ici le *score*, peut-on dire que la probabilité que sa valeur excède 100 équivaut à zéro ? Pourquoi cette réponse contredit-elle l'hypothèse de distribution Normale pour la variable *score* ?

ii. Que constatez-vous au niveau de la queue de distribution à gauche ? La distribution Normale vous parait-elle bien se comporter sur son extrémité gauche ?

EXERCICES SUR ORDINATEUR

C1. On considère les données contenues dans le fichier WAGE1 pour cet exercice.

i. Estimez l'équation

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

Récupérez les résidus et réalisez un histogramme.

ii. Répétez l'étape (i), mais avec log(*wage*) comme variable dépendante.

iii. Diriez-vous que l'hypothèse RLM.6 est plus proche d'être satisfaite pour le modèle en niveau ou en log-niveau ?

C2. Utilisez les données contenues dans GPA2 pour cet exercice.

i. Utilisez les 4 137 observations, estimez l'équation

$$colgpa = \beta_0 + \beta_1 hsperc + \beta_2 sat + u$$

et reportez les résultats sous une forme standard.

ii. Ré-estimez l'équation de la question (i), en n'utilisant que les 2 070 premières observations.

iii. Calculez le ratio des écarts-types estimés du coefficient de *hsperc* dans le cadre des questions (i) et (ii). Comparez-le avec le résultat de l'équation (5.10).

C3. Dans l'équation (4.42) du chapitre 4, utilisez la base de données BWGHT, calculez la statistique *LM* pour tester si *motheduc* et *fatheduc* sont conjointement significatifs. En récupérant les résidus du modèle contraint, prenez garde à ce que le modèle contraint soit estimé avec les seules observations disponibles pour toutes les variables du modèle non contraint (voir l'exemple 4.9).

C4. Plusieurs statistiques sont couramment utilisées pour détecter une éventuelle non normalité dans les distributions de la population sous-jacente. Nous allons nous concentrer dans cet exercice sur l'une d'entre elles qui mesure la quantité d'asymétrie (ou coefficient de « skewness ») de la distribution. Rappelez-vous que toute variable aléatoire caractérisée par une distribution normale est symétrique autour de sa valeur moyenne ; dès lors si nous standardisons une variable aléatoire caractérisée par une distribution symétrique, disons $z = (y - \mu_y)/\sigma_y$, avec $\mu_y = E(y)$ et σ_y, l'écart-type de y, alors z est de moyenne nulle, de variance un, et $E(z^3) = 0$. Soit un échantillon de données observées $\{y_i : i = 1, ..., n\}$, nous pouvons standardiser y_i dans l'échantillon en utilisant $z_i = (y_i - \hat{\mu}_y)/\hat{\sigma}_y$ avec $\hat{\mu}_y$, la moyenne empirique et $\hat{\sigma}_y$, l'écart-type empirique. (Nous ignorons ici le fait que ces estimations sont réalisées sur base de l'échantillon.) Une statistique empirique mesurant le coefficient d'asymétrie est donnée par $n^{-1}\sum_{i=1}^{n} z_i^3$ ou par la même statistique avec $(n - 1)$ remplaçant n de façon à ajuster le nombre de degrés de liberté. Si y est caractérisé par une distribution normale dans la population, le coefficient d'asymétrie mesuré sur base de l'échantillon observé, ne devrait pas être significativement différent de zéro.

i. Utilisez tout d'abord les données contenues dans le fichier 401KSUBS, en ne conservant que les données pour lesquelles *fsize* = 1. Identifiez la mesure du coefficient d'asymétrie pour *inc*. Faites de même

pour log(*inc*). Quelle est, parmi ces deux variables celle dont le coefficient d'asymétrie est le plus élevé et de ce fait, semble le moins correspondre à une variable normalement distribuée ?

ii. Dans un second temps, utilisez les données issues de la base BWGHT2. Identifiez les mesures du coefficient d'asymétrie pour *bwght* et log(*bwght*). Que pouvez-vous en conclure ?

iii. Évaluez l'assertion suivante : « La transformation logarithmique permet à des variables positives de se rapprocher d'une distribution normale. »

iv. Dans la mesure où nous nous intéressons à l'hypothèse de normalité dans le contexte de la régression linéaire, devrions-nous évaluer les distributions non conditionnelles de y et log(y) ? Justifiez.

C5. Reprenez l'analyse présentée dans le cadre de l'exercice sur ordinateur C11 du chapitre 4 sur les données contenues dans le fichier HTV, avec *educ* la variable dépendante de la régression.

i. Combien de valeurs différentes recensez-vous pour la variable *educ* dans l'échantillon ? La variable *educ* est-elle caractérisée par une distribution continue ?

ii. Représentez graphiquement l'histogramme des fréquences empiriques associées à la variable *educ* parallèlement à la représentation d'une distribution normale. La distribution empirique de la variable *educ* vous apparaît-elle proche d'une distribution normale ?

iii. Laquelle des hypothèses de base du modèle de régression linéaire vous semble clairement violée dans le modèle suivant :

$$educ = \beta_0 + \beta_1 motherduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u \ ?$$

En quoi cette violation change-t-elle les procédures d'inférence statistique menées à bien dans l'exercice sur ordinateur C11 du chapitre 4 ?

C6. À partir des données issues de la base ECONMAT, répondez aux questions suivantes.

i. En toute logique, quelles sont les valeurs minimale et maximale que peut prendre la variable *score* ? Quelles sont les valeurs minimale et maximale observées dans l'échantillon ?

ii. Soit le modèle linéaire suivant :

$$score = \beta_0 + \beta_1 colgpa + \beta_2 actmth + \beta_3 acteng + u$$

Pourquoi l'Hypothèse RLM.6 ne peut-elle pas tenir pour le terme d'erreur u ? Quelles conséquences cela entraine-t-il quant à l'utilisation de la statistique de Student standard pour tester $H_0 : \beta_0 = 0$?

iii. Estimez le modèle décrit à la question (ii) et récupérez les statistiques de Student et p-valeur associés au test $H_0 : \beta_0 = 0$. Comment défendriez-vous vos résultats face à quelqu'un vous opposant la remarque suivante : « Vous ne pouvez pas croire dans la p-valeur, car le terme d'erreur de votre modèle ne suit de toute évidence par une distribution Normale ».

ANNEXE 5A
Normalité asymptotique [de l'estimateur] des MCO

Nous donnons ici l'intuition de la preuve de la normalité asymptotique de l'estimateur des MCO [théorème 5.2(i)] dans le cas de la régression simple. Écrivons d'abord le modèle de régression simple à l'instar de celui décrit à l'équation (5.16). *Via* les outils standard d'algèbre linéaire propres à l'analyse des régressions linéaires, il est possible de dériver l'expression suivante :

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = (1 / s_x^2)\left[n^{-1/2} \sum_{i=1}^{n} (x_i - \bar{x}) u_i \right],$$

où s_x^2 désigne la variance empirique des $\{x_i : i = 1, 2, ..., n\}$. Par application de la loi des grands nombres (voir annexe C), $s_x^2 \xrightarrow{p} \sigma_x^2 = \text{Var}(x)$. L'hypothèse RLM.3 élimine la possibilité de colinéarité parfaite entre les variables explicatives, ce qui implique que $\text{Var}(x) > 0$ (x_i varie dans l'échantillon, et de ce fait, x n'est pas constante pour la population). Puis, $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i = n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i + (\mu - \bar{x})\left[n^{-1/2}\sum_{i=1}^{n}u_i\right]$, avec $\mu = \text{E}(x)$, la moyenne de x pour la population. Ensuite, $\{u_i\}$ étant une séquence de variables aléatoires i.i.d. de moyennes nulles et de variances σ^2, $n^{-1/2}\sum_{i=1}^{n}u_i$ converge vers une distribution Normale$(0,\sigma^2)$ à mesure que $n \to \infty$; cette assertion est une simple conséquence du théorème central limite rappelé en annexe C. Par application de la loi des grands nombres, plim $(\mu - \bar{x}) = 0$. Un résultat standard de la théorie asymptotique stipule que si plim$(w_n) = 0$ et z_n suit une distribution asymptotiquement normale, alors plim$(w_n z_n) = 0$. [Voir Wooldridge (2010, chapitre 3) pour plus de détails.] Cela implique que $(\mu - \bar{x})\left[n^{1/2}\sum_{i=1}^{n}u_i\right]$ a une limite en probabilité nulle. Ensuite, $\{(x_i - \mu)u_i : i = 1, 2, ...\}$ est une séquence infinie de variables aléatoires i.i.d. de moyennes nulles – puisque u et x sont décorrélées sous l'hypothèse RLM.4 – et de variances $\sigma^2\sigma_x^2$ du fait de l'hypothèse d'homoscédasticité RLM.5. Il s'ensuit que $n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i$ suit asymptotiquement une distribution Normale$(0, \sigma^2\sigma_x^2)$. Nous venons ainsi de montrer que la différence entre $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$ et $n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i$ a une limite en probabilité nulle. Un résultat de la théorie asymptotique affirme que si z_n est asymptotiquement normalement distribué et que plim$(v_n - z_n) = 0$, alors v_n est caractérisée par la même distribution asymptotique que z_n. Ainsi, $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$ suit également une distribution asymptotiquement Normale$(0, \sigma^2\sigma_x^2)$. En combinant tous ces éléments bout à bout, on obtient :

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = (1/\sigma_x^2)\left[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i\right] + [(1/s_x^2) - (1/\sigma_x^2)]\left[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i\right],$$

et puisque plim$(1/s_x^2) = 1/\sigma_x^2$, le second terme est caractérisé par une plim nulle. Dès lors, la distribution asymptotique de $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ est Normale $(0, \{\sigma^2\sigma_x^2\}/\{\sigma_x^2\}^2) =$ Normale $(0, \sigma^2 / \sigma_x^2)$. Ceci complète la preuve dans le cas de la régression simple puisque $a_1^2 = \sigma_x^2$ dans ce cas. Voir Wooldridge (2010, chapitre 4) pour le cas général.

CHAPITRE 6

QUESTIONS ADDITIONNELLES SUR LE MODÈLE DE RÉGRESSION

Traduction de Cédric Heuchenne

Ce chapitre introduit plusieurs autres questions en régression multiple qui n'ont pas été traitées dans les chapitres précédents. Celles-ci sont moins fondamentales que les sujets développés aux chapitres 3 et 4, mais elles sont importantes en vue d'appliquer la régression multiple à une large gamme de problèmes empiriques.

6.1 EFFETS DES CHANGEMENTS DES ÉCHELLES DES DONNÉES SUR LES STATISTIQUES DES MCO

Dans le chapitre 2 sur la régression simple, nous avons brièvement discuté l'effet d'un changement d'unités de mesure sur les estimateurs de pente et d'ordonnée à l'origine. Nous avons également montré que changer les unités de mesure n'affectait pas le R-carré. Nous revenons à la question des changements d'échelles des données (pour la variable dépendante ou les variables indépendantes) et examinons leurs effets sur les écarts-types, les statistiques t, les statistiques F et les intervalles de confiance.

Nous allons découvrir que tout ce qui est supposé se produire se produit bien : un changement d'échelle des variables entraine un changement des coefficients, des écarts-types estimés, des statistiques t, des statistiques F et des intervalles de confiance de manière à préserver les effets mesurés et les sorties des tests. Bien qu'il ne s'agisse pas d'une grande surprise – ce serait en fait inquiétant si c'était le cas –, il est utile de voir ce qui se produit de manière explicite. Souvent, un changement d'échelle des données est utilisé pour des raisons cosmétiques comme par exemple lorsqu'on réduit le nombre de zéros après la virgule dans un coefficient. En choisissant de manière judicieuse les unités de mesure, on peut améliorer l'apparence d'une équation estimée sans changer quoi que ce soit d'essentiel.

Nous pourrions traiter ce problème de manière générale mais il est beaucoup mieux illustré par l'intermédiaire d'exemples. Aussi, il y a peu d'intérêt à introduire ici des notations abstraites.

Nous commençons avec une équation reliant le poids d'un enfant à la naissance avec le nombre de cigarettes fumées par la mère de l'enfant et le revenu familial

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc, \qquad \textbf{[6.1]}$$

où
bwght = poids de l'enfant à la naissance en onces (28,35 grammes).
cigs = nombre de cigarettes fumées par jour, par la mère pendant la grossesse.
faminc = revenu annuel de la famille, en milliers de dollars.

Tableau 6.1 *Effets de changements d'échelles des données*

Variable dépendante	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Variables indépendantes			
cigs	– 0,4634 (0,0916)	– 0,0289 (0,0057)	–
packs	–	–	– 9,268 (1,832)
faminc	0,0927 (0,0292)	0,0058 (0,0018)	0,0927 (0,0292)
constante	116,974 (1,049)	7,3109 (0,0656)	116,974 (1,049)

Variable dépendante	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Observations	1 388	1 388	1 388
R-carré	0,0298	0,0298	0,0298
SCR	557 485,51	2 177,6778	557 485,51
ETR	20,063	1,2539	20,063

Les estimations de cette équation, obtenues en utilisant les données dans BWGHT, sont données dans la première colonne du tableau 6.1. Les écarts-types des estimateurs (erreurs types) sont indiqués entre parenthèses. L'estimation du coefficient de *cigs* dit que si une femme a fumé 5 cigarettes supplémentaires par jour, le poids prédit à la naissance diminue de 0,4634(5) = 2,317 onces. La statistique *t* pour *cigs* vaut −5,06, donc cette variable est statistiquement très significative.

Supposons maintenant qu'on veuille mesurer le poids à la naissance en livres plutôt qu'en onces. Soit *bwghtlbs* = *bwght*/16, le poids en livres à la naissance. Que deviennent nos statistiques des MCO si nous utilisons cette variable comme variable dépendante dans notre équation ? Il est facile de dégager l'effet sur les estimations des coefficients par simple manipulation de l'équation (6.1). En divisant toute cette équation par 16, on obtient

$$\widehat{bwght} / 16 = \hat{\beta}_0 / 16 + (\hat{\beta}_1 / 16)\, cigs + (\hat{\beta}_2 / 16)\, faminc.$$

Puisque le côté gauche de cette équation correspond au poids à la naissance en livres, il s'ensuit que chaque nouveau coefficient sera l'ancien coefficient divisé par 16. Pour vérifier ceci, la régression de *bwghtlbs* sur *cigs* et *faminc* est rapportée en colonne (2) du tableau 6.1. Jusqu'au quatrième chiffre, l'ordonnée à l'origine et les pentes sont juste celles de la colonne (1) divisées par 16. Par exemple, le coefficient de *cigs* est maintenant −0,0289 ; cela signifie que si *cigs* augmentait de 5, le poids à la naissance diminuerait de 0,0289(5) = 0,1445 livres. En termes d'onces, nous obtenons 0,1445(16) = 2,312, ce qui est légèrement différent de 2,317 (obtenu ci-avant) et dû à l'erreur d'arrondi. En fait, une fois les coefficients transformés dans les mêmes unités, on obtient évidemment la même réponse, peu importe la manière avec laquelle la variable dépendante est mesurée.

Et qu'en est-il de la significativité statistique des variables ? Comme on peut s'y attendre, transformer la variable dépendante d'onces en livres n'a aucun effet sur l'importance statistique des variables indépendantes. Les erreurs types dans la colonne (2) sont 16 fois plus petites que dans la colonne (1). Quelques calculs rapides montrent que les statistiques *t* dans la colonne (2) sont identiques à celles de la colonne (1). Les bornes des intervalles de confiance dans la colonne (2) sont les bornes de la colonne (1) divisées par 16. Les intervalles de confiance changent par le même facteur que les erreurs types. [Rappel : l'intervalle de confiance de niveau 95 % est ici $\hat{\beta}_j \pm 1{,}96\sigma(\hat{\beta}_j)$.

En matière d'ajustement, les *R*-carrés des deux régressions sont identiques, comme attendu. Remarquons que la somme des carrés des résidus, SCR, et l'estimation de l'écart-type des résidus de la régression, ETR, diffèrent selon la régression. Ces différences peuvent aisément s'expliquer. Prenons le résidu de l'observation *i* dans l'équation (6.1). Le résidu quand *bwghtlbs* est la variable dépendante est simplement $\hat{u}_i/16$. Donc, le résidu au carré dans la seconde équation est $(\hat{u}_i /16)^2 = \hat{u}_i^2/256$. C'est pourquoi la somme des carrés des résidus dans la colonne (2) est égale à SCR dans la colonne (1) divisée par 256.

Puisque ETR = $\hat{\sigma} = \sqrt{\text{SCR} / (n - k - 1)} = \sqrt{\text{SCR} / 1\,385}$, ETR dans la colonne (2) est 16 fois plus petit que dans la colonne (1). Une autre manière de comprendre ce résultat est de voir que l'erreur dans l'équation avec *bwghtlbs* comme variable dépendante a un écart-type 16 fois plus petit que celui de l'erreur

initiale. Cela ne signifie pas que nous ayons réduit l'erreur en changeant la manière de mesurer le poids à la naissance ; un ETR plus petit traduit simplement une différence d'unités de mesure.

Ensuite, reprenons la variable dépendante avec ses unités initiales : *bwght* est mesuré en onces. Changeons plutôt l'unité de mesure d'une des variables indépendantes, *cigs*. Définissons *packs*, le nombre de paquets de cigarettes fumés par jour. Donc, *packs* = *cigs*/20. Que deviennent alors les coefficients et les autres statistiques des MCO ? Nous pouvons écrire

$$\widehat{bwght} = \hat{\beta}_0 + (20\hat{\beta}_1)\,(cigs / 20) + \hat{\beta}_2\ faminc = \hat{\beta}_0 + (20\hat{\beta}_1)\ packs + \hat{\beta}_2\ faminc.$$

Donc, l'ordonnée à l'origine et le coefficient de *faminc* restent inchangés mais le coefficient de *packs* est 20 fois celui de *cigs*. Les résultats de la régression de *bwght* sur *packs* et *faminc* se trouvent en colonne (3) du tableau 6.1. Rappelons à ce propos qu'inclure les deux variables *cigs* et *packs* dans la même équation n'aurait aucun sens ; cela entraînerait une parfaite multicolinéarité et n'aurait aucune signification intéressante.

Dans la colonne (3), une autre statistique diffère également de la colonne (1) : l'écart-type estimé du coefficient de *packs* est 20 fois plus grande que celle du coefficient de *cigs* dans la colonne (1). Cela signifie que la statistique *t* pour tester l'effet de fumer des cigarettes est la même qu'on utilise comme étalon de mesure le nombre de cigarettes ou le nombre de paquets. Ceci est bien naturel.

L'exemple précédent explique clairement la plupart des possibilités qui surviennent quand les variables dépendante et indépendantes subissent des changements d'échelles. En économie, un changement d'échelle est souvent effectué avec des montants en dollars, en particulier quand ces montants sont très grands.

Dans le chapitre 2, si la variable dépendante apparaît sous forme logarithmique, nous avons montré que changer son unité de mesure n'affectait pas le coefficient de pente. Similairement ici, changer l'unité de mesure de la variable dépendante quand celle-ci apparaît sous forme logarithmique n'affecte aucune estimation de coefficient de pente. Ceci vient du fait que $\log(c_1 y_i) = \log(c_1) + \log(y_i)$ pour n'importe quelle constante $c_1 > 0$. La nouvelle ordonnée à l'origine sera $\log(c_1) + \hat{\beta}_0$. Similairement, changer l'unité de mesure de n'importe quel x_j, où $\log(x_j)$ apparaît dans la régression, affecte seulement l'ordonnée à l'origine. Ceci est cohérent avec ce que nous connaissons des changements en pourcentage et, en particulier, des élasticités : ces dernières sont invariantes aux unités de mesure que ce soit de y ou de x_j. Par exemple, si on avait spécifié la variable dépendante dans (6.1) comme étant log(*bwght*), estimé l'équation et ensuite ré-estimé celle-ci avec log(*bwghtlbs*) comme variable dépendante, les coefficients de *cigs* et de *faminc* auraient été les mêmes dans les deux régressions ; seule l'ordonnée à l'origine aurait été différente.

Pour aller plus loin 6.1

Dans l'équation du poids à la naissance (6.1), supposons que *faminc* soit mesuré en dollars plutôt qu'en milliers de dollars. Donc, définissons la variable *fincdol* = 1,000·*faminc*. Comment les statistiques des MCO vont-elles changer si *faminc* est remplacé par *fincdol* ? Afin de présenter au mieux les résultats de la régression, pensez-vous qu'il soit préférable de mesurer le revenu en dollars ou en milliers de dollars ?

Coefficients Beta

Parfois, dans les applications économétriques, une variable clé est mesurée sur une échelle difficile à interpréter. Les économistes du travail insèrent souvent des scores de tests dans les équations du travail, et l'échelle sur laquelle ces tests ont été quantifiés est souvent arbitraire et difficile à interpréter (au moins pour les économistes !). Dans presque tous les cas, on s'intéresse à la différence entre le score d'un individu en particulier et celui de la population. Donc au lieu de se poser la question de l'effet sur le salaire par heure si, disons, le score d'un test a augmenté de 10 points, on se demande plutôt ce qui se passe quand le score du test a augmenté d'un *écart-type*.

Rien ne nous interdit d'observer la variation de la variable dépendante quand une variable indépendante dans un modèle estimé augmente d'un certain nombre d'écarts-types de cette variable dans l'échantillon ; on suppose qu'on peut bien calculer cet écart-type (ce qui est aisé avec la plupart des packages sur la régression). C'est souvent une bonne idée. Ainsi, quand on regarde par exemple l'effet du score d'un test comme le score SAT sur la GPA obtenue à l'université (moyenne globale des résultats obtenus dans le système scolaire américain), on peut calculer l'écart-type de ce score et observer les variations consécutives à une augmentation du score SAT d'une ou deux fois cet écart-type.

Parfois, il est utile d'obtenir les résultats de la régression quand toutes les variables considérées, dépendantes ou indépendantes, ont été standardisées. Une variable est standardisée dans l'échantillon quand on lui soustrait sa moyenne dans l'échantillon et qu'on la divise par son écart-type dans l'échantillon (voir l'annexe C). Cela signifie qu'on calcule le z-score pour chaque variable dans l'échantillon. Ensuite, on effectue la régression en utilisant ces z-scores.

Pourquoi la standardisation est-elle utile ? Le plus simple est de commencer par l'équation des MCO initiale avec les variables sous leur forme originale :

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i. \qquad \textbf{[6.2]}$$

Nous avons inclus l'indice i pour insister sur le fait que notre standardisation est appliquée à toutes les valeurs de l'échantillon. Si on calcule la moyenne (de l'échantillon) de (6.2), utilise le fait que la moyenne (de l'échantillon) des $\hat{u}_i$ est nulle et soustrayons le résultat de (6.2), nous obtenons

$$y_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_k(x_{ik} - \bar{x}_k) + \hat{u}_i.$$

Soit $\hat{\sigma}_y$, l'écart-type de l'échantillon pour la variable dépendante, $\hat{\sigma}_1$, l'écart-type de l'échantillon des x_{i1}, $\hat{\sigma}_2$, celui des x_{i2} et ainsi de suite. Alors, de simples manipulations algébriques mènent à

$$(y_i - \bar{y}) / \hat{\sigma}_y = (\hat{\sigma}_1 / \hat{\sigma}_y)\hat{\beta}_1[(x_{i1} - \bar{x}_1) / \hat{\sigma}_1] + \ldots + (\hat{\sigma}_k / \hat{\sigma}_y)\hat{\beta}_k[(x_{ik} - \bar{x}_k) / \hat{\sigma}_k] + (\hat{u}_i / \hat{\sigma}_y). \qquad \textbf{[6.3]}$$

Chaque variable dans (6.3) a été standardisée en la remplaçant par son z-score, faisant ainsi apparaître de nouveaux coefficients de pente. Par exemple, le coefficient de pente de $(x_{i1} - \bar{x}_1) / \hat{\sigma}_1$ est $(\hat{\sigma}_1 / \hat{\sigma}_y)\hat{\beta}_1$. Il s'agit simplement du coefficient initial, $\hat{\beta}_1$, multiplié par le rapport des écarts-types estimés de x_1 et de y. L'ordonnée à l'origine a quant à elle disparu.

Il est utile de réécrire (6.3) en omettant l'indice i

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_k z_k + erreur, \qquad \textbf{[6.4]}$$

où z_y correspond au z-score de y, z_1 au z-score de x_1, et ainsi de suite. Les nouveaux coefficients sont

$$\hat{b}_j = (\hat{\sigma}_j / \hat{\sigma}_y)\hat{\beta}_j, \text{ for } j = 1, \ldots, k. \qquad \textbf{[6.5]}$$

Ces $\hat{b}_j$ sont habituellement appelés **coefficients de la régresssion réduits** ou **coefficients beta**. (Ce dernier nom est plus utilisé, ce qui peut porter à confusion parce que nous avons utilisé beta surmonté d'un accent circonflexe pour noter les estimations par MCO habituelles.)

Une signification intéressante pour les coefficients beta peut être tirée de l'équation (6.4) : si x_1 augmente d'un écart-type, alors $\hat{y}$ change de $\hat{b}_1$ écarts-types. Donc, nous mesurons l'effet non pas en termes des unités initiales de y ou de x_j, mais en unités d'écart-type. Puisque la standardisation rend l'échelle des régresseurs non pertinente, cette équation met tous les régresseurs sur un même pied. Dans une équation des MCO standard, il n'est pas possible de simplement regarder la taille des différents coefficients et de conclure que la variable explicative avec le plus grand coefficient est « la plus importante ». On vient juste de voir que les

valeurs des coefficients peuvent changer à volonté en changeant les unités de mesure des x_j. Mais quand chaque x_j a été standardisé, comparer les valeurs des coefficients beta résultants devient pertinent. Quand l'équation de la régression a une seule variable explicative, x_1, son coefficient de la régression réduit est simplement le coefficient de corrélation (de l'échantillon) entre y et x_1, ce qui signifie qu'il doit se trouver entre -1 et 1.

Même dans les situations où les coefficients sont aisément interprétables – disons que les variables dépendantes et indépendantes sont sous forme logarithmique, les coefficients d'intérêt issus des MCO étant ainsi des estimations d'élasticités – les coefficients beta peuvent aussi avoir leur intérêt. Bien que les élasticités soient indépendantes des unités de mesure, changer une variable explicative de, disons, 10 % peut représenter un plus grand ou un plus petit changement sur la gamme de valeurs de cette variable que changer des mêmes 10 % une autre variable explicative. Par exemple, dans un état avec une large variation des revenus mais une variation relativement faible des dépenses par étudiant, comparer des élasticités de performance par rapport aux revenus et aux dépenses pourrait ne pas être pertinent. Comparer des valeurs de coefficients beta peut alors s'avérer utile.

Afin d'obtenir les coefficients beta, on peut toujours standardiser y, x_1, ..., x_k et ensuite effectuer la régression des z-scores de y contre les z-scores des x_1, ..., x_k – où il n'est pas nécessaire d'inclure une ordonnée à l'origine, puisque celle-ci vaudra zéro. En pratique, cette procédure peut s'avérer fastidieuse, notamment si l'on considère beaucoup de variables indépendantes. Heureusement, certains packages de régression fournissent les coefficients beta via une simple commande. L'exemple suivant illustre l'utilisation des coefficients beta.

EXEMPLE 6.1
Effets de la pollution sur le prix du logement

Nous utilisons les données de l'exemple 4.5 (dans le fichier HPRICE2) pour illustrer l'utilisation des coefficients beta. Rappelons que la variable indépendante clé est *nox*, une mesure d'oxyde d'azote dans l'air sur chaque communauté. Une manière de comprendre l'importance de l'effet de la pollution – sans entrer dans les concepts scientifiques sous-tendant l'effet de l'oxyde d'azote sur la qualité de l'air – est de calculer les coefficients beta. (Une approche alternative est utilisée dans l'exemple 4.5 : nous avons obtenu une élasticité du prix par rapport à *nox* en utilisant *price* et *nox* sous leurs formes logarithmiques.)

L'équation de la population est le modèle suivant

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + \beta_4 dist + \beta_5 stratio + u,$$

où toutes les variables excepté *crime* ont été définies dans l'exemple 4.5 ; *crime* est le nombre de crimes répertoriés par personne. Les coefficients beta sont rapportés dans l'équation suivante (chaque variable a été convertie en z-score) :

$$zprice = 0{,}340\ znox - 0{,}143\ zcrime + 0{,}514\ zrooms - 0{,}235\ zdist - 0{,}270\ zstratio$$

Cette équation montre qu'une augmentation d'un écart-type de *nox* diminue *price* de 0,34 écart-type ; une augmentation d'un écart-type de *crime* diminue *price* de 0,14 écart-type. Donc, une même variation relative de pollution dans la population a un plus grand effet sur le prix du logement que le crime. La taille de la maison, mesurée par le nombre de pièces (*rooms*), a le plus grand effet standardisé. Si nous voulons connaître l'effet de chaque variable indépendante sur la valeur en dollars du prix médian d'une maison, nous devons utiliser les variables non standardisées.

Que nous utilisions des variables standardisées ou non n'affecte pas le caractère statistiquement significatif : les statistiques t sont les mêmes dans les deux cas.

6.2 COMPLÉMENTS SUR LA FORME FONCTIONNELLE

Dans plusieurs des exemples précédents, nous avons rencontré l'outil le plus populaire en économétrie pour permettre des relations non linéaires entre les variables expliquées et explicatives : utiliser des logarithmes pour les variables dépendantes et indépendantes. Nous avons également observé des modèles contenant des formes quadratiques pour certaines variables explicatives, mais nous avons encore à fournir une manière systématique de les traiter. Dans cette section, nous couvrons certaines variations et extensions des formes fonctionnelles souvent rencontrées en pratique.

Compléments concernant l'utilisation de formes fonctionnelles logarithmiques

Commençons par revoir comment interpréter les paramètres dans le modèle

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u, \qquad \textbf{[6.6]}$$

où ces variables sont tirées de l'exemple 4.5. Rappelons que tout au long du texte, $\log(x)$ est le logarithme *naturel* x. Le coefficient β_1 est l'élasticité de *price* par rapport à *nox* (pollution). Le coefficient β_2 est le changement de $\log(price)$ quand $\Delta rooms = 1$; comme nous avons vu à de nombreuses reprises, multiplié par 100, il s'agit du changement approximatif de *price* en pourcentage. Rappelons que $100 \cdot \beta_2$ est parfois appelé la semi-élasticité de *price* par rapport à *rooms*.

Après estimation en utilisant les données dans HPRICE2, nous obtenons

$$\begin{gathered} \log(price) = 9{,}23 - 0{,}718 \log(nox) + 0{,}306\ rooms \\ \qquad\quad (0{,}19)\ \ (0{,}066)\qquad\quad (0{,}019) \\ n = 506,\ R^2 = 0{,}514 \end{gathered} \qquad \textbf{[6.7]}$$

Donc, quand *nox* augmente de 1 %, *price* diminue de 0,718 %, en gardant *rooms* fixé. Quand *rooms* augmente de un, *price* augmente d'approximativement 100(0,306) = 30,6 %.

L'estimation indiquant qu'une pièce supplémentaire augmente le prix d'environ 30,6 % est assez imprécise pour cette application. L'erreur d'approximation se produit car, à mesure que le changement de $\log(y)$ augmente, l'approximation $\%\Delta y \approx 100 \cdot \Delta\log(y)$ devient de plus en plus imprécise. Heureusement, un calcul simple est disponible pour calculer le changement en pourcentage exact.

Pour décrire la procédure, considérons le modèle estimé général

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 x_2.$$

(Ajouter des variables indépendantes supplémentaires ne change pas la procédure.) Pour x_1 fixé, nous avons $\Delta\widehat{\log(y)} = \hat{\beta}_2 \Delta x_2$. En utilisant les propriétés algébriques des fonctions logarithme et exponentielle, on obtient le changement en pourcentage *exact* de l'y prédit :

$$\%\Delta\hat{y} = 100\,[\exp(\hat{\beta}_2 \Delta x_2) - 1], \qquad \textbf{[6.8]}$$

où la multiplication par 100 transforme le changement en pourcentage. Quand $\Delta x_2 = 1$,

$$\%\Delta\hat{y} = 100 \cdot [\exp(\hat{\beta}_2) - 1]. \qquad \textbf{[6.9]}$$

Appliqué à l'exemple du prix du logement avec x_2 = rooms et $\hat{\beta}_2 = 0{,}306$, $\%\Delta\widehat{price} = 100[\exp(0{,}306) - 1] = 35{,}8$ %, ce qui est clairement plus grand que le pourcentage approximé, 30,6 %, obtenu directement par (6.7). {À ce propos, il ne s'agit pas d'un estimateur non biaisé puisque $\exp(\cdot)$ est une fonction non linéaire ;

il s'agit cependant d'un estimateur convergent de $100[\exp(\beta_2) - 1]$. La raison en est que la limite en probabilité passe dans les fonctions continues alors que l'opérateur espérance ne le fait pas. Voir l'annexe C.}

Le calcul (6.8) n'est pas crucial pour des petits changements. Par exemple, quand on introduit le rapport étudiant-enseignant dans l'équation (6.7), son coefficient estimé est –0,052, ce qui signifie que si *stratio* augmente de un, *price* diminue d'environ 5,2 %. Le changement en pourcentage exact est $100[\exp(-0,052) - 1] \approx 100(-0,051)$, soit –5,1 %. D'un autre côté, si on augmente *stratio* de cinq, alors le changement de prix en pourcentage approximé est –26 %, tandis que le changement exact obtenu par l'équation (6.8) est $100\ [\exp(-0,26) - 1] \approx -22,9$ %.

L'approximation logarithmique du changement en pourcentage possède un avantage qui justifie son utilisation même quand le changement en pourcentage est grand. Pour décrire cet avantage, considérons encore l'effet sur le prix d'un changement du nombre de pièces d'une unité. L'approximation logarithmique est simplement le coefficient de *rooms* dans l'équation (6.7) multiplié par 100, à savoir, 30,6 %. Nous avons également calculé une estimation du changement en pourcentage exact pour un *accroissement* de un du nombre de pièces : 35,8 %. Mais qu'en est-il si nous voulons estimer le changement en pourcentage pour une diminution de un du nombre de pièces ? Dans l'équation (6.8), nous avons $\Delta x_2 = -1$, $\hat{\beta}_2 = 0,306$ et ainsi $\%\widehat{\Delta price} = 100[\exp(-0,306) - 1] = -26,4$, soit une chute de 26,4 %. Remarquons que l'approximation basée sur l'utilisation du coefficient de *rooms* se trouve entre 26,4 et 35,8. En d'autres mots, utiliser simplement le coefficient (multiplié par 100) nous donne une estimation qui se trouve toujours entre les valeurs absolues des estimations pour une augmentation et une diminution. Si nous sommes spécifiquement intéressés par une augmentation ou une diminution, nous pouvons utiliser le calcul basé sur l'équation (6.8).

Ce dernier point à propos du calcul des changements en pourcentage est essentiellement le même que celui traité en introduction à l'économie lorsqu'il s'agit de calculer, disons, l'élasticité-prix de la demande basée sur de grands changements de prix : le résultat dépend du fait qu'on utilise le prix et la demande au début ou à la fin quand on calcule les changements en pourcentage. Utiliser l'approximation logarithmique se fait dans le même esprit que calculer une élasticité d'arc de la demande, où les moyennes des prix et des quantités sont utilisées aux dénominateurs quand on calcule les changements en pourcentage.

Nous avons vu que l'utilisation des logarithmes naturels mène à des interprétations intéressantes des coefficients. De plus, nous pouvons ignorer les unités de mesure des variables apparaissant sous forme logarithmique parce que les coefficients de pente sont invariants aux changements d'échelle. Il y a plusieurs autres raisons qui justifient une telle utilisation des logarithmes naturels dans les travaux appliqués. D'abord, quand $y > 0$, les modèles utilisant $\log(y)$ comme variable dépendante satisfont les hypothèses des modèles linéaires classiques de manière souvent plus proche que les modèles utilisant le niveau de y (forme linéaire). Les variables strictement positives ont souvent des distributions conditionnelles asymétriques ou hétéroscédastiques ; prendre le log peut atténuer, sinon éliminer les deux problèmes.

Un autre intérêt potentiel quand on utilise le log d'une variable est souvent que la gamme de valeurs de celle-ci rétrécit. C'est en particulier vrai pour les variables qui peuvent avoir de grandes valeurs monétaires, comme les ventes annuelles de firmes ou les salaires des joueurs de baseball. Les variables de la population tendent aussi à présenter de grandes variations. Rétrécir la gamme de valeurs des variables dépendantes ou indépendantes peut rendre les estimations par MCO moins sensibles aux valeurs extrêmes ; on peut reprendre à ce propos la question des observations aberrantes au Chapitre 9.

Cependant, on ne doit pas utiliser la fonction logarithmique aveuglément car dans certains cas, elle peut vraiment créer des valeurs extrêmes. C'est la cas, par exemple, quand une variable y se trouve entre zéro et un (comme une proportion) et prend des valeurs proches de zéro ; $\log(y)$ (qui est nécessairement négatif) peut alors avoir des valeurs (absolues) très grandes alors que la variable initiale, y, est bornée entre zéro et un.

Il y a certaines règles automatiques pour choisir de prendre les logs bien qu'aucune ne soit gravée dans la pierre. Quand la variable est un montant en dollars positif, le log est souvent utilisé. Nous l'avons entre

autres vu pour des variables comme le salaire, les ventes et la valeur marchande d'une firme. Des variables relatives à une population comme le nombre total d'employés ou d'inscriptions à l'école apparaissent souvent sous forme logarithmique ; elles ont cette caractéristique commune d'être des grandes valeurs entières.

Des variables qui sont mesurées en années – comme le niveau d'instruction, l'expérience, la durée de la période probatoire au travail, l'âge… – apparaissent habituellement sous leur forme originale. Une variable de proportion ou de pourcentage – comme le taux de chômage, le taux de participation à un plan de pension, le pourcentage d'étudiants passant un examen standardisé ou le taux d'arrestations par rapport au nombre de crimes recensés – peut apparaître soit sous sa forme originale, soit sous la forme logarithmique, bien qu'il y ait une tendance à l'utiliser sous la première forme. Les coefficients de la régression avec les variables originales – qu'elles soient dépendantes ou indépendantes – ont en effet une interprétation en termes de changement en points de pourcentage. (Voir l'annexe A pour une revue de la distinction entre changement en pourcentage et changement en points de pourcentage.) Si on utilise, disons, log(*unem*) dans une régression, où *unem* est le pourcentage d'individus sans emploi, on doit être attentif à distinguer le changement en points de pourcentage du changement en pourcentage. Pour rappel, si *unem* augmente de 8 à 9, il s'agit d'une augmentation d'un point de pourcentage mais de 12,5 % par rapport au niveau de pourcentage initial d'individus sans emploi. Utiliser le log signifie que nous regardons le changement en pourcentage du taux de chômage : $\log(9) - \log(8) \approx 0{,}118$ ou 11,8 % qui est l'approximation logarithmique de la réelle augmentation de 12,5 %.

Une limitation du log est qu'il ne peut être utilisé si une variable prend la valeur zéro ou des valeurs négatives. Dans les cas où la variable y est positive mais peut prendre la valeur zéro, $\log(1+y)$ est parfois utilisé. Les interprétations de changement en pourcentage sont souvent conservées, excepté pour les changements commençant à $y=0$ (où le changement en pourcentage n'est pas défini). Généralement, utiliser $\log(1 + y)$ et ensuite interpréter les estimations comme si on avait utilisé $\log(y)$ est acceptable quand les données correspondant à y contiennent relativement peu de zéros. À titre d'exemple, y pourrait être le nombre d'heures de formation par employé fournies par les entreprises industrielles si une large fraction de ces firmes fournit de la formation à au moins un employé. Cependant, techniquement, $\log(1 + y)$ ne peut être distribué normalement (bien qu'il puisse être moins hétéroscédastique que y). Des alternatives utiles, bien que plus avancées, comme les modèles Tobit et de Poisson sont développées au Chapitre 17.

Pour aller plus loin 6.2

Supposons que le nombre annuel d'arrestations pour conduite en état d'ivresse soit déterminée par

$$\text{Log}(arrests) = \beta_0 + \beta_1 \log(pop) + \beta_2\, age\ 16\_25 + autres\ facteurs$$

où *age 16_25* est la proportion de la population entre 16 et 25 ans. Montrez que β_2 a l'interprétation suivante (ceteris paribus) : il s'agit du changement en pourcentage de *arrests* quand le pourcentage de la population âgée de 16 à 25 ans augmente d'un point de pourcentage.

Un défaut de l'utilisation d'une variable dépendante sous forme logarithmique est la difficulté à prédire la variable originale. Le modèle original nous permet de prédire $\log(y)$, pas y. Néanmoins, il est assez aisé de transformer une prévision pour $\log(y)$ en une prévision pour y (voir section 6.4). Une remarque liée à ce point réside dans le fait qu'il *n'est pas* cohérent de comparer les R-carrés de modèles où la variable dépendante est y dans un cas et $\log(y)$ dans l'autre cas. Ces mesures expliquent les variations de variables différentes. On explique comment calculer des mesures d'ajustement comparables en section 6.4.

Modèles quadratiques

Les fonctions quadratiques sont également souvent utilisées en économie appliquée pour capturer des effets marginaux croissants ou décroissants. Les propriétés des fonctions quadratiques sont revues dans l'annexe A.

Dans le cas le plus simple, y dépend d'un seul facteur observé, mais d'une manière quadratique :

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

Par exemple, prenons $y = wage$ et $x = exper$. Comme discuté dans le chapitre 3, ce problème sort d'une analyse de régression simple mais peut être facilement traité dans le cadre de la régression multiple.

Il est important de constater que β_1 ne mesure pas le changement de y par rapport à x ; il est insensé de changer x *en gardant* x^2 fixé. Si nous écrivons l'équation estimée comme

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 , \qquad \textbf{[6.10]}$$

alors on obtient l'approximation

$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x, \text{ donc } \Delta\hat{y} / \Delta x \approx \hat{\beta}_1 + 2\hat{\beta}_2 x. \qquad \textbf{[6.11]}$$

Ceci indique que la pente de la relation entre x et y dépend de la valeur de x, la pente estimée étant $\hat{\beta}_1 + 2\hat{\beta}_2 x$. Si nous fixons $x = 0$, on observe que $\hat{\beta}_1$ peut être interprété comme une approximation de la pente quand on va de $x = 0$ à $x = 1$. Après, le second terme, $2\hat{\beta}_2 x$, doit être pris en compte.

Si on veut seulement calculer le changement prédit en y pour une valeur initiale donnée de x et un changement en x, on peut utiliser (6.10) directement : il n'y a pas de raison d'utiliser l'approximation. Cependant, on veut souvent résumer rapidement l'effet de x sur y, et l'interprétation de $\hat{\beta}_1$ et de $\hat{\beta}_2$ dans l'équation (6.11) fournit ce résumé. On pourrait y insérer la valeur moyenne de x de l'échantillon ou d'autres valeurs d'intérêt comme les valeurs de la médiane ou des quartiles inférieur ou supérieur.

Dans de nombreuses applications, $\hat{\beta}_1$ est positif et $\hat{\beta}_2$ est négatif. Par exemple, quand on utilise les données de salaire dans WAGE1, on obtient

$$\widehat{wage} = 3{,}73 + 0{,}298\ exper - 0{,}0061\ exper^2$$
$$(0{,}35)\ (0{,}041)\qquad (0{,}0009)$$
$$n = 526,\ R^2 = 0{,}093 \qquad \textbf{[6.12]}$$

Cette équation estimée implique que *exper* a un effet sur *wage* qui diminue. La première année, elle vaut à peu près 30¢ (cent) par heure ($0,298). La seconde année, elle vaut moins [environ 0,298 – 2(0,0061)(1) ≈ 0,286, soit 28,6¢, selon l'approximation (6.11) avec $x = 1$]. En allant de 10 à 11 ans d'expérience, la prévision de l'augmentation de *wage* est d'environ 0,298 – 2(0,0061)(10) = 0,176, soit 17,6¢.

Quand le coefficient de x est positif et celui de x^2 négatif, la courbe quadratique a une forme parabolique. Il y a toujours une valeur positive de x pour laquelle l'effet de x sur y est nul ; avant ce point, x a un effet positif sur y ; après ce point, x a un effet négatif sur y. En pratique, il peut être important de savoir où se trouve ce point de changement de signe dans l'effet.

Dans l'équation estimée (6.10) avec $\hat{\beta}_1 > 0$ et $\hat{\beta}_2 < 0$, ce point (ou le maximum de la fonction) est toujours réalisé pour une valeur de x égale à :

$$x^* = |\hat{\beta}_1 / (2\hat{\beta}_2)|. \qquad \textbf{[6.13]}$$

Dans l'exemple des salaires, $x^* = exper^*$ est 0,298/[2(0,0061)] ≈ 24,4. (On a juste abandonné le signe moins dans –0,0061.) Cette relation quadratique est illustrée à la figure 6.1.

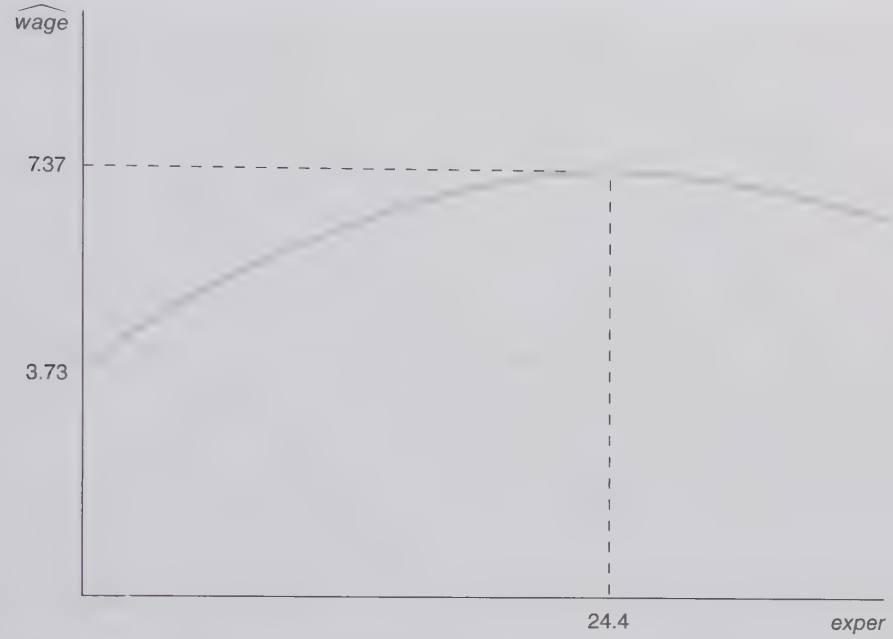


Figure 6.1 *Relation quadratique entre* $\widehat{wage}$ *et exper*

Dans l'équation des salaires (6.12), le gain dû à l'expérience devient nul après environ 24,4 ans. Que pouvons-nous faire de ce résultat ? Il existe au moins trois explications possibles. Premièrement, il se peut que peu de gens dans l'échantillon possèdent plus de 24 ans d'expérience et donc la partie à droite de 24 peut être ignorée. L'utilisation du modèle quadratique pour calculer les effets décroissants a un coût : la forme quadratique se retourne finalement en un point. Si un petit pourcentage de l'échantillon se trouve au-delà de ce point, alors cela n'a pas beaucoup d'importance. Mais dans les données WAGE1, environ 28 % des individus de l'échantillon ont plus de 24 ans d'expérience ; c'est un pourcentage trop grand pour l'ignorer.

Il est possible que l'effet de *exper* devienne réellement négatif à partir d'un certain point, mais il est difficile de croire que cela se produise après 24 ans d'expérience. Une autre possibilité plus vraisemblable est que l'estimation de l'effet de *exper* sur *wage* est biaisé car aucun autre facteur n'est considéré ou parce que la forme fonctionnelle entre *wage* et *exper* dans l'équation (6.12) n'est pas entièrement correcte. L'exercice C2 demande d'explorer cette possibilité en tenant compte de l'influence du niveau d'instruction ; il demande également de considérer log(*wage*) comme variable dépendante.

Quand un modèle a une variable dépendante sous forme logarithmique et une variable explicative sous forme quadratique, il faut être prudent dans l'interprétation des effets partiels. L'exemple suivant montre que la forme quadratique peut avoir une forme en U, plutôt qu'une forme parabolique. Une forme en U se produit dans l'équation (6.10) si $\hat{\beta}_1$ est négatif et $\hat{\beta}_2$ positif ; elle capture alors un effet croissant de x sur y.

EXEMPLE 6.2
Effets de la pollution sur le prix du logement

Nous modifions le modèle de prix de l'exemple 4.5 pour introduire un terme quadratique pour *rooms* :

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + u. \quad \textbf{[6.14]}$$

L'estimation du modèle utilisant les données HPRICE2 est

$$\begin{aligned}\log(price) &= 13{,}39 - 0{,}902\log(nox) - 0{,}087\log(dist)\\ &\quad (0{,}57)\ \ (0{,}115)\qquad\quad (0{,}043)\\ &\quad - 0{,}545\,rooms + 0{,}062\,rooms^2 - 0{,}048\,stratio\\ &\quad (0{,}165)\qquad (0{,}013)\qquad\quad (0{,}006)\\ &\quad n = 506,\ R^2 = 0{,}603\end{aligned}$$

Le terme $rooms^2$ a une statistique t d'environ 4,77 et est donc très significatif. Mais comment interpréter l'effet de *rooms* sur log(*price*) ? Au début, l'effet semble étrange. Parce que le coefficient de *rooms* est négatif et celui de $rooms^2$ positif, cette équation implique littéralement que pour des faibles valeurs de *rooms*, une pièce supplémentaire a un effet *négatif* sur log(*price*). À partir d'un certain point, l'effet devient positif et la forme quadratique signifie que la semi-élasticité de *price* par rapport à *rooms* augmente quand *rooms* augmente. Cette situation est illustrée en figure 6.2.

On obtient la valeur de retournement de *rooms* en utilisant l'équation (6.13) (même si $\hat{\beta}_1$ est négatif et $\hat{\beta}_2$ positif). La valeur absolue du coefficient de *rooms*, 0,545, divisée par deux fois le coefficient de $rooms^2$, 0,062, donne $rooms^* = 0{,}545/[2(0{,}062)] \approx 4{,}4$; il s'agit du point identifié sur la Figure 6.2.

Peut-on réellement penser que passer de trois à quatre pièces diminue la valeur attendue d'une maison ? Certainement pas. Seulement cinq des 506 communautés dans l'échantillon ont des maisons comportant en moyenne 4,4 pièces ou moins, soit environ 1 % de l'échantillon. Cette quantité est suffisamment petite pour ignorer la partie à gauche de 4,4. À droite de 4,4, on voit qu'ajouter une pièce a un effet croissant sur le changement en pourcentage du prix : $\Delta\widehat{\log(price)} \approx \{-0{,}545 + 2\,(0{,}062)rooms\}\,\Delta rooms$, et donc

$$\begin{aligned}\%\Delta price &\approx 100\,\{-0{,}545 + 2(0{,}062)rooms\}\Delta rooms\\ &= \{-54{,}5 + 12{,}4rooms)\}\Delta rooms\end{aligned}$$

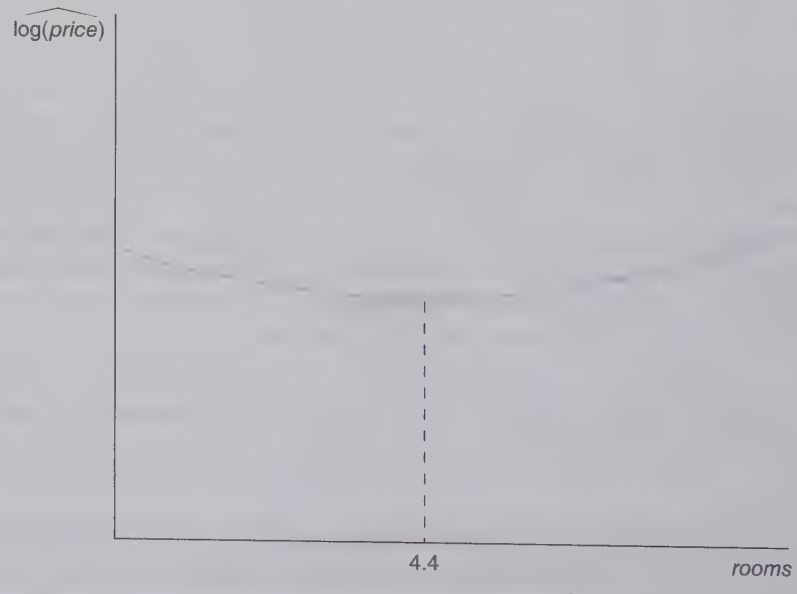


Figure 6.2 Log(*price*) comme fonction quadratique de *rooms*.

Donc, une augmentation du nombre de pièces de, disons, 5 à 6, augmente le prix d'environ –54,5 + 12,4(5) = 7,5 % ; une augmentation de 6 à 7 pièces augmente le prix d'environ –54,5 + 12,4(6) = 19,9 %. Il s'agit d'une croissance d'effet conséquente.

Cette croissance de l'effet de *rooms* sur log(*price*) dans cet exemple illustre une leçon importante : on ne peut pas simplement regarder le coefficient du terme quadratique – dans ce cas, 0,062 – et conclure uniquement sur base de sa valeur absolue que celui-ci est trop petit pour le conserver. Dans de nombreuses applications avec des termes quadratiques, le coefficient de la variable au carré contient un ou plusieurs zéros après la virgule : après tout, ce coefficient mesure comment la pente change quand x (*rooms*) change. Un changement de coefficient apparemment petit peut avoir des conséquences en pratique importantes, ainsi que nous venons de le voir. En général, on doit calculer l'effet partiel et voir comment il varie avec x pour déterminer si le terme quadratique est important en pratique. Ce faisant, il est utile de comparer la pente variable provenant du modèle quadratique avec la pente constante obtenue à partir du modèle contenant seulement un terme linéaire. Si on ne prend pas en compte $rooms^2$, le coefficient de *rooms* devient environ 0,255, ce qui implique que chaque pièce supplémentaire – ajoutée à n'importe quel nombre de pièces – augmente le prix d'environ about 25,5 %. C'est très différent du modèle quadratique pour lequel l'effet est de 25,5 % pour une valeur de *rooms* = 6,45 mais change rapidement quand *rooms* augmente ou diminue. Par exemple, pour *rooms* = 7, l'effet de la pièce additionnelle est d'environ 32,3 %.

Que se passe-t-il généralement si les coefficients des termes linéaire et quadratique ont le *même* signe (soit tous deux négatifs, soit tous deux positifs) et la variable explicative est positive ou nulle (comme dans le cas de *rooms* ou *exper*) ? Dans chaque cas, il n'y a pas de point de retournement pour des valeurs de $x > 0$. Par exemple, si β_1 et β_2 sont tous les deux positifs, la plus petite valeur attendue de y se trouve en $x = 0$, et les augmentations de x ont toujours un effet positif et croissant sur y. (C'est aussi vrai si $\beta_1 = 0$ et $\beta_2 > 0$, ce qui signifie que l'effet partiel est zéro en $x = 0$ et croissant quand x augmente.) De manière similaire, si β_1 et β_2 sont tous les deux négatifs, la plus grande valeur attendue de y se trouve en $x = 0$, et les augmentations en x ont un effet négatif sur y, la valeur absolue de cet effet augmentant avec x.

La formule générale pour le point de retournement d'une forme quadratique est $x^* = -\hat{\beta}_1/(2\hat{\beta}_2)$, ce qui mène à une valeur positive si $\hat{\beta}_1$ et $\hat{\beta}_2$ ont des signes opposés et une valeur négative si $\hat{\beta}_1$ et $\hat{\beta}_2$ ont le même signe. Connaître cette simple formule est utile dans les cas où x peut prendre des valeurs positives et négatives ; on peut calculer le point de retournement et voir si celui-ci a du sens en prenant en compte toute la gamme des valeurs possibles de x dans l'échantillon.

Il y a de nombreuses autres possibilités d'utilisation des formes quadratiques avec des logarithmes. Par exemple, une extension de (6.14) permet une élasticité non constante entre *price* et *nox* :

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 [\log(nox)]^2 + \beta_3 crime + \beta_4 rooms + \beta_5 rooms^2 + \beta_6 stratio + u \qquad \textbf{[6.15]}$$

Si $\beta_2 = 0$, alors β_1 est l'élasticité de *price* par rapport à *nox*. Sinon, cette élasticité dépend du niveau de *nox*. Pour comprendre ceci, on peut combiner les arguments pour les effets partiels dans les formes quadratiques et les modèles logarithmiques afin de montrer que

$$\%\ \Delta price \approx [\beta_1 + 2\ \beta_2 \log(nox)]\ \%\ \Delta nox\ ; \qquad \textbf{[6.16]}$$

donc, l'élasticité de *price* par rapport à *nox* est $\beta_1 + 2\ \beta_2 \log(nox)$ de telle sorte qu'elle dépende de log(*nox*).

Enfin, d'autres termes polynomiaux peuvent être insérés dans les modèles de régression. La forme quadratique est la plus souvent rencontrée mais parfois, un terme cubique voire un terme d'ordre 4 peuvent apparaître. Une forme fonctionnelle souvent raisonnable pour une fonction de coût total est

$$cost = \beta_0 + \beta_1\, quantity + \beta_2\, quantity^2 + \beta_3 quantity^3 + u.$$

Estimer un tel modèle ne pose pas de problème. Interpréter les paramètres est plus délicat (bien qu'évident en utilisant les calculs) ; nous n'étudierons pas ces modèles ici.

Modèles avec termes d'interaction

Parfois, il est naturel pour un effet partiel, l'élasticité ou la semi-élasticité de la variable dépendante par rapport à une variable explicative, de dépendre d'une autre variable explicative. Par exemple, dans le modèle

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \cdot bdrms + \beta_4 bthrms + u,$$

l'effet partiel de *bdrms* sur *price* (toutes les autres variables étant fixées) est

$$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft. \qquad \textbf{[6.17]}$$

Si $\beta_3 > 0$, alors (6.17) implique qu'une chambre supplémentaire entraîne une augmentation du prix du logement plus grande pour les plus grandes maisons. En d'autres mots, il y a un effet d'interaction entre la superficie et le nombre de chambres. Pour résumer l'effet de *bdrms* sur *price*, on doit évaluer (6.17) en des valeurs d'intérêt de *sqrft* comme la moyenne ou les quartiles inférieur ou supérieur dans l'échantillon. Que β_3 soit nul ou non, il s'agit de quelque chose de facile à tester.

Les paramètres des variables originales peuvent ne pas être aisés à interpréter quand on considère un terme d'interaction. Par exemple, dans l'équation du prix du logement précédente, l'équation (6.17) montre que β_2 est l'effet de *bdrms* sur *price* pour une maison de superficie nulle ! Cet effet n'a clairement pas beaucoup d'intérêt. Nous devons plutôt être attentifs à insérer dans la version estimée de l'équation (6.17), des valeurs d'intérêt de *sqrft* comme la moyenne ou la médiane dans l'échantillon.

Souvent, il est utile de reparamétriser un modèle de telle sorte que les coefficients des variables originales ont une signification d'intérêt. Considérons le modèle avec deux variables explicatives et un terme d'interaction :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u.$$

Comme précédemment mentionné, β_2 est l'effet partiel de x_2 sur y quand $x_1 = 0$. Souvent, ceci n'a pas d'intérêt. On peut plutôt reparamétriser le modèle de la manière suivante :

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u,$$

où μ_1 est la moyenne de x_1 dans la population et μ_2 celle de x_2. On peut voir facilement que le coefficient de x_2, δ_2, est maintenant l'effet partiel de x_2 sur y à la valeur moyenne de x_1. (En développant le terme d'interaction dans la seconde équation et en comparant les coefficients, on peut facilement montrer que $\delta_2 = \beta_2 + \beta_3 \mu_1$. Le paramètre δ_1 a une interprétation similaire.) Donc, si nous soustrayons les moyennes des variables – en pratique, celles-ci seraient les moyennes dans l'échantillon – avant de créer le terme d'interaction, les coefficients des variables originales ont une interprétation utile. De plus, nous obtenons immédiatement les erreurs types pour les effets partiels aux valeurs moyennes. Rien ne nous interdit ensuite de remplacer μ_1 ou μ_2 par d'autres valeurs des variables explicatives qui peuvent avoir un intérêt. L'exemple suivant illustre comment on peut utiliser les termes d'interaction.

EXEMPLE 6.3
Effets de la participation sur la performance lors d'un examen final

Un modèle pour expliquer le résultat standardisé d'un examen final (*stndfnl*) en fonction du pourcentage de participation (présence) aux cours, de la moyenne des résultats obtenus précédemment avant à l'université et du sore ACT peut être défini par

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 + \beta_5 ACT^2 + \beta_6 priGPA\ atndrte + u \qquad \textbf{[6.18]}$$

(On utilise le résultat standardisé à l'examen pour les raisons discutées à la section 6.1 : il est ainsi plus facile d'interpréter la performance de l'étudiant par rapport au reste de la classe.) En plus des termes quadratiques pour *priGPA* et *ACT*, un terme d'interaction entre *priGPA* et le taux de participation est aussi présent dans le modèle. L'idée est que le taux de participation pourrait avoir un effet différent pour les étudiants ayant obtenu des résultats différents dans le passé (résultats mesurés par *priGPA*). On s'intéresse aux effets de la participation sur les résultats à l'examen : $\Delta stndfnl/\Delta atndrte = \beta_1 + \beta_6 priGPA$.

En utilisant les 680 observations dans ATTEND, pour des étudiants inscrits à un cours de principes microéconomiques, l'équation estimée est

$$\widehat{stndfnl} = 2{,}05 - 0{,}0067\ atndrte - 1{,}63\ priGPA - 0{,}128\ ACT \quad [6.19]$$
$$(1{,}36)\ (0{,}0102) \qquad (0{,}48) \qquad (0{,}098)$$
$$+ 0{,}296\ priGPA^2 + 0{,}0045\ ACT^2 + 0{,}0056\ priGPA\ atndrte$$
$$n = 680,\ R^2 = 0{,}229,\ \overline{R^2} = 0{,}222$$

Nous devons interpréter cette équation avec une extrême attention. Si on regarde simplement les coefficients de *atndrte*, on va conclure de manière incorrecte que la participation a un effet négatif sur le résultat de l'examen final. Mais ce coefficient mesure l'effet sans intérêt correspondant à *priGPA* = 0 (dans cet échantillon, le plus petit *priGPA* est d'environ 0,86). On doit aussi prendre garde à ne pas regarder séparément les estimations de β_1 et β_6 en concluant que parce que chaque statistique t est non significative, on ne peut rejeter $H_0 : \beta_1 = 0,\ \beta_6 = 0$. En fait, la p-valeur du test F de cette hypothèse jointe est 0,014, de telle sorte que nous rejetons certainement H_0 au seuil 5 %. C'est un bon exemple du fait que regarder les statistiques t séparément quand on teste une hypothèse jointe peut mener à des erreurs.

Comment doit-on estimer l'effet partiel de *atndrte* sur *stndfnl* ? On doit calculer l'effet partiel pour des valeurs d'intérêt de *priGPA*. La valeur moyenne de *priGPA* dans l'échantillon est 2,59, de telle sorte qu'à la moyenne de *priGPA*, l'effet de *atndrte* sur *stndfnl* est $-0{,}0067 + 0{,}0056(2{,}59) \approx 0{,}0078$. Que cela signifie-t-il ? Parce que *atndrte* est mesuré en pourcentage, cela signifie qu'une augmentation de 10 points de pourcentage de *atndrte* augmente *stndfnl* de 0,078 écart-type au-dessus de la moyenne du résultat de l'examen final.

Pour aller plus loin 6.3

Si nous ajoutons le terme $\beta_7 ACT \cdot atndrte$ à l'équation (6.18), quel est l'effet partiel de *atndrte* sur *stndfnl* ?

Comment pouvons-nous dire si l'estimation 0,0078 est statistiquement différente de zéro ? Il faut à nouveau faire tourner la régression pour laquelle nous remplaçons *priGPA*·*atndrte* par (*priGPA* – 2,59)·*atndrte*. Cela fournit donc, comme nouveau coefficient de *atndrte*, l'effet estimé à *priGPA* = 2,59 ainsi que son écart-type estimé type ; rien d'autre ne change dans la régression. (Ce calcul est décrit en section 4.4.) La nouvelle régression donne alors l'écart-type estimé de $\hat{\beta}_1 + \hat{\beta}_6(2{,}59) = 0{,}0078$: 0,0026, ce qui entraîne $t = 0{,}0078/0{,}0026 = 3$. Donc, à la moyenne de *priGPA*, nous concluons que la participation a un effet positif significatif sur le résultat de l'examen final.

Il est plus compliqué de trouver l'effet de *priGPA* sur *stndfnl* à cause du terme quadratique $priGPA^2$. Pour obtenir l'effet aux valeurs moyennes de *priGPA* et de *atndrte* (82), $priGPA^2$ est remplacé par $(priGPA - 2{,}59)^2$ et *priGPA atndrte* par *priGPA* (*atndrte* – 82). Le coefficient de *priGPA* devient alors l'effet partiel aux valeurs moyennes et l'écart-type estimé est ainsi obtenue. (Voir l'exercice C7.)

Calculer des effets partiels moyens

Comprendre des effets partiels qui dépendent des valeurs d'une ou plusieurs variables explicatives est une caractéristique des modèles avec des termes quadratiques, des interactions et d'autres formes fonctionnelles non linéaires. Par exemple, on vient de voir dans l'exemple 6.3 que l'effet de *atndrte* dépend de la valeur de *priGPA*. Il est aisé d'observer que l'effet partiel de *priGPA* dans l'équation (6.18) est

$$\beta_2 + 2\beta_4 priGPA + \beta_6 atndrte$$

(on peut le vérifier avec du calcul simple ou simplement en combinant les formules quadratiques et d'interactions). Ces raffinements dans l'équation (6.18) peuvent être utiles afin de voir comment la force des associations entre *stndfnl* et chaque variable explicative change avec les valeurs de toutes les variables explicatives. La flexibilité apportée par un modèle comme (6.18) a un coût : il est difficile de décrire les effets partiels des variables explicatives sur *stndfnl* avec un seul nombre.

Souvent, on désire une seule valeur pour décrire la relation entre la variable dépendante y et chaque variable explicative. L'effet partiel moyen (EPM), appelé aussi *effet marginal moyen* est une mesure agrégée populaire. L'idée derrière l'EPM est simple pour des modèles comme (6.18). Après le calcul des effets partiels et l'insertion des paramètres estimés, on effectue la moyenne des effets estimés pour chaque unité dans l'échantillon. Ainsi, l'effet partiel estimé de *atndrte* sur *stndfnl* est

$$\hat{\beta}_1 + \hat{\beta}_6 priGPA_i.$$

On ne désire pas reporter cet effet partiel pour chacun des 680 étudiants dans notre échantillon. On effectue plutôt la moyenne de ces effets partiels pour obtenir

$$EPM_{stndfnl} = \hat{\beta}_1 + \hat{\beta}_6 \overline{priGPA},$$

où $\overline{priGPA}$ est la moyenne échantillon de *priGPA*. Le nombre $EPM_{stndfnl}$ est l'EPM estimé. L'EPM de *priGPA* est juste un peu plus compliqué :

$$EPM_{priGPA} = \hat{\beta}_2 + 2\hat{\beta}_4 \overline{priGPA} + \hat{\beta}_6 \overline{atndrte}.$$

$EPM_{stndfnl}$ et EPM_{priGPA} nous informent sur l'importance des effets partiels en moyenne. Le centrage des variables explicatives autour de leur moyenne avant de créer des termes quadratiques ou d'interactions force les coefficients sur les niveaux à être les EPM. Certains packages communément utilisés calculent les EPM avec une simple commande après l'estimation par les MCO. Des erreurs types propres sont calculées en utilisant le fait qu'un EPM est une combinaison linéaire des coefficients MCO. Par exemple, les EPM et leurs erreurs types pour des modèles avec des termes quadratiques et d'interactions, comme dans l'exemple 6.3, sont faciles à obtenir.

Les EPM sont aussi utiles dans les modèles non linéaires en les paramètres, ce qui est traité au chapitre 17. Nous y revisiterons la définition et le calcul des EPM.

6.3 COMPLÉMENTS SUR L'AJUSTEMENT ET LA SÉLECTION DES RÉGRESSEURS

Jusqu'à présent, l'accent n'a pas été porté sur la valeur du R-carré dans l'évaluation des modèles de régression car les étudiants débutants ont tendance à accorder trop d'importance à celui-ci. Ainsi qu'il sera brièvement abordé, choisir un ensemble de variables explicatives sur base du R-carré peut mener à des modèles insensés. Au chapitre 10, des R-carrés rendus artificiellement hauts pour des régressions sur des séries temporelles mèneront à des conclusions erronées.

Les hypothèses du modèle linéaire classique ne demandent pas que le *R*-carré se trouve au-dessus d'une certaine valeur ; le *R*-carré constitue simplement un estimateur de la manière dont la variation de y est expliquée $x_1, x_2, \ldots, x_k$ dans la population. Il n'est pas rare d'observer des régressions avec de petits *R*-carrés ; si plusieurs facteurs affectant y n'ont pas été pris en compte, cela ne signifie pas pour autant que ces facteurs dans u soient corrélés avec les variables indépendantes. L'hypothèse de moyenne conditionnelle nulle RLM.4 permet de déterminer si les estimateurs des effets des variables indépendantes (toutes les autres choses restant égales) sont non biaisés ; cette propriété n'est aucunement influencée par la valeur du *R*-carré.

Un petit *R*-carré implique que la variance de l'erreur est grande par rapport à celle de y, ce qui signifie que des difficultés pourraient survenir pour estimer précisément les β_j. Mais, rappelons qu'en section 3.4, nous avons vu qu'une grande taille d'échantillon peut compenser le problème d'une grande variance de l'erreur : s'il y a suffisamment de données, on peut estimer précisément les effets partiels même si beaucoup de facteurs non observés n'ont pas été pris en compte. Avoir des estimations suffisamment précises dépend des applications. Par exemple, supposons que des bourses pour acheter de l'équipement informatique soient allouées aléatoirement à des étudiants entrant dans une grande université. Si le montant de la bourse est vraiment déterminé de manière aléatoire, on peut estimer (toutes autres choses restant égales) l'effet du montant de la bourse sur la moyenne des points obtenus ensuite à l'université en utilisant la régression. (À cause de l'allocation aléatoire, tous les autres facteurs qui affectent les résultats des étudiants sont non corrélés avec le montant de la bourse.) Il semble raisonnable de penser que le montant de la bourse explique peu la variation des résultats des étudiants ; ainsi le *R*-carré d'une telle régression est probablement très petit. Mais avec une grande taille d'échantillon, une estimation raisonnable de l'effet de la bourse pourrait malgré tout être obtenu.

Une autre illustration du fait qu'un faible pouvoir d'explication n'a rien avoir avec une estimation non biaisée des β_j est fournie par l'analyse des données APPLE. Contrairement aux autres ensembles de données que nous avons utilisés, les variables explicatives clés dans APPLE ont été établies expérimentalement – c'est-à-dire sans se préoccuper des autres variables qui pourraient affecter la variable dépendante. La variable à expliquer, *ecolbs*, est le nombre de livres (hypothétique) de pommes écologiquement "friendly" (avec label écologique) correspondant à la demande d'une famille. On présente à chaque famille (ou plutôt au chef de famille) une description des pommes avec un label écologique ainsi que les prix des pommes classiques (*regprc*) et des pommes à label écologique hypothétique (*ecoprc*). Puisque les paires de prix ont été allouées aléatoirement à chaque famille, elles ne sont pas corrélées aux autres facteurs observés (comme le revenu de la famille) et non observés (comme le désir d'un environnement propre). Donc, la régression de *ecolbs* par rapport à *ecoprc*, *regprc* (sur tous les échantillons générés de cette manière) fournit des estimateurs non biaisés des effets des prix. Néanmoins, le *R*-carré de la régression est seulement de 0,0364 : les variables prix expliquent seulement environ 3,6 % de la variation totale de *ecolbs*. Voici donc un cas où très peu de variation de y est expliqué ; encore une fois, il s'agit de cette situation rare pour laquelle on sait que les données ont été générées de manière à pouvoir obtenir des estimateurs non biaisés des β_j. (À ce propos, ajouter des caractéristiques de la famille a un effet très faible sur le pouvoir explicatif. Voir l'exercice C11.)

Rappelons-nous cependant que le *changement* relatif du *R*-carré quand des variables sont ajoutées à une équation est très utile : la statistique F (4,41) pour tester la significativité jointe dépend crucialement de la différence des *R*-carrés entre les modèles restreints et non restreints.

Comme nous le verrons dans la section 6.4, une conséquence importante d'un faible *R*-carré est qu'il est difficile d'obtenir de bonnes prédictions. Parce que la plus grande partie de la variation de y est expliquée par des facteurs non observés (ou au moins des facteurs qui ne sont pas introduits dans le modèle), il est difficile de prédire les réalisations futures de y avec l'équation des MCO sur base d'un ensemble de valeurs des variables explicatives.

R-carré ajusté

La plupart des packages sur la régression fournissent en plus du *R*-carré une statistique appelée le **adjusted *R*-squared** (***R*-carré ajusté**). Puisqu'il est rapporté dans de nombreux travaux appliqués et parce qu'il possède de caractéristiques intéressantes, il est détaillé dans cette sous-section.

Afin de comprendre comment il peut être ajusté, il est utile d'écrire le R^2 comme

$$R^2 = 1 - (\text{SCR/n})/(\text{SCT/n}), \qquad \textbf{[6.20]}$$

où SCR est la somme des résidus au carré et SCT est la somme des carrés totaux ; comparé à l'équation (3.28), tout ce que nous avons fait est diviser SCR et SCT par n. Cette expression révèle ce que le *R*-carré estime réellement. Définissons σ_y^2, la variance de y dans la population et σ_u^2 la variance dans la population du terme d'erreur u. (Jusqu'à présent, σ^2 a été utilisé pour σ_u^2 mais il est utile d'être plus spécifique ici.) Le ***R*-carré de la population** est défini comme $\rho^2 = 1 - \sigma_u^2 / \sigma_y^2$; il s'agit de la proportion de la variation de y dans la population expliquée par les variables indépendantes. C'est ce que le *R*-carré est supposé estimer.

Le *R*-carré estime σ_u^2 par SCR/n, que l'on sait être biaisé. Pourquoi ne pas donc remplacer SCR/n par SCR/$(n - k - 1)$? Aussi, SCT/$(n - 1)$ peut être utilisé à la place de SCT/n, puisque ce dernier est un estimateur non biaisé de σ_y^2. En utilisant ces formules, on arrive à la formule du *R*-carré ajusté :

$$\overline{R}^2 = 1 - [\text{SCR} / (n - k - 1)] / [\text{SCT} / (n - 1)] = 1 - \hat{\sigma}^2 / [\text{SCT}/(n - 1)] \qquad \textbf{[6.21]}$$

parce que $\hat{\sigma}^2 = \text{SCR} / (n - k - 1)$. De part la notation utilisée pour ce *R*-carré ajusté, il est parfois appelé *R-bar carré*.

Il est également parfois appelé *R*-carré corrigé, mais cette appellation n'est pas nécessairement indiquée dans le sens où elle suppose en quelque sorte que le $\overline{R}^2$ est meilleur que le *R*-carré. Malheureusement, $\overline{R}^2$ n'est pas connu pour être un meilleur estimateur. Il s'agit d'une tentative qui suggère que $\overline{R}^2$ corrige le biais dans $\overline{R}^2$ en tant qu'estimateur de ρ^2 mais ce n'est pas le cas : le rapport de deux estimateurs non biaisés n'est pas un estimateur non biaisé.

La première qualité de $\overline{R}^2$ est qu'il impose une pénalité pour chaque ajout de variable indépendante dans le modèle. On sait que le *R*-carré ne peut jamais diminuer quand on ajoute une variable indépendante à une équation de régression : SCR n'augmente jamais quand on ajoute des variables indépendantes. La formule de $\overline{R}^2$, en revanche, dépend explicitement de k, le nombre de variables indépendantes. Il en résulte que l'ajout d'une variable indépendante à la régression diminue la valeur de SCR mais également le nombre de degrés de liberté *ddl* dans la régression $n - k - 1$. Le ratio SCR/$(n - k - 1)$ peut donc aussi bien augmenter que diminuer suite à l'ajout d'une variable indépendante à la régression.

Une caractéristique algébrique intéressante est la suivante : quand une variable indépendante est ajoutée à une équation de régression, $\overline{R}^2$ augmente si et seulement si la statistique t pour la nouvelle variable est plus grande que un en valeur absolue. (Une extension de ce $\overline{R}^2$ augmente quand un groupe de variables est ajouté à la régression si et seulement si la statistique F pour la significativité jointe des nouvelles variables est plus grande que l'unité.) Ainsi, utiliser le $\overline{R}^2$ pour décider si une certaine variable indépendante (ou un ensemble de variables) appartient à un modèle donne une réponse différente de celle des tests t ou F (parce qu'une statistique t ou F unitaire n'est pas significative aux niveaux de test habituels).

Il est parfois utile d'avoir une formule pour le $\overline{R}^2$ en fonction de R^2 :

$$\overline{R}^2 = 1 - (1 - R^2)(n - 1) / (n - k - 1). \qquad \textbf{[6.22]}$$

Par exemple, si $R^2 = 0{,}30$, $n = 51$ et $k = 10$, alors $\overline{R}^2 = 1 - 0{,}70(50)/40 = 0{,}125$. Ainsi, pour des petits n et des grands k, $\overline{R}^2$ peut se trouver substantiellement sous le R-carré. En fait, si le R-carré habituel et $n - k - 1$ sont petits, $\overline{R}^2$ peut même être négatif ! Par exemple, $R^2 = 0{,}10$, $n = 51$ et $k = 10$ entraîne $\overline{R}^2 = -0{,}125$. Un $\overline{R}^2$ négatif indique un ajustement du modèle très pauvre par rapport au nombre de degrés de liberté.

Le $\overline{R}^2$ est parfois donné avec le R^2 habituel et parfois remplace ce R^2. Il est important de se rappeler que c'est le R^2, pas le $\overline{R}^2$, qui apparaît dans la statistique F dans (4.41). La même formule avec $\overline{R}_r^2$ et $\overline{R}_{ur}^2$ n'est pas valide.

Utiliser le R-carré ajusté pour sélectionner des modèles non emboîtés

Dans la section 4.5, nous avons étudié comment calculer une statistique F pour tester la significativité jointe d'un groupe de variables ; ceci permet de décider à un seuil de test particulier si au moins une variable dans le groupe affecte la variable dépendante. Ce test ne permet pas de décider laquelle de ces variables a un effet. Dans certains cas, on veut choisir un modèle sans variable indépendante redondante et le $\overline{R}^2$ peut aider dans ce cas.

Dans l'exemple des salaires en ligue de baseball de la section 4.5, on a vu que ni *hrunsyr* ni *rbisyr* n'étaient individuellement significatifs. Ces deux variables sont hautement corrélées et donc on pourrait vouloir choisir entre les modèles

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + u$$

et

$$\log(salary) = \beta_0 + \beta_1 \text{year} + \beta_2 gamesyr + \beta_3 bavg + \beta_4 rbisyr + u.$$

Ces deux équations sont des modèles non emboîtés car aucun d'eux n'est un cas particulier de l'autre. Les statistiques F étudiées au chapitre 4 permettent seulement de comparer des modèles emboîtés : le modèle restreint est un cas particulier du modèle non restreint. Voir les équations (4.32) et (4.28) pour des exemples de modèles restreints et non restreints. Une possibilité est de créer un modèle "composite" qui contient *toutes* les variables des modèles originaux et de tester ensuite chaque modèle contre le modèle général en utilisant le test F. Le problème avec cette procédure est que soit les deux modèles peuvent être rejetés, soit ils peuvent tous les deux ne pas être rejetés (comme dans l'exemple de la section 4.5 sur les salaires en ligue de baseball). Ainsi, cette procédure ne fournit pas toujours la possibilité de choisir entre deux modèles avec des régresseurs non emboîtés.

Dans l'exercice de régression sur le salaire des joueurs de baseball, le $\overline{R}^2$ pour la régression contenant *hrunsyr* est 0,6211, et le $\overline{R}^2$ pour la régression contenant *rbisyr* est 0,6226. Donc en se basant sur le R-carré ajusté, il y a une très faible préférence pour le modèle avec *rbisyr*. Mais cette différence reste de faible ampleur et nous pourrions obtenir en pratique une réponse différente en contrôlant pour certaines variables dans l'exercice C5 du chapitre 4. (Parce que les deux modèles non emboîtés contiennent cinq paramètres, le R-carré habituel peut être utilisé pout tirer les mêmes conclusions.)

Comparer le $\overline{R}^2$ pour choisir parmi les différents ensembles non emboîtés de variables indépendantes peut s'avérer d'un grand intérêt quand ces variables représentent des formes fonctionnelles différentes. Considérons deux modèles reliant l'investissement dans la R&D aux ventes des firmes :

$$rdintens = \beta_0 + \beta_1 log(sales) + u. \qquad \textbf{[6.23]}$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u. \qquad \textbf{[6.24]}$$

Le premier modèle capture l'effet de *sales* en l'incluant sous forme logarithmique ; le second modèle le fait en utilisant une forme quadratique. Donc, le second modèle contient un paramètre de plus que le premier.

Quand l'équation (6.23) est estimée en utilisant les 32 observations de firmes chimiques dans RDCHEM, le *R*-carré est 0,061 ; le *R*-carré pour l'équation (6.24) est quant à lui 0,148. Ainsi, le modèle quadratique semble s'ajuster beaucoup mieux aux données. Mais la comparaison des *R*-carrés dans ce cas n'est pas indiquée puisque le second modèle contient un paramètre de plus que le premier. Cela signifie que (6.23) est un modèle plus parcimonieux que (6.24).

Toutes autres choses restant égales, les modèles les plus simples sont les meilleurs. Puisque le *R*-carré habituel ne pénalise pas les modèles plus compliqués, il est donc préférable d'utiliser le $\bar{R}^2$. Celui-ci pour (6.23) est égal à 0,030, tandis qu'il vaut 0,090 pour (6.24). Donc même après ajustement pour la différence en degrés de liberté, le modèle quadratique gagne. Celui-ci est d'ailleurs également préféré quand la marge bénéficiaire est ajoutée à chaque régression.

Il y a une importante limitation dans l'utilisation du $\bar{R}^2$ pour choisir parmi des modèles non emboîtés : on ne peut l'utiliser pour choisir entre différentes formes fonctionnelles pour la variable dépendante. Décider sur base d'une mesure d'ajustement si y ou $\log(y)$ (ou peut-être une autre transformation) doit être utilisé comme variable dépendante est une question d'intérêt pour laquelle ni le $\bar{R}^2$ ni le *R*-carré ne peuvent être utilisés. La raison en est simple : ces quantités mesurent la proportion expliquée de la variation totale d'une variable dépendante que nous utilisons dans la régression et différentes fonctions de la variable dépendante auront différents montants de variation à expliquer. Par exemple, les variations totales en y et $\log(y)$ ne sont pas les mêmes et sont même souvent très différentes. Comparer les $\bar{R}^2$ de régressions avec ces différentes formes de la variable dépendante ne dit rien quant au modèle qui s'ajuste le mieux aux données ; chacun ajuste une variable dépendante différente.

Pour aller plus loin 6.4

Expliquez pourquoi choisir un modèle en maximisant le *R*-carré est équivalent à le choisir en minimisant $\hat{\sigma}$ (ou l'erreur standard de la régression, c'est-à-dire l'écart-type des résidus).

EXEMPLE 6.4
Indemnité du chef d'entreprise et performance de la firme

Considérons deux modèles estimés reliant l'indemnité du chef d'entreprise à la performance de la firme :

$$\widehat{salary} = 830{,}63 + 0{,}0163\,sales + 19{,}63\,roe$$
$$(223{,}90) \quad (0{,}0089) \qquad (11{,}08) \qquad \textbf{[6.25]}$$
$$n = 209,\ R^2 = 0{,}029,\ \bar{R}^2 = 0{,}020$$

et

$$\widehat{salary} = 4{,}36 + 0{,}275\,lsales + 0{,}0179\,roe$$
$$(0{,}29) \quad (0{,}033) \qquad (0{,}0040) \qquad \textbf{[6.26]}$$
$$n = 209,\ R^2 = 0{,}282,\ \bar{R}^2 = 0{,}275$$

où *roe* est le rendement des capitaux propres discuté au chapitre 2. Par simplicité, *lsalary* et *lsales* correspondent aux logs naturels de *salary* et de *sales* respectivement. Nous savons déjà comment interpréter ces différentes estimations d'équations. Mais peut-on dire qu'un modèle s'ajuste mieux aux données que l'autre ?

Le R^2 pour l'équation (6.25) montre que *sales* et *roe* expliquent seulement environ 2,9 % de la variation du salaire du chef d'entreprise dans l'échantillon. *sales* et *roe* ont par ailleurs des significativités statistiques marginales.

L'équation (6.26) montre que log(*sales*) et *roe* expliquent environ 28,2 % de la variation de log(*salary*). En termes d'ajustement, ce *R*-carré beaucoup plus élevé semble impliquer que le modèle (6.26) est bien meilleur. Malheureusement, ce n'est pas nécessairement le cas. La somme totale des carrés pour *salary* dans l'échantillon est 391 732,982, tandis que la somme totale des carrés pour log(*salary*) est seulement de 66,72. Donc, il y a beaucoup moins de variation qui doit être expliquée dans log(*salary*). À ce stade, on peut utiliser d'autres caractéristiques que le *R*-carré ou le $\bar{R}^2$ pour choisir entre ces modèles. Par exemple, log(*sales*) et *roe* sont beaucoup plus statistiquement significatifs dans (6.26) que *sales* et *roe* dans (6.25), et les coefficients dans (6.26) sont probablement d'un plus grand intérêt. Pour en être sûr, il faut cependant développer une comparaison d'ajustement valide.

Dans la section 6.4, une mesure d'ajustement sera développée pour comparer des modèles où soit *y*, soit log(*y*) apparaît.

Prendre en compte l'influence de trop de facteurs dans une analyse de régression

Dans beaucoup d'exemples rencontrés, et certainement dans la discussion sur le biais lié aux variables omises au chapitre 3, la question de l'omission dans un modèle de facteurs importants éventuellement corrélés avec la variable indépendante a été abordée. Il est aussi possible qu'on prenne en compte l'influence de trop de variables dans une analyse de régression.

Si on se focalise trop sur l'ajustement, le risque est de considérer, dans un modèle de régression, l'influence de facteurs dont on ne devrait pas tenir compte. Pour éviter cette erreur, on doit se souvenir de ce que signifie « toutes autres choses restant égales » (*ceteris paribus*) dans l'interprétation des modèles de régression multiple.

Pour illustrer cette question, supposons que nous menions une étude afin d'évaluer l'impact des taxes sur la bière (par état) sur les accidents de la route mortels. L'idée est qu'une taxe sur la bière plus haute réduira la consommation d'alcool, donc la conduite en état d'ivresse et donc le nombre d'accidents de la route mortels. Pour mesurer cet effet des taxes *ceteris paribus*, on peut modéliser *fatalities* comme une fonction de plusieurs facteurs, incluant *tax*, la taxe sur la bière :

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 miles + \beta_3 percmale + \beta_4 perc\ 16\_21 + \ldots,$$

où

miles = nombre total de miles parcourus.
percmale = pourcentage de la population mâle de l'état.
perc16_21 = pourcentage de la population entre 16 et 21 ans, et ainsi de suite…

Notons qu'aucune variable mesurant la consommation de bières n'a été introduite. Commettons-nous une erreur en omettant une telle variable ? La réponse est non. Si nous tenions compte de l'influence de la consommation de bière dans cette équation, comment les taxes sur la bière affecteraient-elles les accidents mortels ? Dans l'équation

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 beercons + \ldots,$$

β_1 mesure la différence en accidents mortels due à une augmentation d'un point de pourcentage de *tax*, en gardant *beercons* fixé. Il est difficile de comprendre en quoi ceci pourrait être intéressant. On ne doit pas

tenir compte de l'influence des différences de consommation de bière entre états, à moins que nous voulions tester un certain effet indirect des taxes sur la bière. D'autres facteurs comme le genre ou la distribution de l'âge devraient par contre pris en compte être compte.

Comme second exemple, supposons que pour un pays en voie de développement, nous voulions estimer les effets de l'usage d'un pesticide par des fermiers sur les dépenses de santé de leur famille. En plus des quantités de pesticide utilisées, doit-on inclure le nombre de visites du docteur comme variable explicative ? Non. Les dépenses de santé incluent les visites du docteur, et nous voudrions capturer tous les effets de l'utilisation du pesticide sur les dépenses de santé. Si nous incluons le nombre de visites du docteur comme variable explicative, alors, on mesure seulement les effets de l'utilisation du pesticide sur les dépenses de santé autres que les visites du docteur. Il est plus sensé d'utiliser le nombre de visites du docteur comme variable dépendante en fonction des quantités de pesticide dans une régression séparée.

Les exemples précédents constituent un excès de prise en compte de facteurs en régression multiple (« **over controlling** »). Ce phénomène résulte souvent d'une volonté de supprimer tout biais éventuel provenant d'une variable explicative manquante. Il est néanmoins important de se souvenir de la nature des termes « toutes autres choses restant égales » (*ceteris paribus*) en régression multiple. Dans certains cas, garder fixés des facteurs alors qu'ils changent lorsque la variable étudiée (*tax* ci-dessus) change n'a aucun sens.

Malheureusement, la question de savoir si oui ou non on doit tenir compte de l'influence d'un certain facteur n'est pas toujours tranchée. Par exemple, Betts (1995) étudie l'effet de la qualité de l'école secondaire (lycée) sur les gains ultérieurs. Il met en évidence le fait que si une plus grande qualité de l'école mène à un plus haut niveau d'instruction, alors tenir compte du niveau d'instruction dans une régression avec des mesures de qualité sous-estimera les gains dus à la qualité. Betts effectue l'analyse avec et sans les années d'instruction dans l'équation pour obtenir une gamme d'effets estimés de la qualité de l'école.

Afin de voir explicitement comment rechercher les plus hauts R-carrés peut mener à des problèmes, considérons l'exemple du prix du logement de la section 4.5 qui illustre le problème de tests d'hypothèses multiples. Dans cet exemple, on voulait tester la rationalité des évaluations des prix des maisons. log(*price*) a été étudié en fonction de log(*assess*), log(*lotsize*), log(*sqrft*) et *bdrms* et on a testé si ces trois dernières variables avaient des coefficients nuls dans la population alors que log(*assess*) avait un coefficient unitaire. Mais qu'en est-il si nous changeons le but de l'analyse et que nous estimons un *modèle de prix hédonique* qui permet d'obtenir les valeurs marginales de certains attributs de la maison ? Devons-nous inclure log(*assess*) dans l'équation ? Le $\bar{R}^2$ de la régression avec log(*assess*) est 0,762, tandis qu'il vaut 0,630 si on exclut log(*assess*). Si on se base uniquement sur cet ajustement, on devrait inclure log(*assess*). Mais ce n'est pas correct si notre but est de déterminer les effets de la taille du lotissement, de la superficie et du nombre de chambres sur la valeur d'une maison. Inclure log(*assess*) dans l'équation revient à garder une mesure de valeur fixée et ensuite à demander de combien une chambre supplémentaire changerait une autre mesure de valeur. Cela n'a pas de sens si on veut évaluer les attributs d'une maison.

Si on se souvient que différents modèles servent différents objectifs, et si on se concentre sur la signification de « toutes autres choses restant égales » (*ceteris paribus*) en régression, alors on n'inclura pas les mauvais facteurs dans le modèle de régression.

Ajouter des régresseurs pour réduire la variance de l'erreur

On vient de voir des exemples où certaines variables indépendantes ne devaient pas être inclues dans un modèle de régression même si celles-ci étaient corrélées avec la variable dépendante. Nous savons (voir chapitre 3) aussi qu'ajouter une variable indépendante peut introduire des problèmes de multicolinéarité. D'un

autre côté, dès qu'on retire quelque chose du terme d'erreur, ajouter une nouvelle variable réduit la variance de cette erreur. De manière générale, on ne peut savoir quel effet domine.

Cependant, un cas est très clair : on doit toujours inclure les variables dépendantes qui affectent y et ne sont corrélées avec aucune variable indépendante. Pourquoi ? Car ajouter une telle variable n'induit jamais de colinéarité dans la population (et donc la multicolinéarité dans l'échantillon devrait être négligeable) mais elle réduit la variance de l'erreur. En grands échantillons, les erreurs types de tous les estimateurs des MCO seront réduites.

Par exemple, considérons l'estimation de la demande individuelle de bière comme une fonction du prix moyen de la bière (par conté). On peut supposer que les caractéristiques individuelles ne sont pas corrélées avec les prix et donc une simple régression de la consommation de bière en fonction du prix suffit pour estimer l'effet du prix sur la demande individuelle. Mais il est possible d'obtenir une estimation plus précise de l'élasticité-prix de la demande de bière en incluant des caractéristiques individuelles telles que l'âge ou le niveau d'instruction. Si ces facteurs affectent la demande et ne sont pas corrélés au prix, alors l'écart-type estimé du coefficient du prix sera plus petite, au moins en grands échantillons.

Comme second exemple, considérons celui des bourses pour de l'équipement informatique donné au début de la section 6.3. Si, en plus de la variable sur le montant de la bourse, on tient compte de l'influence d'autres facteurs qui peuvent expliquer les résultats moyens à l'université (GPA), on peut probablement obtenir une estimation plus précise de l'effet de la bourse. Des mesures de résultats moyens et de classements dans le secondaire (lycée), les scores SAT et ACT ainsi qu'une famille de variables d'acquis antérieurs sont de bons candidats. Puisque les montants des bourses sont alloués aléatoirement, toutes les variables supplémentaires dont on peut tenir compte sont non corrélées avec le montant de la bourse ; dans l'échantillon, la multicolinéarité entre ces variables supplémentaires et le montant de la bourse doit être minimale. Par contre, l'ajout de ces variables pourrait significativement diminuer la variance de l'erreur, menant à une estimation plus précise de l'effet de la bourse. Souvenons-nous, il n'est pas ici question de problème de biais : on obtient un estimateur non biaisé et convergent qu'on ajoute ou non, les variables relatives aux performances dans le secondaire ou la famille de variables d'acquis antérieurs. La question est donc ici plutôt d'obtenir un estimateur avec une plus petite variance (estimée sur base de l'échantillon).

Liée à ce point, la situation dans laquelle l'affectation d'une mesure est aléatoire ne nécessite pas de savoir si certaines variables explicatives sont endogènes, étant donné que ces variables ne sont pas affectées par la mesure en question. Par exemple, quand on étudie l'effet du nombre d'heures d'un programme de formation sur les salaires, on peut inclure comme variable explicative le nombre d'années d'études avant le programme de formation. On ne doit pas s'inquiéter d'une possible corrélation entre l'éducation et certains facteurs omis tels que la « compétence », car on n'essaie pas ici d'estimer le rendement de l'éducation. On essaie d'estimer l'effet d'un programme de formation, et on peut inclure n'importe quelle variable de contrôle qui n'est pas affectée par cette formation sans biaiser l'effet de la formation. Ce que nous devons éviter est l'inclusion d'une variable comme le nombre d'années d'études après le programme de formation puisqu'une personne peut décider de se former plus en fonction du nombre d'heures qui lui ont été assignées dans le programme de formation.

Malheureusement, les cas où on dispose d'information sur des variables explicatives supplémentaires qui sont non corrélées avec les variables explicatives étudiées sont assez rares en sciences humaines. Mais il convient de se souvenir que quand ces variables sont disponibles, elles peuvent être inclues dans un modèle pour réduire la variance de l'erreur sans induire de multicolinéarité.

6.4 ANALYSE DES RÉSIDUS ET PRÉDICTION

Au chapitre 3, les valeurs prédites ou ajustées par les MCO ainsi que les résidus des MCO ont été définis. Ces **prévisions** sont certainement utiles mais sujettes aux variations d'échantillonnage puisqu'elles sont obtenues

en utilisant les estimateurs des MCO. Ainsi, on montre dans cette section comment obtenir des intervalles de confiance pour une prédiction à partir de la droite de régression des MCO.

Des chapitres 3 et 4, nous savons que les résidus sont utilisés pour obtenir la somme des carrés des résidus et le *R*-carré ; ils sont ainsi importants pour l'ajustement et les tests. Parfois, les économistes étudient les résidus d'observations particulières pour obtenir des informations sur les individus (ou les firmes, les maisons…) de l'échantillon.

Intervalles de confiance pour prédictions

Supposons que nous ayons l'équation estimée

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \qquad \textbf{[6.27]}$$

Pour des valeurs particulières des variables indépendantes, on obtient une prévision pour y, c'est-à-dire une estimation de la *valeur attendue* de y étant donné les valeurs particulières des variables explicatives. Soit $c_1, c_2, \ldots, c_k$ des valeurs particulières pour chacune des k variables indépendantes ; celles-ci correspondent ou non à une donnée réelle de l'échantillon. Le paramètre que nous voudrions estimer est

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_k c_k = \mathrm{E}(y | x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k). \qquad \textbf{[6.28]}$$

L'estimateur de θ_0 est

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \ldots + \hat{\beta}_k c_k. \qquad \textbf{[6.29]}$$

Bien qu'il soit facile de la calculer, cette quantité ne nous fournit pas de mesure d'incertitude sur la valeur prédite. Il est donc naturel de construire un intervalle de confiance pour θ_0 ; celui-ci sera centré sur $\hat{\theta}_0$.

Pour obtenir cet intervalle de confiance, l'écart-type estimé de $\hat{\theta}_0$ est nécessaire. Ensuite, avec un grand *ddl* (nombre de degrés de liberté), on peut construire un intervalle de confiance de niveau 95 % : $\hat{\theta}_0 \pm 2\,\sigma(\hat{\theta}_0)$ (comme toujours, on peut utiliser les percentiles exacts de **la distribution *t***.)

Comment obtenir l'écart-type estimé de $\hat{\theta}_0$? Il s'agit du même problème que celui rencontré en section 4.4 : on doit obtenir une écart-type estimé pour une combinaison linéaire d'estimateurs des MCO. Ici, le problème est même plus compliqué puisque tous les estimateurs des MCO apparaissent généralement dans $\hat{\theta}_0$ (à moins que certains c_j soient zéro). Cependant, la même astuce que celle utilisée en section 4.4 fonctionne ici. Ecrivons $\beta_0 = \theta_0 - \beta_1 c_1 - \ldots - \beta_k c_k$ et insérons-le dans l'équation

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

pour obtenir

$$y = \theta_0 + \beta_1 (x_1 - c_1) + \beta_2 (x_2 - c_2) + \ldots + \beta_k (x_k - c_k) + u. \qquad \textbf{[6.30]}$$

En d'autres termes, on soustrait la valeur c_j de chaque x_j, et on fait tourner la régression de

$$y_i \text{ contre } (x_{i1} - c_1), \ldots, (x_{ik} - c_k), \; i = 1, 2, \ldots, n. \qquad \textbf{[6.31]}$$

EXEMPLE 6.5
Intervalle de confiance pour la gpa prédite à l'université

En utilisant les données dans GPA2, on obtient l'équation suivante pour la GPA :

$$\widehat{colgpa} = 1{,}493 + 0{,}00149\,sat - 0{,}01386\,hsperc$$
$$\quad (0{,}075) \quad (0{,}00007) \quad (0{,}00056)$$
$$-0{,}06088\,hsize + 0{,}00546\,hsize^2 \qquad [6.32]$$
$$(0{,}01650) \qquad (0{,}00227)$$
$$n = 4\,137,\ R^2 = 0{,}278,\ \overline{R}^2 = 0{,}277,\ \hat{\sigma} = 0{,}560,$$

où les estimations sont présentées avec plusieurs chiffres après la virgule pour réduire les erreurs d'arrondi. Quelle est la GPA prédite à l'université, quand $sat = 1\,200$, $hsperc = 30$ et $hsize = 5$ (qui signifie 500) ? Il suffit d'insérer ces valeurs dans l'équation (6.32) : $\widehat{colgpa} = 2{,}70$ (arrondi à deux chiffres après la virgule). Malheureusement, on ne peut utiliser l'équation (6.32) directement pour obtenir un intervalle de confiance pour la valeur attendue de *colgpa* aux valeurs données des variables indépendantes. Une manière simple d'en obtenir un est de définir un nouvel ensemble de variables indépendantes : $sat0 = sat - 1\,200$, $hsperc0 = hsperc - 30$, $hsize0 = hsize - 5$ et $hsizesq0 = hsize^2 - 25$. En effectuant la régression de *colgpa* contre ces nouvelles variables indépendantes, on obtient

$$\widehat{colgpa} = 2{,}700 + 0{,}00149\,sat0 - 0{,}01386\,hsperc0$$
$$(0{,}020)(0{,}00007)(0{,}00056)$$
$$-0{,}06088\,hsize0 + 0{,}00546\,hsize0$$
$$(0{,}01650) \qquad (0{,}00227)$$
$$n = 4137,\ R^2 = 0{,}278,\ \overline{R}^2 = 0{,}277,\ \hat{\sigma} = 0{,}560$$

L'unique différence entre cette régression et (6.32) est l'ordonnée à l'origine ; il s'agit de la prévision désirée avec son écart-type estimé 0,020. Ce n'est pas un accident si les coefficients de pente, leurs erreurs standards, les R-carré etc. sont les mêmes qu'avant ; ceci permet de vérifier que les bonnes transformations ont été effectuées. On peut alors construire un intervalle de confiance de niveau 95 % pour la GPA attendue à l'université : $2{,}70 \pm 1{,}96(0{,}020)$ ou environ de 2,66 à 2,74. Cet intervalle de confiance est assez étroit, dû à la très grande taille de l'échantillon.

La valeur prédite dans (6.29) et, plus important, son écart-type estimé, sont obtenues par l'*ordonnée à l'origine* (ou terme constant ou intercept) dans la régression (6.31).

En guise d'exemple, on obtient ci-dessous un intervalle de confiance pour une prévision dans le problème de régression sur les résultats moyens à l'université (GPA), où on utilise l'information de l'école secondaire (lycée).

Puisque la variance de l'estimateur de l'ordonnée à l'origine est la plus petite quand chaque variable explicative a une moyenne (dans l'échantillon) nulle (voir question 2.5 pour le cas de la régression simple), il s'ensuit de (6.31) que la variance de la prévision est la plus petite à la valeur moyenne des x_j. ($c_j = \overline{x}_j$ for all j.) Ce résultat n'est pas surprenant puisque la droite de régression est la plus « stable » près du milieu des données. Plus les valeurs de c_j s'éloignent de $\overline{x}_j$, plus $Var\,(\hat{y})$ est grande.

La méthode précédente nous permet de construire un intervalle de confiance autour de l'estimation par MCO de $E(y|x_1, \ldots, x_k)$ pour n'importe quelles valeurs des variables explicatives. En d'autres termes, il s'agit d'un intervalle de confiance pour la valeur *moyenne* de y pour la sous-population correspondant à un ensemble donné de covariables. Mais un intervalle de confiance pour une personne moyenne dans la population n'est pas un intervalle de confiance pour une unité particulière (individu, famille, firme…) de la population. En construisant un intervalle de confiance pour une réalisation inconnue y, une autre source importante de variation doit être prise en compte : la variance de l'erreur traduisant notre ignorance sur les facteurs non observés affectant y.

Soit y^0 la valeur pour laquelle on voudrait construire un intervalle de confiance, aussi appelé **intervalle de prédiction/prévision**. y^0 peut par exemple représenter une personne ou une firme dans l'échantillon original. Soit $x_1^0, \ldots, x_k^0$, les nouvelles valeurs observées des variables indépendantes et soit u^0 l'erreur non observée. Nous avons donc

$$y^0 = \beta_0 + \beta_1 x_1^0 + \beta_2 x_{2+\ldots+} \beta_k x_k^0 + u^0. \qquad \textbf{[6.33]}$$

Notre meilleure prévision de y^0 est toujours la valeur attendue de y^0 (étant donné les variables explicatives) que nous estimons avec la droite de régression des MCO : $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \hat{\beta}_2 x_2^0 + \ldots + \hat{\beta}_k x_k^0$. L'**erreur de prédiction/prévision** résultante est

$$\hat{e}^0 = y^0 - \hat{y}^0 = \beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0 + u^0 - \hat{y}^0. \qquad \textbf{[6.34]}$$

Bien sûr, $E(\hat{y}^0) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x_1^0 + E(\hat{\beta}_2)x_2^0 + \ldots + E(\hat{\beta}_k)x_k^0 = \beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0$, puisque les $\hat{\beta}_j$ sont non biaisés. (Comme précédemment, ces espérances sont toutes conditionnelles aux valeurs des variables indépendantes dans l'échantillon.) Puisque u^0 a une moyenne nulle, $E(\hat{e}^0) = 0$: la moyenne de l'erreur prédiction est donc nulle.

Pour trouver la variance de $\hat{e}^0$, notons que l'erreur u^0 n'est pas corrélée avec les $\hat{\beta}_j$ puisqu'elle n'est pas corrélée avec les erreurs dans l'échantillon utilisées pour obtenir les $\hat{\beta}_j$. Par les propriétés de la covariance (voir l'annexe B), u^0 et $\hat{y}^0$ sont donc non corrélées. Ainsi la **variance de l'erreur de prédiction** (conditionnelle à toutes les valeurs des variables indépendantes dans l'échantillon) est la somme des variances :

$$\text{Var}(\hat{e}^0) = \text{Var}(\hat{y}^0) + \text{Var}(u^0) = \text{Var}(\hat{y}^0) + \sigma^2 \qquad \textbf{[6.35]}$$

où $\sigma^2 = \text{Var}(u^0)$ est la variance de l'erreur. Il y a deux sources de variation de $\hat{e}^0$. La première est l'erreur d'échantillonnage dans $\hat{y}^0$ qui apparaît car nous avons estimé les β_j. Parce que chaque $\hat{\beta}_j$ a une variance proportionnelle à $1/n$, où n est la taille de l'échantillon, $\text{Var}(\hat{y}^0)$ est proportionnelle à $1/n$. Pour de grands échantillons, $\text{Var}(\hat{y}^0)$ peut donc être très petite. Par contre, la seconde source de variation de $\hat{e}^0$, σ^2, est la variance de l'erreur dans la population ; elle ne change pas avec la taille d'échantillon. Dans beaucoup d'exemples, σ^2 sera le terme dominant dans (6.35).

EXEMPLE 6.6
Intervalle de confiance pour la gpa future à l'université

Supposons que nous voulions un intervalle de confiance (**IC**) de niveau 95 % pour la GPA future à l'université d'un étudiant de l'enseignement secondaire (lycée) avec un score *sat* = 1,200, *hsperc* = 30 et *hsize* = 5. Dans l'exemple 6.5, un IC de niveau 95 % a été obtenu pour la GPA moyenne à l'université des étudiants ayant les caractéristiques *sat* = 1,200, *hsperc* = 30 et *hsize* = 5. On veut à présent un IC de niveau 95 % pour n'importe quel étudiant *en particulier* avec ces caractéristiques. Cet intervalle de prédiction de niveau 95 % doit tenir compte des caractéristiques non observées qui affectent la performance à l'université. Pour construire cet IC pour *colgpa*, nous avons : $\hat{\sigma}(\hat{y}^0) = 0{,}020$, $\hat{\sigma} = 0{,}560$ et donc, avec (6.36), $\hat{\sigma}(\hat{e}^0) = [(0{,}020)^2 + (0{,}560)^2]^{1/2} \approx 0{,}560$. Remarquons que $\hat{\sigma}(\hat{y}^0)$ est petit par rapport à $\hat{\sigma}$: pratiquement toute la variation dans $\hat{e}^0$ vient de la variation dans u^0. Cet IC de niveau 95 % CI est 2,70 ± 1,96(0,560) ou environ de 1,60 à 3,80. Il s'agit d'un large IC qui montre que, basé sur les facteurs que nous avons introduits, on ne peut définir exactement ce que sera la GPA future d'un individu à l'université. (Dans un sens, c'est une bonne nouvelle puisque cela signifie que le classement obtenu dans le secondaire et la performance au SAT d'un étudiant ne prédestinent pas sa performance à l'université.) De toute évidence, les caractéristiques non observées qui affectent la GPA à l'université varient largement selon les individus qui ont le même score SAT et le même classement dans le secondaire.

Sous les hypothèses du modèle linéaire classique, les $\hat{\beta}_j$ et u^0 sont distribués normalement, et donc les $\hat{e}^0$ le sont aussi (conditionnellement à toutes les valeurs des variables explicatives de l'échantillon). Précédemment, des estimateurs non biaisés pour y^0 et σ^2 (chapitre 3) ont été obtenus. En utilisant ceux-ci, on peut définir l'écart-type estimé de $\hat{e}^0$ comme

$$\hat{\sigma}(\hat{e}^0) = \{[\hat{\sigma}(\hat{y}^0)]^2 + \hat{\sigma}^2\}^{1/2}. \qquad \textbf{[6.36]}$$

En utilisant le même raisonnement que pour les statistiques t des $\hat{\beta}_j$, $\hat{e}^0 / \hat{\sigma}(\hat{e}^0)$ a une distribution t avec $n - (k + 1)$ degrés de liberté. Donc,

$$P[-t_{0,025} \leq \hat{e}^0 / \hat{\sigma}(\hat{e}^0) \leq t_{0,025}] = 0,95,$$

où $t_{0,025}$ est le 97,5ème percentile dans la distribution t_{n-k-1}. Pour de grandes valeurs de $n - k - 1$, $t_{0,025} \approx 1,96$. Un intervalle de prédiction de niveau 95 % pour y^0 est donc :

$$\hat{y}^0 \pm t_{0,025}\hat{\sigma}(\hat{e}^0) ; \qquad \textbf{[6.37]}$$

comme d'habitude, excepté pour de petits *ddl*, cet intervalle équivaut à $\hat{y}^0 \pm 2\hat{\sigma}(\hat{e}^0)$. Il est plus large que l'intervalle de confiance pour y^0 à cause de $\hat{\sigma}^2$ dans (6.36) ; il est même souvent beaucoup plus large pour traduire les facteurs dans u^0 dont on n'a pas tenu compte.

Analyse des résidus

Parfois, il est utile d'examiner les observations individuelles pour voir si la valeur réelle de la variable dépendante est au-dessus ou en-dessous de la valeur prédite ; c'est-à-dire qu'on examine les résidus pour les observations individuelles. Cette pratique est appelée **analyse de résidus**. Un exemple de travail classique de l'économiste est l'examen des résidus de la régression dans un but de conseil pour l'achat d'une maison. L'exemple des prix des maisons suivant illustre l'analyse des résidus. Le prix d'une maison est lié à de nombreuses caractéristiques observables de l'habitation. On peut lister toutes les caractéristiques que l'on considère importantes : la taille, le nombre de chambres, de salles de bain... On peut utiliser un échantillon de maisons pour estimer une relation entre le prix et les attributs ; on obtient ainsi une valeur prédite et une valeur réelle pour chaque maison. On construit alors les résidus $\hat{u}_i = y_i - \hat{y}_i$. La maison avec le résidu le plus négatif est la plus sous-évaluée par rapport à ses caractéristiques observées, pour le moins dans une situation où on ne considère que les facteurs dont on a tenu compte de l'influence. Bien sûr, un prix de vente substanciellement sous sa valeur prédite peut indiquer une caractéristique indésirable qu'on n'a pas pris en compte (et qui est donc contenue dans l'erreur). Il est également utile de calculer un IC pour le prix de vente futur d'une telle maison en utilisant la méthode décrite en (6.37).

Dans les données HPRICE1, une régression de *price* contre *lotsize*, *sqrft* et *bdrms* est appliquée. Dans l'échantillon des 88 maisons, le résidu le plus négatif est –120 206, pour la 81ème maison. Ainsi, le prix demandé pour cette maison est $120 206 sous son prix prédit.

Beaucoup d'autres utilisations de l'analyse des résidus existent. Une manière de classer les écoles de droit est d'effectuer une régression du salaire médian en début de carrière contre des caractéristiques des étudiants (telles que les scores LSAT médians, la GPA médiane de la classe de première année...) et d'obtenir ainsi une valeur prédite et un résidu pour chaque école de droit. L'école de droit avec le plus grand résidu a la plus haute valeur ajoutée (prédite). (Bien sûr, il y a encore beaucoup d'incertitude à propos de la manière avec laquelle le salaire d'un débutant doit être comparé à la médiane pour l'école entière.) Ces résidus peuvent alors être comparés aux coûts d'inscription de chaque école pour déterminer un décompte approprié des gains futurs.

L'analyse des résidus joue aussi un rôle dans les affaires juridiques. Un article du *New York Times* intitulé "Judge Says Pupil's Poverty, Not Segregation, Hurts Scores" (28/06/95) décrit une affaire judiciaire

importante. La question était de savoir si les faibles performances aux tests standardisés dans les écoles du district de Hartford (par rapport aux performances des banlieues voisines) étaient dues à la faible qualité de ces écoles à forte ségrégation. La juge a conclu que « la disparité dans les scores aux tests n'indique pas que Hartford fait du travail d'éducation de ses étudiants pauvre ou inadéquat ou que ses écoles font défaut, car les scores basés sur des facteurs socioéconomiques pertinents se trouvent environ aux niveaux attendus ». Cette conclusion est basée sur l'analyse de régression du score médian ou moyen contre les caractéristiques socioéconomiques des divers districts (comprenant les écoles) du Connecticut. La conclusion du juge suggère que, étant donné les niveaux de pauvreté des étudiants dans les écoles de Hartford, les scores aux tests réels étaient similaires à ceux prédits par une analyse de régression : le résidu pour Hartford n'était pas suffisamment négatif pour conclure que les écoles elles-mêmes étaient la cause des faibles scores aux tests.

Pour aller plus loin 6.5

Comment utiliseriez-vous l'analyse des résidus pour déterminer quels sont les acteurs de cinéma professionnels qui sont surpayés ou sous-payés par rapport à leur performance ?

Prédire y quand log(y) est la variable dépendante

Puisque la transformation logarithmique naturelle est souvent utilisée en économie empirique, cette sous-section est consacrée à la question de prédire y quand $\log(y)$ est la variable dépendante. En complément, une mesure de l'ajustement pour le modèle log sera obtenue et pourra être comparée au R-carré du modèle sans log.

Pour obtenir un prédiction/prévision, il est utile de définir $logy = \log(y)$; on insiste ici sur le fait qu'il s'agit d'un y qui est prédit dans le modèle

$$logy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \qquad \textbf{[6.38]}$$

Dans cette équation, les x_j peuvent être des transformations d'autres variables ; par exemple, on pourrait avoir $x_1 = \log(sales)$, $x_2 = \log(mktval)$, $x_3 = ceoten$ dans l'exemple du salaire des chefs d'entreprises.

Étant donné les estimateurs des MCO, on sait comment prédire $logy$ pour n'importe quelle valeur des variables indépendantes :

$$\widehat{logy} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \qquad \textbf{[6.39]}$$

Ainsi, notre premier candidat pour prédire y est simplement l'exponentielle de la valeur prédite pour $\log(y)$: $\hat{y} = \exp(\widehat{logy})$. En fait, cette technique sous-estime systématiquement la valeur attendue de y. Si (6.38) suit les hypothèses du modèle linéaire classique (LMC) RLM.1 à RLM.6, on peut montrer que

$$E(y|\mathbf{x}) = \exp(\sigma^2/2)\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k),$$

où $\mathbf{x}$ est le vecteur des variables indépendantes et σ^2 est la variance de u. [Si u ~ Normal (0, σ^2), alors la moyenne de $\exp(u)$ est $\exp(\sigma^2/2)$.] Ainsi, une estimation suffit pour prédire y :

$$\hat{y} = \exp(\hat{\sigma}^2 / 2)\exp(\widehat{logy}), \qquad \textbf{[6.40]}$$

où $\hat{\sigma}^2$ est simplement l'estimateur non biaisé de σ^2. Puisque $\hat{\sigma}$, l'écart-type estimé de l'erreur, est toujours reporté, il est facile d'obtenir des valeurs prédites pour y. $\exp(\hat{\sigma}^2 / 2) > 1$ car $\hat{\sigma}^2 > 0$. Pour de grands $\hat{\sigma}^2$, ce facteur d'ajustement peut être substanciellement plus grand que l'unité.

La prédiction dans (6.40) est biaisée mais convergente. Il n'y a pas de prédiction non biaisée pour y et, dans beaucoup de cas, (6.40) marche bien. Cependant, cette relation s'appuie sur la normalité du terme d'erreur. Dans le chapitre 5, on a montré des propriétés intéressantes des estimateurs des MCO même quand u n'est pas distribuée selon une normale. Donc, il est utile d'avoir une prédiction qui ne se repose pas sur la normalité. Si nous supposons juste que u est indépendante des variables explicatives,

$$E(y|\mathbf{x}) = \alpha_0 \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k), \qquad [6.41]$$

où α_0 est l'espérance de $\exp(u)$ (qui doit être plus grande que l'unité). Pour une estimation $\hat{\alpha}_0$, on peut prédire y de la manière suivante :

$$\hat{y} = \hat{\alpha}_0 \exp(\widehat{logy}), \qquad [6.42]$$

qui demande une fois de plus d'appliquer simplement la fonction exponentielle à la valeur prédite par le modèle log et de multiplier par $\hat{\alpha}_0$.

Deux approches permettent d'estimer α_0 sans l'hypothèse de normalité. La première est basée sur $\alpha_0 = E[\exp(u)]$. Pour estimer α_0, on remplace l'espérance dans la population par une moyenne dans l'échantillon et les erreurs non observées, u_i, par les résidus des MCO, $\hat{u}_i = \log(y_i) - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}$. L'estimateur s'écrit

$$\hat{\alpha}_0 = n^{-1} \sum_{i=1}^{n} \exp(\hat{u}_i). \qquad [6.43]$$

Il est donc obtenu par la méthode des moments (voir l'annexe C). $\hat{\alpha}_0$ est un estimateur convergent de α_0 mais est biaisé parce qu'on a remplacé u_i par $\hat{u}_i$ à l'intérieur d'un fonction non linéaire. Cette version de $\hat{\alpha}_0$ est un cas particulier de "**smearing estimate**" (ainsi appelé par Duan, 1983). Parce que les résidus des MCO ont une moyenne (dans l'échantillon) nulle, on peut montrer que pour n'importe quel ensemble de données, $\hat{\alpha}_0 > 1$ (Techniquement, $\hat{\alpha}_0$ serait égal à un si tous les résidus valaient zéro, mais cela n'arrive jamais dans les applications intéressantes.) Enfin, cet $\hat{\alpha}_0$ sera nécessairement plus grand que un puisque $\alpha_0 > 1$.

Un autre estimateur pour $\hat{\alpha}_0$ est basé sur la régression avec une ordonnée à l'origine nulle. Définissons $m_i = exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})$, de telle sorte, que par (6.41), $E(y_i|m_i) = \alpha_0 m_i$. Si on pouvait observer les m_i, on pourrait obtenir un estimateur non biaisé de α_0 en effectuant une régression de y_i contre m_i sans ordonnée à l'origine. Nous devons remplacer les β_j par leurs estimations (obtenues par les MCO) et obtenons m_i $mi = \exp(\widehat{logy}_i)$, où bien sûr, les $\widehat{logy}_i$ sont les valeurs ajustées de la régression des $logy_i$ contre $x_{i1}, \ldots, x_{ik}$ (avec une ordonnée à l'origine). Alors $\breve{\alpha}_0$ [à distinguer de $\hat{\alpha}_0$ dans (6.43)] est l'estimation de la pente des MCO dans la régression simple des y_i contre les $\hat{m}_i$ (pas d'ordonnée à l'origine) :

$$\breve{\alpha}_0 = \left(\sum_{i=1}^{n} \hat{m}_i^2 \right)^{-1} \left(\sum_{i=1}^{n} \hat{m}_i y_i \right). \qquad [6.44]$$

On appellera $\breve{\alpha}_0$, l'estimateur de régression de α_0. Comme $\hat{\alpha}_0$, $\breve{\alpha}_0$ est convergent mais biaisé. De manière intéressante, $\breve{\alpha}_0$ n'est pas nécessairement plus grand que un bien qu'il le soit dans la plupart des applications. Si $\breve{\alpha}_0$ est plus petit que un et surtout s'il est beaucoup plus petit que un, il est vraisemblable que l'hypothèse d'indépendance entre u et les x_j soit violée. Si $\breve{\alpha}_0 < 1$, une possibilité est de simplement utiliser l'estimateur (6.43), même si cela peut simplement masquer un problème concernant le modèle linéaire pour $\log(y)$. On peut résumer la (les) méthode(s) selon les étapes suivantes.

Prédire y quand la variable dépendante est log(y) :

1. Obtenir les valeurs ajustées, $\widehat{logy}_i$, et les résidus, $\hat{u}_i$, de la régression de *logy* contre $x_1, \ldots, x_k$.
2. Obtenir $\hat{\alpha}_0$ selon l'équation (6.43) ou $\tilde{\alpha}_0$ selon (6.44).
3. Pour des valeurs données de $x_1, \ldots, x_k$, obtenir $\widehat{logy}$ de (6.42).
4. Obtenir la prédiction $\hat{y}$ de (6.42) (avec $\hat{\alpha}_0$ ou $\tilde{\alpha}_0$).

On montre maintenant comment prédire les salaires des chefs d'entreprises avec cette procédure.

EXEMPLE 6.7
Prédire les salaires des chefs d'entreprises

Le modèle d'intérêt est

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \log(mktval) + \beta_3 ceoten + u,$$

où β_1 et β_2 sont des élasticités et $100 \cdot \beta_3$ est une semi-élasticité. L'estimation de cette équation en utilisant CEOSAL2 est

$$\widehat{salary} = 4{,}504 + 0{,}163\, lsales + 0{,}109\, lmktval + 0{,}0117\, ceoten \quad \textbf{[6.45]}$$
$$(0{,}257)\ \ (0{,}039)\qquad (0{,}050)\qquad (0{,}0053)$$
$$n = 117,\ R^2 = 0{,}318$$

où, pour raison de clarté, *lsalary* désigne le log de *salary* (et similairement pour *lsales* et *lmktval*). On obtient ensuite $m_i = \exp(\widehat{salary}_i)$ pour chaque observation dans l'échantillon.

L'estimateur "smearing" (6.43) vaut $\hat{\alpha}_0 = 1{,}136$ et l'estimateur de régression (6.44), $\tilde{\alpha}_0 = 1{,}117$. On peut utiliser ces estimations pour estimer *salary* pour n'importe quelle valeur de *sales*, *mktval* et *ceoten*. Pour *sales* = 5 000 (ce qui signifie 5 milliards de dollars parce que *sales* est en millions), *mktval* = 10 000 (ou 10 milliards de dollars) et *ceoten* = 10, la prédiction pour *lsalary* est par (6.45), $4{,}504 + 0{,}163 \log(5\,000) + 0{,}109 \log(10\,000) + 0{,}0117(10) \approx 7{,}013$, et $\exp(7{,}013) \approx 1110{,}983$. En utilisant l'estimation (6.43), le salaire prédit est environ 1 262 077 dollars ou \$1 262 077. L'estimateur (6.44) fournit une estimation du salaire d'à peu près \$1 240 968. Ces estimations diffèrent l'une de l'autre beaucoup moins que ce qu'elles diffèrent de la prédiction naïve \$1 110 983.

On utilise la méthode précédente pour déterminer comment le modèle avec log(*y*) explique la variable dépendante *y*. Nous disposons déjà de mesures quand *y* est la variable dépendante : le *R*-carré et le *R*-carré ajusté. Le but est de trouver une mesure d'ajustement du modèle avec log(*y*) qui puisse être comparée avec le *R*-carré d'un modèle où *y* est la variable dépendante.

Il y a différentes manières de définir une mesure d'ajustement dans le but de prédire *y* après transformation du modèle pour log(*y*). Ici, nous présentons une approche facile à implémenter qui donne la même valeur que nous estimions α_0 avec (6.40), (6.43) ou (6.44). Pour motiver cette mesure, rappelons que dans l'équation de la régression linéaire estimée par MCO,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k, \quad \textbf{[6.46]}$$

et que le *R*-carré habituel est simplement le carré de la corrélation entre y_i et $\hat{y}_i$ (voir section 3.2). Si maintenant nous calculons les valeurs ajustées avec (6.42) – $\hat{y}_i = \hat{\alpha}_0 m_i$, pour toute observation *i* –, il est alors logique d'utiliser comme *R*-carré le carré de la corrélation entre les y_i et ces valeurs ajustées. Parce que la corrélation n'est pas affectée si on multiplie par une constante, l'estimateur de α_0 utilisé importe peu. En fait, ce *R*-carré pour *y* [pas log(*y*)] est juste le carré du coefficient de corrélation entre y_i et $\hat{m}_i$. On peut directement comparer celui-ci avec

le *R*-carré correspondant à l'équation (6.46). [Puisque le calcul du *R*-carré ne dépend pas de l'estimation de α_0, il ne permet pas de choisir (6.40), (6.43) ou (6.44). Mais nous savons que (6.44) minimise la somme des carrés des résidus $y_i - \hat{m}_i$, sans terme constant (intercept). En d'autres termes, étant donné les $\hat{m}_i$, $\check{\alpha}_0$ est choisi pour produire le meilleur ajustement basé sur la somme des carrés des résidus. On veut ici choisir le modèle linéaire pour y ou celui pour log(y), et donc une mesure de *R*-carré qui ne dépend pas de comment α_0 est estimé peut convenir.]

La mesure de corrélation au carré ne dépend pas de la manière avec laquelle on estime α_0. Une seconde approche consiste à calculer un *R*-carré pour y basé sur une somme de résidus au carré. Concrètement, supposons qu'on utilise l'équation (6.43) pour estimer α_0. Alors, le résidu pour la prédiction de y_i est

$$\hat{r}_i = y_i - \hat{\alpha}_0 \exp(\widehat{logy}_i), \qquad \textbf{[6.47]}$$

et on peut utiliser ces résidus pour calculer une somme de résidus au carré. En utilisant la formule en régression linéaire pour le *R*-carré, on a

$$1 - \frac{\sum_{i=1}^{n} \hat{r}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad \textbf{[6.48]}$$

comme mesure d'ajustement alternative qui peut être comparée avec le *R*-carré de la régression linéaire pour y. Remarquons qu'on peut calculer une telle mesure pour les estimations alternatives de α_0 dans les équations (6.40) et (6.44) en insérant ces estimations à la place de $\hat{\alpha}_0$ dans (6.47). Contrairement à la corrélation au carré entre y_i et $\hat{m}_i$, le *R*-carré dans (6.48) dépendra de la manière avec laquelle on estime α_0. L'estimation qui minimise $\sum_{i=1}^{n} \hat{r}_i^2$ est celle de l'équation (6.44), mais cela ne signifie pas qu'elle doit être préférée (certainement pas si $\check{\alpha}_0 < 1$). Le but n'est pas de choisir un estimateur particulier de α_0 mais plutôt de trouver une mesure d'ajustement comparable avec le modèle linéaire pour y.

EXEMPLE 6.8
Prédire les salaires des chefs d'entreprises

Après avoir obtenu les $\hat{m}_i$, on trouve le coefficient de corrélation entre $salary_i$ et $\hat{m}_i$; il vaut 0,493. Son carré est à peu près 0,243 et correspond à la manière avec laquelle le modèle log explique la variation de *salary* (et non log(*salary*)). [Le *R*-carré de (6.45), 0,318, affirme quant à lui que le modèle log explique 31,8 % de la variation de log(*salary*).]

Comme modèle linéaire concurrent, supposons que nous estimions un modèle avec toutes les variables en niveaux (c'est-à-dire non transformées par une fonction log ou autre) :

$$salary = \beta_0 + \beta_1 sales + \beta_2 mktval + \beta_3 ceoten + u \qquad \textbf{[6.49]}$$

La variable dépendante est maintenant *salary*. On pourrait utiliser le log de *sales* ou de *mktval* dans le membre de droite, mais il semble raisonnable de considérer toutes les valeurs en dollars si *salary* apparaît à gauche. Le *R*-carré de cette équation estimée utilisant les mêmes 177 observations vaut 0,201. Le modèle log explique plus de variation de *salary*, et donc, on le préfère à (6.47) au point de vue de l'ajustement. Le modèle log est aussi préféré car il semble plus réaliste et ses paramètres sont plus faciles à interpréter.

Si on maintient l'ensemble complet d'hypothèses du modèle linéaire classique dans le modèle (6.38), on peut facilement obtenir des intervalles de prédiction pour $y^0 = \exp(\beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0 + u^0)$ quand on a estimé un modèle linéaire pour log(y). Rappelons que $x_1^0, x_2^0, \ldots, x_k^0$ sont des valeurs connues et que u^0 est l'erreur non observée qui détermine y^0 partiellement. Par l'équation (6.37), un intervalle de prédiction de niveau 95 %

pour $logy^0 = \log(y^0)$ est simplement $\widehat{logy}^0 \pm t_{0,025} \cdot \hat{\sigma}(\hat{e}^0)$, où $\hat{\sigma}(\hat{e}^0)$ est obtenu par la régression de $\log(y)$ contre $x_1, \ldots, x_k$ en utilisant les n observations originales. Soit $c_l = \widehat{logy}^0 - t_{0,025} \cdot \hat{\sigma}(\hat{e}^0)$ et $c_u = \widehat{logy}^0 + t_{0,025} \cdot \hat{\sigma}(\hat{e}^0)$, les bornes supérieure et inférieure de l'intervalle de prédiction pour $logy^0$: $P(c_l \leq logy^0 \leq c_u) = 0,95$. Puisque la fonction exponentielle est strictement croissante, il est clair que $P[\exp(c_l) \leq \exp(logy^0) \leq \exp(c_u)] = 0,95$. Donc, on peut prendre $\exp(c_l)$ et $\exp(c_u)$ comme bornes respectivement inférieure et supérieure d'un intervalle de prédiction de niveau 95 % pour y^0. Pour n grand, $t_{0,025} = 1,96$, et donc, par exemple, la borne inférieure d'un intervalle de prédiction de niveau 95 % pour y^0 vaut $\exp[-1,96 \cdot \hat{\sigma}(\hat{e}^0)]\exp(\hat{\beta}_0 + x^0\hat{\beta})$, où $x^0\hat{\beta}$ est un raccourci pour $\hat{\beta}_1 x_1^0 + \ldots + \hat{\beta}_k x_k^0$. Pour rappel, les $\hat{\beta}_j$ et $\hat{\sigma}(\hat{e}^0)$ sont obtenus par la régression avec $\log(y)$ comme variable dépendante. Puisqu'on suppose la normalité de u dans (6.38), on peut utiliser (6.40) pour obtenir une prédiction pour y^0. Contrairement à (6.37), cette prédiction ne se situera pas à égale distance des bornes $\exp(c_l)$ et $\exp(c_u)$. On peut aussi obtenir d'autres intervalles de prédiction de niveau 95 % en choisissant des quantiles différents dans la distribution t_{n-k-1}. Si $q_{\alpha 1}$ et $q_{\alpha 2}$ sont des quantiles tels que $\alpha_2 - \alpha_1 = 0,95$ on peut alors choisir $c_l = \widehat{logy}^0 + q_{\alpha 1}\hat{\sigma}(\hat{e}^0)$ et $c_u = \widehat{logy}^0 + q_{\alpha 2}\hat{\sigma}(\hat{e}^0)$.

Comme exemple, considérons la régression traitant les salaires des chefs d'entreprises, pour laquelle on cherche la prédiction aux mêmes valeurs de *sales*, *mktval* et *ceoten* que dans l'exemple 6.7. L'écart-type des résidus dans (6.45) vaut environ 0,505 et l'écart-type estimé de $\widehat{logy}^0$, 0,075. Donc, en utilisant l'équation (6.36), $\hat{\sigma}(\hat{e}^0) \approx 0,511$; comme dans l'exemple sur la GPA, la variance de l'erreur submerge l'estimation de l'erreur sur les paramètres, même si la taille d'échantillon est ici de 177. Un intervalle de prédiction de niveau 95 % pour $salary^0$ va de $\exp[-1,96 \cdot (0,511)]\exp(7,013)$ à $\exp[1,96 \cdot (0,511)]\exp(7,013)$, soit environ de 408,071 à 3 024,678, ou de \$408 071 à \$3 024 678. Cet intervalle de prédiction de niveau 95 % pour le salaire des chefs d'entreprise en des valeurs données de *sales*, *mktval* et *ceoten* est très large ; cela montre que beaucoup d'information déterminante pour le salaire n'a pas été inclue dans la régression. À ce propos, la prédiction pour le salaire, en utilisant (6.40), vaut environ \$1 262 075, ce qui est plus haut que les prédictions utilisant les autres estimations de α_0 et plus proche de la borne inférieure que de la borne supérieure de l'intervalle de prédiction de niveau 95 %.

RÉSUMÉ

Dans ce chapitre, nous avons traité d'importantes questions sur la régression multiple.

La section 6.1 a montré qu'un changement des unités de mesure d'une variable indépendante change les coefficients obtenus par les MCO de la manière attendue : si x_j est multiplié par c, son coefficient est divisé par c. Si la variable dépendante est multipliée par c, *tous* les coefficients obtenus par les MCO sont multipliés par c. Ni les statistiques t, ni les statistiques F ne sont affectées par un changement d'unités de mesure de variable (dépendante ou indépendante).

Nous avons discuté les coefficients beta, qui mesurent l'effet des variables indépendantes sur les variables dépendantes en unités d'écart-type. Les coefficients beta sont obtenus à partir d'une régression standard (avec MCO) après transformation des variables dépendantes et indépendantes en z-scores.

Nous avons fourni une discussion détaillée de la forme fonctionnelle, incluant la transformation logarithmique, les termes quadratiques et d'interaction. Il est utile de résumer certaines de nos conclusions.

CONSIDÉRATIONS PAR RAPPORT À L'UTILISATION DES LOGARITHMES

1. Les coefficients ont une interprétation en termes de changement en pourcentage. On peut ignorer les unités de mesure de toute variable apparaissant sous la forme logarithmique, et changer les unités de, disons, dollars à milliers de dollars n'a aucun effet sur le coefficient d'une variable apparaissant sous forme logarithmique.

2. Les logs sont souvent utilisés pour des montants en dollars (qui sont toujours positifs) ou des populations, surtout lorsqu'il y a beaucoup de variation. Ils sont moins souvent utilisés pour des variables mesurées en années, telles que les études, l'âge et l'expérience. Les logs sont plus rarement utilisés pour des variables qui sont déjà des pourcents ou des proportions, telles que le taux de chômage ou le taux de réussite à un test.

3. Les modèles avec $\log(y)$ comme variable dépendante satisfont souvent mieux les hypothèses du modèle linéaire classique. Par exemple, ces modèles ont de meilleures chances d'être linéaires, de satisfaire l'homoscédasticité, et la normalité est souvent plus plausible.

4. Dans de nombreux cas, prendre le log réduit largement la variation d'une variable, rendant les estimations par MCO moins sensibles à l'influence des valeurs extrêmes. Cependant, dans les cas où y est une fraction proche de zéro pour beaucoup d'observations, $\log(y_i)$ peut avoir beaucoup plus de variabilité que y_i. Pour des valeurs de y_i très proches de zéro, $\log(y_i)$ est un nombre négatif très grand en valeur absolue.

5. Si $y \geq 0$ mais $y = 0$ est possible, on ne peut pas utiliser $\log(y)$. Parfois $\log(1 + y)$ est utilisé, mais l'interprétation des coefficients est difficile.

6. Pour de grands changements dans une variable explicative, on peut calculer une estimation plus précise de l'effet de changement en pourcentage.

7. Il est plus difficile mais possible de prédire y quand on a estimé un modèle avec $\log(y)$.

CONSIDÉRATIONS QUAND ON UTILISE DES TERMES QUADRATIQUES

1. Une fonction quadratique pour une variable explicative permet d'avoir des effets croissants et décroissants.

2. Le point de retournement d'une fonction quadratique est facilement obtenu et doit être calculé afin de voir s'il a un sens.

3. Les fonctions quadratiques où les coefficients ont des signes opposés ont un point de retournement strictement positif ; si les signes sont les mêmes, le point de retournement correspond à une valeur négative de x.

4. Un coefficient du carré d'une variable visiblement petit peut en pratique être important au niveau de l'effet sur un changement de pente. Un test t peut être utilisé pour voir si le terme quadratique est statistiquement significatif, et on peut calculer la pente à différentes valeurs de x pour voir si ce terme est important en pratique.

5. Pour un modèle quadratique en la variable x, le coefficient de x mesure l'effet partiel à partir de $x = 0$, comme on peut le voir dans l'équation (6.11). Si zéro n'est pas une valeur de x possible ou intéressante, on peut, avant de calculer le carré, centrer x autour d'une valeur plus intéressante comme la moyenne dans l'échantillon. L'exercice 6.12 sur ordinateur fournit un exemple.

CONSIDÉRATIONS QUAND ON UTILISE DES TERMES D'INTERACTION

1. Les termes d'interaction permettent à l'effet partiel d'une variable explicative, disons x_1, de dépendre d'une autre variable, disons x_2, – et vice versa.

2. Interpréter les modèles avec interactions peut être délicat. Le coefficient de x_1, disons β_1, mesure l'effet partiel de x_1 sur y quand $x_2 = 0$, effet qui peut être irréalisable ou inintéressant. Centrer x_1 et x_2 autour de valeurs intéressantes avant de construire le terme d'interaction mène à une équation virtuellement plus interprétable.

3. Un test t standard peut être utilisé pour déterminer si un terme d'interaction est statistiquement significatif. Calculer les effets partiels pour différentes valeurs des variables explicatives peut être effectué pour juger de l'importance des interactions en pratique.

On a introduit le R-carré ajusté, $\bar{R}^2$, comme une alternative au R-carré habituel pour mesurer l'ajustement. Alors que le R-carré ne diminue jamais quand on ajoute une variable à la régression, le $\bar{R}^2$ pénalise le nombre de régresseurs et peut diminuer lorsqu'une variable indépendante est ajoutée. Ceci rend le $\bar{R}^2$ préférable pour choisir parmi des modèles non emboîtés avec des nombres différents de variables explicatives. Ni le R-carré, ni le $\bar{R}^2$ ne peuvent être utilisés pour comparer des modèles avec des variables dépendantes différentes. Néanmoins, comme montré en section 6.4, on peut obtenir une mesure d'ajustement pour choisir un modèle avec y ou $\log(y)$ comme variable dépendante.

Dans la section 6.3, certains problèmes subtiles liés à un excès de confiance dans le R-carré ou le $\bar{R}^2$ pour arriver à un modèle final ont été commentés : il est possible de prendre en compte l'influence de trop de facteurs dans un modèle de régression. Il est important, pour cette raison, de penser plus loin la spécification de modèle, particulièrement la nature "*ceteris paribus*" de l'équation de régression multiple. Les variables explicatives qui affectent y et ne sont pas corrélées avec les autres variables explicatives peuvent être utilisées pour réduire la variance de l'erreur sans induire de multicolinéarité.

Dans la section 6.4, on a démontré comment obtenir un intervalle de confiance pour une prédiction/prévision construite à partir de la droite de régression des MCO. On a aussi montré comment un intervalle de confiance peut être construit pour une valeur future inconnue de y.

Parfois, on veut prédire y quand $\log(y)$ est utilisé comme variable dépendante dans un modèle de régression : la section 6.4 explique à cette fin une méthode simple. Enfin, on peut s'intéresser au signe et à la valeur absolue de résidus pour des observations particulières. L'analyse des résidus peut être utilisée pour déterminer si des individus de l'échantillon ont des valeurs prédites bien au-dessus ou bien en-dessous des valeurs réelles.

MOTS-CLÉS

PROBLÈMES

1. L'équation suivante a été estimée en utilisant les données dans CEOSAL1 :

$$\widehat{\log(salary)} = 4{,}322 + 0{,}276\log(sales) + 0{,}0215\,roe - 0{,}00008\,roe^2$$
$$(0{,}324)\ \ (0{,}033)\qquad\qquad (0{,}0129)\qquad (0{,}00026)$$
$$n = 209,\ R^2 = 0{,}282.$$

Cette équation permet à *roe* d'avoir un effet sur log(*salary*) qui diminue. Cette caractéristique est-elle nécessaire ? Expliquez pourquoi ou pourquoi pas.

2. Soit $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, les estimations par MCO de la régression de y_i contre x_{i1}, ..., x_{ik}, i=1, 2, ..., n. Pour des constantes non nulles c_0, c_1, ..., c_k, montrez que l'ordonnée à l'origine et les pentes obtenues par MCO dans la régression de $c_0 y_i$ contre $c_1 x_{i1}$, ..., $c_k x_{ik}$, $i = 1, 2, \ldots, n$, sont données par $\tilde{\beta}_0 = c_0\hat{\beta}_0, \tilde{\beta}_1 = (c_0 / c_1)\hat{\beta}_1, \ldots, \tilde{\beta}_k = (c_0 / c_k)\hat{\beta}_k$. [*indice* : utilisez le fait que les $\hat{\beta}_j$ résolvent les conditions de premier ordre dans (3.13), et que les $\tilde{\beta}_j$ doivent résoudre les conditions de premier ordre induisant les variables dépendantes et indépendantes dont l'échelle a été transformée.]

3. En utilisant les données dans RDCHEM, l'équation suivante a été obtenue en utilisant les MCO :

$$\widehat{rdintens} = 2{,}613 + 0{,}00030\,sales - 0{,}0000000070\,sales^2$$
$$(0{,}429)\ \ (0{,}00014)\qquad (0{,}0000000037)$$
$$n = 32,\ R^2 = 0{,}1484.$$

i. En quel point l'effet marginal de *sales* sur *rdintens* devient-il négatif ?

ii. Faut-il garder le terme quadratique dans le modèle ? Expliquez.

iii. Définissons *salesbil* les ventes mesurées en milliards de dollars : *salesbil* = *sales*/1 000. Réécrivez l'équation estimée avec *salesbil* et *salesbil*^{2} comme variables indépendantes. Rapportez les erreurs types et le *R*-carré. [*indice* : remarquez que $salesbil^2 = sales^2/(1\,000)^2$.]

iv. Pour des raisons de rapport des résultats, quelle équation préférez-vous ?

4. Le modèle suivant permet au rendement du niveau d'instruction de dépendre du montant total du niveau d'instruction des deux parents, appelé *pareduc* :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ.pareduc + \beta_3 exper + \beta_4 tenure + u$$

i. Montrez que le rendement d'une année d'instruction dans ce modèle est

$$\Delta\log(wage) / \Delta educ = \beta_1 + \beta_2 pareduc.$$

À quel signe peut-on s'attendre pour b_2 ? Pourquoi ?

ii. En utilisant les données de WAGE2, l'équation estimée est

$$\widehat{\log(wage)} = 5{,}65 + 0{,}047\ educ + 0{,}00078\ educ.pareduc +$$
$$(0{,}13)\ \ (0{,}010)\qquad\qquad (0{,}00021)$$
$$0{,}019\ exper + 0{,}010\ tenure$$
$$(0{,}004)\qquad (0{,}003)$$
$$n = 722,\ R^2 = 0{,}169.$$

(Seulement 722 observations contiennent une information complète sur le niveau d'instruction des parents.) Interprétez le coefficient de ce terme d'interaction. Cela peut aider de choisir deux valeurs spécifiques pour *pareduc* – par exemple, *pareduc* = 32 si les deux parents ont une formation universitaire ou *pareduc* = 24 si les deux parents ont une formation secondaire (lycée) – et de comparer les rendements estimés de *educ*.

iii. Quand *pareduc* est ajouté au modèle, nous obtenons :

$$\widehat{\log(wage)} = 4{,}94 - 0{,}097\ educ + 0{,}033\ pareduc - 0{,}0016\ educ.pareduc$$
$$(0{,}13)\ (0{,}010)\qquad (0{,}00021)\qquad (0{,}0012)$$
$$0{,}020\ exper + 0{,}010\ tenure$$
$$(0{,}004)\qquad\quad (0{,}003)$$
$$n = 722, R^2 = 0{,}174.$$

Le rendement du niveau d'instruction dépend-il toujours positivement du niveau d'instruction des parents ? Testez l'hypothèse nulle que le rendement du niveau d'instruction ne dépend pas du niveau d'instruction des parents.

5. Dans l'exemple 4.2, où la variable dépendante est le pourcentage d'étudiants obtenant une note sur 10 suffisante pour réussir un examen de mathématiques (*math*10), est-il pertinent d'introduire *sci11* – le pourcentage d'élèves en dernière année de lycée (enseignement secondaire) passant un examen de sciences – comme variable explicative additionnelle ?

6. Lorsque $atndrte^2$ et $ACT \cdot atndrte$ sont ajoutées à l'équation estimée dans (6.19), le *R*-carré devient 0,232. Ces variables additionnelles sont-elles conjointement significatives au seuil de test 10 % ? Les incluriez-vous dans le modèle ?

7. Les 3 équations suivantes ont été estimées à l'aide des 1 534 observations contenues dans 401K :

$$\widehat{prate} = 80{,}29 + 5{,}44\ mrate + 0{,}269\ age - 0{,}00013\ totemp$$
$$(0{,}78)\ (0{,}52)\qquad (0{,}045)\qquad (0{,}00004)$$
$$R^2 = 0{,}100,\ \bar{R}^2 = 0{,}098$$

$$\widehat{prate} = 97{,}32 + 5{,}02\ mrate + 0{,}314\ age - 2{,}66\ \log(totemp)$$
$$(1{,}95)\ (0{,}51)\qquad (0{,}044)\qquad (0{,}28)$$
$$R^2 = 0{,}144,\ \bar{R}^2 = 0{,}142$$

$$\widehat{prate} = 80{,}62 + 5{,}34\ mrate + 0{,}290\ age - 0{,}00043\ totemp$$
$$(0{,}78)\ (0{,}52)\qquad (0{,}045)\qquad (0{,}00009)$$
$$+0{,}0000000039\ totemp^2$$
$$0{,}0000000010$$
$$R^2 = 0{,}108,\ \bar{R}^2 = 0{,}106$$

Lequel de ces modèles a votre préférence ? Pourquoi ?

8. Supposons que nous voulions estimer les effets de la consommation d'alcool (*alcohol*) sur la moyenne des notes obtenues pendant des études universitaires (*colGPA*). En plus de collecter des données sur la moyenne des notes de chaque étudiant et sur sa consommation d'alcool, nous obtenons également des informations concernant son assiduité (le pourcentage des cours de son cursus auxquels il assiste). Les résultats d'un test standardisé d'entrée à l'université (SAT) ainsi que le score réalisé à la sortie du secondaire (lycée) sont également disponibles.

i. Devrions-nous inclure les deux variables *attend* et *alcohol* comme variables explicatives d'un modèle de régression multiple ? (Réfléchissez à la manière dont vous interpréteriez $b_{alcohol}$).

ii. Faudrait-il inclure *SAT* et *hsGPA* comme variables explicatives additionnelles ? Expliquez.

9. Si nous commençons avec (6.38) sous les hypothèses du modèle linéaire classique, supposons n suffisamment grand et ignorons l'erreur d'estimation de $\hat{\beta}_j$, un intervalle de prédiction au niveau 95 % pour y^0 est $[\exp(-1{,}96\hat{\sigma})\exp(\widehat{logy^0}), \exp(1{,}96\hat{\sigma})\exp(\widehat{logy^0})]$.

La prédiction ponctuelle pour y^0 est $\hat{y}^0 = \exp(\hat{\sigma}^2 / 2)\exp(\widehat{logy^0})$.

i. Pour quelles valeurs de $\hat{\sigma}$ cette prédiction ponctuelle sera-t-elle dans l'intervalle de prédiction au niveau 95 % ? Vous semble-t-il probable que cette condition soit respectée dans la plupart des applications ?

ii. Vérifiez que la condition calculée dans la question (i) est satisfaite dans l'exemple du salaire du PDG.

10. Les deux équations suivantes ont été estimées en utilisant les données MEAPSINGLE. La variable explicative clé est *lexppp*, le logarithme des dépenses par étudiant au niveau de l'école.

$$\widehat{math4} = \underset{(59,24)}{24,49} + \underset{(4,04)}{9,01}\ lexppp - \underset{(0,071)}{0,422}\ free - \underset{(5,358)}{0,752}\ lmedinc - \underset{(0,161)}{0,274}\ pctsgle$$

$$n = 229,\ R^2 = 0,472,\ \bar{R}^2 = 0,462$$

$$\widehat{math4} = \underset{(41,7)}{149,38} + \underset{(2,82)}{1,93}\ lexppp - \underset{(0,054)}{0,060}\ free - \underset{(3,76)}{10,78}\ lmedinc - \underset{(0,111)}{0,397}\ pctsgle$$

$$+ \underset{(0,042)}{0,667}\ read4$$

$$n = 229,\ R^2 = 0,749,\ \bar{R}^2 = 0,743$$

i. En tant que décideur politique essayant d'estimer l'effet causal des dépenses par étudiant sur les performances aux tests de mathématiques, expliquez pourquoi la première équation est plus pertinente que la seconde ?

ii. Ajouter *read*4 à la régression a-t-il des effets particuliers sur les coefficients autres que β_{lexppp} et leur significativité ?

iii. Comment expliquer avec des notions basiques de régression pourquoi, dans ce cas, on préfère l'équation avec le plus petit R-carré ajusté ?

EXERCICES SUR ORDINATEUR

C1. Utilisez les données contenues dans KIELMC, seulement pour l'année 1981, afin de répondre aux questions suivantes. Ces données concernent des maisons vendues pendant l'année 1981 dans la ville de North Andover, Massachussets ; l'année 1981 correspond également à la date à laquelle a commencé la construction d'un incinérateur local.

i. Afin d'étudier les effets de la proximité de l'incinérateur sur les prix de l'immobilier, considérez le modèle de régression simple suivant :

$$\log(price) = \beta_0 + \beta_1 \log(dist) + u$$

où la variable *price* désigne le prix de chaque propriété en dollars, et la variable *dist* désigne la distance mesurée en pieds séparant la propriété de l'incinérateur. En interprétant cette équation de manière causale, à quel signe vous attendez-vous pour le coefficient β_1 si la présence de l'incinérateur a un impact négatif sur le prix de l'immobilier ? Estimez cette équation et interprétez les résultats obtenus.

ii. Ajoutez maintenant les variables log(*intst*), log(*area*), log(*land*), *rooms*, *baths* et *age* au modèle de régression simple présenté dans (i), où *intst* désigne la distance séparant la maison de l'autoroute, *area* la superficie de la maison, *land* la superficie du terrain, *rooms* le nombre total de pièces, *baths* le nombre de salles de bains, et *age* l'âge de la maison en nombre d'années. Quelle est votre nouvelle conclusion concernant l'impact de la proximité de l'incinérateur ? Expliquez pourquoi (i) et (ii) donnent des résultats contradictoires.

iii. Ajoutez $[\log(intst)]^2$ au modèle présenté dans la question (ii). Quelle est la conséquence de cet ajout ? Qu'en concluez-vous quant à l'importance de la forme fonctionnelle ?

iv. Le carré de log(*dist*) est-il significatif lorsqu'il est ajouté au modèle de la question (iii) ?

C2. Utilisez les données de WAGE1 pour cet exercice.

i. Utilisez la méthode des MCO pour estimer l'équation suivante :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

et présentez les résultats obtenus sous leur forme habituelle.

ii. La variable $exper^2$ est-elle statistiquement significative au seuil de test 1 % ?

iii. En utilisant l'approximation

$$\%\Delta\widehat{wage} \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 exper)\Delta exper,$$

trouvez le rendement approximatif de la 5ème année d'expérience professionnelle. Quel est le rendement approximatif de la 20ème année d'expérience professionnelle ?

iv. À partir de quelle valeur de la variable *exper* l'expérience additionnelle commence-t-elle à avoir un impact négatif sur la valeur prédite de la variable log(*wage*) ? Combien de personnes ont-elles plus d'expérience professionnelle que cette valeur charnière dans l'échantillon considéré ?

C3. Considérez un modèle où le rendement du niveau d'instruction dépend de l'expérience professionnelle (et vice versa) :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ.exper + u$$

i. Montrez que le rendement d'une année supplémentaire d'instruction (en format décimal), pour une valeur donnée de la variable *exper*, est $\beta_1 + \beta_3 exper$.

ii. Ecrivez l'hypothèse nulle qui affirme que le rendement du niveau d'instruction ne dépend pas du niveau de la variable *exper*. Quelle est selon vous l'alternative appropriée ?

iii. Utilisez les données présentes dans WAGE2 afin de tester l'hypothèse nulle de (ii) contre l'alternative que vous avez proposée.

iv. Nous définissons θ_1 comme désignant le rendement du niveau d'instruction (en format décimal) lorsque $exper = 10$: $\theta_1 = \beta_1 + 10\beta_3$. Calculez $\hat{\theta}_1$ et un intervalle de confiance de niveau 95 % pour θ_1. (*Indice* : écrivez $\beta_1 = \theta_1 - 10\beta_3$ et insérez ce terme dans l'équation estimée ; réarrangez. Cela vous donne la régression permettant d'obtenir l'intervalle de confiance pour θ_1).

C4. Utilisez les données présentes dans GPA2 pour cet exercice.

i. Estimez le modèle

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u,$$

où la variable *hsize* désigne la taille de la promotion (en centaines d'élèves), et présentez les résultats sous leur forme usuelle. Le terme quadratique est-il statistiquement significatif ?

ii. En utilisant l'équation estimée dans la question (i), déterminez la taille « optimale » d'une promotion dans l'enseignement secondaire (lycéens). Justifiez votre réponse.

iii. Cette analyse est-elle représentative du niveau de performance académique de *tous* les élèves en dernière année de l'enseignement secondaire (lycée) ? Expliquez.

iv. Calculez la taille optimale estimée d'une promotion d'élèves dans l'enseignement secondaire (lycéens), en utilisant log(*sat*) comme variable dépendante. Le résultat est-il très différent de ce que vous avez obtenu dans la question (ii) ?

C5. Utilisez les données du prix de l'immobilier présentes dans HPRICE1 pour cet exercice.

i. Estimez le modèle

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + u$$

et présentez vos résultats dans le format des MCO habituel.

ii. Calculez la valeur prédite de log(*price*) lorsque *lotsize* = 20 000, *sqrft* = 2 500 et *bdrms* = 4. En utilisant les méthodes présentées dans la section 6.4, trouvez la valeur prédite de la variable *price* pour les mêmes valeurs des variables explicatives.

iii. Afin d'expliquer les variations de *price*, préférez-vous le modèle présenté dans (i) ou le modèle ci-dessous ?

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u,$$

C6. Utilisez les données présentées dans VOTE1 pour cet exercice.

i. Considérez un modèle comprenant un terme d'interaction entre les dépenses :

$$voteA = \beta_0 + \beta_1 prtystrA + \beta_2 expendA + \beta_3 expendB + \beta_4 expendA \cdot expendB + u.$$

Quel est l'effet partiel de *expendB* sur *voteA*, en maintenant *prtystrA* et *expendA* fixés ? Quel est l'effet partiel de *expendA* sur *voteA* ? Le signe attendu pour le coefficient β_4 est-il évident ?

ii. Estimez l'équation présentée dans la question (i) et présentez les résultats sous la forme habituelle. Le terme d'interaction est-il statistiquement significatif ?

iii. Trouvez la valeur moyenne de la variable *expendA* dans l'échantillon considéré. Fixez *expendA* à 300 (i.e. \$300 000). Quel est l'effet estimé d'une augmentation de \$100 000 des dépenses du candidat B sur *voteA* ? Cet effet est-il important ?

iv. Maintenant fixez *expendB* à 100. Quel est l'effet estimé de $\Delta expendA = 100$ sur *voteA* ? Ce résultat a-t-il du sens ?

v. Estimez à présent un modèle remplaçant le terme d'interaction par *shareA*, le pourcentage représenté par les dépenses réalisées par le candidat A dans le montant total des dépenses de campagne. Cela a-t-il du sens de maintenir à la fois *expendA* et *expendB* fixes tout en faisant varier *shareA* ?

vi. (Demande des calculs) En considérant le modèle présenté dans la question (v), trouvez l'effet partiel d'*expendB* sur *voteA*, en maintenant *prtystrA* et *expendA* fixés. Evaluez ce dernier à *expendA* = 300 et *expendB* = 0, et commentez les résultats obtenus.

C7. Utilisez les données présentes dans ATTEND pour cet exercice.

i. Dans le modèle de l'exemple 6.3, montrez que

$$\Delta stndfnl / \Delta priGPA \approx \beta_2 + 2\beta_4 priGPA + \beta_6 atndrte$$

Utilisez l'équation (6.19) pour estimer cet effet partiel quand *priGPA* = 2,59 et *atndrte* = 82. Interprétez cette estimation.

ii. Montrez que l'équation peut être écrite de la manière suivante :

$$stndfnl = \theta_0 + \beta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4 (priGPA - 2{,}59)^2 + \beta_5 ACT^2 + \beta_6 priGPA(atndrte - 82) + u,$$

où $\theta_2 = \beta_2 + 2\beta_4(2{,}59) + \beta_6(82)$ (Notez que l'ordonnée à l'origine a été modifiée, mais ce changement est sans importance.) Utilisez cette expression afin de calculer l'écart-type estimé de $\hat{\theta}_2$ de la question (i).

iii. Supposez qu'au lieu de *priGPA*(*atndrte* – 82), vous écriviez (*priGPA* – 2,59).(*atndrte* – 82). De quelle manière interprétez-vous à présent les coefficients associés aux variables *atndrte* et *priGPA* ?

C8. Utilisez les données présentes dans HPRICE1 pour cet exercice.

i. Estimez le modèle

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

et présentez vos résultats sous la forme habituelle, en incluant l'écart-type de la régression. Calculez la valeur prédite du prix lorsque *lotsize* = 10 000, *sqrft* = 2 300 et *bdrms* = 4 ; arrondissez ce prix à l'entier le plus proche.

ii. Effectuez une régression vous permettant de calculer un intervalle de confiance de niveau 95 % pour la valeur prédite dans la question (i). Notez que votre prédiction sera légèrement affectée par l'erreur d'arrondi.

iii. Définissons à présent la variable $price^0$ comme le futur prix de vente non déterminé de la maison présentant les caractéristiques énumérées dans les questions (i) et (ii). Calculez un intervalle de confiance de niveau 95 % pour $price^0$ et commentez la largeur de cet intervalle de confiance.

C9. La base de données NBASAL contient des informations relatives aux salaires et aux statistiques de carrière de 269 joueurs de la National Basketball Association (NBA).

i. Estimez un modèle reliant les points réalisés par un joueur au cours d'un match (*points*) à ses années d'ancienneté dans la ligue (*exper*), son âge (*age*) et le nombre d'années pendant lesquelles il a joué dans le circuit universitaire (*coll*). Incluez également un terme quadratique pour la variable *exper*, les autres variables figurant sous forme linéaire. Présentez les résultats obtenus sous la forme habituelle.

ii. En maintenant les années d'université et l'âge fixés, à partir de quel niveau d'expérience observe-t-on un impact négatif d'une année d'expérience supplémentaire sur les points réalisés lors d'un match ? Ce résultat a-t-il du sens ?

iii. Proposez une explication concernant le coefficient négatif et statistiquement significatif associé à la variable *coll* (Indice : les joueurs de la NBA peuvent être recrutés avant la fin de leur cursus universitaire, et même parfois directement à la sortie du secondaire (lycée).)

iv. Ajoutez un terme quadratique pour la variable *age* dans l'équation estimée. Cela est-il nécessaire ? Qu'est-ce que cela semble impliquer concernant les effets de l'âge, une fois qu'on a tenu compte des effets de l'expérience et des années passées à l'université ?

v. Régressez à présent log(*wage*) contre *points*, *exper*, $exper^2$, *age* et *coll*. Présentez les résultats obtenus sous la forme habituelle.

vi. Testez si les variables *age* et *coll* sont conjointement significatives dans la régression de la question (v). Qu'est-ce que cela implique concernant les éventuels effets distincts de l'âge et des années passées à l'université sur les salaires, une fois que l'on a tenu compte de la productivité et de l'ancienneté ?

C10. Utilisez les données présentes dans BWGHT2 pour cet exercice.

i. Estimez l'équation

$$\log(bwght) = \beta_0 + \beta_1 npvis + \beta_2 npvis^2 + u$$

par la méthode des MCO, et présentez les résultats sous la forme habituelle. Le terme quadratique est-il statistiquement significatif ?

ii. Montrez que, en se basant sur l'équation de la question (i), le nombre de visites prénatales maximisant log(*bwght*) est estimé à 22. Combien de femmes ont eu au moins 22 visites prénatales dans l'échantillon ?

iii. La prédiction que le poids à la naissance diminue au-delà de 22 visites prénatales a-t-elle du sens ? Expliquez.

iv. Ajoutez l'âge de la mère à l'équation estimée, en utilisant une forme fonctionnelle quadratique. En maintenant *npvis* fixé, pour quel âge de la mère atteint-on la valeur maximale de poids à la naissance ? Quelle fraction de l'échantillon correspond à des femmes plus âgées que l'âge « optimal » ?

v. Diriez-vous que l'âge de la mère ainsi que le nombre de visites prénatales expliquent une part importante des variations de log(*bwght*) ?

vi. En utilisant une forme quadratique pour *npvis* et *age*, décidez laquelle de la forme log ou de la forme linéaire est la meilleure pour prédire *bwght*.

C11. Utilisez APPLE afin de vérifier certaines affirmations faites dans la section 6.3.

i. Régressez *ecolbs* contre *ecoprc*, *regprc* et présentez les résultats sous leur forme habituelle, en incluant le *R*-carré et le *R*-carré ajusté. Interprétez les coefficients associés aux variables de prix, et commentez leur signe ainsi que leur importance.

ii. Les variables de prix sont-elles statistiquement significatives ? Donnez les *p*-valeurs pour les tests *t* individuels.

iii. Quelle est la gamme de valeurs ajustées pour *ecolbs* ? Quelle fraction de l'échantillon donne-t-elle *ecolbs* = 0 ? Commentez.

iv. Pensez-vous que les variables de prix considérées conjointement permettent d'expliquer une part importante des variations de *ecolbs* ? Expliquez.

v. Ajoutez les variables *faminc*, *hhsize* (taille du ménage), *educ* et *age* à la régression de la question (i). Trouvez la *p*-valeur pour leur significativité jointe. Qu'en concluez-vous ?

vi. Procédez à deux régressions distinctes de *ecolbs* contre *ecoprc* puis de *ecolbs* contre *regprc*. Comparez les coefficients obtenus par ces régressions simples à ceux obtenus par la régression multiple de la question (i). Calculez le coefficient de corrélation entre *ecoprc* et *regprc* afin d'expliquer vos résultats.

C12. Utilisez le sous-échantillon extrait de 401KSUBS avec *fsize* = 1 ; cela restreint l'analyse aux ménages composés d'une seule personne ; voir également l'exercice sur ordinateur C8 du chapitre 4.

i. Quel est l'âge minimal des personnes présentes dans cet échantillon ? Combien de personnes ont-elles cet âge ?

ii. Dans le modèle

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 age^2 + u,$$

quelle est l'interprétation littérale de β_2 ? Est-il d'un grand intérêt considéré isolément ?

iii. Estimez le modèle de la question (ii) et présentez vos résultats sous la forme usuelle. Vous semble-t-il préoccupant que le coefficient associé à la variable *age* soit négatif ? Expliquez.

iv. Puisque les personnes les plus jeunes de l'échantillon ont 25 ans, il semble logique de penser que pour un niveau donné de revenus, le montant moyen minimal d'actifs financiers nets soit atteint à 25 ans. Rappelez-vous que l'effet partiel de *age* sur *nettfa* est $\beta_2 + 2\beta_3 age$, donc l'effet partiel à l'âge de 25 ans est $\beta_2 + 2\beta_3(25) = \beta_2 + 50\beta_3$; notez cette expression θ_2. Trouvez $\hat{\theta}_2$ et calculez la *p*-valeur pour un test bilatéral de $H_0 : \theta_2 = 0$. Vous devriez en conclure que $\hat{\theta}_2$ est petit et statistiquement très non significatif. [*Indice :* une manière d'obtenir ce résultat est d'estimer le modèle $nettfa = \alpha_0 + \beta_1 inc + \theta_2 age + \beta_3(age - 25)^2 + u$ où l'ordonnée à l'origine, α_0, est différente de β_0. Il existe également d'autres manières.]

v. Puisque que les arguments contre $H_0 : \theta_2 = 0$ sont très faibles, fixez ce dernier à zéro et estimez le modèle :

$$nettfa = \alpha_0 + \beta_1 inc + \beta_3(age - 25)^2 + u.$$

Au point de vue de l'ajustement, ce modèle s'ajuste-t-il mieux que celui présenté dans la question (ii) ?

vi. Considérez l'équation estimée dans la question (v), fixez $inc = 30$ (correspondant grossièrement à sa valeur moyenne) et représentez graphiquement la relation entre *nettfa* et *age*, mais seulement pour $age \geq 25$. Décrivez les résultats obtenus.

vii. Vérifiez si l'introduction d'un terme quadratique pour *inc* est nécessaire.

C13. Utilisez les données contenues dans MEAP00 pour répondre aux questions suivantes.

i. Estimez le modèle

$$math4 = \beta_0 + \beta_2 lexppp + \beta_2 lenroll + \beta_3 lunch + u$$

au moyen de la méthode des MCO, et présentez vos résultats sous leur forme habituelle. Toutes les variables explicatives sont-elles statistiquement significatives au seuil 5 % ?

ii. Obtenez les valeurs ajustées de la régression dans la question (i). Quelle est la gamme des valeurs ajustées ? Comparez cette dernière à la gamme des valeurs réelles de *math*4.

iii. Obtenez les résidus de la régression dans la question (i). Quelle est l'école qui a le plus grand résidu (positif) ? Donnez une interprétation de ce résidu.

iv. Ajoutez des formes quadratiques pour toutes les variables explicatives à l'équation, et testez leur significativité conjointe. Les laisseriez-vous dans le modèle ?

v. En considérant à nouveau le modèle de la question (i), divisez la variable dépendante et chaque variable explicative par son écart-type dans l'échantillon, et procédez à nouveau à la régression. (Incluez également une ordonnée à l'origine, à moins de soustraire au préalable sa moyenne à chaque variable.) En termes d'unités d'écart-type, quelle variable explicative a-t-elle l'effet le plus important sur le taux de réussite à l'examen de mathématiques ?

C14. Utilisez les données de BENEFITS pour répondre aux questions suivantes. Il s'agit d'un échantillon de données par école concernant le salaire moyen des enseignants et les avantages octroyés. Voir l'exemple 4.10 pour une contextualisation.

i. Régressez *lavgsal* contre *bs* et présentez les résultats sous la forme habituelle. Est-il possible de rejeter $H_0 : \beta_{bs} = 0$ contre une alternative bilatérale ? Est-il possible de rejeter $H_0 : \beta_{bs} = -1$ contre $H_0 : \beta_{bs} > -1$. Donnez les *p*-valeurs pour chacun des deux tests.

ii. Soit $lbs = \log(bs)$. Trouvez la gamme de valeurs pour *lbs* et calculez son écart-type. Comparez ces deux dernières grandeurs à la gamme de valeurs et à l'écart-type de *bs*.

iii. Régressez *lavgsal* contre *lbs*. Cette régression s'ajuste-t-elle mieux que celle de la question (i) ?

iv. Estimez l'équation

$$lavgsal = \beta_0 + \beta_1 bs + \beta_2 lenroll + \beta_3 lstaff + \beta_4 lunch + u$$

et présentez les résultats sous leur forme usuelle. Qu'arrive-t-il au coefficient de *bs* ? Est-il à présent statistiquement différent de zéro ?

v. Interprétez le coefficient de *lstaff*. Pourquoi est-il négatif selon vous ?

vi. Ajoutez $lunch^2$ à l'équation de la question (iv). Cette variable est-elle statistiquement significative ? Calculez le point de retournement (valeur minimale) de la forme quadratique, et montrez qu'il est dans l'intervalle des valeurs réelles prises par la variable *lunch*. Combien de valeurs réelles de la variable *lunch* sont-elles supérieures au point de retournement obtenu ?

vii. En vous basant sur les résultats de la question (vi), décrivez de quelle manière les salaires des enseignants sont liés aux taux de pauvreté des écoles. En termes salariaux, et en gardant fixés les autres facteurs, est-il plus avantageux d'enseigner dans une école avec *lunch* = 0 (pas d'élève sous le seuil de pauvreté), *lunch* = 50 ou *lunch* = 100 (tous les élèves sont éligibles pour le programme de déjeuners subventionnés) ?

ANNEXE 6A

6A. Une brève introduction aux techniques de bootstrap

Dans de nombreux cas où les formules pour les erreurs types sont difficiles à obtenir mathématiquement, ou lorsqu'elles sont considérées comme de médiocres approximations de la véritable variation de l'estimateur due à l'échantillonnage, il est possible de recourir à une méthode de ré-échantillonnage. L'idée générale est de traiter les données observées comme une population à partir de laquelle on peut extraire des échantillons. La technique la plus commune de ré-échantillonnage est le bootstrap. (Il existe en fait plusieurs versions du bootstrap, mais la plus générale et la plus facilement appliquée, est appelée bootstrap non paramétrique, et c'est celle que nous décrivons ici).

Supposons que nous disposions d'une estimation, $\hat{\theta}$, d'un paramètre d'une population, θ. Nous avons obtenu cette estimation, qui pourrait être également une fonction d'estimations obtenues au moyen de la méthode des MCO (ou d'estimations que nous traiterons dans des chapitres ultérieurs), à partir d'un échantillon aléatoire de taille n. Nous aimerions obtenir un écart-type estimé pour $\hat{\theta}$ que nous pourrions utiliser afin de construire des statistiques t ou des intervalles de confiance. Il est en fait possible d'obtenir un écart-type estimé valable en calculant des estimations à partir de différents échantillons aléatoires tirés des données d'origine.

L'implémentation est simple. Si nous listons nos observations de *1 à n*, nous tirons n nombres, avec remise, de cette liste. Cela produit une nouvelle base de données (de taille n), composée des données d'origine, mais avec de nombreuses observations apparaissant plusieurs fois (sauf dans le cas exceptionnel où nous ré-échantillonnons à l'identique la base de données originelle). À chaque fois que nous tirons un échantillon aléatoire des données d'origine, nous pouvons estimer θ au moyen de la même méthode que celle que nous avions utilisée sur les données d'origine. Soit $\hat{\theta}^{(b)}$ l'estimation dans l'échantillon bootstrap b. Maintenant, si nous répétons le ré-échantillonnage et l'estimation m fois, nous avons m nouvelles estimations, $\{\hat{\theta}^{(b)}: b=1,2,\ldots,m\}$. L'écart-type estimé bootstrap de $\hat{\theta}$ consiste simplement en l'écart-type des valeurs de $\hat{\theta}^{(b)}$, soit

$$\text{bse}(\hat{\theta}) = \left((m-1)^{-1} \sum_{b=1}^{m} (\hat{\theta}^{(b)} - \bar{\hat{\theta}})^2 \right)^{1/2}, \qquad \textbf{[6.50]}$$

où $\bar{\hat{\theta}}$ est la moyenne des estimations bootstrap.

Si l'obtention d'une estimation de θ à partir d'un échantillon de taille n ne nécessite que peu de temps de calcul, comme dans le cas de la méthode des MCO et de tous les autres estimateurs que nous allons évoquer dans ce texte, nous pouvons nous permettre de choisir un grand m, i.e. d'effectuer un grand nombre de réplications bootstrap. Une valeur habituelle est $m = 1\ 000$, mais même $m = 500$ ou une valeur relativement plus petite peut produire un écart-type estimé satisfaisant. Notez que la taille de m – le nombre de fois que nous ré-échantillonnons les données d'origine – n'a rien à voir avec la taille de l'échantillon n. (Pour certains problèmes d'estimation hors du champ de ce texte, un grand n peut contraindre à un nombre plus restreint de réplications bootstrap.) De nombreux logiciels statistiques et économétriques ont des commandes bootstrap intégrées, ce qui simplifie grandement le calcul d'erreurs types bootstrap, en particulier comparé au travail important souvent requis afin d'obtenir une formule analytique pour un écart-type asymptotique.

Dans la plupart des cas, il est en fait possible d'obtenir de meilleurs résultats en utilisant l'échantillon bootstrap pour le calcul de p-valeurs de statistiques t (et de statistiques F) ou pour l'obtention d'intervalles de confiance, qu'en obtenant un écart-type bootstrap à utiliser pour la construction de ces statistiques t ou de ces intervalles de confiance. Voir Horowitz (2001) pour un traitement plus exhaustif de cette question.

CHAPITRE 7

MODÈLE DE RÉGRESSION MULTIPLE AVEC VARIABLES QUALITATIVES : VARIABLES BINAIRES OU INDICATRICES

Traduction de Sophie Béreau

Dans les chapitres précédents, les variables indépendantes et dépendantes de nos modèles de régression multiple avaient une interprétation exclusivement *quantitative*. Les quelques exemples que nous avons traités jusqu'à présent incluent le nombre d'heures travaillées quotidiennement, le nombre d'années d'étude, la moyenne obtenue durant les premières années d'études à l'université, le niveau de la pollution de l'air, les niveaux des ventes des entreprises, ou encore le nombre d'arrestations. Dans chacun des cas, l'information relativement à l'étendue ou l'ampleur de ces variables nous apporte de l'information pertinente. En outre, dans les travaux empiriques, nous devons parfois introduire des facteurs *qualitatifs* dans nos modèles de régression. À titre d'exemple, le genre ou l'origine d'un individu, l'industrie auquel se rattache une entreprise (secteur manufacturier, vente de détail, etc.), la région des États-Unis dont une ville est issue (Sud, Nord, Ouest, etc.) sont autant de facteurs dits qualitatifs.

Ce chapitre est presque entièrement consacré à l'étude des variables qualitatives *indépendantes*. Après avoir discuté des manières les plus adaptées pour décrire l'information qualitative dans la section 7.1, nous montrons comment l'information qualitative peut être facilement incorporée dans des modèles de régression dans les sections 7.2, 7.3, et 7.4. Ces sections couvrent la majorité des approches traditionnelles permettant de faire usage des variables qualitatives indépendantes dans les analyses en coupes instantanées.

Dans la section 7.5, nous étudions le cas de la variable binaire dépendante, qui est un cas particulier de variable qualitative dépendante. Le modèle de régression multiple présente dans ce cas précis une interprétation intéressante et porte le nom de modèle à probabilités linéaires. Bien que très vivement critiqué par certains économètres, la simplicité du modèle à probabilités linéaires le rend utile dans de nombreux contextes empiriques. Nous présentons un certain nombre de critiques dont il fait l'objet dans la section 7.5, mais celles-ci sont souvent de second ordre dans la plupart des travaux empiriques.

7.1 DÉCRIRE L'INFORMATION QUALITATIVE

L'information qualitative est la plupart du temps, décrite sous forme binaire : une personne est un homme ou une femme, une personne possède ou ne possède pas d'ordinateur personnel ; une entreprise offre ou non un certain type de plan de retraite ; un État américain pratique ou non la peine de mort. Dans chacun de ces exemples, l'information pertinente peut être saisie au moyen d'une **variable binaire** ou **variable dichotomique.** En économétrie, les variables binaires sont communément appelées **variables indicatrices** ou **indicatrices**, bien que cette appellation ne soit pas particulièrement informative.

Tableau 7.1 *Descriptif partiel des données contenues dans WAGE1*

Individu	Salaire	Educ	Expér	Femme	Marié
1	3,10	11	2	1	0
2	3,24	12	22	1	1
3	3,00	11	2	0	0
4	6,00	8	44	0	1
5	5,30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11,56	16	5	0	1
526	3,50	14	5	1	0

En définissant une variable indicatrice, il convient de décider quel événement sera associé à la valeur unitaire et quel autre à la valeur nulle. Par exemple, dans le cadre de l'étude sur les déterminants du salaire, nous pourrions définir *femme* comme une variable binaire prenant la valeur un si l'individu est une femme, zéro sinon. Le nom de la variable indique l'événement associé à la valeur unitaire. La même information peut être saisie en définissant la variable *homme* comme prenant la valeur un si l'individu est un homme et zéro sinon. L'une ou l'autre de ces variables est préférable à la variable *genre* en raison du caractère équivoque de l'interprétation associée à la valeur unitaire de la variable : est-ce que *genre* = 1 correspond à l'item homme ou femme ? Si la manière dont nous dénommons nos variables n'influence pas les résultats de nos régressions, un choix judicieux permet de faciliter l'exploitation de nos résultats.

Pour aller plus loin 7.1

Supposons que dans une étude comparant les résultats d'élections entre les candidats démocrates et républicains, vous souhaitez indiquer le parti de chacun des candidats en présence. Pensez-vous que le nom *parti* soit un choix judicieux pour votre variable binaire dans ce cas ? Quel pourrait être une meilleure appellation ?

Revenons à l'exemple relatif à l'équation de salaire et supposons que nous avons choisi de dénommer *femme* notre variable indicatrice relative au genre des individus de notre étude. Nous définissons maintenant une autre variable indicatrice, *marié* égale à un si l'individu est marié et zéro sinon. Le tableau 7.1 illustre le type de données qui pourraient être issues de cette catégorisation. Nous observons que l'individu 1 est une femme non mariée, l'individu 2 une femme mariée, l'individu 3 un homme non marié, etc.

Pourquoi utiliser des valeurs binaires qui prennent pour valeurs zéro ou un pour décrire l'information qualitative ? Dans un sens, ces valeurs sont arbitraires : n'importe quel couple de valeurs différentes pourraient faire l'affaire. Le vrai bénéfice de saisir l'information qualitative au moyen de valeurs binaires tient à l'interprétation des paramètres découlant de l'estimation des modèles de régression associés comme nous allons le voir maintenant.

7.2 CAS D'UNE UNIQUE VARIABLE INDICATRICE INDÉPENDANTE

Comment incorporer de l'information binaire dans les modèles de régression ? Dans les cas les plus simples, il suffit d'adjoindre aux variables explicatives du modèle, une variable indicatrice dans l'équation. Par exemple, si l'on considère le modèle suivant relatif aux déterminants du taux de salaire horaire :

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u \qquad [7.1]$$

δ_0 désigne ici le paramètre associé à la variable indicatrice *female* de façon à la distinguer des autres paramètres du modèle (β). Par suite, nous ferons potentiellement l'usage d'autres notations selon le contexte.

Dans ce modèle (7.1), seuls deux facteurs sont supposés influencer le salaire à savoir le genre et le niveau d'éducation. Puisque la variable *female* prend la valeur 1 lorsque l'individu est une femme, et 0 lorsque l'individu est un homme, le paramètre δ_0 a l'interprétation suivante : δ_0 correspond à la différence de salaire horaire entre les femmes et les hommes, pour un même niveau d'éducation *donné* (et un même terme d'erreur u). De ce fait, le coefficient δ_0 détermine s'il existe de la discrimination envers les femmes : si $\delta_0 < 0$ alors toutes choses égales par ailleurs, les femmes gagnent en moyenne un salaire moindre que les hommes.

En termes d'espérance, si nous faisons l'hypothèse de la nullité de l'espérance conditionnelle du terme d'erreur $E(u|female, educ) = 0$, alors :

$$\delta_0 = E(wage|female = 1, educ) - E(wage|female = 0, educ)$$

Puisque *femme* = 1 correspond à la situation où l'individu est une femme et *female* = 0 à celle où il s'agit d'un homme, nous pouvons réécrire cette équation comme suit :

$$\delta_0 = \mathrm{E}(wage|female, educ) - \mathrm{E}(wage|male, educ) \quad [7.2]$$

L'élément clé ici tient au fait que le niveau d'éducation est le même dans le calcul des deux espérances ; la différence, saisie par le paramètre δ_0, est ainsi uniquement due au genre.

Cette situation peut être décrite graphiquement par un **glissement de l'ordonnée à l'origine** entre la droite de régression obtenue pour les hommes et les femmes. La figure 7.1 fait état du cas où $\delta_0 < 0$, de sorte que les hommes gagnent un montant horaire fixe supérieur à celui des femmes. La différence ne dépend pas du niveau d'éducation et cela explique pourquoi les droites reflétant le lien entre salaire et éducation sont parallèles pour les hommes et les femmes.

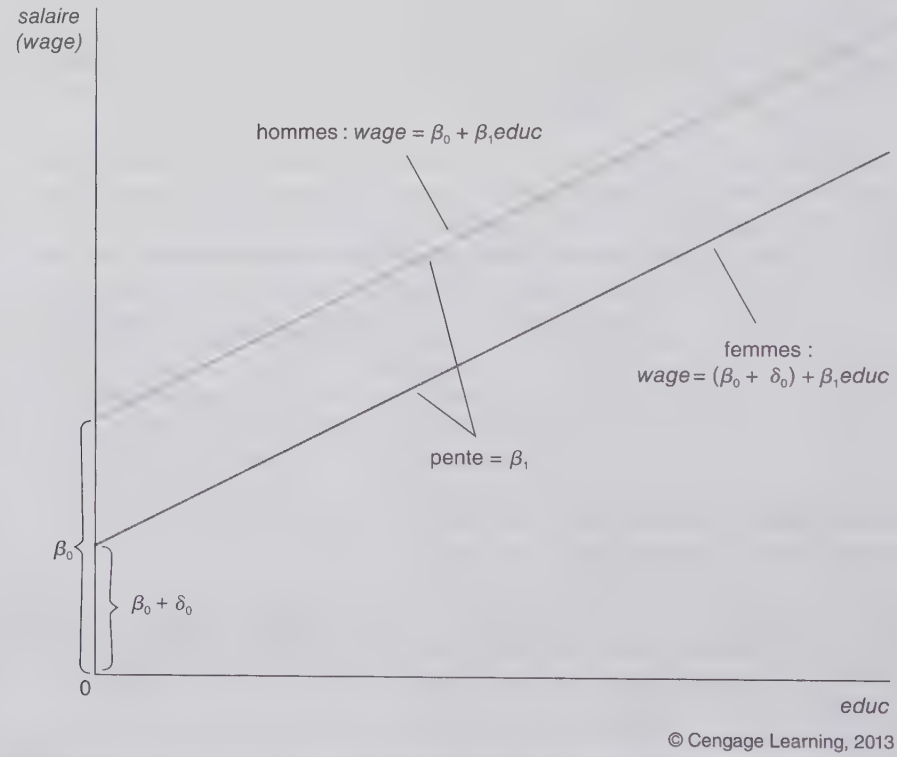


Figure 7.1 *Représentation graphique de* $wage = \beta_0 + \delta_0\, female + \beta_1\, educ$ *pour* $\delta_0 < 0$.

À ce stade de l'analyse, vous devez vous demander pourquoi nous n'avons pas également introduit dans l'équation (7.1) une variable indicatrice, mettons *male* (homme en anglais), qui prendrait la valeur 1 si l'individu est un homme et zéro sinon, la réponse étant que cela serait redondant. Dans (7.1), la constante pour les hommes est donnée par β_0, et celle pour le groupe des femmes par $\beta_0 + \delta_0$. Dans la mesure où il n'y a que deux groupes distincts, nous n'avons besoin que de deux constantes différentes. Cela implique, qu'en plus de β_0, nous n'avons besoin que d'une *unique* variable indicatrice ; celle que nous avons choisie étant la variable *female*. Utiliser deux variables indicatrices introduirait de la colinéarité parfaite dans la mesure où *female* + *male* = 1, ce qui implique que la constante du modèle est une combinaison linéaire parfaite des variables *male* et *female*. Inclure les variables indicatrices pour les deux genres constitue l'exemple le plus simple que ce que l'on nomme « **trappe à variables indicatrices** », et qui survient

lorsque l'on introduit un trop grand nombre de variables indicatrices décrivant les différents groupes en présence. Ce problème sera traité par après.

Dans (7.1), nous avons choisi les hommes comme étant le **groupe de référence** ou le **groupe témoin**, c'est-à-dire, le groupe à partir duquel les comparaisons sont établies. C'est pourquoi le paramètre β_0 correspond à la constante pour les hommes, et δ_0 à la *différence* entre les constantes valant pour les groupes des femmes et des hommes. Il est possible de choisir les femmes comme étant le groupe de référence en écrivant le modèle suivant :

$$wage = \alpha_0 + \gamma_0 male + \beta_1 educ + u,$$

où la constante pour les femmes est donnée par α_0 et celle pour les hommes par $\alpha_0 + \gamma_0$ ce qui implique que $\alpha_0 = \beta_0 + \delta_0$ et $\alpha_0 + \gamma_0 = \beta_0$. Quelle que soit l'application, le choix du groupe de référence n'apparaît pas primordial, bien qu'il soit essentiel pour l'analyse de garder à l'esprit la trace des choix effectués.

Une autre possibilité est de supprimer la constante du modèle et d'introduire une variable indicatrice pour chacun des groupes. L'équation de notre exemple devient alors : $wage = \beta_0\ male + \alpha_0\ female + \beta_1\ educ + u$, où la constante pour les hommes est β_0 et celle pour les femmes α_0. Dans ce cas, nous évitons le problème de « trappe à variables indicatrices » puisque la constante est absente du modèle. Néanmoins, cette formulation présente un désavantage par rapport aux précédentes dans la mesure où tester la différence entre les constantes est rendu plus difficile. Par ailleurs, il n'y a en général pas de méthode communément admise pour le calcul du *R*-carré en l'absence de constante. Pour cette raison, nous introduirons systématiquement une constante dans notre modèle en sus des variables indicatrices.

La démarche est similaire lorsque l'on accroît le nombre de variables explicatives. En prenant les hommes comme groupe de référence, un modèle qui tient compte de l'influence de l'expérience (*exper*) et de la titularisation (*tenure*) en sus du niveau d'éducation s'écrit alors comme suit :

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u \qquad [7.3]$$

Si *educ*, *exper*, and *tenure* sont toutes des caractéristiques pertinentes de la productivité, l'hypothèse nulle d'*absence* de différence de salaire entre hommes et femmes est donnée par $H_0 : \delta_0 = 0$, l'alternative étant l'existence de discrimination salariale envers les femmes soit $H_1 : \delta_0 < 0$.

EXEMPLE 7.1
Équation du salaire horaire

À partir des données contenues dans WAGE1, nous estimons le modèle (7.3). Pour le moment nous utilisons *wage*, plutôt que log(*wage*), comme variable dépendante :

$$\widehat{wage} = -1,57 - 1,81 female + 0,572 educ + 0,25 exper + 0,141 tenure \qquad [7.4]$$
$$(0,72)\ (0,26)\ (0,049)\ (0,012)\ (0,021)$$
$$n = 526,\ R^2 = 0,364$$

La valeur négative obtenue pour la constante – la constante pour les hommes dans ce cas – n'est pas très représentative ici puisqu'aucun individu de l'échantillon ne présente de valeur nulle simultanément pour *educ*, *exper*, et *tenure*. Le coefficient associé à *female* est quant à lui digne d'intérêt puisqu'il mesure la différence moyenne entre le salaire horaire des hommes et des femmes, présentant les mêmes niveaux pour *educ*, *exper*, et *tenure*. Prenons le cas d'un homme et d'une femme présentant les mêmes niveaux d'éducation, d'expérience ainsi que les mêmes caractéristiques en matière de titularisation, il apparaît qu'une femme gagne en moyenne 1,81 USD de moins par heure qu'un homme. (Rappelons ici qu'il s'agit des salaires de 1976.)

Il est important de rappeler que parce que nous avons réalisé des régressions multiples et tenu compte de l'influence des niveaux d'*educ*, *exper*, et *tenure*, la différence de 1,81 USD de salaire ne peut pas être expliquée par des niveaux d'éducation, d'expérience ou de titularisation différents entre hommes et femmes. Nous pouvons donc conclure que ce différentiel de 1,81 USD est dû au genre ou à des facteurs qui lui sont associés que nous n'avons pas introduits dans notre régression. [En dollars de 2003, l'écart salarial est de $3{,}23(1{,}81) \approx 5{,}85$ USD.]

Il est utile de comparer le coefficient de *female* dans l'équation (7.4) aux estimations obtenues lorsque toutes les autres variables explicatives sont omises dans l'équation :

$$\widehat{wage} = 7{,}10 - 2{,}51\, female \quad \text{[7.5]}$$
$$(0{,}21)\ \ (0{,}30)$$
$$n = 526,\ R^2 = 0{,}116$$

Le coefficient de l'équation (7.5) a une interprétation simple. La constante correspond au salaire moyen des hommes (lorsque *female* = 0), ce qui implique que les hommes gagnent en moyenne 7,10 USD par heure. Le coefficient associé à *female* représente la différence entre les salaires horaires moyens des hommes et des femmes. Ainsi, la moyenne pour les femmes de notre échantillon est de 7,10 – 2,51 = 4,59, ou 4,59 USD par heure. (À titre d'information, l'échantillon comprend ici 274 hommes et 252 femmes.)

L'équation (7.5) permet de façon simple d'établir un *test de comparaison des moyennes* entre les deux groupes, à savoir les hommes et les femmes dans notre exemple. La différence estimée est ici de – 2,51 et présente une statistique de Student ou statistique *t* de – 8,37, ce qui apparaît très significatif (et bien évidemment, 2,51 USD est également très significatif du point de vue économique). De façon générale, procéder à une régression simple sur une constante et une variable indicatrice est une manière simple de comparer des moyennes entre groupes. Pour que le test *t* standard soit valide nous devons supposer que l'hypothèse d'homoscédasticité prévaut, ce qui implique que la variance du salaire au sein de la population des hommes est la même que celle des femmes.

Le différentiel de salaire estimé entre hommes et femmes est plus important dans (7.5) que dans (7.4) car (7.5) ne tient pas compte des différences en matière d'éducation, d'expérience, et de titularisation alors que ces caractéristiques sont en effet en moyenne plus basses chez les femmes que chez les hommes dans cet échantillon. L'équation (7.4) donne donc une meilleure estimation du différentiel de salaire entre hommes et femmes toutes choses égales par ailleurs, celui-ci demeurant substantiel.

Comment tester l'existence de discrimination salariale ? La réponse est simple : estimer le modèle par les MCO, à l'instar de ce qui a été fait précédemment et recourir au test *t*. Rien de la mécanique des MCO ou de la théorie statistique ne change lorsque certaines des variables indépendantes sont des variables indicatrices. La seule différence avec les cas que que nous avons traités jusqu'à maintenant tient à l'interprétation des coefficients associés aux variables indicatrices.

EXEMPLE 7.2
Des effets de la détention d'un ordinateur personnel sur les notes obtenues en licence

Dans le but de déterminer les effets de la détention d'un ordinateur personnel sur les notes obtenues en moyenne durant les premières années d'études à l'université, nous estimons le modèle suivant :

$$colGPA = \beta_0 + \delta_0 PC + \beta_1 hsGPA + \beta_2 ACT + u,$$

où la variable indicatrice *PC* vaut un si l'étudiant dispose d'un ordinateur personnel, zéro sinon. Différentes explications peuvent prévaloir pour justifier l'impact potentiel de la détention d'un PC sur la moyenne des résultats (universitaires) des étudiants, saisis ici par la variable *colGPA*. La qualité des travaux réalisés par

l'étudiant pourrait être accrue lorsqu'il les réalise au moyen d'un ordinateur, et du temps pourrait être gagné s'il n'a pas à se déplacer en salle informatique pour cela. Bien évidemment, un étudiant disposant d'un ordinateur personnel pourrait également être tenté de jouer à des jeux vidéo ou de surfer sur internet, il n'est donc pas évident que la valeur nette de δ_0 soit positive. Les variables *hsGPA* (soient les résultats moyens obtenus au lycée) et d'*ACT* (ou *achievement test score*) sont introduites comme variables de contrôle : il est attendu que des étudiants présentant des valeurs élevées pour ces deux variables, traduisant ainsi un niveau scolaire élevé à l'issue du lycée, soient aussi les plus enclins à détenir un ordinateur personnel. Nous introduisons ces variables de contrôle dans l'équation car nous voulons mesurer ici l'impact moyen de la possession d'un ordinateur personnel sur les résultats universitaires pour un étudiant pris au hasard dans la population.

À partir des données contenues dans GPA1, nous obtenons les résultats suivants :

$$\widehat{colGPA} = 1{,}26 + 0{,}157\,PC + 0{,}447\,hsGPA + 0{,}0087\,ACT \quad [7.6]$$
$$(0{,}33)\ (0{,}057)\quad (0{,}094)\quad (0{,}0105)$$
$$n = 141,\ R^2 = 0{,}219$$

Ces résultats impliquent qu'un étudiant qui détient un ordinateur personnel présente des résultats de l'ordre 0,16 points plus élevés qu'un étudiant aux caractéristiques similaires ne possédant pas d'ordinateur (rappelez-vous ici que *colGPA* et *hsGPA* sont calculés sur une échelle à 4 points). L'effet mesuré ici est en outre très significatif puisque $t_{PC} = 0{,}157/0{,}057 \approx 2{,}75$.

Que se passe-t-il si nous retirons de l'équation les variables *hsGPA* et *ACT* ? S'il est attendu que le retrait d'*ACT* n'entraîne que peu de changements, son coefficient et sa statistique étant très faibles, celui de *hsGPA* qui apparaît très significatif pourrait affecter l'estimation du coefficient β_{pc}. Ici, la régression de *colGPA* sur *PC* donne une estimation du coefficient associé à *PC* d'environ 0,170, avec un écart-type de 0,063 ; dans ce cas, la valeur de $\hat{\beta}_{pc}$ et sa statistique *t* ne sont donc pas grandement modifiées.

Dans les exercices proposés à la fin de ce chapitre, il vous sera demandé d'introduire des variables de contrôle pour d'autres facteurs dans l'équation de façon à voir si l'impact de la détention d'un ordinateur personnel disparaît ou est tout du moins amoindri.

En règle générale, les variables indicatrices indépendantes reflètent les choix des individus ou des unités économiques considérées (par opposition aux aspects prédéterminés tel que le genre). Dans de telles situations, la question de la causalité apparaît à nouveau centrale. Dans l'exemple qui suit, nous souhaitons mieux évaluer si la détention d'un ordinateur personnel *cause* l'obtention, en moyenne, de meilleures notes durant les premières années d'études à l'université.

Chacun des exemples précédents peut être perçu comme pertinent pour l'**analyse de politique économique**. Dans le premier exemple, nous nous intéressions à la discrimination homme/femme sur le marché du travail. Dans le second exemple, nous traitions de l'impact de la détention d'un ordinateur personnel sur les résultats universitaires durant les premières années d'études. Un cas particulier de l'analyse de politique économique concerne l'**évaluation des politiques publiques**, dans le cadre de laquelle on s'intéresse à l'impact d'un ensemble de mesures économiques et sociales sur les individus, entreprises, voisinage, villes, etc.

Dans le cas le plus simple, on distingue deux groupes d'individus. Le **groupe de contrôle** qui n'est pas soumis aux réformes testées et le **groupe de traitement** qui est quant à lui soumis à ces mêmes réformes. Ces appellations sont inspirées de la littérature en sciences expérimentales mais ne doivent pas être interprétées littéralement. À de rares exceptions près, le choix des groupes de contrôle et de traitement n'est pas aléatoire. Pour autant, dans de nombreux cas, l'analyse d'un modèle de régression multiple permet de tenir compte d'un nombre suffisant de facteurs pour évaluer l'impact causal d'une réforme.

EXEMPLE 7.3
Des effets des subventions pour la formation des employés sur les heures de formation

À partir de données datant de 1988 relatives aux entreprises du secteur manufacturier de l'État du Michigan contenus dans JTRAIN, nous obtenons les résultats d'estimation suivants :

$$\widehat{hrsemp} = 46{,}67 + 26{,}25\,grant - 0{,}98\log(sales) - 6{,}07\log(employ) \quad [7.7]$$
$$(43{,}41)\quad (5{,}59)\qquad (3{,}54)\qquad\qquad (3{,}88)$$
$$n = 105,\ R^2 = 0{,}237.$$

La variable dépendante est le nombre d'heures de formation par employé en moyenne par entreprise. La variable *grant* est une variable indicatrice égale à un si l'entreprise perçoit une aide de l'État pour la formation de ses employés en 1988 et zéro sinon. Les variables *sales* et *employ* représentent quant à elles le niveau des ventes annuelles et le nombre d'employés, respectivement. Nous ne pouvons pas introduire *hrsemp* sous forme logarithmique car *hrsemp* prend la valeur nulle pour 29 des 105 entreprises considérées dans notre échantillon.

La variable *grant* apparaît très significative avec $t_{grant} = 4{,}70$. En tenant compte du niveau des ventes et du nombre d'employés, les entreprises qui perçoivent la subvention pour la formation de son personnel, ont formé leurs employés en moyenne 26,25 heures de plus que les autres. Dans la mesure où le nombre d'heures consacrées à la formation du personnel est en moyenne de 17 heures sur l'échantillon d'entreprises considérées, avec une valeur maximale de 164, *grant* apparaît avoir ici un impact important sur la formation, comme cela était attendu.

Le coefficient associé à log(*sales*) est très faible et peu significatif. Le coefficient relatif à log(*employ*) s'interprète comme suit. Si une entreprise détient une masse salariale de 10 % supérieure à la moyenne, alors elle formera en moyenne 0,61 heures de moins. La statistique *t* associée est de – 1,56, ce qui implique que l'effet est peu significatif.

À l'instar des estimations réalisées avec n'importe quelles autres variables dépendantes, nous devons nous poser la question de la nature causale de l'effet obtenu. Dans l'équation (7.7), la différence de temps consacré à la formation entre les entreprises ayant reçu la subvention et celles ne l'ayant pas reçu est-elle due à la subvention elle-même ou révèle-t-elle simplement l'existence d'un mécanisme autre ? Il se pourrait que les entreprises ayant reçu la subvention consacrent en moyenne un temps plus important à la formation de leur personnel, ceci même en l'absence de subvention. Rien dans l'analyse ne permet d'établir si ce que nous estimons est effectivement un effet causal ; nous devons pour ce faire connaître les mécanismes sous-jacents qui expliquent le fait que certaines entreprises ont bénéficié de subventions. Au mieux, nous pouvons espérer avoir pris en compte suffisamment de facteurs susceptibles de déterminer l'obtention d'une subvention ainsi que le nombre d'heures consacrées à la formation au sein des entreprises.

Nous reviendrons plus loin à l'évaluation des politiques publiques au moyen de variables indicatrices dans la section 7.6, ainsi que dans les derniers chapitres de cet ouvrage.

Interpréter des coefficients associés aux variables indicatrices explicatives lorsque la variable dépendante est log(y)

Une spécification usuelle dans les travaux appliqués consiste à introduire la variable dépendante sous une forme logarithmique, avec une ou plusieurs variables indicatrices comme variables indépendantes. Comment doit-on alors interpréter les coefficients associés à ces variables indicatrices ? Sans surprise, les coefficients ont une interprétation en termes de *pourcentage*.

EXEMPLE 7.4
Régression sur le prix de l'immobilier

À partir des données contenues dans HPRICE1, nous procédons à l'estimation du modèle suivant :

$$\begin{aligned}\widehat{\log(price)} = &-1{,}35 + 0{,}168\log(lotsize) + 0{,}707\log(sqrft) \\ &\quad (0{,}65)\ (0{,}038)\qquad\quad (0{,}093) \\ &+ 0{,}027\ bdrms + 0{,}54\ colonial \\ &\quad (0{,}29)\qquad\ (0{,}045) \\ n = 88,\ &R^2 = 0{,}649\end{aligned} \qquad [7.8]$$

Toutes les variables trouvent une interprétation évidente (*lotsize*, *sqrft*, et *bdrms* désignent la superficie du terrain, la taille de la maison en pieds carrés et le nombre de chambres resp.) à l'exception de *colonial*, qui est ici une variable binaire prenant la valeur un si la maison est de style colonial. Quelle interprétation peut-on faire des coefficients estimés de la variable *colonial* ? Pour des niveaux données de *lotsize*, *sqrft*, et *bdrms*, la différence du log des prix, $\widehat{\log(price)}$, entre une maison de style colonial et une autre d'un autre type est de 0,054. Cela signifie qu'il est attendu qu'une maison de style colonial se vende à un prix en moyenne 5,4 % plus cher qu'une maison d'un style différent, toutes choses égales par ailleurs.

Cet exemple nous montre que lorsque $\log(y)$ constitue la variable dépendante du modèle, le coefficient associé à la variable indicatrice multiplié par 100 peut être interprété comme une différence de pourcentages pour y, toutes choses égales par ailleurs. Lorsque ce coefficient suggère une différence importante en pourcentage pour y, l'exacte différence peut être obtenue en procédant comme pour le calcul de semi-élasticité explicité dans la section 6.2.

EXEMPLE 7.5
Équation du log salaire horaire

Nous nous proposons maintenant de ré-estimer l'équation de salaire horaire donnée dans l'exemple 7.1, en utilisant cette fois $\log(wage)$ comme variable dépendante ainsi que des formes quadratiques pour *exper* et *tenure* :

$$\begin{aligned}\widehat{\log(wage)} = &\ 0{,}417 - 0{,}297\ female + 0{,}080\ educ + 0{,}029 \\ &\quad (0{,}99)\ (0{,}036)\qquad (0{,}007)\qquad (0{,}005) \\ &- 0{,}00058\ exper^2 + 0{,}32\ tenure - 0{,}00059\ tenure^2 \\ &\quad (0{,}00010)\qquad\ (0{,}007)\qquad\ (0{,}00023) \\ n = 526,\ &R^2 = 0{,}441\end{aligned} \qquad [7.9]$$

En utilisant la même approximation que celle exposée dans l'exemple 7.4, le coefficient obtenu pour la variable *female* implique qu'à niveaux d'*educ*, *exper*, et *tenure* équivalents, les femmes gagnent environ 100(0,297) = 29,7% de moins que les hommes. Nous pouvons aller plus loin en calculant le taux de variation exact de la prédiction des salaires. Ce que nous voulons évaluer, c'est la différence de salaires relative entre hommes et femmes, toutes choses égales par ailleurs : $(\widehat{wage_F} - \widehat{wage_M})/\widehat{wage_M}$. À partir de (7.9) nous pouvons calculer :

$$\widehat{\log(wage_F)} - \widehat{\log(wage_M)} = -0{,}297$$

Il suffit alors de prendre l'exponentiel de ce résultat et d'y soustraire un pour obtenir :

$$(\widehat{wage_F} - \widehat{wage_M})/\widehat{wage_M} = \exp(-0{,}297) - 1 \approx -0{,}257$$

Cette évaluation plus précise implique que le salaire des femmes est en moyenne 25,7 % inférieur à celui d'un homme aux compétences similaires.

Si nous avions procédé à la même correction dans l'exemple 7.4, nous aurions obtenu exp(0,054) – 1 ≈ 0,0555, soit environ 5,6 %. Cette correction a un impact moindre dans l'exemple 7.4 que dans l'illustration relative aux salaires car la magnitude du coefficient affectant la variable indicatrice est beaucoup plus faible dans (7.8) que dans (7.9).

En général, si $\hat{\beta}_1$ est le coefficient associé à la variable indicatrice, mettons x_1, lorsque log(y) est la variable dépendante, le taux de variation en pourcentage de la variable prédite y lorsque $x_1 = 1$ contre $x_1 = 0$ est donné par :

$$100 \cdot [\exp(\hat{\beta}_1 - 1)]. \qquad \textbf{[7.10]}$$

Le coefficient estimé, $\hat{\beta}_1$, peut être positif ou négatif, il est par ailleurs important de préserver le signe lors du calcul du taux de variation. (7.10).

L'approximation logarithmique présente l'avantage de fournir une estimation intermédiaire entre les magnitudes obtenues en utilisant l'un ou l'autre des deux groupes comme référence. En particulier, bien que l'équation (7.10) nous donne une meilleure estimation que 100 $\hat{\beta}_1$ de la proportion avec laquelle y augmente quand $x_1 = 1$ plutôt que $x_1 = 0$, (7.10) n'est plus une bonne estimation si nous changeons de groupe de référence.

Dans l'exemple 7.5, nous pouvons estimer de combien en pourcentage le salaire d'un homme excède celui d'une femme, à compétences équivalentes, et cette estimation est donnée par [exp ($-\hat{\beta}_1$) – 1] = 100. [exp (0,297) – 1] ≈ 34,6. L'approximation reposant sur le calcul de 100. $\hat{\beta}_1$, 29,7, est comprise entre 25,7 et 34,6 (et proche de la valeur moyenne). Il apparaît donc légitime de reporter que « la différence de salaire prédit entre les hommes et les femmes est d'environ 29,7 % » sans avoir à préciser le groupe de référence.

7.3 UTILISER DES VARIABLES INDICATRICES À CATÉGORIES MULTIPLES

Il est possible d'utiliser plusieurs variables indicatrices indépendantes dans la même équation. Par exemple, nous pourrions introduire la variable indicatrice *married* à l'équation (7.9). Le coefficient associé à *married* donne alors une approximation du taux de variation des salaires entre les individus mariés et non mariés, toutes choses égales par ailleurs, c'est-à-dire à niveaux d'*educ*, *exper*, et *tenure* fixés. Lorsque nous estimons ce modèle, le coefficient associé à la variable *married* (avec les écarts-types estimés entre parenthèses) est de 0,053 (0,041), et le coefficient associé à *female* devient – 0,290 (0,036). De fait, la « prime » associée au mariage est estimée à environ 5,3 %, mais n'apparaît pas significativement différentes de zéro (t = 1,29). Une limite importante du modèle est que la prime de mariage est supposée être la même pour les hommes et les femmes, hypothèse qui sera assouplie dans l'exemple qui suit.

EXEMPLE 7.6
Équation du log salaire horaire

Estimons maintenant un modèle dans lequel on distingue les différences de salaires pour quatre groupes différents soient : les hommes mariés, les femmes mariées, les hommes non mariés et les femmes non mariées. Pour ce faire, nous devons identifier un groupe de référence ; nous choisissons celui des hommes non mariés. Puis, nous devons définir des variables indicatrices pour chacun des groupes restants. Nous les appelons *marrmale*, *marrfem*, et *singfem*. Nous introduisons ces trois variables dans l'équation (7.9) (mais pas *female*, puisque cela serait redondant), son estimation donne :

$$\widehat{\log(wage)} = 0,321 + 0,213\, marrmale - 0,198\, marrfem$$
$$(0,100)\ (0,055)\qquad\qquad (0,58)$$

$$\begin{aligned} &- 0{,}110\ singfem + 0{,}079\ educ + 0{,}027\ exper - 0{,}00054\ exper^2 \\ &\quad (0{,}056) \qquad (0{,}007) \qquad (0{,}005) \qquad (0{,}00011) \\ &+ 0{,}29\ tenure - 0{,}00053\ tenure^2 \\ &\quad (0{,}007) \qquad (0{,}00023) \\ &n = 526,\ R^2 = 0{,}461 \end{aligned} \qquad \textbf{[7.11]}$$

Tous les coefficients à l'exception de *singfem*, présentent des statistiques *t* bien supérieures à deux en valeur absolue. La statistique *t* associée à *singfem* est quant à elle d'environ – 1,96, ce qui est à peine significatif au seuil de 5 % contre l'alternative de test bilatérale.

Pour interpréter les coefficients associés aux variables indicatrices nous devons nous rappeler que le groupe de référence est celui des hommes non mariés. De ce fait, les estimations relatives aux trois variables indicatrices mesurent les taux de variations de salaires *relativement* à ceux des hommes non mariés. Par exemple, il ressort que les hommes mariés touchent un salaire d'environ 21,3 % de plus que les hommes non mariés, à niveaux d'éducation, expérience, et de titularisation fixés. [Une estimation plus précise de (7.10) est d'environ 23,7 %.] Une femme mariée quant à elle touche un salaire prédit de 19,8 % moindre qu'un homme non marié, à caractéristiques équivalentes.

Du fait que le groupe de référence choisi est représenté par la constante dans (7.11), nous n'avons inclus que trois variables indicatrices pour quatre groupes. Si nous avions ajouté une variable indicatrice pour les hommes non mariés dans l'équation (7.11), nous aurions rencontré un problème de « trappe à variables indicatrices » dû à la présence de colinéarité parfaite entre les variables du modèle. Certains logiciels corrigent automatiquement ce problème, d'autres vous le signalent au moyen d'un message d'erreur mentionnant la colinéarité parfaite de certaines de vos variables. Il est préférable de spécifier correctement votre modèle et vos variables indicatrices de façon à correctement interpréter vos estimations.

Alors même que les hommes non mariés constituent le groupe de référence dans (7.11), nous pouvons utiliser cette équation pour calculer les différences de salaires estimées entre chacune des paires. À noter que dans la mesure où la constante du modèle est identique pour chacun des groupes, nous pouvons l'ignorer pour le calcul des différences. De fait, les taux de variation de salaire estimés entre le groupe de femmes mariées et non mariées est donné par – 0,110 – (– 0,198) = 0,088, ce qui implique que les femmes non mariées gagnent en moyenne environ 8,8 % de plus que les femmes mariées. Malheureusement, nous ne pouvons nous référer à l'équation (7.11) pour tester si cette différence est statistiquement significative. Connaissant les écarts-types associés aux variables *marrfem* et *singfem* n'est en effet pas suffisant pour mettre en œuvre le test (voir section 4.4). Le plus simple consiste à choisir un de ces groupes comme le groupe de référence et de ré-estimer l'équation. Rien de substantiel donc mais cela permet d'obtenir directement les écarts-types estimés requis pour la mise en œuvre du test. Lorsque nous choisissons le groupe des femmes mariées comme groupe de référence, nous obtenons les résultats suivants :

$$\begin{aligned} \widehat{\log(wage)} &= 0{,}123 + 0{,}411\ marrmale + 0{,}198\ singmale + 0{,}088\ singfem + \ldots, \\ &\quad (0{,}106)\ (0{,}056) \qquad (0{,}58) \qquad (0{,}52) \end{aligned}$$

où bien évidemment aucun des coefficient et écart-type non reportés n'ont changé. L'estimation du coefficient associé à *singfem* est comme attendu de 0,088. Nous disposons donc maintenant de l'écart-type requis pour mettre en place le test précédent. La statistique *t* calculée sous l'hypothèse nulle d'absence de différence de salaire entre les populations des femmes mariées et non mariées est de $t_{singfem} = 0{,}088/0{,}052 \approx 1{,}69$. Ce résultat constitue une preuve marginale à l'encontre de l'hypothèse nulle. Nous notons en outre que la différence entre les hommes et les femmes mariés est quant à elle très significative ($t_{marrmale} = 7{,}34$).

L'exemple précédent illustre un principe général valant pour l'introduction de variables indicatrices relatives à différents groupes : si le modèle de régression contient différentes variables indicatrices associées à *g* groupes ou catégories, il convient d'introduire *g* – 1 variables indicatrices dans le modèle en plus de la

constante. La constante du modèle capture alors l'effet moyen pour le groupe de référence, et le coefficient associé à un groupe particulier représente la différence estimée entre les effets moyens pour ce groupe et le groupe de référence. Inclure g variables indicatrices en plus de la constante aura pour conséquence d'introduire de la multicolinéarité parfaite. Une alternative serait d'introduire g variables indicatrices et d'exclure la constante du modèle. Inclure g indicatrices sans constante est parfois utile mais présente au moins deux désavantages. En premier lieu, il devient plus complexe de tester les différences relativement au groupe de référence. Par ailleurs, les logiciels changent en règle générale la manière dont le *R-carré* est calculé lorsqu'une constante n'est pas incluse dans le modèle de régression. En particulier, dans la formule du $R^2 = 1 - SCR/SCT$, la somme des carrés totaux, *SCT*, est remplacée par une somme des carrés non centrée en la moyenne des y_i, mettons, $SCT_0 = \sum_{i=1}^{n} y_i^2$. Le *R-carré* résultant de ce calcul, notons le $R_0^2 = 1 - SCR/SCT_0$, est parfois appelé le ***R-carré* non centré**. Malheureusement, R_0^2 n'est pas une bonne mesure de l'asjutement du modèle. S'il est toujours vrai que $SCT_0 \geq SCT$ l'égalité ne tenant que lorsque $\bar{y} = 1$, souvent, SCT_0 est bien plus élevé que *SCT*, ce qui implique que R_0^2 est bien plus élevé que R^2. À titre d'exemple, si dans les illustrations précédentes, nous avions régressé log(*wage*) sur *marrmale*, *singmale*, *marrfem*, *singfem*, et les autres variables explicatives sans la constante, le *R-carré* reporté par Stata, soit R_0^2, eût été 0,948. Ce *R-carré* très élevé est un artéfact de calcul dû à l'absence de centrage de la somme des carrés totaux. La vraie valeur du *R-carré* est donnée dans l'équation (7.11) comme étant 0,461. Un certain nombre de logiciels, à l'instar de Stata, proposent une option pour forcer le calcul du *R-carré* centré quand bien même une constante n'a pas été introduite dans le modèle ; recourir à cette option est en général conseillé. Dans la grande majorité des cas, tous les *R-carrés* dérivés à partir de la comparaison des *SCR* et *SCT* devraient avoir des *SCT* calculés sur base des y_i centrés autour de $\bar{y}$. Nous pouvons interpréter *SCT* comme étant la somme des carrés des résidus qui serait obtenue si nous utilisions la moyenne simple pour prédire chaque observation y_i. Notons que notre ambition est alors très faible puisque nous mesurons la performance prédictive du modèle relativement à celle d'une constante. Dans le cas d'un modèle sans constante dont la performance serait médiocre, il est possible que $SCR > SCT$, ce qui impliquerait un R^2 négatif. Le *R-carré* non centré sera toujours compris entre zéro et un, ce qui explique pourquoi il est habituellement choisi par défaut par la plupart des logiciels lorsqu'une constante n'est pas introduite dans le modèle.

Pour aller plus loin 7.2

Dans les données relatives aux salaires des joueurs professionnels de baseball exposées dans MLB1, les joueurs peuvent occuper l'un des six postes suivants : *frstbase* (première base), *scndbase* (deuxième base), *thrdbase* (troisième base), *shrtstop* (arrêt court), *outfield* (joueur de champ extérieur), ou *catcher* (receveur). Pour mesurer les différences de salaires selon les postes occupés, et en définissant les joueurs de champ extérieur comme le groupe de référence, quelles variables indicatrices convient-il d'introduire dans le modèle ?

Introduire de l'information ordinale via les variables indicatrices

Nous souhaitons maintenant estimer l'effet des notations (ou *ratings*) de crédits des municipalités sur le taux d'intérêt des obligations municipales – soient les obligations émises par ces dernières. Plusieurs compagnies financières telles que *Moody's Investors Service* et *Standard and Poor's*, évaluent la qualité de la dette pour les autorités locales au travers de notations qui dépendent de critères telles que la probabilité de défaut. (Les autorités locales préfèrent des taux d'intérêt bas dans le but de réduire le coût d'emprunt). Supposons pour simplifier que ces notes s'étendent sur une échelle de zéro à quatre, zéro étant la note la moins bonne et quatre la meilleure. Ceci est une exemple de **variable ordinale.** Dénommons cette variable *CR* (*credit ratings*). La question que l'on se pose ici est de savoir comment incorporer la variable *CR* dans le modèle pour expliquer *MBR* (*municipality bond rate*) ?

Une première possibilité serait d'inclure directement *CR* dans le modèle à l'instar des autres variables explicatives :

$$MBR = \beta_0 + \beta_1 CR + autres\ facteurs$$

où nous n'explicitons pas ici quels sont les autres facteurs du modèle. Par la suite, β_1 s'interprète comme la variation en pourcentage de *MBR* lorsque *CR* augmente d'une unité, toutes choses égales par ailleurs. Malheureusement, il s'avère difficile d'interpréter l'augmentation d'une unité de *CR*. Si nous pouvons interpréter facilement une année additionnelle de formation ou un dollar supplémentaire dépensé par étudiant, les notions telles que les notations ont une interprétation purement ordinale. Nous savons qu'un *CR* de quatre est meilleur d'un *CR* de trois, mais peut-on dire que la différence entre trois et quatre équivaut celle entre zéro et un ? Si cela n'est pas le cas, il est alors incorrect de supposer que l'augmentation d'une unité de *CR* a un effet constant sur *MBR*.

Une meilleure approche que nous pouvons mettre en oeuvre en raison du peu de valeurs possibles prises par *CR* serait de définir des variables indicatrices pour chaque réalisation potentielle de *CR*. Posons alors $CR_1 = 1$ si $CR = 1$, et $CR_1 = 0$ sinon ; $CR_2 = 1$ si $CR = 2$, et $CR_2 = 0$ sinon ; et ainsi de suite. De fait, nous considérons ici les cinq catégories de notations et les transformons en variables. Nous pouvons alors estimer le modèle suivant :

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + autres\ facteurs \qquad [7.12]$$

Selon notre règle d'inclusion des variables indicatrices, nous n'introduisons que quatre des variables précédemment définies puisque nous sommes en présence ici de cinq catégories. La catégorie omise ici est le rating de zéro et constitue le groupe de référence de notre étude. (C'est pour cela que nous n'avons pas besoin de définir une variable indicatrice pour cette catégorie). Les coefficients sont faciles à interpréter : δ_1 correspond à la différence de *MBR* (toutes choses égales par ailleurs) entre une commune présentant une notation de 1 et une autre associée à un rating nulle ; δ_2 à la différence de *MBR* entre une commune notée deux et une autre à zéro ; etc. L'impact du passage d'une note à une autre n'est ici pas supposé constant. De fait, recourir à l'équation (7.12) permet une plus grande flexibilité que lorsque l'on introduit *CR* directement comme variable explicative. Une fois les variables indicatrices définies, l'estimation de (7.12) est aisée.

EXEMPLE 7.7
Des effets de l'attractivité physique sur le salaire

Hamermesh et Biddle (1994) ont utilisé des mesures de l'attractivité physique pour l'estimation d'équations de salaires. (À noter que le fichier BEAUTY contient un plus petit nombre de variables mais un plus grand nombre d'observations que celui utilisé par Hamermesh et Biddle. Voir exercice pratique C12.) Chaque individu de l'échantillon est classé par l'enquêteur selon son degré d'attractivité physique en cinq catégories (du physique ingrat, ordinaire, dans la moyenne, charmant, à la beauté spectaculaire). Dans la mesure où très peu de personnes correspondaient aux descriptifs des deux catégories extrêmes, les auteurs ont considéré les individus relatifs aux trois catégories intermédiaires pour leurs régressions qu'ils ont requalifiées : beauté moyenne, beauté inférieure à la moyenne, beauté supérieure à la moyenne, avec comme groupe de référence le groupe moyen labellisé *average*. À partir de données du *« Quality of Employment Survey »* datant de 1977 et après avoir introduit un certain nombre de variables de contrôle relatives à la productivité, Hamermesh et Biddle ont estimé l'équation suivante pour le groupe des hommes :

$$\widehat{\log(wage)} = \hat{\beta}_0 - \underset{(0,046)}{0,164}\, belavg + \underset{(0,033)}{0,016}\, abvavg + autres\ facteurs$$

$$n = 700,\ R^2 = 0,403$$

et pour les femmes :

$$\widehat{\log(wage)} = \hat{\beta}_0 - \underset{(0,006)}{0,124\,belavg} + \underset{(0,049)}{0,035\,abvavg} + autres\ facteurs$$
$$n = 409,\ R^2 = 0,330$$

Les autres variables de contrôle incluent les niveaux d'éducation, d'expérience, la titularisation, le statut marital ainsi que les origines ethniques ; on se référera au tableau 3 de l'article de Hamermesh et Biddle pour une liste plus complète. Dans le souci de gagner de l'espace les coefficients associés à ces variables de contrôle ne sont pas reportés dans le corps du texte de l'article, de même que celui de la constante du modèle.

Concernant les hommes, on estime que ceux dont la beauté est jugée inférieure à la moyenne gagnent en moyenne 16,4 % de moins qu'un homme d'attractivité moyenne possédant les même caractéristiques (notamment en matière d'éducation, d'expérience, de titularisation, de statut marital et d'origines ethniques). L'effet apparaît statistiquement significatif, avec $t = -3,57$. À l'inverse, les résultats d'estimation suggèrent que les hommes présentant une attractivité physique jugée supérieure à la moyenne gagnent en moyenne 1,6 % plus, mais l'effet n'apparaît pas ici statistiquement significatif ($t < 0,5$).

Les femmes présentant un physique jugé inférieur à la moyenne gagnent quant à elle en moyenne 12,4 % de moins que leurs collègues féminines jugées dans la moyenne, avec $t = -1,88$. Comme cela était le cas pour les hommes, l'effet estimé pour *abvavg* n'apparaît pas ici statistiquement significatif.

Dans une étude similaire, Biddle et Hamermesh (1998) revisitent les effets de l'apparence sur les bénéfices en ayant recours à un groupe plus homogène à savoir les diplômés d'une faculté de droit particulière. Les auteurs confirment l'impact de l'attractivité physique sur les bénéfices annuels, un résultat qui paraîtra sans doute peu surprenant pour les personnes pratiquant le droit.

Pour aller plus loin 7.3

Dans le modèle (7.12), comment procéderiez-vous pour tester l'hypothèse nulle d'absence d'impact de la notation de crédit sur *MBR* ?

L'équation (7.12) inclut le cas particulier des effets partiels constants. Une manière d'écrire les trois restrictions impliquant la constance des effets partiels est de poser : $\delta_2 = \delta_1$, $\delta_3 = 2\delta_1$, $\delta_3 = 3$, δ_1 et $\delta_4 = 4\delta_1$. Lorsque l'on introduit ces équations dans (7.12) il suit : $MBR = \beta_0 + \delta_1 (CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + autres\ facteurs$. Le paramètre δ_1 est bien associé à la variable *CR* telle qu'initialement introduite. Afin de calculer la statistique de Fischer *F* associée au test de constance des effets partiels, c'est-à-dire à la validation des restrictions posées plus haut, nous estimons les modèles contraint d'une part, correspondant à la régression de *MBR* sur *CR* ainsi que les autres variables de contrôle, et non-contraint (7.12) d'autre part puis, récupérons les *R-carrés* associés. La statistique *F* est ainsi obtenue comme dans l'équation (4.1) avec $q = 3$.

Dans certains cas, la variable ordinale prend un trop grand nombre de valeurs possibles, de sorte qu'il n'est pas possible d'inclure une variable indicatrice par catégorie. À titre d'exemple, le fichier LAWSCH85 contient des données sur les salaires médians à l'entrée des jeunes diplômés des écoles de commerce[1] sur le marché du travail. Une des variables explicatives clé est le classement de l'école. Comme chaque école dispose d'un classement différent, nous ne pouvons décemment pas introduire autant de variables indica-

1 NdT : Nous avons délibérément traduit ici « *law school* » par « école de commerce », pour nous rapprocher au mieux de la problématique traitée dans l'exemple, à savoir, l'impact du classement des écoles sur les salaires à l'embauche des jeunes diplômés.

trices que de réalisations possibles de la variable initiale. Si nous ne souhaitons pas introduire directement la variable *rank* dans l'équation, nous pouvons regrouper différentes sous-catégories comme le suggère l'exemple ci-dessous.

EXEMPLE 7.8
Des effets du classement des écoles de commerce sur les salaires à l'entrée dans la vie active

Définissons les variables indicatrices *top10*, *r11_25*, *r26_40*, *r41_60*, *r61_100* comme prenant la valeur unitaire lorsque la variable *rank* s'étend entre les bornes correspondantes. Nous choisissons les écoles classées en deçà des 100 premières comme le groupe de référence. Les résultats d'estimation sont alors les suivants :

$$\begin{aligned}\widehat{\log(salary)} &= \underset{(0,41)}{9,17} + \underset{(0,053)}{0,700}\,top10 + \underset{(0,039)}{0,594}\,r11\_25 + \underset{(0,034)}{0,375}\,r26\_40 \\ &\quad + \underset{(0,028)}{0,263}\,r41\_60 + \underset{(0,021)}{0,132}\,r61\_100 + \underset{(0,0031)}{0,0057}\,LSAT \\ &\quad + \underset{(0,074)}{0,014}\,GPA + \underset{(0,026)}{0,036}\log(libvol) + \underset{(0,0251)}{0,0008}\log(cost) \\ n &= 136,\ R^2 = 0,911,\ \bar{R}^2 = 0,905\end{aligned} \qquad [7.13]$$

Il apparaît très clairement que les variables indicatrices associées aux différents groupes décrits précédemment sont toutes très significatives. L'estimation du coefficient associé à *r61_100* signifie qu'à niveaux de *LSAT*, *GPA*, *libvol*, et *cost* fixés, le salaire médian pour un jeune diplômé issu d'une école de commerce classée de la 61e à la 100e place touchera en moyenne un salaire 13,2 % plus élevé qu'un diplômé issu d'une école classée en deçà des 100 premières. La différence de salaire entre un jeune diplômé issue d'une école classée parmi les 10 premières (*top10*) et une autre classée en deçà des 100 premières est relativement importante. En recourant au calcul de l'effet exact au moyen de l'équation (7.10) nous obtenons $\exp(0,700) - 1 \approx 1,014$, impliquant que le salaire médian est plus de 100 % supérieur pour un jeune diplômé issu de l'une des 10 premières écoles du classement comparativement à celui d'un jeune diplômé issu d'une école classée en deçà des 100 premières.

Pour mesurer l'apport éventuel du regroupement de *rank* en différentes catégories, nous pouvons comparer les *R-carrés* ajustés de l'équation estimée (7.13) et de celle où *rank* serait introduit en une seule variable : le premier est donné par 0,905 alors que le second est de 0,836. Dans ce cas une plus grande flexibilité du modèle est récompensée (7.13).

Il est à noter qu'une fois que la variable *rank* est introduite (reconnaissons le, de façon arbitraire) via différentes catégories, toutes les autres variables deviennent alors non significatives. Le test de significativité jointe des coefficients associés à *LSAT*, *GPA*, log(*libvol*), et log(*cost*) donne une *p*-valeur de 0,055, impliquant le non rejet de l'hypothèse nulle au seuil de 5 % (l'hypothèse n'est rejetée qu'au seuil de 10 %). Lorsque la variable *rank* est introduite dans sa forme originale, la *p*-valeur associée au test de significativité jointe est de 0 à quatre décimales près.

Enfin, il est à noter qu'en dérivant les propriétés de l'estimateur des moindres carrés ordinaires, nous avions fait l'hypothèse que nous avions affaire à un échantillon aléatoire. Or, les applications que nous considérons ici violent cette hypothèse et cela tient à la façon dont la variable *rank* a été construite. En effet, le classement d'une école dépend nécessairement du classement des autres écoles de l'échantillon, de ce fait les données ne peuvent donc pas s'assimiler à des tirages indépendants de la population des écoles de commerce. Pour autant, cela n'a pas d'implication sérieuse, tant que le terme d'erreur n'apparaît pas corrélé avec les variables explicatives.

7.4 VARIABLES D'INTERACTION IMPLIQUANT DES VARIABLES INDICATRICES

À l'instar des variables à interprétation quantitative, les variables indicatrices peuvent être associées à des variables d'interaction dans les modèles de régression. Nous avons déjà étudié un cas dans le cadre de l'exemple 7.6, lors duquel nous avions défini quatre catégories selon le statut marital et le genre. En effet, il est possible de réécrire le modèle en ajoutant un **terme d'interaction** entre les variables *female* et *married* de façon à modéliser les cas où *female* et *married* apparaissent séparément. En procédant de la sorte, nous pouvons mettre en lumière une prime liée au mariage dépendante du genre à l'instar de celle que nous avions estimée dans l'équation (7.11). À des fins de comparaison, le modèle estimé avec le terme d'interaction *female-married* est reproduit ci-dessous :

$$\begin{aligned}\widehat{\log(wage)} &= 0{,}321 - 0{,}110\,female + 0{,}213\,married \\ &\qquad (0{,}100)\ (0{,}056)\qquad (0{,}055) \\ &\quad - 0{,}301\,female\,married + \ldots, \\ &\qquad (0{,}072)\end{aligned} \qquad \textbf{[7.14]}$$

où les autres variables explicatives sont identiques à celles de l'équation (7.11). Les résultats de l'estimation de l'équation (7.14) confirment qu'il existe bien une interaction statistiquement significative entre le genre et le statut marital. Ce modèle permet en outre d'obtenir les différentiels de salaires pour les quatre groupes, en prenant soin de bien reporter correctement les combinaisons de zéro et uns correspondant aux différents cas en présence.

Nous fixons *female* = 0 et *married* = 0 comme le groupe des hommes non mariés, soit le groupe de référence puisque cela permet d'éliminer les variables *female*, *married*, et *female·married*. La constante valant pour les hommes mariés est obtenue en fixant *female* = 0 et *married* = 1 dans l'équation (7.14) ; ce qui donne une valeur de 0,321 + 0,213 = 0,534, et ainsi de suite.

EXEMPLE 7.9
Des effets de l'utilisation de l'outil informatique sur les salaires

Krueger (1993) estime les effets de l'utilisation de l'outil informatique sur les salaires. Il définit pour cela une variable indicatrice que nous nommerons *compwork*, égale à un si l'individu utilise un ordinateur sur son lieu de travail. Une autre variable indicatrice *comphome*, est égale à un si l'individu utilise un ordinateur à son domicile. À partir des données relatives à 13 379 personnes issues du recensement [« *Current Population Survey* »] de 1989, Krueger (1993, tableau 4) obtient les résultats suivants :

$$\begin{aligned}\widehat{\log(wage)} &= \hat{\beta}_0 + 1{,}77\,compwork + 0{,}070\,comphome \\ &\qquad (0{,}009)\qquad\quad 0{,}019) \\ &\quad + 0{,}017\,compwork{\cdot}comphome + autres\ facteurs \\ &\qquad (0{,}23)\end{aligned} \qquad \textbf{[7.15]}$$

(Les autres facteurs sont les régresseurs habituels des équations de salaire tels que les niveaux d'éducation, d'expérience, le genre ainsi que le statut marital ; se référer à l'étude de Krueger pour une liste détaillée.) Krueger ne reporte pas les résultats relatifs à la constante du modèle en raison de son peu d'intérêt ici ; tout ce qui nous importe est de savoir que le groupe de référence consiste en l'ensemble des individus n'utilisant d'ordinateur ni sur le lieu de travail ni dans le cadre privé. Nous notons que le rendement estimé de l'utilisation de l'outil informatique au travail (et non dans le privé) est d'environ 17,7 %. (Une estimation plus précise

est donnée par 19,4 %.) À l'inverse, les personnes utilisant des ordinateurs chez eux mais pas sur leur lieu de travail obtiennent une prime salariale estimée d'environ 7 % comparativement à ceux n'utilisant pas du tout d'ordinateur. La différence de salaires entre ceux utilisant un ordinateur à la fois sur leur lieu de travail et dans le privé comparativement à ceux qui n'en ont pas du tout l'usage est d'environ 26,4 % (ce chiffre est obtenu en sommant chacun des trois coefficients puis en multipliant le résultat par 100), soit au moyen d'une estimation plus fine à partir de l'équation (7.10), de l'ordre de 30,2 %.

Le terme d'interaction dans (7.15), s'il n'apparaît pas statistiquement significatif, ni même pertinent du point de vue économique, n'est pas dommageable et est donc conservé dans le modèle.

L'équation (7.14) est une approche alternative d'identification des différentiels de salaires selon toutes les combinaisons possibles de genre et de statut marital. Il nous est alors possible de tester l'hypothèse nulle que le différentiel de salaire entre hommes et femmes ne dépend pas du statut marital (ou de façon équivalente, que le différentiel de salaires entre individus mariés et non mariés ne dépend pas du genre). L'équation (7.11) est la plus adaptée pour mettre en oeuvre le test entre n'importe lesquels des groupes et celui des hommes non mariés.

Relâcher l'hypothèse d'homogénéité des pentes

Nous avons maintenant passé en revue un ensemble d'exemples relatifs à l'incorporation de variables indicatrices associées à un ou plusieurs groupes d'individus dans les modèles de régression multiple. Nous avons en outre étudié certains cas d'interactions entre variables indicatrices permettant de **relâcher l'hypothèse d'homogénéité des pentes**. Pour poursuivre sur l'exemple des déterminants des salaires, supposons que l'on souhaite tester si le rendements de l'éducation sont les mêmes pour les hommes et les femmes, sous l'hypothèse d'un différentiel de salaire constant entre hommes et femmes (différence dont nous avons empiriquement validé l'existence précédemment). Par soucis de simplicité, nous nous proposons de n'inclure que les déterminants relatifs au niveau d'éducation et au genre dans le modèle. La question qui se pose est de savoir quel type de modèle permettrait de rendre compte de telles différences dans les rendements de l'éducation ? Considérons le modèle suivant :

$$\log(wage) = \beta_0 + \delta_0 (female) + (\beta_1 + \delta_1 female)\, edu + u \qquad \textbf{[7.16]}$$

Si nous fixons $female = 0$ dans (7.16), il apparaît que la constante pour les hommes correspond au paramètre β_0, et la pente associée à la variable *educ* à β_1. Pour les femmes, nous fixons $female = 1$; dès lors la constante du modèle pour les femmes est donnée par $\beta_0 + \delta_0$ et la pente associée à la variable explicative par $\beta_1 + \delta_1$. Il suit que, δ_0 mesure la différence entre les constantes du modèle lorsqu'il est estimé pour les hommes et les femmes et δ_1 la différence en matière de rendements d'éducation entre ces deux mêmes groupes. Deux des quatre cas possibles relativement aux signes de δ_0 et δ_1 sont présentés à la figure 7.2.

Le graphique (a) illustre le cas où la constante du modèle pour les femmes est située sous celle des hommes, alors que la pente de la droite est plus faible pour les femmes que pour les hommes. Ces résultats impliquent que les femmes gagnent en moyenne moins que les hommes quelque soient les niveaux d'éducation, cette différence ayant tendance à croître avec le niveau d'éducation, i.e. lorsque *educ* prend des valeurs plus grandes. Sur le graphique (b), la constante du modèle pour les femmes est située sous celle des hommes, alors qu'à l'inverse, la pente relative au niveau d'éducation apparaît supérieure pour les femmes. Ces résultats impliquent cette fois que les femmes gagnent en moyenne moins que les hommes pour des niveaux d'éducation faibles mais que cette différence a tendance à diminuer avec l'accroissement du niveau d'éducation. Au-delà d'un certain seuil, une femme gagne plus qu'un homme, pour un même niveau d'éducation (cette valeur peut être facilement évaluée à partir des résultats estimés).

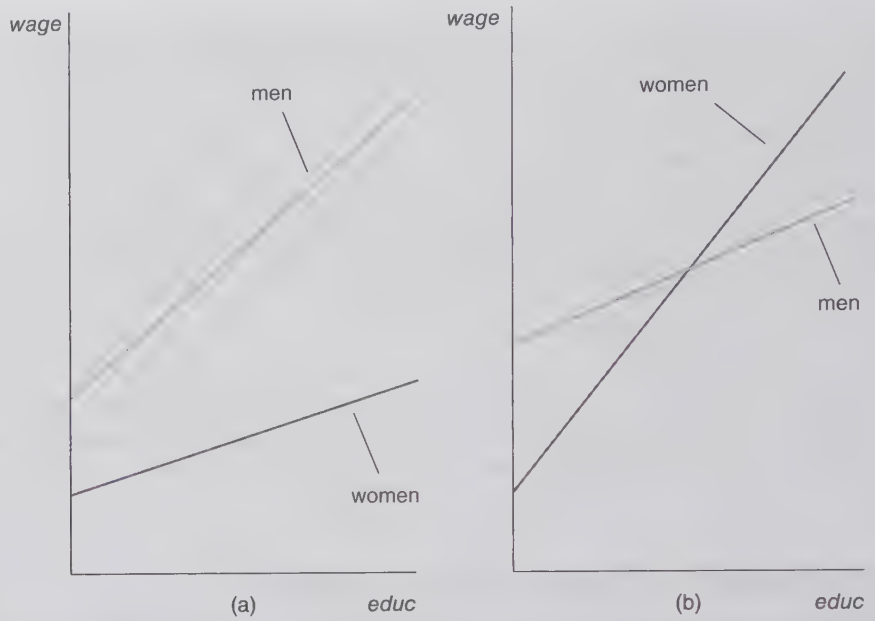


Figure 7.2 *Représentation graphique de l'équation (7.16) : (a) $\delta_0 < 0$, $\delta_1 < 0$; (b) $\delta_0 < 0$, $\delta_1 > 0$.*

Comment peut-on estimer le modèle décrit par l'équation (7.16) ? Avant d'appliquer les MCO, nous devons réécrire le modèle au moyen d'une variable d'interaction entre *female* et *educ* :

$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u \qquad [7.17]$$

Ces paramètres peuvent maintenant être estimés à partir de la régression de log(*wage*) sur *female*, *educ*, et *female·educ*. La construction du terme d'interaction au moyen des logiciels d'économétrie classiques ne pose pas de difficulté majeure. Ne soyez pas surpris par la nature étrange de la variable *female·educ*, qui prend la valeur zéro pour tous les hommes et égale au niveau d'éducation des femmes de notre échantillon.

Une hypothèse clé ici tient à l'existence de rendements d'éducation identiques pour les hommes et les femmes. En termes de modélisation, cela implique de tester à partir de l'équation (7.17), que $H_0 : \delta_1 = 0$ ce qui signifie que la pente de log(*wage*) pour la variable *educ* est la même pour les hommes et les femmes. Notons que cette hypothèse n'implique pas de restrictions sur la différence des constantes δ_0. Une différence de salaire entre les hommes et les femmes est rendue possible sous l'hypothèse nulle, mais doit être identique quelque soit le niveau d'éducation. Cette situation est décrite à la figure 7.1.

EXEMPLE 7.10
L'équation du log salaire horaire

Nous ajoutons des termes quadratiques pour les variables relatives à l'expérience et la titularisation à l'équation (7.17) :

$$\widehat{\log(wage)} = \underset{(0,119)}{0,389} - \underset{(0,168)}{0,227}\, female + \underset{(0,008)}{0,082}\, educ - \underset{(0,0131)}{0,0056}\, female.educ + \underset{(0,005)}{0,029}\, exper - \underset{(0,00011)}{0,00058}\, exper^2 \qquad [7.18]$$

$$+ 0{,}032\ tenure - 0{,}00059\ tenure^2$$
$$(0{,}007) \qquad (0{,}00024)$$
$$n = 526,\ R^2 = 0{,}441$$

Les rendements de l'éducation estimés pour les hommes dans cette équation sont donnés par 0,082, soit 8,2 %. Pour les femmes, ils sont de 0,082 – 0,0056 = 0,0764, soient environ 7,6 %. La différence de – 0,56 % représente à peine plus d'un demi point de pourcentage en moins pour les femmes et n'apparaît donc pas très importante du point de vue économique, ni même statistiquement significative puisque la statistique de Student associée est de $-0{,}0056/0{,}0131 \approx -0{,}43$. Sur base de ces résultats, nous pouvons conclure qu'il n'apparaît pas ici de preuve tangible à l'encontre de l'hypothèse d'égalité des rendements de l'éducation pour les hommes et les femmes.

Les coefficients associés à *female*, alors qu'ils restent économiquement importants, n'apparaissent plus statistiquement significatifs aux seuils conventionnels ($t = -1{,}35$). Les coefficient et statistique t de l'équation sans terme d'interaction étaient de –0,297 et –8,25, respectivement [voir équation (7.9)]. Devrions-nous pour autant conclure à l'absence de preuve statistique en faveur de l'hypothèse d'une discrimination salariale à l'encontre des femmes, pour des niveaux équivalents d'*educ*, *exper*, et *tenure* ? Ceci serait une grave erreur. En effet, nous avons ajouté la variable d'interaction *female·educ* à l'équation, le coefficient associé à *female* est maintenant estimé de façon moins précise qu'il ne l'était dans l'équation (7.9) : l'écart-type associé au coefficient estimé a quasiment été multiplié par cinq ($0{,}168/0{,}036 \approx 4{,}67$) en raison de la très forte corrélation entre *female* et *female·educ* dans l'échantillon. Cet exemple fournit une illustration utile de la multicolinéarité : dans l'équation (7.17) ainsi que l'équation plus générale estimée dans (7.18), δ_0 mesure le différentiel de salaire entre les hommes et les femmes lorsque *educ* = 0. Peu d'individus dans l'échantillon présentent des niveaux très faibles d'éducation, il n'est donc pas surprenant que nous soyons confrontés à de grandes difficultés pour estimer ce différentiel lorsque *educ* = 0 (cette évaluation n'apporte d'ailleurs que peu d'information pertinente). Il serait en revanche plus intéressant d'estimer le différentiel entre les hommes et les femmes pour, mettons, un niveau d'éducation moyen dans l'échantillon (environ 12,5). Pour ce faire, nous pourrions remplacer *female·educ* par *female·*(*educ* – 12,5) et ré-estimer l'équation ; ce qui aura pour seul effet de modifier la valeur estimée du coefficient de *female* et son écart-type estimé associé. (Voir exercice pratique C7.)

Lorsque nous calculons la statistique de Fisher, F, pour $H_0 : \delta_0 = 0,\ \delta_1 = 0$, nous obtenons $F = 34.33$, ce qui représente une valeur relativement élevée pour une variable aléatoire, avec comme degrés de liberté $ddl = 2$ et 518 : la p-valeur associée est de zéro à 10^{-4} près. À l'aune de ces résultats, nous pouvons conclure que le modèle (7.9) est préférable, ce qui confirme l'hypothèse d'un différentiel de salaire constant entre les hommes et les femmes.

Nous nous intéressons également à l'hypothèse selon laquelle la moyenne des salaires est identique pour les hommes et les femmes présentant un même niveau d'éducation. Cela implique que les paramètres δ_0 et δ_1 sont *tous deux* égaux à zéro sous l'hypothèse nulle. Dans l'équation (7.17), nous devons recourir à un test de Fisher de façon à tester $H_0 : \delta_0 = 0,\ \delta_1 = 0$. Dans le modèle avec une constante pour mesurer la différence, nous rejetons cette hypothèse du fait que $H_0 : \delta_0 = 0$ est clairement rejetée au profit de l'alternative $H_1 : \delta_0 < 0$

Un exemple plus complexe d'utilisation des variables d'interaction est donné par l'étude des effets des origines ethniques d'une part et de la composition ethnique de la ville d'autre part sur les salaires des joueurs de la ligue de baseball professionnel.

Pour aller plus loin 7.4

Quelles variables additionnelles introduiriez-vous pour relâcher l'hypothèse de constance des rendements à la titularisation entre les hommes et les femmes dans l'équation (7.18) ?

EXEMPLE 7.11
Des effets des origines ethniques sur les salaires des joueurs de baseball professionnels

À partir des données contenues dans MLB1, nous estimons l'équation suivante pour les 330 principaux joueurs de la ligue majeure de baseball pour lesquels des données relatives à la structure ethnique de la ville sont également disponibles. Les variables *black* et *hispan* désignent des indicatrices binaires pour les joueurs individuels. (Le groupe de référence étant constitué des joueurs de type caucasien.) La variable *percblck* correspond au pourcentage d'africains américains dans la ville de l'équipe alors que *perchisp* correspond au pourcentage d'hispaniques. Les autres variables mesurent la productivité et la longévité des joueurs. Ici nous nous intéressons aux effets des origines ethniques une fois prises en compte un ensemble de variables de contrôle.

En plus des variables *black* et *hispan*, nous ajoutons dans l'équation les interactions *black·percblck* et *hispan·perchisp*. L'équation estimée est alors :

$$\begin{aligned}\widehat{\log(salary)} = {} & 10{,}34 + 0{,}0673\ years + 0{,}0089\ gamesyr \\ & \quad (2{,}18)\ (0{,}0129) \qquad (0{,}0034) \\ & + 0{,}00095\ bavg + 0{,}0146\ hrunsyr + 0{,}0045\ rbisyr \\ & \quad (0{,}00151) \qquad (0{,}0164) \qquad (0{,}0076) \\ & + 0{,}0072\ runsyr + 0{,}0011\ fldperc + 0{,}0075\ allstar \\ & \quad (0{,}0046) \qquad (0{,}0021) \qquad (0{,}0029) \\ & - 0{,}198\ black - 0{,}190\ hispan + 0{,}0125\ black{\cdot}percblck \\ & \quad (0{,}125) \qquad (0{,}153) \qquad (0{,}0050) \\ & + 0{,}0201\ hispan{\cdot}perchisp \\ & \quad (0{,}0098) \\ n = 330,\ & R^2 = 0{,}638.\end{aligned} \qquad \textbf{[7.19]}$$

Dans un premier temps, il convient de tester la significativité des quatre variables associées aux origines ethniques, *black*, *hispan*, *black·percblck*, et *hispan·perchisp*. En estimant sur le même échantillon de 330 joueurs, le modèle sans les variables indicatrices, nous obtenons un *R-carré* de 0,626. Dans la mesure où nous considérons quatre restrictions ici et que $ddl = 330 - 13$ dans le modèle non contraint, la statistique de Fisher est évaluée à 2,63, ce qui correspond à une *p*-valeur de 0,034. Les quatre variables sont donc conjointement significatives au seuil de 5 % (mais pas au seuil de 1 %).

Comment pouvons nous interpréter les coefficients estimés des variables relatives aux origines ethniques ? Dans la discussion qui suit, tous les facteurs de productivité sont supposés constants. Dans un premier temps, nous étudions la situation des joueurs africains américains, à composition ethnique au sein de la ville (*perchisp*) fixée. Le coefficient de −0,198 associé à *black* implique que si un joueur est africain américain dans une ville qui n'en compte aucun, (*percblck* = 0), alors il gagne environ 19,8 % de moins qu'un joueur de type caucasien aux qualités comparables. À mesure que *percblck* augmente – ce qui signifie que la population de type caucasien décroît proportionnellement puisque *perchisp* est maintenu constante – le salaire des joueurs africains américains augmente relativement à ceux de type caucasien. Dans une ville contenant 10 % de personnes africaines américaines, la valeur de log(*salary*) pour les africains américains comparativement à celle des joueurs de type caucasien est donnée par $-0{,}198 + 0{,}0125(10) = -0{,}073$, le salaire des joueurs africains américains est donc évalué à environ 7,3 % de moins que pour ceux de type caucasien dans une telle ville. Lorsque *percblck* = 20, les africains américains gagnent environ 5,2 % de plus que les joueurs de type caucasien. Notons que la proportion de personnes africaines américaines la plus élevée dans l'échantillon est de 74 % (Détroit).

De façon similaire, notre raisonnement peut être appliqué dans les villes où la communauté hispanique est faiblement représentée. Nous pouvons facilement identifier la valeur de *perchisp* qui réduit à zéro le différentiel de salaire entre communautés hispaniques et celle des personnes de type caucasien : elle est solution de

l'équation $-0{,}190 + 0{,}0201\ perchisp = 0$, ce qui correspond à $perchisp \approx 9{,}45$. Pour les villes dans lesquelles le taux d'hispaniques dans la population est inférieure à 9,45 %, il est attendu que ceux-ci touchent un salaire inférieur aux joueurs de type caucasien (pour une population de personnes d'origine africaine américaine donnée), l'inverse est vrai lorsque cette proportion dépasse le seuil de 9,45 %. Douze des 22 villes représentées dans l'échantillon comprennent des communautés hispaniques représentant moins de 9,45 % de la population totale. Le pourcentage le plus élevé est d'environ 31 %.

Comment pouvons-nous interpréter ces résultats ? Il serait réducteur d'affirmer qu'il existe de la discrimination à l'encontre des africains américains et des hispaniques puisque nos estimations mettent en évidence que les joueurs de type caucasien gagnent moins que les africains américains et hispaniques dans les villes où ces minorités sont très fortement représentées dans les populations associées. L'importance de la composition ethnique de la ville sur les salaires pourrait être dus aux préférences des joueurs : peut-être les meilleurs joueurs africains américains vivent-ils de façon disproportionnée dans des villes où leur communauté est dominante, même chose pour les latino-américains. Les estimations en (7.19) nous permette de mettre à jour certains mécanismes mais ne nous autorisent pas à trancher entre ces deux hypothèses.

Tester les différences de spécifications entre groupes

Les exemples précédents illustrent l'intérêt des variables d'interaction lorsqu'elles sont introduites conjointement aux variables indépendantes du modèle. Il arrive parfois que l'on veuille tester l'hypothèse nulle selon laquelle deux populations ou groupes suivent une même spécification homogène, contre l'alternative qu'une ou plusieurs pentes diffèrent. Nous étudierons également des exemples dans le chapitre 13 dans lequel nous abordons les données de panel.

Supposons que nous voulions tester si un même modèle de régression permet de décrire la moyenne des résultats universitaires obtenus par des étudiants en premier cycle, sportifs de haut niveau, homme ou femme. L'équation est donnée par :

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

avec *sat* correspondant au score SAT[2], *hsperc* le quantile dans lequel se classe le lycée dont est issu l'étudiant, et *tothrs* le nombre total d'heures de cours à l'université. Nous savons que pour permettre une éventuelle différence de niveau moyen (constante) entre les hommes et les femmes de notre échantillon, nous pouvons inclure une variable indicatrice. Par ailleurs, si nous souhaitons que l'un des coefficients associés aux variables explicatives du modèle dépende du genre, nous devons croiser la variable correspondante avec, par exemple, *female*, et l'inclure dans l'équation.

Si nous souhaitons tester l'existence d'une différence *quelconque* entre les hommes et les femmes nous devons assouplir le modèle de façon à ce que chacun des coefficients associés aux variables explicatives ainsi que la constante, diffèrent entre groupes :

$$\begin{aligned} cumgpa = {} & \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female{\cdot}sat + \beta_2 hsperc \\ & + \delta_2 female{\cdot}hsperc + \beta_3 tothrs + \delta_3 female{\cdot}tothrs + u \end{aligned} \qquad \textbf{[7.20]}$$

Le paramètre δ_0 correspond à la différence entre les constantes valant pour les hommes et les femmes, δ_1 correspond à la différence entre les coefficients associés à la variable *sat* valant pour les hommes et les femmes, etc. L'hypothèse nulle que *cumgpa* suit le même modèle pour les hommes et les femmes s'écrit alors comme suit :

$$H_0 : \delta_0 = 0,\ \delta_1 = 0,\ \delta_2 = 0,\ \delta_3 = 0 \qquad \textbf{[7.21]}$$

2 NdT : Le SAT pour « *Scholastic Aptitude Test* » est un examen standardisé utilisé sur un base nationale pour l'admission à l'université aux États-Unis. Source : Wikipédia (consulté en septembre, 2014).

Si l'un des paramètres δ_j apparaît significativement différent de zéro, alors nous rejetons l'hypothèse d'homogénéité de la spécification valant pour les deux groupes, hommes et femmes.

À partir des données relatives aux résultats du 2nd semestre contenues dans le fichier GPA3, l'estimation du modèle complet donne les résultats suivants :

$$\begin{aligned}\widehat{cumgpa} = {} & 1{,}48 - 0{,}353\ female + 0{,}0011\ sat + 0{,}00075\ female.sat \\ & \quad (0{,}21)\ (0{,}411) \qquad (0{,}0002) \quad (0{,}00039) \\ & - 0{,}0085\ hsperc - 0{,}00055\ female{\cdot}hsperc + 0{,}0023\ tothrs \\ & \quad (0{,}0014) \qquad (0{,}00316) \qquad (0{,}0009) \\ & - 0{,}00012\ female{\cdot}tothrs \\ & \quad (0{,}00163) \\ n = {} & 366,\ R^2 = 0{,}406,\ \overline{R}^2 = 0{,}394.\end{aligned} \qquad \textbf{[7.22]}$$

Il est à noter qu'aucun des quatre coefficients impliquant la variable indicatrice *female* n'apparaît significatif ; seul le coefficient associé à la variable d'interaction *female* · *sat* présente une statistique *t* proche de deux. Mais dans le but de tester la nullité jointe des coefficients à l'instar de (7.21), il convient plutôt de calculer la statistique de Fischer, *F*, pour laquelle nous devons estimer le modèle contraint obtenu en retirant *female* ainsi que toutes les variables d'interaction précédemment introduites ; ce qui donne un R^2 (contraint) d'environ 0,352, soit une statistique *F* d'environ 8,14 ; la *p*-valeur est alors de zéro à 10^{-5} près, ce qui nous invite à rejeter très clairement l'hypothèse nulle jointe (7.21). Il ressort de ces résultats que les athlètes hommes ou femmes sont gouvernés par des modèles de GPA distincts, alors même que chacun des termes de l'équation (7.22) qui permettaient la différenciation des effets entres hommes et femmes lorsqu'ils étaient considérés un à un n'apparaissait pas significatif au seuil de 5 %.

Les écarts-types associés aux coefficients de la variable *female* ainsi que ceux des termes d'interaction apparaissent très élevés et nous invitent donc à la plus grande prudence quant à l'interprétation de l'équation (7.22) puisque les différences entre hommes et femmes dépendent toutes des termes d'interaction. En nous concentrant sur la variable *female*, nous pourrions conclure à tort que *cumgpa* est environ 0,353 moindre pour les femmes que pour les hommes, toutes choses égales par ailleurs. Cette valeur correspond à la différence estimée lorsque *sat*, *hsperc*, et *tothrs* sont fixés à zéro, ce qui paraît peu plausible. Pour *sat* = 1100, *hsperc* = 10, et *tothrs* = 50, la différence de GPA prédite entre un homme et une femme est alors de $-0{,}353 + 0{,}00075(1100) - 0{,}00055(10) - 0{,}00012(50) \approx 0{,}461$. De fait, l'athlète féminine de cet exemple présente un niveau de GPA prédit d'environ un point et demi plus élevé que son homologue masculin présentant des caractéristiques similaires.

Dans un modèle à trois variables, *sat*, *hsperc*, et *tothrs*, il est assez simple d'ajouter toutes les interactions possibles pour tester d'éventuelles différences entre groupes. De ce fait, de nombreuses variables explicatives peuvent être impliquées et il s'avère alors utile de recourir à une autre méthode pour le calcul de la statistique de test. En effet, la statistique *F* calculée à partir des sommes des carrés des résidus issus des différents modèles peut être facilement évaluée même lorsque de nombreuses variables indépendantes sont impliquées.

Dans le modèle général avec *k* variables explicatives et une constante, on suppose que l'on distingue deux groupes ; soient $g = 1$ et $g = 2$. Nous souhaiterions tester si les constantes et pentes sont identiques pour les deux groupes. Nous écrivons alors le modèle comme suit :

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + \ldots + \beta_{g,k}x_k + u, \qquad \textbf{[7.23]}$$

pour $g = 1$ et $g = 2$. L'hypothèse que chaque beta dans (7.23) est identique pour les deux groupes implique $k + 1$ restrictions (dans l'exemple relatif au test GPA, $k + 1 = 4$). Le modèle non contraint proposé présente une variable indicatrice faisant référence à l'un des deux groupes, ainsi que *k* termes d'interaction en plus

de la constante et des variables seules, soit $n - 2(k + 1)$ degrés de liberté. [Dans l'exemple du test GPA, $n - 2(k + 1) = 366 - 2(4) = 358$.] Jusqu'alors rien de nouveau. L'élément clé ici est que la somme des carrés des résidus du modèle non contraint peut être obtenu à l'aide de deux régressions *séparées*, à raison d'une par groupe. Soit SCR_1 la somme des carrés des résidus obtenus à l'issue de l'estimation de (7.23) pour le premier groupe ; impliquant n_1 observations. Soit SCR_2 la somme des carrés des résidus obtenus à l'issue de l'estimation du même modèle pour le second groupe (n_2 observations). Dans l'exemple précédent, si le groupe 1 correspond aux femmes, alors $n_1 = 90$ et $n_2 = 276$. À présent, la somme des carrés des résidus pour le modèle non contraint est simplement donné par $SCR_{nc} = SCR_1 + SCR_2$. La somme des carrés des résidus pour le modèle contraint quant à elle correspond simplement à la SCR issues de la concaténation des groupes et l'estimation d'une seule régression, soit SCR_p. Une fois que nous disposons de ces deux éléments, nous pouvons calculer la statistique F comme suit :

$$F = \frac{[SCR_p - (SCR_1 + SCR_2)]}{SCR_1 + SCR_2} \cdot \frac{[n - 2(k + 1)]}{k + 1} \qquad \textbf{[7.24]}$$

avec n le nombre *total* d'observations. Cette statistique F particulière est généralement dénommée **statistique de Chow** en économétrie. Il est à noter que du fait que le test de Chow s'assimile à un test de Fisher, il n'est valide qu'en présence d'erreurs homoscédastiques et en particulier, sous l'hypothèse que les variances des erreurs pour les deux groupes sont égales. Comme précédemment, la normalité n'est pas requise pour l'analyse asymptotique.

Pour appliquer le test de Chow à l'exemple précédent du test GPA, nous devons connaître la SCR issue de la régression sur l'échantillon complet provenant du regroupement des hommes et des femmes soit $SCR_p = 85{,}515$. La SCR valant pour le groupe des 90 femmes de notre échantillon est donnée par $SCR_1 = 19{,}603$, et celle valant pour les hommes $SCR_2 = 58{,}752$. Dès lors, $SCR_{nc} = 19{,}603 + 58{,}752 = 78{,}355$. La statistique F est alors de $[(85{,}515 - 78{,}355)/78{,}355](358/4) \approx 8{,}18]$; et bien évidemment, sujette à des erreurs d'arrondis, c'est pourquoi nous continuons à utiliser la statistique calculée à partir des *R-carrés* des modèles contraint et non contraint (avec variables d'interaction). (Quelques précisions : il n'existe pas de calcul simple de la statistique à partir des *R-carrés* pour le test si des régressions séparées sont réalisées pour chacun des groupes ; la forme *R-carré* du test ne peut être utilisée que dans le cas où des variables d'interaction ont été introduites pour créer le modèle non contraint.)

Une limite importante du test de Chow traditionnel, quelque soit la méthode retenue pour l'implémentation de sa statistique, tient au fait que sous l'hypothèse nulle aucune différence n'est autorisée entre les groupes. Dans de nombreux cas, il apparaît plus intéressant d'autoriser les constantes à différer entre les groupes et de tester ensuite l'existence de différences entre les pentes à l'instar de ce que nous avions réalisé pour l'équation de salaire dans l'exemple 7.10. Il existe deux approches permettant d'autoriser les constantes à différer sous l'hypothèse nulle. Une première possibilité consiste à introduire une variable indicatrice ainsi que l'ensemble des termes d'interaction, comme dans l'équation (7.22), et de procéder au test de significativité jointe des seuls termes d'interaction. La seconde approche, qui aboutira à la même valeur de la statistique, consiste à calculer une statistique F à partir des SCR comme dans l'équation (7.24), avec cette fois la SCR du modèle contraint, appelée « SCR_p » dans l'équation (7.24), qui sera obtenue à partir d'une régression contenant seulement des constantes hétérogènes. Comme nous testons k restrictions, plutôt que $k + 1$, la statistique F devient :

$$F = \frac{[SCR_p - (SCR_1 + SCR_2)]}{SCR_1 + SCR_2} \cdot \frac{[n - 2(k + 1)]}{k}$$

En appliquant cette démarche à l'exemple du test GPA, nous otenons SCR_p à partir de la régression de *cumgpa* sur *female*, *sat*, *hsperc*, et *tothrs* en utilisant les données complètes, pour les athlètes hommes et femmes.

Dans la mesure où il y a relativement peu de variables explicatives dans cet exemple, il est aisé d'estimer (7.20) et de tester $H_0 : \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$ (avec δ_0 non contraint sous l'hypothèse nulle). La statistique F pour les trois contraintes d'exclusion donne une p-valeur égale à 0,205, on ne rejette donc pas l'hypothèse nulle, même au seuil de significativité de 20 %.

L'échec du rejet de l'hypothèse nulle impliquant la non significativité des termes d'interaction suggère que le meilleur modèle est celui autorisant l'hétérogénéité des seules constantes :

$$\begin{aligned}\widehat{cumgpa} &= 1,39 + 1,310\,female + 10,0012\,sat - 0,0084\,hsperc \\ &\quad (0,18)\ (0,059) \qquad\quad (0,0002) \qquad (0,0012) \\ &\quad + 0,0025\,tothrs \\ &\quad (0,0007) \\ n &= 5\,366,\ R^2 = 0,398,\ \bar{R}^2 = 0,392.\end{aligned} \qquad \textbf{[7.25]}$$

Les coefficients associés aux variables explicatives dans (7.25) sont proches de ceux estimés pour le groupe de référence (celui des hommes) dans (7.22) ; retirer les variables d'interaction entraîne peu de changement. Ceci étant, la variable indicatrice *female* dans (7.25) apparaît hautement significative puisque sa statistique de Student est de plus de 5. Les estimations indiquent par ailleurs que pour des niveaux données de *sat*, *hsperc*, et *tothrs*, une athlète féminine présente un niveau prédit pour le test GPA de 0,31 point plus élevé qu'un athlète homme, ce qui en pratique, représente une différence substantielle.

7.5 LE CAS DES VARIABLES BINAIRES DÉPENDANTES : LE MODÈLE À PROBABILITÉS LINÉAIRES

Jusqu'à présent nous avons passé en revue les propriétés et conditions d'application d'un certain nombre de modèles de régression linéaire. Dans les dernières sections de ce chapitre, nous avons étudié comment, au travers l'utilisation de variables binaires indépendantes, nous pouvions introduire de l'information qualitative dans le modèle de régression multiple. Dans tous ces modèles, la variable dépendante y avait une interprétation *quantitative* (par exemple, y est exprimé en dollars, ou correspond à un score obtenu à un test, ou un pourcentage ou ces mêmes grandeurs prises en logarithmes). La question que l'on se pose maintenant est de savoir ce qu'il se passe lorsque l'on utilise le modèle de régression linéaire multiple pour *expliquer* un événement qualitatif.

Parmi les cas les plus simples, le plus courant d'entre eux en pratique consiste à expliquer un événement binaire. En d'autres termes, notre variable dépendante y, ne prend que deux valeurs possibles : zéro ou un. À titre d'exemple, y peut indiquer si un adulte a suivi des études secondaires au lycée ou non ; si un étudiant de premier cycle a eu recours à des stupéfiants durant l'une de ses années d'études en particulier ; ou encore si une entreprise a été rachetée par une autre durant l'année. Dans chacun de ces exemples, nous pouvons définir $y = 1$ comme la réalisation d'un des deux événements et $y = 0$ l'événement alternatif.

Quel sens donne-t-on au modèle de régression linéaire

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \qquad \textbf{[7.26]}$$

lorsque y est une variable binaire ? Dans la mesure où y ne peut prendre que deux valeurs, β_j ne peut être interprété comme le changement de y consécutif à l'accroissement d'une unité de x_j, toutes choses égales par ailleurs car soit y change de zéro à un soit de un à zéro (ou ne change pas). Pour autant, les β_j conservent de leur intérêt. Si nous faisons l'hypothèse que la condition relative à l'exogénéité des régresseurs (MLR.4) est valide, c'est-à-dire que $E(u|x_1, \ldots, x_k) = 0$, alors nous avons bien comme toujours :

$$E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

où **x** désigne le vecteur regroupant l'ensemble des variables explicatives du modèle.

L'élément clé ici est que puisque y est une variable binaire, prenant donc toujours les valeurs zéro ou un, alors la probabilité de « succès » $P(y = 1|\mathbf{x})$ est égale à $E(y|\mathbf{x})$ l'espérance conditionnelle de y, soit $P(y = 1|\mathbf{x}) = E(y|\mathbf{x})$ est toujours vraie. Il vient :

$$P(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \qquad [7.27]$$

qui nous dit que la probabilité de succès que l'on note $p(\mathbf{x}) = P(y = 1|\mathbf{x})$, est une fonction linéaire des x_j. L'équation (7.27) est un exemple de modèle de choix binaire, et $P(y = 1|\mathbf{x})$ est également appelée la **probabilité de réponse**. (Nous traiterons d'autres modèles de choix binaire dans le chapitre 17.) Du fait que la somme des probabilités doit valoir 1, $P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x})$ est aussi une fonction linéaire des x_j.

Le modèle de régression multiple avec variable binaire dépendante est appelé **modèle à probabilités linéaires (MPL)** du fait que la probabilité de réponse est une fonction linéaire des paramètres β_j. Dans le MPL, β_j mesure le changement dans la probabilité de succès lorsque x_j change, toutes choses égales par ailleurs :

$$\Delta P(y = 1|\mathbf{x}) = \beta_j \Delta x_j. \qquad [7.28]$$

Le modèle de régression multiple peut alors nous permettre d'estimer les effets de différentes variables explicatives sur des événements qualitatifs. La mécanique des MCO demeure identique. Soit l'équation estimée :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k,$$

nous devons nous souvenir que $\hat{y}$ correspond ici à la probabilité de succès estimée. De ce fait, $\hat{\beta}_0$ s'interprète comme la probabilité de succès estimée lorsque chacun des x_j est fixé à zéro, ce qui peut avoir du sens ou pas. Le coefficient (de pente) $\hat{\beta}_1$ mesure le changement prédit dans la probabilité de succès lorsque x_1 augmente d'une unité.

Pour interpréter correctement le modèle à probabilités linéaires, nous devons savoir ce que constitue un « succès ». Un bon moyen de procéder consiste à désigner la variable par le nom décrivant l'événement $y = 1$. À titre d'exemple, soit *inlf* (« *in the labor force* ») une variable binaire indiquant la participation au marché du travail d'une femme mariée durant l'année 1975 : *inlf* = 1 si la femme rapporte avoir exercé une activité salariée durant l'année, zéro sinon. Nous faisons l'hypothèse que le taux de participation au marché du travail dépend des autres sources de revenus et notamment des ressources du mari (capturées par la variable *nwifeinc*, mesurée en milliers de dollars), du nombre d'années d'études (*educ*), du nombre d'années d'expérience professionnelle (*exper*), de l'âge (*age*), du nombre d'enfant de moins de 6 ans (*kidslt6*), et du nombre d'enfants entre 6 et 18 ans (*kidsge6*). À partir des données contenus dans MROZ reprises de Mroz (1987), nous estimons le modèle à probabilités linéaires suivant, avec 428 des 753 femmes de l'échantillon reportant avoir exercé une activité salariée durant l'année 1975 :

$$\begin{aligned} \widehat{inlf} &= 0{,}586 - 0{,}0034\,nwifeinc + 0{,}038\,educ + 0{,}039\,exper \\ &\quad (0{,}154)\ (0{,}0014)\qquad (0{,}007)\qquad (0{,}006) \\ &\quad - 0{,}00060\,exper^2 - 0{,}016\,age - 0{,}262\,kidslt6 + 0{,}013\,kidsge6 \\ &\quad (0{,}00018)\qquad (0{,}002)\qquad (0{,}034)\qquad (0{,}013) \\ n &= 753,\ R^2 = 0{,}264. \end{aligned} \qquad [7.29]$$

En faisant l'usage des statistiques t traditionnelles, toutes les variables de (7.29) à l'exception de *kidsge6* apparaissent statistiquement significatives et leurs effets correspondent aux attentes de la théorie économique (ou du bon sens).

Pour interpréter ces estimations, nous devons nous souvenir qu'un changement de la variable indépendante a pour conséquence une modification de la probabilité que *inlf* = 1. Par exemple, le coefficient relatif à *educ* implique que, toutes choses égales par ailleurs dans (7.29), une nouvelle année d'études augmente la probabilité

de participation au marché du travail de 0,038. Si nous prenons cette estimation au pied de la lettre, 10 années d'études supplémentaires augmenterait la probabilité de participation au marché du travail de 0,038(10) = 0,38, ce qui s'apparente ici à une hausse substantielle de la probabilité estimée. La relation entre la probabilité de participation au marché du travail et la variable *educ* est caractérisée par le graphique de la figure 7.3. En guise d'illustration, les autres variables indépendantes sont fixées aux niveaux suivants : *nwifeinc* = 50, *exper* = 5, *age* = 30, *kidslt6* = 1, et *kidsge6* = 0. La probabilité prédite est négative jusqu'à ce que le nombre d'années d'études atteigne la valeur de 3,84 années. Cela ne devrait pas trop nous inquiéter puisque dans l'échantillon, aucune femme ne présente une valeur inférieure à cinq ans. Le niveau le plus élevé pour cette variable est par ailleurs de 17 années, avec une probabilité associée de 0,5. Si nous modifions les valeurs allouées pour les autres variables explicatives, l'étendue des probabilités prédites s'en trouve modifiée. Pour autant, l'effet marginal d'une année d'étude supplémentaire sur la probabilité de participation du marché du travail est toujours de 0,038.

Le coefficient relatif à *nwifeinc* implique que, si $\Delta nwifeinc = 10$ (ce qui correspond ici à un accroissement de 10 000 USD de revenus du mari), la probabilité qu'une femme participe au marché du travail chute de 0,034. Cet effet n'est pas particulièrement important compte tenu du fait qu'une augmentation de salaire de 10 000 USD est substantielle en dollars de 1975. La variable relative à l'expérience a été introduite sous forme quadratique de façon à ce que l'effet de l'expérience passée sur la participation au marché de travail puisse diminuer avec le nombre d'années. Toutes choses égales par ailleurs, la variation de probabilité estimée est évaluée à $0,039 - 2(0,0006)exper = 0,039 - 0,0012\ exper$. La valeur seuil au-delà de laquelle l'expérience n'a plus d'impact sur la probabilité de participation au marché du travail est donnée par 0,039/0,0012 = 32,5, ce qui représente un niveau élevé d'expérience : seules 13 des 753 femmes de notre échantillon présentent un niveau d'expérience supérieur à 32 ans.

Contrairement au nombre d'enfants âgés de 6 à 18 ans, le nombre d'enfants en bas âge a un impact important sur la participation des femmes au marché du travail. Avoir un enfant de moins de 6 ans supplémentaire réduit la probabilité de participation de l'ordre de – 0,262, toutes choses égales par ailleurs. Dans notre échantillon, un peu moins de 20 % des femmes ont au moins un enfant en bas âge.

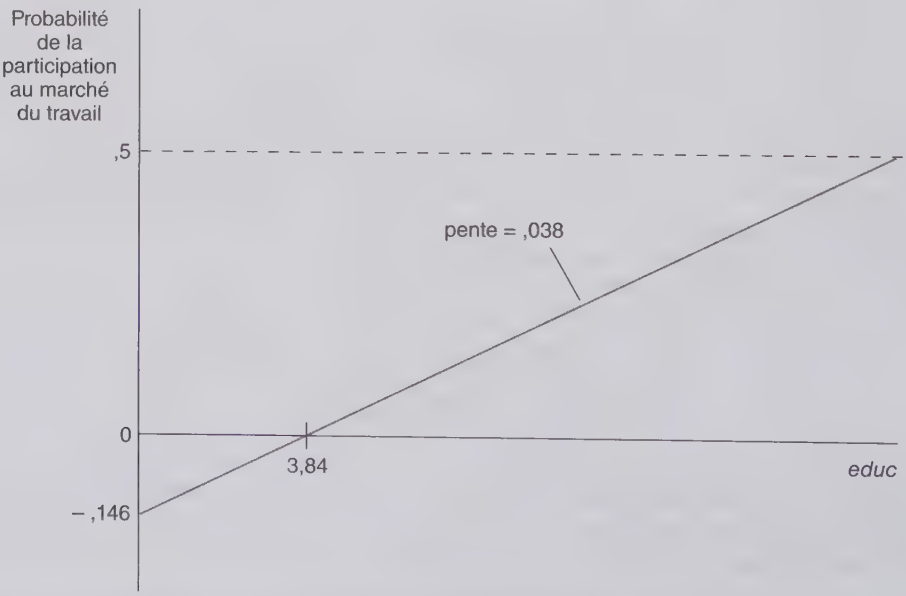


Figure 7.3 *Relation estimée entre la probabilité de participation au marché du travail et les années d'études, toutes choses égales par ailleurs.*

Cet exemple illustre à la fois la grande facilité d'estimation et d'interprétation des modèles à probabilités linéaires mais met également en lumière un certain nombre de ses limites. En premier lieu, il est aisé de voir que lorsque l'on introduit certaines combinaisons de valeurs pour les variables explicatives dans (7.29), il en résulte des valeurs soit négatives soit plus grandes que un pour la variable dépendante. Puisque celle-ci renvoie à une probabilité et qu'une probabilité par définition est bornée entre zéro et un, cela peut s'avérer quelque peu problématique. Par exemple, quel sens cela a-t-il de dire qu'une femme participe à la force de travail avec une probabilité de – 0,10 ? En fait, des 753 femmes de notre échantillon, 16 des valeurs ajustées issues de (7.29) sont inférieures à zéro, et 17 autres plus grandes que l'unité.

Un problème connexe tient à ce que selon le modèle, la probabilité est liée linéairement aux variables indépendantes quelque soient les valeurs prises par ces dernières. Par exemple, (7.29) prédit que passer de zéro à un enfant en bas âge réduit la probabilité de participation au marché du travail de 0,262. C'est également la diminution prédite dans le cas où la femme passe d'un à deux enfants en bas âge. Or, il semble plus réaliste de considérer que le premier enfant en bas âge ait l'impact le plus fort sur la probabilité de participation et que le second enfant ait ensuite un effet marginal inférieur sur cette même probabilité. De fait, pousser le raisonnement à l'extrême (7.29) implique que passer de zéro à quatre enfants en bas âge réduirait la probabilité de participation au marché du travail de $\Delta \widehat{inlf} = 0{,}262\ (\Delta kidslt6) = 0{,}262\ (4) = 0{,}262\ (4) = 1{,}0481$, ce qui est évidemment impossible.

Malgré ces problèmes, le modèle à probabilités linéaires est utile et souvent utilisé en économie. Il fonctionne en général assez bien pour des valeurs de variables indépendantes proches des valeurs moyennes de l'échantillon. Dans l'exemple de la participation à la force de travail, aucune des femmes de l'échantillon n'avait plus de quatre enfants ; de fait, seules trois femmes étaient mères de trois jeunes enfants. Plus de 96 % des femmes étaient sans enfant ou avaient un enfant en bas âge, impliquant par là qu'il était raisonnable de se concentrer sur ce cas particulier pour l'analyse de nos résultats d'estimation.

Les probabilités prédites supérieures à l'unité sont quelque peu troublantes lorsque nous voulons réaliser des prédictions. Pour autant il est possible de faire usage des probabilités estimées (même si certaines d'entre elles sont négatives ou supérieures à un) pour prédire un résultat binaire. Comme précédemment, supposons que $\hat{y}$ désigne les valeurs ajustées de la variable dépendante – qui potentiellement peuvent ne pas être comprises entre zéro et un. Définissons la valeur prédite (corrigée) comme $\hat{y} = 1$ si $\hat{y} \leq 0{,}5$, et $\tilde{y}_i = 0$ si $\tilde{y}_i < 0{,}5$. Nous disposons maintenant d'un ensemble de valeurs prédites, qui, à l'instar de y_i, prennent soit la valeur zéro soit un. Nous pouvons alors utiliser les données relatives à $\hat{y}_i$ et $\tilde{y}_i$ pour obtenir les fréquences de prédictions correctes de $y_i = 1$ et $y_i = 0$, de même que la proportion de prédictions correctes dans l'ensemble. Cette dernière mesure, lorsqu'elle est exprimée en pourcentage est une mesure très standard de la bonne adéquation d'un modèle à variables dépendantes binaires que l'on appelle le **pourcentage des prédictions correctes**. Un exemple est détaillé dans l'exercice sur ordinateur C9(v), et de plus amples développements dans le contexte de modèles plus avancés peuvent être consultés dans la section 17.1.

En raison de la nature binaire de y, le modèle à probabilités linéaires viole l'une des hypothèses de Gauss-Markov. En effet, lorsque y est une variable binaire, sa variance conditionnellement à $\mathbf{x}$, est donnée par :

$$\text{Var}(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})], \qquad \textbf{[7.30]}$$

avec $p(\mathbf{x})$ la probabilité de succès : $p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. Cela signifie qu'à l'exception du cas où la probabilité ne dépend d'aucune des variables explicatives du modèle, il **doit** y avoir de *l'hétéroscédasticité* dans le modèle à probabilités linéaires. Nous savons depuis le chapitre 3 que cela ne génère pas de biais pour les estimateurs des β_j par la méthode des MCO. Mais comme abordé dans le cadre des chapitres 4 et 5 l'hypothèse d'homoscédasticité est cruciale pour justifier le recours aux tests de Student et de Fischer, même en présence de grand échantillon. Du fait que les écarts-types estimés dans (7.29) ne sont en général pas valides, nous devons y avoir recours avec précaution. Nous verrons plus tard comment corriger les écarts-types estimés en présence d'hétéroscédasticité dans le chapitre 8. Il s'avère que dans de nombreuses

applications, les statistiques usuelles associées aux estimateurs des MCO ne sont pas trop éloignées de leurs vraies valeurs. Il est donc acceptable dans la plupart des travaux appliqués d'avoir recours à l'analyse par les MCO du modèle à probabilités linéaires.

Nous pouvons également combiner variables indicatrices dépendantes et indépendantes dans un même modèle de régression. Les coefficients associés à ces dernières mesurent alors la différence de probabilité prédite relativement au groupe de référence. Par exemple, si nous ajoutons deux variables indicatrices capturant les origines ethniques, *black* et *hispan*, à l'équation relative aux arrestations en 1986, nous obtenons les résultats suivants :

$$\begin{aligned}\widehat{arr86} &= 0{,}380 + 0{,}152\,pcnv + 0{,}0046\,avgsen - 0{,}0026\,tottime \\ &\quad (0{,}019)\ (0{,}021)\qquad (0{,}0064)\qquad (0{,}0049) \\ &\quad - 0{,}024\,ptime86 - 0{,}038\,qemp86 + 0{,}170\,black + 0{,}096\,hispan \\ &\quad (0{,}005)\qquad (0{,}005)\qquad (0{,}024)\qquad (0{,}021) \\ n &= 2\,725,\ R^2 = 0{,}0682.\end{aligned} \qquad [7.32]$$

Le coefficient estimé associé à la variable *black* implique que, toutes choses égales par ailleurs, un homme africain américain aura une probabilité de 0,17 fois plus élevée d'être arrêté en moyenne qu'un homme de type caucasien (le groupe de référence dans cet exemple). Une autre façon d'interpréter ce résultat est de relever que la probabilité d'arrestation en 1986 est de 17 points de pourcentage plus grand pour les africains américains que pour les personnes de type caucasien. Cette différence apparaît statistiquement significative. De façon équivalente, un homme issu de la communauté hispanique aura une probabilité moyenne 0,096 fois plus élevée d'être arrêté qu'un homme de type caucasien.

Pour aller plus loin 7.5

Quelle est la probabilité prédite d'arrestation d'un homme africain américain sans antécédent de condamnation par la justice – de sorte que *pcnv*, *avgsen*, *tottime*, et *ptime86* sont toutes fixées à zéro – ayant exercé une activité professionnelle durant les quatre trimestres de 1986 ? Ce résultat vous paraît-il plausible ?

EXEMPLE 7.12
Un modèle à probabilités linéaires du nombre d'arrestations

Soit *arr86* une variable binaire égale à un si un individu a été arrêté durant 1986 et zéro sinon. La population considérée ici correspond à un groupe d'hommes jeunes, de Californie nés entre 1960 et 1961 ayant été arrêtés au moins une fois avant 1986. Un modèle à probabilités linéaires pour expliquer la variable *arr86* est donné par :

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u,$$

où

pcnv = la proportion d'arrestations antérieures ayant mené à une condamnation.
avgsen = la peine moyenne consécutive aux condamnations passées (exprimée en mois).
tottime = le nombre de mois passés en prison depuis l'âge de 18 ans avant 1986.
ptime86 = le nombre de mois passés en prison en 1986.
qemp86 = le nombre de trimestres (de 0 à 4) durant lequel l'individu était employé légalement en 1986.

Les données utilisées pour cet exemple sont issues de CRIME1, soit la même base que celle de l'exemple 3.5. Ici, nous introduisons une variable dépendante binaire car seuls 7,2 % des individus de l'échantillon ont été arrêtés plus d'une fois. Environ 27,7 % des hommes ont été arrêtés au moins une fois durant l'année 1986. Les résultats d'estimation sont les suivants :

$$\begin{aligned}\widehat{arr86} &= 0{,}441 - 0{,}162\,pcnv + 0{,}0061\,avgsen - 0{,}0023\,tottime \\ &\quad (0{,}017)\ (0{,}021)\qquad (0{,}0065)\qquad (0{,}0050)\end{aligned}$$

$$- 0{,}022\ ptime86 - 0{,}043\ qemp86 \qquad [7.31]$$
$$(0{,}005) \qquad (0{,}005)$$
$$n = 2\ 725,\ R^2 = 0{,}0474.$$

La constante, 0,441, correspond à la probabilité prédite d'arrestation d'un individu n'ayant ni été condamné précédemment (puisque *pcnv* et *avgsen* sont alors fixées à zéro), ni été en prison avant l'âge de 18 ans ou durant l'année 1986 et qui était sans emploi durant l'année entière. Les variables *avgsen* et *tottime* s'avèrent non significatives à la fois sur base des tests de significativité individuelle et jointe (à la statistique F du test de Fisher est associée une p-valeur de 0,347) alors que le signe du coefficient *avgsen* s'avère contre-intuitif si l'on admet le principe que des sentences plus lourdes sont supposées avoir un effet dissuasif sur le crime. Grogger (1991), étudie cette question sur une base de données similaire mais plus étendue en ayant recours à différentes techniques économétriques et montre que *tottime* a un effet significativement *positif* sur le nombre d'arrestations et conclut alors que *tottime* s'assimile à une mesure de capital humain accumulée dans le cadre de l'activité criminelle.

Si, à la lumière de nos résultats, l'accroissement de la probabilité de condamnation semble avoir pour effet de décroître la probabilité d'arrestation en 1986, nous devons demeurer très prudent quant à l'interprétation de la valeur des coefficients estimés. La variable *pcnv* est une proportion comprise entre zéro et un. Dans ce cadre, passer d'une valeur de zéro à un s'interprète comme passer de la certitude de ne pas être condamnée à celle de l'être. Alors qu'en cas de variation forte la probabilité d'arrestation diminue de seulement 0,162 ; accroître *pcnv* de 0,5 décroît la probabilité d'arrestation durant l'année 1986 de 0,081.

L'effet d'interaction est capturé par le coefficient relatif à la variable *ptime86*. Si un individu est en prison, il ne peut pas être arrêté. Puisque la variable *ptime86* est mesurée en mois, six mois supplémentaires passés en prison réduisent la probabilité d'arrestation ultérieure de 0,022(6) = 0,132. L'équation (7.31) donne un autre exemple des limites du modèle à probabilités linéaires et de son caractère erroné pour certaines réalisations particulières des variables indépendantes. Si un homme est en prison durant les 12 mois de l'année 1986, il ne peut bien évidemment pas être arrêté en 1986. Or, en fixant toutes les autres variables à zéro, la probabilité prédite d'une arrestation en 1986 lorsque *ptime86* = 12 est de 0,441 – 0,022(12) = 0,177, ce qui est différent de zéro. Néanmoins, si nous effectuons ce même calcul à partir de l'expression de la probabilité non conditionnelle d'une arrestation en 1986, 0,277, 12 mois de prison réduisent la probabilité à environ zéro : 0,277 – 0,022(12) = 0,013.

Enfin, l'emploi réduit la probabilité d'arrestation de façon significative. Toutes choses égales par ailleurs, un individu ayant travaillé durant les quatre trimestres présente 0,172 chances de moins d'être arrêté qu'un autre n'ayant exercé aucune activité durant la même période.

7.6 POUR ALLER PLUS LOIN EN MATIÈRE D'ÉVALUATION DES POLITIQUES PUBLIQUES

Nous avons passé en revue un certain nombre de modèles contenant des variables indicatrices pouvant être utiles pour l'évaluation des politiques publiques, à l'instar de l'exemple 7.3 où seul un certain nombre d'entreprises recevaient des subventions pour la formation de leur personnel.

Comme nous l'avons mentionné plus haut, nous devons nous montrer extrêmement prudent quant à l'évaluation des politiques publiques du fait que, comme dans la plupart des exercices empiriques réalisés en sciences sociales, les groupes de contrôle et de traitement ne sont pas sélectionnés de façon aléatoire. Revenons à l'étude de Holzer et al. (1993) et intéressons-nous maintenant à l'effet des subventions publiques à la formation du personnel sur la productivité des travailleurs (par opposition au volume d'heures de formation professionnelle). L'équation d'intérêt est maintenant donnée par :

$$\log(scrap) = \beta_0 + \beta_1 grant + \beta_2 \log(sales) + \beta_3 \log(employ) + u,$$

où *scrap* est le taux de rebut de l'entreprise, les deux autres variables étant des variables de contrôle. La variable binaire *grant* indique si une entreprise a bénéficié d'une subvention en 1988 pour la formation professionnelle de son personnel.

Avant d'analyser les résultats d'estimation, nous pourrions nous inquiéter de l'impact éventuel d'un certain nombre de facteurs non observés sur la productivité des travailleurs – tels que les niveaux moyens d'éducation, d'habileté, d'expérience ou d'ancienneté après titularisation – qui pourraient être corrélés avec la capacité de l'entreprise à recevoir une subvention. Holzer et al. ont mis en évidence que les subventions furent octroyées sur base d'un critère du type « premier arrivé, premier servi ». Mais cela ne revient pour autant pas au même que de distribuer ces subventions de façon purement aléatoire. Il se pourrait en effet que les entreprises dont les employés sont les moins productifs, aient vu là une opportunité à saisir pour accroître leur productivité et aient été les plus rapides à se manifester pour obtenir une subvention.

À partir des données contenues dans JTRAIN pour 1988 – à partir du moment où les entreprises ont été en mesure de percevoir les subventions – nous obtenons les résultats d'estimation suivants :

$$\begin{aligned} \widehat{\log(scrap)} &= 4{,}99 - 0{,}052\, grant - 0{,}455 \log(sales) \\ &\quad (4{,}66)\ (0{,}431) \qquad (0{,}373) \\ &\quad + 0{,}639 \log(employ) \\ &\quad (0{,}365) \\ n &= 50,\ R^2 = 0{,}072. \end{aligned} \qquad \textbf{[7.33]}$$

(17 des 50 entreprises ont reçu une subvention pour la formation de leur personnel et le taux moyen de rebut est évalué à 3,47 en moyenne sur l'ensemble des entreprises.) L'estimation (ponctuelle) du coefficient associé à la variable *grant* est ici de – 0,052 et implique qu'à niveaux donnés pour *sales* et *employ*, les entreprises percevant une subvention, présentent un taux de rebut d'environ 5,2 % plus faible que les entreprises n'en ayant pas bénéficié. Ceci est conforme aux effets attendus, si les subventions sont efficaces, mais les statistiques t associées sont très faibles. De fait, à partir de ces résultats en coupe, nous devons conclure que les subventions n'ont pas d'effet significatif sur la productivité des entreprises. Nous reviendrons sur cet exemple dans le cadre du chapitre 9 et montrerons comment l'ajout d'information relative à l'année antérieure à l'étude nous mène à des conclusions bien différentes.

Même dans des cas où l'évaluation de politique publique n'implique pas d'assigner certaines observations à un groupe de contrôle ou de traitement, nous devons demeurer très prudents et inclure tous facteurs qui pourraient être systématiquement liés à notre variable dépendante binaire d'intérêt. Un bon exemple est donné par l'étude du phénomène de discrimination raciale. Les origines ethniques sont une caractéristique qui n'est par essence pas déterminée par un individu ou un gouvernement et à ce titre, pourrait constituer l'illustration parfaite d'un facteur exogène dans la mesure où celles-ci sont déterminées à la naissance. Cependant, pour des raisons historiques, les origines ethniques sont souvent liées à d'autres facteurs pertinents : il existe des différences systématiques en matière d'origines familiales et sociales entre groupes ethniques, et ces différences doivent être prises en compte dans l'évaluation des discriminations actuelles.

À titre d'exemple, étudions maintenant le phénomène de discrimination ethnique dans le cadre spécifique de l'octroi de crédits. Supposons que nous ayons la possibilité de collecter des données relatives aux candidatures individuelles pour une hypothèque, nous pouvons alors définir une variable indicatrice indépendante nommée *approved* valant un si une candidature à l'hypothèque a été approuvée, zéro sinon. Une différence systématique des taux d'approbation entre personnes d'origines ethniques distinctes est une indication de discrimination éventuelle. Cependant, dans la mesure où l'approbation du dossier de candidature dépend de nombreux autres facteurs, notamment du revenu, de la richesse, du risque crédit évalué sur base d'un score, ainsi que de la capacité d'ensemble à rembourser le prêt octroyé, nous devons tenir compte de ces effets et étudier s'il existe des différences systématiques pour ces facteurs au regard des origines ethniques.

Un modèle à probabilités linéaires permettant de tester le phénomène de discrimination sur critère ethnique pourrait être le suivant :

$$approved = \beta_0 + \beta_1 nonwhite + \beta_2 income + \beta_3 wealth + \beta_4 credrate + autres\ facteurs.$$

La discrimination à l'encontre des minorités sera caractérisée dans notre exemple par le rejet de l'hypothèse nulle $H_0 : \beta_1 = 0$ en faveur de $H_1 : \beta_1 < 0$ puisque β_1 correspond au différentiel de probabilités d'obtention d'une hypothèque entre les personnes de type caucasien et les autres (la population de type caucasien étant le groupe de référence) toutes choses égales par ailleurs. Si les variables *income*, *wealth*, etc. prennent des valeurs systématiquement différentes selon les origines ethniques, alors il est important de tenir compte de ces facteurs dans nos régressions.

Un autre problème qui émerge régulièrement lorsque l'on souhaite évaluer des mesures de politique publique tient au fait que les individus (entreprises ou villes) concernés sont libres de participer ou non à ces programmes. Par exemple, les individus choisissent de consommer des drogues ou de boire de l'alcool. Si nous voulons examiner les effets de tels comportements sur l'emploi de ces personnes, leurs revenus ou leurs comportements criminels, nous devons tenir compte du fait que la consommation de drogue peut être corrélée avec d'autres facteurs qui peuvent à leur tour avoir des effets sur l'emploi et la criminalité. Les enfants éligibles pour des programmes tels que le programme « *Head Start* »[3] aux États-Unis peuvent participer sur autorisation parentale. Dans la mesure où les origines familiales sont prises en considération par « *Head Start* » pour l'intégration au programme et affecte les résultats des étudiants retenus, nous devrions considérer ces facteurs comme des variables de contrôle dans notre analyse des effets du programme « *Head Start* » [voir à titre d'exemple, Currie and Thomas (1995)]. Les individus (salariés ou gouvernements) sélectionnés par des agences pour participer à des programmes de formation peuvent ou non participer audit programme et il est peu probable que cette décision soit aléatoire [voir, par exemple, Lynch (1992)]. Les villes ou États peuvent décider d'implémenter ou non une législation en faveur du contrôle des armes à feu, et il est probable que cette décision soit systématiquement liée à d'autres facteurs influant sur le taux de criminalité [voir à ce titre, Kleck and Patterson (1993)].

Le paragraphe précédent détaillait un certain nombre d'exemples de ce qu'il est commun de nommer en économie des problèmes d'**auto-sélection**. L'expression vient du fait que les individus choisissent leur groupe d'appartenance relativement à certains comportements ou programmes : l'adoption d'un comportement ou d'un programme n'est donc pas déterminée de façon aléatoire. Ce problème survient lorsqu'une variable binaire de participation peut être systématiquement reliée à des facteurs inobservés. En effet, soit le modèle de régression simple suivant :

$$y = \beta_0 + \beta_1\ partic + u, \qquad \textbf{[7.34]}$$

où y correspond à la variable expliquée mesurant un résultat et *partic* une variable binaire prenant la valeur unitaire lorsque l'individu, l'entreprise ou la ville adopte un comportement, participe à un programme ou met en place un certain type de législation. Nous nous inquiétons de ce que la valeur moyenne des u dépende de la participation c'est-à-dire : $E(u|partic = 1) \neq E(u|partic = 0)$. Comme nous le savons, cette propriété a pour conséquence de biaiser l'estimateur des MCO de β_1, nous empêchant alors d'identifier le véritable effet de la participation sur notre variable expliquée. De ce fait, le problème d'auto-sélection est une autre source potentielle d'endogénéité des variables explicatives (*partic* dans notre cas).

Nous savons que le modèle de régression multiple peut dans une certaine mesure, nous permettre d'éviter le problème d'auto-sélection. Les déterminants inclus dans le terme d'erreur de l'équation (7.34)

3 NdT : « *Head Start* » est un programme du Département de la Santé, de l'éducation et des services sociaux des États-Unis qui fournit une éducation complète, des services d'implication parentale, de santé, de nutrition, aux enfants à faibles revenus et à leurs familles. Source : Wikipédia (consulté en septembre 2014).

corrélés avec *partic* peuvent être inclus dans un modèle de régression multiple, en faisant l'hypothèse qu'il est possible d'obtenir des données pour ces facteurs bien évidemment. Malheureusement, dans de nombreux cas cela n'est pas possible, et en particulier, un certain nombre de facteurs non-observables sont liés à la participation, induisant un biais d'estimation des paramètres du modèle de régression multiple.

Dans le cadre des modèles de régression multiples traditionnels en coupes instantanées, nous devons être conscients du risque d'identifier des effets fallacieux lorsque nous évaluons l'impact d'un programme ou d'une mesure de politique publique du fait du problème d'auto-sélection. Un bon exemple est celui décrit par Currie and Cole (1993). Ces auteurs examinent l'effet de la participation au programme AFDC (« *Aid to Families with Dependent Children* »[4]) aux États-Unis sur le poids de naissance des nourrissons. Même après avoir introduit un certain nombre de variables de contrôle relatives aux origines familiales et sociales des individus, les auteurs obtiennent des estimations par les MCO, impliquant que la participation au programme AFDC a pour effet de réduire le poids moyen des nourrissons à la naissance. Comme les auteurs le soulignent, il est difficile de croire que la participation au programme seule *cause* un amoindrissement du poids à la naissance. [Voir Currie (1995) pour des exemples additionnels.] À l'aide d'une méthode économétrique alternative, que nous présenterons au chapitre 15, Currie et Cole obtiennent des résultats en faveur d'une absence d'effet ou d'un effet *positif* de la participation au programme AFDC sur le poids des nourrissons à la naissance.

Pour corriger le problème d'auto-sélection qui entraîne un biais dans les résultats d'estimation des paramètres du modèle de régression multiple, et lorsqu'il n'est pas possible d'introduire un nombre suffisant de variables de contrôle, les méthodes plus avancées présentées dans les chapitres 13, 14, et 15 sont requises.

7.7 INTERPRÉTER DES RÉSULTATS DE RÉGRESSION AVEC DES VARIABLES DÉPENDANTES DISCRÈTES

Une réponse binaire est la forme la plus extrême que puisse prendre les réalisations d'une variable de choix discret puisque ladite variable prend alors seulement deux valeurs possibles à savoir zéro ou un. Comme nous l'avons discuté précédemment dans la section 7.5, les paramètres du modèle à probabilités linéaires peuvent être interprétés comme mesurant le changement dans la *probabilité* que $y = 1$ résultant de l'accroissement d'une unité des variables explicatives associées, toutes choses égales par ailleurs. Nous avions également souligné qu'en raison de la nature binaire de y, l'égalité suivante était vraie : $P(y = 1) = E(y)$, même lorsque nous la conditionnons sur les réalisations d'un ensemble de variables explicatives.

D'autres variables dépendantes discrètes peuvent en outre être introduites en pratique comme nous l'avons déjà vu dans le cadre d'exemples tels que celui relatif à la modélisation du nombre d'arrestations (exemple 3.5). Des études relatives aux facteurs influençant la fertilité ont souvent recours au nombre d'enfants en vie comme variable dépendante dans des modèles de régression. À l'instar du nombre d'arrestations, le nombre d'enfants en vie peut prendre un petit nombre de valeurs entières, et parmi elles, la valeur nulle de façon relativement fréquente. Les données issues de FERTIL2 qui contiennent des informations relatives à un grand échantillon de femmes du Botswana en est un exemple. La plupart du temps, les démographes s'intéressent à la mesure des effets de l'éducation sur la fertilité, et essaient en particulier de déterminer l'existence d'un lien causal de la première sur la seconde. De tels exemples soulèvent la question de savoir comment interpréter les résultats d'estimation des coefficients d'un modèle de régression : après tout, personne ne peut avoir des « fractions » d'enfants.

4 NdT : AFDC était un programme fédéral d'aide aux familles défavorisées mis en place entre 1935 et 1996 dans le cadre du *Social Security Act* aux États-Unis et géré par le ministère de la santé et des services sociaux américain. Source : Wikipédia, consulté en septembre 2014.

De façon à illustrer cette question, nous nous proposons d'étudier les résultats d'estimation du modèle de régression sur les données issues de FERTIL2 ci-après :

$$\widehat{children} = -1{,}997 + 0{,}175\,age - 0{,}090\,educ \quad [7.35]$$
$$(0{,}094)\ (0{,}003)\qquad (0{,}006)$$
$$n = 4\,361,\ R^2 = 0{,}560.$$

À ce stade, nous ignorons si cette régression introduit suffisamment de variables de contrôle pour correctement estimer les effets de chacune des variables explicatives sur la fertilité. Ici, nous nous préoccupons plutôt de l'interprétation des coefficients estimés.

Considérons le principal coefficient d'intérêt à savoir $\hat{\beta}_{educ} = -0{,}090$. Si nous l'interprétons de façon littérale, cela implique que chaque année d'études additionnelle réduit le nombre d'enfants estimé de 0,090 – ce qui est vraisemblablement impossible pour toute femme. Un problème équivalent survient lorsque l'on tente d'interpréter le coefficient relatif à l'âge $\hat{\beta}_{age} = 0{,}175$. Comment tirer de l'information qui ait du sens à partir de ces résultats d'estimation ?

Il est utile de se souvenir que l'interprétation des résultats des MCO dans le cas général, qui s'applique même lorsque y est une variable discrète avec un faible nombre de valeurs possibles, fait référence aux effets de x_j sur la valeur *attendue* (ou *moyenne*) de y. En général, sous les hypothèses MLR.1 et MLR.4,

$$\mathrm{E}(y|x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \quad [7.36]$$

De ce fait, β_j correspond à l'effet d'une hausse de x_j sur la valeur attendue de y toutes choses égales par ailleurs. Comme nous l'avons discuté précédemment dans la section 6.4, pour un ensemble de valeurs possibles de x_j, nous interprétons la valeur prédite $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$ comme une estimation de l'espérance conditionnelle de y sachant $\boldsymbol{x}$, soit $\mathrm{E}(y|x_1, x_2, \ldots, x_k)$. Dès lors, β_j correspond à l'estimation de l'ampleur avec laquelle la *moyenne* de y change lorsque $\Delta x_j = 1$ (toutes choses étant égales par ailleurs).

Vu sous cet angle, nous pouvons à nouveau nous atteler à l'interprétation de nos résultats d'estimation pour le modèle des déterminants de la fertilité décrits dans l'équation (7.35). Le coefficient $\hat{\beta}_{educ} = -0{,}090$ signifie que nous estimons que la fertilité chute *en moyenne* d'environ 0,09 enfants pour une année d'étude supplémentaire. Une manière élégante d'interpréter ce résultat consiste à dire que si chaque femme parmi un groupe de 100 individus, obtient une année d'étude supplémentaire, nous estimons à neuf le nombre d'enfants en moins pour l'ensemble d'entre elles.

Ajouter des variables indicatrices dans des modèles de régression lorsque y elle-même est discrète ne pose pas de problèmes lorsque nous estimons les effets estimés en termes de valeurs moyennes. À partir des données issues de FERTIL2 nous obtenons les résultats d'estimation suivants :

$$\widehat{children} = -2{,}071 + 0{,}177\,age - 0{,}079\,educ - 0{,}362\,electric \quad [7.37]$$
$$(0{,}095)\ (0{,}003)\qquad (0{,}006)\qquad (0{,}068)$$
$$n = 4\,358,\ R^2 = 0{,}562,$$

avec *electric* une variable indicatrice prenant la valeur un si la femme vit dans une maison avec électricité. Bien évidemment, il ne peut être vraisemblable qu'une femme bénéficiant de l'électricité ait 0,362 enfant de moins qu'une femme aux caractéristiques similaires dans la situation contraire. Mais il est possible de dire que lorsque l'on compare 100 femmes bénéficiant de l'électricité à domicile à 100 autres femmes ne bénéficiant pas de l'électricité – à niveaux d'âge et d'éducation similaires – on estime que le premier groupe aura 36 enfants de moins.

Il est à noter que, lorsque y est une variable discrète, le modèle linéaire n'est pas toujours le plus apte à fournir les meilleures estimations des effets marginaux sur $\mathrm{E}(y|x_1, x_2, \ldots, x_k)$. Le chapitre 17 passe en

revue un ensemble de modèles ainsi que des méthodes d'estimation plus avancés permettant de mieux expliquer les données lorsque l'étendue des valeurs possibles pour y est limitée de façon substantielle. Ceci étant dit, un modèle linéaire estimé par la méthode des MCO, fournit une bonne approximation des vrais effets marginaux, au moins en moyenne.

RÉSUMÉ

Dans ce chapitre, nous avons appris comment utiliser de l'information qualitative dans des modèles de régression. Dans les cas les plus simples, une variable indicatrice est définie pour distinguer les individus appartenant à différents groupes, et les coefficients associés à ces variables indicatrices correspondent aux différences entre deux groupes toutes choses égales par ailleurs. Il est possible de considérer plus de deux groupes distincts. Pour ce faire, il convient de définir un ensemble de variables indicatrices : en présence de g groupes, on définit $g - 1$ variables indicatrices qui seront incluses dans le modèle. L'ensemble des estimations relatives aux variables indicatrices sont interprétées relativement à un groupe de référence (le groupe pour lequel la variable indicatrice est absente du modèle).

Les variables indicatrices sont également utiles pour incorporer de l'information ordinale dans des modèles de régressions, telle que l'évaluation du risque de crédit. Nous définissons simplement un ensemble de variables indicatrices représentant les différents résultats possibles de la variable ordinale en désignant l'une des catégories comme étant le groupe de référence.

Les variables indicatrices peuvent être croisées avec des variables quantitatives pour permettre aux pentes de différer selon l'appartenance à tel ou tel groupe. Dans le cas extrême, nous pouvons autoriser chaque groupe à être caractérisé par des coefficients particuliers pour chacune des variables, de même qu'une constante spécifique. Le test de Chow peut être utilisé pour détecter d'éventuelles différences entre groupes. Dans de nombreux cas, il est plus intéressant de tester si, après avoir relâché l'hypothèse d'une constante commune pour tous les groupes, les pentes pour deux groupes différents peuvent être considérées comme identiques. Un test de Fisher standard permet alors de valider empiriquement cette assertion, dans un modèle non contraint incluant toutes les interactions possible entre les variables se rapportant aux différents groupes et les autres variables explicatives du modèle.

Le modèle à probabilités linéaires, qui est simplement estimé par la méthode des MCO, nous permet d'expliquer une variable dichotomique en ayant recours à la technique de la régression linéaire. Les estimations par les MCO sont interprétées comme les changements dans la probabilité de « succès » ($y = 1$), consécutif à l'accroissement d'une unité de la variable explicative correspondante. Le MPL souffre toutefois de quelques limites. Il peut produire des probabilités prédites négatives ou supérieures à l'unité, il implique un effet marginal constant pour chacune des variables explicatives du modèle dans sa forme originale, et contient par construction de l'hétéroscédasticité. Les deux premiers problèmes ne sont en général pas cruciaux lorsque l'on se concentre sur l'estimation des effets partiels des variables explicatives pour des valeurs moyennes de l'échantillon. L'hétéroscédasticité par ailleurs invalide les résultats des MCO relatifs aux écarts-types estimés ainsi qu'aux statistiques de tests, mais comme nous le verrons dans le prochain chapitre, ces limites peuvent être facilement dépassées dans le cas d'échantillons de grande taille.

La section 7.6 étudie comment les variables binaires sont utilisées pour évaluer l'impact des programmes et politiques publiques. Comme dans tout exercice de régression, nous devons nous souvenir que toute participation à un programme, ou toute autre variable binaire ayant des implications de politique économique, peut être corrélée avec un certain nombre de facteurs non observés qui affectent la variable dépendante avec pour conséquence un biais de variable omise.

Nous clôturons ce chapitre avec une discussion générale sur la manière d'interpréter les modèles de régression lorsque la variable dépendante est de nature discrète. L'élément clé est de se souvenir que les coefficients peuvent être interprétés comme les effets estimés sur la valeur attendue de la variable dépendante.

MOTS-CLÉS

Analyse de politique économique p. 281
Auto-sélection p. 305
Différence des pentes p. 297
Évaluation de politique publique p. 281
Groupe de contrôle p. 281
Groupe de traitement p. 281
Modèle à probabilités linéaires (MPL) p. 299
Pourcentage de prédictions correctes p. 301
Probabilité de réponse p. 299
R-carré non centré p. 286
Terme d'interaction p. 290
Test de Chow p. 297
Trappe à variables indicatrices p. 278
Variable binaire p. 276
Variable dichotomique p. 276
Variable indicatrice p. 276
Variable ordinale p. 286

EXERCICES

1. À partir des données contenues dans SLEEP75 (voir également l'exercice 3 du chapitre 3), nous obtenons les résultats d'estimation suivants :

$$\widehat{sleep} = 3,840 - 0,163\,totwrk - 11,71\,educ - 8,70\,age$$
$$(235,11)\quad (0,018)\qquad (5,86)\qquad (11,21)$$
$$+ 0,128\,age^2 + 87,75\,male$$
$$(0,134)\qquad (34,33)$$
$$n = 706,\ R^2 = 0,123,\ \bar{R}^2 = 0,117.$$

La variable *sleep* correspond au nombre total de minutes de sommeil nocturne par semaine, *totwrk* au nombre total de minutes consacrées au travail par semaine, *educ* et *age* sont mesurées en années, et *male* correspond à la variable indicatrice indiquant le genre.

i. Toutes choses égales par ailleurs, peut-on prouver que les hommes dorment plus que les femmes ? Ce résultat est-il robuste ?

ii. Observe-t-on un arbitrage statistiquement significatif entre temps de sommeil et temps de travail ? À combien est-il estimé ?

iii. Quel autre modèle de régression devez-vous estimer pour tester l'hypothèse nulle selon laquelle, l'âge n'a pas d'effet sur le sommeil toutes choses égales par ailleurs ?

2. Les modèles de régression suivants ont été estimés sur les données contenues dans BWGHT :

$$\widehat{\log(bwght)} = 4,66 - 0,0044\,cigs + 0,0093\log(faminc) + 0,016\,parity$$
$$(0,22)\ (0,0009)\qquad (0,0059)\qquad (0,006)$$
$$+ 0,027\,male + 0,055\,white$$
$$(0,010)\qquad (0,013)$$
$$n = 1\,388,\ R^2 = 0,0472$$

et

$$\widehat{\log(bwght)} = 4{,}65 - 0{,}0052\,cigs + 0{,}0110\log(faminc) + 0{,}017\,parity$$
$$(0{,}38)\ (0{,}0010)\qquad (0{,}0085)\qquad (0{,}006)$$
$$+ 0{,}034\,male + 0{,}045\,white - 0{,}0030\,motheduc + 0{,}0032\,fatheduc$$
$$(0{,}011)\qquad (0{,}015)\qquad (0{,}0030)\qquad (0{,}0026)$$
$$n = 1\,191,\ R^2 = 0{,}0493.$$

Les variables sont définies comme dans l'exemple 4.9, nous y avons ajouté deux variables indicatrices permettant de préciser si l'enfant est un garçon et s'il est de type caucasien.

i. Dans la première équation, interprétez le coefficient associé à la variable *cigs*. Déterminez plus précisément quel est l'effet sur le poids de naissance de l'enfant de la consommation de plus de 10 cigarettes par jour par la mère.

ii. Combien de kg de plus s'attend-t-on à ce qu'un enfant de type caucasien pèse à la naissance comparativement aux autres enfants au regard des résultats prédits par le modèle, toutes choses égales par ailleurs ? Cette différence est-elle statistiquement significative ?

iii. Commentez les estimation et significativité statistique du coefficient associé à *motheduc*.

iv. À partir des résultats estimés, expliquez pourquoi vous n'êtes pas en mesure de calculer la statistique de Fisher relative au test de significativité jointe de *motheduc* et *fatheduc* ? Comment devriez-vous procéder pour pouvoir la calculer ?

3. À partir des données de GPA2, on estime le modèle suivant :

$$\widehat{sat} = 1028{,}10 + 19{,}30\,hsize - 2{,}19\,hsize^2 - 45{,}09\,female$$
$$(6{,}29)\ (3{,}83)\qquad (0{,}53)\qquad (4{,}29)$$
$$- 169{,}81\,black + 62{,}31\,female{\cdot}black$$
$$(12{,}71)\qquad (18{,}15)$$
$$n = 4\,137,\ R^2 = 0{,}0858.$$

La variable *sat* désigne le score SAT, *hsize* la taille de la classe du lycée dont est issu l'étudiant, en centaines, *female* une variable indicatrice capturant le genre, et *black* une variable indicatrice égale à un si l'individu est africain américain, zéro sinon.

i. Les résultats estimés confirment-ils que $hsize^2$ est une variable pertinente du modèle ? À partir de ces résultats, quelle est la taille optimale d'une classe dans le cycle secondaire ?

ii. Supposons *hsize* fixé, quelle est la différence estimée de score SAT entre les hommes et les femmes n'appartenant pas à la communauté africaine américaine ? Quelle est la significativité statistique de cette différence ?

iii. Quelle est la différence estimée en termes de score SAT entre les hommes africains américains et non africains américains ? Testez l'hypothèse nulle d'absence de différences de scores contre l'alternative d'une différence significative.

iv. Quelle est la différence estimée de score SAT entre les femmes africaines américaines et non africaines américaines ? Que devriez-vous entreprendre pour tester l'existence d'une différence significative ?

4. Un modèle expliquant le salaire du directeur général d'une entreprise est donné par :

$$\widehat{\log(salary)} = 4{,}59 + 0{,}257\log(sales) + 0{,}011\,roe + 0{,}158\,finance$$
$$(0{,}30)\ (0{,}032)\qquad (0{,}004)\qquad (0{,}089)$$
$$+ 1{,}181\,consprod - 0{,}283\,utility$$
$$(0{,}085)\qquad (0{,}099)$$
$$n = 209,\ R^2 = 0{,}357.$$

Les données utilisées ici sont contenues dans le fichier CEOSAL1, avec *finance*, *consprod*, et *utility* des variables binaires identifiant les industries financières, de biens de consommation et de services respectivement, l'industrie non reportée étant celle relative aux transports.

i. Calculez la valeur approchée du pourcentage de différence de salaire estimé entre les industries de services et de transports pour des niveaux de *sales* et *roe* fixés. Cette différence est-elle statistiquement significative au seuil de 1 % ?

ii. Aidez-vous de l'équation (7.10) pour calculer le pourcentage exact de différence de salaire estimé entre les industries de services et des transports et comparez votre résultat à celui obtenu dans la question précédente.

iii. Quelle est la valeur approchée du pourcentage de différence de salaire estimé entre les industries financières et de biens de consommation ? Écrivez un modèle permettant de tester si cette différence est statistiquement significative.

5. Dans l'exemple 7.2, posons *noPC* une variable indicatrice égale à un si l'étudiant ne possède pas son propre ordinateur et zéro sinon.

i. Si *noPC* est introduit à la place de *PC* dans le modèle décrit par l'équation (7.6), quelle conséquence cela entraîne-t-il pour la valeur de la constante estimée ? Quelle sera alors la valeur du coefficient associé à *noPC* ? (*Astuce* : Écrivez $PC = 1 - noPC$ et introduisez cette expression dans l'équation $\widehat{colGPA} = \hat{\beta}_0 + \hat{\delta}_0 PC + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT$.)

ii. Quel impact ce changement aura-t-il sur la valeur du *R-carré* ?

iii. Les variables *PC* et *noPC* devraient-elles être incluses toutes les deux dans le modèle ? Justifiez.

6. Afin d'évaluer la performance d'un programme de formation professionnelle sur le niveau des salaires des employés, nous posons le modèle suivant :

$$\log(wage) = \beta_0 + \beta_1 train + \beta_2 educ + \beta_3 exper + u,$$

avec *train* une variable binaire égale à un si un travailleur a effectivement bénéficié du programme. Raisonnez en considérant que le terme d'erreur *u* contient des éléments non observables mesurant le degré d'habileté des employés. À supposer que les employés les moins habiles aient plus de chance de bénéficier du programme, que peut-on dire du biais résultant de l'estimation du paramètre β_1 par la méthode des MCO ? (*Astuce* : Référez-vous au chapitre 3.)

7. Dans l'exemple de l'équation (7.29), on suppose que l'on définit *outlf* comme valant un si la femme se retire du marché du travail, zéro sinon.

i. Si nous régressons *outlf* sur l'ensemble des variables indépendantes de l'équation (7.29), qu'adviendra-t-il des constantes et pentes estimées ? (*Astuce* : $inlf = 1 - outlf$. Introduisez cette expression dans l'équation $inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \ldots$ et réarrangez les termes.)

ii. Qu'adviendra-t-il des écarts-types estimés, des constantes et pentes estimées ?

iii. Quel impact ce changement aura-t-il sur le *R-carré* ?

8. Imaginez que vous collectez des données issues d'une enquête sur les salaires, le niveau d'éducation, l'expérience et le genre. Vous récoltez en plus de l'information relative à la consommation de marijuana. La question posée est plus exactement : « À quelles occasions distinctes avez-vous fumé de la marijuana ce mois dernier ? »

i. Écrivez un modèle vous permettant d'estimer les effets de la consommation de marijuana sur le salaire, en tenant compte des autres facteurs. Vous devriez être capable d'établir des conclusions telles que « Fumer de la marijuana cinq fois de plus par mois a un impact estimé sur le salaire de *x* %. »

ii. Écrivez un modèle vous permettant de tester si la consommation de drogue a un impact différencié sur le salaire entre hommes et femmes. Comment testeriez-vous l'absence de différence d'impact de la consommation de drogue entre hommes et femmes ?

iii. À supposer qu'il est préférable de mesurer la consommation de marijuana en distinguant quatre catégories de personnes : non-consommateur, consommateur occasionnel (1 à 5 fois par mois), consommateur modéré (6 à 10 fois par mois) et consommateur régulier (plus de 10 fois par mois), écrivez maintenant le modèle vous permettant d'estimer les effets de la consommation de marijuana sur le salaire.

iv. Utilisez le modèle développé en (iii) et expliquez en détails comment vous testeriez l'hypothèse nulle que la consommation de marijuana n'a pas d'effet sur le salaire. Soyez très précis dans votre réponse et détaillez la liste des degrés de liberté.

v. Quels sont les problèmes potentiels que vous pourriez rencontrer en tentant de tirer des conclusions causales à partir des données d'enquête collectées précédemment ?

9. Soit d une variable indicatrice et z une variable quantitative. On considère le modèle suivant :

$$y = \beta_0 + \delta_0 d + \beta_1 z + \delta_1 d \cdot z + u ;$$

Il s'agit là de la version générale du modèle avec interaction entre une variable indicatrice et une variable quantitative. [Un cas particulier est traité à l'équation (7.17).]

i. Sans perdre en généralité, on fixe le terme d'erreur à zéro, soit $u = 0$. Par la suite, lorsque $d = 0$ on peut écrire la relation entre y et z comme une fonction. Écrivez la même relation lorsque $d = 1$, en notant $f_1(z)$ à gauche de l'équation pour nommer la fonction linéaire de z.

ii. Supposons que $\delta_1 \neq 0$ (ce qui implique que les deux droites ne sont pas parallèles), montrez que la valeur de z^* telle que $f_0(z^*) = f_1(z^*)$, correspondant au point où les deux droites se coupent [comme dans la figure 7.2(b)], est donné par $z^* = z^* = -\delta_0 / \delta_1$. Montrez alors que z^* est positif si et seulement si δ_0 et δ_1 ont des signes opposés.

iii. À partir des données contenues dans le fichier TWOYEAR, on procède à l'estimation du modèle suivant :

$$\widehat{\log(wage)} = -2{,}289 - 0{,}357\, female + 0{,}50\, totcoll + 0{,}030\, female \cdot totcoll$$
$$(0{,}011)\ (0{,}015) \qquad (0{,}003) \qquad (0{,}005)$$
$$n = 6{,}763,\ R^2 = 0{,}202,$$

où tous les coefficients et écarts-types ont été arrondis à 10^{-3} près. À partir de ces résultats, trouvez la valeur de *totcoll* telle que la valeur prédite de log(*wage*) soit la même pour les hommes et les femmes.

iv. À partir de l'équation mentionnée en (iii), les femmes peuvent-elles avoir de façon réaliste, accumulé suffisamment d'années d'études pour rattraper le niveau de revenu des hommes ? Justifiez votre réponse.

10. Considérons un enfant i résidant dans un quartier spécifique, soit $voucher_i$ une variable indicatrice égale à un si l'enfant est sélectionné pour participer à un programme d'allocations d'études[5] et $score_i$ le score obtenu par cet enfant à un examen standard réalisé ultérieurement. Supposons que la variable relative à la participation, $voucher_i$, est complètement aléatoire, dans le sens où ses réalisations sont indépendantes des facteurs à la fois observés et inobservés pouvant affecter le score obtenu lors du test.

5 NdT : Certains programmes d'allocations d'études aux États-Unis impliquent la distribution de « bons » ou « tickets » (*school voucher*) permettant aux parents des enfants scolarisés de s'acquitter de tout ou partie des frais de scolarité des écoles de leur choix.

i. Si vous régressez $score_i$ sur $voucher_i$ sur un échantillon aléatoire de taille n, l'estimateur des MCO vous permet-il d'obtenir un estimateur sans biais des effets du programme d'allocations d'études sur le score ?

ii. Supposez que vous êtes en mesure de collecter des données additionnelles, tels que le niveau de revenu de la famille, la structure familiale (par ex. si l'enfant vit avec ses deux parents, etc.), ainsi que les niveaux d'éducation des parents. Avez-vous besoin d'introduire ces facteurs dans votre modèle pour obtenir un estimateur sans biais des effets du programme sur le score ? Justifiez.

iii. Pourquoi devriez-vous inclure une variable capturant les origines familiales dans la régression ? Y-aurait-il des situations pour lesquelles vous pourriez ne pas tenir compte de ces facteurs ?

11. Les équations qui suivent ont été estimées à partir des données issues de la base ECONMATH, les écarts-types estimés figurant sous les coefficients entre parenthèses. Le résultat moyen de la classe (capturé par la variable *score*), mesuré en pourcentage, est d'environ 72,2. Par ailleurs, il est à noter qu'exactement la moitié des étudiants sont de sexe masculin et que la moyenne de la variable *colgpa* (correspondant au résultat moyen en début de semestre) est d'environ 2,81.

$$\widehat{score} = 32.31 + 14.32\ colgpa$$
$$(2.00)\ \ (0.70)$$
$$n = 856,\ R^2 = .329,\ \bar{R}^2 = .328$$

$$\widehat{score} = 29.66 + 3.83\ male + 14.57\ colgpa$$
$$(2.04)\ (0.74)\ \ \ (0.69)$$
$$n = 856,\ R^2 = .349,\ \bar{R}^2 = .348$$

$$\widehat{score} = 30.36 + 2.47\ male + 14.33\ colgpa + 0.479\ male \cdot colgpa$$
$$(2.86)\ (3.96)\ \ \ (0.98)\ \ \ (1.383)$$
$$n = 856,\ R^2 = .349,\ \bar{R}^2 = .347$$

$$\widehat{score} = 30.36 + 3.82\ male + 14.33\ colgpa + 0.479\ male \cdot (colgpa - 2.81)$$
$$(2.86)\ (0.74)\ \ \ (0.98)\ \ \ (1.383)$$
$$n = 856,\ R^2 = .349,\ \bar{R}^2 = .347$$

i. Interprétez le coefficient associé à la variable *male* dans la deuxième équation et construisez un intervalle de confiance à 95 % pour β_{male}. Cet intervalle exclut-il la valeur zéro ?

ii. Dans la deuxième équation, comment se fait-il que le coefficient associé à la variable *male* soit si imprécis ? Devrions-nous conclure qu'il n'existe aucune différence de score entre les hommes et les femmes après avoir tenu compte de l'influence de *colgpa* ? [Indice : Vous pourriez envisager de calculer une statistique de Fisher pour tester l'hypothèse nulle d'absence de différence entre les hommes et les femmes en introduisant une variable d'interaction.]

iii. Comparativement aux résultats de la troisième équation, comment se fait-il que le coefficient associé à la variable *male* de la dernière équation apparaisse si proche de celui estimé à la deuxième équation tant en termes d'ordre de grandeur que de précision ?

EXERCICES SUR ORDINATEUR

C1. Pour cet exercice, nous utilisons les données contenues dans GPA1.

i. Ajoutez les variables *mothcoll* et *fathcoll* à l'équation (7.6) et reportez les résultats d'estimation. Qu'observez-vous pour le coefficient de la variable relative à la possession d'un PC ? *PC* apparaît-il toujours statistiquement significatif ?

ii. Testez pour la significativité jointe des coefficients associés aux variables *mothcoll* et *fathcoll* issues de l'équation en (i) en mentionnant les *p*-valeurs.

iii. Ajoutez $hsGPA^2$ au modèle décrit en (i), jugez-vous cette généralisation pertinente ?

C2. Pour cet exercice, nous utilisons les données contenues dans WAGE2.

i. Estimez le modèle suivant :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u$$

et reportez les résultats. Quelle est la différence approximative de salaire mensuel entre les africains américains et les non africains américains, toutes choses égales par ailleurs ? Cette différence est-elle statistiquement significative ?

ii. Ajoutez les variables $exper^2$ et $tenure^2$ à l'équation précédente et montrez que les coefficients associés ne sont pas conjointement statistiquement significatifs au seuil de 20 %.

iii. Amendez le modèle original pour permettre aux rendements de l'éducation de dépendre des origines ethniques et testez si effectivement les rendements de l'éducation en dépendent.

iv. À nouveau, partez du modèle de base et amendez le ensuite en autorisant les salaires à différer selon les groupes : mariés et africains américains, mariés et non africains américains, non mariés et africains américains, non mariés et non africains américains. Quelle est la différence de salaire estimée entre les mariés africains américains et non africains américains ?

C3. Un modèle permettant de caractériser les salaires des joueurs de la ligue majeure de baseball aux États-Unis est donné par :

$$\begin{aligned}\log(salary) = {} & \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr \\ & + \beta_5 rbisyr + \beta_6 runsyr + \beta_7 fldperc + \beta_8 allstar \\ & + \beta_9 frstbase + \beta_{10} scndbase + \beta_{11} thrdbase + \beta_{12} shrtstop \\ & + \beta_{13} catcher + u,\end{aligned}$$

avec *outfield* (joueur de champ extérieur) le groupe de référence.

i. Établissez formellement l'hypothèse nulle selon laquelle, toutes choses égales par ailleurs, attrapeurs (*catcher*) et joueurs de champ extérieur gagnent en moyenne le même revenu. Testez ensuite cette hypothèse à partir des données contenues dans la base MLB1 et commentez l'étendue du différentiel de salaire.

ii. Établissez formellement et testez l'hypothèse selon laquelle il n'y a aucune différence de salaire moyen selon les postes occupés, une fois l'ensemble des variables de contrôle prises en compte.

iii. Les résultats tirés des questions (i) et (ii) sont-ils cohérents ? Dans le cas contraire, expliquez le mécanisme sous-jacent.

C4. Nous utilisons maintenant les données contenues dans GPA2.

i. On considère le modèle suivant :

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u,$$

avec *colgpa* la moyenne des notes obtenues en premier cycle à l'université, *hsize* la taille de la classe au lycée, en centaines, *hsperc* le quantile dans lequel se situe l'étudiant à l'issue de son parcours universitaire, *sat* le score SAT, *female* une variable binaire relative au genre, et *athlete* une variable binaire prenant la valeur un si l'étudiant est un athlète. Quelles sont vos attentes relativement aux valeurs des différents coefficients du modèle ? Quels sont ceux pour lesquels vous avez des doutes ?

ii. Estimez l'équation mentionnée en question (i) et reportez les résultats de façon standard. Quel est le différentiel de GPA estimé entre les athlètes et les non athlètes ? Cette différence est-elle statistiquement significative ?

iii. Enlevez *sat* du modèle et ré-estimez l'équation. Quel est maintenant l'effet estimé du statut d'athlète ? Discutez les raisons pour lesquelles ces résultats diffèrent de ceux présentées à la question (ii).

iv. Dans le modèle décrit à la question (i), autorisez l'impact du statut d'athlète à différer selon le genre et testez l'hypothèse nulle d'absence de différence entre les femmes et les hommes athlètes, toutes choses égales par ailleurs.

v. L'effet de *sat* sur *colgpa* diffère-t-il selon le genre ? Justifiez votre réponse.

C5. Dans l'exercice 2 du chapitre 4, nous avions ajouté les rendements du cours boursier de l'entreprise, *ros*, au modèle expliquant le salaire du PDG ; celle-ci s'avérant par la suite non significative. Définissons maintenant une variable indicatrice *rosneg*, égale à un si ros < 0 et zéro si $ros \geq 0$. À partir des données de CEOSAL1 estimez le modèle :

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 rosneg + u.$$

Interprétez la valeur estimée et la significativité de β_3

C6. Cet exercice repose sur les données contenues dans SLEEP75. Le modèle d'intérêt est le suivant :

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u.$$

i. Estimez l'équation séparément pour les hommes et les femmes et reportez vos résultats de façon standard. Observez-vous des différences notables entre les deux équations estimées ?

ii. Calculez la statistique du test de Chow pour tester l'égalité des paramètres de l'équation expliquant le temps de sommeil des hommes et des femmes. Utilisez la forme du test qui repose sur l'introduction de la variable *male* ainsi que des termes d'interaction *male·totwrk*, …, *male·yngkid* en utilisant l'intégralité des observations à disposition. Quels sont les degrés de liberté *ddl* associés à la statistique de test ? Rejetez-vous l'hypothèse nulle au seuil de 5 % ?

iii. Laissez maintenant la possibilité de distinguer les constantes pour les groupes des hommes et des femmes et déterminez si les termes d'interaction impliquant *male* sont conjointement significatifs.

iv. Compte-tenu des résultats mentionnés aux questions (ii) et (iii), quel serait le modèle finalement retenu ?

C7. Nous nous référons aux données contenues dans le fichier WAGE1 pour cet exercice.

i. Utilisez l'équation (7.18) pour estimer le différentiel de salaire entre hommes et femmes lorsque le niveau d'éducation est fixé à *educ* = 12,5. Comparez ces résultats lorsque *educ* = 0.

ii. Estimez le modèle permettant d'obtenir les résultats rapportés dans l'équation (7.18), avec *female*·(*educ* – 12,5) à la place de *female*·*educ*. Comment interprétez-vous le coefficient associé à la variable *female* ?

iii. Le coefficient associé à *female* de la question (ii) est-il statistiquement significatif ? Comparez vos résultats avec ceux mentionnés dans l'équation (7.18) et interprétez.

C8. Nous considérons les données contenues dans LOANAPP pour cet exercice. La variable binaire à expliquer *approve*, est égale à un si un prêt hypothécaire est accordé à l'individu. La variable explicative clé dans notre modèle est *white*, une variable indicatrice valant un si le candidat à l'emprunt est de type caucasien. Les autres candidats de la base de données sont soit africains américains soit d'origine hispanique.

Pour tester l'existence de discrimination à l'encontre des minorités sur le marché du crédit hypothécaire, un modèle à probabilités linéaires peut être estimé :

$$approve = \beta_0 + \beta_1 white + autres\ facteurs.$$

i. En présence de discrimination à l'encontre des minorités et en supposant que l'ensemble des variables de contrôle pertinentes ont été prises en compte dans le modèle, quel est le signe attendu du paramètre β_1 ?

ii. Régressez *approve* sur *white* et reportez les résultats obtenus sous une forme usuelle. Interprétez le coefficient associé à la variable *white*. Est-il statistiquement significatif ? Est-il économiquement significatif ?

iii. Ajoutez les variables de contrôle *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2*, et *vr*. Qu'advient-il du coefficient associé à la variable *white* ? Existe-t-il encore des preuves de discrimination à l'encontre des minorités ?

iv. Introduisez maintenant des variables d'interaction impliquant les origines ethniques et la variable *obrat* mesurant les autres engagements en pourcentage du revenu. Ce terme d'interaction apparaît-il significatif ?

v. À partir du modèle décrit à la question (iv), mesurer l'impact de l'origine ethnique sur la probabilité d'obtention du crédit hypothécaire lorsque *obrat* = 32, soit approximativement la valeur moyenne sur l'ensemble de l'échantillon considéré. Calculez un intervalle de confiance à 95 % pour cette mesure.

C9. La question de savoir si la mise à disposition du système d'épargne par capitalisation nommé « plan 401(k) » auprès de nombreux travailleurs américains accroîtrait l'épargne net, a suscité un vif intérêt. La base de données 401KSUBS contient de l'information relative aux actifs financiers nets (*nettfa*), revenu familial (*inc*), une variable binaire mentionnant l'éventuelle éligibilité pour un plan 401(k) (*e401k*), ainsi que d'autres variables.

i. Parmi l'ensemble des familles considérées dans notre échantillon, quelle proportion d'entre elles est éligible pour participer au plan 401(k) ?

ii. Estimez un modèle à probabilités linéaires expliquant l'éligibilité au plan 401(k) en fonction du revenu, de l'âge et du genre. Introduisez ensuite le revenu et l'âge sous formes quadratiques et reportez les résultats de façon standard.

iii. Diriez-vous que l'éligibilité au plan 401(k) est indépendante du revenu et de l'âge ? Et du genre ? Justifiez.

iv. Calculez les valeurs prédites du modèle à probabilités linéaires estimées à la question (ii). Certaines valeurs prédites sont-elles négatives ou supérieures à un ?

v. Utilisez les valeurs prédites $\widehat{e401k}_i$ obtenues en (iv), définissez ensuite $\widehat{e401k}_i = 1$ si $\widehat{e401k}_i \geq 0{,}5$ et $\widehat{e401k}_i = 0$ si $\widehat{e401k}_i < 0{,}5$. Parmi les 9,275 familles, combien selon le modèle, sont caractérisées comme éligible pour un plan 401(k) ?

vi. Parmi les 5 638 familles non éligibles pour un plan 401(k), et à partir de la règle mentionnée supra $\widehat{e401k}_i$, quel pourcentage d'entre elles avaient été (correctement) caractérisées par le modèle comme non éligible à un plan 401(k). Parmi les 3 637 familles éligibles pour un plan 401(k), combien d'entre elles avaient été correctement caractérisées par le modèle comme éligible à un tel programme. (Il est utile d'avoir recours ici à la commande « *tabulate* » de votre logiciel économétrique.)

vii. Le pourcentage total de valeurs correctement prédites est d'environ 64,9 %. Pensez-vous que cela constitue une description complète de la bonne performance du modèle compte-tenu de votre réponse en (vi) ?

viii. Ajoutez la variable *pira* comme variable explicative de votre modèle à probabilités linéaires. Toutes choses égales par ailleurs, si un des membres d'une famille dispose d'un compte épargne retraite, de combien la probabilité d'éligibilité de la famille augmente-t-elle pour un plan 401(k) ? Cet effet est-il statistiquement différent de zéro au seuil de 10 % ?

C10. On utilise les données contenues dans NBASAL pour cet exercice.

i. Estimez un modèle de régression linéaire expliquant les points obtenus par match en fonction de l'expérience au sein de la ligue et la position (défenseur, ailier ou pivot). Introduisez l'expérience sous forme quadratique et utilisez le centre comme catégorie de référence. Reportez les résultats sous une forme usuelle.

ii. Pourquoi ne pas introduire les trois positions comme variables indicatrices dans l'équation décrite en (i) ?

iii. À niveau d'expérience fixé, un défenseur marque-t-il plus qu'un ailier ? De combien ? Cette différence est-elle statistiquement significative ?

iv. Ajoutons maintenant le statut marital à l'équation. À positions et niveaux d'expérience fixés, les joueurs mariés sont-ils plus productifs (sur la base des points par match) ?

v. Ajoutons maintenant des variables d'interaction impliquant le statut marital avec chacune des deux variables relatives à l'expérience. Dans ce modèle étendu, parvient-on à montrer de façon claire que le statut marital influence le nombre de points par match ?

vi. Estimez le modèle à partir de l'équation (iv) en remplaçant la variable dépendante par le nombre de passes décisives par match. Observe-t-on des différences notables comparativement aux résultats obtenus en (iv) ? Etayez votre réponse.

C11. On utilise les données contenues dans 401KSUBS pour cet exercice.

i. Calculez la moyenne, l'écart-type, les valeurs minimum, et maximum de la variable *nettfa* pour l'échantillon considéré.

ii. Testez l'hypothèse nulle selon laquelle la valeur moyenne de *nettfa* reste inchangée quelque soit le statut d'éligibilité pour un plan 401(k) ; utilisez un test d'hypothèses bilatéral. Quelle est en dollar, la différence estimée ?

iii. À partir des éléments en (ii) de l'exercice sur ordinateur C9, il apparaît clairement que *e401k* n'est pas exogène dans le modèle de régression ; au mieux, sa valeur change en fonction du revenu et de l'âge. Estimez un modèle de régression multiple pour *nettfa* incluant le revenu, l'âge et *e401k* comme variables explicatives. Les variables relatives au revenu et l'âge doivent être introduites sous forme quadratique. Réévaluez maintenant l'effet en dollars de l'éligibilité au plan 401(k).

iv. Au modèle estimé en question (iii), ajoutez les variables d'interaction $e401k \cdot (age - 41)$ et $e401k \cdot (age - 41)^2$. Notez que l'âge moyen sur l'échantillon est évalué à 41 de sorte que dans la nouvelle

version du modèle, le coefficient associé à *e401k* correspond à l'effet estimé de l'éligibilité au plan 401(k) à l'âge moyen. Quel terme d'interaction apparaît significatif ?

v. Comparez les estimations des questions (iii) et (iv), les effets estimés de l'éligibilité au plan 401(k) diffèrent-ils grandement ? Justifiez.

vi. Retirez maintenant les variables d'interaction du modèle et définissez cinq variables indicatrices relatives à la taille de la famille soient : *fsize1*, *fsize2*, *fsize3*, *fsize4*, et *fsize5*. La variable *fsize5* vaut 1 pour des familles comportant cinq membres ou plus. Introduisez les variables indicatrices dans le modèle proposé à la question (iii) ; en prenant soin d'identifier un groupe de référence. Les variables indicatrices sont-elles significatives au seuil de 1 % ?

vii. Réalisez maintenant un test de Chow pour le modèle suivant :

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc1 + \beta_3 age + \beta_4 age^2 + \beta_5 e401k + u$$

pour les cinq catégories précédemment mentionnées. La somme des carrés des résidus du modèle contraint, SCR_c, est obtenue à l'issue de l'estimation du modèle décrit à la question (vi) puisque ce modèle fait l'hypothèse que l'ensemble des pentes sont identiques. La somme des carrés des résidus du modèle non contraint est donnée par $SCR_{nc} = SCR_1 + SCR_2 + \ldots + SCR_5$, avec SCR_f la somme des carrés des résidus estimés de l'équation relative pour la taille de famille *f*. Prenez conscience du fait que 30 paramètres sont à estimer dans le modèle non contraint (5 constantes plus 25 coefficients de pente) et 10 dans le modèle contraint (5 constantes plus 5 pentes). De ce fait le nombre de restrictions à tester est de $q = 20$, et le nombre de degrés de liberté associés *ddl* pour le modèle non contraint $9275 - 30 = 9245$.

C12. Utilisez les données contenues dans la base BEAUTY. Celle-ci comprend un sous-ensemble des variables utilisées par Hamermesh and Biddle (1994) (mais un plus grand nombre d'observations exploitables que dans les exercices de régressions menés par les auteurs).

i. Identifiez les proportions d'hommes et de femmes catégorisés comme ayant un physique jugé supérieur à la moyenne. Y-a-t-il dans l'échantillon considéré plus de personnes dont le physique est jugé supérieur ou inférieur à la moyenne ?

ii. Testez l'hypothèse nulle selon laquelle les proportions de populations masculines et féminines jugées au-dessus de la moyenne sont identiques. Reportez les *p*-valeurs associées au test unilatéral de l'hypothèse nulle selon laquelle cette proportion est plus élevée pour les femmes. (*Astuce* : L'approche la plus simple serait d'estimer un modèle à probabilités linéaires.)

iii. Estimez le modèle suivant :

$$\log(wage) = \beta_0 + \beta_1 belavg + \beta_2 abvavg + u$$

séparément pour les hommes puis les femmes et reportez les résultats sous la forme usuelle. Dans les deux cas, interprétez la valeur du coefficient associé à la variable *belavg*. Expliquez avec vos propres mots la signification de l'hypothèse de test $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 < 0$, et déterminez les *p*-valeurs associées pour les groupes des hommes puis des femmes.

iv. Vos résultats confirment-ils que les femmes avec un physique jugé au-delà de la moyenne gagnent en moyenne plus que les femmes au physique jugé dans la moyenne ? Etayez votre réponse.

v. Ajoutez pour les deux groupes, les variables explicatives suivantes *educ*, *exper*, $exper^2$, *union*, *goodhlth*, *black*, *married*, *south*, *bigcity*, *smllcity*, et *service*. Les effets des variables associées au physique changent-ils dans des proportions importantes ?

vi. Utilisez la formulation de la somme des carrés des résidus (SCR) de l'expression de la statistique du test de Chow et testez si les pentes des régressions de la question (v) diffèrent entre hommes et femmes. Assurez-vous d'autoriser une différence au niveau de la constante sous l'hypothèse nulle.

C13. On utilise les données contenues dans le fichier APPLE pour répondre à cette question.

i. Définissez une variable binaire *ecobuy* = 1 si *ecolbs* > 0 et *ecobuy* = 0 si *ecolbs* = 0. En d'autres termes, *ecobuy* indique si, à prix donnés, une famille consomme des pommes issues de l'agriculture biologique. Combien de familles affirment acheter des pommes labélisées bio ?

ii. Estimez le modèle à probabilités linéaires suivant :

$$ecobuy = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc + \beta_4 hhsize + \beta_5 educ + \beta_6 age + u,$$

et reportez les résultats sous une forme usuelle. Interprétez avec soin les coefficients associés aux variables de prix.

iii. Y-a-t-il des variables non associées aux prix qui soient conjointement significatives dans le modèle ? (Utilisez les statistiques de Fisher usuelles même si celles-ci ne sont pas valides en présence d'hétéroscédasticité.) Quelle variable explicative autre que les variables de prix semblent avoir l'effet le plus important sur la décision de consommer des pommes issues de l'agriculture biologique ? Cela vous paraît-il avoir du sens ?

iv. À partir du modèle discuté dans la question (ii), remplacez *faminc* par log(*faminc*). Lequel des deux modèles vous semble le mieux expliquer vos données celui introduisant *faminc* ou log(*faminc*) ? Interprétez le coefficient associé à log(*faminc*).

v. Sur base des estimations réalisées en question (iv), combien des probabilités estimées s'avèrent négatives ? Combien s'avèrent supérieures à l'unité ? En quoi cela devrait-il vous interpeler ?

vi. Revenons à l'estimation de la question (iv), calculez le pourcentage de prédictions correctes pour chacune des valeurs possible de la variable indépendante c'est-à-dire *ecobuy* = 0 puis *ecobuy* = 1. Quel résultat s'avère le mieux prédit par le modèle ?

C14. Utilisez les données contenues dans le fichier CHARITY pour répondre à cette question. La variable *respond* est une variable indicatrice égale à un si un individu a répondu par un don à la sollicitation par mail la plus récente d'une association caritative. La variable *resplast* est une variable indicatrice égale à un si un individu a répondu à la sollicitation par mail précédente, *avggift* est la moyenne des dons passés (en florins néerlandais) et *propresp* est le temps relatif passé par la personne à répondre aux sollicitations passées.

i. Estimez un modèle à probabilités linéaires expliquant *respond* en fonction de *resplast* et *avggift*. Reportez les résultats sous une forme usuelle et interprétez le coefficient de *resplast*.

ii. La valeur moyenne des dons passés semble-t-elle affecter la probabilité de réponse ?

iii. Ajoutez la variable *propresp* au modèle et interprétez son coefficient. (Soyez attentif ici : une hausse de une unité de *propresp* est le changement le plus important possible.)

iv. Qu'est-il advenu du coefficient de *resplast* lorsque *propresp* a été ajouté au modèle de régression ? Cela a-t-il du sens ?

v. Ajoutez *mailsyear*, le nombre de mails envoyés par an, au modèle. Evaluez l'effet associé ? Pourquoi n'est-ce sans doute pas une mesure correcte de l'effet causal de l'envoi ciblé de mails (ou « publipostage ») sur la probabilité de réponse ?

C15. Utilisez les données contenues dans FERTIL2 pour répondre à cette question.

i. Identifiez les valeurs minimale et maximale de la variable *children* dans l'échantillon. Quelle est la valeur moyenne associée à cette variable ? Est-il possible que certaines des femmes aient exactement le nombre d'enfants moyen ?

ii. Parmi les femmes de l'échantillon, combien sont celles possédant l'électricité à domicile ?

iii. Calculez le nombre moyen d'enfants (*children*) parmi ceux ne bénéficiant pas de l'électricité, faites de même pour ceux bénéficiant de l'électricité. Commentez vos résultats. Testez la différence de valeurs moyenne au moyen d'une régression simple.

iv. À partir des éléments de la question (iii), pouvez-vous conclure que le fait de bénéficier de l'électricité « cause » une fertilité moindre des femmes ? Justifiez.

v. Estimez un modèle de régression multiple dans la veine de celui explicité à l'équation (7.37), en ajoutant age^2, *urban*, ainsi que trois variables indicatrices renseignant l'obédience religieuse. Réévaluez les effets de l'électricité et comparez vos résultats avec ceux obtenus dans la question (iii) ? L'effet est-il statistiquement significatif ?

vi. À partir de l'équation en (v), ajoutez une variable d'interaction entre *electric* et *educ*. Les coefficients associés sont-ils statistiquement significatifs ? Qu'advient-il du coefficient associé à la variable *electric* ?

vii. La valeur médiane et mode pour la variable *educ* est de 7. À partir de l'équation en (vi), utilisez le terme d'interaction centré $electric \cdot (educ - 7)$ à la place de $electric \cdot educ$. Qu'advient-il du coefficient de *electric* comparativement à celui obtenu en (vi) ? Pourquoi ? Comment se comporte le coefficient de *electric* comparativement à celui obtenu en (v) ?

C16. À partir de données de la base CATHOLIC répondez aux questions suivantes.

i. Considérez l'échantillon complet et identifiez quel pourcentage d'étudiants sont régulièrement inscrit dans un établissement secondaire d'obédience catholique. Quelle est la moyenne de la variable *math12* ?

ii. Réalisez une régression simple de *math12* sur *cathhs* et reportez les résultats selon les normes usuelles. Interprétez vos résultats.

iii. Ajoutez maintenant les variables *lfaminc*, *motheduc*, et *fatheduc* au modèle de régression précédent. De combien d'observations disposez-vous pour cette régression ? Qu'advient-il du coefficient de la variable *cathhs*, apparaît-il significatif ?

iv. Reprenez le cas d'une régression simple de *math12* sur *cathhs*, en restreignant les observations à celles utilisées dans la régression multiple de la question (iii). Certaines de vos conclusions en sont-elles modifiées ?

v. À partir du modèle de régression étudié en question (iii), ajoutez des variables d'interaction entre *cathhs* et chacune des autres variables explicatives du modèle. Ces variables d'interaction sont-elles individuellement et/ou conjointement significatives ?

vi. Qu'advient-il du coefficient de *cathhs* dans la régression de la question (v). Expliquez pourquoi ce coefficient n'est pas très intéressant à étudier.

vii. Calculez l'effet marginal moyen de *cathhs* dans le model estimé en question (v). Comparez les résultats obtenus avec les coefficients de *cathhs* estimés dans la parties (iii) et (v).

CHAPITRE 8

HÉTÉROSCÉDASTICITÉ

Traduction de Jean-Yves Gnabo

L'hypothèse d'homoscédasticité introduite au chapitre 3 dans le cadre du modèle de régression linéaire multiple (régression multiple), suppose que la variance des erreurs non observées, u, conditionnellement aux variables explicatives, est constante. Cette hypothèse n'est plus valide dès lors que la variance des facteurs non observés contenus dans u, diffère selon l'appartenance à tel ou tel segment de la population, ceux-ci étant déterminés par les différentes réalisations des variables explicatives. Dans le cas d'une équation d'épargne par exemple, nous serons en présence d'hétéroscédasticité si la variance des facteurs non observés affectant l'épargne, contenus dans le terme d'erreur, augmente avec le revenu.

Dans les chapitres 4 et 5, nous avons eu recours à l'hypothèse d'homoscédasticité dans le cadre du modèle de régression multiple, y compris en présence de grands échantillons [dans le cadre asymptotique], pour dériver des tests de Student, des tests de Fisher, ou encore des intervalles de confiance qui soient valides, l'objectif étant de procéder à de l'inférence statistique sur les estimateurs issus de la méthode des moindres carrés ordinaires (MCO). Que faire maintenant si cette hypothèse n'est pas respectée ? Dans ce chapitre, nous examinons les possibles solutions à adopter en présence d'hétéroscédasticité. Nous discutons également des tests à réaliser pour la détecter. Avant d'aborder ces questions, nous passons en revue brièvement les conséquences de l'hétéroscédasticité sur les propriétés des estimateurs des MCO.

8.1 CONSÉQUENCES DE L'HÉTÉROSCÉDASTICITÉ POUR LES MCO

Considérons à nouveau le modèle de régression multiple :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \qquad \textbf{[8.1]}$$

Dans le chapitre 3, nous avons prouvé que les estimateurs des paramètres du modèle par la méthode des MCO, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ sont sans biais sous les quatre premières hypothèses de Gauss-Markov, de RLM.1 à RLM.4. Dans le chapitre 5, ce résultat a été étendu à la convergence des estimateurs sous ces mêmes hypothèses. L'hypothèse additionnelle, RLM.5, relative à l'homoscédasticité suppose que la variance conditionnelle des termes d'erreur est constante. Formellement, on l'écrit de la façon suivante : $\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$. Ainsi, contrairement à certains problèmes de spécification du modèle, tel l'omission d'une variable importante, la présence d'hétéroscédasticité n'introduit pas de biais dans les estimateurs des paramètres β_j obtenus par les MCO. Les estimateurs conservent également leur propriété de convergence.[1]

De même, la présence d'hétéroscédasticité n'a pas d'incidence sur la qualité d'ajustement du modèle, telle que mesurée par le coefficient de détermination, R^2 ou sa version ajustée, $\bar{R}^2$. Comment peut-on expliquer ce résultat ? Pour rappel (voir section 6.3), les mesures du R carré et du R carré ajusté, calculés à partir de l'échantillon, sont différentes manières d'estimer le R carré de la population. Ce dernier peut également s'écrire $1 - \sigma_u^2/\sigma_y^2$, où σ_u^2 et σ_y^2 désignent respectivement la variance du terme d'erreur et celle de la variable dépendante, y dans la population. Dans la mesure où ces deux variances sont des mesures inconditionnelles, nos indicateurs de qualité du modèle ne sont donc pas affectés par la présence d'hétéroscédasticité dans $\text{Var}(u|x_1, x_2, \ldots, x_k)$. De plus, SCR/$n$ et SCT/n demeurent des estimateurs convergents des paramètres de variance σ_u^2 et σ_y^2, indépendamment du fait que $\text{Var}(u|x_1, x_2, \ldots, x_k)$ soit constante ou non. Ces propriétés restent valables si l'on tient compte de l'ajustement lié aux degrés de liberté. Par conséquent, que l'hypothèse d'homoscédasticité soit vérifiée ou non, le coefficient de détermination, R^2, et le coefficient de détermination ajusté, $\bar{R}^2$, demeurent tous deux des estimateurs convergents du R carré.

1 Lorsque le terme de « convergence » est utilisé dans le texte, nous ne précisons pas à chaque fois qu'il s'agit d'une convergence vers la valeur vraie du paramètre. En anglais, les économètres parlent de « consistency » dont la traduction littérale en français est « consistance ».

Si le relâchement de l'hypothèse d'homoscédasticité n'occasionne pas de biais d'estimation ou de non convergence des estimateurs des MCO, on peut légitimement se demander pour quelles raisons l'avoir introduite parmi les hypothèses de Gauss-Markov. Un premier élément de réponse tient à son rôle crucial dans l'établissement des propriétés relatives à l'estimateur de la *variance* des paramètres estimés, $\text{Var}(\hat{\beta}_j)$, puisque celui-ci apparaît biaisé en l'absence d'homoscédasticité (voir chapitre 3). Dans la mesure où les écarts-types estimés des estimateurs des MCO découlent directement de ces mesures de variances, les valeurs obtenues pour les intervalles de confiance, de même que celles des statistiques t, ne sont plus valides.[2] Plus spécifiquement, en présence d'hétéroscédasticité, les statistiques de Student habituelles issues de l'application des MCO ne suivent plus une distribution de Student, même en présence de grands échantillons. Nous abordons ce point dans le cadre des régressions linéaire simples (régression simple) à la section suivante. Cette section sera également l'occasion de dériver la variance des estimateurs des MCO des coefficients de pente sous l'hypothèse d'hétéroscédasticité et de proposer une approche alternative valide. De façon similaire, lorsque l'hypothèse d'homoscédasticité des erreurs est relâchée, nous verrons que la statistique F ne suit plus de distribution de Fisher et que celle du multiplicateur de Lagrange (LM) ne suit plus une distribution asymptotique du chi-deux. En conséquence, c'est l'ensemble des statistiques dont nous avions l'usage pour l'inférence statistique standard dans le cadre des hypothèses de Gauss-Markov, qui ne sont plus valables en présence d'hétéroscédasticité.

Qu'en est-il de l'efficience de l'estimateur des MCO ? D'après le théorème de Gauss-Markov, nous savons que, sous les hypothèses du même nom (notamment celle d'homoscédasticité des erreurs), l'estimateur des MCO est le meilleur estimateur linéaire sans biais, soit l'estimateur *BLUE*. Si nous relâchons cette hypothèse et autorisons $\text{Var}(u|\mathbf{x})$ à ne plus être constante, cette propriété d'efficience disparaît aussi bien en échantillon fini (l'estimateur des MCO n'est plus *BLUE*) qu'au niveau asymptotique (l'estimateur n'est plus asymptotiquement efficace dans la classe des estimateurs décrite par le théorème 5.3). Comme nous le verrons dans la section 8.4, il est possible de trouver des estimateurs alternatifs plus efficients que l'estimateur des MCO en présence d'hétéroscédasticité (même si ceci implique de connaître la forme de l'hétéroscédasticité). En grand échantillon, néanmoins le problème relatif à l'efficience devient moins aigu. Dans la section suivante, nous montrons également comment modifier les statistiques habituelles utilisées dans le cadre des MCO afin d'obtenir des statistiques valides en présence d'hétéroscédasticité, au moins asymptotiquement.

8.2 INFÉRENCE ROBUSTE À L'HÉTÉROSCÉDASTICITÉ APRÈS ESTIMATION PAR LES MCO

La mise en œuvre de tests d'hypothèses joue un rôle crucial dans l'analyse économétrique. Dès lors, si les estimateurs des MCO et les statistiques dérivées pour l'inférence statistique ne sont pas valides en présence d'hétéroscédasticité, pourquoi devrions-nous continuer à les utiliser ? Fort heureusement, au cours de ces 20 dernières années, de nombreuses méthodes ont visé à corriger les écarts-types estimés, ainsi que les statistiques traditionnelles des test de Student, de Fisher et du multiplicateur de Lagrange, dans le but ultime de les rendre valides en présence d'**hétéroscédasticité de forme inconnue**, en grand échantillon tout au moins. Ces approches sont naturellement très utiles puisqu'elles préservent la validité des tests d'inférence statistique, quelle que soit la nature de l'hétéroscédasticité présente dans la population. On parle généralement de statistiques *robustes à l'hétéroscédasticité*. (Dans la suite de ce chapitre, la robustesse désignera par défaut la robustesse à l'hétéroscédasticité).

2 Par souci de concision, nous emploierons souvent la terminologie « **écarts-types** » pour désigner les **écarts-types estimés** sur un échantillon (les « *standard errors* » en anglais) en opposition aux **écarts-types sur la population** (les « *standard deviations* » en anglais) que nous désignerons explicitement de cette manière à chaque fois pour éviter toute ambiguïté.

Commençons par présenter, dans les grandes lignes, la démarche qui nous permet d'estimer correctement la variance, $\text{Var}(\hat{\beta}_j)$, en présence d'hétéroscédasticité. Cette section n'a pas pour ambition de proposer une dérivation complète des résultats théoriques, ce qui irait bien au-delà de l'objectif de ce chapitre. Elle vise plutôt à mettre en lumière les éléments les plus importants. En pratique, l'application des méthodes robustes est simple car la plupart des logiciels économétriques proposent des options qui y sont dédiées.

Par souci de clarté, nous partons du modèle de régression simple. L'indice i indique que les réalisations des valeurs de y, x et u sont celles de l'individu i :

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

Dans ce chapitre, nous allons toujours considérer que les quatre premières hypothèses de Gauss-Markov sont vérifiées. Si les erreurs sont hétéroscédastiques, alors

$$\text{Var}(u_i|x_i) = \sigma_i^2,$$

où l'indice i de σ_i^2 indique que la variance de l'erreur dépend de la réalisation particulière de x_i.

Nous avons vu auparavant que l'estimateur MCO peut s'écrire comme suit

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Sur base des hypothèses RLM.1 à RLM.4 (sans l'hypothèse d'homoscédasticité) et en raisonnant conditionnellement aux valeurs prises par x_i dans l'échantillon, nous pouvons utiliser les mêmes arguments que ceux développés dans le chapitre 2 pour montrer que :

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_i^2}{SCT_x^2}, \qquad \textbf{[8.2]}$$

où $SCT_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ est la somme des carrés totaux des x_i. Dans le cas d'une régression simple, on constate tout de suite que, si $\sigma_i^2 = \sigma^2$ pour tout i. Cette expression se ramène à la formule habituelle, σ^2/SCT_x. Dans la situation contraire où les erreurs sont hétéroscédastiques, l'équation (8.2) montre également que la formule de la variance dérivée sous l'hypothèse d'homoscédasticité n'est plus valide.

Dans la mesure où l'écart-type de, $\hat{\beta}_1$, repose directement sur l'estimation de $\text{Var}(\hat{\beta}_1)$, nous devons trouver un moyen d'estimer l'équation (8.2) correctement en présence d'hétéroscédasticité. Une proposition dans ce sens a été faite par White (1980) qui a développé un estimateur de $\text{Var}(\hat{\beta}_1)$ valide *quelle que soit* la forme de l'hétéroscédasticité présente dans les erreurs, et même lorsqu'il n'y en a pas (dans le cas d'homoscédasticité). Pour le vérifier, considérons le résidu, $\hat{u}_i$, issu de la régression par les MCO de y sur x. L'estimateur s'écrit :

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{u}_i^2}{SCT_x^2}, \qquad \textbf{[8.3]}$$

Cet estimateur peut être aisément calculé après l'application de la méthode des MCO sur les données de l'échantillon.

Avant d'aller plus loin, demandons-nous pour quelle raison (8.3) fournit un estimateur valide de $\text{Var}(\hat{\beta}_1)$. La réponse n'est pas simple et ne sera pas entièrement développée ici. Nous soulignerons simplement qu'il est possible de montrer que lorsque l'équation (8.3) est multipliée par la taille n de l'échantillon, l'expression converge en probabilité vers $\text{E}[(x_i - \mu_x)^2 u_i^2]/(\sigma_x^2)^2$, qui est la limite en probabilité de n multiplié par (8.2). Ce résultat, nous permet ainsi d'obtenir un estimateur utile pour construire les intervalles de confiance et les statistiques de Student. On notera aussi que la loi des grands nombres et le théorème central limite jouent un rôle clé dans l'établissement de ces résultats de convergence. Pour plus de détails sur cet estimateur, on peut se référer à la publication originale de White, en gardant à l'esprit toutefois que ce document reste assez technique. Voir aussi Wooldridge (2010, chapitre 4).

Une formule similaire peut s'appliquer au cas plus général de la régression multiple :

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u.$$

Il est possible de montrer que, sous les hypothèses RLM.1 à RLM.4, un estimateur valide de $\text{Var}(\hat{\beta}_j)$, s'écrit :

$$\widehat{\text{Var}}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SCR_j^2}, \qquad \textbf{[8.4]}$$

où $\hat{r}_{ij}$ est le résidu, pour la $i^{\text{ème}}$ observation, issu de la régression auxiliaire de x_j sur les autres variables indépendantes et SCR_j, la somme des carrés des résidus de cette même régression (voir la section 3.2 pour la représentation des estimateurs MCO en termes d'effets *nets* ou *purgés*). La racine carrée de l'expression (8.4) est **appelée écart-type estimé robuste à l'hétéroscédasticité** de $\hat{\beta}_j$. En économétrie, ces écarts-types sont généralement attribués à White (1980). Des travaux antérieurs en statistiques, notamment ceux de Eicker (1967) et Huber (1967), avaient toutefois déjà évoqué la possibilité d'obtenir des estimateurs robustes. Dans les travaux appliqués, ceux-ci sont parfois appelés *écarts-types de White, Huber, ou Eicker* (ou une combinaison de ces noms avec un trait d'union). Par la suite, nous les appellerons *écarts-types robustes à l'hétéroscédasticité*, voire simplement *écarts-types robustes* lorsque le contexte n'est pas équivoque.

Il arrive parfois de corriger (8.4) pour tenir compte du nombre de degrés de liberté en multipliant l'expression par $n/(n - k - 1)$ avant de calculer la racine carrée. Cet ajustement permet de retrouver les valeurs des écarts-types habituels des MCO lorsque les résidus $\hat{u}_i^2$ sont identiques pour toutes les observations i (correspondant à la forme la plus forte d'homoscédasticité). D'autres variantes de (8.4) sont examinées dans MacKinnon et White (1985). Toutes ces approches sont équivalentes lorsque la taille de l'échantillon tend vers l'infini ; elles ne sont d'ailleurs valides que sur le plan asymptotique. Par conséquent, aucun estimateur robuste des écarts-types n'est unanimement préféré aux autres. En pratique, les écarts-types robustes sont calculés en fonction des options proposées par le logiciel économétrique.

Après avoir calculé l'écart-type robuste à l'hétéroscédasticité, il suffit de remplacer l'écart-type standard des MCO par sa version robuste dans (8.5) pour construire une **statistique de Student robuste à l'hétéroscédasticité**. Rappelons la forme générale de la statistique de Student :

$$t = \frac{\textit{estimation} - \textit{valeur hypothétique}(\textit{sous}\, H_0)}{\textit{écart-type estimé}} \qquad \textbf{[8.5]}$$

Étant donné que nous utilisons toujours les estimateurs des MCO et que la même valeur hypothétique est fixée, la seule différence entre la statistique de Student habituelle et la statistique de Student robuste réside dans le calcul de la valeur située au dénominateur de l'expression. Revenons maintenant brièvement sur les facteurs susceptibles d'influencer les valeurs des écarts-types robustes. Le terme, SCR_j, dans l'équation (8.4) peut être remplacé par $\text{SCT}_j (1 - R_j^2)$, où SCT_j représente la somme des carrés des x_j et R_j^2 est le coefficient de

détermination issu de la régression auxiliaire de x_j sur les autres variables explicatives. [Nous avons implicitement utilisé cette équivalence en dérivant l'équation (3.51).] Cette décomposition permet de montrer que les écarts-types robustes sont élevés lorsque les x_j varient peu au sein de l'échantillon ou que les x_j sont fortement corrélés au reste des variables explicatives (en présence d'une forte multicolinéarité). Cette discussion fait écho à celle de la section 3.4 portant sur les facteurs influençant les valeurs des écarts-types traditionnels des MCO.

EXEMPLE 8.1
Estimation d'une équation de salaire à l'aide d'écarts-types robustes à l'hétéroscédasticité

Nous estimons le modèle de l'exemple 7.6 en reportant cette fois-ci les valeurs des écarts-types robustes à l'hétéroscédasticité en dessous des écarts-types habituels des MCO. Certaines estimations sont reportées avec plusieurs décimales afin de faciliter la comparaison des résultats :

$$\begin{aligned}\widehat{\log(wage)} = {} & 0{,}321 + 0{,}213\ marrmale - 0{,}198\ marrfem - 0{,}110\ singfem \\ & \quad (0{,}100)\ \ (0{,}055) \qquad\quad (0{,}058) \qquad\quad (0{,}056) \\ & \quad [0{,}109]\ \ [0{,}057] \qquad\quad [0{,}058] \qquad\quad [0{,}057] \\ & + 0{,}0789\ educ + 0{,}0268\ exper - 0{,}00054\ exper^2 \\ & \quad (0{,}0067) \qquad (0{,}0055) \qquad (0{,}00011) \\ & \quad [0{,}0074] \qquad [0{,}0051] \qquad [0{,}00011] \\ & + 0{,}0291\ tenure - 0{,}00053\ tenure2 \\ & \quad (0{,}0068) \qquad (0{,}00023) \\ & \quad [0{,}0069] \qquad [0{,}00024] \\ & n = 526,\ R^2 = 0{,}461. \end{aligned} \qquad \textbf{[8.6]}$$

Les écarts-types des MCO habituels ainsi que les écarts-types robustes sont reportés respectivement entre parenthèses et entre crochets, sous les coefficients correspondants. Les chiffres entre crochets sont les seuls éléments nouveaux du tableau puisque l'équation est toujours estimée par les MCO. Plusieurs points intéressants ressortent de l'équation (8.6). Tout d'abord, les résultats des tests statistiques de significativité sont, dans cette application, insensibles à la méthode de calcul des écarts-types : l'ensemble des variables statistiquement significatives avec l'une le reste avec l'autre. Ceci s'explique assez facilement par la faible différence constatée entre les valeurs d'écarts-types robustes et non robustes. (Si nous avions à calculer les p-valeurs correspondantes, elles seraient légèrement différentes puisque les deux séries de statistiques de Student ne sont pas strictement identiques.) Le changement le plus marqué concerne l'écart-type du coefficient de la variable *educ* puisqu'on passe d'une valeur de 0,0067, à 0,0074 en appliquant l'approche robuste. En dépit de cette légère augmentation, la statistique de Student de la variable *educ* reste très élevée, se situant aux alentours de 10.

L'équation (8.6) montre également que l'application d'une méthode robuste peut aussi bien conduire les écarts-types à augmenter qu'à diminuer. Par exemple, dans le cas de la variable *exper*, on obtient une valeur égale à 0,0051 par l'approche robuste contre 0,0051 initialement. Il n'est dès lors pas possible de savoir à l'avance quel sera l'impact de la correction sur la valeur finale de l'écart-type. En général, on notera toutefois que les écarts-types robustes sont plus grands.

Avant de clore cette discussion, nous devons souligner l'absence de test préalable pour s'assurer de la présence ou de l'absence d'hétéroscédasticité au niveau du modèle sur la population (8.6). En plus des écarts-types habituels, nous avons ici décidé de reporter directement ceux qui restent valides (asymptotiquement) en présence d'hétéroscédasticité. Au bout du compte, comme souvent dans les travaux appliqués, ce type de correction a eu peu d'incidence sur les conclusions de l'analyse. Dans certains cas néanmoins, les différences peuvent être substantielles et les conclusions peuvent être opposées. L'exemple sur ordinateur C2 illustre bien ce second cas de figure.

À ce stade, on peut légitimement s'interroger sur l'utilité des écarts-types habituels étant donné la disponibilité d'une version robuste, valide tant en présence qu'en l'absence d'hétéroscédasticité. La réponse est simple. Si les erreurs sont normalement distribuées et homoscédastiques, la statistique de Student provenant des MCO habituels suit exactement une loi de Student, indépendamment de la taille de l'échantillon (voir chapitre 4). À l'inverse, l'approche robuste requiert de travailler en grand échantillon pour obtenir une statistique qui suive une loi de Student. En présence d'un échantillon de petite taille, les statistiques de Student calculées à partir des écarts-types robustes peuvent avoir une distribution très éloignée de la distribution de Student, même sous les hypothèses MLC ; l'inférence statistique qui en découle n'est dès lors plus fiable.

En grand échantillon, en revanche, il est cohérent de reporter systématiquement les écarts-types robustes à l'hétéroscédasticité dans les études empiriques. Cette pratique est d'ailleurs de plus en plus courante. Il est également assez courant de reporter simultanément les versions habituelles et robustes des écarts-types, comme proposé dans l'équation (8.6). Procéder de la sorte a le mérite de laisser le lecteur évaluer par lui-même la sensibilité des conclusions au mode de calcul des écarts-types.

Nous savons maintenant comment procéder à des tests d'hypothèses sur un paramètre unique en présence d'hétéroscédasticité. Qu'en est-il des tests d'hypothèses jointes ? La bonne nouvelle est qu'il est également possible d'obtenir des versions robustes à l'hétéroscédasticité de forme inconnue ou arbitraire de la statistique du multiplicateur de Lagrange (*LM*) ainsi que de la statistique *F*. La **statistique de Fisher robuste à l'hétéroscédasticité** (ou une simple transformation de celle-ci) est aussi appelée *statistique de Wald robuste à l'hétéroscédasticité*. Une présentation générale de la statistique de Wald ne sera pas faite ici car elle nécessite de recourir à des notions d'algèbre matricielle. Le lecteur intéressé pourra se reporter à l'annexe E qui en développe certains éléments (voir également Wooldridge (2010, chapitre 4) pour un traitement plus détaillé). En pratique, il n'est pas compliqué d'utiliser les statistiques robustes dans les tests d'exclusions multiples (où plusieurs variables peuvent être exclues simultanément du modèle) : leur implémentation dans la plupart des logiciels économétriques est prévue.

EXEMPLE 8.2
Statistique *F* robuste à l'hétéroscédasticité

Le fichier GPA3 contient des données sur la réussite scolaire au second semestre dans l'enseignement secondaire. Il nous permet d'estimer l'équation suivante :

$$\begin{aligned}\widehat{cumgpa} = {} & 1{,}47 + 0{,}00114\ sat - 0{,}00857\ hsperc + 0{,}00250\ tothrs \\ & \underset{[0{,}22]}{(0{,}23)}\ \underset{[0{,}00019]}{(0{,}00018)}\ \underset{[0{,}00140]}{(0{,}00124)}\ \underset{[0{,}00073]}{(0{,}00073)} \\ & + 0{,}303\ female - 0{,}128\ black - 0{,}059\ white \\ & \underset{[0{,}059]}{(0{,}059)}\ \underset{[0{,}118]}{(0{,}147)}\ \underset{[0{,}110]}{(0{,}141)} \\ & n = 366,\ R^2 = 0{,}4006,\ \overline{R}^2 = 0{,}3905. \end{aligned} \qquad \textbf{[8.7]}$$

On constate, de nouveau, peu de différences entre les écarts-types habituels et leur version robuste. Les tests individuels de significativité obtenus avec ces mesures respectives aboutissent à des conclusions similaires. Il en va de même des résultats des tests joints de significativité. Supposons maintenant que nous voulions tester si l'origine ethnique n'a pas d'impact sur la variable *cumgpa*, étant donné les variables de contrôle incluses dans l'équation (8.7). Formellement cette hypothèse s'écrit de la façon suivante : $H_0 : \beta_{black} = 0, \beta_{white} = 0$. La statistique *F* habituelle se calcule facilement à partir du *R* carré du modèle contraint qui est égal à 0,3983 : $F = [(0{,}4006 - 0{,}3983) / (1 - 0{,}4006)] (359/2) \approx 0{,}69$. En présence d'hétéroscédasticité, nous avons vu que cette version

du test n'est pas valide. La version robuste à l'hétéroscédasticité n'a pas de forme simple, mais elle peut tout de même être calculée à l'aide de certains logiciels économétriques. Dans notre exemple, on trouve une statistique F robuste égale à 0,75. Cette valeur reste très proche de celle obtenue initialement. Ainsi, la p-valeur du test robuste, p-valeur = 0,474, se situe bien au-delà des seuils de significativité traditionnels. À la vue de ces résultats, nous ne pouvons donc pas rejeter l'hypothèse nulle.

En raison de l'invalidité de la statistique de Fisher en présence d'hétéroscédasticité, nous devons rester prudents lorsque nous réalisons un test de Chow pour tester l'égalité de coefficients entre deux groupes distincts. La forme de la statistique dans l'équation (7.24) n'est plus valide en présence d'hétéroscédasticité, ce qui inclut le cas simple où la variance de l'erreur diffère entre les deux groupes. En lieu et place de cette statistique, nous pouvons obtenir un test de Chow robuste à l'hétéroscédasticité en incluant une variable muette différenciant les deux groupes et en incluant des variables d'interaction, créées en combinant cette variable muette avec les autres variables explicatives du modèle. Il est alors possible de tester s'il existe des différences dans les deux spécifications – en testant que les coefficients affectant les variables muettes et les variables d'interaction sont tous nuls – ou simplement de tester si les pentes sont toutes similaires, ce qui revient à ne pas contraindre le coefficient de la variable muette. Voir l'Exercice sur ordinateur C14 pour un exemple.

Calcul du test LM robuste à l'hétéroscédasticité

Les logiciels économétriques n'offrent pas tous la possibilité de calculer directement des statistiques F robustes à l'hétéroscédasticité. Par conséquent, la réalisation des tests d'exclusion multiple requiert parfois l'application d'une autre procédure. Dans un tel cas de figure, la **statistique *LM* robuste à l'hétéroscédasticité** offre une bonne alternative puisqu'elle peut être construite très facilement à partir de n'importe quel logiciel.

Pour aller plus loin 8.1

Commenter la phrase suivante : « les écarts-types robustes à l'hétéroscédasticité sont toujours plus grands que les écarts-types habituels ».

Pour illustrer le calcul de la statistique *LM* robuste, considérons le modèle suivant

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u.$$

Supposons par ailleurs le test suivant : H0 : $\beta_4 = 0$, $\beta_5 = 0$. Pour calculer la statistique *LM* habituelle, il nous faut dans un premier temps estimer le modèle contraint (c'est-à-dire le modèle duquel sont exclues les variables x_4 et x_5) afin d'en extraire la série de résidus, $\tilde{u}$. Celle-ci fait ensuite l'objet d'une régression auxiliaire sur l'ensemble des variables indépendantes. La statistique *LM* résulte alors directement de l'application de la formule, $LM = n \cdot R_{\tilde{u}}^2$, où $R_{\tilde{u}}^2$ est le coefficient de détermination de la régression auxiliaire.

Le calcul de la version robuste de la statistique requiert un peu plus de travail. Une façon de procéder consiste à réaliser des régressions par les MCO de manière séquentielle. Pour commencer, nous devons régresser, sur l'ensemble des variables restantes, chaque variable à exclure du modèle sous l'hypothèse nulle. Dans notre cas, on notera $\tilde{r}_1$, le résidu de la régression auxiliaire de x_4 sur x_1, x_2, x_3, et $\tilde{r}_2$, celui de x_5 sur x_1, x_2, x_3. L'étape suivante peut sembler quelque peu étrange, mais il s'agit d'une étape intermédiaire nécessaire en vue d'obtenir notre statistique. Elle demande d'effectuer la régression auxiliaire de

$$1 \text{ sur } \tilde{r}_1\tilde{u}, \tilde{r}_2\tilde{u}, \qquad \textbf{[8.8]}$$

sans inclure de constante. On remarquera la nature particulière de notre variable dépendante, définie comme une série de « un ». Elle est ensuite régressée sur les produits $\tilde{r}_1\tilde{u}$ et $\tilde{r}_2\tilde{u}$. La statistique *LM* robuste peut enfin

être calculée comme suit : $LM = n - SCR_1$, où SCR_1 est la somme habituelle des carrés des résidus de la régression (8.8). Une fois la statistique obtenue, on applique la règle de décision usuelle.

La raison pour laquelle cette procédure nous permet de construire une statistique robuste est un peu technique. De manière générale, cela revient à répéter pour le test *LM* ce qui est fait pour les écarts-types robustes du test de Student. [Voir Wooldridge (1991b) ou Davidson et MacKinnon (1993) pour une discussion plus détaillée.]

Nous résumons maintenant le calcul de la statistique *LM* robuste à l'hétéroscédasticité dans le cas général.

Étapes de la construction d'une statistique LM *robuste à l'hétéroscédasticité :*

1. Estimer le modèle contraint afin de récupérer les résidus $\tilde{u}$.

2. Régresser chacune des variables indépendantes exclues du modèle sous l'hypothèse nulle sur l'ensemble des variables indépendantes restantes du modèle. À partir de ces régressions auxiliaires, récupérer ainsi q séries de résidus $(\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_q)$, avec q le nombre de contraintes d'exclusion sous H_0.

3. Calculer les produits entre chaque $\tilde{r}_j$ et $\tilde{u}$ (pour toutes les observations).

4. Régresser une série dont toutes les valeurs sont égales à 1 sur $\tilde{r}_1\tilde{u}, \tilde{r}_2\tilde{u}, \ldots, \tilde{r}_q\tilde{u}$, sans introduire de constante. La statistique *LM* robuste à l'hétéroscédasticité se calcule alors comme suit : $LM = n - SCR_1$, où SCR_1 est la somme des carrés des résidus habituelle de cette dernière régression. Sous H_0, la statistique *LM* suit approximativement une loi du χ^2_q.

Une fois le calcul de la statistique *LM* robuste réalisé, nous pouvons déduire la *p*-valeur en se référant aux valeurs tabulées (voir section 5.2), puis conclure en appliquant la règle de décision habituelle.

EXEMPLE 8.3
Test *LM* robuste à l'hétéroscédasticité

Utilisons les données du fichier CRIME1 pour vérifier si la durée moyenne des peines liées à des condamnations antérieures (*avgsen*) affecte le nombre d'arrestations au cours d'une année (*narr86*). (L'année de référence est 1986 dans la base de données). Les résultats du modèle sont reportés ci-dessous

$$\widehat{narr86} = 0{,}567 - 0{,}136\ pcnv + 0{,}0178\ avgsen - 0{,}00052\ avgsen2$$
$$(0{,}036)\ (0{,}040)\quad (0{,}0097)\quad (0{,}00030)$$
$$[0{,}040]\ [0{,}034]\quad [0{,}0101]\quad [0{,}00021]$$
$$- 0{,}0394\ ptime86 - 0{,}0505\ qemp86 - 0{,}00148\ inc86$$
$$(0{,}0087)\quad (0{,}0144)\quad (0{,}00034)$$
$$[0{,}0062]\quad [0{,}0142]\quad [0{,}00023]$$
$$+ 0{,}325\ black + 0{,}193\ hispan$$
$$(0{,}045)\quad (0{,}040)$$
$$[0{,}058]\quad [0{,}040]$$
$$n = 2\,725,\ R^2 = 0{,}0728, \qquad \textbf{[8.9]}$$

Dans cet exemple, les différences entre les valeurs d'écarts-types sont plus marquées qu'auparavant. Par exemple, la statistique de Student de la variable $avgsen^2$ voit sa valeur diminuer notablement en passant de – 1,73, à – 2,48 lorsque la correction est appliquée. La significativité de la variable est donc accrue par la prise en compte de l'hétéroscédasticité.

L'effet de *avgsen* sur la variable d'intérêt, *narr86*, s'avère un peu plus difficile à identifier. La relation étant quadratique, il est possible de déterminer un point de retournement à partir duquel l'effet change de signe. Ce point se situe autour de 0,0178/[2 (0,00052)] ≈ 17,12. Rappelons que cette variable est mesurée en mois. Littéralement, cela signifie que la variable *narr86* est positivement corrélée à *avgsen* tant que *avgsen* reste inférieure à 17 mois ; l'effet dissuasif n'apparaît qu'une fois ce seuil dépassé, c'est-à-dire lorsque la moyenne des peines antérieures excède 17 mois.

Testons maintenant si cet effet est statistiquement significatif. Pour y parvenir, nous devons procéder au test joint suivant : $H_0 : \beta_{avgsen}=0, \beta_{avgsen^2}=0$. La statistique *LM* habituelle nous donne *LM* = 3,54 (voir la section 5.2). La valeur tabulée correspondante pour une distribution du chi-deux avec 2 *ddl* donne une *p*-valeur égale à 0,170. Sur base de ce résultat, nous ne pouvons pas rejeter l'hypothèse nulle, même à un seuil de 15 %. La statistique *LM* robuste à l'hétéroscédasticité est légèrement supérieure, soit *LM* = 4,00 (arrondi à deux décimales). Bien que légèrement inférieure, la *p*-valeur correspondante, égale à 0,135, ne nous permet toujours pas de rejeter H_0. Nous pouvons en déduire que *avgsen* n'a pas d'effet significatif sur *narr86*. [Notons en passant que, lorsque la variable *avgsen* est incluse sans le terme quadratique dans le modèle (8.9), la statistique standard de Student est égale à 0,658 contre une statistique robuste de 0,592.]

8.3 TESTER LA PRÉSENCE D'HÉTÉROSCÉDASTICITÉ

Dans la section précédente, nous avons vu que les écarts-types robustes à l'hétéroscédasticité fournissent une méthode simple pour calculer les statistiques de Student distribuées asymptotiquement selon une distribution de Student, tant en présence qu'en l'absence d'hétéroscédasticité. De la même manière, nous avons montré qu'il existe pour les tests joints des versions robustes des statistiques *F* et *LM*. Un des avantages majeurs de ces tests réside dans le fait qu'ils ne nécessitent pas de savoir si les erreurs du modèle sont réellement hétéroscédastiques. Étant donné les propriétés désirables de ces statistiques, est-il toujours nécessaire de se soucier de la présence de l'hétéroscédasticité ? Ne suffit-il pas simplement d'appliquer ces corrections par défaut ? En réalité, tester la présence d'hétéroscédasticité dans un modèle peut se révéler utile, pour plusieurs raisons. Comme nous l'avons évoqué dans la section précédente, sous les hypothèses du modèle linéaire classique, les statistiques de Student habituelles suivent une distribution exacte, et non asymptotique, à la différence des versions robustes. C'est la raison pour laquelle les statistiques classiques des tests d'inférence sont encore utilisées par beaucoup d'économistes lorsqu'il n'y a pas d'indication d'hétéroscédasticité dans les erreurs. Par ailleurs, lorsque les erreurs sont hétéroscédastiques, l'estimateur des MCO perd sa qualité de meilleur estimateur linéaire sans biais. Or, si la forme de l'hétéroscédasticité est connue, il est possible d'obtenir un meilleur estimateur, comme nous le verrons dans la section 8.4. L'utilité des tests d'hétéroscédasticité n'est donc pas nulle.

Au cours des dernières années, de nombreux tests ont été proposés pour détecter l'hétéroscédasticité dans le terme d'erreur. Certains d'entre eux permettent d'atteindre cet objectif sans pour autant être capable d'identifier clairement les facteurs qui en sont à l'origine. Dans la suite de cette section, nous n'évoquerons pas cette catégorie de tests afin de nous concentrer sur ceux capables de déterminer les causes du problème. Se restreindre à ce type de tests permet également de conserver un cadre d'analyse uniforme.

Comme d'habitude, nous commençons par le modèle linéaire suivant :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \qquad \textbf{[8.10]}$$

Nous travaillons dans cette section sous les hypothèses RLM.1 à RLM.4. En particulier, nous supposons $E(u|x_1, x_2, \ldots, x_k) = 0$, de sorte que l'estimateur MCO soit sans biais et convergent.

L'hypothèse nulle à tester considère l'hypothèse RLM.5 comme vraie :

$$H_0 : Var(u|x_1, x_2, \ldots, x_k) = \sigma^2 \qquad \textbf{[8.11]}$$

Le principe général du test consiste à considérer que l'hypothèse d'homoscédasticité est valide dans un premier temps, pour ensuite la confirmer ou l'infirmer à la lumière des données. Si nous ne pouvons pas rejeter (8.11) à un seuil de significativité suffisamment petit, nous devrons conclure que l'hétéroscédasticité n'est pas un problème dans notre modèle. Dans le cas contraire, il faudra la prendre en compte dans le choix des méthodes d'estimation et des procédures d'inférence statistique. Comme d'habitude, il convient de garder à l'esprit que nous n'acceptons jamais H_0 ; ne pas rejeter H_0 signifie tout simplement que nous n'avons pas suffisamment d'éléments empiriques à notre disposition pour la rejeter.

Conformément à nos hypothèses de départ, le terme d'erreur u est d'espérance conditionnelle nulle, donc $Var(u|\mathbf{x}) = E(u^2|\mathbf{x})$. L'hypothèse nulle d'homoscédasticité peut alors s'écrire :

$$H_0 : E(u^2|x_1, x_2, \ldots, x_k) = E(u^2) = \sigma^2.$$

L'écriture de l'hypothèse nulle sous cette forme permet de montrer que le test d'homoscédasticité revient à tester si la valeur attendue de u^2 dépend d'une ou plusieurs variables explicatives du modèle. Ainsi, si H_0 est fausse, la valeur attendue de u^2, compte tenu des variables indépendantes, peut être représentée par n'importe quelle fonction des x_j. Une approche simple pour réaliser ce test consiste à assumer comme on le fait souvent une fonction linéaire :

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + v, \qquad \textbf{[8.12]}$$

où v est un terme d'erreur de moyenne nulle conditionnellement aux valeurs de x_j. Il est important de faire particulièrement attention à la nature de la variable dépendante dans cette équation qui, pour rappel, n'est pas le u de (8.10) mais bien le *carré* de l'erreur, u^2. L'hypothèse nulle d'homoscédasticité s'écrit alors :

$$H_0 : \delta_1 = \delta_2 = \ldots \delta_k = 0. \qquad \textbf{[8.13]}$$

Sous l'hypothèse nulle, il est souvent raisonnable de supposer l'erreur de (8.12), v, indépendante des variables indépendantes $x_1, x_2, \ldots, x_k$. Nous ferons de même ici. Si l'on s'attarde maintenant sur la formulation de l'hypothèse nulle (8.13), on constate clairement la nature du test d'homoscédasticité : il s'agit d'un test de significativité globale du modèle (8.12). Nous savons depuis la section 5.2 que les statistiques F et LM peuvent être utilisées à cette fin. Toutes deux, par ailleurs, restent valident asymptotiquement même si u^2 n'est pas distribué normalement (Par exemple, si u est normalement distribué, u^2/σ^2 suivra une loi χ^2_1.) La possibilité d'observer u^2 dans l'échantillon nous permettrait de calculer la statistique très simplement en régressant u^2 sur $x_1, x_2, \ldots, x_k$ par la méthode des MCO sur l'ensemble des n observations de l'échantillon avant d'appliquer un test de significativité globale.

Comme nous l'avons déjà souligné, cependant, nous ne connaitrons jamais les erreurs réelles dans la population. On peut en revanche s'appuyer sur des valeurs estimées de ces erreurs : les résidus MCO, $\hat{u}_i$. L'équation que nous pouvons estimer prend la forme

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + erreur \qquad \textbf{[8.14]}$$

Elle nous permet de calculer les statistiques F ou LM pour tester la significativité jointe des $x_1, x_2, \ldots, x_k$. Bien qu'il soit assez difficile de le démontrer, l'utilisation des résidus estimés par les MCO, à la place des erreurs, n'affecte pas la distribution des statistiques F ou LM en grand échantillon.

Revenons maintenant au calcul des statistiques d'intérêt. Les statistiques F et LM dépendent toutes les deux du R carré de la régression auxiliaire (8.14). Nous l'appellerons, $R^2_{\hat{u}^2}$, afin de le distinguer du R carré de l'équation originale (8.10). La statistique F s'écrit

$$F = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(n - k - 1)}, \qquad \textbf{[8.15]}$$

où k est le nombre des régresseurs dans (8.14). On retrouve évidemment le même nombre de variables indépendantes que dans (8.10). Le calcul de (8.15) est réalisé automatiquement dans la plupart des logiciels économétriques pour évaluer la significativité globale d'une régression. Cette statistique F suit (approximativement) une distribution $F_{k,n-k-1}$ sous l'hypothèse nulle d'homoscédasticité.

La statistique LM, s'obtient comme d'habitude en effectuant le produit de la taille de l'échantillon avec le coefficient de détermination de (8.14) :

$$LM = n \cdot R^2_{\hat{u}^2}. \qquad \textbf{[8.16]}$$

Sous l'hypothèse nulle, LM suit asymptotiquement une loi χ^2_k. Cette statistique s'obtient également très facilement après l'estimation de (8.14).

La version LM du test porte généralement le nom de **test de Breusch-Pagan pour l'hétéroscédasticité** ou plus simplement de **test BP**. Breusch et Pagan (1979) ont également proposé une forme alternative du test qui repose sur l'hypothèse de la normalité des erreurs. Les travaux de Koenker (1981) indiquent qu'il est préférable de garder la forme de la statistique LM en (8.16). Dans la pratique, le test de Breusch-Pagan est effectivement réalisé à partir de l'équation (8.16), en raison notamment de sa plus grande simplicité de mise en œuvre.

Résumons maintenant la marche à suivre en vue de tester l'hétéroscédasticité à l'aide du test BP :

Étapes du test d'hétéroscédasticité de Breusch-Pagan :

1. Estimer le modèle (8.10), comme d'habitude, en appliquant les MCO. Sauver les résidus élevés au carré, $\hat{u}^2$.

2. Effectuer la régression auxiliaire (8.14) et récupérer le coefficient de détermination de cette régression, $R^2_{\hat{u}^2}$.

3. À partir des valeurs obtenues, calculer les statistiques F ou bien LM. Rejeter l'hypothèse nulle d'homoscédasticité si la p-valeur correspondante est suffisamment faible, c'est-à-dire inférieure au seuil de significativité retenu. (Dans le premier cas, les p-valeurs sont issues des valeurs tabulées de la distribution $F_{k,n-k-1}$; dans le second cas, elles le sont de la distribution χ^2_k).

Si le test BP nous amène à rejeter l'hypothèse nulle, certaines mesures correctives doivent être prises. Une première stratégie possible consiste à conserver les estimateurs MCO et à substituer les écarts-types habituels par une forme robuste à l'hétéroscédasticité en vue de procéder à l'inférence statistique (voir section précédente). La seconde stratégie consiste à abandonner les MCO au profit d'une autre méthode d'estimation mieux adaptée à la présence d'hétéroscédasticité. Cette seconde stratégie sera examinée à la section 8.4.

EXEMPLE 8.4
Tester la présence d'hétéroscédasticité dans un modèle de détermination du prix des maisons

Utilisons les données du fichier HPRICE1 pour tester la présence d'hétéroscédasticité dans un modèle de détermination du prix des maisons. L'équation estimée sur les variables prises en niveau s'écrit

$$\widehat{price} = -21,77 + 0,00207 lotsize + 0,123 sqrft + 13,85 bdrms$$
$$(29,48)\ (0,00064)\qquad (0,013)\qquad (9,01) \qquad \textbf{[8.17]}$$
$$n = 88,\ R^2 = 0,672.$$

Cette expression ne nous permet pas de savoir si le modèle contient des erreurs hétéroscédastiques sur la population. Pour cela, il nous faut mettre en œuvre les procédures de test vues précédemment. Dans un premier temps, on régresse le carré des résidus MCO sur les variables indépendantes. Puis, on calcule le *R* carré de la régression auxiliaire de $\hat{u}^2$ sur *lotsize*, *sqrft*, et *bdrms*. On obtient $R^2_{\hat{u}^2} = 0,1601$. Notre échantillon contient $n = 88$ observations et notre modèle inclut trois variables indépendantes, soit $k = 3$. Ces informations nous permettent de calculer la statistique $F = [0,1601 / (1 \text{ à } 0,1601)]\ (84/3) \approx 5,34$. En consultant les tables de la $F_{3,84}$, on voit que la *p*-valeur correspondante est égale à 0,002. Ce résultat plaide fortement en faveur du rejet de l'hypothèse nulle. Nous réalisons maintenant le même exercice pour la statistique *LM*. Le nombre d'observations et la valeur du R carré de la régression auxiliaire nous permettent de calculer la statistique telle que $LM = 88\ (0,1601) \approx 14.09$. En consultant les tables de la χ^2_3, nous obtenons une *p*-valeur correspondante de 0,0028. Ce dernier résultat vient conforter les conclusions précédentes. On en déduit que les écarts-types habituels reportés dans (8.17) ne sont pas fiables.

Dans le chapitre 6, nous avons vu qu'une transformation logarithmique de la variable dépendante permettait souvent d'atténuer l'ampleur de l'hétéroscédasticité dans les erreurs. Étant donné le problème que l'on vient d'identifier, l'application de cette transformation aux variables *price*, *lotsize*, et *sqrft* est une bonne idée. Cet ajustement permet par ailleurs de mesurer directement l'élasticité de *price* en fonction de *lotsize*, et *sqrft*. Les résultats de l'estimation sont reportés ci-dessous

$$\widehat{\log(price)} = -1,30 + 0,168 \log(lotsize) + 0,700 \log(sqrft) + 0,037 bdrms$$
$$(0,65)\ (0,038)\qquad (0,093)\qquad (0,028) \qquad \textbf{[8.18]}$$
$$n = 88,\ R^2 = 0,643.$$

La régression du carré des résidus de (8.18) obtenus par les MCO, sur les variables indépendantes log(*lotsize*), log(*sqrft*), et *bdrms*, nous donne : $R^2_{\hat{u}^2} = 0,0480$. À partir de là, nous pouvons directement obtenir les statistiques de tests et leur *p*-valeur correspondante : $F = 1,41$ ($p = 0,245$) et $LM = 4,22$ ($p = 0,239$). Nous ne sommes maintenant plus en mesure de rejeter l'hypothèse nulle d'homoscédasticité. Ce résultat vient confirmer le constat partagé par bon nombre d'études empiriques quant à l'effet modérateur de la transformation logarithmique sur l'hétéroscédasticité d'un modèle.

Il est facile d'ajuster le test de Breusch-Pagan dans l'hypothèse où l'hétéroscédasticité ne proviendrait que de certaines variables explicatives : il suffit de régresser $\hat{u}^2$ sur le groupe de variables qui sont supposées être à la source du problème. Les statistiques *F* ou *LM* sont ensuite calculées comme d'habitude. Rappelez-vous qu'il convient dans ce cas d'ajuster le nombre de degrés de liberté puisque ceux-ci sont fonction du nombre de régresseurs du modèle auxiliaire. Le nombre de variables indépendantes de l'équation (8.10) n'est donc pas pertinent dans ce cas.

Pour aller plus loin de 8.2

Considérons l'équation de salaire (7.11). Si nous pensons que la variance conditionnelle de *log(wage)* ne dépend pas des variables *educ*, *exper*, et *tenure* mais qu'elle diffère en fonction des quatre groupes démographiques de l'échantillon (constitués des hommes mariés, femmes mariées, hommes célibataires et femmes célibataires), quelle régression devons-nous effectuer pour tester l'hétéroscédasticité ? Quel est le nombre de degrés de liberté du test *F* ?

Si les résidus au carré sont régressés sur une seule variable indépendante, le test d'hétéroscédasticité se ramène à un test de Student habituel sur la variable en question. Une statistique de Student significative suggère la présence d'hétéroscédasticité dans les erreurs du modèle.

Le test de White pour l'hétéroscédasticité

Le test de White repose sur le même principe que le test BP ; il en constitue même une extension. Dans le chapitre 5, sous les hypothèses de Gauss-Markov, nous avons montré la validité asymptotique des écarts-types des MCO ainsi que celle des tests statistiques habituels. Il s'avère toutefois que l'hypothèse d'homoscédasticité, notée formellement $\text{Var}(u_1|x_1, \ldots, x_k) = \sigma^2$, peut être remplacée par une hypothèse moins stricte. Celle-ci suppose l'absence de corrélation entre, d'une part, le terme d'erreur au carré, u^2, et, d'autre part, toutes les variables indépendantes (x_j), leurs carrés (x_j^2) et les doubles produits ($x_j x_h$ pour $j \neq h$). Cette observation a conduit White (1980) à proposer un test d'hétéroscédasticité tenant compte des carrés et des produits croisés de toutes les variables indépendantes. Le test est explicitement destiné à tester la présence de formes additionnelles d'hétéroscédasticité susceptibles d'invalider aussi bien les écarts-types MCO que les statistiques de tests habituelles.

Lorsque le modèle contient $k = 3$ variables indépendantes, le test de White repose sur l'estimation du modèle suivant :

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + \textit{erreur} \qquad \textbf{[8.19]}$$

Nous pouvons remarquer que cette équation présente six régresseurs de plus que celle du test de Breusch-Pagan. La mise en œuvre du **test de White pour détecter la présence d'hétéroscédasticité** repose sur le calcul d'une statistique *LM* sous l'hypothèse nulle d'absence de significativité de tous les paramètres δ_j de l'équation (8.19), à l'exception de la constante. À l'instar du test BP, la dernière étape du test de White peut également être réalisée à partir d'une statistique *F* à la place de la statistique *LM*.

Avec seulement trois variables indépendantes dans le modèle d'origine, l'équation (8.19) contient neuf paramètres à estimer, en plus de la constante. Avec six variables indépendantes, le nombre passe à 27 (à moins que certaines variables soient redondantes). Cette abondance de paramètres à estimer constitue une des principales faiblesses de ce test qui est très gourmand en degrés de liberté, même pour un nombre modéré de régresseurs dans le modèle d'origine.

Une variante plus parcimonieuse du test de White a été proposée pour répondre à ce problème. Rappelons d'abord que la différence entre les tests de Breusch-Pagan et celui de White réside dans l'ajout des carrés et des produits croisés des variables indépendantes. Une manière de pallier le problème de perte de degrés de liberté, tout en préservant l'esprit du test, consiste à remplacer les variables indépendantes par les valeurs ajustées de la variable dépendante. Celles-ci sont définies, pour chaque observation *i*, par

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}.$$

Par définition, les valeurs ajustées sont fonction des variables indépendantes. Nous pouvons également les élever au carré pour obtenir une fonction particulière des carrés des variables indépendantes et de tous les produits croisés. L'information contenue dans les variables explicatives précédentes est alors « synthétisée » dans les valeurs ajustées. Dans ce cas, nous pouvons réécrire le test comme suit

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \textit{erreur}, \qquad \textbf{[8.20]}$$

où $\hat{y}$ désigne les valeurs ajustées. Il est évidemment important de ne pas confondre $\hat{y}$ avec y dans cette équation. Nous utilisons les valeurs ajustées parce qu'elles sont uniquement fonction des variables indépendantes

(ainsi que des paramètres estimés). La substitution de $\hat{y}$ par y dans (8.20) ne permettrait pas d'obtenir un test valide d'hétéroscédasticité.

Comme précédemment, les statistiques F ou LM peuvent être employées indifféremment pour réaliser le test de l'hypothèse nulle, $H_0 : \delta_1 = 0, \delta_2 = 0$, dans l'équation (8.20). L'avantage de cette version synthétique du test de White est qu'elle n'implique que deux restrictions, indépendamment du nombre de variables indépendantes du modèle original. Elle est, par ailleurs, très simple à appliquer.

En réalité, cette version synthétique correspond à un cas particulier du test de White. Il faut garder à l'esprit que la variable $\hat{y}$ est une estimation de la valeur attendue de y, compte tenu des x. L'utilisation de (8.20) pour tester l'hétéroscédasticité se justifie dans les cas où nous suspectons que la variance change avec le niveau des valeurs attendues, $E(y|\mathbf{x})$. Par conséquent, le test de (8.20) peut être considéré comme un cas particulier du test de White puisque l'équation (8.20) impose des restrictions sur les paramètres de l'équation (8.19).

Étapes du cas particulier du test d'hétéroscédasticité de White :

1. Estimer le modèle (8.10) par les MCO, comme d'habitude. Sauver les résidus et les valeurs ajustées. Élever au carré les résidus MCO, $\hat{u}^2$, ainsi que les valeurs ajustées, $\hat{y}^2$.

2. Effectuer la régression auxiliaire (8.20). Puis, récupérer le R carré de cette seconde régression, $R^2_{\hat{u}^2}$.

3. À partir des valeurs obtenues, calculer la statistique F ou LM. Rejeter l'hypothèse nulle d'homoscédasticité si la p-valeur correspondante est suffisamment faible, c'est-à-dire inférieure au seuil de significativité retenu. (Dans le premier cas, les p-valeurs sont issues des valeurs tabulées de la distribution $F_{2,n-k}$; dans le second cas, elles le sont de la distribution χ^2_2).

EXEMPLE 8.5
Tester la présence d'hétéroscédasticité dans un modèle de détermination du prix des maisons à l'aide du test particulier de White

Nous appliquons maintenant la version parcimonieuse du test de White à l'équation (8.18). Le test est réalisé à l'aide de la statistique LM. Soulignons de nouveau que sous l'hypothèse nulle, la distribution du chi-deux de la statistique du test ne possède que deux degrés de liberté. La régression des carrés des résidus, $\hat{u}^2$, sur les variables $\widehat{lprice}$, $(\widehat{lprice})^2$, où $\widehat{lprice}$ désigne les valeurs ajustées de (8.18), nous donne un coefficient de détermination égal à 0,0392, soit $R^2_{\hat{u}^2} = 0{,}0392$. Ce résultat nous permet de calculer la statistique LM qui est égale à 3,45, soit $LM = 88\ (0{,}0392) \approx 3{,}45$. En consultant les tables de la distribution χ^2_2, cela correspond à une p-valeur de 0,178. Cette p-valeur est légèrement plus faible que celle obtenue précédemment à partir du test de Breusch-Pagan. Ceci ne suffit toutefois pas à rejeter l'hypothèse nulle d'homoscédasticité, même au seuil de 15 %.

Avant de clore cette section, revenons sur un point important concernant l'interprétation à donner du rejet de l'hypothèse nulle des tests de Breusch-Pagan et des tests de White. Jusqu'à maintenant, ces tests ont été spécifiés sous les hypothèses RLM.1 à RLM.4. En particulier, nous avons supposé que la forme fonctionnelle de $E(y|\mathbf{x})$ était correctement spécifiée. Or, si ce n'est pas le cas, nous pouvons être amenés à rejeter l'hypothèse nulle d'homoscédasticité alors que $Var(y|\mathbf{x})$ est constante. Par exemple, si un terme quadratique dans la régression est omis ou si l'estimation se fait en niveau alors qu'elle devrait inclure les variables en logarithme, le résultat du test indiquera à tort qu'il faut rejeter l'hypothèse nulle d'homoscédasticité. Pour cette raison, certains chercheurs considèrent les tests d'hétéroscédasticité comme des tests plus généraux

d'erreurs de spécification de la forme fonctionnelle. Il existe néanmoins des tests plus adaptés à ce problème de spécification, comme nous le verrons dans la section 9.1. En pratique, il est préférable de débuter une étude empirique par l'analyse de la forme fonctionnelle à l'aide de tests spécialisés. En d'autres termes, il convient de tester l'hétéroscédasticité après avoir déterminé la forme fonctionnelle.

8.4 ESTIMATION PAR LES MOINDRES CARRÉS PONDÉRÉS

Que devons-nous faire si les tests statistiques concluent à la présence d'hétéroscédasticité ? Depuis la section 8.2, nous savons qu'il est possible d'estimer le modèle par les MCO et de corriger ensuite les écarts-types et les tests d'inférence statistique. Avant le développement de cette méthode, il était nécessaire de spécifier la forme de l'hétéroscédasticité pour ensuite estimer le modèle par la méthode des moindres carrés pondérés (MCP) que nous allons décrire dans cette section. En supposant que la forme de la variance du terme d'erreur ait été correctement spécifiée (en fonction de variables explicatives), la méthode des MCP présente l'avantage d'être plus efficiente que les MCO. En outre, les MCP permettent d'obtenir des statistiques de tests, t et F, qui suivent exactement les distributions de Student et de Fisher. Néanmoins, il arrive que la variance soit mal spécifiée. Dans cette section, nous discuterons des implications qu'une mauvaise spécification de la variance peut avoir sur l'estimation du modèle par les MCP.

Hétéroscédasticité connue à une constante multiplicative près

Soit $\mathbf{x}$ l'ensemble des variables explicatives de l'équation (8.10). Supposons par ailleurs que

$$\text{Var}(u|\mathbf{x}) = \sigma^2 h(\mathbf{x}), \qquad \textbf{[8.21]}$$

où $h(\mathbf{x})$ est une fonction des variables explicatives qui détermine la forme de l'hétéroscédasticité. Étant donné que la variance doit être positive, l'inégalité $h(\mathbf{x}) > 0$ doit être respectée pour toutes les valeurs possibles des variables indépendantes. Pour l'instant, nous supposons $h(\mathbf{x})$ connue. Le paramètre de la population, σ^2, reste inconnu, mais nous serons en mesure de l'estimer à partir d'un échantillon de données.

Pour une observation prise au hasard dans la population, on aura $\sigma_i^2 = \text{Var}(u_i|\mathbf{x}_i) = \sigma^2 h(\mathbf{x}_i) = \sigma^2 h_i$, où $\mathbf{x}_i$ désigne les valeurs de *l'ensemble* des variables indépendantes pour l'observation i. On notera que h_i change pour chaque observation en fonction des valeurs $\mathbf{x}_i$. Par exemple, considérons le cas d'une fonction simple de l'épargne

$$sav_i = \beta_0 + \beta_1 inc_i + u_i \qquad \textbf{[8.22]}$$

$$\text{Var}(u_i|inc_i) = \sigma^2 inc_i. \qquad \textbf{[8.23]}$$

On obtient $h(x) = h(inc) = inc$; la variance des erreurs est donc proportionnelle au niveau du revenu. Cela signifie qu'une augmentation du revenu accroit la variabilité de l'épargne (Si $\beta_1 > 0$, la valeur attendue de l'épargne augmente également avec le revenu.) Comme la variable *inc* est toujours positive, il est possible de calculer la racine carrée de la variance de l'équation (8.23) pour obtenir l'écart-type de u_i, conditionnellement à inc_i. On écrit : $\sigma\sqrt{inc_i}$.

La question que nous nous posons maintenant est de savoir comment utiliser l'information de l'équation (8.21) pour estimer les β_j. Considérons une nouvelle fois le modèle initial, sous les hypothèses RLM.1 à RLM.4 :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i, \qquad \textbf{[8.24]}$$

Les erreurs sont maintenant hétéroscédastiques. Avant d'estimer le modèle, nous devons transformer cette équation en rendant les erreurs homoscédastiques dans le but ultime de valider les cinq hypothèses de Gauss-Markov. Comme h_i est une fonction de $\mathbf{x}_i$, l'espérance conditionnelle de $u_i/\sqrt{h_i}$ (par rapport à $\mathbf{x}_i$) est égale à zéro. Comme $\text{Var}(u_i|\mathbf{x}_i) = \text{E}(u_i^2|\mathbf{x}_i) = \sigma^2 h_i$, la variance conditionnelle de $u_i/\sqrt{h_i}$ (par rapport à $\mathbf{x}_i$) est σ^2 :

$$\text{E}\left((u_i/\sqrt{h_i})^2\right) = \text{E}(u_i^2)/h_i = (\sigma^2 h_i)/h_i = \sigma^2,$$

Notez bien que, pour des raisons de clarté, nous avons omis le conditionnement par rapport à $\mathbf{x}_i$ dans cette expression. Nous pouvons maintenant diviser l'équation d'origine (8.24) par $\sqrt{h_i}$ afin d'obtenir le modèle transformé suivant

$$y_i/\sqrt{h_i} = \beta_0/\sqrt{h_i} + \beta_1(x_{i1}/\sqrt{h_i}) + \beta_2(x_{i2}/\sqrt{h_i}) + \ldots + \beta_k(x_{ik}/\sqrt{h_i}) + (u_i/\sqrt{h_i}) \qquad \textbf{[8.25]}$$

ou

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \ldots + \beta_k x_{ik}^* + u_i^*, \qquad \textbf{[8.26]}$$

Notez que la constante a été remplacée par une variable, $x_{i0}^* = 1/\sqrt{h_i}$, dont les valeurs peuvent maintenant changer en fonction de i. Les autres variables « étoiles » sont celles de l'équation originale, standardisées par $\sqrt{h_i}$.

Les variables transformées de l'équation (8.26) n'ont pas vraiment de sens économique. À ce stade, il est important de retenir que cette transformation vise à obtenir des estimateurs plus efficients que les MCO en présence d'hétéroscédasticité. En outre, l'interprétation des paramètres de pente reste identique à celle du modèle original (8.24).

Illustrons le principe de cette transformation à l'aide de l'exemple relatif à l'équation d'épargne. Le modèle transformé s'écrit

$$sav_i/\sqrt{inc_i} = \beta_0(1/\sqrt{inc_i}) + \beta_1\sqrt{inc_i} + u_i^*,$$

Pour simplifier l'expression, nous utilisons le fait que $inc_i/\sqrt{inc_i} = \sqrt{inc_i}$. Comme évoqué plus haut, les paramètres gardent leur interprétation initiale. Ainsi, β_0 reste la constante et β_1 représente la propension marginale à épargner en fonction du revenu telle que définie dans l'équation (8.22).

L'équation (8.26) satisfait toujours les hypothèses RLM.1 à RLM.4. L'équation, linéaire dans ses paramètres, reste issue d'un échantillonnage aléatoire et possède des termes d'erreur u_i^* de moyenne nulle conditionnellement aux valeurs prises par $\mathbf{x}_i^*$. En revanche, la variance conditionnelle est maintenant égale à σ^2. Si l'équation originale satisfait bien les quatre premières hypothèses de Gauss-Markov, cela signifie que l'équation transformée (8.26) les vérifie toutes. Enfin, si u_i est distribuée normalement, u_i^* suit également une distribution normale de variance σ^2 et l'équation transformée satisfait les hypothèses du MLC (de RLM.1 à RLM.6).

Puisque nous savons que les MCO possèdent de bonnes propriétés sous les hypothèses de Gauss-Markov (l'estimateur est *BLUE*, par exemple), la discussion du paragraphe précédent suggère que nous estimions les paramètres de l'équation (8.26) par les MCO. Ces estimateurs, $\beta_0^*, \beta_1^*, \ldots, \beta_k^*$, seront différents des estimateurs des MCO appliqués sur l'équation originale. Les paramètres β_j^* sont des exemples d'**estimateurs des moindres carrés généralisés (MCG)**. Dans ce chapitre, l'estimateur MCG est utilisé pour tenir compte de l'hétéroscédasticité dans les erreurs. Dans le chapitre 12, nous verrons que les estimateurs MCG sont également utiles dans d'autres circonstances.

Comme l'équation (8.26) satisfait à l'ensemble des hypothèses de Gauss-Markov, les écarts-types et les statistiques de tests, de Student ou de Fisher, peuvent tous être obtenus à partir de l'estimation du modèle transformé. Par ailleurs, la somme des carrés des résidus de (8.26), rapportée au nombre de degrés de liberté, fournit un estimateur sans biais de σ^2. Étant donné que les estimateurs MCG sont les meilleurs estimateurs linéaires sans biais de β_j, ils sont nécessairement plus précis que les estimateurs MCO, $\hat{\beta}_j$, appliqués à l'équation non transformée. En pratique, après avoir transformé les variables, on suit simplement la procédure d'analyse standard des MCO. Au moment d'interpréter les résultats, il faut toutefois garder à l'esprit que les paramètres estimés s'interprètent en fonction de l'équation d'origine.

Le R carré de l'estimation (8.26) s'avère utile dans le calcul de la statistique F. Il l'est en revanche beaucoup moins pour évaluer la qualité d'ajustement du modèle initial. En effet, sa valeur nous renseigne sur la part des variations de y^* expliquées par les x_j^*. Le R carré *ne* nous renseigne *pas* sur la part des variations de y, notre variable d'intérêt, expliquées par les x_j. Par conséquent, le R carré est rarement instructif dans la pratique.

Les estimateurs MCG visant à corriger l'hétéroscédasticité sont appelés les **estimateurs des moindres carrés pondérés (MCP)**. Ce nom vient de la méthode d'estimation qui consiste à obtenir x_j^* en minimisant la somme pondérée des carrés des résidus, chaque résidu au carré étant pondéré par $1/h_i$. L'idée qui sous-tend cette transformation est d'accorder moins de poids aux observations ayant une variance de l'erreur plus élevée. À l'inverse, les MCO accordent le même poids à chaque observation, ce qui permet d'obtenir les estimateurs les plus précis lorsque la variance de l'erreur est constante pour toutes les sous-parties de la population. Mathématiquement, les estimateurs MCP permettent d'obtenir les valeurs de b_j qui minimisent

$$\sum_{i=1}^{n}(y_i - b_0 - b_1x_{i1} - b_2x_{i2} - \ldots - b_kx_{ik})^2/h_i. \qquad \textbf{[8.27]}$$

Si nous introduisons maintenant la racine carrée de $1/h_i$ à l'intérieur des parenthèses, nous pouvons observer que la somme pondérée des carrés des résidus est égale à la somme des carrés des résidus des variables transformées :

$$\sum_{i=1}^{n}(y_i^* - b_0x_{i0}^* - b_1x_{i1}^* - b_2x_{i2}^* - \ldots - b_kx_{ik}^*)^2.$$

Puisque les MCO minimisent la somme des carrés des résidus (quelle que soit la définition de la variable dépendante et des variables indépendantes), il s'ensuit que les estimateurs MCP qui minimisent (8.27) sont tout simplement les estimateurs MCO de (8.26). On notera également que les carrés des résidus dans (8.27) sont pondérés par $1/h_i$ tandis que les variables transformées en (8.26) sont pondérés par $1/\sqrt{h_i}$.

Un estimateur des MCP peut être défini pour n'importe quel ensemble de poids positifs. Les MCO peuvent dès lors être considérés comme un cas particulier des MCG, pour lequel un même poids est accordé à toutes les observations. La procédure d'estimation efficiente, celle des MCG, pondère chaque résidu au carré par l'*inverse* de la variance conditionnelle de u_i sachant $\mathbf{x}_i$.

En pratique, il peut s'avérer fastidieux de devoir transformer les variables de l'équation (8.25) en vue d'effectuer manuellement les MCO. Ces étapes peuvent également être source d'erreur. Heureusement, la plupart des logiciels économétriques offrent la possibilité d'appliquer automatiquement les MCP. En règle générale, dans un logiciel standard, l'estimation du modèle (8.27) par les MCP nécessite simplement que l'utilisateur spécifie la fonction de pondération, $1/h_i$, en plus des variables dépendantes et indépendantes du modèle d'origine. Autrement dit, il suffit d'indiquer que des pondérations proportionnelles à l'inverse de la variance doivent être utilisées. Non seulement cette procédure limite le risque de commettre une erreur mais elle nous permet d'interpréter les estimateurs des moindres carrés pondérés en fonction du modèle

original. En fait, nous pouvons écrire l'équation estimée de la manière habituelle. Les estimateurs et les écarts-types seront différents des MCO, mais la façon dont nous *interprétons* les estimateurs, les écarts-types et les statistiques de tests, reste la même.

Les logiciels économétriques qui disposent de routines pour réaliser des régressions pondérées seront à même de reporter une valeur de R-carré (et de R-carré ajusté) en sus des coefficients estimés et des écarts-types estimés associés. En effet, le R-carré obtenu à l'issue d'une régression par la méthode des MCP peut être calculé en minimisant l'équation (8.27), ainsi qu'une somme des carrés totaux (SCT) pondérés obtenus à l'aide du même vecteur de poids, mais en fixant toutes les pentes de l'équation (8.27), $b_1, b_2, \ldots, b_k$, à zéro. En tant que mesure d'ajustement du modèle aux données, ce R-carré n'est pas d'une grande utilité dans la mesure où il quantifie la variation expliquée de y_i^* plutôt que celle de y_i. Néanmoins, les R-carrés calculés tel que décrit plus haut sont valides pour réaliser des tests de contraintes linéaires multiples de Fisher (si tant est que nous ayons correctement spécifié la fonction de variance au préalable). En effet, tout comme dans le cadre des moindres carrés ordinaires, le terme de SCT disparaît et nous pouvons alors calculer une statistique de Fisher sur la seule base des SCR pondérés.

Le R-carré issu de la régression par les MCO de l'équation (8.26) n'est en revanche pas une mesure pertinente de l'ajustement du modèle aux données dans la mesure où le calcul de la SCT fait peu de sens ici : il est en effet attendu que soit exclue la constante de la régression, ce qui amène mécaniquement les logiciels statistiques à calculer la SCT sans centrer préalablement les données y_i^*. Ceci constitue une raison supplémentaire pour privilégier le recours à des packages pré-programmés pour l'estimation du modèle par les MCP de façon à avoir la garantie que le R-carré reporté soit correctement calculé et constitue alors un outil d'analyse utile pour comparer le modèle estimé complet avec un autre ne contenant qu'une constante. Du fait de la disparition du terme de SCT lorsque l'on procède à des tests de restrictions multiples, le fait de ne pas correctement calculer ce terme n'affecte pas la valeur de la statistique de Fisher calculée à partir des R-carrés des régressions auxiliaires. Il convient néanmoins de rester prudent et de garder à l'esprit que cette valeur de R-carré ne permet pas de juger correctement de l'adéquation du modèle aux données.

EXEMPLE 8.6
Modélisation du patrimoine financier

Lorsque *educ* = 8, le salaire horaire estimé à partir de (2.27) est égal à 3,42 USD en 1976. Que vaudrait ce salaire horaire en 2003 ? (Nous estimons une équation qui explique le patrimoine financier net total (*nettfa*, mesuré en millier de dollars) en fonction du revenu (*inc*, également mesuré en millier de dollars) et d'un ensemble de variables de contrôle, comme l'âge, le sexe, et un indicateur d'admissibilité au plan de retraite 401(k). Nous disposons des données du fichier 401KSUBS pour estimer ce modèle. Seules les données relatives aux personnes seules (*fsize* = 1) sont considérées dans cette étude. Dans l'exercice sur ordinateur C12 du chapitre 6, nous avons transformé la variable *age* en appliquant une fonction quadratique spécifique : $(age - 25)^2$. Cette transformation a pour vertu de faciliter l'interprétation de l'effet de l'âge sur le patrimoine financier net, puisque l'âge minimum des individus au sein de l'échantillon est de 25 ans. De ce fait, *nettfa* est une fonction croissante de l'âge à partir de 25 ans.

Les résultats de l'estimation sont indiqués dans le tableau 8.1. Comme nous pouvons raisonnablement soupçonner la présence d'hétéroscédasticité dans les erreurs du modèle, les écarts-types robustes à l'hétéroscédasticité sont directement indiqués en dessous des estimations des paramètres obtenues par les MCO. Nous calculons également les estimations des MCP et leurs écarts-types respectifs sous l'hypothèse suivante : $\text{Var}(u|inc) = \sigma^2 inc$.

Si nous ne tenons pas compte de facteurs tiers, les résultats de l'estimation par les MCO indiquent qu'un dollar supplémentaire de revenu augmente le patrimoine financier net, *nettfa*, d'environ 82 centimes ; lorsque l'estimateur des MCP est utilisé, l'effet diminue légèrement, à hauteur de 79 ¢. Cet écart reste toutefois limité. Le constat est différent au sujet des écarts-types. En effet, l'écart-type issu des MCP est 40 % inférieur à celui des MCO, en supposant naturellement que le modèle, $\text{Var}(nettfa|inc) = \sigma^2 inc$, soit correct.

Que se passe-t-il si nous incluons les variables de contrôle dans le modèle ? On note d'abord une légère diminution de la valeur des deux coefficients estimés de la variable *inc*. Les MCP fournissent toujours l'estimateur de β_{inc} le plus petit et le plus précis. L'âge a un effet positif à partir de 25 ans. Son ampleur reste plus importante sur base de l'estimateur MCO mais l'estimateur MCP est en revanche plus précis, comme précédemment. Le sexe n'a pas d'effet statistiquement significatif sur *nettfa*, à l'inverse de l'admissibilité au plan 401(k). L'estimation par les MCO nous indique qu'une personne admissible possédera en moyenne un patrimoine financier net total supérieur à une personne non éligible de $ 6 890, indépendamment des revenus, de l'âge et du sexe. De manière générale, les valeurs des estimateurs MCP sont nettement plus petites que celles des MCO. Ce constat laisse à penser que la forme fonctionnelle de l'équation est mal spécifiée. (Une manière de résoudre ce problème est d'introduire dans le modèle une variable d'interaction entre *e401k* et *inc*. Voir exercice sur ordinateur C11.)

Analysons maintenant la significativité jointe des variables du modèle. Pour appliquer le test F aux trois variables $(age - 25)^2$, *male*, et *e401k*, nous devons calculer les R carrés des modèles contraints et non contraints estimés par les MCP. Ensuite, à l'aide de cette information, nous pouvons calculer la statistique F. Sa valeur est égale à $F = 30{,}8$. Pour des degrés de liberté égaux à 2 et 2 012, la p-valeur que nous obtenons est extrêmement petite, puisqu'elle vaut zéro jusqu'à la 15e décimale. Ce résultat n'est pas surprenant, étant donné les valeurs élevées des statistiques de Student pour les variables *age* et *e401k*.

Tableau 8.1 *Variable dépendante :* nettfa

Variables indépendantes	(1) MCO	(2) MCP	(3) MCO	(4) MCP
inc	0,821(,104)	0,787 (0,063)	0,771 (0,100)	0,740 (0,064)
$(age - 25)^2$	–	–	0,0251 (0,0043)	0,0175 (0,0019)
male	–	–	2,48 (2,06)	1,84 (1,56)
e401k	–	–	6,89 (2,29)	5,19 (1,70)
constant	– 10,57 (2,53)	– 9,58 (1,65)	– 20,98 (3,50)	– 16,70 (1,96)
Observations	2 017	2 017	2 017	2 017
R carré	0,0827	0,0709	0,1279	0,1115

Pour aller plus loin 8.3

Vous régressez $\hat{u}^2$ sur *inc* en vous basant sur les résidus des MCO calculés pour la régression de la colonne (1) du tableau 8.1. La statistique de Student que vous obtenez pour cette régression auxiliaire est égale à 2,96. Sur base de ce résultat, considérez-vous que l'hétéroscédasticité est un problème ?

Dans l'exemple sur le patrimoine financier net, nous avons appliqué les MCP en supposant que la variance des termes d'erreur était proportionnelle au revenu. Ce choix reste fondamentalement arbitraire. Dans la plupart des cas, il n'existe pas de règle objective de détermination du poids qu'il convient d'adopter dans l'équation de la variance. Une exception concerne l'utilisation de données moyennes par groupe ou par région géographique, à la place de données à l'échelle des individus. En guise d'illustration, examinons le lien entre, d'une part, la contribution d'un travailleur à son plan de retraite (*contrib*) *et*, d'autre part, la participation complémentaire de l'employeur (*mrate*). Soit *i* une entreprise particulière et *e* un employé en son sein. Un modèle simple expliquant la contribution au plan de retraite s'écrit comme suit

$$contrib_{i,e} = \beta_0 + \beta_1 earns_{i,e} + \beta_2 age_{i,e} + \beta_3 mrate_i + u_{i,e}, \qquad [8.28]$$

où $contrib_{i,e}$ est la contribution annuelle au plan de retraite de l'employé, *e*, au sein de l'entreprise *i* ; $earns_{i,e}$, représente les revenus annuels de cet employé ; et $age_{i,e}$, son âge. Le variable $mrate_i$ est le montant de la participation de l'employeur au plan d'épargne retraite pour chaque dollar versé par le salarié.

Si (8.28) satisfait les hypothèses de Gauss-Markov, nous pouvons l'estimer sur un échantillon composé de salariés provenant de différentes entreprises. Supposons, maintenant, que nous ne disposions plus de données à l'échelle des individus ; les seules valeurs disponibles sont des valeurs moyennes, à l'échelle des entreprises, sur le montant des cotisations, le revenu et l'âge des employés. Par $\overline{contrib_i}$, $\overline{earns_i}$, et $\overline{age_i}$, nous désignons les valeurs moyennes pour l'entreprise *i*. Le nombre d'employés de l'entreprise *i* est représenté par m_i. Il est considéré comme connu. Nous prenons la moyenne de l'équation (8.28) sur l'ensemble des employés de *i* afin d'obtenir une relation entre les variables d'intérêt à l'échelle de l'entreprise :

$$\overline{contrib_i} = \beta_0 + \beta_1 \overline{earns_i} + \beta_2 \overline{age_i} + \beta_3 mrate_i + \bar{u}_i, \qquad [8.29]$$

où $\bar{u}_i = m_i^{-1} \sum_{e=1}^{m_i} u_{i,e}$ représente l'erreur moyenne sur l'ensemble des salariés au sein de l'entreprise *i*. Si nous avons *n* entreprises dans notre échantillon, alors (8.29) correspond à un modèle de régression multiple traditionnel, qui peut être estimé par les MCO. Les estimateurs sont sans biais si le modèle original (8.28) satisfait les hypothèses de Gauss-Markov et possède des erreurs individuelles, $u_{i,e}$, indépendantes de la taille de l'entreprise, m_i. [Dans ce cas, la valeur attendue de $\bar{u}_i$, compte tenu des variables explicatives (8.29), est égale à zéro].

Si l'équation à l'échelle des individus (8.28) satisfait l'hypothèse d'homoscédasticité et que les erreurs au sein de la firme *i* ne sont pas corrélées entre employés, nous pouvons montrer que l'équation à l'échelle des entreprises (8.29) contient une forme particulière d'hétéroscédasticité. Plus précisément, si $\text{Var}(u_{i,e}) = \sigma^2$, pour tout *i* et *e*, et $\text{Cov}(u_{i,e}, u_{i,g}) = 0$, pour chaque paire d'employés $e \neq g$ dans l'entreprise *i*, alors $\text{Var}(\bar{u}_i) = \sigma^2 / m_i$; nous retrouvons la formule habituelle de la variance d'une moyenne de variables aléatoires non corrélées et de même variance. En d'autres termes, la variance du terme d'erreur $\bar{u}_i$ diminue avec la taille de l'entreprise : $h_i = 1/m_i$. Dans ce cas, la procédure d'estimation la plus efficiente est celle des MCP, avec un poids égal au nombre de salariés de l'entreprise ($1/h_i = m_i$). De cette manière, nous pouvons estimer de manière efficiente les paramètres à l'échelle des individus à partir de données moyennes à l'échelle des entreprises.

La question est maintenant de savoir si une pondération similaire peut être appliquée dans le cadre de données individuelles observées au niveau d'une ville, d'un département, d'une région, ou d'un pays. Si le modèle au niveau désagrégé (exprimé à l'échelle des individus, pour des données par tête) satisfait les hypothèses de Gauss-Markov, les termes d'erreur de cette équation possèdent une variance égale à la variance de la population divisée par sa taille. Par conséquent, il est approprié de recourir aux MCP en utilisant des poids égaux à la taille de la population. Illustrons cette approche en considérant des données, disponibles à l'échelle d'une ville, sur la consommation de bière par habitant (*beerpc*, en onces), le pourcentage de personnes de moins de 21 ans dans la population (*perc21*), le nombre d'années d'éducation moyen par adulte (*avgeduc*), le niveau de revenu moyen (*incpc*), et le prix de la bière (*price*). Le modèle à l'échelle communale s'écrit

$$beerpc = \beta_0 + \beta_1 perc21 + \beta_2 avgeduc + \beta_3 incpc + \beta_4 price + u.$$

Nous pouvons estimer ce modèle par les MCP en utilisant la population de chaque ville comme pondération.

Les avantages liés à l'utilisation d'une pondération par la taille (qu'elle soit relative aux nombres de salariés d'une entreprise, au nombre d'habitants d'une ville ou à tout autre niveau d'agrégation) reposent sur l'hypothèse d'homoscédasticité du modèle sous-jacent à l'échelle des individus. En présence d'hétéroscédasticité, la pondération appropriée ne dépend plus simplement du facteur « taille » mais également de la forme de l'hétéroscédasticité. De plus, en présence d'erreurs corrélées au sein d'un même groupe (par exemple, au sein d'une entreprise), la relation entre les variances n'est plus vérifiée : $\text{Var}(\bar{u}_i) \neq \sigma^2/m_i$ (voir l'exercice 7). Dans les modèles tels que (8.29), l'incertitude qui existe autour de la forme exacte de $\text{Var}(\bar{u}_i)$, conduit en général à estimer l'équation par les MCO en se limitant à rendre robustes les écarts-types estimés, ainsi que les statistiques de tests, en présence d'hétéroscédasticité. Une alternative consiste à appliquer les MCP en se servant des tailles des groupes comme pondération, puis à calculer des écarts-types estimés robustes. Si le modèle à l'échelle des individus satisfait les hypothèses de Gauss-Markov, cette démarche permet d'obtenir des estimateurs efficients et de procéder à des tests d'inférence robustes en présence aussi bien d'hétéroscédasticité à l'échelle des individus que d'erreurs corrélées au sein des groupes.

Estimation de la fonction d'hétéroscédasticité : les moindres carrés quasi généralisés (MCQG)

Dans la sous-section précédente, nous avons considéré des exemples pour lesquels l'hétéroscédasticité est connue, même sous une forme multiplicative. Dans la plupart des cas cependant, la forme exacte de l'hétéroscédasticité n'est pas évidente à identifier. En d'autres termes, il est difficile de trouver la fonction $h(x_i)$ évoquée dans la section précédente, qui nous a permis de construire des estimateurs efficaces en présence d'hétéroscédasticité. Il est néanmoins possible, dans de nombreux cas, de modéliser cette fonction puis d'en estimer les paramètres inconnus à l'aide de données. Il en résulte une estimation de chaque h_i, noté $\hat{h}_i$. La substitution de h_i par $\hat{h}_i$ dans la procédure d'estimation par les MCG permet d'obtenir un nouvel estimateur, appelé **estimateur des moindres carrés quasi généralisés (MCQG)**. On emploiera également les acronymes FGLS (*feasible generalized least squares*) ou EGLS (*estimated* GLS), tirés de la terminologie anglo-saxonne.

Il existe de nombreuses façons de modéliser l'hétéroscédasticité. Dans cette section, nous allons nous concentrer sur une approche particulièrement flexible. Supposons que :

$$\text{Var}(u|\mathbf{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k), \qquad \textbf{[8.30]}$$

où $x_1, x_2, \ldots, x_k$ désignent les variables indépendantes du modèle de régression [voir équation (8.1)] et les δ_j correspondent aux paramètres inconnus. Les variables x_j peuvent apparaître sous d'autres formes dans le modèle, mais nous nous concentrons ici sur la forme définie dans l'équation (8.30), comme évoqué précédemment. Conformément aux notations de la section précédente, on peut écrire : $h(\mathbf{x}) = \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k)$.

À ce stade, vous devez certainement vous interroger sur le choix de la fonction exponentielle dans (8.30). À l'occasion de l'application du test de Breusch-Pagan, nous avions d'ailleurs modélisé l'hétéroscédasticité comme une fonction linéaire des x_j. Les formes linéaires telles que (8.12) sont tout à fait appropriées lorsqu'il s'agit de *tester* la présence d'hétéroscédasticité. En revanche, elles le sont beaucoup moins lorsqu'il s'agit de *corriger* l'hétéroscédasticité à l'aide de la méthode des MCP. Nous avons déjà évoqué la raison qui justifie ce choix : le risque d'obtenir une variance prédite négative à partir d'un modèle linéaire n'est pas nul. Or, la variance doit toujours être positive.

Si nous avions connaissance de la valeur des paramètres δ_j, il nous suffirait d'appliquer les MCP, à l'instar de ce qui a été fait dans la sous-section précédente. Ce cas de figure n'est cependant pas très réaliste. En pratique, nous devons généralement utiliser des données pour estimer ces paramètres et construire ensuite les poids sur base de ces estimations. Comment peut-on concrètement estimer les paramètres δ_j ? La procédure d'estimation consiste essentiellement à transformer l'expression initiale sous forme linéaire pour parvenir ensuite à l'estimer par les MCO.

Sous l'hypothèse (8.30), nous pouvons écrire :

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k) v,$$

où la moyenne conditionnelle de v, sachant $\mathbf{x} = (x_1, x_2, \ldots, x_k)$, est égale à l'unité. Si v est indépendant de $\mathbf{x}$, nous pouvons écrire que

$$\log(u^2) = a_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + e, \qquad \textbf{[8.31]}$$

où e est indépendant de $\mathbf{x}$ et son espérance est nulle. La constante du modèle diffère de celle de (8.30). Ceci n'a toutefois pas d'incidence sur la mise en œuvre des MCP. Suite à cette transformation, la variable dépendante devient le logarithme de l'erreur quadratique. Comme (8.31) satisfait les hypothèses de Gauss-Markov, nous pouvons obtenir des estimateurs sans biais à partir des MCO.

Comme d'habitude, il faut remplacer les termes d'erreur non observés du modèle, u, par les résidus issus de la régression par les MCO. Nous pouvons ensuite estimer la régression suivante :

$$\log(\hat{u}^2) \text{ on } x_1, x_2, \ldots, x_k. \qquad \textbf{[8.32]}$$

Elle nous permet de calculer les valeurs ajustées, appelées $\hat{g}_i$. Nous en déduisons directement les estimateurs de h_i :

$$\hat{h}_i = \exp(\hat{g}_i) \qquad \textbf{[8.33]}$$

Enfin, nous pouvons appliquer les MCP en substituant les poids de l'équation (8.27), $1/h_i$, par les valeurs estimées, $1/\hat{h}_i$. Nous résumons ci-dessous les étapes requises.

Procédure de correction des estimateurs par les MCGF en présence d'hétéroscédasticité

1. Régresser y sur $x_1, x_2, \ldots, x_k$ et récupérer le résidu, $\hat{u}$.

2. Calculer $\log(\hat{u}^2)$ en élevant les termes du résidu MCO au carré avant de prendre le logarithme népérien.

3. Effectuer la régression de l'équation (8.32) afin d'obtenir les valeurs ajustées, $\hat{g}$.

4. Prendre l'exponentiel des valeurs ajustées de (8.32) : $\hat{h} = \exp(\hat{g})$.

5. Estimer l'équation $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$ à l'aide des poids prédits, $1/\hat{h}$, issus des étapes précédentes. Autrement dit, nous remplaçons h_i par $\hat{h}_i$ dans l'équation (8.27). Pour rappel, il convient de pondérer le résidu

au *carré* de l'observation i par $1/\hat{h}_i$. Par contre, si nous transformons toutes les variables avant d'appliquer les MCO, chaque variable, y compris la constante, doit être multipliée par $1/\sqrt{\hat{h}_i}$.

En résumé, dans la section précédente, nous avons mis en évidence les propriétés désirables dont disposent les estimateurs obtenus par les MCP, construits à partir de h_i. Ils correspondent aux meilleurs estimateurs linéaires sans biais (*BLUE*). Comme h_i est rarement connu, ce modèle est difficilement applicable dans la pratique. Il est possible de remplacer h_i dans la procédure d'estimation par l'estimateur $\hat{h}_i$ obtenu à partir des mêmes données. Malheureusement, cette étape supplémentaire affecte les propriétés des estimateurs car l'estimateur des MCQG n'est plus sans biais ; il ne peut pas être *BLUE*. Les MCQG fournissent néanmoins des estimateurs convergents et plus efficients que les MCO sur le plan asymptotique. Ce résultat s'avère difficile à démontrer en raison de l'estimation des paramètres de variance. Si nous ignorons cet aspect du problème (ce que nous ferons ici), la preuve est identique à celle de l'efficience asymptotique de l'estimateur des MCO (voir le théorème 5.3). Les MCQG offrent par conséquent en présence d'hétéroscédasticité une alternative intéressante aux MCO, pour des échantillons de grande taille.

Il est important de rappeler que les estimateurs des MCQG portent sur les paramètres du modèle de la population habituelle :

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u.$$

À l'image des estimateurs des MCO, ils mesurent l'impact marginal de chaque variable x_j sur y. En grand échantillon, on aura tendance à appliquer les MCQG plutôt que les MCO en raison de leur plus grande efficience et du fait que les statistiques de tests qui en découlent suivent les distributions traditionnelles de Student et de Fisher. En cas de doute sur la variance spécifiée dans l'équation (8.30), il est possible d'utiliser les écarts-types estimés et les statistiques robustes à l'hétéroscédasticité dans l'équation transformée.

À l'instar du raisonnement ayant conduit à la construction de la version parcimonieuse du test de White, une autre façon d'estimer h_i consiste à remplacer les variables indépendantes de la régression (8.32) par les valeurs ajustées et leurs carrés, obtenus par les MCO. En d'autres termes, nous obtenons $\hat{g}_i$ à partir des valeurs prédites de

$$\log(\hat{u}^2) \text{ on } \hat{y}, \hat{y}^2. \qquad \textbf{[8.34]}$$

Les valeurs de $\hat{h}_i$ peuvent alors être calculées de la même manière que dans l'équation (8.33). À l'exception de l'étape (3), la procédure mise en œuvre pour déterminer $\hat{h}_i$ est identique à la précédente.

S'il est possible d'utiliser la régression (8.32) pour estimer la fonction de variance, pourquoi ne pas y recourir pour tester l'hétéroscédasticité (à l'aide d'un test F ou LM) ? Ce fut la proposition de Park (1966). Malheureusement, les propriétés du test de Park s'avèrent moins intéressantes que celles des tests présentés dans la section 8.3. Tout d'abord, l'hypothèse nulle est plus restrictive puisqu'elle requiert à la fois l'homoscédasticité des termes d'erreur et l'indépendance de u et $\mathbf{x}$, contrairement aux tests de Breusch-Pagan et de White. Ensuite, la substitution de u par les résidus MCO, $\hat{u}$, dans (8.32) peut conduire la statistique F à s'écarter de la distribution de Fisher, même en grand échantillon. Aucun de ces problèmes n'avait été rencontré précédemment. Pour ces raisons, le test de Park n'est pas recommandé pour détecter la présence d'hétéroscédasticité. En revanche, la régression (8.32) reste utile pour appliquer les MCP puisqu'elle nous fournit des estimateurs convergents des δ_j.

Lors de la mise en œuvre de tests multiples à partir d'une équation estimée par les MCP, il est nécessaire de faire attention au mode de calcul de la statistique F. (Cette remarque s'applique aux statistiques F construites aussi bien à partir de la somme des carrés des résidus qu'à partir du R carré.) Plus précisément, il convient de bien veiller à utiliser les mêmes poids lors des estimations des modèles contraints et non contraints. Ce dernier doit d'abord être estimé par les MCO. Une fois estimés, les poids sont utilisés dans l'estimation du modèle contraint. Sinon, la statistique F se calcule comme d'habitude. De nombreux logiciels nous évitent

de devoir réaliser cette procédure manuellement : il existe des commandes simples qui permettent de tester automatiquement ce type d'hypothèses jointes par les MCP.

EXEMPLE 8.7
Consommation de cigarettes

Nous disposons des données du fichier SMOKE pour expliquer la consommation quotidienne de cigarettes. Comme la majorité des gens ne fument pas, la plupart des observations de *cigs* sont égales à zéro. En réalité, le choix d'un modèle linéaire n'est pas idéal car il ne garantit pas que les valeurs prédites soient toutes cohérentes avec la nature de la variable d'intérêt, c'est-à-dire positives ou nulles dans le cas d'espèce. L'utilisation d'un modèle linéaire reste malgré tout intéressante lorsqu'il s'agit d'identifier les facteurs influençant la consommation de cigarettes.

Nous estimons l'équation par les MCO. Les écarts-types estimés par la méthode des MCO sont reportés entre parenthèses :

$$\begin{aligned}\widehat{cigs} &= -3{,}64 + 0{,}880\log(income) - 0{,}751\log(cigpric) \\ &\qquad (24{,}08)(0{,}728) \qquad\qquad (5{,}773) \\ &\quad -0{,}501 educ + 0{,}771 age - 0{,}0090 age^2 - 2{,}83 restaurn \\ &\quad (0{,}167) \qquad (0{,}160) \qquad (0{,}0017) \qquad (1{,}11) \\ &\qquad n = 807,\ R^2 = 0{,}0526\end{aligned} \qquad [8.35]$$

où *cigs* = nombre de cigarettes fumées par jour ; *income* = le revenu annuel ; *cigpric* = le prix des cigarettes par paquet (en cents) ; *educ* = le nombre d'années d'étude ; *age* = mesurée en années ; *restaurn* = un indicateur binaire égal à 1 si la personne réside dans un État ayant adopté une législation interdisant ou limitant la consommation du tabac dans les restaurants.

Comme nous allons estimer le modèle par les MCP un peu plus tard, nous n'indiquons pas les écarts-types corrigées pour la présence l'hétéroscédasticité à ce stade de l'analyse. (Notez en passant que 13 des 807 valeurs prédites par le modèle sont inférieures à zéro, c'est-à-dire moins de 2 %. La nature linéaire du modèle ne doit donc pas être forcément considérée comme une source de préoccupation majeure.)

Ni le revenu, ni le prix des cigarettes n'apparaissent comme statistiquement significatifs dans (8.35). L'ampleur de leur effet *ceteris paribus* est d'ailleurs très faible. Par exemple, si le revenu augmente de 10 %, *cigs* devrait augmenter en moyenne de (0,880/100) (10) = 0,088, soit moins d'un dixième de cigarette par jour. L'effet du prix sur la consommation relève du même ordre de grandeur.

Chaque année d'éducation supplémentaire permet en moyenne de réduire la consommation quotidienne d'une demi-cigarette. Cette fois-ci, l'effet est statistiquement significatif. La consommation de cigarette s'avère également liée à l'âge, de façon quadratique. Elle augmente avec l'âge jusque 0,771/[2 (0,009)] ≈ 42,83 ans, puis l'effet s'inverse et elle diminue au fil des années. On notera que les deux termes de la forme quadratique sont statistiquement significatifs, confirmant l'existence d'un effet non linéaire sur les paramètres de la population. Enfin, l'adoption d'une législation restreignant la consommation de tabac dans les restaurants, a l'effet négatif escompté puisqu'elle diminue la consommation d'environ trois cigarettes en moyenne par jour.

Que pouvons-nous dire de la présence d'hétéroscédasticité dans le terme d'erreur du modèle sous-jacent (8.35) ? Pour répondre à cette question, nous appliquons le test de Breusch-Pagan. Si nous estimons l'équation (8.14), le $R^2_{\hat{u}^2}$ de la régression auxiliaire des résidus au carré sur les variables indépendantes de (8.35) est égal 0,040. Le pouvoir explicatif du modèle est très faible. A priori, ce résultat semblerait indiquer que l'erreur est homoscédastique. Pour pouvoir conclure, il est impératif de procéder à des tests statistiques formels, à l'aide

des statistiques *F* ou *LM*. Si l'échantillon est de grande taille, gardez à l'esprit qu'une faible valeur du $R^2_{\hat{u}^2}$ peut malgré tout conduire à un rejet incontestable de l'hypothèse d'homoscédasticité. À partir de la valeur estimée du *R* carré, nous pouvons calculer la statistique *LM*, soit *LM* = 807 (0,040) = 32,28. Sous l'hypothèse nulle, cette statistique suit une loi du χ^2_6. En consultant les tables de la χ^2_6, cela correspond à une *p*-valeur inférieure à 0,000015. Ce résultat constitue une indication solide de présence d'hétéroscédasticité dans le modèle.

Les conclusions du test d'hétéroscédasticité nous conduisent à estimer l'équation (8.32) par les moindres carrés quasi-généralisés (MCQG). Les résultats de l'estimation sont reportés ci-dessous :

$$\begin{aligned} \widehat{cigs} &= 5{,}64 + 1{,}30\ \log(income) - 2{,}94\log(cigpric) \\ &\quad (17{,}80)(0{,}44) \qquad\qquad (4{,}46) \\ &\quad -0{,}463\,educ + 0{,}482\,age - 0{,}0056\,age^2 - 3{,}46\,restaurn \\ &\quad (0{,}12) \qquad (0{,}097) \qquad (0{,}0009) \qquad (0{,}80) \\ &\quad n = 807,\ R^2 = 0{,}1134 \end{aligned} \qquad \textbf{[8.36]}$$

En comparant les résultats de (8.35) et (8.36), nous constatons plusieurs évolutions. Tout d'abord, l'effet du revenu devient statistiquement significatif et son ampleur est plus grande. L'impact du prix sur la consommation de cigarette augmente également, bien qu'il ne soit toujours pas statistiquement significatif. [Ce résultat s'explique certainement par la faible variance de *cipric* : le prix des cigarettes ne change que d'un État à un autre. La variable log(*cigpric*) est donc beaucoup plus stable que les autres variables propres à chaque individu, soit log (*income*), *educ*, et *age*.]

Les résultats relatifs aux autres variables du modèle ont légèrement évolué, sans pour autant altérer les principaux enseignements de nos estimations précédentes : le nombre d'années d'étude tend à diminuer la consommation de cigarette ; l'âge a un impact positif puis négatif au-delà d'un certain seuil ; enfin, la mise en place d'une législation visant à limiter la consommation de tabac dans les restaurants a l'effet négatif attendu.

L'exemple 8.7 est l'illustration d'un problème auquel on est parfois confronté lors de l'application des MCP : celui de la divergence entre les valeurs des estimateurs des MCO et des MCP. Dans l'exemple sur la consommation de cigarettes, le problème reste limité. Les signes des estimations sont les mêmes. Seules les variables non significatives présentent des différences notables. En règle générale, il est tout à fait normal d'observer des différences entre estimateurs MCO et MCP, ne serait-ce qu'en raison de l'erreur d'échantillonnage. La question est de savoir si l'ampleur de ces différences est susceptible d'affecter en profondeur les conclusions de l'étude.

Pour aller plus loin 8.4

Considérez $\hat{u}_i$, les résidus *non* pondérés des MCP appliqués à (8.36), ainsi que $\widehat{cigs}_i$, les valeurs ajustées. (La formule qui permet d'obtenir ces valeurs ajustées est similaire à celle que nous utilisons pour les MCO. La seule différence tient au choix de l'estimateur des β_j). Un moyen de savoir si l'hétéroscédasticité a correctement été éliminée consiste à appliquer un nouveau test d'hétéroscédasticité sur les valeurs des résidus standardisés (ou transformés), soit $\hat{u}_i^2/\hat{h}_i = (\hat{u}_i/\sqrt{\hat{h}_i})^2$. [Si $h_i = Var(u_i|x_i)$, les résidus transformés ne devraient pas souffrir d'hétéroscédasticité] Il existe plusieurs approches pour conduire ce test. L'une d'entre elles consiste à appliquer, au modèle transformé, la version parcimonieuse du test de White. Autrement dit, le résidu transformé et élevé au carré, soit $\hat{u}_i^2/\hat{h}_i$, est régressé sur les variables prédites du modèle transformé et leur carré, soit $\widehat{cigs}_i/\sqrt{\hat{h}_i}$ et $\widehat{cigs}_i^2/\hat{h}_i$ (en veillant à ajouter une constante). Sur base des données du fichier SMOKE, nous obtenons une statistique *F* du test joint égale à 11,15. Peut-on conclure à l'élimination de l'hétéroscédasticité suite à l'application de cette correction ?

Si d'importantes différences sont constatées entre les valeurs des estimateurs MCO et MCP (comme celles liées à un changement de signe, par exemple), il est préférable de considérer les résultats avec une extrême prudence. En règle générale, cette instabilité provient de la violation d'une autre hypothèse de Gauss-Markov, comme l'hypothèse RLM.4. Si l'espérance conditionnelle de l'erreur n'est pas nulle, $E(y|\mathbf{x}) \neq \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ si bien que les estimateurs MCO et MCP ont des valeurs attendues et des limites en probabilité différentes. En effet, pour que les estimateurs MCP des β_j soient convergents, il ne suffit pas que u ne soit pas corrélé à chaque x_j. Il faut également que l'hypothèse RLM.4 soit respectée. Une différence marquée entre les valeurs des deux estimateurs peut donc indiquer que la forme fonctionnelle de $E(y|\mathbf{x})$ est mal spécifiée. Il reste naturellement à déterminer ce qu'est une différence significative. Le test d'Hausman peut être utilisé à cette fin [Hausman (1978)]. Ce test permet de comparer formellement les valeurs des deux estimateurs MCO et MCP dans le but d'évaluer si la différence constatée dépasse celle due à l'erreur d'échantillonnage. L'explication de ce test dépasse le cadre de cet ouvrage. Dans de nombreux cas, il suffit de jeter un simple « coup d'œil » sur les valeurs estimées pour détecter le problème.

Que faire si la fonction d'hétéroscédasticité présumée est fausse ?

Nous venons d'évoquer l'incidence d'une mauvaise spécification de la moyenne conditionnelle, $E(y|\mathbf{x})$, sur les valeurs des estimateurs MCO et MCP. Nous pouvons maintenant nous interroger sur l'effet d'une mauvaise spécification de la fonction de la variance sur les propriétés des estimateurs MCP. En d'autres termes, qu'advient-il des estimateurs MCP si $\text{Var}(y|\mathbf{x}) \neq \sigma^2 h(\mathbf{x})$? En réalité, les propriétés d'absence de biais et de convergence ne sont pas affectées par une erreur de spécification dans $h(\mathbf{x})$, du moins sous l'hypothèse RLM.4. Rappelons que, si $E(u|\mathbf{x}) = 0$, alors toute fonction de $\mathbf{x}$ n'est pas corrélée avec u. L'erreur pondérée, $u/\sqrt{h(\mathbf{x})}$, ne doit donc pas être corrélée avec les régresseurs pondérés par une quelconque fonction $h(\mathbf{x})$ positive. Par conséquent, les écarts importants entre estimateurs MCO et MCP sont le signe d'une mauvaise spécification de la forme fonctionnelle. Notez que le fait de devoir estimer les paramètres de la fonction, $h(x, \hat{\delta})$, introduit un biais dans les estimateurs MCP, même s'ils restent généralement convergents (que la fonction de la variance soit ou non correctement spécifiée).

Certes, l'estimateur des MCP conserve sa propriété de convergence sous les hypothèses de RLM.1 à RLM.4, mais une mauvaise spécification de la fonction de la variance reste préoccupante pour deux raisons. La première est particulièrement importante : même en présence de grands échantillons, les écarts-types des MCP, ainsi que les tests statistiques habituels calculés sous l'hypothèse $\text{Var}(y|\mathbf{x}) = \sigma^2 h(\mathbf{x})$, ne sont plus valides. Par exemple, la fiabilité des estimateurs MCP et de leurs écarts-types estimés, reportés dans la colonne (4) du tableau 8.1, dépendent de la validité de l'hypothèse d'une variance conditionnelle proportionnelle au revenu : $\text{Var}(nettfa|inc, age, male, e401k) = \text{Var}(nettfa|inc) = \sigma^2 inc$. Si ce n'est pas le cas, les écarts-types estimés ne sont pas valides, ce qui fausse également toutes les statistiques qui en dépendent. Heureusement, il existe une solution simple à ce problème. À l'instar des écarts-types estimés robustes calculés pour les estimateurs MCO à partir d'une forme inconnue d'hétéroscédasticité, il est possible de calculer des écarts-types estimés pour les MCP qui permettent à la fonction de variance d'être mal spécifiée. Nous pouvons aisément le comprendre. Nous pouvons écrire l'équation transformée sous la forme

$$y_i/\sqrt{h_i} = \beta_0(1/\sqrt{h_i}) + \beta_1(x_{i1}/\sqrt{h_i}) + \ldots + \beta_k(x_{ik}/\sqrt{h_i}) + u_i/\sqrt{h_i}.$$

Si $(u_i|\mathbf{x}_i) \neq \sigma^2 h_i$, l'erreur pondérée est hétéroscédastique. Dans ce cas, nous pouvons simplement appliquer les écarts-types estimés robustes à l'hétéroscédasticité (de forme inconnue) après avoir estimé cette équation par la méthode des MCO. Souvenez-vous que les MCP ne sont rien d'autre que les MCO utilisés sur l'équation pondérée.

Pour illustrer l'effet de cette correction sur l'inférence statistique des estimateurs MCP, nous comparons les écarts-types robustes et non robustes à l'hétéroscédasticité. La colonne (1) du tableau 8.2 est identique à la dernière colonne du tableau 8.1 : elle reprend les écarts-types non robustes des MCP. Les écarts-types des

MCP robustes à l'inégalité $\text{Var}(u_i|\mathbf{x}_i) \neq \sigma^2 inc_i$, sont indiqués dans la colonne (2). En d'autres termes, les écarts-types estimés de la colonne (2) sont correctement estimés même si la fonction de la variance est mal spécifiée.

Pour les variables relatives au revenu et à l'âge, les corrections augmentent la taille des écarts-types estimés. Les variables *male* et *e401k* illustrent l'effet inverse, avec des écarts-types estimés robustes plus petits que les écarts-types estimés habituels. Comme dans le cas des MCO, les écarts-types robustes calculés après estimation par les MCP ne sont pas systématiquement plus petits ou plus grands que ceux issus de l'approche standard.

Même lorsque nous utilisons une forme fonctionnelle de la variance aussi générale que celle de (8.30), nous n'avons aucune garantie qu'il s'agisse du bon modèle. La forme exponentielle est séduisante pour modéliser la présence d'hétéroscédasticité, notamment en raison de sa flexibilité, mais elle reste malgré tout un cas particulier. Par conséquent, il est toujours conseillé de calculer les écarts-types robustes, ainsi que les statistiques de tests correspondantes, après avoir estimé le modèle par la méthode des MCP.

Au cours de ces dernières années, la méthode des MCP a fait l'objet de critiques en raison du risque de mal spécifier l'équation de la variance. Rien ne garantit, en effet, que les estimateurs MCP soient plus efficients que les MCO. Si $\text{Var}(y|\mathbf{x})$ n'est ni constante ni égale à $\sigma^2 h(\mathbf{x})$, nous ne pouvons plus classer les estimateurs MCO et MCP en fonction de leurs variances (sachant que $h(\mathbf{x})$ représente la fonction retenue pour modéliser l'hétéroscédasticité). Cela est également vrai lorsque les paramètres de la variance doivent être estimés et que la comparaison porte sur les variances asymptotiques. Cette critique s'appuie néanmoins sur des éléments théoriques et omet un argument pratique important : en présence d'une forte hétéroscédasticité, il est souvent préférable d'appliquer les MCP en utilisant une mauvaise spécification de l'hétéroscédasticité plutôt que d'appliquer les MCO en l'ignorant totalement. Des modèles tels que (8.30) peuvent approximer de nombreuses formes d'hétéroscédasticité et permettre d'obtenir des estimateurs avec une variance (asymptotique) plus petite. L'exemple 8.6, où nous faisions l'hypothèse d'une hétéroscédasticité de forme simple, telle que $\text{Var}(nettfa|\mathbf{x}) = \sigma^2 inc$, en est une bonne illustration puisque les écarts-types robustes obtenus à l'aide des MCP étaient effectivement plus petits que leurs homologues des MCO. (En comparant les écarts-types robustes des deux estimateurs, nous les mettons sur un pied d'égalité : ni l'hypothèse d'absence d'hétéroscédasticité, ni celle de la présence d'une forme particulière d'hétéroscédasticité, du type $\sigma^2 inc$ pour la variance, n'est nécessaire). Par exemple, l'écart-type robuste de l'estimateur des MCP est 0,075 environ. Cette valeur est 25 % inférieure à celle des MCO (environ 0,100). Le constat est similaire en ce qui concerne la variable $(age - 25)^2$, puisque l'écart-type des MCP (environ 0,0026) est 40 % inférieur à celui des MCO (environ 0,0043).

Tableau 8.2 *Estimation par les mcp de l'équation* nettfa

Variables indépendantes	Écarts-types non robustes	Écarts-types robustes
inc	0,740 (0,064)	0,740 (0,075)
$(age - 25)^2$	0,0175 (0,0019)	0,0175 (0,0026)
male	1,84 (1,56)	1,84 (1,31)
e401k	5,19 (1,70)	5,19 (1,57)
constante	– 16,70 (1,96)	– 16,70 (2,24)

Variables indépendantes	Écarts-types non robustes	Écarts-types robustes
Observations	2 017	2 017
R carré	0,1115	0,1115

Prévisions et intervalles de prévision en présence d'hétéroscédasticité

Si nous partons du modèle linéaire standard sous les hypothèses RLM.1 à RLM.4, la présence d'hétéroscédasticité de la forme $\text{Var}(y|\mathbf{x}) = h(\mathbf{x})$ [voir l'équation (8.21)], sera susceptible d'affecter les prédictions ponctuelles de y uniquement à travers l'estimation de β_j. Il est naturel de chercher à estimer ce modèle par les MCP sur un échantillon de taille n afin d'obtenir $\hat{\beta}_j$. La prédiction de la valeur non observée, y^0, compte tenu des valeurs connues des variables explicatives $\mathbf{x}^0$, s'obtient en appliquant la formule de la section 6.4 : $\hat{y}^0 = \hat{\beta}_0 + \mathbf{x}^0\hat{\beta}$. Par exemple, le fait de connaitre $E(y|\mathbf{x})$ nous permet d'utiliser cette information pour calculer la valeur de notre prédiction. La structure de $\text{Var}(y|\mathbf{x})$, en revanche, ne joue aucun rôle direct dans ce calcul.

Il en va différemment du calcul des *intervalles de confiance des valeurs prédites* (ou *intervalles de prévisions*) puisqu'il dépend directement de la nature de $\text{Var}(y|\mathbf{x})$. Rappelons que dans la section 6.4, l'intervalle de prévision a été construit sous les hypothèses du MLC. Supposons maintenant que toutes les hypothèses MLC tiennent, à l'exception de l'hypothèse d'homoscédasticité, RLM.5, que nous remplaçons par (8.21). Nous savons que les estimateurs des MCP sont *BLUE* et suivent une distribution (conditionnelle) normale. Nous pouvons obtenir $\hat{\sigma}(\hat{y}^0)$ en appliquant la méthode présentée à la section 6.4, à l'aide des estimateurs des MCP. [Une approche simple consiste à écrire $y_i = \theta_0 + \beta_1(x_{i1} - x_1^0) + \ldots + \beta_k(x_{ik} - x_k^0) + u_i$, où x_j^0 désigne les valeurs des variables explicatives pour lesquelles nous souhaitons obtenir une valeur prédite de y. Nous pouvons estimer cette équation par les MCP, puis obtenir $\hat{y}_0 = \hat{\theta}_0$ et $\hat{\sigma}(\hat{y}^0) = \hat{\sigma}(\hat{\theta}_0)$.] Il nous faut également estimer l'écart-type de u^0, qui correspond à la partie non observée de y^0. Vu que $\text{Var}(u^0|\mathbf{x} = \mathbf{x}^0) = \sigma^2 h(\mathbf{x}^0)$, alors $\hat{\sigma}(u^0) = \hat{\sigma}\sqrt{h(\mathbf{x}^0)}$, avec $\hat{\sigma}$ désignant l'écart-type de la régression (ETR) par les MCP. Par conséquent, l'intervalle de prévision à 95 % s'écrit :

$$\hat{y}^0 \pm t_{0,025} \cdot \hat{\sigma}(\hat{e}^0). \qquad \textbf{[8.37]}$$

avec

$$\hat{\sigma}(\hat{e}^0) \pm \left\{[\hat{\sigma}(\hat{y}^0)]^2 + \hat{\sigma}^2 h(\mathbf{x}^0)\right\}^{1/2}.$$

Cet intervalle est exact, à condition de ne pas devoir estimer la fonction de la variance. Par contre, s'il est nécessaire de passer par une phase d'estimation (à l'instar du modèle 8.30), il nous est impossible d'obtenir un intervalle exact. En réalité, la prise en compte des erreurs d'estimations de $\hat{\beta}_j$ et $\hat{\delta}_j$ (variance des paramètres) rend le calcul des intervalles de confiance extrêmement ardu. Dans la section 6.4, nous avons rencontré deux exemples pour lesquels la taille de l'erreur d'estimation des paramètres était très petite, voire négligeable, par rapport à celle du terme non observable, u^0. Nous pourrions donc appliquer l'équation (8.37) en ne remplaçant que $h(\mathbf{x}^0)$ par $\hat{h}(\mathbf{x}^0)$. Si nous décidons d'ignorer entièrement l'erreur d'estimation des paramètres, nous pouvons simplifier l'expression de $\hat{\sigma}(\hat{e}^0)$ en négligeant le terme $\hat{\sigma}(\hat{y}^0)$. [Rappelez-vous que $\hat{\sigma}(\hat{y}^0)$ converge vers zéro au taux $1/\sqrt{n}$, alors que $\hat{\sigma}(\hat{u}^0)$ est à peu près constant.]

Nous pouvons également obtenir une prévision de y à partir du modèle suivant :

$$\log(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \qquad \textbf{[8.38]}$$

où u est hétéroscédastique. Nous supposons que u possède une distribution conditionnelle normale et faisons par ailleurs l'hypothèse d'une forme d'hétéroscédasticité spécifique. Plus précisément, nous adoptons la forme exponentielle de l'équation (8.30), en y ajoutant l'hypothèse de normalité des erreurs :

$$u|x_1, x_2, \ldots, x_k \sim \text{Normal}[0, \exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k)]. \qquad \textbf{[8.39]}$$

La notation peut être simplifiée si nous écrivons le modèle de la variance sous la forme $\exp(\delta_0 + \mathbf{x}\boldsymbol{\delta})$. La distribution conditionnelle de $\log(y)$, étant donné $\mathbf{x}$, suit une distribution normale de moyenne $\beta_0 + \mathbf{x}\boldsymbol{\delta}$ et de variance $\exp(\delta_0 + \mathbf{x}\boldsymbol{\delta})$. Il s'en suit que

$$E(y|\mathbf{x}) = \exp(\boldsymbol{\delta}_0 + \mathbf{x}\boldsymbol{\delta} + \sigma^2 \exp(\delta_0 + \mathbf{x}\boldsymbol{\delta})/2. \qquad \textbf{[8.40]}$$

Nous pouvons maintenant estimer $\hat{\beta}_j$ et δ_j en appliquant les MCP sur (8.38). Dans un premier temps, comme nous l'avons vu au préalable, les MCO permettent à la fois d'obtenir les résidus et d'estimer l'équation (8.32) afin de récupérer les valeurs prédites,

$$\hat{g}_i = \hat{\alpha}_0 + \hat{\delta}_1 x_{i1} + \ldots + \hat{\delta}_k x_{ik}. \qquad \textbf{[8.41]}$$

Ces prédictions permettent de calculer la valeur de $\hat{h}_i$ comme spécifié dans l'équation (8.33). À partir des $\hat{h}_i$ nous sommes en mesure d'obtenir les coefficients estimés par les MCP, $\hat{\beta}_j$, ainsi que de calculer le terme de variance $\hat{\sigma}^2$ à partir des résidus pondérés au carré. Comparativement au modèle initial de $Var(u \mid \boldsymbol{x})$, on a $\delta_0 = \alpha_0 + \log(\sigma^2)$, dès lors on obtient : $Var(u|\boldsymbol{x}) = \sigma^2 \exp(\alpha_0 + \delta_1 x_1 + \ldots + \delta_k x_k)$. La variance estimée est alors donnée par : $\sigma^2 \exp(\hat{g}_i) = \hat{\sigma}^2 \hat{h}_i$, et les valeurs prédites de y_i par :

$$\hat{y}_i = \exp(\widehat{logy}_i + \hat{\sigma}^2 \hat{h}_i / 2). \qquad \textbf{[8.42]}$$

Nous pouvons utiliser ces valeurs prédites en vue de calculer une mesure de R carré, tel que décrit à la section 6.4. À cette fin, nous utilisons le coefficient de corrélation porté au carré entre y_i et $\hat{y}_i$.

Pour toutes les valeurs des variables explicatives $\mathbf{x}^0$, nous pouvons estimer $E(y|\mathbf{x} = \mathbf{x}^0)$ comme suit :

$$\hat{E}(y|\mathbf{x} = \mathbf{x}^0) = \exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}} + \hat{\sigma}^2 \exp(\hat{\alpha}_0 + \mathbf{x}^0\hat{\boldsymbol{\delta}})/2), \qquad \textbf{[8.43]}$$

avec l'estimateur des MCP $\hat{\beta}_j$, la constante $\hat{\alpha}_j$ de (8.41), et les coefficients de pente $\hat{\delta}_j$ de la même régression.

Si la procédure de calcul des valeurs prédites reste relativement simple, le calcul d'écarts-types estimés pour l'équation (8.42) s'avère beaucoup plus ardu. En l'absence de formule analytique simple et à l'instar de ce qui a été fait dans la section 6.4, il est en général fréquent d'appliquer une méthode de ré-échantillonnage, comme le *bootstrap* décrite à l'annexe 6A.

En présence d'hétéroscédasticité, il est également difficile de calculer les intervalles de confiance pour les prévisions ; le calcul est lourd et compliqué. Il est toutefois possible de se simplifier la tâche. En effet, dans la section 6.4, nous avons vu deux exemples pour lesquels la variance de l'erreur était bien plus importante que l'erreur d'estimation des paramètres. Ignorer cette dernière n'a donc que peu d'incidence dans ces circonstances. En utilisant des arguments similaires à ceux de la section 6.4, les valeurs des bornes d'un intervalle de prévision à 95 % (pour les grandes tailles d'échantillon) sont $\exp[-1{,}96 \cdot \hat{\sigma}\sqrt{\hat{h}(\mathbf{x}^0)}]\exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$ et $\exp[1{,}96 \cdot \hat{\sigma}\sqrt{\hat{h}(\mathbf{x}^0)}]\exp(\hat{\beta}_0 + \mathbf{x}^0\hat{\boldsymbol{\beta}})$. La valeur prédite de la variance pour la valeur $\mathbf{x}^0$ est donnée par $\hat{h}(\mathbf{x}^0)$, soit $\hat{h}(\mathbf{x}^0) = \exp(\hat{\alpha}_0 + \hat{\delta}_1 x_1^0 + \ldots + \hat{\delta}_k x_k^0)$. À l'instar de la section 6.4, nous obtenons cet intervalle en calculant simplement l'exponentielle de ces bornes.

8.5 LE MODÈLE DE PROBABILITÉ LINÉAIRE REVISITÉ

Comme nous l'avons vu dans la section 7.5, les erreurs du modèle sont hétéroscédastiques lorsque la variable dépendante y est une variable binaire, à moins que tous les paramètres de pente soient nuls. Nous sommes maintenant en mesure de faire face à ce type de problème.

La façon la plus simple de traiter l'hétéroscédasticité dans le modèle de probabilité linéaire (MPL) consiste simplement à appliquer l'estimateur des MCO et de calculer les écarts-types robustes à l'hétéroscédasticité nécessaires à la réalisation de tests d'inférence statistique. Cette procédure ignore le fait que nous connaissons la forme précise que prend l'hétéroscédasticité dans le cas du MPL. Néanmoins, l'estimation du MPL par les MCO fournit souvent des résultats satisfaisants et d'autre part simple à réaliser.

EXEMPLE 8.8
Taux d'activité des femmes mariées

Dans l'exemple traitant de la participation des femmes mariées au marché du travail que nous avons étudié dans la section 7.5 [voir l'équation (7.29)], nous avions utilisé les écarts-types habituels estimés par la méthode des MCO. Nous complétons l'analyse en calculant les écarts-types estimés, robustes à l'hétéroscédasticité. Ceux-ci sont reportés entre crochets, juste en dessous des écarts-types estimés habituels :

$$\begin{aligned}\widehat{inlf} &= 0{,}586 + 0{,}0034\,nwifeinc + 0{,}038\,educ + 0{,}039\,exper \\ &\quad (0{,}154)\ (0{,}0014)\qquad (0{,}007)\qquad (0{,}006) \\ &\quad [0{,}151]\ [0{,}0015]\qquad [0{,}007]\qquad [0{,}006] \\ &\quad -00060\,exper^2 - 0{,}016\,age - 0{,}262\,kidslt6 + 0{,}0130\,kidsge6 \\ &\quad (0{,}00018)\qquad (0{,}002)\qquad (0{,}034)\qquad (0{,}0132) \\ &\quad [0{,}00019]\qquad [0{,}002]\qquad [0{,}032]\qquad [0{,}0135] \\ &\quad n = 753,\ R^2 = 0{,}264 \qquad\qquad [8.44]\end{aligned}$$

Pour plusieurs paramètres, la valeur des écarts-types robustes n'est pas différente que celle issue des MCO. En dépit de son importance sur le plan théorique, l'impact de l'hétéroscédasticité peut s'avérer marginal en pratique ; c'est effectivement le cas dans cet exemple. Les écarts-types standards estimés par les MCO et les tests statistiques correspondants donnent souvent des résultats assez similaires aux approches robustes. Il n'est pas compliqué de calculer les deux versions, standards et robustes, dans les logiciels économétriques.

En règle générale, les estimateurs des MCO appliqués au MPL ne sont pas efficients. Rappelons que la variance conditionnelle de la variable y dans le cas de ce modèle s'écrit :

$$\text{Var}(y|\mathbf{x}) = p(\mathbf{x})\,[1 - p(\mathbf{x})], \qquad [8.45]$$

où

$$p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \qquad [8.46]$$

est la probabilité de réponse ou probabilité de succès ($y = 1$). Il paraît naturel à première vue d'utiliser les MCP pour estimer le modèle. Pour y parvenir cependant, un certain nombre de petites difficultés sont à surmonter. Par exemple, la probabilité $p(\mathbf{x})$ dépend de paramètres inconnus de la population. Nous savons cependant que l'application des MCO nous permet d'obtenir des estimateurs sans biais de ces paramètres. Une fois appliqués à l'équation (8.46), nous obtenons des valeurs prédites pour chaque observation i, grâce auxquelles nous pouvons calculer $\text{Var}(y_i|\mathbf{x}_i)$, soit

$$\hat{h}_i = \hat{y}_i(1 - \hat{y}_i), \qquad [8.47]$$

où $\hat{y}_i$ désigne la valeur prédite par les MCO pour l'observation i. Nous sommes maintenant en mesure d'appliquer les MCQG, à l'image de ce qui a été fait dans la section 8.4.

Malheureusement, le fait d'avoir estimé h_i pour chaque i ne signifie pas que nous pouvons passer directement à l'application des MCP. Le problème qu'il nous reste à résoudre a été brièvement évoqué dans la section 7.5. Pour que la procédure d'estimation puisse fonctionner, les valeurs prédites, $\hat{y}_i$, doivent appartenir à l'intervalle se situant entre 0 et 1. Si $\hat{y}_i < 0$ ou bien $\hat{y}_i > 1$, l'équation (8.47) montre que $\hat{h}_i$ est négatif. Si tel est le cas, il n'est pas possible d'appliquer les MCP puisqu'ils requièrent le calcul de la racine carré de $\hat{h}_i$. Sans elle, il est impossible de calculer la pondération de l'observation i, soit $1/\sqrt{\hat{h}_i}$. Rien ne garantit que les prédictions $\hat{y}_i$ obtenues par les MCO soient comprises entre 0 et 1.

Si la contrainte $0 < \hat{y}_i < 1$ est respectée pour tout i, les MCP peuvent être employés pour estimer le MPL. Il est cependant très fréquent de trouver des valeurs prédites en dehors de l'intervalle unité, notamment lorsque les observations sont nombreuses et que les probabilités de succès ou d'échec sont faibles. Dans un tel cas de figure, il est plus simple d'abandonner les MCP au profit d'un simple report des statistiques robustes à l'hétéroscédasticité, comme dans l'exemple portant sur la participation au marché du travail de (8.44). Une alternative consiste à ajuster les valeurs prédites qui sont inférieures à zéro ou supérieures à l'unité, avant d'appliquer les MCP. Par exemple, il est possible de corriger les valeurs prédites comme suit : $\hat{y}_i = 0{,}01$ si $\hat{y}_i < 0$ et $\hat{y}_i = 0{,}99$ si $\hat{y}_i > 1$. Malheureusement, cela requiert un choix arbitraire : pourquoi ne pas utiliser 0,001 et 0,999 à la place de 0,01 et 0,99 comme valeurs prédites, en l'occurrence ? Si de nombreuses valeurs prédites sont en dehors de l'intervalle unité, l'ajustement de ces valeurs peut affecter le résultat final. Dans ce cas, il est à nouveau préférable de se contenter des écarts-types robustes basés sur les MCO.

L'estimation du MPL par les MCP

1. Estimer le modèle par les MCO pour obtenir les valeurs prédites $\hat{y}$.

2. Repérer si des valeurs ajustées sont en dehors de l'intervalle unité. Si tel est le cas, opérer un ajustement afin de contraindre toutes les valeurs prédites du modèle dans l'intervalle unité. Sinon, passer directement à l'étape (3).

3. Calculer les variances estimées de l'équation (8.47).

4. Estimer l'équation

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

par MCP, en utilisant des poids, $1/\hat{h}_i$.

EXEMPLE 8.9
S'équiper d'un ordinateur personnel à l'université

Nous utilisons les données du fichier GPA1 pour estimer la probabilité pour un étudiant de posséder un ordinateur à l'université. Soit *PC*, une variable binaire égale à 1 si l'étudiant possède un ordinateur, et 0 sinon. La variable *hsGPA* désigne la moyenne des notes obtenues à la fin du cycle de l'enseignement secondaire ; *ACT* indique la note acquise par l'étudiant à l'issue d'un test d'évaluation des connaissances à l'entrée de l'université ; *Parcoll* représente une variable binaire égale à 1 si au moins un des parents est allé à l'université, et 0 sinon. (En raison de leur forte corrélation, les niveaux d'éducation de la mère et du père ne sont pas statistiquement significatifs lorsqu'ils sont inclus séparément dans le modèle.)

L'estimation par les MCO donne les résultats suivants :

$$\widehat{PC} = -0{,}0004 + 0{,}065\ hsGPA + 0{,}0006\ ACT + 0{,}221\ parcoll$$
$$(0{,}4905)\quad (0{,}137)\quad (0{,}0155)\quad (0{,}093)$$
$$[0{,}4888]\quad [0{,}139]\quad [0{,}0158]\quad [0{,}087]$$
$$n = 141,\ R^2 = 0{,}0415, \qquad \textbf{[8.48]}$$

À l'instar de l'exemple 8.8, il n'y a aucune différence notable entre les écarts-types standards et les écarts-types robustes. Nous pouvons malgré tout décider d'estimer ce modèle par les MCP. Comme toutes les valeurs prédites par les MCO se trouvent à l'intérieur de l'intervalle unité, aucun ajustement n'est requis. Nous obtenons :

$$\widehat{PC} = 0{,}026 + 0{,}033\ hsGPA + 0{,}0043\ ACT + 0{,}215\ parcoll$$
$$(0{,}477)\quad (0{,}130)\quad (0{,}0155)\quad (0{,}086)$$
$$n = 141,\ R^2 = 0{,}0464, \qquad \textbf{[8.49]}$$

Nous ne notons aucune différence substantielle entre les estimations des écarts-types MCP et MCO. Si nous analysons le coefficient de la seule variable statistiquement significative du modèle, soit le niveau d'éducation des parents (*Parcoll*), nous constatons que la probabilité de posséder un *PC* est d'environ 22 points de pourcentage plus élevée si au moins un des parents a fréquenté l'université.

RÉSUMÉ

Dans ce chapitre, nous avons commencé par examiner les propriétés des moindres carrés ordinaires en présence d'hétéroscédasticité. La présence d'hétéroscédasticité n'introduit pas de biais dans les estimateurs des paramètres du modèle, obtenus par les MCO. Les estimateurs conservent également leur propriété de convergence. En revanche, les écarts-types standards estimés par les MCO, ainsi que les tests d'inférence statistique, ne sont plus valables. Pour faire face à ce problème, nous avons montré qu'il était possible de corriger les écarts-types estimés pour les rendre « robustes » à la présence d'hétéroscédasticité dans les erreurs. La plupart des logiciels d'économétrie proposent des commandes qui permettent de calculer et d'afficher tant les écarts-types estimés que les statistiques robustes à l'hétéroscédasticité, y compris pour la statistique *F*.

Comment pouvons-nous savoir si le modèle étudié possède des erreurs hétéroscédastiques ? Nous avons présenté deux approches fréquemment utilisées pour tester la présence d'hétéroscédasticité : le test de Breusch-Pagan et un cas particulier du test de White. Dans les deux tests, les résidus des MCO, élevés au *carré*, caractérisent la variable dépendante de la régression auxiliaire. Dans le premier test, il s'agit de les régresser sur les variables x_j du modèle ; dans le second, ces résidus au carré sont régressés sur les valeurs prédites de ces variables x_j, en niveau et au carré. Dans les deux cas, l'étape suivante consiste à effectuer un test de Fisher d'exclusion globale, asymptotiquement valide. Il existe également une version alternative du test basée sur le multiplicateur de Lagrange.

En présence d'hétéroscédasticité, l'estimateur des MCO n'est plus le meilleur estimateur linéaire sans biais. Lorsque la forme d'hétéroscédasticité est connue, l'estimateur des moindres carrés *généralisés* (MCG) peut être utilisé. Cela revient à appliquer les moindres carrés pondérés (MCP) en vue d'obtenir un estimateur *BLUE*. Les statistiques de tests de l'estimateur par les MCP sont soit exactes lorsque le terme d'erreur est normalement distribué, soit asymptotiquement valides dans le cas contraire.

En règle générale, cependant, la forme de l'hétéroscédasticité est inconnue. Il convient alors d'estimer un modèle visant à caractériser l'hétéroscédasticité *avant* d'appliquer les MCP. L'estimateur des MCG qui en résulte est appelé l'estimateur des moindres carrés quasi généralisés (MCQG) : il est biaisé mais reste

convergent et asymptotiquement efficace. Les statistiques habituelles obtenues par application des MCP sont asymptotiquement valides. Cette méthode n'est valide que si les variances estimées sont strictement positives pour toutes les observations. Ce n'est parfois pas le cas en pratique : il est alors nécessaire de corriger les données.

Comme nous l'avons vu dans le chapitre 7, le modèle de probabilité linéaire (MPL), dans lequel la variable dépendante est binaire, possède nécessairement un terme d'erreur hétéroscédastique. Une façon simple de résoudre ce problème consiste à calculer des statistiques robustes à l'hétéroscédasticité. Notons également que, si toutes les valeurs prédites (c'est-à-dire les probabilités estimées) sont strictement comprises entre zéro et un, les MCP peuvent être utilisés pour obtenir des estimateurs asymptotiquement efficients.

MOTS-CLÉS

Écart-type robuste à l'hétéroscédasticité p. 325
Estimateur des moindres carrés généralisés (MCG) p. 337
Estimateur des moindres carrés quasi généralisés (MCQG) p. 342
Estimateur des moindres carrés pondérés (MCP) p. 338
Hétéroscédasticité de forme inconnue p. 323
Statistique *F* robuste à l'hétéroscédasticité p. 327
Statistique *LM* robuste à l'hétéroscédasticité p. 328
Statistique *t* robuste à l'hétéroscédasticité p. 325
Test de Breusch-Pagan (test BP) p. 332
Test de White p. 334

EXERCICES

1. Parmi les propositions suivantes, quelles sont celles susceptibles d'être causées par la présence d'hétéroscédasticité ?

i. Les estimateurs des MCO, $\hat{\beta}_j$, ne sont pas convergents.

ii. La statistique *F* habituelle ne suit plus une distribution F.

iii. Les estimateurs des MCO ne sont plus *BLUE*.

2. Considérons le modèle linéaire suivant pour expliquer la consommation mensuelle de bière :

$$beer = \beta_0 + \beta_1 inc + \beta_2\, price + \beta_3 educ + \beta_4\, female + u$$
$$\mathrm{E}(u|inc, price, educ, female) = 0$$
$$\mathrm{Var}(u|inc, price, educ, female) = \sigma^2 inc^2.$$

Écrivez le modèle transformé avec erreurs homoscédastiques.

3. Vrai ou faux : les MCP sont préférables aux MCO lorsqu'une variable importante du modèle a été omise.

4. Le modèle suivant a été estimé à partir des données du fichier GPA3, sur un échantillon composé d'étudiants inscrits aux premier et second semestres :

$$\widehat{trmgpa} = -2{,}12 + 0{,}900\, crsgpa + 0{,}193\, cumgpa + 0{,}0014\, tothrs$$
$$(0{,}55)\ (0{,}175)\qquad (0{,}064)\qquad (0{,}0012)$$
$$[0{,}55]\ [0{,}166]\qquad [0{,}074]\qquad [0{,}0012]$$

$$+ 0{,}0018\,sat - 0{,}0039\,hsperc + 0{,}351\,female - 0{,}157\,season$$
$$(0{,}0002) \quad (0{,}0018) \quad (0{,}085) \quad (0{,}098)$$
$$[0{,}0002] \quad [0{,}0019] \quad [0{,}079] \quad [0{,}080]$$
$$n = 269,\ R^2 = 0{,}465$$

La variable *trmgpa* correspond à la moyenne des notes obtenues pour les cours suivis durant un semestre à l'université ; *crsgpa* représente la moyenne pondérée pour *tous* les cours suivis depuis le début du parcours universitaire ; *tothrs* correspond à la moyenne des notes obtenues pour les cours suivis *au préalable* (avant le semestre) ; *sat* est le résultat obtenu à un test standardisé (qui est utilisé à l'entrée de l'université) ; *hsperc* donne le quantile dans lequel se situe l'étudiant à l'issue de ses études secondaires ; *female* est une variable muette relative au sexe ; *season* est également une variable muette égale à 1 si le sport de l'étudiant se pratique au premier semestre, et 0 sinon. Les estimations des écarts-types standards et des écarts-types robustes à l'hétéroscédasticité sont respectivement indiquées entre parenthèses et entre crochets.

i. Les variables *crsgpa*, *cumgpa*, et *tothrs*, ont-elles les effets estimés attendus ? Lesquelles de ces trois variables sont statistiquement significatives au seuil de 5 % ? Le résultat est-il sensible au choix des écarts-types estimés ?

ii. L'hypothèse H0 : $\beta_{crsgpa} = 1$ fait-elle sens ? Si oui, pourquoi ? Testez cette hypothèse contre l'hypothèse alternative bilatérale au seuil de 5 %, à l'aide des deux estimations d'écarts-types. Expliquez vos conclusions.

iii. Toujours à l'aide des deux écarts-types estimés, testez si la variable *season* influe sur les résultats. Est-ce que le niveau de significativité auquel l'hypothèse nulle peut être rejetée dépend de l'écart-type retenu ?

5. La variable *smokes* est une variable binaire, égale à 1 si une personne fume, et 0 sinon. À l'aide des données du fichier SMOKE, nous estimons un modèle de probabilité linéaire (MPL) qui explique le fait d'être fumeur :

$$\widehat{smokes} = 0{,}656 - 0{,}69\,\log(cigpric) + 0{,}12\,\log(income) - 0{,}029\,educ$$
$$(0{,}855) \quad (0{,}204) \quad (0{,}026) \quad (0{,}006)$$
$$[0{,}856] \quad [0{,}207] \quad [0{,}026] \quad [0{,}006]$$
$$+ 0{,}020\,age - 0{,}0026\,age^2 - 0{,}101\,restaurn - 0{,}026\,white$$
$$(0{,}006) \quad (0{,}00006) \quad (0{,}039) \quad (0{,}052)$$
$$[0{,}005] \quad [0{,}00006] \quad [0{,}038] \quad [0{,}050]$$
$$n = 807,\ R^2 = 0{,}062$$

Le variable *white* prend la valeur 1 si le répondant est de type caucasien, et 0 sinon. Les autres variables indépendantes sont définies dans l'exemple 8.7. Les estimations des écarts-types habituels et des écarts-types robustes à l'hétéroscédasticité sont indiqués respectivement entre parenthèses et entre crochets, sous les coefficients correspondants.

i. Observez-vous des différences importantes entre les deux méthodes de calcul des écarts-types ?

ii. Toutes choses égales par ailleurs (*ceteris paribus*), quel est l'effet attendu d'un prolongement des études égal à 4 ans, sur la probabilité de fumer ?

iii. À partir de quel moment la probabilité de fumer diminue avec l'âge ?

iv. Interprétez le coefficient de la variable binaire *restaurn* (égale à 1 si la personne vit dans un État ayant une législation qui restreint la consommation de cigarettes dans les restaurants ; et 0 sinon).

v. La 206ème personne dans la base de données présente les caractéristiques suivantes : *cigpric* = 67,44, *income* = 6 500, *educ* = 16, *age* = 77, *restaurn* = 0, *white* = 0, et *smokes* = 0. Calculez la probabilité prédite que cette personne fume. Commentez le résultat.

6. Il existe différentes façons de mettre en œuvre les tests de Breusch-Pagan et de White pour détecter la présence d'hétéroscédasticité. Une possibilité, qui n'est pas abordée dans cet ouvrage, consiste à effectuer la régression de

$$\hat{u}_i^2 \text{ sur } x_{i1}, x_{i2}, \ldots, x_{ik}, \hat{y}_i^2, i = 1, \ldots, n,$$

où $\hat{u}_i^2$ sont les résidus MCO et $\hat{y}_i^2$ les valeurs prédites. Nous pouvons ensuite tester la significativité jointe de x_{i1}, x_{i2}, …, x_{ik} et $\hat{y}_i^2$. (La régression doit évidemment contenir une constante.)

i. Quel est le nombre de degrés de liberté du test F utilisé pour détecter la présence d'hétéroscédasticité ?

ii. Expliquez pourquoi le R carré de la régression ci-dessus sera toujours inférieur à celui calculé lors des procédures de tests BP et White.

iii. Le point (ii) signifie-t-il que le nouveau test fournit toujours une p-valeur inférieure à celle obtenue à partir des statistiques de BP ou de White ? Expliquez.

iv. Supposons que quelqu'un suggère d'ajouter $\hat{y}_i$ parmi les régresseurs du nouveau test. Que doit-on en penser ?

7. Considérons un modèle appliqué à un ensemble d'employés

$$y_{i,e} = \beta_0 + \beta_1 x_{i,e,1} + \beta_2 x_{i,e,2} + \ldots + \beta_k x_{i,e,k} + f_i + v_{i,e},$$

où la variable non observée, f_i, capture « l'effet entreprise », c'est-à-dire l'effet que les caractéristiques propres de l'entreprise i peuvent avoir sur la variable dépendante. Le terme d'erreur, $v_{i,e}$, est spécifique à l'employé e de l'entreprise i. L'erreur *composite*, $u_{i,e} = f_i + v_{i,e}$, comme dans l'équation (8.28).

i. Supposons que $\text{Var}(f_i) = \sigma_f^2$, $\text{Var}(v_{i,e}) = \sigma_v^2$, et que f_i et $v_{i,e}$ ne sont pas corrélés. Montrez alors que $\text{Var}(u_{i,e}) = \sigma_f^2 + \sigma_v^2$. Appelez cette variance σ^2.

ii. Supposons maintenant que $v_{i,e}$ et $v_{i,g}$ ne sont pas corrélés, pour $e \neq g$,. Montrez que $\text{Cov}(u_{i,e}, u_{i,g}) = \sigma_f^2$.

iii. Soit $\bar{u}_i = m_i^{-1} \sum_{\hat{e}=1}^{m_i} u_{i,e}$, la moyenne des erreurs composites au sein d'une entreprise. Montrez que $\text{Var}(\bar{u}_i) = \sigma_f^2 + \sigma_v^2 / m_i$.

iv. Discutez l'intérêt de la partie (iii) pour l'application de la technique MCP à des données agrégées par entreprise, où le poids utilisé pour l'observation i est fonction de la taille de l'entreprise.

8. Les équations suivantes ont été estimées sur les données décrites dans la base de données ECONMATH. La première équation a été réalisée sur les données relatives aux seuls hommes et la deuxième aux seules femmes. Les troisièmes et quatrièmes équations quant à elles combinent les données pour les deux sexes.

$$\widehat{score} = 20.52 + 13.60\ colgpa + 0.670\ act$$
$$(3.72)\quad (0.94)\qquad (0.150)$$
$$n = 406,\ R^2 = .4025,\ SSR = 38,781.38$$

$$\widehat{score} = 13.79 + 11.89\ colgpa + 1.03\ act$$
$$(4.11)\quad (1.09)\qquad (0.18)$$
$$n = 408,\ R^2 = .3666,\ SSR = 48,029.82$$

$$\widehat{score} = 15.60 + 3.17\ male + 12.82\ colgpa + 0.838\ act$$
$$(2.80)\quad (0.73)\qquad (0.72)\qquad (0.116)$$
$$n = 814,\ R^2 = .3946,\ SSR = 87{,}128.96$$

$$\widehat{score} = 13.79 + 6.73\ male + 11.89\ colgpa + 1.03\ act + 1.72\ male \cdot colgpa - 0.364\ male \cdot act$$
$$(3.91)\ (5.55)\qquad (1.04)\qquad (0.17)\qquad (1.44)\qquad (0.232)$$
$$n = 814,\ R^2 = .3968,\ SSR = 86{,}811.20$$

i. Calculez la statitistique du test de Chow usuelle permettant de tester l'hypothèse nulle selon laquelle les modèles de régression sont identiques pour les hommes et les femmes du panel. Identifiez la *p*-valeur du test.

ii. Calculez la statistique du test de Chow usuelle permettant de tester l'hypothèse nulle d'égalité des paramètres de pente des équations estimées pour les groupes d'hommes et de femmes. Reportez la *p*-valeur du test.

iii. Disposez-vous de suffisamment d'information que pour calculer des versions robustes à l'hétéroscédasticité des statistiques de tests calculées aux questions (ii) et (iii) ? Justifiez.

EXERCICES SUR ORDINATEUR

C1. Considérons le modèle suivant pour expliquer les facteurs influençant les heures de sommeil :

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u.$$

i. Écrivez un modèle qui permette à la variance de u de différer entre hommes et femmes. Cet écart ne doit pas dépendre d'autres facteurs.

ii. Utilisez les données du fichier SLEEP75 afin d'estimer les paramètres du modèle d'hétéroscédasticité. (Ceci requiert l'estimation de l'équation par les MCO afin d'obtenir les résidus.) La variance estimée de u est-elle plus élevée pour les hommes ou pour les femmes ?

iii. La variance de u est-elle statistiquement différente entre hommes et femmes ?

C2. Utilisez les données du fichier HPRICE1 pour obtenir les écarts-types estimés robustes à l'hétéroscédasticité de l'équation (8.17). Discutez des différences entre ces valeurs et celles des écarts-types estimés habituels.

ii. Répétez la partie (i) pour l'équation (8.18).

iii. Qu'est-ce que cet exemple suggère concernant la présence d'hétéroscédasticité et l'effet de la transformation de la variable dépendante ?

C3. Appliquez le test de White complet [voir l'équation (8.19)] à l'équation (8.18) afin de détecter la présence d'hétéroscédasticité. Utilisez la forme chi-carré de la statistique, puis déterminez la *p*-valeur. Que peut-on conclure ?

C4. Utilisez le fichier VOTE1 pour cet exercice.

i. Estimez un modèle expliquant la variable, *voteA*, à l'aide des variables indépendantes, *prtystrA*, *demoCA*, log(*expendA*) et log(*expendB*). Calculez les résidus MCO, $\hat{u}_i$; régressez-les sur toutes les variables indépendantes. Expliquez la raison pour laquelle vous obtenez une valeur du R^2 égale à 0.

ii. Ensuite, calculez le test d'hétéroscédasticité de Breusch-Pagan. Utilisez pour cela la version du test construite à partir de la statistique *F*. Reportez la *p*-valeur.

iii. Appliquez le cas particulier du test d'hétéroscédasticité de White à l'aide de la formule utilisant la statistique *F*. Quelle certitude peut-on avoir quant à la présence d'hétéroscédasticité ?

C5. Utilisez les données du fichier PNTSPRD pour cet exercice.

i. *sprdcvr* est une variable binaire égale à 1 si l'écart de points observé en faveur de l'équipe favorite, lors d'une rencontre de basketball universitaire, franchit le seuil du *Las Vegas point spread*, qui mesure l'écart de points *attendu par les pronostiqueurs* ; elle prend la valeur 0, dans le cas contraire. La *valeur attendue* de *sprdcvr*, notée u, mesure donc la probabilité d'observer, lors d'une rencontre prise au hasard, un écart de points supérieur au *Las Vegas point spread*. Testez $H_0 : u = 0{,}5$ contre $H_1 : u \neq 0{,}5$ à un seuil de significativité de 10 %. Discutez le résultat obtenu. (*Astuce* : Il suffit d'effectuer un test de Student après avoir régressé *sprdcvr* sur une constante.)

ii. Parmi les 553 rencontres de l'échantillon, combien d'entre elles ont été jouées sur terrain neutre ?

iii. Estimez le modèle de probabilité linéaire

$$sprdcvr = \beta_0 + \beta_1 favhome + \beta_2 neutral + \beta_3 fav25 + \beta_4 und25 + u$$

Indiquez les résultats comme d'habitude, en précisant les estimations obtenues pour les écarts-types standards des MCO et les écarts-types robustes à l'hétéroscédasticité. Quelle est la variable la plus importante d'un point de vue statistique ? Qu'en est-il d'un point de vue pratique ?

iv. Expliquez la raison pour laquelle il n'y a pas d'hétéroscédasticité dans le modèle sous l'hypothèse nulle $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

(c) Utilisez la statistique *F* habituelle afin de tester l'hypothèse de la partie (iv). Que peut-on conclure ?

vi. Étant donné le résultat de l'analyse précédente, peut-on dire qu'il est possible de prédire systématiquement le franchissement du seuil du *Las Vegas spread* à partir de l'information disponible avant la rencontre ?

C6. Dans l'exemple 7.12, nous avons estimé un modèle de probabilité linéaire en vue d'expliquer les arrestations de jeunes gens en 1986 :

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u.$$

i. Estimez le modèle par les MCO à partir des données du fichier CRIME1. Vérifiez ensuite que toutes les valeurs prédites se trouvent bien entre 0 et 1. Quelles sont les valeurs minimales et maximales prédites ?

ii. Estimez l'équation par les moindres carrés pondérés (MCP), comme expliqué dans la section 8.5.

iii. Utilisez les estimateurs MCP pour déterminer si les variables *avgsen* et *tottime* sont conjointement significatives au seuil de 5 %.

C7. Utilisez les données du fichier LOANAPP pour cet exercice.

i. Estimez l'équation de la partie (iii) des exercices sur ordinateur C8, du chapitre 7. Calculez ensuite les écarts-types estimés robustes à l'hétéroscédasticité. Comparez l'intervalle de confiance de 95 % à celui obtenu à partir d'une approche non robuste.

ii. Calculez les valeurs prédites de la régression de la partie (i). Certaines de ces valeurs sont-elles inférieures à zéro ou supérieures à un ? En cas de réponse positive, quelles sont les implications pour la mise en œuvre des MCP ?

C8. Utilisez les données du fichier GPA1 pour cet exercice.

i. Appliquez les MCO afin d'estimer un modèle expliquant *colGPA* par *hsGPA*, *ACT*, *skipped*, et *PC*. Sauver les résidus.

ii. Calculez la version particulière du test d'hétéroscédasticité de White. À cette fin, calculez les valeurs prédites, $\hat{h}_i$, à partir de la régression auxiliaire de $\hat{u}_i^2$ sur $\widehat{colGPA}_i$, $\widehat{colGPA}_i^2$.

iii. Vérifiez que les valeurs prédites de la partie (ii) sont toutes strictement positives. Estimez ensuite le modèle par les MCP à l'aide des poids $1/\hat{h}_i$. À partir de ces résultats, comparez les effets respectifs de l'absentéisme et de la possession d'un PC sur la réussite scolaire selon que l'estimation est réalisée par les MCP ou les MCO. Qu'en est-il de la significativité statistique de ces variables ?

iv. Dans l'estimation MCP de la partie (iii), calculez les écarts-types estimés robustes à l'hétéroscédasticité. Cette approche permet de tenir compte d'une mauvaise spécification de la variance estimée dans la partie (ii) (voir question 8.4.) Est-ce que les écarts-types estimés changent beaucoup par rapport à la partie (iii) ?

C9. Dans l'exemple 8.7, nous avons estimé les paramètres d'une équation de demande de cigarettes à l'aide des MCO, avant d'utiliser l'estimateur des MCP.

i. Reportez les estimateurs MCO de l'équation (8.35).

ii. Calculez la variable $\hat{h}_i$ utilisée dans l'estimation des MCP de l'équation (8.36). Ensuite, répliquez les résultats de l'équation (8.36). À partir de cette équation, obtenez les résidus non pondérés ainsi que les valeurs prédites. Désignez-les respectivement par $\hat{u}_i$ et $\hat{y}_i$. (Par exemple, dans le logiciel Stata, les résidus non pondérés et les valeurs prédites sont fournis automatiquement.)

iii. Soit $\breve{u}_i = \hat{u}_i/\sqrt{\hat{h}_i}$ et $\breve{y}_i = \hat{y}_i/\sqrt{\hat{h}_i}$ les valeurs pondérées. Appliquez le cas particulier du test d'hétéroscédasticité de White en régressant $\breve{u}_i^2$ sur $\breve{y}_i$, $\breve{y}_i^2$, sans oublier, comme toujours, d'inclure une constante dans le modèle. Trouvez-vous de l'hétéroscédasticité dans les résidus pondérés ?

iv. Quelle conclusion tirez-vous des résultats de la partie (iii) concernant la forme de l'hétéroscédasticité utilisée pour obtenir (8.36) ?

v. Pour les estimateurs des MCP, estimez les écarts-types qui restent valides en cas de mauvaise spécification de la variance.

C10. Utilisez l'ensemble des données du fichier 401KSUBS pour cet exercice.

i. Appliquez les MCO afin d'estimer un modèle de probabilité linéaire (MPL) expliquant *e401k* à l'aide de *inc*, inc^2, *age*, age^2, et *male*. Estimez les écarts-types des MCO habituels et les écarts-types robustes à l'hétéroscédasticité. Existe-t-il des différences importantes ?

ii. Dans le version particulière du test d'hétéroscédasticité de White, les termes des résidus MCO élevés au carré sont régressés sur une forme quadratique des valeurs prédites, tel que $\hat{u}_i^2$ sur $\hat{y}_i$, $\hat{y}_i^2$, $i = 1, \ldots, n$. Montrez que la limite en probabilité du coefficient de $\hat{y}_i$ doit être égale 1, que celle du coefficient $\hat{y}_i^2$ doit être égale à – 1, et enfin que la limite en probabilité de la constant doit être 0. {*Astuce* : Rappelons que $\text{Var}(y|x_1, \ldots, x_k) = p(x)\,[1 - p(x)]$, où $p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$.}

iii. Pour le modèle estimé à partir de la partie (i), calculez le test de White, puis examinez si les estimations des coefficients correspondent (plus ou moins) aux valeurs théoriques décrites dans la partie (ii).

iv. Après avoir vérifié que les valeurs prédites de la partie (i) sont toutes comprises entre 0 et 1, obtenez les estimateurs des moindres carrés pondérés du PML. Diffèrent-ils de façon importante des estimateurs des MCO ?

C11. Utilisez les données du fichier 401KSUBS pour cette question, en restreignant l'échantillon à *fsize* = 1.

i. Ajoutez le terme d'interaction *e401k·inc* au modèle dont les estimations sont indiquées dans le tableau 8.1. Estimez l'équation par les MCO et calculez les écarts-types estimés habituels ainsi que

les écarts-types estimés robustes. Que peut-on conclure au sujet de la significativité statistique du terme d'interaction ?

ii. Estimez maintenant le modèle le plus complet du tableau 8.1 en appliquant les MCP à partir de la pondération, $1/inc_i$. Calculez les écarts-types estimés habituels et les écarts-types estimés robustes de l'estimateur des MCP. Le terme d'interaction est-il statistiquement significatif lorsque vous utilisez la méthode robuste ?

iii. Discutez la valeur du coefficient de *e401k* obtenu par application des MCP sur le modèle le plus complet. Cette valeur revêt-elle un intérêt particulier ? Expliquez.

iv. Estimez de nouveau le modèle par MCP en incluant cette fois-ci la variable d'interaction, $e401k \cdot (inc - 30)$, parmi les régresseurs. Sachant que le revenu moyen dans l'échantillon est d'environ 29,44, interprétez le coefficient de *e401k*.

C12. Utilisez les données du fichier MEAP00 afin de répondre aux questions suivantes.

i. Estimez le modèle

$$math4 = \beta_0 + \beta_1 lunch + \beta_2 \log(enroll) + \beta_3 \log(exppp) + u$$

par les MCO, puis calculez les écarts-types estimés habituels et robustes. Quelle est généralement la différence attendue entre les deux ?

ii. Appliquez le cas particulier du test d'hétéroscédasticité de White. Quelle est la valeur du test F ? Quelle conclusion pouvez-vous tirer ?

iii. Obtenez $\hat{g}_i$ en calculant la valeur prédite à partir de la régression auxiliaire de $\log(\hat{u}_i^2)$ sur $\widehat{math4}_i$, $\widehat{math4}_i^2$ où $\widehat{math4}_i$ sont les valeurs MCO prédites et $\hat{u}_i$ est le résidu MCO. Soit $\hat{h}_i = \exp(\hat{g}_i)$. Utilisez $\hat{h}_i$ pour appliquer les MCP. Existe-il des différences importantes avec les coefficients MCO ?

iv. Estimez les écarts-types des MCP, qui restent valides en dépit d'une mauvaise spécification de la variance. Diffèrent-ils beaucoup des écarts-types standards des MCP ?

v. On cherche à analyser les effets des dépenses sur la variable *math4*. Quelle méthode d'estimation, entre les MCO et les MCP, est susceptible de fournir les estimateurs les plus précis ?

C13. Utilisez les données du fichier FERTIL2 afin de répondre aux questions suivantes.

i. Estimez le modèle

$$children = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educ + \beta_4 electric + \beta_5 urban + u.$$

Indiquez les estimations des écarts-types habituels et robustes à l'hétéroscédasticité. Ces derniers sont-ils toujours les plus élevés ?

ii. Ajoutez les trois variables muettes relatives à la religion afin de tester si elles sont conjointement significatives. Quelles sont les p-valeurs des statistiques robustes ? Même question pour les statistiques non robustes.

iii. Calculez les valeurs prédites, $\hat{y}$ ainsi que les termes du résidu, $\hat{u}$ à partir de la régression de la partie (ii). Puis, régressez $\hat{u}^2$ sur $\hat{y}$, $\hat{y}^2$ afin de tester la significativité jointe des deux variables explicatives. Peut-on conclure que l'hétéroscédasticité du modèle est liée à la variable *children* ?

iv. Peut-on dire que l'hétéroscédasticité trouvée dans la partie (iii) est importante en pratique ?

C14. Utilisez les données du fichier BEAUTY pour cette question.

i. Utilisez les données regroupant les hommes et les femmes pour estimer l'équation

$$lwage = \beta_0 + \beta_1 belavg + \beta_2 abvavg + \beta_3 female + \beta_4 educ + \beta_5 exper + \beta_6 exper^2 + u,$$

Reportez les résultats en indiquant la valeur des écarts-types estimés robustes à l'hétéroscédasticité sous les coefficients correspondants. Est-ce que le signe et la taille des coefficients sont surprenants ? Est-ce que le coefficient de la variable, *female*, est important sur les plans pratique et statistique ?

ii. Dans le modèle de la partie (i), ajoutez des variables d'interactions entre *female* et toutes les autres variables explicatives (soit cinq interactions en tout). Calculez le test F habituel de significativité jointe des cinq interactions ; faites de même avec une version robuste à l'hétéroscédasticité du test. L'utilisation de la version robuste change-t-elle le résultat d'une façon importante ?

iii. Dans le modèle complet intégrant des interactions, déterminez si celles décrivant la beauté physique, soit *female·belavg* et *female·abvavg*, sont conjointement significatives. Leurs coefficients sont-ils de petite taille sur le plan pratique ?

CHAPITRE 9

COMPLÉMENTS SUR LA SPÉCIFICATION ET LA QUESTION DES DONNÉES

Traduction de Cédric Heuchenne

Dans le chapitre 8, nous avons étudié le cas de la violation d'une des cinq hypothèses de Gauss-Markov, celle d'homoscédasticité. Si l'hétéroscédasticité dans les erreurs peut être vue comme un problème de modélisation, ce dernier est relativement mineur. La présence d'hétéroscédasticité n'introduit pas de biais dans les estimateurs des moindres carrés ordinaires (MCO), qui conservent également leur propriété de convergence. La présence d'hétéroscédasticité ne complexifie pas réellement le calcul des intervalles de confiance et des statistiques t et F ; il est relativement aisé d'estimer des écarts-types robustes, après estimation par les MCO, ou de calculer des estimateurs plus efficaces en utilisant les moindres carrés pondérés (MCP).

Dans ce chapitre, nous revenons au problème plus grave de corrélation entre l'erreur et une ou plusieurs des variables explicatives. Pour rappel (chapitre 3), si u est, pour une raison quelconque, corrélée avec la variable explicative x_j, nous disons que x_j est une **variable explicative endogène**. Nous discuterons en détail des raisons pour lesquelles une variable explicative est endogène ; dans certains cas, nous serons capables de régler ce problème d'endogénéité.

Nous avons déjà vu aux chapitres 3 et 5 que l'omission d'une variable importante peut introduire de la corrélation entre l'erreur et certaines variables explicatives, ce qui généralement biaise tous les estimateurs des MCO et les prive de leur propriété de convergence. Dans le cas particulier où la variable omise est fonction d'une ou plusieurs variables explicatives *déjà présentes* dans le modèle, ce dernier souffre d'une **erreur de spécification de la forme fonctionnelle**.

Nous débutons la première section par l'étude des conséquences d'une erreur de spécification dans la forme fonctionnelle, que nous apprenons également à détecter. Dans la section 9.2, nous démontrons que l'utilisation d'une variable de substitution peut résoudre, ou pour le moins atténuer, le biais dû à une variable omise. Dans la section 9.3, nous expliquons et calculons le biais de l'estimateur des MCO qui peut provenir d'une **erreur de mesure**. D'autres problèmes spécifiques, liés à l'utilisation de données, sont abordés dans la section 9.4.

Toutes les procédures de ce chapitre sont basées sur l'estimation par les MCO. Comme nous le verrons, certains problèmes, qui génèrent de la corrélation entre les erreurs et certaines variables explicatives, ne peuvent pas être résolus en recourant simplement aux MCO. Nous étudierons ces méthodes d'estimation alternatives dans la partie 3.

9.1 ERREUR DE SPÉCIFICATION DE LA FORME FONCTIONNELLE

Un modèle de régression linéaire multiple (RLM) souffre d'erreur de spécification de la forme fonctionnelle quand il ne tient pas correctement compte de la relation entre la variable dépendante, d'une part, et les variables explicatives, d'autre part. Par exemple, si le **salaire horaire** est déterminé par $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$, mais que le terme d'expérience au carré, $exper^2$, est omis, alors nous commettons une erreur de spécification de la forme fonctionnelle. Dans le chapitre 3, nous avons déjà vu que cela conduisait généralement à des estimateurs biaisés pour β_0, β_1, et β_2 (Nous ne pouvons pas estimer β_3 puisque $exper^2$ est exclu du modèle.) Donc, le fait de mal spécifier la manière avec laquelle *exper* affecte log(*wage*) entraîne généralement un estimateur biaisé du rendement du niveau d'instruction β_1, l'ampleur de ce biais dépendant de la taille de β_3 et des corrélations entre *educ*, *exper* et $exper^2$.

L'estimation du rendement de l'expérience est particulièrement affectée par cette mauvaise spécification du modèle : même si nous pouvions obtenir un estimateur sans biais de β_2, nous ne serions pas capables d'estimer le rendement de l'expérience puisqu'il vaut $\beta_2 + 2\beta_3 exper$ (sous forme décimale). Le fait d'utiliser uniquement l'estimateur biaisé de β_2 peut mener à des erreurs, surtout lorsque la valeur de *exper* est élevée.

Un autre exemple concerne log(*wage*) dans l'équation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female + \beta_5 female.educ + u, \quad [9.1]$$

où *female* est une variable binaire. Si nous omettons le terme d'interaction, *female·educ* alors nous commettons une erreur de spécification de la forme fonctionnelle. En général, les autres paramètres seront biaisés. Par ailleurs, sans considérer le terme d'interaction, la signification du rendement du niveau d'instruction n'est pas claire, puisque ce rendement dépend du genre de l'individu.

Omettre des fonctions de variables indépendantes, comme $exper^2$ ou *female·educ*, n'est pas la seule manière d'obtenir un modèle dont la forme fonctionnelle est mal spécifiée. Par exemple, si (9.1) est le vrai modèle qui satisfaisait les quatre premières hypothèses de Gauss-Markov, mais que *wage* est utilisé comme variable dépendante plutôt que log(*wage*), alors nous n'obtiendrons pas d'estimateur sans biais ou convergent des effets *ceteris paribus*. Les tests que nous allons aborder dans cette section, peuvent détecter ce type de problème, mais il existe des tests plus appropriés, que nous mentionnerons dans une autre sous-section et qui comparent des modèles non imbriqués.

Une mauvaise spécification de la forme fonctionnelle d'un modèle peut avoir de lourdes conséquences. Le problème n'est pas si difficile à surmonter pour autant : notez bien que nous avons à notre disposition toutes les variables nécessaires à l'identification de la bonne forme fonctionnelle. Dans la prochaine section, nous verrons qu'il est plus ardu de corriger une forme fonctionnelle lorsqu'une variable clé est omise et qu'aucune donnée ne peut être collectée pour la mesurer.

Nous avons déjà un outil très puissant pour détecter les formes fonctionnelles mal spécifiées : le test *F* qui permet de formuler des contraintes d'exclusion jointes. Par exemple, il est souvent important d'ajouter les termes quadratiques des variables significatives d'un modèle et d'effectuer un test de significativité jointe. Si les termes quadratiques additionnels sont significatifs, ils peuvent être ajoutés au modèle (au prix d'une complication dans l'interprétation du modèle). Cependant, la significativité de termes quadratiques peut être symptomatique d'autres problèmes de forme fonctionnelle, tels que celui lié à l'utilisation de la variable en niveau alors que son logarithme serait plus approprié (ou vice versa). Il peut être difficile d'identifier la raison pour laquelle une forme fonctionnelle est mal spécifiée. Heureusement, dans beaucoup de cas, utiliser les logarithmes de certaines variables ou ajouter des termes quadratiques est suffisant pour détecter un bon nombre de relations non linéaires importantes en économie.

EXEMPLE 9.1
Modélisation du nombre d'arrestations

Le tableau 9.1 contient les estimations par MCO d'un modèle sur le risque de récidive (voir les exemples 3.5 et 8.3). Le modèle est d'abord estimé sans terme quadratique ; ces résultats se trouvent en colonne (1).

Dans la colonne (2), les carrés de *pcnv*, *ptime86* et *inc86* sont ajoutés ; ces variables sont ajoutées car chaque terme linéaire (de niveau) est significatif en colonne (1). Comme la variable *qemp86* est une variable discrète qui ne prend que cinq valeurs différentes (de zéro à quatre ; elle représente le nombre de trimestres durant lesquels l'individu a travaillé), son carré n'est pas considéré dans la colonne (2).

Chaque terme au carré est significatif et, ensemble, ils sont également très significatifs ($F = 31{,}37$, avec $ddl = 3$ et 2 713 ; la *p*-valeur vaut essentiellement zéro). Il apparaît donc que le modèle initial néglige certaines non-linéarités potentielles.

La présence de plusieurs termes quadratiques rend l'interprétation du modèle quelque peu compliquée. Par exemple, *pcnv* n'a plus maintenant d'effet dissuasif strictement positif : la relation entre *narr86* et *pcnv* est positive jusqu'à $pcnv = 0{,}365$; ensuite, elle devient négative. On pourrait conclure que l'effet dissuasif est limité lorsque les valeurs de *pcnv* sont faibles ; l'effet démarrerait seulement pour des taux de condamnation

plus élevés dans le passé. Pour vérifier cette conclusion, il serait néanmoins plus prudent d'utiliser d'autres formes fonctionnelles plus sophistiquées (que des termes quadratiques). Il se peut aussi que *pcnv* ne soit pas complètement exogène. Par exemple, les hommes qui n'ont pas été condamnés dans le passé (donc avec *pcnv* = 0), sont sans doute des criminels occasionnels ; ils seraient donc moins susceptibles d'être arrêtés au cours de l'année 1986. Cette réalité peut biaiser les estimations.

Pour aller plus loin 9.1

Pourquoi ne pas inclure les carrés de *black* et de *hispan* dans la colonne (2) du tableau 9.1 ?

Y aurait-il du sens à ajouter des termes d'interactions de *black* et *hispan* avec certaines autres variables rapportées dans la table ?

D'une manière semblable, la relation entre *narr86* et *ptime86* est positive jusqu'à *ptime86* = 4,85 (presque cinq mois en prison) et devient négative par la suite. La grande majorité des hommes dans cet échantillon n'ont pas été en prison en 1986 ; à nouveau, il faut être prudent dans l'interprétation de ces résultats [et du point de retournement, puisque la relation négative qui intervient au-delà du seuil de 4,85 ne concerne pas nécessairement beaucoup de personnes].

Le revenu légal a un effet négatif sur *narr86* jusqu'à *inc86* = 242,85 ; comme le revenu est mesuré en centaines de dollars, cela signifie un revenu annuel de $24 285. Seulement 46 hommes dans l'échantillon ont des revenus supérieurs. Nous pouvons donc conclure que la relation entre *narr86* et *inc86* est négative mais décroissante.

Tableau 9.1 *Variable dépendante :* narr86

Variables indépendantes	(1)	(2)
pcnv	−0,133 (0,040)	0,533 (0,154)
$pcnv^2$	–	−0,730 (0,156)
avgsen	−0,011 (0,012)	−0,017 (0,012)
tottime	0,012 (0,009)	0,012 (0,009)
ptime86	0,041 (0,009)	0,287 (0,004)
$ptime86^2$	–	−0,0296 (0,0039)
qemp86	−0,051 (0,014)	−0,014 (0,017)
inc86	−0,0015 (0,0003)	−0,0034 (0,0008)
$inc86^2$	–	−0,000007 (0,000003)
black	0,327 (0,045)	0,292 (0,045)

Variables indépendantes	(1)	(2)
hispan	0,194 (0,040)	0,164 (0,039)
constante	0,596 (0,036)	0,505 (0,037)
Observations	2 725	2 725
R carré	0,0723	0,1035

L'exemple 9.1 constitue un problème de forme fonctionnelle relativement ardu, étant donné la nature de la variable dépendante (qui ne prend que quelques valeurs entières). D'autres approches sont théoriquement mieux adaptées pour modéliser ce type de variables dépendantes. Nous les aborderons brièvement dans le chapitre 17.

RESET : un test général pour les erreurs de spécification de la forme fonctionnelle

Certains tests ont été proposés pour détecter des erreurs de spécification de la forme fonctionnelle. Le test **d'erreur de spécification de la régression** de Ramsey (1969) (« **regression specification error test** » – **RESET** –) est utile dans ce contexte.

L'idée derrière le RESET est assez simple. Si le modèle original

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \qquad [9.2]$$

satisfait l'hypothèse RLM.4, alors aucune fonction non linéaire des variables indépendantes ne devrait être significative une fois ajoutée à l'équation (9.2). Dans l'exemple 9.1, nous avons ajouté des termes quadratiques aux variables explicatives qui étaient significatives. Bien que cette technique détecte souvent des problèmes de forme fonctionnelle, le désavantage est qu'elle consomme un grand nombre de degrés de liberté lorsque le modèle original contient beaucoup de variables explicatives (comme dans le cas du test complet de White pour l'hétéroscédasticité). Ensuite, l'ajout de termes quadratiques ne capture pas d'autres formes de non-linéarité souvent négligées. C'est la raison pour laquelle le RESET ajoute des polynômes calculés sur base des valeurs ajustées des MCO dans l'équation (9.2), l'objectif étant de détecter des erreurs de spécification de forme fonctionnelle générales.

Pour implémenter le RESET, il est nécessaire de déterminer le nombre de fonctions non linéaires à inclure dans la régression étendue. Il n'existe pas de réponse toute faite à cette question, mais les carrés et les cubes se sont avérés utiles dans la plupart des applications.

Soit $\hat{y}$, les valeurs ajustées obtenues en estimant (9.2) par les MCO. Considérons l'équation étendue

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + \textit{erreur}. \qquad [9.3]$$

Cette équation semble un peu bizarre, car plusieurs fonctions des valeurs ajustées de l'équation de départ apparaissent dorénavant comme variables explicatives. En réalité, notre intérêt ne porte pas sur les estimations des paramètres β_j de l'équation (9.3) ; cette équation ne sert qu'à vérifier s'il manque des non-linéarités importantes dans (9.2). Notez bien que $\hat{y}^2$ et $\hat{y}^3$ correspondent à des fonctions non linéaires des x_j.

L'hypothèse nulle du RESET est que (9.2) est correctement spécifiée. Le RESET est basé sur la statistique F. L'objectif est de tester $H_0: \delta_1 = 0, \delta_2 = 0$ dans le modèle étendu (9.3). Une statistique F significative suggère un problème de forme fonctionnelle. En grands échantillons, sous l'hypothèse nulle (et les hypothèses de Gauss-Markov), la distribution de la statistique F suit approximativement une distribution $F_{2,n-k-3}$. Les degrés de liberté (*ddl*) dans l'équation étendue (9.3) sont $n-k-1-2 = n-k-3$. Une

version *LM* est également disponible ; sa distribution du chi-carré a un *ddl* égal à deux. Enfin, le test peut être rendu robuste à l'hétéroscédasticité en utilisant les méthodes discutées dans la section 8.2.

EXEMPLE 9.2
Équation du prix des maisons

Estimons deux modèles pour le prix des maisons. Le premier comprend toutes les variables sous forme linéaire :

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u. \quad [9.4]$$

Le second utilise les logarithmes de toutes les variables sauf *bdrms* :

$$lprice = \beta_0 + \beta_1 llotsize + \beta_2 lsqrft + \beta_3 bdrms + u. \quad [9.5]$$

En utilisant $n = 88$ maisons disponibles dans HPRICE1, la statistique RESET pour l'équation (9.4) vaut 4,67, soit la valeur d'une variable aléatoire $F_{2,82}$ ($n = 88$, $k = 3$). La p-valeur associée est 0,012. Ce résultat indique que la spécification de la forme fonctionnelle dans (9.4) n'est pas la bonne.

La statistique RESET dans (9.5) vaut 2,56 avec une p-valeur = 0,084. Le modèle (9.5) n'est donc pas rejeté à un seuil de significativité égal à 5 % (à un niveau de 10 %, il le serait). Sur base du RESET, le modèle log-log (9.5) a notre préférence.

Dans l'exemple précédent, nous avons testé deux modèles dont l'objectif était d'expliquer le prix de biens immobiliers. Le premier a été rejeté par le RESET, mais pas le second (à un seuil de 5 %, à tout le moins). Dans certains cas, les résultats ne sont pas aussi catégoriques. Le RESET n'indique pas non plus de réelle direction à suivre en cas de rejet d'un modèle. Le rejet de (9.4) par le RESET ne nous a pas suggéré qu'il fallait passer à la spécification de (9.5). L'équation (9.5) a été estimée pour la simple raison que les modèles à élasticité constante sont faciles à interpréter et jouissent de propriétés statistiques intéressantes.

Le RESET est parfois considéré comme un test *global* de détection d'erreurs dans la spécification du modèle, capable d'identifier les problèmes d'hétéroscédasticité ou de variables non mesurables (et donc) omises du modèle. En réalité, le RESET n'est pas aussi puissant. Tout d'abord, le RESET est incapable de détecter le biais lié à l'omission d'une variable non observable lorsque cette variable est une fonction *linéaire* des variables indépendantes déjà incluses dans le modèle. [Voir Wooldridge (1995) pour l'énoncé précis]. Ensuite, si la forme fonctionnelle s'avère être correctement spécifiée, le RESET sera incapable de détecter la présence d'hétéroscédasticité. Le RESET est un test portant sur la forme fonctionnelle ; rien de moins, rien de plus.

Tests de modèles non emboîtés

Lorsqu'il s'agit de détecter d'autres formes d'erreur de spécification (en comparant directement des modèles linéaire et logarithmique, par exemple), le RESET ne convient pas ; nous devons utiliser d'autres méthodes. Par exemple, supposons que nous cherchions à comparer la forme fonctionnelle du modèle

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \quad [9.6]$$

à celle du modèle

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u. \quad [9.7]$$

Comme il s'agit de **modèles non emboités** ou non imbriqués (voir chapitre 6), un test F classique ne convient donc pas. Deux approches ont été suggérées. La première consiste à construire un modèle englobant, dont les modèles (9.6) et (9.7) sont des cas particuliers, et à tester ensuite les restrictions qui mènent à chacun des deux. Dans notre exemple, le modèle englobant est

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log(x_1) + \gamma_4 \log(x_2) + u. \quad [9.8]$$

Pour tester (9.6), nous pouvons écrire $H_0 : \gamma_3 = 0, \gamma_4 = 0$. Pour le modèle particulier (9.7), $H_0 : \gamma_1 = 0, \gamma_2 = 0$. Cette approche a été suggérée par Mizon et Richard (1986).

Une autre approche a été proposée par Davidson et MacKinnon (1981). Ils considèrent que, si (9.6) est le vrai modèle, alors les valeurs ajustées de l'autre modèle, (9.7), ne doivent pas être significatives dans (9.6). Par conséquent, pour tester la validité de (9.6), nous devons d'abord estimer le modèle (9.7) par les MCO pour en obtenir les valeurs ajustées, soit $\check{y}$. Ensuite, le **test de Davidson-MacKinnon** consiste à tester la significativité du paramètre de $\check{y}$ dans l'équation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \check{y} + erreur.$$

Une statistique t significative (contre une alternative bilatérale) entraîne le rejet de (9.6).

De manière similaire, si $\hat{y}$ correspond aux valeurs ajustées du modèle (9.6), le test de la forme fonctionnelle de (9.7) porte sur la statistique t de $\hat{y}$ dans le modèle

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \theta_1 \hat{y} + erreur.$$

Une statistique t significative conduit à un rejet de (9.7). Notez bien que ces approches ne sont valides que si les deux modèles non emboîtés partagent la même variable dépendante, exprimée de la même manière [dans les deux modèles, (9.6) et (9.7), y est en niveau].

Nous pouvons néanmoins faire face à quelques problèmes lorsqu'il s'agit de tester des modèles non emboîtés. Le premier problème survient lorsqu'il est impossible d'identifier le meilleur modèle parmi les deux modèles candidats. Par exemple, nous pouvons obtenir deux rejets ou deux non rejets. Dans le dernier cas, nous pouvons éventuellement utiliser le R carré ajusté pour effectuer notre choix. Par contre, si les deux modèles sont rejetés, il reste du travail à accomplir car la forme fonctionnelle doit être repensée. Notons qu'il est important de bien saisir les conséquences que peut avoir, sur le plan pratique, le choix d'une forme fonctionnelle en particulier. Dans certains cas, les variables significatives dans les deux modèles ne seront pas les mêmes. Par contre, si les effets des variables indépendantes clés sur la variable y sont similaires dans les deux modèles, alors le choix final de la forme fonctionnelle importe peu.

Le deuxième problème peut être énoncé comme suit : la décision de rejeter (9.6), sur base du test de Davidson-MacKinnon par exemple, *ne* signifie *pas* que (9.7) est nécessairement le bon modèle. Le modèle (9.6) peut être rejeté en raison d'une spécification particulièrement médiocre de la forme fonctionnelle, ce qui n'implique pas que la spécification du modèle alternatif est remarquable.

Enfin, les tests non emboîtés ne sont pas directement applicables lorsque les modèles concurrents ont des variables dépendantes différentes. Le cas le plus courant est y sur $\log(y)$. Dans le chapitre 6, nous avons vu que les mesures de qualité d'ajustement [comme le R carré ajusté] ne pouvaient pas être directement comparées dans un tel cas de figure. Des tests plus élaborés permettent de résoudre ce problème, mais ils sont au-delà de la portée de cet ouvrage. [Voir Wooldridge (1994a) pour un test dont l'implémentation et l'interprétation sont aisées.]

9.2 UTILISATION DE VARIABLES DE SUBSTITUTION

Un problème plus difficile à résoudre surgit lorsqu'une variable clé ne peut pas être incluse dans un modèle parce qu'il est impossible de la mesurer et donc de récolter des données à son sujet. Considérons l'équation de salaire dans laquelle les aptitudes « innées » (*abil*) affectent le salaire sous forme logarithmique, $\log(wage)$:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u. \qquad \textbf{[9.9]}$$

Ce modèle indique clairement qu'il est important de mesurer l'effet *ceteris paribus* (ou effet partiel) des variables *educ* et *exper* en tenant compte des aptitudes. Par exemple, si la variable *educ* est corrélée avec *abil*, le fait d'ignorer la variable *abil* en la laissant dans le terme d'erreur, entraîne un biais dans l'estimateur des MCO de β_1 (et de β_2). Nous avons rencontré ce problème à plusieurs reprises dans les chapitres précédents.

Dans l'équation (9.9), notre attention se porte avant tout sur les paramètres de pente β_1 et β_2. Obtenir un estimateur sans biais ou convergent de l'ordonnée à l'origine (β_0) n'est pas très intéressant ; en règle générale, c'est d'ailleurs impossible. Nous ne pouvons pas non plus estimer β_3 puisque la variable *abil* n'est pas observée. Les aptitudes d'un individu étant un concept pour le moins vague, il n'y aurait de toute façon pas grand intérêt à estimer et interpréter β_3.

Comment pouvons-nous faire disparaître ou atténuer le biais provenant de l'omission d'un facteur important, que nous ne pouvons pas observer ? Une stratégie est d'identifier une **variable de substitution**, qui permet de remplacer la variable omise dans le modèle. Pour faire simple, une variable de substitution correspond à un facteur proche de celui que nous aimerions prendre en considération dans notre analyse. Dans l'équation de salaire, nous pourrions utiliser les résultats à un test de quotient intellectuel (QI) comme variable de substitution aux aptitudes. Le QI ne doit pas nécessairement correspondre parfaitement aux aptitudes ; Il est par contre nécessaire que le QI soit suffisamment corrélé aux aptitudes innées, comme nous allons maintenant l'expliquer.

Considérons un modèle à trois variables indépendantes, dont seules deux peuvent être observées :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u. \qquad \textbf{[9.10]}$$

Nous supposons que les données sont disponibles pour y, x_1 et x_2 (dans l'exemple du salaire, celles-ci correspondent à log(*wage*), *educ* et *exper*, respectivement). La variable explicative x_3^* n'est pas observée mais nous disposons d'une variable de substitution, soit x_3.

Que devons-vous exiger de la part de x_3 ? Qu'il existe au minimum une relation entre elle et x_3^*. Cette relation peut être synthétisée à l'aide d'une régression simple, soit

$$x_3^* = \delta_0 + \delta_3 x_3 + v_3, \qquad \textbf{[9.11]}$$

où v_3 est une erreur provenant du fait que x_3^* et x_3 ne sont pas identiques. Le paramètre δ_3 mesure la relation entre x_3^* et x_3. Par exemple, dans le cas typique, x_3^* et x_3 seront positivement liées, de telle sorte que $\delta_3 > 0$. Si $\delta_3 = 0$ alors x_3 n'est pas une variable de substitution valable pour x_3^*. L'ordonnée à l'origine δ_0 de (9.11) peut être positive ou négative ; elle permet simplement de capturer l'effet lié à l'utilisation d'une échelle de mesure différente pour x_3^* et x_3. (Par exemple, le QI et l'aptitude non observée d'un individu n'ont pas nécessairement la même valeur moyenne dans la population des États-Unis d'Amérique.)

Comment pouvons-nous utiliser x_3 pour obtenir des estimateurs sans biais (ou au moins convergents) de β_1 et β_2 ? Une proposition évidente consiste à considérer que x_3 et x_3^* sont identiques et de régresser

$$y \text{ sur } x_1, x_2, x_3. \qquad \textbf{[9.12]}$$

Cette méthode revient donc à **remplacer la variable omise par la variable de substitution**. Dans notre exemple, avant de calculer les MCO, nous insérons x_3^* au lieu de x_3. Si x_3 est véritablement liée à x_3^*, cette méthode est censée. Cependant, comme x_3 et x_3^* ne sont pas identiques, nous devons être capable d'identifier les circonstances dans lesquelles cette procédure conduit effectivement à des estimateurs convergents de β_1 et β_2.

Les deux hypothèses sur lesquelles repose la propriété de convergence des estimateurs de β_1 et β_2 concernent u, d'une part, et v_3, d'autre part.

(1) L'erreur u n'est pas corrélée avec x_1, x_2 et x_3^*. Cette hypothèse correspond à une hypothèse classique du modèle (9.10). Ce n'est pas tout : le terme u n'est pas non plus corrélé avec x_3. Cette dernière hypothèse signifie juste que, si les variables x_1, x_2, et x_3^* sont incluses dans le modèle, x_3 est superflue. Par définition, cela doit être vrai : x_3 n'est qu'une variable de substitution ; c'est x_3^* qui affecte directement y. Par conséquent, cette hypothèse d'absence de corrélation entre u et les variables x_1, x_2, x_3^* et x_3, n'est pas controversée. (Cette hypothèse peut être énoncée de la manière suivante : étant donné toutes ces variables, l'espérance de u est nulle.)

(2) L'erreur v_3 n'est pas corrélée avec x_1, x_2 et x_3. Pour que v_3 ne soit pas corrélée avec x_1 et x_2, il faut que x_3 soit une « bonne » variable de substitution de x_3^*. Pour mieux le comprendre, reformulons cette hypothèse en utilisant l'espérance conditionnelle :

$$\mathrm{E}(x_3^*|x_1,x_2,x_3) = \mathrm{E}(x_3^*|x_3) = \delta_0 + \delta_3 x_3. \qquad \textbf{[9.13]}$$

La première égalité est la plus importante ; elle stipule que l'espérance de x_3^* ne dépend pas de x_1 ou x_2 lorsque l'influence de x_3 est prise en compte. Autrement dit, une fois que l'effet de x_3 sur x_3^* a été isolé, la corrélation entre x_3^* et les variables x_1 et x_2 est nulle.

Dans l'équation du salaire (9.9), où le QI sert de variable de substitution aux aptitudes, la condition (9.13) devient

$$\mathrm{E}(abil|educ,exper,IQ) = \mathrm{E}(abil|IQ) = \delta_0 + \delta_3 IQ.$$

Par conséquent, le niveau moyen des aptitudes ne change qu'avec le QI ; lorsque le QI est pris en compte, ce niveau moyen ne dépend pas des variables *educ* et *exper*. Est-ce une hypothèse raisonnable ? Cette hypothèse n'est sans doute pas tout à fait exacte, mais la vérité ne doit pas en être très éloignée non plus. En tout cas, cela vaut très certainement la peine d'inclure le QI dans l'équation de salaire si nous cherchons à mieux estimer le « rendement » du niveau d'instruction.

Il est aisé de comprendre la raison pour laquelle la satisfaction de ces deux hypothèses permet d'obtenir des estimateurs convergents après insertion de la variable de substitution dans le modèle. Si nous insérons l'équation (9.11) dans l'équation (9.10), une simple manipulation algébrique permet d'obtenir

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3.$$

Soit $e = u + b_3 v_3$, l'erreur composite de l'équation. Elle dépend de deux erreurs : celle du modèle d'intérêt (9.10), u, et celle de l'équation de la variable de substitution, v_3. Puisque u et v_3 ont une moyenne nulle et ne sont pas corrélées avec x_1, x_2, et x_3, c'est également vrai pour e. Ecrivons cette équation sous la forme

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e,$$

où $\alpha_0 = (\beta_0 + \beta_3\delta_0)$ est la nouvelle ordonnée à l'origine, et $\alpha_3 = \beta_3\delta_3$ est le paramètre de pente de la variable de substitution x_3. Comme nous l'avions déjà laissé entendre, la régression (9.12) ne donne pas d'estimateurs sans biais de β_0 et de β_3 ; par contre, nous obtiendrons des estimateurs sans biais (à tout le moins, convergents) de α_0, β_1, β_2, et α_3. Comme notre attention porte avant tout sur *educ* et *exper*, le plus important est d'obtenir de bons estimateurs des paramètres β_1 et β_2.

Dans la plupart des cas, l'estimation de α_3 est plus riche d'enseignements que celle de β_3 [en supposant qu'il eut été possible d'obtenir une estimation de β_3]. Par exemple, dans l'équation du salaire, α_3 mesure le « rendement » (ou gain salarial en pourcentage) d'un point supplémentaire de QI.

EXEMPLE 9.3
Le QI comme variable de substitution pour les aptitudes

Le fichier WAGE2 contient les données de Blackburn et Neumark (1992) sur le revenu mensuel (*wage*), le niveau d'instruction (*educ*), plusieurs variables démographiques, et le QI (*IQ*) pour 935 hommes en 1980. Pour tenir compte du biais lié à l'omission de l'aptitude (innée) des travailleurs, nous incluons la variable *IQ* dans l'équation logarithmique du salaire. Les résultats sont indiqués dans le tableau 9.2.

Tableau 9.2 *Variable dépendante : log*(wage)

Variables indépendantes	(1)	(2)	(3)
educ	0,065 (0,006)	0,054 (0,007)	0,018 (0,041)
exper	0,014 (0,003)	0,014 (0,003)	0,014 (0,003)
tenure	0,012 (0,002)	0,011 (0,002)	0,011 (0,002)
married	0,199 (0,039)	0,200 (0,039)	0,201 (0,039)
south	−0,091 (0,026)	−0,080 (0,026)	−0,080 (0,026)
urban	0,184 (0,027)	0,182 (0,027)	0,184 (0,027)
black	−0,188 (0,038)	−0,143 (0,039)	−0,147 (0,040)
IQ	–	0,0036 (0,0010)	−0,0009 (0,0052)
educ·IQ	–	–	0,00034 (0,00038)
constante	5,395 (0,113)	5,176 (0,128)	5,648 (0,546)
Observations *R* carré	935 0,253	935 0,263	935 0,263

Notre intérêt porte essentiellement sur le rendement estimé du niveau d'instruction. La colonne (1) contient les estimations du modèle qui n'inclut pas *IQ* comme variable de substitution. Le rendement estimé du niveau d'instruction est 6,5 %. Si nous pensons que les aptitudes d'un individu sont corrélées positivement avec son niveau d'instruction, alors l'estimateur du coefficient de *educ* est biaisé vers le haut, ce qui conduit à une estimation trop élevée du « rendement de l'éducation ». (Pour être plus précis, c'est la moyenne des estimations obtenues à partir de tous les échantillons aléatoires, qui sera trop élevée.) Quand la variable *IQ* est incluse dans l'équation, le rendement du niveau d'instruction tombe à 5,4 %, ce qui semble confirmer notre intuition quant à la présence d'un biais lié à l'omission des aptitudes.

L'effet de *IQ* sur les variables socio-économiques a été documenté dans un livre controversé, de Herrnstein et Murray (1994), intitulé *The Bell Curve*. La colonne (2) montre que l'effet *ceteris paribus* de *IQ* sur le revenu mensuel est positif et statistiquement significatif. Toutes autres choses étant égales (après avoir pris en compte l'influence de plusieurs autres facteurs), une augmentation de 10 points de QI augmente le revenu mensuel de 3,6 % en moyenne. Comme l'écart-type du QI dans la population des USA vaut 15 points, une augmentation de cette ampleur est associée à une augmentation du revenu mensuel de 5,4 %. Cela équivaut à l'effet sur le revenu mensuel d'une année supplémentaire d'instruction. La colonne 2 indique clairement que le niveau d'instruction joue un rôle important dans l'augmentation du revenu mensuel, même si cet effet est moins important qu'il ne l'était en l'absence de la variable *IQ*.

Certaines observations intéressantes ressortent des colonnes (1) et (2). L'inclusion de *IQ* dans l'équation a un effet mineur sur le *R* carré, qui passe de 0,253 à 0,263. Une grande partie de la variation dans log(*wage*) reste donc inexpliquée par le modèle de la colonne (2). Ensuite, l'ajout de *IQ* n'élimine pas la différence de revenu mensuel estimé entre les afro-américains et les américains de type caucasien : en moyenne, la différence de revenu mensuel entre un « noir » et un « blanc » sera statistiquement très significative et égale à 14,3 %. Notez qu'il s'agit bien d'une différence *ceteris paribus*. Ces deux individus sont identiques par ailleurs : ils ont le même QI, le même niveau d'instruction, la même expérience, etc.

Pour aller plus loin 9.2

Dans la colonne (3) du tableau 9.2, que faites-vous du coefficient de la variable *educ*, qui est de faible ampleur et n'est pas statistiquement significatif ? (*Astuce :* quand *educ.IQ* se trouve dans l'équation, quelle interprétation donnez-vous au coefficient de *educ* ?)

La colonne (3) du tableau 9.2 inclut le terme d'interaction *educ.IQ*. Ce terme permet de tenir compte de l'effet combiné que peuvent avoir les variables *educ* et *abil* sur la détermination de log(*wage*). Par exemple, nous pourrions penser que « le rendement » du niveau d'instruction est d'autant plus grand que les aptitudes sont fortes ; ce n'est pas le cas, le terme d'interaction n'étant pas significatif. L'ajout du terme d'interaction complique donc inutilement le modèle ; il rend également les variables *educ* et *IQ* individuellement non significatives. Les estimations de la colonne (2) sont donc préférables.

Dans cet exemple, nous pourrions inclure d'autres variables de substitution. Par exemple, la base de données WAGE2 contient le score que chaque individu dans l'échantillon a obtenu au test "*Knowledge of the World of Work*" (KWW). Le degré d'aptitude pourrait donc être appréhendé sur cette base ; cette nouvelle variable de substitution pourrait être utilisée avec ou sans la variable *IQ* pour améliorer l'estimation du « rendement de l'éducation » (voir l'exercice assisté par ordinateur C2).

Nous pouvons aisément comprendre la raison pour laquelle l'utilisation d'une variable de substitution ne garantit pas la disparition du biais d'omission. Imaginons qu'au lieu de (9.11), la variable non observée, x_3^*, est liée à toutes les autres variables observées, soit

$$x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3, \qquad \textbf{[9.14]}$$

où v_3 a une moyenne nulle et n'est pas corrélée avec x_1, x_2, et x_3. L'équation (9.11) implique que les paramètres δ_1 et δ_2 sont, tous les deux, nuls. En insérant l'équation (9.14) dans (9.10), on obtient

$$y = (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3, \qquad \textbf{[9.15]}$$

dont nous déduisons que $\text{plim}(\hat{\beta}_1) = \beta_1 + \beta_3\delta_1$ et $\text{plim}(\hat{\beta}_2) = \beta_2 + \beta_3\delta_2$. [En effet, l'erreur dans (9.15), $u + \beta_3 v_3$, a une moyenne nulle et n'est pas corrélé avec x_1, x_2 et x_3.] Dans l'exemple précédent où $x_1 = educ$ et $x_3^* = abil$, nous avions $\beta_3 > 0$. Par conséquent, si la corrélation partielle entre *educ* et *abil* est positive

(soit $\delta_1 > 0$), alors il y a un biais positif (et une absence de convergence) dans l'estimateur de β_1. Si *IQ* n'est pas une bonne variable de substitution de *abil* (car elle est corrélée avec les autres variables explicatives), son inclusion peut ne pas résoudre le problème de biais dans l'estimateur du « rendement de l'éducation ». Nous pouvons néanmoins penser que ce biais serait plus important si les aptitudes étaient purement et simplement ignorées.

Une critique souvent émise à propos de l'inclusion d'une variable de substitution est qu'elle introduit de la multicolinéarité et aboutit à une estimation moins précise du paramètre de la variable d'intérêt. Dans l'exemple précédent, *IQ* est ajouté à *educ*, ce qui rendrait l'estimation de β_{educ} moins précise. Cette critique néglige deux points importants. En premier lieu, l'inclusion de *IQ* réduit la variance de l'erreur, car la part des aptitudes expliquée par *IQ* est retirée de l'erreur. En règle générale, l'écart-type estimé de la régression sera donc plus petit (en notant que l'ajustement des degrés de liberté peut contrecarrer cette baisse). En second lieu et surtout, l'augmentation de la multicolinéarité est un mal nécessaire si nous désirons obtenir un estimateur de β_{educ} moins biaisé qu'auparavant. La raison pour laquelle *educ* et *IQ* sont corrélées est précisément celle qui explique que *educ* et *abil* le sont également ; *IQ* n'est qu'une variable de remplacement pour *abil*. Si nous avions pu observer *abil*, nous l'aurions inclus dans la régression et il y aurait eu, de toute façon, de la multicolinéarité causée par la corrélation entre *educ* et *abil*.

Les variables de substitution peuvent également prendre la forme d'une variable binaire. Dans l'exemple 7.9 [voir l'équation (7.15)], nous avons discuté des estimations de Krueger (1993) quant à l'impact, en terme salarial, de l'utilisation d'un ordinateur sur le lieu de travail. Dans cette étude, une variable binaire indiquait également si le travailleur disposait d'un ordinateur à la maison (sans oublier le terme d'interaction entre ces deux variables binaires). En 1993, le fait de savoir si un ordinateur était utilisé à la maison permettait d'appréhender « les aptitudes techniques » qui peuvent directement affecter le salaire. Cette variable binaire servait donc de variable de substitution.

Une variable dépendante retardée comme variable de substitution

Dans certains cas, comme dans l'exemple sur le salaire, nous avons une idée, même vague, du facteur non observé dont nous aimerions tenir compte. Le choix des variables de substitution en est facilité. Dans d'autres applications, nous pouvons deviner qu'un facteur non observé est absent de l'équation sans avoir la moindre idée du type de variable de substitution qui pourrait être utilisée. Dans de telles situations, nous pouvons inclure la valeur de la variable dépendante observée dans le passé pour tenir compte de ce facteur non observé. Cette stratégie est particulièrement utilisée dans les analyses de politiques publiques.

L'utilisation d'une **variable dépendante retardée** dans une équation requiert plus de données mais elle fournit un moyen simple de tenir compte de facteurs historiques qui expliquent le niveau *actuel* de la variable dépendante et que nous ne pourrions pas prendre en compte autrement. Par exemple, certaines villes ont enregistré des taux de criminalité élevés dans le passé. De nombreux facteurs non observés contribuent aux niveaux de la criminalité, tant dans le passé et que dans le présent. De la même manière, certaines universités attirent historiquement de meilleurs académiques que d'autres. Ces effets d'inertie sont également capturés en prenant des valeurs de *y* retardées.

Considérons une équation simple pour expliquer les taux de criminalité des villes :

$$crime = \beta_0 + \beta_1 unem + \beta_2 expend + \beta_3 crime_{-1} + u, \qquad \textbf{[9.16]}$$

où *crime* est une mesure de la criminalité par habitant, *unem* le taux de chômage de la ville, *expend* les dépenses par habitant consacrées aux forces de l'ordre, et $crime_{-1}$ le taux de criminalité de la ville mesuré lors d'une période précédente (soit l'année passée, soit il y a plusieurs années). Notre intérêt porte sur les estimateurs des coefficients de *unem* et *expend*.

Quel est l'intérêt d'inclure $crime_{-1}$ dans l'équation ? Nous pouvons très certainement anticiper que $\beta_3 > 0$, en raison de l'inertie dans l'évolution de la criminalité. L'inclusion de la variable dépendante retardée se justifie par le fait que les villes qui ont souffert d'une criminalité élevée dans le passé peuvent logiquement dépenser plus pour la contrecarrer. Par conséquent, ne pas tenir compte de ces facteurs historiques non observés, qui expliquent le taux de criminalité actuel dans une ville, reviendrait à ignorer la corrélation de ces facteurs avec les variables *expend* (et *unem*). Si nous avions utilisé une analyse en coupe instantanée pure, dans laquelle toutes les variables sont observées au même moment, nous aurions très vraisemblablement obtenu des estimateurs biaisés de l'effet causal de *unem* et *expend* sur la criminalité urbaine. Par contre, en incluant $crime_{-1}$ dans l'équation, nous pouvons au moins envisager le scénario suivant : si deux villes enregistrent actuellement le même taux de chômage et qu'elles ont souffert d'une criminalité identique dans le passé, alors l'estimateur de β_2 mesure l'effet, sur la criminalité, d'un dollar supplémentaire dépensé en faveur des forces de l'ordre.

EXEMPLE 9.4
Taux de criminalité urbaine

Nous estimons une version à élasticité constante du modèle de criminalité décrit dans l'équation (9.16). Comme la variable *unem* est donnée en pourcentage, elle est laissée sous sa forme linéaire. Les données de CRIME2 concernent 46 villes et sont observées en 1987. Le taux de criminalité est également disponible en 1982 ; nous l'utilisons comme variable indépendante afin de prendre en compte l'influence de facteurs non observables, qui affectent la criminalité et peuvent être corrélés avec les dépenses actuellement consacrées aux forces de l'ordre. Le tableau 9.3 présente les résultats de ce modèle.

Sans le taux de criminalité retardé ($crmrte_{82}$) dans l'équation, les effets du taux de chômage ($unem_{87}$) et des dépenses actuellement consacrées aux forces de l'ordre ($lawexpc_{87}$) sont contre-intuitifs ; aucun des deux paramètres n'est statistiquement significatif, bien que la statistique t sur $\log(lawexpc_{87})$ soit 1,17. Ces résultats peuvent s'expliquer par le fait qu'une augmentation des forces de l'ordre permet de mieux constater et répertorier les différentes infractions, conduisant à une augmentation de la criminalité effectivement rapportée. Mais il est également probable que des villes dont les niveaux de de criminalité récents sont élevés dépensent davantage pour assurer le respect de la loi.

Tableau 9.3 *Variable dépendante : log(crmrte$_{87}$)*

Variables indépendantes	(1)	(2)
$unem_{87}$	– 0,029 (0,032)	0,009 (0,020)
$\log(lawexpc_{87})$	0,203 (0,173)	– 0,140 (0,109)
$\log(crmrte_{82})$	–	1,194 (0,132)
constante	3,34 (1,25)	0,076 (0,821)
Observations *R* carré	46 0,057	46 0,680

L'inclusion du logarithme du taux de criminalité enregistré cinq ans plus tôt modifie sensiblement le coefficient de la variable $lawexpc_{87}$. L'élasticité du taux de criminalité par rapport aux dépenses consacrées aux forces de la loi devient égale à – 0,14, avec $t = -1,28$. Elle n'est pas très significative ; cela suggère néanmoins qu'un modèle plus sophistiqué, portant sur un plus grand nombre de villes, pourrait aboutir à des résultats significatifs.

Il n'est pas surprenant de constater que le taux de criminalité observé actuellement est fortement lié au taux de criminalité observé dans le passé. L'estimation indique que si le taux de criminalité en 1982 avait été supérieur de 1 %, alors le taux de criminalité prédit en 1987 aurait augmenté de 1,19 % environ. L'hypothèse selon laquelle l'élasticité du taux de criminalité actuel par rapport à son niveau passé est égale à l'unité, n'est pas rejetée [$t = (1{,}194 - 1) / 0{,}132 \approx 1{,}47$]. L'inclusion du taux de criminalité passé [en tant que variable explicative de substitution] augmente sensiblement le pouvoir explicatif de la régression, comme nous pouvions nous y attendre. Notez bien que la présence du taux de criminalité passé dans la régression est motivée par l'objectif d'obtenir une meilleure estimation de l'effet *ceteris paribus* de $\log(lawexpc_{87})$ sur $\log(crmrte_{87})$.

La pratique consistant à introduire une variable dépendante retardée en tant que variable explicative (dans le but d'incorporer l'influence de facteurs non observés) est loin d'être parfaite. Elle peut néanmoins nous aider à obtenir de meilleures estimations des effets de certaines politiques sur différents résultats.

Il existe d'autres manières d'exploiter des données observées sur deux années différentes afin de prendre en considération l'influence de facteurs omis. Dans les chapitres 13 et 14, nous verrons en effet que les données observées sur les mêmes coupes instantanées, à différents moments dans le temps, peuvent être exploitées dans des modèles de données en panel.

Un point de vue différent sur la régression multiple

La discussion sur les variables de substitution, que nous venons de tenir dans cette section, suggère qu'il est possible d'interpréter différemment les résultats d'une régression multiple. Jusqu'à présent, à chaque fois qu'une variable explicative n'était pas observable, nous avons spécifié un modèle de population dont l'erreur était additive, à l'image de l'équation (9.9). Les conclusions de notre discussion à ce sujet dépendaient de la qualité de la variable de substitution sélectionnée pour capturer l'effet de la variable explicative non observée (le QI servant de variable de substitution pour les aptitudes d'un individu dans ce cas précis ; nous aurions pu également choisir d'autres tests).

Une approche plus pragmatique de la régression multiple consisterait à renoncer à l'idéal d'élaboration de modèles capables de capturer les effets de facteurs non observés. La démarche consiste tout simplement à reconnaître que seul un ensemble de variables explicatives observables est disponible : la variable d'intérêt premier, comme les années d'instruction, et d'autres variables de contrôles, comme les résultats de différents tests de compétence observables. L'espérance de y n'est conditionnelle qu'à cet ensemble de variables explicatives observées. Par exemple, dans l'exemple sur le salaire où *lwage* correspond à $\log(wage)$, nous pouvons estimer $E(lwage \mid educ, exper, tenure, south, urban, black, IQ)$, ce qui correspond exactement aux résultats du tableau 9.2. Nos objectifs sont à présent plus modestes. Plutôt que d'introduire le concept nébuleux d'aptitude dans l'équation (9.9), nous fixons dès le début un objectif plus réaliste, celui d'estimer l'effet *ceteris paribus* du niveau d'instruction, en gardant la variable *IQ* (et les autres facteurs observés) inchangés. Il n'est plus nécessaire de déterminer si *IQ* est effectivement une bonne variable de substitution pour les aptitudes individuelles. En renonçant à répondre à la question plus ambitieuse de l'équation (9.9), portant sur l'effet des aptitudes sur le salaire, nous cherchons à répondre à une autre question plus pragmatique : si deux personnes ont les mêmes niveaux de *IQ* (et les mêmes valeurs pour l'expérience, etc.), quelle est la différence attendue en pourcentage entre leurs salaires si l'une bénéficie d'une année d'instruction en plus que l'autre ?

Prenons un autre modèle dans lequel la variable dépendante correspond aux résultats scolaires obtenus par les élèves d'un établissement à un test standardisé. Si nous incluons le taux de pauvreté comme variable explicative dans la régression, nous devons admettre que ce taux ne capture que sommairement les multiples différences économiques et sociodémographiques qui peuvent exister entre les parents et leurs enfants au sein de chaque établissement. D'un autre côté, il s'agit souvent de la seule information disponible et il est préférable de prendre en considération l'influence de la pauvreté que faire l'inverse. Même si nous ne disposons pas de

variable de substitution valable pour « l'aptitude des étudiants » ou « le sentiment d'implication des parents », nous parvenons à mieux estimer l'effet *ceteris paribus* des dépenses réalisées par l'établissement scolaire sur la performance des élèves, en incluant le taux de pauvreté dans la régression plutôt qu'en l'excluant.

Dans certains cas, la régression n'est utilisée que pour générer la meilleure prédiction possible de y, étant donné un ensemble de variables explicatives $(x_1, \ldots, x_k)$. Dans un tel scénario, se préoccuper de la présence d'un biais dans certains estimateurs provenant de l'omission de facteurs non observés n'a pas beaucoup de sens. Le but de l'exercice est d'identifier le meilleur modèle de prédiction en s'assurant que tous les régresseurs peuvent être observés lors du calcul de la prédiction. Par exemple, le responsable du département des admissions d'une grande université peut chercher à prédire la réussite des candidats à l'entrée, sur base de variables qui seront disponibles au moment de la soumission des dossiers de candidature. Ces variables peuvent inclure les résultats obtenus au lycée (pour des cours spécifiques ou sous la forme d'une moyenne globale), les résultats obtenus à des tests standardisés, la participation à des activités variées (clubs de mathématiques, etc.), et même des informations liées au cadre familial. Par contre, une variable mesurant le taux de présence aux cours délivrés à l'université ne peut pas être incluse, cette information n'étant pas disponible au moment de l'inscription. Le responsable du département n'a que faire du biais éventuel qu'introduit l'omission de cette variable dans son modèle : il n'est pas intéressé par l'estimation de l'effet *ceteris paribus* de la performance au lycée du candidat sur sa probabilité de réussite à l'université (étant donné un taux de présence aux cours délivrés à l'université). De même, il n'a pas à s'inquiéter du biais dont pourraient souffrir les coefficients de son modèle et qui proviendrait de l'impossibilité de mesurer certains facteurs, tels que la motivation du candidat. Evidemment, disposer d'une telle information améliorerait substantiellement la prévision, mais la présence d'un biais lié à son absence ou l'estimation de l'effet *ceteris paribus* d'une variable spécifique n'a pas d'importance. L'objectif consiste simplement à élaborer le meilleur modèle en utilisant le plus grand nombre de variables explicatives *observables* [et pertinentes].

9.3 MODÈLES À PENTES ALÉATOIRES

Dans notre analyse de la régression multiple jusqu'à présent, nous avons considéré que les coefficients de pente étaient identiques pour tous les individus de la population ; seules des caractéristiques mesurables pouvaient impliquer une variation de la pente, auquel cas des termes d'interaction étaient introduits dans la régression multiple. Par exemple, dans la section 7.4, nous avons introduit un terme d'interaction entre le niveau d'instruction et le genre des individus pour constater que « le rendement de l'éducation » pouvait effectivement varier entre les hommes et les femmes.

La question à laquelle nous aimerions répondre dans cette section est liée, mais différente malgré tout : qu'advient-il de l'effet *ceteris paribus* d'une variable si cet effet dépend de facteurs non observés propres à *chaque* unité de la population ? En ne considérant qu'une seule variable explicative, soit x, nous pouvons écrire le modèle général suivant, qui repose sur un tirage aléatoire, i, de la population :

$$y_i = a_i + b_i x_i, \qquad \textbf{[9.17]}$$

où a_i est l'ordonnée à l'origine pour l'unité i et b_i, la pente. Dans le modèle de régression simple du chapitre 2, nous avons supposé que $b_i = \beta$ et que $a_i = u_i$. Le modèle (9.17) est parfois appelé **modèle à coefficients aléatoires** ou **à pente aléatoire**, car le coefficient de pente, b_i, est considéré comme provenant d'un tirage aléatoire au sein d'une population, comme le sont les données observées, (x_i, y_i), et l'ordonnée à l'origine non observée, a_i. Par exemple, si $y_i = \log(wage_i)$ et $x_i = educ_i$, alors (9.17) permet au rendement du niveau d'instruction, b_i, de varier en fonction de chaque personne. Si, disons, b_i dépend du niveau d'aptitude non mesurée (comme cela pourrait être le cas de a_i), alors l'effet *ceteris paribus* d'une année d'études supplémentaire peut être biaisé.

Sur base d'un échantillon aléatoire de taille n, nous pouvons tirer (implicitement) n valeurs de b_i et n valeurs de a_i (ainsi que les données observées de x et y). Naturellement, il nous est impossible

d'estimer une pente – ou une ordonnée à l'origine – pour chaque i. Nous pouvons néanmoins estimer une pente moyenne (et une ordonnée à l'origine moyenne), où la moyenne se définit sur base de toute la population. Si nous définissons $\alpha = E(a_i)$ et $\beta = E(b_i)$, alors β est la moyenne de l'effet *ceteris paribus* de x sur y. Le paramètre β mesure l'**effet partiel moyen (EPM)**, ou **l'effet marginal moyen (EMM)**. [On parle également d'un effet *ceteris paribus* moyen.] Dans l'équation du log des salaires, β mesure l'EPM du niveau d'instruction sur le salaire en pourcentage. Autrement dit, β mesure le « rendement moyen de l'éducation » dans la population (pour une année d'études supplémentaire).

Si nous écrivons $a_i = \alpha + c_i$ et $b_i = \beta + d_i$, alors d_i correspond à l'écart par rapport à l'EPM propre à chaque individu i. Par construction, $E(c_i) = 0$ et $E(d_i) = 0$. En substituant dans (9.17),

$$y_i = \alpha + \beta x_1 + c_i + d_i x_i \equiv \alpha + \beta x_i + u_i, \qquad \textbf{[9.18]}$$

où $u_i = c_i + d_i x_i$. (Pour alléger la notation, α et β représentent les valeurs moyennes de a_i et b_i, respectivement pour l'ordonnée à l'origine et la pente.) En d'autres mots, le modèle à coefficients aléatoires peut afficher des coefficients constants à condition que les erreurs contiennent un terme d'interaction entre d_i, l'écart non observable par rapport à l'EPM, et x_i, la variable explicative observée.

Sous quelles hypothèses pouvons-nous obtenir une régression simple de y_i sur x_i dont les estimateurs de α et β sont sans biais ? Nous pouvons exploiter le résultat du chapitre 2. Si $E(u_i|x_i) = 0$, alors les estimateurs des MCO sont en général sans biais. Si $u_i = c_i + d_i x_i$, alors les deux conditions qui suffisent à garantir l'absence de biais sont $E(c_i|x_i) = E(c_i) = 0$ et $E(d_i|x_i) = E(d_i) = 0$. En les exprimant sur base de l'ordonnée à l'origine et de la pente spécifiques à l'unité i, cela donne :

$$E(a_i|x_i) = E(a_i) \text{ et } E(b_i|x_i) = E(b_i). \qquad \textbf{[9.19]}$$

Dans ce cas, les espérances de a_i et b_i sont indépendantes de la variable x_i. Si nous cherchons à faire varier la pente en fonction de facteurs propres aux unités de la population, nous obtiendrons un estimateur des MCO convergent de la moyenne de cette pente (au sein de la population) à la condition que cette moyenne soit indépendante de la variable explicative. (Voir le problème 6 pour envisager un ensemble plus faible de conditions qui garantissent la convergence des estimateurs des MCO.)

Le terme d'erreur dans (9.18) contient très certainement de l'hétéroscédasticité. En fait, si $\text{Var}(c_i|x_i) = \sigma_c^2$, $\text{Var}(d_i|x_i) = \sigma_d^2$ et $\text{Cov}(c_i, d_i|x_i) = 0$, alors

$$\text{Var}(u_i|x_i) = \sigma_c^2 + \sigma_d^2 x_i^2. \qquad \textbf{[9.20]}$$

Le terme d'erreur u_i souffre d'hétéroscédasticité, sauf si $\sigma_d^2 = 0$ (exigeant que $b_i = \beta + d_i = \beta$ pour tout i). Nous avons appris à corriger les écarts-types estimés dans de telles circonstances. Nous pouvons utiliser les MCO et rendre les écarts-types estimés robustes à l'hétéroscédasticité ; nous pouvons également estimer la fonction de variance (9.20) et appliquer les moindres carrés pondérés (MCP). Comme cette dernière stratégie vise à rendre homoscédastiques l'ordonnée à l'origine et la pente, toutes deux aléatoires, il convient de s'assurer que les MCP sont effectivement parfaitement robustes aux violations de (9.20).

L'équation (9.20) conforte certains auteurs dans l'idée que la présence d'hétéroscédasticité dans les erreurs provient de l'existence de coefficients aléatoires de pente. Notons néanmoins que la forme de l'équation (9.20) est très particulière : elle n'autorise aucune hétéroscédasticité dans a_i ou b_i. On ne peut donc pas faire la différence entre un modèle à pente aléatoire, dans lequel l'ordonnée à l'origine et la pente sont indépendantes de x_i, et un modèle à pente constante, qui souffrirait d'une hétéroscédasticité dans a_i.

Le développement du modèle à coefficients aléatoires est similaire dans le cadre de la régression multiple. De manière générale, écrivons

$$y_i = a_i + \beta_{i1} x_{i1} + \beta_{i2} x_{i2} + \ldots + \beta_{ik} x_{ik}. \qquad \textbf{[9.21]}$$

Ensuite, comme $a_i = \alpha + c_i$ et $b_{ij} = \beta_j + d_{ij}$, nous obtenons

$$y_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i, \qquad [9.22]$$

où $u_i = c_i + d_{i1}x_{i1} + \ldots + d_{ik}x_{ik}$. Si nous conservons les hypothèses d'indépendance des espérances, $E(a_i|\mathbf{x}_i) = E(a_i)$ et $E(b_{ij}|\mathbf{x}_i) = E(b_{ij})$, $j = 1, \ldots, k$, alors $E(y_i|x_i) = \alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$. Lorsque les échantillons sont tirés aléatoirement, les MCO produisent des estimateurs sans biais de α et des β_j. Comme dans la régression simple, $\text{Var}(u_i|\mathbf{x}_i)$ est très certainement hétéroscédastique.

Les coefficients b_{ij} peuvent dépendre de variables explicatives observables ou non observables. Par exemple, supposons, avec $k = 2$, que l'effet de x_{i2} dépende de x_{i1} ; écrivons $b_{i2} = \beta_2 + \delta_1(x_{i1} - \mu_1) + d_{i2}$, où $\mu_1 = E(x_{i1})$. Si $E(d_{i2}|\mathbf{x}_i) = 0$ (et similairement pour c_i et d_{i1}), alors $E(y_i|x_{i1}, x_{i2}) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \delta_1(x_{i1} - \mu_1)x_{i2}$, qui inclut un terme d'interaction entre x_{i1} et x_{i2}. En soulignant que nous avons soustrait la moyenne μ_1 de l'autre variable explicative x_{i1}, β_2 mesure l'effet partiel moyen de x_{i2}.

L'idée centrale de cette section est que l'introduction de pentes aléatoires dans une régression est très simple lorsque ces pentes, ou leur moyenne à tout le moins, sont indépendantes des variables explicatives. Par ailleurs, il n'est pas compliqué de modéliser les pentes comme des fonctions de variables exogènes, en recourant à l'utilisation de termes d'interaction ou de termes au carré. Bien sûr, dans le chapitre 6, nous avons vu que ce type de modèles pouvait être utile sans qu'il soit nécessaire d'introduire la notion de pente aléatoire. La spécification de pentes aléatoires est une autre manière de justifier l'utilisation de ce type de modèles. L'estimation devient considérablement plus difficile lorsque l'ordonnée à l'origine et les pentes aléatoires sont corrélées avec certains régresseurs. Nous abordons le problème lié à la présence de variables explicatives endogènes dans le chapitre 15.

9.4 PROPRIÉTÉS DES ESTIMATEURS DES MCO EN PRÉSENCE D'ERREURS DE MESURE

Dans certaines applications économiques, il est impossible d'obtenir des informations sur une variable qui affecte sensiblement le comportement économique. Pensons, par exemple, au taux marginal d'imposition sur le revenu annuel, qu'une famille aimerait connaître avant d'envoyer des dons à différentes œuvres caritatives. Ce taux marginal peut être difficile à obtenir pour chaque famille et à résumer en un seul chiffre étant donné les différents niveaux de revenus [au sein de la famille]. Nous pourrions considérer à la place le taux moyen d'imposition, basé sur le revenu total et les taxes versées.

Lorsqu'une variable économique est mesurée de manière imprécise dans une régression, le modèle contient des erreurs de mesure. Dans cette section, nous évaluons les conséquences que la présence de telles erreurs peut avoir sur les estimateurs des MCO. Sous certaines conditions, les estimateurs des MCO resteront (sans biais et) convergents. Lorsque ce n'est pas le cas, il est parfois possible de calculer la taille du biais asymptotique.

Comme nous le verrons, le problème d'erreur de mesure s'insère dans une structure statistique similaire à celle du problème de « variable omise – variable de substitution », que nous avons étudié dans la section précédente ; ils restent néanmoins différents sur le plan conceptuel. Dans le cas de la variable de substitution, nous cherchons une variable qui est, d'une certaine manière, associée à un facteur non observé. Dans le cas d'une erreur de mesure, la variable non observée est bien définie sur le plan quantitatif (comme le taux marginal d'imposition ou le revenu annuel réel) mais les mesures disponibles de cette variable sont imprécises et contiennent des erreurs. Par exemple, le revenu annuel déclaré sert à mesurer le revenu annuel réel, alors que le QI est une variable de substitution pour l'aptitude.

Une autre différence importante entre les problématiques de substitution et d'erreur de mesure est que, dans le dernier cas, l'intérêt porte souvent sur la variable indépendante, qui est mal mesurée. Dans

le cas d'une variable de substitution, il est rare que l'estimation de l'effet *ceteris paribus* de la variable omise constitue l'élément central. En règle générale, elle joue le rôle de variable de contrôle et ce sont plutôt les effets *ceteris paribus* des autres variables indépendantes, qui importent.

Avant d'approfondir cette discussion, notez bien que cette question des erreurs de mesure ne se pose que dans les cas où les variables à la disposition de l'économètre ne correspondent pas aux variables qui influencent réellement les décisions des individus, familles, firmes, etc.

Erreur de mesure dans la variable dépendante

Nous commençons par le cas simple où l'erreur de mesure n'affecte que la variable dépendante. Soit y^*, la variable (dans la population, comme toujours) que nous aimerions expliquer. Par exemple, y^* pourrait correspondre à l'épargne annuelle d'une famille, effectivement réalisée. Le modèle de régression multiple prend la forme habituelle

$$y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \qquad \textbf{[9.23]}$$

en supposant qu'il respecte les hypothèses de Gauss-Markov. Soit y, la mesure observable de y^*. Dans le cas de l'épargne, y correspond à l'épargne déclarée. Malheureusement, les familles n'évaluent pas leur épargne de manière parfaite ; il est facile d'oublier une catégorie de dépense ou de surestimer une contribution à l'épargne. En règle générale, nous pouvons anticiper une différence entre y et y^*, au moins pour un sous-ensemble de familles dans la population.

L'**erreur de mesure** (dans la population) est définie comme la différence entre les valeurs observée et réelle :

$$e_0 = y - y^*. \qquad \textbf{[9.24]}$$

Pour un tirage aléatoire i dans la population, nous pouvons écrire $e_{i0} = y_i - y_i^*$. La grande interrogation porte sur la manière avec laquelle l'erreur de mesure est liée aux autres facteurs. Pour pouvoir estimer ce modèle, nous devons insérer $y^* = y - e_0$ dans l'équation (9.23) [puisque y^* ne peut pas être observée]. En réarrangeant, nous obtenons

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u + e_0. \qquad \textbf{[9.25]}$$

Le terme d'erreur dans l'équation (9.25) est $u + e_0$. Étant donné que $y, x_1, x_2, \ldots, x_k$ sont des variables observées, ce modèle peut être estimé par les MCO, comme d'habitude. Jusqu'à présent, nous avons ignoré le fait que y est une mesure imparfaite de y^*.

Lorsque y^* est replacé par y, la méthode des MCO parvient-elle encore à générer des estimateurs convergents des β_j ? Étant donné que le modèle original (9.23) satisfait les hypothèses de Gauss-Markov, le terme d'erreur u a une moyenne nulle et n'est pas corrélé avec les x_j. En règle générale, il est naturel de supposer que la moyenne de l'erreur de mesure est nulle. Si cela devait ne pas être le cas, la seule conséquence serait l'introduction d'un biais dans l'estimateur de l'ordonnée à l'origine, β_0 ; en présence d'une erreur de mesure, ce paramètre n'est pas au centre de nos préoccupations de toute manière. Par contre, la nature de la relation entre l'erreur de mesure e_0 et les variables explicatives x_j revêt une très grande importance. L'hypothèse la plus courante est de considérer que cette erreur de mesure dans y est statistiquement indépendante de chacune des variables explicatives présentes dans le modèle. Si cette hypothèse est vraie, alors les estimateurs des MCO de (9.25) sont sans biais et convergents. Les procédures habituelles d'inférence statistique (basées sur les t, F, et multiplicateur de Lagrange) sont également valides.

Si e_0 et u ne sont pas corrélées par ailleurs, comme il est coutume de le supposer, alors $\text{Var}(u + e_0) = \sigma_u^2 + \sigma_0^2 > \sigma_u^2$. L'erreur de mesure dans la variable dépendante conduit à une plus grande variance du terme d'erreur et, donc, à une plus grande variance des estimateurs des MCO. Nous pouvions

logiquement nous y attendre : la seule manière d'atténuer ce problème est d'obtenir des données plus précises. L'essentiel est de constater que les estimateurs des MCO conservent leurs propriétés désirables, à condition que l'erreur de mesure ne soit pas corrélée avec les variables indépendantes.

EXEMPLE 9.5
Fonction d'épargne et erreur de mesure

Considérons la fonction d'épargne

$$sav^* = \beta_0 + \beta_1 inc + \beta_2 size + \beta_3 educ + \beta_4 age + u,$$

où l'épargne réelle (sav^*) peut ne pas correspondre à l'épargne déclarée (sav). La question est de savoir si l'erreur de mesure dans sav est systématiquement liée aux variables explicatives. À première vue, il semble raisonnable de supposer que l'erreur de mesure n'est pas corrélée avec *inc*, *size*, *educ*, et *age*. D'un autre côté, nous pourrions penser que les familles dont le niveau d'instruction ou les revenus sont les plus élevés, parviennent à mieux estimer leur épargne. Au bout du compte, il est impossible de le savoir sans information disponible sur sav^*. Pour chaque observation, l'erreur de mesure est : $e_{i0} = sav_i - sav_i^*$.

Lorsque la variable dépendante prend la forme logarithmique, la variable dépendante correspond à $\log(y^*)$; l'équation de l'erreur de mesure prend alors la forme

$$\log(y) = \log(y^*) + e_0. \qquad \textbf{[9.26]}$$

Ceci correspond à une **erreur de mesure multiplicative** pour y, soit $y = y^* a_0$, où $a_0 > 0$ et $e_0 = \log(a_0)$.

EXEMPLE 9.6
Erreur de mesure dans les taux de rebus

Dans la section 7.6, nous avons cherché à savoir si l'octroi de subventions destinées à améliorer la formation professionnelle permettait de réduire le taux de rebus dans les entreprises manufacturières. Nous pouvons aisément considérer que les taux de rebus que les firmes déclarent sont mesurés avec une certaine marge d'erreur. (En réalité, la plupart des firmes reprises dans l'échantillon ne déclarent aucun rebus.) Dans le cadre d'une régression simple, nous avons

$$\log(scrap^*) = \beta_0 + \beta_1 grant + u,$$

où $scrap^*$ est le vrai taux de rebus et *grant* est une variable binaire égale à 1 lorsque l'entreprise a bénéficié d'une subvention. L'erreur de mesure est modélisée par

$$\log(scrap) = \log(scrap^*) + e_0.$$

Peut-on considérer que l'erreur de mesure, e_0, est indépendante du fait qu'une firme reçoive une subvention ? Une personne cynique (ou réaliste) pourrait penser que le taux de rebus déclaré aura tendance à être plus faible pour les entreprises qui bénéficient d'une subvention dans le but de justifier l'octroi de la subvention. Si tel est le cas, le modèle que nous devons estimer est

$$\log(scrap) = \beta_0 + \beta_1 grant + u + e_0,$$

et l'erreur $u + e_0$ est négativement corrélée avec *grant*. Le coefficient β_1 affiche alors un biais négatif vers le bas, qui rend le programme de formation plus efficace qu'il ne l'est en réalité. (Notez bien qu'un coefficient β_1 plus négatif indique que le programme de formation est plus efficace, puisqu'une productivité par travailleur accrue est associée à un taux de rebus plus faible.)

Le point essentiel à retenir est qu'une erreur de mesure dans la variable dépendante *peut* introduire un biais dans les estimateurs des MCO *si* cette erreur de mesure est liée à une ou plusieurs variables explicatives. En réalité, l'hypothèse la plus fréquente est de considérer que cette erreur de mesure est introduite aléatoirement et indépendamment des variables explicatives ; sous cette hypothèse, le recours aux estimateurs des MCO ne pose aucun problème particulier.

Erreur de mesure dans la variable explicative

En règle générale, l'introduction d'une erreur de mesure dans la variable explicative constitue un problème beaucoup moins épineux que celui provenant d'une erreur de mesure dans la variable dépendante. Nous allons maintenant le démontrer en débutant par le cas de la régression simple, soit

$$y = \beta_0 + \beta_1 x_1^* + u, \qquad \textbf{[9.27]}$$

en considérant que ce modèle respecte au moins les quatre premières hypothèses de Gauss-Markov. Cela implique que l'estimation de (9.27) par les MCO conduira à des estimateurs sans biais et convergents de β_0 et β_1. Le problème est que x_1^* ne peut pas être observé. À la place, nous ne disposons que d'une mesure imparfaite de x_1^* ; appelons-la x_1. Par exemple, x_1^* pourrait être le revenu réel et x_1 le revenu déclaré.

L'erreur de mesure dans la population est tout simplement égale à

$$e_1 = x_1 - x_1^*, \qquad \textbf{[9.28]}$$

qui peut être positive, négative ou nulle. On suppose que l'erreur de mesure dans la population est *en moyenne* nulle, soit $E(e_1) = 0$. Il s'agit d'une hypothèse naturelle qui ne modifie en rien les conclusions que nous allons tirer. Une autre hypothèse à laquelle nous allons également recourir est l'absence de corrélation entre u et les *deux* variables x_1^* et x_1. Sur base des espérances conditionnelles, nous pouvons écrire $E(y|x_1^*, x_1) = E(y|x_1^*)$. Cette égalité montre que x_1 n'affecte pas y lorsque l'influence de x_1^* est prise en considération. Cette hypothèse a également été utilisée dans le cas des variables de substitution. Elle se vérifie par définition (ou presque) ; elle ne fait donc pas l'objet de controverse.

Notre objectif est d'identifier les propriétés des estimateurs des MCO dans le cadre d'une régression simple de y sur x_1 sachant que x_1 remplace x_1^*. Ces propriétés reposent prioritairement sur des hypothèses qui touchent à l'erreur de mesure. Dans la littérature économétrique, nous rencontrons souvent deux hypothèses « opposées » à son sujet. La première hypothèse implique que e_1 n'est pas corrélée avec la mesure *observée*, x_1, soit

$$Cov(x_1, e_1) = 0. \qquad \textbf{[9.29]}$$

Si l'hypothèse (9.29) est vraie, étant donné la relation (9.28), e_1 est obligatoirement corrélée avec la variable non observée x_1^* [Sinon, (9.28) n'a aucun sens]. Pour déterminer les propriétés des estimateurs des MCO dans ce cas, nous pouvons écrire $x_1^* = x_1 - e_1$ et insérer cette égalité dans l'équation (9.27), soit

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1). \qquad \textbf{[9.30]}$$

Étant donné les espérances nulles de u et e_1 et l'absence de corrélation entre ces deux termes d'erreur, le terme $u - \beta_1 e_1$ a également une moyenne nulle et n'est pas corrélé avec x_1. Il s'ensuit que les MCO produisent un estimateur convergent de β_1 (et de β_0), même si x_1 est utilisée à la place de x_1^*. Puisque u n'est pas corrélée avec e_1, la variance de l'erreur dans (9.30) vaut $Var(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$. Donc, hormis le cas où $\beta_1 = 0$, l'introduction d'une erreur de mesure dans la variable explicative augmente la variance de l'erreur, mais cela n'affecte pas les propriétés des estimateurs des MCO (même si les variances des $\hat{\beta}_j$ seront plus grandes que si nous pouvions observer x_1^* directement).

L'hypothèse d'absence de corrélation entre e_1 et x_1 est analogue à l'hypothèse à laquelle nous avons recouru dans la section 9.2 au sujet des variables de substitution. Comme nous venons de le voir, cette première

hypothèse implique que les estimateurs des MCO conservent toutes leurs propriétés désirables. En réalité, la plupart des économètres n'ont pas cette hypothèse en tête lorsqu'ils reconnaissent la présence d'une erreur de mesure dans une variable explicative. Il s'agit plutôt de la seconde hypothèse, celle d'**erreur classique dans les variables** (**ECV**). Elle repose sur l'absence de corrélation entre l'erreur de mesure et la variable explicative *non observée*, soit

$$\text{Cov}(x_1^*, e_1) = 0. \qquad \textbf{[9.31]}$$

Comme nous pouvons écrire que

$$x_1 = x_1^* + e_1,$$

cette hypothèse implique que les deux composantes de x_1 ne sont pas corrélées. (Ceci ne modifie en rien les hypothèses que nous avons posées à l'égard de u, dont la corrélation avec x_1^* *et* x_1 est nulle ; les termes u et e_1 ne sont donc pas corrélés non plus.)

Si l'hypothèse (9.31) est valide, alors x_1 et e_1 *doivent* être corrélées :

$$\text{Cov}(x_1, e_1) = \text{E}(x_1 e_1) = \text{E}(x_1^* e_1) + \text{E}(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2. \qquad \textbf{[9.32]}$$

Par conséquent, la covariance entre x_1 et e_1 est égale à la variance de l'erreur de mesure sous l'hypothèse d'ECV.

En repartant de l'équation (9.30), nous pouvons constater que la corrélation entre x_1 et e_1 va poser problème. Puisque l'erreur u et la variable x_1 ne sont pas corrélées, la covariance entre x_1 et l'erreur composite $u - \beta_1 e_1$ est

$$\text{Cov}(x_1, u - \beta_1 e_1) = -\beta_1 \text{Cov}(x_1, e_1) = -\beta_1 \sigma_{e_1}^2.$$

Par conséquent, dans le cas d'une ECV, l'estimation de la régression de y sur x_1 par les MCO donne un estimateur biaisé et non convergent.

En utilisant les résultats asymptotiques du chapitre 5, nous pouvons déterminer la taille du biais asymptotique résultant de cette absence de convergence. La limite en probabilité de $\hat{\beta}_1$ est égale au coefficient β_1 augmenté du ratio de la covariance entre x_1 et $u - \beta_1 e_1$ sur la variance de x_1 :

$$\begin{aligned} \text{plim}(\hat{\beta}_1) &= \beta_1 + \frac{\text{Cov}(x_1, u - \beta_1 e_1)}{\text{Var}(x_1)} \\ &= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \\ &= \beta_1 \left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right) \\ &= \beta_1 \left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right), \end{aligned} \qquad \textbf{[9.33]}$$

où nous avons utilisé le fait que $\text{Var}(x_1) = \text{Var}(x_1^*) + \text{Var}(e_1)$.

L'équation (9.33) est très intéressante. Le facteur multipliant β_1, qui correspond au rapport $\text{Var}(x_1^*) / \text{Var}(x_1)$, est toujours inférieur à 1. [Il s'agit d'une implication de l'hypothèse d'ECV (9.31)]. Donc, $\text{plim}(\hat{\beta}_1)$ est toujours plus proche de zéro que β_1. Sur le plan asymptotique, il s'agit d'un **biais d'atténuation** dans les estimateurs des MCO dû à l'erreur classique dans les variables. En moyenne (ou en grands échantillons), l'effet de x_1 sur y, estimé par les MCO, sera *atténué* en présence d'une ECV. Par exemple, si β_1 est positif, alors $\hat{\beta}_1$ sera en moyenne plus petit que β_1. Il s'agit d'une conclusion importante, valable dans le contexte d'une ECV.

Notez que, si la variance de x_1^* est élevée par rapport à la variance de l'erreur de mesure e_1, ce biais asymptotique des estimateurs des MCO sera faible, étant donné que $\text{Var}(x_1^*) / \text{Var}(x_1)$ est proche de l'unité lorsque $\sigma_{x_1}^2 / \sigma_{e_1}^2$ est élevé. Par conséquent, une ECV introduit un biais plus ou moins grand en fonction du rapport entre la variation de x_1^* et celle de e_1.

Les choses se compliquent dans le cas de la régression multiple. En guise d'illustration, considérons le modèle

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + \beta_3 x_3 + u, \qquad \textbf{[9.34]}$$

où la première des trois variables explicatives est mesurée avec imprécision. Nous posons l'hypothèse habituelle d'absence de corrélation entre u et les variables x_1^*, x_2, x_3, et x_1. Comme auparavant, l'hypothèse cruciale concerne l'erreur de mesure e_1. Dans presque tous les cas, e_1 est supposée ne pas être corrélée avec les variables explicatives mesurées sans erreur, soit x_2 et x_3. Le point clé est de déterminer si l'erreur e_1 est corrélée avec x_1. Dans le cas où elle ne l'est pas, les MCO de la régression de y sur x_1, x_2, et x_3 produisent des estimateurs convergents. Nous pouvons le vérifier aisément en écrivant la régression sous la forme

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u - \beta_1 e_1, \qquad \textbf{[9.35]}$$

dans laquelle aucune des erreurs, u et e_1, n'est corrélée avec les variables explicatives.

Par contre, si e_1 est corrélée avec x_1 dans l'équation (9.35) sous l'hypothèse d'ECV, les estimateurs des MCO seront biaisés et non convergents. Notez bien que cela signifie, en règle générale, que tous les estimateurs des MCO seront biaisés, pas uniquement $\hat{\beta}_1$. Dans ces conditions, comment peut-on caractériser le biais de l'estimateur $\hat{\beta}_1$ dans (9.33) ? À nouveau, il s'avère qu'un biais asymptotique d'atténuation existe. Nous pouvons montrer que

$$\text{plim}(\hat{\beta}_1) = \beta_1 \left(\frac{\sigma_{r_1^*}^2}{\sigma_{r_1^*}^2 + \sigma_{e_1}^2} \right), \qquad \textbf{[9.36]}$$

où r_1^* est l'erreur de la population dans la régression $x_1^* = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + r_1^*$. La formule (9.36) s'applique également dans le cas général où il y a k variables, à condition que x_1 soit la seule variable mal mesurée naturellement.

Les conclusions sont moins claires concernant les estimateurs des coefficients β_j relatifs aux variables mesurées sans erreur. Dans le cas particulier où x_1^* n'est pas corrélée avec x_2 et x_3, $\hat{\beta}_2$ et $\hat{\beta}_3$ sont des estimateurs convergents. Cette condition est rarement remplie dans la pratique. En règle générale, la présence d'une erreur de mesure dans une variable explicative empêche les estimateurs $\hat{\beta}_j$ de converger. Il est en outre difficile de déterminer la taille, et même la direction, de ces biais.

EXEMPLE 9.7
Erreur de mesure et résultats des étudiants à l'université

Considérons l'effet du revenu familial sur la moyenne des résultats obtenus à l'université. Dans la régression, nous allons utiliser deux variables de contrôle : *hsGPA* (moyenne des points obtenus au lycée) et *SAT* (résultat à un test d'aptitude utilisé pour accéder aux études supérieures). Bien que le revenu familial est sans doute important pour expliquer la performance des étudiants à l'université, il est possible que son effet ne soit qu'indirect [puisque le revenu familial déclaré par l'étudiant ne correspond pas nécessairement au revenu familial réel]. Pour tester cette hypothèse, nous utilisons le modèle

$$colGPA = \beta_0 + \beta_1 faminc^* + \beta_2 hsGPA + \beta_3 SAT + u,$$

où *faminc** est le revenu familial annuel réel. (Cette variable pourrait apparaître sous forme logarithmique ; pour des raisons de simplicité, nous la laissons sous la forme de niveau, linéaire). Des données précises sur *colGPA*, *hsGPA* et *SAT* sont relativement faciles à obtenir. Par contre, le revenu familial, surtout lorsqu'il est déclaré par les étudiants, peut facilement être affecté par une erreur de mesure. Si $faminc = faminc^* + e_1$ et que les hypothèses de l'ECV sont valides, alors l'utilisation du revenu familial déclaré (*faminc*), au lieu du revenu familial réel (*faminc**) conduit à biaiser l'estimateur de β_1 vers zéro. Par conséquent, la probabilité de rejeter $H_0 : \beta_1 = 0$ sera, en moyenne, moins grande qu'elle ne devrait l'être. Nous aurons moins de chance de détecter que $\beta_1 > 0$.

Il arrive naturellement qu'une erreur de mesure soit présente dans plusieurs variables explicatives et parfois dans la variable dépendante également. Nous avons déjà précisé que l'erreur de mesure dans la variable dépendante est souvent considérée comme n'étant pas corrélée avec les variables explicatives, qu'elles soient observées ou non. Il est également compliqué de calculer le biais des estimateurs des MCO sous des extensions des hypothèses de l'ECV ; cela ne mène d'ailleurs à aucun résultat concluant.

Dans certains cas, il apparaît clairement que l'hypothèse d'ECV, telle définie en (9.31) n'est pas appropriée. Considérons une variante de l'exemple 9.7 :

$$colGPA = \beta_0 + \beta_1 fsmoked^* + \beta_2 hsGPA + \beta_3 SAT + u,$$

où *smoked** est le nombre réel de cigarettes de marijuana fumées par un étudiant au cours des 30 derniers jours. La variable *smoked* contient les réponses que les étudiants ont apportées à la question suivante : combien de cigarettes de marijuana avez-vous fumées au cours des 30 derniers jours ? Soit le modèle d'erreur de mesure standard :

$$smoked = smoked^* + e_1.$$

Il est très probable que l'hypothèse d'ECV ne tienne pas dans ce modèle. En effet, les étudiants qui ne fument pas du tout de marijuana, de telle sorte que $smoked^* = 0$, vont très probablement répondre $smoked = 0$. L'erreur de mesure sera donc nulle pour ces étudiants. Par contre, quand $smoked^* > 0$, il est probable que l'étudiant commette des erreurs en comptabilisant le nombre de cigarettes consommées au cours des 30 derniers jours, même si l'étudiant ne craint pas de dire la vérité. Cela signifie que l'erreur de mesure e_1 et la variable non observée *smoked** seront corrélées, ce qui viole l'hypothèse d'ECV décrite en (9.31). Malheureusement, les implications d'une erreur de mesure, qui ne correspond ni à (9.29) ni à (9.31), sont difficiles à identifier ; leur analyse dépasse de toute manière la portée de cet ouvrage.

Avant de conclure cette section, notez que l'hypothèse d'ECV (9.31) demeure une hypothèse forte, tout en étant moins stricte que l'hypothèse (9.29). La vérité est probablement dans un « mélange des deux » et, si e_1 est corrélée à la fois avec x_1 et x_1^*, les estimateurs des MCO ne sont pas convergents. Cela soulève une question importante : doit-on se résigner à utiliser des estimateurs non convergents en présence d'une erreur de mesure dans les variables, qu'elle soit classique ou pas ? Heureusement, la réponse est non. Dans le chapitre 15, nous verrons que, sous certaines hypothèses, certains estimateurs conservent leur propriété de convergence, même en présence d'une erreur de mesure générale. Nous n'abordons pas cette discussion à ce stade car il nous faudrait sortir du contexte de l'estimation par les MCO. (Voir le problème 7 dans lequel différentes méthodes de réduction du biais d'atténuation sont proposées.)

Pour aller plus loin 9.3

Soit *educ**, le niveau d'instruction réel, mesuré en années (éventuellement partielles), et *educ*, le plus haut diplôme obtenu. Pensez-vous que la relation entre *educ* et *educ** peut être caractérisée par un modèle d'erreur classique dans les variables ?

9.5 DONNÉES MANQUANTES, ÉCHANTILLONS NON ALÉATOIRES ET OBSERVATIONS EXTRÊMES

Vu que le problème d'erreur de mesure apparaît lorsqu'il est impossible d'obtenir des données exactes sur la variable d'intérêt, il peut à juste titre être considéré comme un problème de fiabilité des données. Sous l'hypothèse d'ECV, le terme d'erreur composite est corrélé avec la variable indépendante mal mesurée, ce qui viole les hypothèses de Gauss-Markov.

La multicolinéarité est un autre problème, lié à l'utilisation de données, que nous avons fréquemment souligné dans les chapitres précédents. On se souvient que la corrélation entre les variables explicatives ne viole aucune hypothèse. Quand deux variables indépendantes sont hautement corrélées, il s'avère plus difficile d'estimer l'effet *ceteris paribus* de chacune d'entre elles, ce que les statistiques habituelles des MCO reflètent.

Dans cette section, nous étudions plus spécifiquement les problèmes de données qui conduisent à la violation de l'hypothèse d'échantillonnage aléatoire, RLM.2. Dans certains cas, l'utilisation d'un échantillonnage non aléatoire n'a aucun effet sur les estimateurs des MCO. Par contre, dans d'autres, les estimateurs des MCO sont biaisés et ne convergent pas. Nous en discutons plus en détail dans le chapitre 17.

Données manquantes

Le problème des **données manquantes** peut se présenter sous plusieurs formes. Dans bon nombre de cas, l'échantillon aléatoire (comprenant des personnes, des villes, des écoles, etc.) est constitué avant que l'on puisse s'apercevoir de l'absence d'observations (concernant plusieurs unités et variables clés dans l'échantillon). Par exemple, dans le jeu de données BWGHT, 197 observations sur un total de 1 388 ne sont pas disponibles ; ces 197 données manquantes concernent le niveau d'instruction (de la mère, du père, ou des deux). Un autre exemple concerne le jeu de données LAWSCH85, qui reprend les salaires médians que reçoivent les jeunes diplômés dans 156 écoles de droit. En réalité, 6 de ces 156 écoles n'ont pas d'information sur la médiane des résultats obtenus par leurs diplômés au test d'aptitude SAT (LSAT), sans compter d'autres variables également manquantes pour certaines écoles.

Si certaines données sont manquantes pour une variable, qu'il s'agisse de la variable dépendante ou d'une variable explicative, l'observation correspondante ne peut pas être utilisée dans une analyse de régression multiple standard. Si les données manquantes sont clairement indiquées, tous les logiciels économétriques actuels sont capables de les identifier automatiquement et de les ignorer lorsqu'il s'agit d'effectuer une régression. Nous l'avions constaté dans l'exemple 4.9, lorsque 197 observations avaient été écartées en raison d'information manquante concernant le niveau d'instruction des parents.

Dans la littérature sur les données manquantes, un estimateur qui utilise seulement les observations relatives à des données complètes pour y et $x_1, x_2, \ldots, x_k$ est appelé estimateur à cas complets ; comme mentionné plus tôt, cet estimateur est calculé par défaut pour les estimateurs des MCO (et tous les estimateurs couverts dans la suite). Y a-t-il, en plus de diminuer la taille de l'échantillon, d'autres conséquences de l'utilisation des MCO en ignorant les données manquantes ? Si, dans le langage de la littérature sur les données manquantes (voir, par exemple, Little et Rubin, (2002, chapitre 1)), les données sont manquantes complètement au hasard (ce qui est parfois appelé MCAR), alors les données manquantes ne posent pas de problème statistique. L'hypothèse MCAR implique que la raison pour laquelle les données sont manquantes, est indépendante, au sens statistique, des facteurs observés et inobservés affectant y. Ainsi, on peut toujours supposer que les données ont été obtenues par échantillonnage aléatoire de la population de telle sorte que l'hypothèse RLM.2 tienne toujours.

Quand l'hypothèse MCAR tient, il y a des moyens d'utiliser l'information partielle obtenue des unités qui ont été écartées par rapport au cas complet. Par exemple, pour un modèle de régression multiple, prenons

le cas où les données sont toujours disponibles pour y et $x_1, x_2, \ldots, x_{k-1}$ mais sont parfois manquantes pour la variable explicative x_k. Une solution habituelle est de créer deux nouvelles variables. Pour une unité i, la première variable, disons z_{ik}, est définie comme x_{ik} quand x_{ik} est observé, et zéro sinon. La seconde variable est un indicateur de donnée manquante, disons m_{ik}, qui est égal à un quand x_{ik} est manquante et zéro quand x_{ik} est observé. Après définition de ces deux variables, toutes les unités sont utilisées dans la régression de y_i contre $x_{i1}, x_{i2}, \ldots, x_{i,k-1}, z_{ik}, m_{ik}, i = 1, \ldots, n$.

On peut montrer que cette procédure produit des estimateurs non biaisés et convergents de tous les paramètres si le mécanisme de données manquantes pour x_k est MCAR. À ce propos, omettre m_{ik} dans la régression est une idée très pauvre, puisqu'il s'agit de la même chose que de supposer que x_{ik} vaut zéro quand elle est manquante. Remplacer les données manquantes par zéro et ne pas inclure les indicateurs de données manquantes peut causer du biais important dans les estimateurs MCO. Une astuce similaire peut être utilisée quand les données sont manquantes pour plus qu'une variable explicative (mais pas pour y). Le problème 9.10 fournit l'argumentation dans le modèle de régression simple.

L'estimateur qui utilise toutes les données et ajoute les indicateurs de données manquantes est en fait moins robuste que l'estimateur à cas complets. Comme nous le verrons dans la prochaine sous-section, l'estimateur à cas complets est convergent même quand la raison pour laquelle les données sont manquantes est systématiquement liée à $(x_1, \ldots, x_k)$, est une fonction de $(x_1, \ldots, x_k)$, pour autant qu'elle ne dépende pas de l'erreur non observée, u. Il existe des techniques plus compliquées pour utiliser l'information partielle qui sont basées sur le « remplissage » des données manquantes, mais celles-ci sont au-delà des objectifs de ce texte. Pour aller plus loin, le lecteur consultera Little et Rubin (2002).

Échantillons non aléatoires

Les données manquantes posent plus de problèmes quand elles résultent d'un **échantillon non aléatoire** de la population. Dans l'exemple 4.9, que se passe-t-il si la probabilité d'obtenir une donnée manquante est plus élevée pour les personnes dont les niveaux d'instruction moyens sont plus bas ? Dans l'exemple de la section 9.2, nous avons utilisé les scores au test de QI pour déterminer le salaire. Cet échantillon de données a été construit en omettant plusieurs personnes pour lesquelles le test de QI n'était pas disponible. S'il s'avère plus facile d'obtenir les résultats au test pour les personnes dont le QI est plus élevé, l'échantillon n'est pas représentatif de la population. L'hypothèse d'échantillonnage aléatoire RLM.2 est alors violée et il importe de se soucier des conséquences que cette violation peut avoir sur les estimations obtenues par les MCO.

Heureusement, certains types d'échantillonnage non aléatoire n'introduisent *pas* de biais dans les estimateurs des MCO ou ne les privent pas de leur propriété de convergence. Sous les hypothèses de Gauss-Markov, sans RLM.2, l'échantillon peut être choisi en fonction des variables *indépendantes*, sans causer de problème statistique. On parle également d'une *sélection de l'échantillon basée sur les variables indépendantes* ; il s'agit d'un exemple d'**échantillonnage exogène**. Afin d'illustrer cette technique, supposons la fonction d'épargne suivante, dans laquelle l'épargne annuelle (*saving*) dépend du revenu (*income*), de l'âge (*age*), de la taille de la famille (*size*), et d'autres facteurs éventuels. Ce modèle simple est

$$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 size + u. \qquad [9.37]$$

Supposons que nos données proviennent d'une enquête qui ne touche que les personnes âgées de 35 ans ou plus ; cela correspond bien à un échantillonnage non aléatoire de la population des adultes. Si cette situation n'est pas idéale, nous pouvons malgré tout utiliser cet échantillon et obtenir des estimateurs sans biais et convergents des paramètres du modèle de la population (9.37). Nous n'allons pas le démontrer sur le plan formel ; précisons simplement que la propriété d'absence de biais est préservée, car la fonction de

régression E(*saving*|*income*, *age*, *size*) est la même pour n'importe quel sous-ensemble de la population décrit par *income*, *age* ou *size*. À condition qu'il y ait suffisamment de variation dans les variables indépendantes de la sous-population, la sélection de l'échantillon qui en résulte ne pose pas de problème particulier (hormis la plus petite taille de l'échantillon).

Dans l'exemple portant sur le QI, la situation n'est pas aussi claire. Il n'y a en effet aucune règle préétablie basée sur le QI qui précise si un individu est inclus ou pas dans l'échantillon. Au lieu de cela, c'est la *probabilité* d'être inclus dans l'échantillon qui augmente avec le QI. Si les autres variables explicatives de l'équation du salaire, qui déterminent également la sélection de l'échantillon, sont indépendantes du terme d'erreur, alors nous sommes à nouveau en présence d'un échantillonnage exogène ; les estimateurs des MCO conservent toutes les propriétés désirables sous les autres hypothèses de Gauss-Markov.

La situation est très différente lorsque la sélection est basée sur la variable dépendante, *y*. Dans ce cas, on parle d'*échantillonnage basé sur la variable dépendante* ; il s'agit d'un exemple d'**échantillonnage endogène**. Dans ce cas, l'échantillonnage est effectué sur base de certaines valeurs de *y*, limitées par une borne inférieure ou supérieure. Les estimateurs des MCO du modèle de population seront toujours biaisés. Supposons, par exemple, que nous désirions estimer la relation entre le patrimoine des individus (*wealth*) et trois facteurs explicatifs (*educ*, *exper*, *age*) pour la population des adultes :

$$wealth = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + u. \qquad \textbf{[9.38]}$$

Si, par exemple, les personnes dont le patrimoine est supérieur à 250 000 dollars sont exclues de l'échantillon, nous ne pouvons pas considérer cet échantillon comme tiré aléatoirement de la population d'intérêt ; il dépend effectivement des valeurs que peut prendre la variable dépendante. L'utilisation d'un échantillon n'incluant que les patrimoines inférieurs à 250 000 dollars conduit à des estimateurs biaisés et non convergents des paramètres de (9.38). Pour faire bref, cette situation résulte du fait que l'espérance dans la population, soit E(*wealth*|*educ*, *exper*, *age*), n'est plus la même si elle devient conditionnelle à *wealth* pour des valeurs inférieures ou égales à 250 000 dollars.

D'autres méthodes d'échantillonnage, dont l'utilisation est souvent délibérée, aboutissent à la constitution d'échantillons non aléatoires. Une méthode répandue de collecte de données est l'**échantillonnage stratifié** qui implique une division de la population en « strates », c'est-à-dire en groupes exhaustifs et disjoints. Ces groupes sont ensuite échantillonnés sans nécessairement respecter les fréquences imposées par la population. Par exemple, dans certaines enquêtes, des groupes minoritaires ou à faibles revenus peuvent être délibérément surreprésentés dans l'échantillon. C'est la nature de la stratification qui détermine si l'emploi de méthodes spécifiques est nécessaire : la stratification peut être exogène (basée sur des variables explicatives) ou endogène (basée sur la variable dépendante). Supposons qu'une enquête surpondère les femmes dans l'échantillon afin d'étudier les facteurs qui influencent leur rémunération dans l'armée. (Dans les échantillons stratifiés, le suréchantillonnage est assez fréquent lorsque le groupe d'intérêt est relativement petit.) Comme les hommes sont également présents dans l'échantillon, nous pouvons utiliser les MCO dans l'échantillon stratifié pour estimer les différences salariales entre genres, ainsi que les rendements de l'expérience et du niveau d'instruction pour tout le personnel militaire. (Nous supposons que les rendements de l'expérience et du niveau d'instruction ne sont pas spécifiques au genre.) Comme la stratification se fait par rapport à une variable explicative (à savoir, le genre), la stratification est exogène et les estimateurs des MCO sont sans biais et convergents.

Par contre, si le personnel militaire sous-payé n'est pas suffisamment représenté dans l'échantillon stratifié, alors les MCO ne permettront pas d'obtenir des estimateurs convergents pour l'équation du salaire, car la stratification est endogène. Dans un tel cas de figure, l'utilisation de méthodes économétriques plus sophistiquées est requise [voir Wooldridge (2010, chapitre 19)].

L'échantillonnage stratifié est une forme manifeste d'échantillonnage non aléatoire. D'autres méthodes d'échantillonnage sont plus subtiles. Dans plusieurs exemples, nous avons estimé les effets de diverses variables, comme le niveau d'instruction et l'expérience, sur le salaire horaire. Par exemple, la base de données WAGE1 provient d'un échantillonnage aléatoire d'individus *qui travaillent*. Les économistes du travail s'intéressent souvent à l'estimation de l'effet du niveau d'instruction sur l'*offre* de salaire. L'idée est la suivante : chaque personne en âge de travailler se trouve face à une offre de salaire horaire ; elle peut soit travailler pour ce niveau de salaire offert, soit ne pas travailler. Pour la personne qui travaille, l'offre de salaire est le salaire gagné ; par contre, pour la personne qui ne travaille pas, l'offre de salaire ne peut pas être observée. Bien que l'équation d'offre de salaire

$$\log(wage^{o}) = \beta_0 + \beta_1 educ + \beta_2 exper + u \qquad \textbf{[9.39]}$$

soit valable pour la population de toutes les personnes en âge de travail, nous ne pouvons pas l'estimer en utilisant un échantillon aléatoire de la population. En effet, nous ne pouvons obtenir de données sur l'offre de salaire que pour les personnes qui travaillent (même si les données sur *educ* et *exper* sont disponibles pour les personnes qui ne travaillent pas). Si nous utilisons malgré tout un échantillon aléatoire de travailleurs pour estimer (9.39), pouvons-nous obtenir des estimateurs sans biais ? Il est impossible d'apporter une réponse univoque à cette question. Notons tout d'abord que nous ne sommes pas dans le même cas que le précédent : l'échantillon n'est pas constitué en excluant les petits salaires, ce qui aurait constitué un exemple manifeste de stratification endogène. L'échantillon est ici dépendant de la décision de travailler, qui peut être considérée comme exogène. Néanmoins, il faut être prudent : la décision de travailler peut être liée à des facteurs non observés qui affectent le salaire offert. Dans un tel cas de figure, la décision de travailler devient endogène, ce qui introduit un biais de sélection d'échantillonnage dans les estimateurs des MCO. Dans le chapitre 17, nous abordons plusieurs méthodes qui permettent de tester et de corriger ce biais de sélection.

Pour aller plus loin 9.4

Vous cherchez à estimer les effets des dépenses de campagne électorale des élus sortants sur les résultats du scrutin. Sachant que vous ne pouvez pas observer ces deux variables pour des élus sortants qui choisissent de ne pas se représenter, faites-vous face à un problème de sélection endogène de l'échantillon ?

Observations aberrantes

Dans certaines applications, les estimations par MCO sont sensibles à l'inclusion d'une ou plusieurs observations, particulièrement en présence de petits échantillons. Un traitement approfondi de l'impact que peut avoir l'inclusion d'**observations aberrantes**, dépasse la portée de cet ouvrage ; cela exigerait notamment des développements plus pointus en algèbre matriciel. Pour faire simple, une observation est aberrante si son exclusion conduit à une variation « importante » des estimations par MCO. La notion reste assez vague, car elle compare des valeurs de variables pour une observation avec celles du reste de l'échantillon. Il convient néanmoins de se tenir à l'affût de telles observations, tant elles peuvent affecter les estimations obtenues par les MCO.

Les estimateurs des MCO sont sensibles à ces observations pour la simple raison que la méthode des MCO vise à minimiser la somme des carrés des résidus (SCR) : les grands résidus (positifs ou négatifs) reçoivent un poids plus important dans le problème de minimisation des moindres carrés. Si les estimations subissent de grands changements suite à une légère modification de l'échantillon, il est important de s'en préoccuper.

Quand les statisticiens et économètres étudient les aspects théoriques de ce problème, ils considèrent tantôt que les données viennent d'un échantillon aléatoire issu d'une population donnée (en

considérant néanmoins une distribution inhabituelle pouvant produire de telles valeurs extrêmes), tantôt que les observations aberrantes viennent d'une population différente. Sur le plan pratique, l'existence d'observations aberrantes peut s'expliquer de deux façons. Le cas le plus facile à traiter est celui où une erreur s'est glissée au moment de l'encodage des données. Ajouter des zéros ou déplacer la virgule peut, en effet, détériorer sensiblement les estimations par MCO, particulièrement dans les échantillons de petite taille. Il est toujours utile de calculer des statistiques sommaires, en particulier les minima et maxima, afin de détecter ce type d'erreurs d'encodage. Malheureusement, ce n'est pas toujours évident de les repérer.

Des données extrêmes peuvent également apparaître lorsque l'échantillon est tiré d'une petite population au sein de laquelle une ou plusieurs unités sont très différentes du reste de la population, en raison de caractéristiques pertinentes. La décision de garder ou de laisser tomber de telles observations est difficile à prendre, et les propriétés statistiques des estimateurs qui résultent de ce choix sont compliquées. Il est vrai que les observations extrêmes peuvent apporter une information précieuse en augmentant la variation des variables explicatives (ce qui réduit les écarts-types estimés). En règle générale, si les résultats des MCO changent substantiellement en fonction de l'inclusion ou de l'exclusion de certaines données, il convient de l'indiquer et de comparer les résultats avec soin.

EXEMPLE 9.8
Dépenses en r&d et taille de la firme

Nous supposons que les dépenses en R&D (*rdintens*, en pourcentage du chiffre d'affaires) sont liées au chiffre d'affaires (*sales*, en millions) et aux profits (*profmarg*, en pourcentage du chiffre d'affaires) :

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 profmarg + u. \qquad \textbf{[9.40]}$$

En incluant les 32 entreprises chimiques reprises dans la base de données RDCHEM, la régression estimée par les MCO est

$$\widehat{rdintens} = 2{,}625 + 0{,}000053\ sales + 0{,}0446\ profmarg$$
$$(0{,}586)\quad (0{,}000044)\qquad (0{,}0462)$$
$$n = 32,\ R^2 = 0{,}0761,\ \bar{R}^2 = 0{,}0124.$$

Aucune variable explicative n'est statistiquement significative, même à un seuil de 10 %.

Parmi les 32 entreprises, 31 enregistrent un chiffre d'affaires inférieur à 20 milliards de dollars. Une d'entre elles affiche un chiffre d'affaires supérieur à 40 milliards de dollars. La figure 9.1 montre l'écart qui existe entre cette entreprise et le reste de l'échantillon. Cette entreprise enregistre un chiffre d'affaires, au minimum, deux fois supérieur aux autres ; il est donc intéressant d'estimer le modèle sans elle. Nous obtenons

$$\widehat{rdintens} = 2{,}297 + 0{,}000186\ sales + 0{,}0478\ profmag$$
$$(0{,}592)\ (0{,}000084)\qquad (0{,}0445)$$
$$n = 31,\ R^2 = 0{,}1728,\ \bar{R}^2 = 0{,}1137.$$

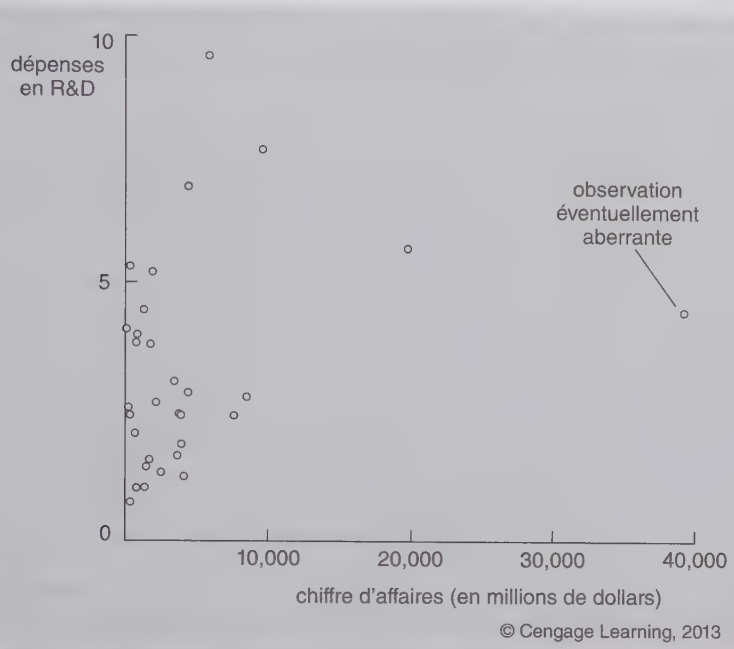


Figure 9.1 *Nuage de points de « l'intensité » en R&D sur le chiffre d'affaires de l'entreprise.*

Lorsque l'entreprise la plus importante est exclue de l'échantillon, le coefficient de *sales* est multiplié par un facteur supérieur à trois, et sa statistique *t* est désormais supérieure à deux. Sur base de l'échantillon composé des 31 plus petites firmes, il existe un effet positif et significatif du chiffre d'affaires sur sa part consacrée à la R&D. Autrement dit, l'effet *ceteris paribus* de la « taille » de l'entreprise sur son « intensité » en R&D est positif et significatif. Par contre, le coefficient de *profmarg* ne change pas beaucoup et la variable n'est toujours pas significative.

Dans certains cas, les observations sont considérées comme aberrantes en fonction de la taille du résidu que l'on obtient par les MCO en prenant en compte *toutes* les observations. En règle générale, ce *n*'est *pas* une bonne idée pour la simple raison que la méthode des MCO tend précisément à minimiser la SCR. Dans l'exemple précédent, le fait d'inclure la plus grande firme aplatit considérablement la droite de régression des MCO, ce qui « masque » la nature spécifique de cette firme. Lorsque les 32 entreprises sont incluses, le résidu ne vaut que – 1,62. Cette valeur ne vaut même pas une fois l'écart-type estimé, soit $\hat{\sigma} = 1{,}82$ (en considérant une moyenne des résidus égale à zéro par construction).

Il existe une meilleure idée, celle qui consiste à calculer le **résidu de Student** (ou « résidu studentisé »). Les résidus de Student sont calculés en divisant les résidus originaux par l'estimation de leur écart-type (conditionnelle aux variables explicatives dans l'échantillon). La formule des résidus de Student requiert l'utilisation de l'algèbre matriciel ; il existe néanmoins une astuce qui permet de calculer ce résidu pour n'importe quelle observation. Définissons une variable binaire égale à 1 pour une observation *h* ; incluons cette variable dans la régression (en utilisant toutes les observations), au côté de toutes les autres variables explicatives. Le coefficient de la variable binaire a une interprétation utile : il s'agit du résidu de l'observation *h* calculé à partir de la droite de régression en tenant compte de toutes les *autres* observations. Par conséquent, le coefficient de la variable binaire permet d'évaluer l'éloignement de l'observation *h* par rapport à la droite de régression qui est obtenue sans faire intervenir cette observation dans le calcul de minimisation de la SCR. Mieux même, la statistique *t* pour le coefficient de cette variable binaire correspond au résidu de Student

pour l'observation h. Sous les hypothèses du modèle de régression linéaire classique (MRLC), la statistique t suit une distribution t_{n-k-2}. Une grande valeur de la statistique t (en valeur absolue) implique donc un grand résidu par rapport à son écart-type estimé.

Lorsque nous ajoutons, dans la régression de l'exemple 9.8, une variable explicative binaire pour la plus grande firme (soit la dixième observation dans le fichier de données), le coefficient estimé vaut – 6,57 ; cette observation est donc très éloignée de la droite de régression calculée en minimisant la SCR des *autres* observations. Néanmoins, le résidu de Student ne vaut que – 1,82. La statistique t est marginalement significative (puisque la p-valeur du test bilatéral est égale à 0,08). En réalité, ce résidu de Student n'est pas le plus grand dans l'échantillon. Si nous utilisons la même méthode pour l'entreprise qui affiche l'intensité en R&D la plus élevée (soit la première observation, avec *rdintens* ≈ 9,42), le coefficient de la variable binaire vaut 6,72 avec une statistique t égale à 4,56. Si on écarte la première observation, le coefficient de *profmarg* devient plus grand et significatif, même s'il est vrai que le coefficient de *sales* change peu (passant de 0,000053 à 0,000051). Nous pourrions donc conclure que la première observation est plus « inhabituelle » que la dixième. Faut-il la considérer comme « aberrante » pour autant ? Cette analyse montre le dilemme auquel nous faisons face lorsqu'il s'agit d'écarter une observation de l'analyse de régression ; ce choix est cornélien, même lorsque la base de données est petite. En réalité, la valeur absolue d'un résidu de Student ne reflète pas nécessairement le degré d'influence que peut avoir une observation sur l'estimation des pentes obtenues pas les MCO.

Pour calculer le résidu de Student d'une observation en particulier, toutes les *autres* observations servent à estimer la droite de régression, ce qui peut poser problème. Quand le résidu de Student est calculé pour la première observation, la dixième observation a été utilisée pour estimer la pente et l'ordonnée à l'origine. Étant donné les petites valeurs des coefficients de pente (en valeur absolue) qui résultent de l'inclusion de la plus grande entreprise (soit la dixième observation), il n'est pas surprenant que la première observation soit éloignée de la droite de régression puisqu'elle correspond à une valeur très élevée de *rdintens*.

Bien sûr, nous pouvons ajouter deux variables binaires dans la régression, l'une pour la première et l'autre pour la dixième donnée. Le nombre d'observations restantes pour estimer la droite de régression tombe à 30. Sans les première et dixième observations, les résultats de la régression deviennent

$$\widehat{rdintens} = 1{,}939 + 0{,}000160\,sales + 0{,}0701\,profmarg$$
$$(0{,}459)\quad (0{,}000065)\qquad (0{,}0343)$$
$$n = 30,\ \bar{R}^2 = 0{,}2711,\ \bar{R}^2 = 0{,}2171.$$

Nous aurions obtenu les mêmes estimations pour ces trois coefficients en utilisant les 32 observations *et* en ajoutant les deux variables binaires. Pour la première variable binaire (relative à la première observation), le coefficient de la variable binaire vaut 6,47 (t = résidu de Student = 4,58) ; pour la seconde variable binaire (relative à la dixième observation), il vaut – 5,41 (t = résidu de Student = – 1,95). Les deux coefficients de *sales* et de *profmarg* sont statistiquement significatifs ; celui de *profmarg* est significatif à un seuil de 5 % environ, pour un test bilatéral (p-valeur = 0,051). Même dans cette régression, subsistent deux observations dont les résidus de Student sont plus grands que deux ; elles correspondent à deux entreprises dont l'intensité en R&D est supérieure à six (pourcents).

Certaines formes fonctionnelles sont moins sensibles aux observations extrêmes. Dans la section 6.2, nous avons indiqué que, pour la plupart des variables économiques, la transformation logarithmique permet de « lisser les données », c'est-à-dire d'en restreindre significativement la dispersion. Cette transformation permet d'obtenir des formes fonctionnelles qui peuvent s'appliquer à une plus large gamme de problèmes, à l'image du modèle à élasticité constante.

EXEMPLE 9.9
Intensité de la R&D

Utilisons un modèle à élasticité constante pour tester si « l'intensité » en R&D (mesurée par la part que représentent les dépenses en R&D dans le chiffre d'affaires) augmente avec « la taille » de l'entreprise (mesurée par son chiffre d'affaires). Le modèle sous-jacent est

$$rd = sales^{\beta_1} \exp(\beta_0 + \beta_2\, profmarg + u). \qquad [9.41]$$

Toutes choses étant égales par ailleurs, l'intensité de la R&D augmente avec *sales* si et seulement si $\beta_1 > 1$. En utilisant la transformation du logarithme népérien pour linéariser le modèle par rapport à ses paramètres, nous obtenons

$$\log(rd) = \beta_0 + \beta_1 \log(sales) + \beta_2 profmarg + u. \qquad [9.42]$$

Sur base des 32 firmes, la régression estimée donne

$$\widehat{\log(rd)} = -4{,}378 + 1{,}084 \log(sales) + 0{,}0217\, profmarg,$$
$$(0{,}468)\ (0{,}060) \qquad (0{,}0128)$$
$$n = 32,\ R^2 = 0{,}9180,\ \bar{R}^2 = 0{,}9123,$$

Si nous excluons la « plus grande » entreprise de l'échantillon (soit la dixième observation dans le fichier de données), nous avons

$$\widehat{\log(rd)} = -4{,}404 + 1{,}088 \log(sales) + 0{,}0218\, profmarg,$$
$$(0{,}511)\ (0{,}067) \qquad (0{,}0130)$$
$$n = 31,\ R^2 = 0{,}9037,\ \bar{R}^2 = 0{,}8968.$$

Sur le plan pratique, les résultats sont équivalents. Dans les deux cas, l'hypothèse nulle, $H_0 : \beta_1 = 1$ contre $H_1 : \beta_1 > 1$, n'est *pas* rejetée. (Pouvez-vous le vérifier ?)

Dans certains cas, nous pouvons, dès le début, identifier le caractère atypique de certaines observations par rapport au reste de l'échantillon. Cela arrive souvent lorsqu'il s'agit de données très agrégées, au niveau des villes, régions ou états. L'exemple suivant en donne une illustration.

EXEMPLE 9.10
Taux de mortalité infantile par état

Dans le recueil « *Statistical Abstract of the United States* » de 1990, nous pouvons obtenir des données, agrégées au niveau de chaque état, pour la mortalité infantile, le revenu par tête, et différentes mesures de soins de santé. Nous réalisons ici une analyse simple dans le but d'illustrer l'effet que peuvent avoir des observations aberrantes. Les données portent sur l'année 1990 et couvrent les 50 états des États-Unis, ainsi que le District de Columbia (D.C.). La variable *infmort* correspond au nombre de décès survenus au cours de la première année d'existence (pour 1000 naissances) ; *pcinc* donne le revenu par tête, en moyenne dans l'État ; *physic* est le nombre de médecins pour 100 000 habitants ; et *popul* représente la population civile (en milliers d'habitants). En utilisant la base données INFMRT et en introduisant toutes les variables indépendantes sous forme logarithmique, nous obtenons :

$$\widehat{infmort} = 33{,}86 - 4{,}68 \log(pcinc) + 4{,}15 \log(physic)$$
$$(20{,}43)\ (2{,}60) \qquad (1{,}51)$$
$$-0{,}088 \log(popul) \qquad [9.43]$$
$$(0{,}287)$$
$$n = 51,\ R^2 = 0{,}139,\ \bar{R}^2 = 0{,}084.$$

Toutes choses étant égales par ailleurs, une augmentation du revenu par tête conduit à une diminution du taux de mortalité. Si ce résultat était attendu, le suivant l'est beaucoup moins : une augmentation du nombre de médecins est associée à une augmentation de la mortalité. Par contre, la taille de la population ne semble pas liée à la mortalité infantile.

En réalité, le District de Columbia est très atypique, en ce sens qu'il contient à la fois des zones d'extrême pauvreté et des quartiers de grande opulence, le tout confiné dans de petits espaces. Ce district enregistre le taux de mortalité le plus élevé de l'échantillon (soit 20,7), le taux immédiatement inférieur valant 12,4. Ce district contient également le plus grand nombre de médecins, soit 615 pour 100,000 membres de la population civile, bien au-delà des 337 médecins pour l'État suivant. Ensemble, ces deux données atypiques (pour les médecins et la mortalité infantile) peuvent certainement influencer les résultats. Si nous excluons ce district de la régression, nous avons :

$$\begin{aligned} \widehat{infmort} &= 23{,}95 - 0{,}57\log(pcinc) - 2{,}74\log(physic) \\ &\quad (12{,}42)\ (1{,}64)\qquad\qquad (1{,}19) \\ &\quad + 0{,}629\log(popul) \\ &\quad (0{,}191) \\ n &= 50,\ R^2 = 0{,}273,\ \overline{R}^2 = 0{,}226. \end{aligned} \qquad \textbf{[9.44]}$$

Désormais, un plus grand nombre de médecins conduit à une diminution, *ceteris paribus*, de la mortalité infantile ; l'estimation est statistiquement différente de zéro à un seuil de 5 %. Par contre, l'effet du revenu par tête chute drastiquement et n'est plus significatif sur le plan statistique. Ensuite, dans l'équation (9.44), le taux de mortalité infantile est plus élevé dans les états plus peuplés, la relation étant statistiquement très significative. Enfin, comme l'indique le R carré, les variations de *infmort* sont beaucoup mieux expliquées lorsque le District de Columbia est écarté de la régression. Il apparaît clairement que les estimations dépendent fortement de la prise en considération du District de Columbia. Au bout du compte, il semble plus opportun de l'exclure de toute analyse ultérieure.

Comme le démontre l'exemple 9.8, il est délicat de distinguer les observations atypiques des observations aberrantes et d'identifier les observations qui impactent le plus les estimations des MCO. La visualisation des données peut aider dans certains cas, mais un traitement plus formel est souvent nécessaire. En recourant à l'algèbre matriciel, Belsley, Kuh et Welsh (1980) ont, par exemple, défini « *l'effet de levier* » que peut avoir une observation ; plus cet effet est élevé, plus grande est l'influence de l'observation sur les estimations obtenues par les MCO. Ces auteurs discutent également en profondeur de l'utilité des résidus standardisés et de Student.

9.6 ESTIMATION PAR MOINDRES DÉVIATIONS ABSOLUES

Il existe une autre manière de se prémunir des méfaits de l'inclusion d'observations aberrantes dans la régression. Cette approche ne vise pas à identifier les observations dont l'influence sur les estimations par les MCO serait injustifiée. Elle consiste plutôt à utiliser une méthode d'estimation moins sensible aux observations aberrantes que les MCO. Cette méthode est celle des **moindres déviations absolues (MDA)**, devenue populaire chez les économètres appliqués. Dans un modèle estimé par les MDA, les estimateurs des coefficients β_j sont calculés en minimisant la somme des valeurs absolues des résidus, soit

$$\min_{b_0, b_1, \ldots b_k} \sum_{i=1}^{n} |y_i - b_0 - b_1 x_{i1} - \ldots - b_k x_{ik}|. \qquad \textbf{[9.45]}$$

Contrairement aux MCO dont l'objectif est de minimiser la SCR, les estimations obtenues par les MDA ne peuvent pas être calculées à partir d'une formule, car il n'existe pas de forme analytique explicite pour les estimateurs. Dans le passé, la résolution du problème de minimisation (9.45) était difficile ; elle l'était d'autant plus que la taille de l'échantillon était grande ou que le nombre de variables explicatives était élevé. Depuis lors, la puissance de calcul des ordinateurs s'est considérablement améliorée et les estimations par MDA sont beaucoup plus faciles à obtenir, même pour de grandes bases de données.

Sur la figure 9.2, la fonction « objectif » des MDA est comparée à celle des MCO. La fonction « objectif » des MDA est linéaire en partant de zéro, si bien qu'elle augmente d'une unité suite à l'augmentation d'un résidu positif de même ampleur. A contrario, la fonction « objectif » des MCO donne une importance croissante aux résidus élevés (en valeurs absolues), ce qui rend les MCO plus sensibles aux observations aberrantes.

Comme les MDA ne donnent pas de poids croissants aux résidus élevés, ils sont moins sensibles que les MCO aux valeurs extrêmes. En fait, les MDA sont conçus pour estimer les paramètres de la **médiane conditionnelle** de y étant donné x_1, x_2, ..., x_k (plutôt que ceux de la moyenne conditionnelle). Étant donné que la médiane n'est pas influencée par de fortes variations dans les observations extrêmes, les estimations par MDA sont plus robustes aux observations aberrantes. (Voir section A.1 pour une brève discussion sur la médiane de l'échantillon.) Pour déterminer les estimations qui minimisent la fonction « objectif », la méthode des MCO utilise chaque résidu au carré, rendant le processus de minimisation très sensible aux observations aberrantes, comme nous l'avons vu dans les exemples 9.8 et 9.10.

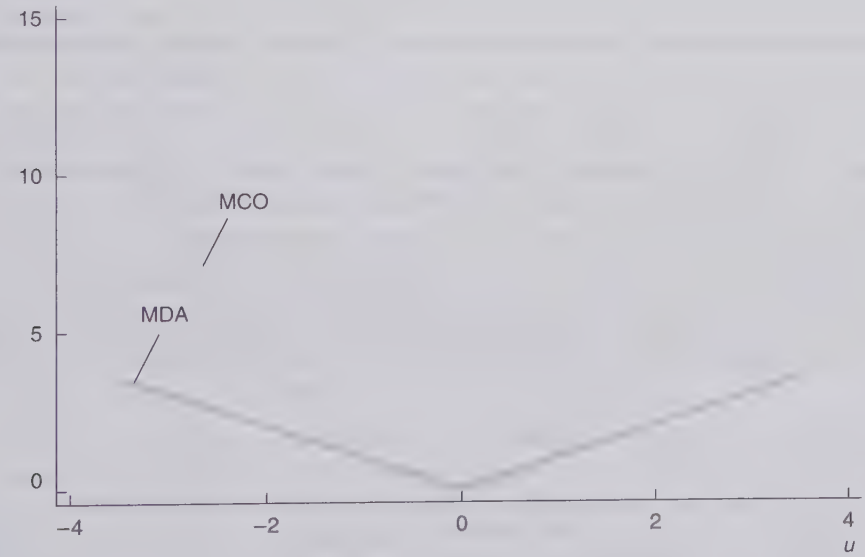


Figure 9.2 *Les fonctions « objectifs » des MCO et des MDA.*

Les MDA ont aussi leurs inconvénients. Nous avons cité le temps de calcul mais il y a également les tests d'hypothèse qui ne sont valides que sur le plan asymptotique (en grands échantillons). [Les formules sont quelque peu compliquées et requièrent des notions d'algèbre matriciel. Nous n'y ferons pas appel dans cet ouvrage. Koenker (2005) en fournit un traitement exhaustif.] Pour rappel, sous les hypothèses du MRLC, les statistiques t et F issues des MCO ont des distributions t et F exactes. Certes, des versions asymptotiques de ces statistiques sont disponibles pour les MDA et sont utilisées automatiquement dans les logiciels économétriques ;

mais leur utilisation ne se justifie qu'en présence de grands échantillons. Cela étant dit, l'absence d'inférence exacte pour les MDA n'est pas très pénalisante, tout comme ne l'est pas le temps de calcul requis. Beaucoup d'applications basées sur les MDA reposent sur des centaines, voire des milliers d'observations. Bien sûr, ce serait sans doute exagéré de recourir à ces approximations asymptotiques dans des exemples similaires à l'exemple 9.8, où $n = 32$. Pourtant, dans un certain sens, ce n'est pas très différent de ce que nous faisons avec les MCO ; bien souvent, lorsqu'une hypothèse du MLRC n'est pas respectée, nous devons recourir aux approximations en grands échantillons pour justifier l'utilisation des tests d'inférence statistique.

Les MDA présentent un autre inconvénient, à la fois plus sérieux et plus subtil. Les MDA n'aboutissent pas toujours à des estimateurs convergents des paramètres de $E(y|x_1, \ldots, x_k)$, qui correspond à la fonction de *moyenne* conditionnelle. Comme nous l'avons indiqué plus haut, les MDA visent à estimer les effets *ceteris paribus* sur la *médiane* conditionnelle. En règle générale, la médiane et la moyenne ne sont pas identiques, sauf dans le cas particulier où la distribution de y, étant donné $x_1, \ldots, x_k$, est symétrique autour de $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. (Autrement dit, lorsque la distribution du terme d'erreur de la population, u, est symétrique autour de zéro.) Pour rappel, les MCO produisent des estimateurs sans biais et convergents des paramètres de la moyenne conditionnelle, quel que soit le degré d'asymétrie de la distribution de l'erreur ; la symétrie n'apparaît pas dans les hypothèses de Gauss-Markov. Lorsque les distributions sont asymétriques, les effets *ceteris paribus* estimés par les MDA ne correspondent généralement pas à ceux estimés par les MCO ; et cette différence entre estimations peut simplement s'expliquer par la différence qui existe entre la moyenne et la médiane, sans que la présence d'observations aberrantes ne soit nécessaire, de quelque manière que ce soit. (Voir l'exercice sur ordinateur C9 pour un exemple.)

Lorsque la symétrie de la distribution de u n'est pas garantie, l'erreur d'échantillonnage ne parvient à expliquer la différence entre les estimations des *pentes* obtenues par MCO et MDA que si l'erreur u du modèle de la population (9.2) est *indépendante* de $(x_1, \ldots, x_k)$. Quant aux estimations des *ordonnées à l'origine* provenant de MCO et des MDA, elles seront généralement différentes car, si la distribution u est asymétrique et que la moyenne est nulle, la médiane sera différente de zéro. Malheureusement, lorsque la méthode des MDA est utilisée, l'hypothèse d'indépendance entre l'erreur et les variables explicatives n'est pas très réaliste. En effet, l'hypothèse d'indépendance exclut l'hétéroscédasticité, alors que ce problème se rencontre fréquemment dans les applications où la distribution est asymétrique.

Par rapport aux MCO, les MDA permettent d'obtenir très facilement les effets *ceteris paribus* des x_j sur y en utilisant des transformations monotones. Prenons le cas de la transformation logarithmique classique. Supposons que $\log(y)$ suive un modèle linéaire dans lequel la *médiane* conditionnelle de l'erreur est nulle :

$$\log(y) = \beta_0 + x\beta + u \qquad \textbf{[9.46]}$$

$$\text{Med}(u|x) = 0, \qquad \textbf{[9.47]}$$

ce qui implique

$$\text{Med}[\log(y)|x] = \beta_0 + x\beta$$

Une caractéristique bien connue de la médiane conditionnelle est qu'elle peut être « transférée » dans des fonctions croissantes. [Voir, par exemple, Wooldridge (2010, chapitre 12]. Par conséquent,

$$\text{Med}(y|\mathbf{x}) = \exp(\beta_0 + x\beta). \qquad \textbf{[9.48]}$$

Il s'ensuit que β_j est la semi-élasticité de $\text{Med}(y|\mathbf{x})$ par rapport à x_j. En d'autres mots, l'effet *ceteris paribus* de x_j dans l'équation linéaire (9.46) peut être utilisé pour obtenir l'effet *ceteris paribus* de x_j dans le modèle non linéaire (9.48). Notez bien qu'il suffit que (9.47) soit vraie pour que ce résultat reste valide, *quelle que soit la distribution de u* ; l'hypothèse d'indépendance entre u et $\mathbf{x}$ n'est pas requise. Si nous utilisons l'espérance plutôt que la médiane, le cheminement n'est pas aussi simple. En effet, il est généralement impossible de retrouver $E(y|\mathbf{x})$ si nous spécifions un modèle linéaire pour $E[\log(y)|\mathbf{x}]$. Il est vrai que nous pouvons

(théoriquement) retrouver E(y|**x**) lorsque la distribution de u étant donné **x** jouit de *toutes* les propriétés désirables. Nous avons étudié un cas spécial dans l'équation (6.40), sous l'hypothèse que log(y) suivait le MRLC. Néanmoins, alors que nous pouvons toujours obtenir Med(y|**x**) à partir de Med[log(y)|**x**], il n'est généralement pas possible de trouver E(y|**x**) à partir d'un modèle pour E[log(y)|**x**]. Dans le problème 9, nous verrons que la présence d'hétéroscédasticité dans un modèle linéaire pour log(y) nous empêche de trouver E(y|**x**).

Les MDA sont un cas particulier des méthodes de *régression robuste*. Malheureusement, l'utilisation du qualificatif « robuste » peut ici prêter à confusion. Dans la littérature statistique, un estimateur robuste est relativement insensible aux observations extrêmes. Effectivement, nous donnons, aux observations dont les résidus sont élevés, un poids plus faible dans le cadre d'une régression robuste que dans le cas d'une régression estimée par les MCO. [Berk (1990) offre une analyse accessible des estimateurs robustes aux observations aberrantes.] Pourtant, dans le jargon économétrique, les MDA ne sont pas considérées comme des estimateurs robustes de la *moyenne* conditionnelle, car leur propriété de convergence (vers la valeur vraie de la moyenne) dépend d'hypothèses supplémentaires. Dans l'équation (9.2), nous avons vu qu'il s'agissait soit de l'hypothèse de symétrie (centrée sur zéro) de la distribution de u étant donné (x_1, ..., x_k), soit de l'hypothèse d'indépendance entre u et (x_1, ..., x_k). Notez bien qu'aucune de ces deux hypothèses n'est requise lorsque la méthode des MCO est utilisée.

Les MDA sont également considérées comme un cas spécifique des méthodes de *régression quantile*. Ce type de régression est utilisée pour estimer l'effet des x_j sur différents « segments » (ou quantiles) de la distribution, pas uniquement la médiane ou la moyenne. Par exemple, si nous visons à étudier l'effet sur le patrimoine de l'accès à un plan de pension via l'employeur (assurance-groupe), il est tout à fait possible que cet effet pour les 10 % des salaires les plus élevés diffère de celui pour les 10 % des salaires les plus bas ; ces deux effets peuvent également différer de l'effet sur les travailleurs dont le salaire est proche de la médiane. Wooldridge (2010, chapitre 12) approfondit l'étude de la régression quantile et en donne plusieurs exemples.

RÉSUMÉ

Nous avons examiné plusieurs problématiques liées à la spécification des modèles et à l'utilisation des données. Ces difficultés se rencontrent fréquemment dans les analyses empiriques en coupe instantanée. Même lorsque toutes les variables requises sont présentes dans le modèle, une mauvaise spécification de la forme fonctionnelle complexifie l'interprétation de l'équation estimée. Il est néanmoins possible de détecter une forme fonctionnelle incorrecte en ajoutant des termes quadratiques, comme dans le RESET, ou en testant le modèle contre une alternative non emboîtée, comme dans le test de Davidson-MacKinnon. La collecte de données supplémentaires n'est pas indispensable.

Il est plus ardu de corriger une forme fonctionnelle lorsqu'une variable clé est omise. Dans la section 9.2, nous avons proposé d'utiliser une variable de substitution, qui permet de remplacer la variable omise. Sous des hypothèses raisonnables, l'inclusion d'une telle variable réduit, voire élimine, le biais présent dans une régression estimée par les MCO. L'inconvénient de cette méthode est que les variables de substitution ne sont pas faciles à identifier. Une règle ad hoc est d'utiliser les observations de la variable dépendante pour la période précédente.

En économie appliquée, les chercheurs sont souvent confrontés à des erreurs de mesure. Sous les hypothèses d'erreur classique dans les variables (ECV), l'utilisation d'une variable dépendante mal mesurée n'a pas d'effet sur les propriétés statistiques des estimateurs des MCO. Par contre, lorsqu'une ECV touche une variable indépendante, l'estimateur des MCO pour le coefficient de la variable mal mesurée, est biaisé vers zéro. Le biais introduit dans les autres coefficients peut être positif ou négatif ; il est difficile de le savoir.

L'utilisation d'un échantillon non aléatoire peut introduire un biais dans les estimateurs des MCO. Dans un échantillonnage endogène, lorsque la sélection de l'échantillon est corrélée avec le terme d'erreur

u, les estimateurs des MCO sont généralement biaisés et ne convergent pas. Par contre, un échantillonnage exogène (qui est soit basé sur les variables explicatives, soit indépendant de u) ne pose pas de problème pour les estimateurs des MCO.

La présence de données aberrantes dans les bases de données peut avoir de lourdes conséquences sur les estimations obtenues par les MCO, surtout en petits échantillons. Il est important d'identifier, au moins de manière informelle, les données extrêmes et d'estimer à nouveau les modèles en excluant les données qui sont susceptibles d'être aberrantes.

L'estimation d'un modèle par les moindres déviations absolues (MDA) constitue une alternative à l'estimation par les MCO. Les MDA sont moins sensibles aux données aberrantes et permettent d'obtenir des estimations convergentes des paramètres de la *médiane* conditionnelle. Grâce à la puissance de calcul des processeurs actuels et une meilleure compréhension des avantages et inconvénients des différentes méthodes d'estimation, l'utilisation des MDA, comme méthode complémentaire aux MCO, est devenue de plus en plus fréquente dans les travaux empiriques.

MOTS-CLÉS

Biais d'atténuation p. 383
Données manquantes p. 386
Échantillonnage endogène p. 388
Échantillonnage exogène p. 387
Échantillon non aléatoire p. 387
Échantillonnage stratifié p. 388
Effet marginal moyen (EMM) p. 378
Effet partiel moyen (EPM) p. 378
Erreur classique dans les variables (ECV) p. 383
Erreur de mesure p. 364, 380
Erreur de mesure multiplicative p. 381
Erreur de spécification de la forme fonctionnelle p. 364
Estimateur à cas complets p. 386
Manquer au hasard p. 386
Manquer complètement au hasard (MCAR) p. 386
Médiane conditionnelle p. 395
Modèle à coefficient (ou pente) aléatoire p. 377
Modèles non emboîtés p. 368
Moindres déviations absolues (MDA) p. 394
Observations aberrantes p. 389
Résidus de Student (ou résidus studentisés) p. 391
Test de Davidson-MacKinnon p. 369
Test d'erreur de spécification de la régression (RESET) p. 367
Variable dépendante retardée p. 374
Variable explicative endogène p. 364
Variable de substitution p. 370

PROBLÈMES

1. Dans le problème 11 du chapitre 4, nous avions obtenu $R^2 = 0{,}353$ suite à l'estimation du modèle

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \log(mktval) + \beta_3 profmarg + \beta_4 coeten + \beta_5 comten + u,$$

sachant que $n = 177$ et que la base de données était CEOSAL2. Lorsque nous ajoutons les variables $ceoten^2$ et $comten^2$, $R^2 = 0{,}375$. Peut-on soupçonner une erreur de spécification de la forme fonctionnelle dans ce modèle ?

2. Nous modifions à présent l'exercice sur ordinateur C4 du chapitre 8 ; cet exercice portait sur les résultats électoraux de 1990 pour les candidats sortants, élus en 1988 lors d'un vote bipartite. Le candidat A, élu en 1988, cherche à se faire réélire en 1990. *voteA90* désigne la part des voix allouées au candidat A lors du vote bipartite de 1990. Le score du candidat en 1988 (*voteA88*) est utilisé comme variable de substitution de la qualité du candidat. Toutes les autres variables concernent l'élection de 1990 ; elles ne sont pas au centre de nos préoccupations dans cet exercice. En utilisant la base de données VOTE2, les équations suivantes ont été estimées :

$$\widehat{voteA90} = 75{,}71 + 0{,}312\, prtystrA + 4{,}93\, democA$$
$$\quad (9{,}25)\ (0{,}046)\qquad (1{,}01)$$
$$\quad -0{,}929 \log(\exp endA) - 1{,}950 \log(\exp endB)$$
$$\quad (0{,}684)\qquad (0{,}281)$$
$$n = 186,\ R^2 = 0{,}495,\ \bar{R}^2 = 0{,}483,$$

et

$$\widehat{voteA90} = 70{,}81 + 0{,}282\, prtystrA + 4{,}52\, democA$$
$$\quad (10{,}01)\ (0{,}052)\qquad (1{,}06)$$
$$\quad -0{,}839 \log(\exp endA) - 1{,}846 \log(\exp endB) + 0{,}067\, voteA88$$
$$\quad (0{,}687)\qquad (0{,}292)\qquad (0{,}053)$$
$$n = 186,\ R^2 = 0{,}499,\ \bar{R}^2 = 0{,}485.$$

i. Interprétez le coefficient associé à la variable *voteA88* et commentez sa significativité statistique.

ii. L'ajout de la variable *voteA88* modifie-t-il substantiellement les coefficients associés aux autres variables ?

3. Soit *math10*, le pourcentage d'étudiants qui sont inscrits dans une école du Michigan et qui ont réussi un test standardisé de mathématiques (voir également l'exemple 4.2). Nous cherchons à estimer l'effet des dépenses par étudiant sur les performances en mathématiques de ces derniers. Un modèle simple est

$$math10 = \beta_0 + \beta_1 \log(expend) + \beta_2 \log(enroll) + \beta_3 poverty + u,$$

où *poverty* correspond au pourcentage d'étudiants vivant sous le seuil de pauvreté.

i. La variable *lnchprg* désigne le pourcentage d'étudiants ayant droit au programme fédéral de repas subventionnés. Pourquoi peut-on considérer cette variable comme une variable de substitution raisonnable pour *poverty* ?

ii. Le tableau ci-dessous contient des estimations obtenues par les MCO, avec et sans *lnchprg* parmi les variables explicatives.

Variables dépendantes : *math10*

Variables indépendantes	(1)	(2)
log(*expend*)	11,13 (3,30)	7,75 (3,04)
log(*enroll*)	0,022 (0,615)	−1,26 (0,58)

Variables indépendantes	(1)	(2)
lnchprg	–	−0,324 (0,036)
ordonnée à l'origine	269,24 (26,72)	223,14 (24,99)
Observations *R* carré	428 0,0297	428 0,1893

Expliquez pourquoi l'effet des dépenses par étudiant sur *math10* est plus faible dans la colonne (2) que dans la colonne (1). L'effet dans la colonne (2) est-il statistiquement plus grand que zéro ?

iii. Les taux de réussites sont-ils plus faibles dans les écoles de plus grande taille, toutes choses étant égales par ailleurs ? Expliquez.

iv. Interprétez le coefficient de *lnchprg* dans la colonne (2).

v. De quelle manière interprétez-vous l'augmentation substantielle du *R* carré lors du passage de la colonne (1) à la colonne (2) ?

4. L'équation suivante explique le nombre d'heures passées par semaine devant la télévision par un enfant, en fonction de son âge, du niveau d'instruction de sa mère, du niveau d'instruction de son père, et du nombre de frères et sœurs :

$$tvhours^* = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 motheduc + \beta_4 fatheduc + \beta_5 sibs + u.$$

Nous nous inquiétons du fait que *tvhours** soit sujet à des erreurs de mesure dans notre sondage. Soit *tvhours*, le nombre d'heures de visionnage déclarées par semaine.

i. Que requièrent les hypothèses d'erreur classique dans les variables (ECV) dans le cadre de ce modèle ?

ii. Les hypothèses d'ECV sont-elles respectées ? Expliquez.

5. Dans l'exemple 4.4., nous avons estimé un modèle associant le nombre de délits commis sur un campus au nombre d'étudiants inscrits à l'université. L'échantillon porte sur l'année 1992 et contient plusieurs universités ; malheureusement, il n'est pas aléatoire, puisque ces données n'étaient pas disponibles pour toutes les universités américaines. Pensez-vous qu'il s'agisse malgré tout d'un échantillonnage exogène ? Expliquez.

6. Dans le modèle (9.17), démontrez que les MCO produisent des estimateurs convergents de α et β à condition que a_i ne soit pas corrélée avec x_i et que b_i ne soit pas corrélée avec x_i et x_i^2. Il s'agit d'hypothèses moins restrictives que pour (9.19). [*Astuce* : écrivez l'équation sous la forme de (9.18) et rappelez-vous que les conditions suffisantes pour la convergence de la pente et de l'ordonnée à l'origine des MCO sont $E(u_i) = 0$ et $Cov(x_i, u_i) = 0$, comme indiqué au chapitre 5.]

7. Soit $y = \beta_0 + \beta_1 x^* + u$, un modèle de régression simple avec erreur de mesure classique dans les variables (ECV), pour lequel nous disposons de m mesures de x^*. Écrivez $z_{ih} = x^* + e_{ih}$, $h = 1..., m$. Supposez que x^* ne soit pas corrélée à u, e_1, ..., e_m, que les erreurs de mesure ne soient pas corrélées deux à deux, et que la variance de ces erreurs soit constante et égale à σ_e^2. Soit $w = (z_1 + ... + z_m) / m$, la moyenne des mesures de x^*, si bien que $w_i = (z_{i1} + ... + z_{im})/m$ représente la moyenne des m mesures pour chaque observation i. Soit $\tilde{\beta}_1$, l'estimateur des MCO de la régression simple de y_i sur 1, w_i, $i = 1, ..., n$, avec échantillonnage aléatoire de données.

i. Montrez que

$$\text{plim}(\bar{\beta}_1) = \beta_1 \left\{ \frac{\sigma_x^{2*}}{[\sigma_x^{2*} + (\sigma_e^2 / m)]} \right\}.$$

$$[\textit{Astuce}: \text{plim}(\bar{\beta}_1) = \text{Cov}(w, y) / \text{Var}(w).]$$

ii. Comparez cette grandeur à la quantité vers laquelle $\bar{\beta}_1$ tend en probabilité lorsqu'une seule mesure est disponible (c.à.d. $m = 1$). Que se passe-t-il lorsque m augmente ? Commentez.

8. L'objectif de cet exercice est de montrer que les tests de formes fonctionnelles ne sont pas fiables lorsqu'il s'agit d'identifier un problème de variable omise. Soit un modèle linéaire de y sur x_1 et x_2, qui, conditionnellement à ces variables explicatives, respecte les hypothèses de Gauss-Markov :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

$$\text{E}(u|x_1, x_2) = 0$$

$$\text{Var}(u|x_1, x_2) = \sigma^2.$$

Afin de rendre le problème intéressant, supposez que $\beta_2 \neq 0$.

Supposez également que x_1 et x_2 soient liées par la relation linéaire simple :

$$x_2 = \delta_0 + \delta_1 x_1 + r$$

$$\text{E}(r|x_1) = 0$$

$$\text{Var}(r|x_1) = \tau^2$$

i. Montrez que

$$\text{E}(y|x_1) = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1.$$

Sous l'hypothèse d'échantillonnage aléatoire, quelle est la limite en probabilité de l'estimateur des MCO de la régression simple de y contre x_1 ? Dans cette régression simple, l'estimateur de β_1 est-il, en général, convergent ? Autrement dit, tend-il en probabilité vers β_1 ?

ii. Si vous effectuez la régression de y sur x_1 et x_1^2, quelle est la limite en probabilité de l'estimateur des MCO du coefficient associé à x_1^2 ? Expliquez.

iii. En utilisant la technique de substitution, montrez qu'il est possible d'écrire

$$y = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1 + u + \beta_2 r.$$

Si nous définissons $v = u + \beta_2 r$, nous pouvons démontrer que $\text{E}(v|x_1) = 0, \text{Var}(v|x_1) = \sigma^2 + \beta_2^2\tau^2$. Que cela implique-t-il concernant la statistique t associée à x_1^2 dans la régression décrite à la question (ii) ?

iv. Quelle conclusion tirez-vous de l'ajout d'une fonction non linéaire de x_1 (la forme quadratique x_1^2, en particulier) dans le but de détecter l'omission de x_2 ?

9. Supposez que $\log(y)$ suive un modèle linéaire avec une forme linéaire d'hétéroscédasticité. Soit

$$\log(y) = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u$$

$$u|x \sim \text{Normal}(0, h(\mathbf{x}))$$

si bien que, conditionnellement à $\mathbf{x}$, u suit une distribution normale de moyenne (et médiane) nulle ; néanmoins, la variance $h(\mathbf{x})$ dépend de $\mathbf{x}$. Puisque $\text{Med}(u|\mathbf{x}) = 0$, l'équation (9.48) est respectée : $\text{Med}(y|x) = \exp(\beta_0 + \mathbf{x}\boldsymbol{\beta})$. De plus, en approfondissant un résultat du chapitre 6, nous pouvons démontrer que

$$\mathrm{E}(y|x) = \exp[\beta_0 + \mathbf{x}\boldsymbol{\beta} + h(\mathbf{x}) / 2].$$

i. Étant donné que $h(\mathbf{x})$ peut correspondre à n'importe quelle fonction positive, est-il possible de conclure que $\partial \mathrm{E}(y|x) / \partial x_j$ est du même signe que β_j ?

ii. Soit $h(\mathbf{x}) = \delta_0 + \mathbf{x}\boldsymbol{\beta}$. Ignorez le problème posé par le fait que les fonctions linéaires ne sont pas nécessairement toujours positives. Montrez qu'une variable particulière, par exemple x_1, peut avoir un effet *négatif* sur Med($y|\mathbf{x}$) et un impact *positif* sur E($y|\mathbf{x}$).

iii. Considérez le cas de la section 6.4, où $h(\mathbf{x}) = \sigma^2$. Comment pouvez-vous prédire la variable y en utilisant une estimation de E($y|\mathbf{x}$) ? Comment pouvez-vous prédire la variable y en utilisant une estimation de Med($y|\mathbf{x}$) ? Quelle prédiction est toujours la plus grande ?

10. Cet exercice montre que dans un modèle de régression simple, ajouter une variable binaire pour les données manquantes produit un estimateur consistent du coefficient de pente si la caractéristique « être manquante » ou non pour une donnée n'est liée ni aux facteurs observables, ni aux facteurs inobservables affectant y. Soit m, une variable telle que $m = 1$ si on n'observe pas x et $m = 0$ si on observe x. On suppose qu'y est toujours observé. Le modèle de population est

$$y = \beta_0 + \beta_1 x + u$$
$$E(u|x) = 0.$$

i. Donnez une interprétation de l'hypothèse plus forte

$$E(u|x, m) = 0$$

En particulier, quel schéma de données manquantes rend cette hypothèse non valide ?

ii. Montrez qu'on peut toujours écrire

$$y = \beta_0 + \beta_1 (1 - m)x + \beta_1 mx + u$$

iii. Soit $\{(x_i, y_i, m_i) : i = 1, \ldots, n\}$, des tirages aléatoires de la population, où x_i est manquante quand $m = 1$. Expliquez la nature de la variable $z_i = (1 - m_i)x_i$. En particulier, que vaut cette variable quand x_i est manquante ?

iv. Soit $\rho = P(m = 1)$ et supposons que m et x sont indépendantes. Montrez que

$$Cov[(1 - m)x, mx] = -\rho(1 - \rho)\mu_x,$$

où $\mu_x = E(x)$. Quelles sont les implications sur l'estimation de β_1 de la régression de y_i contre z_i, $i = 1, \ldots, n$?

v. Si m et x sont indépendantes, on peut montrer que

$$mx = \delta_0 + \delta_1 m + v,$$

où v est non corrélée avec m et $z = (1 - m)x$. Expliquez pourquoi m constitue grâce à cette expression une variable de substitution adéquate pour mx. Que cela signifie-t-il pour le coefficient de z dans la régression

$$y_i \text{ contre } z_i, m_i, \ i = 1, \ldots, n \ ?$$

Soit une population d'enfants, y est le résultat d'un test standardisé obtenu à partir des informations d'écoles, et x est le revenu des familles correspondantes. Ce revenu est donné sur base volontaire par chaque famille (et donc des familles ne délivrent pas leurs revenus). Est-il réaliste de penser que m et x sont indépendantes ? Expliquez.

EXERCICES SUR ORDINATEUR

C1. i. Appliquez le RESET de l'équation (9.3) au modèle estimé dans l'exercice sur ordinateur C5 du chapitre 7. Peut-on soupçonner une erreur de spécification de la forme fonctionnelle dans cette équation ?

ii. Effectuez un RESET robuste à l'hétéroscédasticité. Cela modifie-t-il la réponse apportée à la question (i) ?

C2. Utilisez les données contenues dans WAGE2 pour cet exercice.

i. Utilisez la variable *KWW* (correspondant au score réalisé à un test de « connaissance du monde du travail ») comme variable de substitution pour *IQ* dans l'exemple 9.3. Quel est le rendement estimé du niveau d'instruction dans ce cas-là ?

ii. Utilisez à présent *IQ* et *KWW* conjointement comme variables de substitution. Qu'arrive-t-il à l'estimation du rendement du niveau d'instruction ?

iii. Dans la question (ii), *IQ* et *KWW* sont-ils individuellement significatifs ? Sont-ils conjointement significatifs ?

C3. Utilisez les données contenues dans JTRAIN pour cet exercice.

i. Considérez le modèle de régression simple

$$\log(scrap) = \beta_0 + \beta_1 grant + u,$$

où *scrap* désigne le taux de rebut de l'entreprise et *grant* est une variable binaire indiquant si une entreprise a bénéficié d'une subvention à la formation professionnelle. Pouvez-vous identifier des raisons qui justifieraient une corrélation non nulle entre la variable *grant* et les facteurs non observés contenus dans u ?

ii. Estimez la régression simple en utilisant les données de 1988 (vous devriez avoir 54 observations). L'octroi d'une subvention à la formation professionnelle diminue-t-il significativement le taux de rebut d'une entreprise ?

iii. Ajoutez maintenant $\log(scrap_{87})$ comme variable explicative. De quelle manière cela modifie-t-il l'effet estimé de *grant* ? Interprétez le coefficient de *grant*. Est-il statistiquement significatif à un seuil de 5 %, considérant l'hypothèse alternative unilatérale $H_1 : \beta_{grant} < 0$?

iv. Testez l'hypothèse nulle selon laquelle le paramètre de $\log(scrap_{87})$ est égal à 1 contre une hypothèse alternative bilatérale. Donnez la *p*-valeur associée à ce test.

v. Effectuez à nouveau les questions (iii) et (iv) en utilisant les écarts-types estimés et robustes à l'hétéroscédasticité. Discutez brièvement des différences notables.

C4. Utilisez les données relatives à l'année 1990 contenues dans INFMRT pour cet exercice.

i. Estimez à nouveau l'équation (9.43) en incluant à présent une variable binaire pour l'observation relative au District de Columbia (appelée *DC*). Interprétez le coefficient de *DC* ; commentez sa taille et sa significativité.

ii. Comparez les estimations et les écarts-types estimés de la question (i) à ceux de l'équation (9.44). Que pensez-vous de l'intérêt d'inclure une variable binaire pour une seule observation ?

C5. Utilisez les données contenues dans RDCHEM afin d'examiner plus en détail les effets des observations aberrantes sur les estimations obtenues par MCO et par MDA. L'objectif est d'évaluer la sensibilité de ces méthodes à la présence d'observations aberrantes. Le modèle est

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \beta_3 profmarg + u,$$

dans lequel il est préférable d'exprimer la variable *sales* en milliards de dollars afin de rendre l'interprétation des estimations plus facile.

i. Estimez cette équation en utilisant les MCO, avec et sans prise en compte de l'entreprise réalisant un chiffre d'affaires (*sales*) de près de 40 milliards. Commentez les différences entre les coefficients estimés.

ii. Estimez la même équation en utilisant les MDA (de nouveau, avec et sans prise en compte de l'entreprise réalisant un chiffre d'affaires de près de 40 milliards). Commentez toute différence importante entre les coefficients estimés.

iii. En vous basant sur les résultats précédents, laquelle de ces deux méthodes (MCO ou MDA) est la plus résistante aux observations aberrantes ?

C6. Effectuez à nouveau l'exemple 4.10 en excluant de l'échantillon les écoles dans lesquelles les autres avantages financiers des enseignants représentent moins de 1 % de leur salaire.

i. Combien perdez-vous d'observations ?

ii. L'exclusion de ces observations conduit-elle à modifier l'estimation du compromis entre salaire et pension ?

C7. Utilisez les données contenues dans LOANAPP pour cet exercice.

i. Dans combien de cas les autres engagements de remboursement représentent-ils plus de 40 % du revenu total ? Autrement dit, quelle est la valeur de n lorsque $obrat > 40$?

ii. Estimez à nouveau le modèle de la question (iii) de l'exercice sur ordinateur C8 du chapitre 7, en veillant à exclure de l'échantillon les observations pour lesquelles $obrat > 40$. Que deviennent l'estimation et la statistique t de la variable *white* ?

iii. L'estimation de β_{white} est-elle trop sensible à l'échantillon utilisé ?

C8. Utilisez les données contenues dans TWOYEAR pour cet exercice.

i. La variable *stotal* mesure la performance réalisée à un test standardisé ; elle pourrait être utilisée comme variable de substitution, étant donné que les capacités (innées) d'un individu ne sont pas observées. Trouvez la moyenne et l'écart-type de cette variable pour l'échantillon considéré.

ii. Effectuez des régressions simples de *jc* sur *stotal*, puis de *univ* sur *stotal*. Ces deux variables dépendantes, caractérisant la longueur des études universitaires suivies, sont-elles statistiquement liées à *stotal* ? Expliquez.

iii. Ajoutez *stotal* à l'équation (4.17). Testez l'hypothèse selon laquelle les rendements des cursus universitaires de deux ans (*jc*) et de quatre ans (*univ*) sont les mêmes, contre l'hypothèse alternative unilatérale que le rendement des cursus de 4 ans est supérieur ? Comparez vos résultats à ceux de la section 4.4.

iv. Ajoutez $stotal^2$ à l'équation estimée dans la question (iii). L'inclusion de ce terme quadratique semble-t-elle nécessaire ?

v. Ajoutez les termes d'interaction *stotal·jc* et *stotal·univ* à l'équation de la question (iii). Ces termes sont-ils conjointement significatifs ?

vi. Quel modèle final choisissez-vous pour tenir compte de l'influence des capacités via l'utilisation de *stotal* ? Justifiez votre réponse.

C9. Dans cet exercice, vous devez comparer les estimations obtenues par MCO à celles obtenues par MDA. Il s'agit d'estimer les effets sur le patrimoine net (ou actifs financiers nets) de la participation au plan 401(k), qui est un système d'épargne retraite par capitalisation aux États-Unis. Le modèle est

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 male + \beta_6 e401k + u.$$

i. Utilisez les données contenues dans 401KSUBS afin d'estimer l'équation par MCO et présentez vos résultats sous leur forme habituelle. Interprétez le coefficient de la variable *e401k*.

ii. Utilisez les résidus obtenus par MCO afin de tester la présence d'hétéroscédasticité via le test de Breusch-Pagan. Le terme u est-il indépendant des variables explicatives ?

iii. Estimez l'équation par MDA et présentez les résultats en utilisant le même format que pour la méthode des MCO. Interprétez l'estimation MDA de β_6.

iv. Réconciliez les résultats obtenus dans les questions (i) et (iii).

C10. L'utilisation de deux bases de données, JTRAIN2 et JTRAIN3, est nécessaire pour cet exercice. La première contient les résultats obtenus à une expérience de formation professionnelle. La seconde contient des données observationnelles (non-expérimentales) sur des individus qui ont choisi eux-mêmes de participer (ou non) à une formation professionnelle. Les deux bases de données couvrent la même période.

i. Dans la base de données JTRAIN2, quelle est la proportion d'individus qui ont reçu une formation professionnelle ? Qu'en est-il dans JTRAIN3 ? Comment expliquez-vous cette différence ?

ii. En utilisant les données contenues dans JTRAIN2, effectuez une régression simple de *re78* contre *train*. Quel est l'effet estimé de la participation à une formation professionnelle sur les revenus réels ?

iii. Dans la régression de la question (ii), ajoutez les variables *re74*, *re75*, *educ*, *age*, *black*, et *hisp*. L'effet estimé de la formation professionnelle sur *re78* change-t-il beaucoup ? Comment l'expliquez-vous ? (*Astuce* : il s'agit de données expérimentales.)

iv. Effectuez à nouveau les régressions des questions (ii) et (iii) en utilisant les données contenues dans JTRAIN3. Indiquez les coefficients estimés associés à la variable *train*, ainsi que leurs statistiques *t*. Quel est l'effet lié à la prise en compte de ces facteurs supplémentaires ? Pourquoi ?

v. Soit $avgre = (re74 + re75)/2$. Trouvez les moyennes, écarts-types, ainsi que les valeurs minimales et maximales de cette grandeur dans les deux bases de données. Ces bases de données sont-elles représentatives des mêmes populations pour l'année 1978 ?

vi. Dans la base de données JTRAIN2, 96 % des hommes ont un revenu moyen (*avgre*) inférieur à 10 000 dollars. Effectuez la régression de *re78* sur *train*, *re74*, *re75*, *educ*, *age*, *black*, *hisp*, en ne vous servant que de ces observations. Indiquez l'estimation de l'effet de la formation professionnelle, ainsi que sa statistique *t*. Effectuez la même régression en utilisant les données contenues dans JTRAIN3, en utilisant uniquement les hommes dont $avgre \leq 10$. Pour ce sous-échantillon de travailleurs à bas revenus, comparez les effets estimés de la formation dans les deux bases de données (expérimentale et non-expérimentale).

vii. Utilisez à présent chaque base de données afin d'effectuer une régression simple de *re78* sur *train* pour les hommes qui étaient sans emploi durant les années 1974 et 1975. Comparez à nouveau les effets estimés dans les deux bases de données.

viii. Sur la base de vos résultats obtenus aux points précédents, jugez-vous important de disposer de populations similaires lorsqu'il s'agit de comparer les estimations expérimentales et non-expérimentales ?

C11. Utilisez les données contenues dans MURDER uniquement pour l'année 1993. Notez d'ores et déjà qu'il vous sera nécessaire d'utiliser une variable retardée pour le taux d'homicides, soit $mrdrte_{-1}$.

i. Régressez *mrdrte* sur *exec*, *unem*. Quelles sont les valeurs du coefficient de *exec* et de sa statistique *t* ? Sur base de cette régression, existe-t-il un effet dissuasif lié à une condamnation à la peine capitale ?

ii. Combien d'exécutions sont répertoriées pour l'État du Texas ? (Il s'agit en fait de la somme des exécutions enregistrées pour l'année en cours, soit 1993, et les deux années précédentes). Comparez ce chiffre au nombre d'exécutions qui ont eu lieu dans les autres États. Dans la régression de la question (i), ajoutez une variable binaire pour l'État du Texas. Observez-vous une statistique t particulièrement élevée pour cette variable ? Sur base de ces résultats, peut-on considérer le Texas comme une observation aberrante ?

iii. Ajoutez le taux d'homicides retardé à la régression de la question (i). Que devient $\hat{\beta}_{exec}$? Qu'en est-il sur le plan statistique ?

iv. Dans la régression du point précédent, peut-on considérer l'état du Texas comme une observation aberrante ? Que devient $\hat{\beta}_{exec}$ si l'état du Texas est exclu de la régression ?

C12. Utilisez la base de données ELEM94_95 pour réaliser cet exercice. Consultez également l'exercice sur ordinateur C10 du chapitre 4.

i. En utilisant toutes les données, régressez *lavgsal* sur *bs*, *lenrol*, *lstaff*, et *lunch*. Indiquez le coefficient associé à *bs*, ainsi que les écarts-types estimés (classiques et robustes à l'hétéroscédasticité). Que concluez-vous quant à l'importance de $\hat{\beta}_{bs}$ sur les plan économique et statistique ?

ii. Abandonnez maintenant les quatre observations pour lesquelles $bs > 0{,}5$. Autrement dit, les avantages extra-légaux doivent représenter plus de 50 % du salaire (en moyenne au sein de l'établissement). Quel est le coefficient associé à *bs* ? Est-il statistiquement significatif, en considérant l'écart-type estimé robuste à l'hétéroscédasticité ?

iii. Vérifiez que les quatre observations pour lesquelles $bs > 0{,}5$ sont égales à 68, 1 127, 1 508, et 1 670. Définissez quatre variables binaires, une pour chacune de ces observations. (Vous pouvez les appeler *d*68, *d*1127, *d*1508, et *d*1670, par exemple.) Incluez ces variables binaires dans la régression de la question (i) et vérifiez que les coefficients, ainsi que les écarts-types estimés obtenus par les MCO pour les autres variables, sont identiques à ceux de la question (ii). Laquelle de ces variables binaires est significative sur le plan statistique à un seuil de 5 % ?

iv. Dans l'équation du point (iii), vérifiez que l'observation dont le résidu de Student est le plus élevé (ce qui correspond à la variable binaire dont la statistique t est la plus élevée) a effectivement une influence importante sur les estimations obtenues par les MCO. (Cela revient à effectuer la régression par les MCO en utilisant toutes les observations, sauf celle avec le plus grand résidu de Student). Observez-vous un effet important lié à l'exclusion successive des autres observations pour lesquelles $bs > 0{,}5$?

v. Quelle conclusion tirez-vous de la sensibilité des MCO par rapport à une observation bien spécifique, même lorsque l'échantillon est de grande taille ?

vi. Vérifiez que l'estimateur des MDA n'est pas sensible à l'inclusion de l'observation identifiée au point (iii).

C13. Utilisez la base de données CEOSAL2 afin de réaliser cet exercice.

i. Estimez le modèle

$$lsalary = \beta_0 + \beta_1 lsales + \beta_2 lmktval + \beta_3 ceoten + \beta_4 ceoten^2 + u$$

avec la méthode des MCO en utilisant toutes les observations pour lesquelles les variables *lsalary*, *lsales*, et *lmktval* sont toutes en logarithmes naturels. Présentez vos résultats sous leur forme habituelle en incluant les écarts-types estimés par les MCO. (Vous pouvez vérifier que les écarts-types estimés robustes à l'hétéroscédasticité sont similaires.)

ii. Calculez les résidus de Student pour la régression du point (i). Appelez ces derniers str_i. Combien d'entre eux affichent une valeur absolue supérieure à 1,96 ? Si les résidus de Student provenaient de tirages

indépendants réalisés à partir d'une distribution normale standard, combien d'entre eux devraient être supérieurs à 2 en valeur absolue (parmi les 177 tirages potentiels) ?

iii. Estimez à nouveau l'équation de la question (i) en utilisant la méthode des MCO, uniquement pour les observations où $|str_i| \leq 1{,}96$. Comparez les coefficients obtenus à ceux du point (i).

iv. Estimez l'équation de la question (i) en utilisant la méthode des MDA sur l'échantillon complet. L'estimation de β_1 par les MDA est-elle plus proche de celle des MCO lorsque l'échantillon est complet ou restreint ? Qu'en est-il de β_3 ?

v. Évaluez l'affirmation suivante. « L'exclusion d'observations aberrantes, basée sur l'identification de résidus de Student extrêmes, rend plus proches les estimations obtenues, sur l'échantillon complet, par les MCO et les MDA. »

C14. Utilisez les données ECONMATH pour répondre à cette question. Le modèle de la population est

$$score = \beta_0 + \beta_1 act + u.$$

i. Pour combien d'étudiants la variable *act* est-elle manquante ? Quelle en est la fraction dans l'échantillon ? Définissez une nouvelle variable, *actmiss*, qui vaut un si *act* est manquante et zéro sinon.

ii. Créez une nouvelle variable, disons *act0*, qui vaut *act* quand *act* est fournie et zéro quand *act* est manquante. Trouvez la moyenne de *act0* et comparez-la avec la moyenne de *act*.

iii. Faites tourner une régression linéaire simple de *score* contre *act0* en utilisant uniquement les données complètes. Qu'obtenez-vous pour le coefficient de pente et son erreur type robuste à l'hétéroscédasticité ?

iv. Faites tourner une régression linéaire simple de *score* contre *act0* en utilisant toutes les données. Comparez le coefficient de pente avec celui de (iii) et commentez.

v. Ensuite, utilisez toutes les données et faites tourner la régression

$$score_i \text{ contre } act0_i, actmiss_i.$$

Quelle est l'estimation du coefficient de pente de *act0* ? Comment la compare-t-on avec les réponses aux parties (iii) et (iv) ?

vi. En comparant les régressions (iii) et (v), est-ce qu'utiliser toutes les données et ajouter l'estimateur pour les données manquantes améliorent l'estimation de β_1 ?

vii. Si on ajoute la variable *colgpa* aux régressions dans les parties (iii) et (v), votre réponse à la partie (vi) change-t-elle ?

PARTIE 2

ANALYSE ÉCONOMÉTRIQUE DES SÉRIES TEMPORELLES

Maintenant que nous avons des bases solides en ce qui concerne l'utilisation d'un modèle de régression multiple avec des données en coupe transversale, nous allons nous tourner vers l'analyse économétrique des séries temporelles. La mécanique et l'inférence statistique de la méthode des moindres carrés ordinaires (MCO) ont été vues dans les précédents chapitres, mais comme nous l'avons noté dans le chapitre 1, les séries temporelles possèdent certaines caractéristiques que les données en coupe transversale n'ont pas, et cela peut exiger une attention particulière lors de l'application des MCO.

Le chapitre 10 porte sur l'analyse économétrique de régression simple, et accorde une attention particulière aux problèmes spécifiques à l'analyse des séries temporelles. Le théorème de Gauss-Markov et les hypothèses du modèle linéaire classique pour les séries temporelles seront alors clairement explicités. Les problèmes de forme fonctionnelle, de variables indicatrices, de tendances et de saisonnalité seront également discutés.

Étant donné que certaines séries temporelles violent nécessairement les hypothèses de Gauss-Markov, le chapitre 11 décrit la nature de ces problèmes et présente les propriétés asymptotiques de l'estimateur des MCO. Comme nous ne pouvons plus supposer un échantillonnage aléatoire, il est important de couvrir les conditions qui restreignent la corrélation dans le temps d'une série temporelle, afin de s'assurer que l'analyse asymptotique classique est valide.

Le chapitre 12 présente un autre problème important, celui de l'autocorrélation des termes d'erreur dans l'analyse de régression des séries temporelles. Nous discutons des conséquences, des tests et des méthodes pour faire face à l'autocorrélation. Le chapitre 12 contient également une explication de la façon dont l'hétéroscédasticité peut apparaître dans les modèles de séries temporelles.

CHAPITRE

10

ANALYSE ÉCONOMÉTRIQUE SIMPLE DES SÉRIES TEMPORELLES

Traduction de Alain Durré

Dans ce chapitre, nous commençons par étudier les propriétés des estimateurs des moindres carrés ordinaires (MCO) pour l'analyse économétrique des séries temporelles. Dans la section 10.1, nous discutons de certaines différences conceptuelles entre les données de séries temporelles et les données en coupe transversale. La section 10.2 donne quelques exemples de régressions de séries temporelles souvent utilisées en sciences sociales empiriques. Nous étudions ensuite les propriétés en échantillons finis des estimateurs MCO, en indiquant les hypothèses de Gauss-Markov et les hypothèses du modèle linéaire classique pour les régressions de séries temporelles. Bien que ces hypothèses aient des caractéristiques communes avec les hypothèses concernant l'analyse des données en coupe transversale, il existe aussi quelques différences importantes que nous aurons besoin de mettre en évidence.

De plus, nous reviendrons sur certaines questions que nous avons traitées dans le cadre de l'analyse économétrique des données en coupe transversale, telle que la façon d'utiliser et d'interpréter la forme fonctionnelle logarithmique et les variables indicatrices. L'intégration des tendances et la prise en compte des variations saisonnières dans la régression multiple seront détaillées dans la section 10.5.

10.1 LA NATURE DES SÉRIES TEMPORELLES

Une caractéristique évidente des données de séries temporelles qui les distingue des données en coupe transversale est la dimension temporelle. Par exemple, dans le chapitre 1, nous avons brièvement discuté des données de la série sur l'emploi, le salaire minimum, et d'autres variables économiques à Porto Rico. Dans cet ensemble de données, il est évident que les données de l'année 1970 précèdent immédiatement les données de l'année 1971. Pour l'analyse des données de séries temporelles en sciences sociales, nous devons reconnaître que le passé peut affecter l'avenir, mais pas l'inverse (contrairement à l'univers de la série Star Trek). Pour souligner l'ordre des données de séries temporelles, le tableau 10.1 présente une liste de données à propos du taux d'inflation et du taux chômage aux États-Unis, à partir de différentes éditions du *Rapport économique du Président* (tableaux B-42 et B-64).

Il existe une autre différence plus subtile entre les données en coupe transversale et les séries temporelles. Dans les chapitres 3 et 4, nous avons étudié les propriétés statistiques des estimateurs des MCO, en se basant sur le fait que les échantillons ont été tirés au hasard parmi la population appropriée. Comprendre pourquoi les données en coupe transversale doivent être considérées comme des tirages aléatoires est assez simple : un autre échantillon tiré à partir de la population donnera généralement des valeurs différentes des variables indépendantes et dépendantes (telles que l'éducation, l'expérience, le salaire, et ainsi de suite). Par conséquent, les estimateurs des MCO calculés à partir de différents échantillons aléatoires sont en général différents, et c'est pourquoi nous considérons les estimateurs des MCO comme des variables aléatoires.

Tableau 10.1 *Liste partielle des données sur l'inflation et le taux de chômage aux États-Unis, 1948-2003*

Année	Taux d'inflation	Taux de chômage
1948	8,1	3,8
1949	– 1,2	5,9
1950	1,3	5,3
1951	7,9	3,3
.	.	.
.	.	.

Année	Taux d'inflation	Taux de chômage
1998	1,6	4,5
1999	2,2	4,2
2000	3,4	4,0
2001	2,8	4,7
2002	1,6	5,8
2003	2,3	6,0

Comment devons-nous alors considérer le caractère aléatoire des données de séries temporelles ? Intuitivement, les séries temporelles économiques semblent répondre à cette exigence, en étant les résultats de variables aléatoires. Par exemple, aujourd'hui, nous ne savons pas la valeur que prendra l'indice Dow Jones Industrial Average demain à la fermeture du marché. Nous ne savons pas non plus quelle sera la croissance annuelle de la production au Canada l'année prochaine. Étant donné que les résultats de ces variables ne sont pas connus d'avance, ces variables peuvent être vues comme des variables aléatoires.

Formellement, une suite de variables aléatoires indexées par le temps est appelé **processus stochastique** (« stochastique » est un synonyme d'aléatoire) ou processus de séries chronologiques. Lorsque nous collectons un échantillon de données temporelles, nous obtenons un résultat possible, ou *réalisation*, d'un processus stochastique. Cette réalisation est unique, car nous ne pouvons pas revenir en arrière et recommencer le processus à nouveau (ceci est analogue à l'analyse en coupe **transversale** où il n'est possible de recueillir qu'un seul échantillon aléatoire). Toutefois, si certaines conditions dans l'histoire avaient été différentes, nous aurions généralement obtenu une réalisation différente de ce processus stochastique, et c'est pourquoi nous considérons les séries temporelles comme le résultat de variables aléatoires. L'ensemble de toutes les réalisations possibles d'un processus temporel joue le rôle de la population dans l'analyse en coupe transversale. La taille de l'échantillon d'une série temporelle est le nombre de périodes de temps pendant laquelle nous observons les variables de notre modèle.

10.2 EXEMPLES DE RÉGRESSION DE SÉRIES TEMPORELLES

Dans cette section, nous allons discuter de deux modèles de séries temporelles, utilisés d'un point de vue empirique, et qui sont facilement estimables par la méthode des moindres carrés ordinaires. Nous étudierons ensuite d'autres modèles dans le chapitre 11.

Les modèles statiques

Supposons que nous ayons les séries temporelles concernant deux variables, y et z, où y_t et z_t sont indexées par le temps de manière contemporaine. Un modèle statique expliquant y en fonction de z peut s'écrire sous la forme :

$$y_t = \beta_0 + \beta_1 z_t + u_t, \; t = 1, 2, \ldots, n. \qquad \textbf{[10.1]}$$

Le nom de « modèle statique » vient du fait que nous modélisons une relation contemporaine entre y et z. Habituellement, un modèle statique est utilisé quand un changement de z à l'instant t est censé avoir un effet immédiat sur $\Delta y_t = \beta_1 \Delta z_t$, avec $\Delta u_t = 0$. Le modèle de régression statique est également utilisé lorsque nous sommes intéressés par le rapport de force (ou l'arbitrage) entre y et z.

Un exemple de modèle statique est *la courbe de Phillips statique*, donnée par

$$inf_t = \beta_0 + \beta_1 unem_t + u_t. \qquad [10.2]$$

où inf_t est le taux d'inflation annuel et $unem_t$ est le taux de chômage. Cette forme de la courbe de Phillips suppose que le *taux de chômage naturel* est constant et que les anticipations d'inflation sont constantes, et peut être utilisé pour étudier l'arbitrage contemporain entre le taux d'inflation et le taux chômage. [Voir, par exemple, Mankiw (1994, section 11.2).]

Naturellement, nous pouvons avoir plusieurs variables explicatives dans un modèle de régression statique. Notons $mrdrte_t$ le nombre d'homicides pour 10 000 habitants dans une ville donnée lors de l'année t, $convrte_t$ le taux de condamnation pour meurtre, $unem_t$ le taux de chômage et $yngmle_t$ la fraction de la population d'hommes âgés entre 18 et 25 ans dans cette même ville. Dans cette situation, un modèle statique de régression multiple ayant pour objectif d'expliquer le nombre d'homicides pour 10 000 habitants peut s'écrire :

$$mrdrte_t = \beta_0 + \beta_1 convrte_t + \beta_2 unem_t + \beta_3 yngmle_t + u_t. \qquad [10.3]$$

En utilisant ce modèle, nous pouvons estimer, par exemple, l'effet de l'augmentation du taux de condamnation *convrte* sur le nombre d'homicides pour 10 000 habitants toutes choses égales par ailleurs.

Modèle à retards échelonnés finis

Dans un **modèle à retards échelonnés finis**, nous considérons le fait qu'une ou plusieurs variables explicatives puissent affecter y avec un retard. Par exemple, avec des observations annuelles, considérons le modèle :

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + u_t, \qquad [10.4]$$

où gfr_t est le taux global de fécondité (nombre d'enfants nés pour 1 000 femmes en âge de procréer) et pe_t est la valeur en dollars réel des exemptions d'impôts personnels (ou réduction d'impôts en fonction du nombre d'enfants). L'idée de ce modèle est de voir si, d'un point de vue agrégé, la décision d'avoir des enfants est liée aux valeurs passées des déductions fiscales associées. L'équation (10.4) reconnaît que, pour des raisons biologiques et comportementales, la décision d'avoir des enfants peut dépendre des valeurs passées du montant des allocations familiales.

L'équation (10.4) est un exemple du modèle

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t. \qquad [10.5]$$

Ce modèle est ce que l'on appelle un modèle à retards échelonnés finis *d'ordre 2*. Pour interpréter les coefficients de l'équation (10.5) supposons que z soit constant et égal à c pour toutes les périodes avant t. À la période t, z augmente d'une unité ($z = c + 1$), puis retrouve sa valeur initiale à la période $t + 1$ ($z = c$) (l'augmentation de z est donc temporaire). Plus précisément,

$$\ldots, z_{t-2} = c, z_{t-1} = c, z_t = c + 1, z_{t+1} = c, z_{t+2} = c, \ldots$$

Pour mettre l'accent sur l'effet *ceteris paribus* de z sur y, nous considérons que le terme d'erreur à chaque période est égal à zéro. Donc,

$$\begin{aligned}
y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \\
y_t &= \alpha_0 + \delta_0 (c + 1) + \delta_1 c + \delta_2 c, \\
y_{t+1} &= \alpha_0 + \delta_0 c + \delta_1 (c + 1) + \delta_2 c, \\
y_{t+2} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 (c + 1), \\
y_{t+3} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c,
\end{aligned}$$

et ainsi de suite. En considérant les deux premières équations, nous obtenons $y_t - y_{t-1} = \delta_0$, où δ_0 représente donc le changement immédiat de y suite à la hausse d'une unité de z à la période t. δ_0 est communément appelé le **multiplicateur de court-terme** ou l'**impact de court-terme**.

De la même manière, $\delta_1 = y_{t+1} - y_{t-1}$ correspond à la variation de y une période après le changement temporaire de z, et $\delta_2 = y_{t+2} - y_{t-1}$ est le changement de y deux périodes après le changement temporaire de z. À la période $t + 3$, y reprend sa valeur initiale : $y_{t+3} = y_{t-1}$, et ceci car nous avons supposé dans l'équation (10.5) que seuls deux retards de z pouvaient avoir un impact sur y. Lorsque nous représentons graphiquement δ_j en fonction de j, nous obtenons un graphique de la **distribution des retards**, qui résume l'effet dynamique qu'une augmentation temporaire de z a sur y. Une distribution possible des retards de notre modèle à retards échelonnés finis d'ordre 2 est représentée dans le tableau 10.1 (les paramètres δ_j n'étant pas connus, il est nécessaire d'estimer dans un premier temps δ_j avant de réaliser le graphique de la distribution des retards).

Le graphique de la distribution des retards de la figure 10.1 implique que l'effet le plus important a lieu lors du premier retard. En normalisant la valeur initiale de y à $y_{t-1} = 0$, la distribution des retards retrace toutes les valeurs futures de y suite à une augmentation temporaire de z d'une unité.

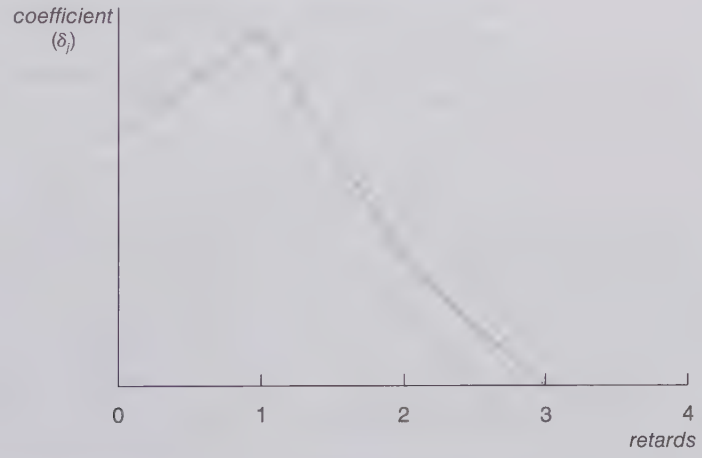


Figure 10.1 *Graphique de la distribution des retards, avec deux retards non nuls. L'effet maximum a lieu pour le premier retard.*

Nous pouvons aussi être intéressé par le changement de y dû à une augmentation *permanente* de z. Considérons alors qu'avant la période t, z est constant et égal à c. À la période t, z augmente de façon permanente d'une unité : $z_s = c$, $s < t$ et $z_s = c + 1$, $s \geq t$. De nouveau, en considérant le terme d'erreur comme nul nous avons :

$$\begin{aligned} y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \\ y_t &= \alpha_0 + \delta_0 (c + 1) + \delta_1 c + \delta_2 c, \\ y_{t+1} &= \alpha_0 + \delta_0 (c + 1) + \delta_1 (c + 1) + \delta_2 c, \\ y_{t+2} &= \alpha_0 + \delta_0 (c + 1) + \delta_1 (c + 1) + \delta_2 (c + 1), \end{aligned}$$

et ainsi de suite. À la suite de l'augmentation permanente de z, et après une période, y a augmenté de $\delta_0 + \delta_1$. Après deux périodes, y a augmenté de $\delta_0 + \delta_1 + \delta_2$. Après la seconde période, il n'y a alors plus d'augmentation de y. La somme des coefficients de z et de ses deux retards, $\delta_0 + \delta_1 + \delta_2$, représente la variation de *long terme* de y à la suite d'un changement permanent de z, et est appelé le **multiplicateur de long-terme** (aussi appelé LRP en anglais pour "long-run propensity") ou **l'impact de long terme**.

Si l'on reprend l'exemple de l'équation (10.4), δ_0 mesure le changement immédiat du taux de fécondité dû à une augmentation de 1 dollar de *pe*. Comme mentionné précédemment, pour des raisons biologiques et/ou comportementales, il y a des raisons de penser que δ_0 est petit, voire même égal à 0. Mais δ_1 et/ou δ_2 peuvent être positifs. Si *pe* augmente de manière permanente de 1 dollar, alors, après deux ans, *gfr* aura augmenté de $\delta_0 + \delta_1 + \delta_2$. Ce modèle suppose donc qu'il n'y a plus d'effet après la seconde année ; la validité de cette hypothèse est une question empirique.

Un modèle à retards échelonnés finis d'ordre q s'écrit :

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \ldots + \delta_q z_{t-q} + u_t. \qquad \textbf{[10.6]}$$

En définissant $\delta_1 = \delta_2 = \ldots = \delta_q = 0$, on retrouve d'ailleurs le modèle statique classique de l'équation (10.1) qui est un cas spécifique d'un modèle à retards échelonnés finis d'ordre q.

Parfois, l'objectif principal de l'estimation d'un modèle à retards échelonnés est de tester si z a un effet retardé sur y. Le multiplicateur d'impact est toujours le coefficient de z contemporain, δ_0. De temps en temps, on omet z_t dans l'équation (10.6), auquel cas le multiplicateur d'impact est égal à zéro.

Dans le cas général, la distribution des retards peut être tracée en représentant graphiquement δ_j (estimé) en fonction de j. Pour tout horizon h, nous pouvons définir l'effet cumulatif $\delta_0 + \delta_1 + \ldots + \delta_h$, qui correspond au changement attendu de y, h périodes après une augmentation permanente d'une unité de x. Une fois les δ_j estimés, on peut tracer les effets cumulatifs estimés en fonction de h. Le multiplicateur de long terme est l'effet cumulatif une fois que tous les changements ont eu lieu ; cela correspond donc simplement à la somme des coefficients des z_{t-j} :

$$\text{LRP} = \delta_0 + \delta_1 + \ldots + \delta_q. \qquad \textbf{[10.7]}$$

En raison de la corrélation souvent importante des différents retards de z – et ce à cause de la multicolinéarité dans (10.6) – il peut être difficile d'obtenir des estimations précises de chaque δ_j. Il est intéressant de souligner que, même si chaque δ_j ne peut pas être estimé avec précision, il est possible d'obtenir une bonne estimation du multiplicateur de long terme, ce que nous verrons ultérieurement dans un exemple.

Pour aller plus loin 10.1

Supposons l'équation suivante, avec des données annuelles :

$$int_t = 1{,}6 + 0{,}48\ inf_t - 0{,}15\ inf_{t-1} + 0{,}32\ inf_{t-2} + u_t,$$

où *int* est le taux d'intérêt et *inf* le taux d'inflation. Quelles sont les valeurs respectives des multiplicateurs de court et long-terme ?

Il est possible que plus d'une variable explicative avec retard soient présentes dans un modèle à retards échelonnés finis, ou bien que ce modèle contienne d'autres variables simultanées. Par exemple, le niveau moyen d'éducation pour les femmes en âge de procréer pourrait être ajouté à (10.4), ce qui nous permettrait de prendre en compte les effets du changement du niveau d'éducation des femmes sur le taux de fécondité.

Convention concernant les indices temporels

Lorsqu'un modèle contient des variables explicatives avec retards (et, comme nous le verrons dans le prochain chapitre, pour les modèles avec retards de la variable dépendante y), il existe un risque de confusion concernant le traitement des premières observations. Par exemple, si dans (10.5) nous supposons que l'équation est valide à partir de $t = 1$, alors les variables explicatives de la première période sont z_1, z_0, et z_{-1}. Par convention, nous définissons ces variables comme étant les valeurs initiales de notre échantillon, de telle façon que

nous commençons toujours notre indice temporel à $t = 1$. En pratique, cela n'a que peu d'importance car les logiciels permettant de réaliser des régressions tiennent automatiquement compte des observations disponibles pour l'estimation de modèles avec retards. Mais pour ce chapitre et les deux suivants, il est important de définir une convention concernant la première période de temps estimée dans nos équations de régressions.

10.3 PROPRIÉTÉS EN ÉCHANTILLON FINI DES MCO SOUS LES HYPOTHÈSES CLASSIQUES

Dans cette section, nous allons voir une liste complète des propriétés des MCO sous les hypothèses classiques, pour les échantillons finis ou de petite taille. Une attention particulière sera portée à la façon dont les hypothèses doivent être modifiées à partir de notre analyse en coupe transversale afin de couvrir les régressions de séries temporelles.

Absence de biais des estimateurs des MCO

La première hypothèse indique simplement qu'un processus de série temporelle suit un modèle dont les paramètres sont linéaires.

Hypothèse TS.1
Linéarité des paramètres

Le processus stochastique $\{(x_{t1}, x_{t2}, \ldots, x_{tk}, y_t) : t = 1, 2, \ldots, n\}$ suit un modèle linéaire

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \qquad [10.8]$$

où $\{u_t : t = 1, 2, \ldots, n\}$ correspond à la série des erreurs ou bruits. Ici, n est le nombre d'observations (périodes de temps).

Dans la notation x_{tj}, t correspond à la période de temps, et j est un indice correspondant à l'une des k variables explicatives. La terminologie utilisée pour les régressions en coupe transversale s'applique aussi ici : y_t est la variable dépendante (ou variable expliquée) ; les x_{tj} sont les variables indépendantes (ou variables explicatives).

Nous pouvons voir l'hypothèse TS.1 comme étant essentiellement la même que l'hypothèse MLR.1 (la première hypothèse vue en coupe transversale), mais nous spécifions désormais un modèle linéaire pour les données de séries temporelles. Les exemples couverts dans la section 10.2 peuvent être transformés sous la forme de (10.8) en définissant de façon appropriée les x_{tj}. Par exemple, l'équation (10.5) est obtenu en définissant $x_{t1} = z_t$, $x_{t2} = z_{t-1}$, et $x_{t3} = z_{t-2}$.

Pour énoncer et discuter certaines des hypothèses restantes, il est utile de définir $\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tk})$ comme l'ensemble des variables explicatives de l'équation à l'instant t. Définissons de plus $\mathbf{X}$, qui représente l'ensemble de toutes les variables explicatives pour toutes les périodes de temps. Il est utile de voir $\mathbf{X}$ comme étant un tableau à n lignes et k colonnes. Cela reflète la façon dont les données de séries temporelles sont stockées dans les logiciels économétriques : la $t^{\text{ième}}$ ligne de $\mathbf{X}$ est $\mathbf{x}_t$, qui est composée de toutes les variables explicatives pour la période de temps t. Par conséquent, la première ligne de $\mathbf{X}$ correspond à $t = 1$, la deuxième ligne à $t = 2$, et la dernière ligne à $t = n$. Un exemple est donné dans le tableau 10.2, en utilisant les variables explicatives de l'équation (10.3) et pour $n = 8$.

Tableau 10.2 *Exemple de X pour les variables explicatives de l'équation (10.3)*

t	*convrte*	*unem*	*yngmle*
1	0,46	0,074	0,12
2	0,42	0,071	0,12
3	0,42	0,063	0,11
4	0,47	0,062	0,09
5	0,48	0,060	0,10
6	0,50	0,059	0,11
7	0,55	0,058	0,12
8	0,56	0,059	0,13

Naturellement, comme avec les régressions en coupe transversale, nous devons écarter toute colinéarité parfaite entre les variables explicatives.

Hypothèse TS.2
Absence de colinéarité parfaite

Dans l'échantillon (et donc dans le processus temporel sous-jacent), aucune variable explicative n'est une constante ou une combinaison linéaire parfaite d'autres variables explicatives.

Nous avons discuté de cette hypothèse en détail dans le contexte des données en coupe transversale (voir chapitre 3). Les enjeux sont essentiellement les mêmes avec des données de séries temporelles. Rappelez-vous, l'hypothèse TS.2 permet que les variables explicatives soient corrélées entre elles, mais exclut la corrélation parfaite dans l'échantillon. La dernière hypothèse d'absence de biais des estimateurs des MCO est analogue à l'hypothèse MLR.4, et permet de se passer de l'hypothèse MLR.2 d'échantillonnage aléatoire.

Hypothèse TS.3
Espérance conditionnelle nulle

Pour chaque t, l'espérance mathématique du terme d'erreur u_t, compte tenu des variables explicatives pour toutes les périodes, est égale à zéro. Mathématiquement,

$$E(u_t|\mathbf{X}) = 0, t = 1, 2, \ldots, n. \qquad \textbf{[10.9]}$$

Il s'agit d'une hypothèse cruciale, et nous avons besoin d'avoir une compréhension intuitive de son sens. Comme dans le cas de données en coupe transversale, il est plus facile de voir cette hypothèse en termes d'absence de corrélation. L'hypothèse TS.3 implique que l'erreur à l'instant t, u_t, n'est pas corrélée avec les variables explicatives pour chaque période de temps. Le fait que cette hypothèse soit exprimée en termes d'espérance conditionnelle signifie que nous devons aussi spécifier correctement la relation fonctionnelle entre la variable y_t et les variables explicatives. Si u_t est indépendante de $\mathbf{X}$ et $E(u_t) = 0$, alors l'hypothèse TS.3 est automatiquement vérifiée.

Compte tenu de l'analyse en coupe transversale du chapitre 3, il n'est pas surprenant que nous exigions que u_t ne soit pas corrélé avec les variables explicatives de la période t. En moyenne conditionnelle, nous avons donc :

$$E(u_t|x_{t1}, \ldots, x_{tk}) = E(u_t|\mathbf{x}_t) = 0. \qquad \textbf{[10.10]}$$

Lorsque (10.10) est vérifié, nous disons que les x_{tj} sont **exogènes de manière contemporaine**. L'équation (10.10) implique ainsi que u_t et les variables explicatives sont non-corrélées de manière contemporaine : $\text{Corr}(x_{tj}, u_t) = 0$, pour tout j.

L'hypothèse TS.3 requiert plus que la simple exogénéité contemporaine : u_t ne doit pas être corrélé avec x_{sj}, même lorsque $s \neq t$. Lorsque les variables explicatives sont exogènes et que TS.3 est vérifié, on dit alors que les variables explicatives sont **strictement exogènes**. Dans le chapitre 11, nous démontrerons que (10.10) est suffisant pour prouver la convergence des estimateurs des MCO. Mais pour montrer que les estimateurs des MCO ne sont pas biaisés, nous avons besoin de l'hypothèse d'exogénéité stricte.

Dans le cas d'une étude en coupe transversale, nous n'avons pas explicitement montré comment le terme d'erreur pour la personne i, u_i, pouvait être relié à des variables explicatives d'autres personnes de l'échantillon. Ce n'était pas nécessaire parce qu'avec l'échantillonnage aléatoire (hypothèse MLR.2), u_i est automatiquement indépendant des variables explicatives pour les observations autres que i. Dans un contexte de séries temporelles, l'échantillonnage aléatoire n'est presque jamais vérifié, et nous devons donc explicitement supposer que la valeur attendue de u_t n'est pas liée à des variables explicatives pour toutes les périodes de temps.

Il est important de voir que l'hypothèse TS.3 ne souffre d'aucune restriction en ce qui concerne la corrélation de la variable indépendante ou la corrélation du terme d'erreur u_t dans le temps. L'hypothèse TS.3 dit simplement que la moyenne des valeurs de u_t ne doit pas être reliée aux variables explicatives pour toutes les périodes de temps.

Tout ce qui implique que les facteurs non observables à l'instant t soient corrélés avec n'importe quelle variable explicative pour n'importe quelle période implique l'invalidation de l'hypothèse TS.3.

Les deux candidats principaux à cet échec sont les variables omises et les erreurs de mesures dans les variables explicatives. Mais l'hypothèse d'exogénéité stricte peut également échouer pour d'autres raisons moins évidentes. Par exemple, dans le modèle de régression statique simple :

$$y_t = \beta_0 + \beta_1 z_t + u_t,$$

l'hypothèse TS.3 exige non seulement que u_t et z_t ne soient pas corrélés, mais aussi que u_t ne soit pas corrélé avec les valeurs passées et futures de z. Cela a deux conséquences. Tout d'abord, z ne peut avoir aucun effet avec retard sur y. Si z a un effet avec retard sur y, alors nous devons utiliser un modèle à retards échelonnés. Un point plus subtil est que l'exogénéité stricte exclut la possibilité que des changements du terme d'erreur aujourd'hui puissent provoquer des changements futurs de z. Ceci exclut donc un effet de y sur les valeurs futures de z. Par exemple, considérons un modèle statique simple pour expliquer le taux d'homicides ($mrdrte_t$) dans une ville en fonction du nombre de policiers par habitant ($polpc_t$) tel que :

$$mrdrte_t = \beta_0 + \beta_1 polpc_t + u_t.$$

Il semble raisonnable dans cet exemple de supposer que u_t n'est pas corrélé avec $polpc_t$, ni même avec les valeurs passées de $polpc_t$, nous entretenons quoi qu'il en soit cette hypothèse dans notre exemple. Supposons que la ville ajuste la taille de sa force de police en fonction des valeurs passées du taux d'homicides. Cela signifie donc que $polpc_{t+1}$ peut alors être corrélé avec u_t (étant donné qu'un u_t plus grand implique une hausse de $mrdrte_t$). Dans ce cas, l'hypothèse TS.3 est généralement violée.

Des considérations similaires existent aussi dans les modèles à retards échelonnés. Habituellement, nous ne nous inquiétons pas du fait que u_t puisse être corrélé avec les valeurs passées de z parce nous contrôlons justement pour les valeurs passées de z dans le modèle. Cependant, l'impact de u sur les futurs z est toujours un problème.

Les variables explicatives qui sont strictement exogènes ne peuvent pas réagir aux variations de y dans le passé. Un facteur tel que la quantité de pluie dans une fonction de production d'une exploitation agricole satisfait cette exigence : les précipitations d'une année future ne sont pas influencées par la production au cours de l'année actuelle ou des années passées. Mais quelque chose comme la quantité de travail peut ne pas être strictement exogène. La quantité de travail est en effet choisie par l'agriculteur, et ce dernier peut ajuster la quantité de travail en fonction des rendements de l'année précédente. Les variables de politique économique, tels que la croissance de la masse monétaire, les aides sociales versées, ou bien encore les limitations de vitesse sur la route sont souvent influencées par les valeurs passées d'autres variables. En sciences sociales, de nombreuses variables explicatives peuvent très facilement violer l'hypothèse d'exogénéité stricte.

Même si l'hypothèse TS.3 semble irréaliste, nous commençons avec celle-ci pour conclure que les estimateurs des MCO sont sans biais. La plupart des traitements de modèles à retards échelonnés statiques et finis supposent que TS.3 est validée en faisant l'hypothèse plus forte que les variables explicatives sont non-aléatoires, ou fixes, dans des échantillons répétés. L'hypothèse supposant que les variables explicatives sont non-aléatoires est évidemment fausse pour les observations de séries temporelles ; l'hypothèse TS.3 a l'avantage d'être plus réaliste quant à la nature aléatoire des x_{tj}, tout en isolant l'hypothèse nécessaire sur le lien entre u_t et les variables explicatives afin que les estimateurs des MCO ne soient pas biaisés.

Théorème 10.1
Absence de biais des MCO

Sous les hypothèses TS.1, TS.2, et TS.3, les estimateurs des MCO sont sans biais, conditionnellement à $\mathbf{X}$, et donc aussi sans biais non-conditionnellement : $E(\hat{\beta}_j) = \beta_j, j = 0, 1, \ldots, k$.

La démonstration de ce théorème est essentiellement la même que celle du Théorème 3.1 du chapitre 3, et nous l'omettons donc ici. Lorsque l'on compare le Théorème 10.1 au Théorème 3.1, nous voyons que nous avons été en mesure de supprimer l'hypothèse d'échantillonnage aléatoire en supposant que, pour tout t, u_t a une moyenne nulle, compte tenu des variables explicatives à toutes les périodes. Si cette hypothèse n'est pas vérifiée, il n'est pas possible de démontrer que les estimateurs des MCO sont sans biais.

Pour aller plus loin 10.2

Dans le modèle à retards échelonnés finis $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + u_t$, que doit-on supposer à propos de la séquence $\{z_0, z_1, \ldots, z_n\}$ afin que l'hypothèse TS.3 soit vérifiée ?

L'analyse du biais causé par les variables omises, que nous avions couverte dans la section 3.3, est essentiellement la même dans le cas des séries temporelles. En particulier, le tableau 3.2 et la discussion qui l'entoure peuvent être utilisés comme précédemment pour déterminer la direction du biais due à des variables omises.

Variance des estimateurs des MCO et théorème de Gauss-Markov

Nous devons ajouter deux hypothèses pour compléter les hypothèses de Gauss-Markov dans le cadre de régressions de séries temporelles. La première fait écho à un concept vu dans le cadre de l'analyse en coupe transversale.

Hypothèse TS.4
Homoscédasticité

Conditionnellement à $\mathbf{X}$, la variance de u_t est la même pour tout t : $\text{Var}(u_t|\mathbf{X}) = \text{Var}(u_t) = \sigma^2$, $t = 1, 2, \ldots, n$.

Cette hypothèse signifie que $\text{Var}(u_t|\mathbf{X})$ ne doit pas dépendre de $\mathbf{X}$ (il suffit pour cela que u_t et $\mathbf{X}$ soient indépendants) et que $\text{Var}(u_t)$ est constante dans le temps. Lorsque TS.4 n'est pas vérifiée, on dit que les erreurs sont *hétéroscédastiques*, comme dans le cadre de l'analyse en coupe transversale. Par exemple, considérons l'équation pour déterminer le taux d'intérêt des bons du Trésor à 3 mois ($i3_t$) en fonction du taux d'inflation (inf_t) et du déficit public en pourcentage du PIB (def_t) :

$$i3_t = \beta_0 + \beta_1 inf_t + \beta_2 def_t + u_t. \qquad \textbf{[10.11]}$$

Entre autres, l'hypothèse TS.4 exige que les effets inobservables affectant le taux d'intérêt aient une variance constante dans le temps. Étant donné que les changements de politique économique sont connus pour affecter la variabilité des taux d'intérêt, cette hypothèse pourrait très bien être fausse. En outre, il se pourrait que la variabilité des taux d'intérêt dépende du niveau de l'inflation ou de la taille relative du déficit. Ce serait également une violation de l'hypothèse d'homoscédasticité.

Lorsque $\text{Var}(u_t|\mathbf{X})$ dépend de $\mathbf{X}$, la variance dépend souvent des variables explicatives à l'instant t, $\mathbf{x}_t$. Dans le chapitre 12, nous verrons comment les tests d'hétéroscédasticité vus dans le chapitre 8 peuvent être utilisés pour les régressions de séries temporelles, tout du moins sous certaines conditions.

L'hypothèse finale de Gauss-Markov pour l'analyse des séries temporelles est quant à elle nouvelle.

Hypothèse TS.5
Absence d'autocorrélation

Conditionnellement à $\mathbf{X}$, les erreurs à deux périodes de temps différentes ne sont pas corrélées entre elles : $\text{Corr}(u_t, u_s|\mathbf{X}) = 0$, pour tout $t \neq s$.

La manière la plus simple de voir cette hypothèse consiste à ignorer dans un premier temps la propriété de conditionnalité à $\mathbf{X}$. Dans ce cas, l'hypothèse TS.5 s'écrit simplement :

$$\text{Corr}(u_t, u_s) = 0, \text{ pour tout } t \neq s. \qquad \textbf{[10.12]}$$

(C'est de cette manière que l'hypothèse d'absence d'autocorrélation est définie lorsque $\mathbf{X}$ n'est pas aléatoire.) Afin de voir si l'hypothèse TS.5 est vérifiée, nous allons nous concentrer sur l'équation (10.12), du fait de sa simplicité d'interprétation.

Lorsque l'équation (10.12) n'est pas vérifiée, nous disons alors que les erreurs de (10.8) souffrent de **corrélation sérielle** ou **d'autocorrélation**, car elles sont corrélées dans le temps. Considérons le cas d'erreurs adjacentes dans le temps. Supposons que lorsque $u_{t-1} > 0$ alors, en moyenne, l'erreur de la période suivante, u_t,

est positive. Dans ce cas, $\text{Corr}(u_t, u_{t-1}) > 0$, et les erreurs souffrent d'autocorrélation. Dans l'équation (10.11), cela signifie que si le taux d'intérêt est plus élevé qu'attendu lors d'une période, alors il sera aussi en moyenne (pour un niveau donné d'inflation et de déficit public) plus élevé à la période suivante. Cela s'avère d'ailleurs être une spécification raisonnable pour les termes d'erreur dans de nombreuses applications de séries temporelles, ce que nous verrons dans le chapitre 12. Pour le moment, supposons que l'hypothèse TS.5 soit vérifiée.

Il est important de noter que l'hypothèse TS.5 ne suppose rien en ce qui concerne la corrélation temporelle des variables indépendantes. Par exemple, dans l'équation (10.11), inf_t est presque certainement corrélée dans le temps. Mais cela n'a rien à voir avec le fait TS.5 soit vérifiée.

Une question naturelle qui se pose est pourquoi, dans les chapitres 3 et 4, n'avons-nous pas supposé que les erreurs pour les différentes observations en coupe transversale ne soient pas corrélées ? La réponse à cette question vient de l'hypothèse d'échantillonnage aléatoire : sous échantillonnage aléatoire, u_i et u_h sont indépendantes pour deux observations i et h. Il peut également être démontré que, sous échantillonnage aléatoire, les erreurs pour les différentes observations sont indépendantes conditionnellement aux variables explicatives de l'échantillon. Ainsi, pour nos besoins, nous considérons l'autocorrélation comme étant un problème potentiel seulement pour les régressions des données de séries temporelles. (Dans les chapitres 13 et 14, la question d'autocorrélation viendra aussi dans le cadre de l'analyse des données de panel.)

Les hypothèses TS.1 à TS.5 sont les hypothèses appropriées de Gauss-Markov pour l'analyse des séries temporelles, mais peuvent aussi être utilisées dans d'autres situations. Parfois, TS.1 à TS.5 sont satisfaites dans les modèles en coupe transversale, même lorsque l'échantillonnage aléatoire n'est pas une hypothèse raisonnable, par exemple lorsque les unités de l'échantillon en coupe transversale sont larges par rapport à la population. Supposons que nous ayons des données en coupe transversale par ville. Il est possible qu'une corrélation existe entre les villes d'un même État pour certaines des variables explicatives, comme les taux de l'impôt foncier ou le niveau des aides sociales par habitant. La corrélation des variables explicatives entre les observations ne cause pas de problèmes pour vérifier les hypothèses de Gauss-Markov, à condition que les termes d'erreur ne soient pas corrélés entre les villes. Cependant, dans ce chapitre, nous nous intéressons principalement à l'application des hypothèses de Gauss-Markov dans le cadre de régressions de série temporelles.

Théorème 10.2
Variance d'échantillonage des estimateurs des MCO

Pour les séries temporelles, et sous les hypothèses de Gauss-Markov TS.1 à TS.5, la variance de $\hat{\beta}_j$ conditionnellement à $\mathbf{X}$, est égale à :

$$\text{Var}(\hat{\beta}_j|\mathbf{X}) = \sigma^2/[\text{SCT}_j(1 - R_j^2)], j = 1, \ldots, k, \qquad \textbf{[10.13]}$$

où SCT_j est la somme des carrés totaux de x_{ij} et R_j^2 est le R-carré issu de la régression de x_j sur les autres variables explicatives.

L'équation (10.13) présente la même variance que celle dérivée au chapitre 3 sous les hypothèses de Gauss-Markov pour des données en coupe transversale. La démonstration étant similaire à celle du Théorème 3.2, nous l'omettons ici. La discussion du chapitre 3 sur les facteurs pouvant être à l'origine d'une variance élevée, y compris la multicolinéarité entre les variables explicatives, s'applique immédiatement au cas des séries temporelles.

Les estimateurs usuels de la variance de l'erreur ne sont pas biaisés sous les hypothèses TS.1 à TS.5, et le théorème de Gauss-Markov est vérifié.

Théorème 10.3
Estimation sans biais de σ^2

Sous les hypothèses TS.1 à TS.5, l'estimateur $\hat{\sigma}^2 = \text{SCR/df}$ est un estimateur sans biais de σ^2 où $df = n - k - 1$.

Théorème 10.4
Le théorème de Gauss-Markov

Sous les hypothèses TS.1 à TS.5, les estimateurs des MCO sont les meilleurs estimateurs linéaires possibles sans biais, conditionnellement à **X**.

L'essentiel ici est que les MCO ont les mêmes propriétés d'échantillons finis souhaitables sous TS.1 à TS.5 que sous MLR.1 à MLR.5.

Pour aller plus loin 10.3

Dans le modèle à retards échelonnés finis $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + u_t$, expliquez la nature de la multicolinéarité des variables explicatives.

Inférence sous les hypothèses classiques d'un modèle linéaire

Afin d'utiliser l'écart-type, la statistique *t* et la statistique F (Cf. chapitre 4) issus des MCO, nous devons ajouter une dernière hypothèse analogue à l'hypothèse de normalité utilisée dans le cadre de l'analyse en coupe transversale.

Hypothèse TS.6
Normalité

Les erreurs u_t sont indépendantes de **X** et indépendamment et identiquement distribuée (i.i.d) selon une loi normale $(0, \sigma^2)$

L'hypothèse TS.6 implique les hypothèses TS.3, TS.4, et TS.5, mais est plus forte de par les hypothèses d'indépendance et de normalité.

Théorème 10.5
Distribution normale

Sous les hypothèses TS.1 à TS.6 (les hypothèses classiques du modèle linéaire de séries temporelles) les estimateurs des MCO sont normalement distribués, conditionnellement à **X**. De plus, sous l'hypothèse nulle, chaque statistique *t* suit une distribution de Student, et chaque statistique F suit une loi de Fisher. La construction habituelle des intervalles de confiance est aussi valide.

Les implications du théorème 10.5 sont d'une importance particulière. Cela implique en effet que, lorsque les hypothèses TS.1 à TS.6 sont vérifiées, tout ce que nous avons vu précédemment à propos des estimations et de l'inférence statistique dans le cadre des régressions en coupe transversale s'appliquent directement

au cas des séries temporelles. Ainsi, la statistique t peut être utilisée pour tester la significativité individuelle des variables explicatives, et la statistique F pour tester la significativité jointe des variables explicatives.

Tout comme dans les cas des données en coupe transversale, les procédures d'inférence habituelles sont seulement aussi bonnes que les hypothèses sous-jacentes. Les hypothèses du modèle linéaire classique pour les données de séries temporelles sont beaucoup plus restrictives que celles des données en coupe transversale. En particulier l'hypothèse de stricte exogénéité et l'hypothèse d'absence d'autocorrélation peuvent être irréalistes. Néanmoins, le cadre du modèle linéaire classique est un bon point de départ pour de nombreuses applications.

EXEMPLE 10.1
Courbe de Phillips statique

Pour déterminer s'il existe un lien, en moyenne, entre le taux de chômage et le taux d'inflation, nous pouvons tester l'hypothèse $H_0 : \beta_1 = 0$ contre l'hypothèse $H_1 : \beta_1 < 0$ dans l'équation (10.2). Si les hypothèses classiques du modèle linéaire sont vérifiées, nous pouvons utiliser le t-stat issu de la méthode des MCO.

Utilisons le fichier PHILLIPS pour estimer l'équation (10.2), en se limitant aux données jusqu'à 1996. (Dans les exercices suivants, par exemple, dans les exercices pratiques sur ordinateur C12 et C10 du chapitre 11, vous devrez utiliser toutes les données jusqu'à 2003 ; dans le chapitre 18, nous utiliserons les données de 1997 à 2003 pour divers exercices de prévision.) Nous obtenons alors :

$$\widehat{inf}_t = 1{,}42 + 0{,}468\ unem_t$$
$$(1{,}72)\ \ (0{,}289)$$
$$n = 49,\ R^2 = 0{,}053,\ \overline{R}^2 = 0{,}033. \qquad \textbf{[10.14]}$$

Cette équation suggère qu'il existe un lien entre *unem* et *inf* : $\hat{\beta}_1 > 0$. La statistique t de $\hat{\beta}_1$ est d'environ 1,62, ce qui nous donne une p-value d'environ 0,11 contre une hypothèse alternative bilatérale. Ainsi, il semble exister une relation positive entre l'inflation et le chômage.

Il existe cependant certains problèmes en ce qui concerne cette analyse que nous ne pouvons pas voir en détail maintenant. Dans le chapitre 12, nous verrons que les hypothèses du modèle linéaire classique ne sont pas respectées dans ce modèle. En outre, la courbe de Phillips statique n'est probablement pas le meilleur modèle pour déterminer s'il existe un arbitrage à court terme entre l'inflation et le chômage. Les macro-économistes préfèrent généralement la courbe de Phillips augmentée des anticipations, dont un exemple simple sera donné dans le chapitre 11.

Dans ce deuxième exemple, nous allons estimer l'équation (10.11) en utilisant des données annuelles concernant l'économie américaine.

EXEMPLE 10.2
Effets de l'inflation et du déficit sur le taux d'intérêt

Les données INTDEF sont issues du Rapport Économique du Président des États-Unis pour l'année 2004 (tableau B-73 et B-79) et couvrent les années 1948 à 2003. La variable *i3* est le taux d'intérêt des bons du Trésor à trois mois, *inf* est le taux annuel d'inflation basé sur l'indice des prix à la consommation (IPC) et *def* est le déficit public en pourcentage du PIB. L'équation estimée est :

$$\widehat{i3}_t = 1{,}73 + 0{,}606\ inf_t + 0{,}513\ def_t$$
$$(0{,}43)\ \ (0{,}082)\ \ \ \ (0{,}118)$$
$$n = 56,\ R^2 = 0{,}602,\ \overline{R}^2 = 0{,}587. \qquad \textbf{[10.15]}$$

Ces estimations montrent qu'une hausse de l'inflation ou du déficit public augmente le taux d'intérêt à court-terme, comme attendu par la théorie économique. Par exemple, toutes choses égales par ailleurs, une augmentation d'un point de pourcentage du taux d'inflation entraîne une augmentation de *i3* de 0,606 points. *inf* et *def* sont tous deux nettement significatifs (en supposant bien sûr que les hypothèses du modèle linéaire classiques soient vérifiées).

10.4 FORME FONCTIONNELLE, VARIABLES BINAIRES ET NOMBRE INDICE

Toutes les formes fonctionnelles vues dans les chapitres précédents peuvent être utilisées dans le cadre des régressions de séries temporelles. La plus importante d'entre elle est le logarithme naturel : les régressions de séries temporelles avec des effets de pourcentage constants apparaissent en effet souvent dans les travaux appliqués.

EXEMPLE 10.3
Taux d'emploi à Porto Rico et salaire minimum

Les données annuelles concernant le taux d'emploi à Porto Rico, le salaire minimum, et d'autres variables, ont été utilisées par Castillo-Freeman et Freeman (1992) pour étudier l'effet du salaire minimum aux USA sur le taux de chômage à Porto Rico. Une version simplifiée de ce modèle peut s'écrire sous la forme :

$$\log(prepop_t) = \beta_0 + \beta_1 \log(mincov_t) + \beta_2 \log(usgnp_t) + u_t, \qquad \textbf{[10.16]}$$

où $prepop_t$ est le taux d'emploi à Porto Rico durant l'année t (ratio du nombre de travailleurs sur la population totale), $usgnp_t$ est le PNB réel américain (en milliards de dollars), et *mincov* mesure l'importance du salaire minimum relativement au salaire moyen. Plus précisément, $mincov = (avgmin/avgwage) \cdot avgcov$, où *avgmin* est le salaire moyen minimum, *avgwage* est le salaire moyen, et *avgcov* est le taux moyen de couverture (la proportion de travailleurs actuellement couverts par la loi sur le salaire minimum).

En utilisant les données du fichier PRMINWGE pour les années 1950 à 1987, nous obtenons

$$\widehat{\log(prepop_t)} = -1{,}05 - 0{,}154 \log(mincov_t) - 0{,}012 \log(usgnp_t)$$
$$(0{,}77)\ \ (0{,}065)\qquad\qquad (0{,}089)$$
$$n = 38,\ R^2 = 0{,}661,\ \bar{R}^2 = 0{,}641. \qquad \textbf{[10.17]}$$

L'élasticité estimée de *prepop* en fonction est de – 0,154, et est statistiquement significative avec $t = -2{,}37$. Par conséquent, une augmentation du salaire minimum réduit le taux d'emploi, en accord avec la théorie économique classique. Le PNB n'est pas statistiquement significatif, mais cela changera dans la section suivante lorsque nous introduirons une tendance temporelle dans le modèle.

Nous pouvons aussi utiliser une forme fonctionnelle logarithmique dans un modèle à retards échelonnés. Par exemple, avec des données trimestrielles, supposons que la demande de monnaie (M_t) et le PIB (GDP_t) soient reliés par la relation :

$$\log(Mt) = \alpha_0 + \delta_0 \log(GDP_t) + \delta_1 \log(GDP_{t-1}) + \delta_2 \log(GDP_{t-2})$$
$$+ \delta_3 \log(GDP_{t-3}) + \delta_4 \log(GDP_{t-4}) + u_t.$$

Le multiplicateur d'impact, δ_0, est aussi appelé élasticité de court terme : il mesure l'effet immédiat d'une hausse de 1 % du PIB sur la variation (en pourcentage) de demande de monnaie. Le multiplicateur de long terme, $\delta_0 + \delta_1 + \ldots + \delta_4$ est parfois appelé **élasticité de long terme** : il mesure la variation en pourcentage de la demande de monnaie après quatre trimestres, à la suite d'une hausse permanente de 1 % du PIB.

Les variables binaires ou indicatrices sont également très utiles pour l'analyse des séries temporelles. Comme l'unité d'observation est le temps, une variable indicatrice indique si, pour chaque période, un certain événement s'est produit. Par exemple, en données annuelles, nous pouvons indiquer pour chaque année si un démocrate ou un républicain était au pouvoir aux États-Unis, en définissant une variable $democ_t$ qui est égale à 1 si le président est un démocrate et à 0 autrement. Ou bien si l'on s'intéresse aux effets de la peine de mort sur le taux d'homicide au Texas, nous pouvons définir une variable indicatrice pour chaque année, égale à 1 si le Texas avait la peine de mort au cours de cette année, et à 0 autrement.

Souvent, les variables indicatrices sont utilisées afin d'isoler certaines périodes de temps qui peuvent être systématiquement différentes des autres périodes couvertes par un ensemble de données.

EXEMPLE 10.4
Effets des déductions fiscales pour enfants à charge

Le taux de fécondité (*gfr*) correspond au nombre d'enfants nés pour 1 000 femmes en âge de procréer. Pour les années 1913 à 1984, considérons l'équation :

$$gfr_t = \beta_0 + \beta_1 pe_t + \beta_2 ww2_t + \beta_3 pill_t + u_t,$$

afin d'expliquer *gfr* en fonction de la valeur moyenne en dollar réel des allocations familiales (*pe*) et de deux variables indicatrices *ww2 et pill*. La variable *ww2* prend la valeur 1 durant les années 1941 à 1945, lorsque les États-Unis étaient impliqués dans la Seconde Guerre Mondiale. La variable *pill* prend la valeur 1 à partir de 1963, date de la commercialisation de la pilule contraceptive.

En utilisant les données du fichier FERTIL3, issues de l'article de Whittington, Alm, et Peters (1990), nous obtenons :

$$\widehat{gfr_t} = 98{,}68 + 0{,}083\, pe_t - 24{,}24\, ww2_t - 31{,}59\, pill_t$$
$$(3{,}21)\ (0{,}030)\qquad (7{,}46)\qquad (4{,}08)$$
$$n = 72,\ R^2 = 0{,}473,\ \bar{R}^2 = 0{,}450. \qquad \textbf{[10.18]}$$

Chaque variable est significative au seuil de 1 % contre une hypothèse alternative bilatéral. Nous remarquons que le taux de fécondité a été plus faible durant la seconde Guerre Mondiale : toutes choses égales par ailleurs, il y a eu environ 24 naissances de moins pour 1 000 femmes en âge de procréer durant cette période, ce qui représente une forte diminution. (Entre 1913 et 1984, *gfr* est compris entre 65 à 127.) De la même manière, le taux de fécondité a fortement baissé depuis l'introduction de la pilule contraceptive.

La variable économique d'intérêt de notre modèle est *pe*. En moyenne, sur la période étudiée, *pe* est égal à 100,40 $, variant de 0 à 243,83 $ ou USD. Le coefficient de *pe* implique qu'une hausse de 12 $ de *pe* augmente *gfr* d'environ une naissance pour 1 000 femmes en âge de procréer. Cet effet n'est guère anodin.
Dans la section 10.2, nous avions noté que le taux de fécondité peut réagir à un changement de *pe* avec retard. En estimant un modèle à retards répartis d'ordre deux, nous obtenons :

$$\widehat{gfr_t} = 95{,}87 + 0{,}073\, pe_t - 0{,}0058\, pe_{t-1} + 0{,}034\, pe_{t-2}$$
$$(3{,}28)\ (0{,}126)\qquad (0{,}1557)\qquad (0{,}126)$$
$$- 22{,}13\, ww2_t - 31{,}30\, pill_t$$
$$(10{,}73)\qquad (3{,}98)$$
$$n = 70,\ R^2 = 0{,}499,\ \bar{R}^2 = 0{,}459. \qquad \textbf{[10.19]}$$

Dans cette régression, nous avons donc 70 observations contre 72 précédemment, car nous utilisons deux retards de la variable *pe*. Les coefficients des variables *pe* sont estimés avec très peu de précision, et individuellement, les coefficients ne sont pas significatifs. Il existe en effet une forte corrélation entre pe_t, pe_{t-1}, et pe_{t-2}, ce qui implique un problème de multicolinéarité et rend difficile l'analyse de l'effet de chaque retard. Cependant, pe_t, pe_{t-1}, and pe_{t-2} sont jointement significatifs : le *F*-stat ayant une *p*-value = 0,012. Ainsi, *pe* a un impact sur *gfr* [comme vu précédemment dans l'équation (10.18)], mais nous n'avons pas d'assez bonnes estimations pour déterminer si cette relation est simultanée ou bien décalée dans le temps (avec un décalage d'un et/ou deux ans). En réalité, pe_{t-1} and $pe_{t\ 2}$ sont conjointement non-significatifs dans cette équation (*p*-value = 0,95), donc à ce stade, nous pourrions justifier l'utilisation d'un modèle statique. Mais à titre d'illustration, nous allons calculer un intervalle de confiance pour le multiplicateur de long terme dans ce modèle.

La valeur estimée du multiplicateur de long terme de l'équation (10.19) est $0{,}073 - 0{,}0058 + 0{,}034 \approx 0{,}101$ Cependant, nous n'avons pas assez d'information dans (10.19) pour obtenir l'écart-type de cet estimateur. Pour obtenir l'écart-type de l'estimateur du multiplicateur de long terme, nous utilisons l'astuce suggérée dans la section 4.4. Définissons $\theta_0 = \delta_0 + \delta_1 + \delta_2$ comme étant le multiplicateur de long terme et écrivons δ_0 en fonction de θ_0, δ_1 et δ_2 comme $\delta_0 = \theta_0 - \delta_1 - \delta_2$.

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + \ldots$$

devient alors :

$$\begin{aligned} gfr_t &= \alpha_0 + (\theta_0 - \delta_1 - \delta_2)pe_t + \delta_1\, pe_{t-1} + \delta_2\, pe_{t-2} + \ldots \\ &= \alpha_0 + \theta_0\, pe_t + \delta_1(pe_{t-1} - pe_t) + \delta_2(pe_{t-2} - pe_t) + \ldots \end{aligned}$$

À partir de cette dernière équation, nous obtenons alors θ_0 et son écart-type en régressant gfr_t sur pe_t, $(pe_{t\ 1} - pe_t)$, $(pe_{t-2} - pe_t)$, $ww2_t$, et $pill_t$. Nous obtenons alors, comme voulu, le coefficient et l'écart-type de pe_t. En lançant cette régression, nous obtenons $\hat{\theta}_0 = 0{,}101$ comme coefficient de pe_t (comme nous avions précédemment et $\text{se}(\hat{\theta}_0) = 0{,}030$ [ce que nous ne pouvions pas calculer à partir de (10.19)]. Ainsi, la statistique *t* de θ_0 est d'environ 3,37, donc $\hat{\theta}_0$ est statistiquement différent de 0 avec un faible seuil de confiance. Bien qu'aucun des $\hat{\delta}_j$ ne soit individuellement significatif, le multiplicateur de long terme est quant à lui nettement significatif. Avec un intervalle de confiance de 95 %, le multiplicateur de long terme se trouve entre 0,041 et 0,160.

Whittington, Alm, et Peters (1990) autorisent l'existence d'un plus grand nombre de retards mais restreignent les coefficients, afin d'atténuer le problème de la multicolinéarité qui entrave l'estimation des δ_j. (Voir problème 6 pour un exemple de la procédure). Pour estimer le multiplicateur de long terme, ce qui semble être l'objectif principal ici, ces restrictions ne sont pas nécessaires. Whittington, Alm, et Peters utilisent quant à eux d'autres variables de contrôles, comme le salaire moyen des femmes et le taux de chômage, afin de vérifier que cette relation est robuste.

Les variables indicatrices sont un élément-clé dans ce qu'on appelle une étude d'évènement. Dans une étude d'événement, l'objectif est de voir si un événement particulier influence certains résultats. Les économistes qui étudient l'organisation industrielle se sont penchés sur les effets de certains événements sur les cours des actions des entreprises. Par exemple, Rose (1985) a étudié les effets de la dérégulation du secteur du transport sur les cours des actions des entreprises de transport.

Une version simple de l'équation utilisée dans une étude d'événement de ce type peut s'écrire sous la forme :

$$R_t^f = \beta_0 + \beta_1\, R_t^m + \beta_2 d_t + u_t$$

où R_t^f est le rendement de l'action de l'entreprise *f* durant la période *t* (habituellement une semaine ou un mois), R_t^m est le rendement du marché (habituellement calculé en considérant un indice du marché), et d_t est

une variable indicatrice indiquant quand l'événement étudié a eu lieu. Par exemple, si l'entreprise est une compagnie aérienne, d_t pourrait désigner si la compagnie aérienne a connu un accident ou un incident pendant la semaine t. L'inclusion de R_m^t dans l'équation permet de contrôler le fait que de larges mouvements de marché pourraient coïncider avec les crashs d'avion. Parfois, plusieurs variables indicatrices sont utilisées. Par exemple, si l'événement est l'imposition d'une nouvelle régulation qui pourrait affecter certaines entreprises, nous pouvons aussi inclure une variable indicatrice pour la semaine précédant l'annonce, et une autre variable indicatrice pour les quelques semaines après l'annonce officielle de la nouvelle régulation. La première variable indicatrice pourrait permettre de détecter la présence d'information privilégiée avant l'annonce officielle.

Avant de donner un exemple d'étude d'événement, nous avons besoin de discuter de la notion d'indice et de la différence entre les variables économiques nominales et réelles. Un indice agrège une vaste quantité d'information sous une seule valeur. Les indices sont régulièrement utilisés dans l'analyse des séries temporelles, principalement en macro-économie. Un exemple d'indice est l'indice de la production industrielle (IPI), calculé chaque mois par le Conseil des Gouverneurs de la Réserve Fédérale. L'IPI est une mesure de la production d'un large éventail d'industries, et, en tant que tel, sa valeur pour une année donnée n'a pas de sens quantitatif. Pour interpréter la grandeur de l'IPI, nous devons connaître la période de base et la valeur de base. Dans le *Rapport Économique du Président* (*REP*) de 1997, l'année de base est 1987, et la valeur de base est égale à 100. (Définir l'IPI comme étant égal à 100 pour l'année de base est une simple convention ; cela aurait autant de sens de définir l'IPI comme étant égal à 1 en 1987 ; certains indices sont d'ailleurs définis avec comme valeur 1 pour l'année de base). Étant donné que l'IPI était égal à 107,7 en 1992, nous pouvons dire que la production industrielle était 7,7 % plus élevée en 1992 qu'en 1987. Il est possible d'utiliser l'IPI pour deux années différentes afin de calculer la variation de production industrielle entre ces deux années. Par exemple, sachant que l'IPI était égal à 61,4 en 1970 et à 85,7 en 1979, la production industrielle a augmenté d'environ 39,6 % durant les années 70.

Il est facile de modifier la période de base pour n'importe quel indice, et il est parfois nécessaire de faire cela pour donner une base commune à différents indices. Par exemple, si nous voulons changer l'année de base de l'IPI de 1987 à 1982, nous devons simplement diviser l'IPI de chaque année par l'IPI de 1982, puis multiplier par 100 pour que la valeur de la période de base soit égale à 100. La formule à utiliser est :

$$newindex_t = 100(oldindex_t/oldindex_{\text{newbase}}), \qquad \textbf{[10.20]}$$

où $oldindex_{\text{newbase}}$ correspond à la valeur originale de l'indice pour l'année de base nouvellement définie. Par exemple, en base 1987, l'IPI de 1992 est égal 107,7 ; en utilisant la base 1982, l'IPI de 1992 devient 100(107.7/81.9) = 131.5 (car l'IPI de 1982 était égal à 81.9).

Un autre exemple important d'indice est l'indice des prix, comme par exemple l'indice des prix à la consommation (IPC). Nous avons d'ailleurs utilisé l'IPC pour calculer l'inflation dans l'exemple 10.1. Tout comme pour l'indice de production industrielle, l'IPC n'a de sens que lorsque l'on compare ses valeurs entre différentes années (ou entre différents mois si nous utilisons des données mensuelles). Par exemple dans le REP de 1997, nous avons IPC = 38,8 en 1970 et IPC = 130,7 en 1990. Ainsi, le niveau général des prix a augmenté de près de 237 % sur cette période de 20 ans (en 1997, l'IPC est défini tel que sa moyenne sur les années 1982, 1983, et 1984 soit égale à 100 ; avec donc comme années de base 1982-1984.)

En plus d'être utilisé pour calculer le taux d'inflation, les indices de prix sont nécessaires afin de transformer des séries temporelles exprimées en dollars nominaux (ou dollars courants) en dollar réels (ou dollars constants). La plupart des comportements économiques sont supposés être influencés par des variables réelles, et non par des variables nominales. Par exemple, l'économie du travail classique suppose que l'offre de travail est basée sur le salaire horaire réel, et non pas sur le salaire horaire nominal. Obtenir le salaire réel

à partir du salaire nominal est une tâche facile lorsque nous disposons d'un indice de prix comme l'Indice des Prix à la Consommation. Nous devons tout d'abord diviser l'IPC par 100, de telle sorte que l'année de base soit égale à 1. Ensuite, si w correspond au salaire horaire moyen nominal en dollars et p = IPC/100, le *salaire réel* est simplement w/p. Ce salaire est mesuré en dollars pour l'année de base de l'IPC. Par exemple, dans le tableau B-45 du REP de 1997, le salaire horaire moyen est reporté en termes nominaux et en dollars de 1982 (ce qui signifie que l'IPC utilisé dans le calcul du salaire réel avait comme année de base 1982). Ce tableau montre que le salaire horaire nominal en 1960 était 2,09 $, mais mesuré en dollars de 1982 (salaire horaire réel), le salaire était de 6,79 $. Le salaire réel a atteint son pic en 1973, à 8,55 $ en dollars de 1982, puis a diminué en termes réels à 7,40 $ en 1995. Il y a donc eu une baisse conséquente du salaire réel durant ces 22 années (si l'on compare le salaire nominal de 1973 de 3,94 $ avec celui 1995 de 11,44 $, nous obtenons une vision très erronée de la réalité. En effet, étant donné la baisse du salaire réel, cette hausse du salaire nominale est entièrement due à l'inflation).

Les mesures habituelles de la production économique sont exprimées en termes réels. La plus importante de ces mesures est le *produit intérieur brut*, ou PIB. Lorsque la presse parle de la croissance du PIB, c'est toujours le PIB *réel* qui est évoqué. Dans le tableau B-2 du REP de 2012, le PIB est reporté en milliards de dollars de 2005. Nous avons d'ailleurs utilisé une mesure similaire de la production et de la production nette globale réelle dans l'exemple 10.3.

Il existe une relation intéressante lorsque des variables en dollars réels sont utilisés en combinaison du logarithme naturel. Supposons par exemple que le nombre moyen d'heures travaillées par semaine soit relié au salaire réel, de telle sorte que :

$$\log(hours) = \beta_0 + \beta_1 \log(w/p) + u.$$

En utilisant $\log(w/p) = \log(w) - \log(p)$, nous pouvons écrire :

$$\log(hours) = \beta_0 + \beta_1 \log(w) + \beta_2 \log(p) + u, \qquad \textbf{[10.21]}$$

avec comme restriction $\beta_2 = -\beta_1$. Ainsi, l'hypothèse que seul le salaire réel influence l'offre de travail impose une restriction en ce qui concerne les paramètres du modèle (10.21). Si $\beta_2 \neq -\beta_1$ alors le niveau des prix a un effet sur l'offre de travail, ce qui peut être le cas si les travailleurs ne comprennent pas parfaitement la différence entre variables réelles et variables nominales.

Il y a de nombreux aspects pratiques concernant le calcul des indices, mais cela sort du cadre de cet ouvrage. Des discussions détaillées concernant les indices de prix peuvent être trouvées dans la plupart des ouvrages de macroéconomie, comme dans Mankiw (1994, Chapitre 2). Pour nous, il est simplement important d'être en mesure d'utiliser les indices dans l'analyse de régression. Comme mentionné précédemment, étant donné que les valeurs des indices ne sont pas particulièrement instructives, les indices apparaissent souvent sous forme logarithmique, de sorte que les coefficients de régression aient des interprétations en tant que pourcentage de variation.

Donnons maintenant un exemple d'une étude d'événement utilisant aussi un indice.

EXEMPLE 10.5
Droits antidumping et importation de produits chimiques

Krupp et Pollard (1996) ont analysé l'effet de la mise en place de droits antidumping sur les importations de divers produits chimiques. Nous nous intéressons ici à un produit chimique industriel spécifique, le chlorure de baryum : un agent de nettoyage utilisé dans divers procédés chimiques et pour la production d'essence.

Les données sont disponibles dans le fichier BARIUM. Au début des années 1980, les producteurs américains de chlorure de baryum pensaient que la Chine vendait le chlorure de baryum à un prix « injustement » bas (pratique connue sous le nom de dumping) et l'industrie américaine de production de chlorure de baryum a porté plainte auprès de la Commission Internationale du Commerce des États-Unis (U.S. International Trade Commission, ITC) en octobre 1983. L'ITC a rendu un jugement favorable aux intérêts des producteurs américains de chlorure de baryum en octobre 1984. Il existe plusieurs questions intéressantes dans ce cas, mais nous allons en traiter seulement quelques-unes. Tout d'abord, les importations étaient-elles anormalement élevées durant la période précédant immédiatement le dépôt de la plainte ? Deuxièmement, les importations ont-elles sensiblement changées après le dépôt de la plainte pour dumping ? Enfin, quelle a été la réduction des importations après l'annonce de la décision en faveur de l'industrie américaine par l'ITC ?

Pour répondre à ces trois questions, et en suivant Krupp et Pollard, nous définissons trois variables indicatrices : *befile6* est égale à 1 durant les six mois précédant la plainte, *affile6* correspond aux 6 mois après la plainte, et *afdec6* correspond aux 6 mois après la décision de l'ITC. La variable dépendante de notre modèle est le volume d'importation de chlorure de baryum en provenance de Chine *chnimp*, que nous utilisons sous sa forme logarithmique. Nous ajoutons comme variables explicatives (sous formes logarithmiques), un indice de la production chimique, *chempi* (pour contrôler la demande globale de chlorure de baryum), le volume de la production d'essence, *gas* (une autre variable de contrôle de la demande), et un indice de taux de change, *rtwex*, qui mesure la force du dollar par rapport à un panier de devises. L'indice de production chimique a été défini à 100 en juin 1977. L'analyse ici diffère quelque peu de celle de Krupp et Pollard car nous utilisons le logarithme naturel de toutes les variables (sauf pour les variables indicatrices bien évidemment), et nous incluons les trois variables indicatrices dans le même modèle.

En utilisant des données mensuelles allant de février 1978 à décembre 1988, nous obtenons :

$$\begin{aligned}\widehat{\log(chnimp)} &= -17{,}80 + 3{,}12\ \log(chempi) + 0{,}196\ \log(gas) \\ &\quad\ (21{,}05)\ \ (0{,}48)\qquad\quad (0{,}907) \\ &\quad + 0{,}983\ \log(rtwex) + 0{,}060\ befile6 - 0{,}032\ affile6 - 0{,}565\ afdec6 \\ &\quad\ \ (0{,}400)\qquad\ (0{,}261)\qquad\ (0{,}264)\qquad\ (0{,}286) \\ n &= 131,\ R^2 = 0{,}305,\ \overline{R}^2 = 0{,}271. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textbf{[10.22]}\end{aligned}$$

Dans l'équation ci-dessus *befile6* n'est pas statistiquement significatif, ce qui montre qu'il n'y a pas eu de hausse inhabituelle des importations durant les six mois précédant le dépôt de la plainte. De plus, bien que le coefficient de *affile6* soit négatif, le coefficient est petit (indiquant une baisse des importations en provenance de Chine de 3,2 %), et clairement non-significatif. Le coefficient de *afdec6* montre quant à lui une baisse significative des importations de chlorure de baryum en provenance de Chine à la suite de la décision de l'ITC en faveur de l'industrie américaine, ce qui n'est pas surprenant. Le coefficient est statistiquement significatif avec un seuil de confiance de 5 %.

Les coefficients des variables de contrôle ont le signe attendu : une hausse globale de la production chimique entraine une hausse de la demande en agent de nettoyage. Bien que cela puisse paraître surprenant, la production d'essence n'affecte pas significativement les importations chinoises. Quant au coefficient de log(*rtwex*), cela montre qu'une hausse de la valeur du dollar relativement aux autres devises entraîne une hausse de la demande d'importation en provenance de Chine, comme prévu par la théorie économique (en réalité, l'élasticité n'est pas significativement différente de 1. Pourquoi ?)

Les interactions entre variables qualitatives et variables quantitatives sont aussi utilisées dans le cadre de l'analyse des séries temporelles. Un exemple pratique de l'utilisation d'un terme d'interaction est présenté ci-après.

EXEMPLE 10.6
Résultat des élections et performance économique

Fair (1996) a travaillé sur l'explication des résultats de l'élection présidentielle en fonction des performances économiques. Il explique la proportion des votes allant vers le candidat démocrate à l'aide des données pour les années 1916 à 1992 pour un total de 20 observations (les élections américaines ayant lieu tous les quatre ans). Nous estimons une version simplifiée du modèle de Fair (en utilisant des noms de variables plus descriptifs que ceux utilisés par Fair) :

$$demvote = \beta_0 + \beta_1\, partyWH + \beta_2 incum + \beta_3\, partyWH \cdot gnews + \beta_4\, partyWH \cdot inf + u,$$

où *demvote* est la proportion des votes allant vers le candidat démocrate. La variable explicative *partyWH* est similaire à une variable indicatrice, mais prend la valeur 1 si le président actuel est un démocrate, et –1 si c'est un républicain. L'usage de cette variable impose la restriction que l'effet qu'un républicain ou qu'un démocrate soit au pouvoir est de même magnitude, mais de sens contraire. Il s'agit d'une restriction naturelle car la somme du pourcentage de vote reçu par les deux partis est égale à 1 par définition. Cela permet aussi de conserver deux degrés de liberté, ce qui est important étant donné le faible nombre d'observations. De la même manière, la variable *incum* est égale à 1 si le président sortant se représente et est un démocrate, –1 si le président sortant se représente et est un républicain, et 0 sinon (dans le cas de nouveaux candidats). La variable *gnews* correspond au nombre de trimestres, durant les 15 premiers trimestres de l'administration, où la croissance trimestrielle de la production réelle par habitant a été supérieure à 2,9 % (en rythme annuel). La variable *inf* est le taux d'inflation annuel moyen durant les 15 premiers trimestres. Voir Fair (1996) pour les définitions précises.

Les économistes sont principalement intéressés par les termes d'interactions *partyWH·gnews* et *partyWH·inf*. Étant donné que *partyWH* est égal à 1 lorsqu'un démocrate est à la Maison Blanche, β_3 mesure l'effet d'une bonne nouvelle macro-économique sur les votes pour le parti en place. Nous attendons donc $\beta_3 > 0$. De la même manière, β_4 mesure l'effet de l'inflation sur les votes pour le parti en place. L'inflation étant considérée durant un mandat comme une mauvaise nouvelle, nous attendons $\beta_4 < 0$.

En utilisant les données du fichier FAIR, l'équation estimée est :

$$\begin{aligned} \widehat{demvote} &= 0{,}481 - 0{,}0435\, partyWH + 0{,}0544\, incum \\ &\quad\ (0{,}012)\ \ (0{,}0405)\qquad\quad (0{,}0234) \\ &\quad + 0{,}0108\, partyWH \cdot gnews - 0{,}0077\, partyWH \cdot inf \\ &\quad\ (0{,}0041)\qquad\qquad\ (0{,}0033) \\ &\quad n = 20,\ R^2 = 0{,}663,\ \overline{R}^2 = 0{,}573. \end{aligned} \qquad \textbf{[10.23]}$$

Tous les coefficients, sauf celui de *partyWH*, sont significatifs avec un seuil de confiance de 5 %. Le fait que le président sortant se représente correspond environ à 5,4 points de pourcentage dans le vote. (Souvenez-vous, *demvote* est mesuré en pourcentage). De plus, la variable économique *gnews* a un effet positif : un trimestre supplémentaire de « bonne nouvelle » rapporte environ 1,1 point de pourcentage de vote. Enfin, et comme prévu, l'inflation a un effet négatif : en moyenne, une inflation annuelle de deux points de pourcentage de plus entraine une baisse de 1,5 point de pourcentage de votes.

Nous aurions pu utiliser cette équation afin de prévoir les résultats de l'élection de 1996 entre le démocrate Bill Clinton et le républicain Bob Dole (le candidat indépendant Ross Peror est exclu du modèle, car l'équation de Fair ne considère que deux partis). Étant donné que Clinton était le président sortant *partyWH* = 1 et *incum* = 1. Pour estimer les résultats de l'élection présidentielle, nous avons donc besoin des variables *gnews* et *inf*. Durant les 15 trimestres du mandat de Bill Clinton, la croissance trimestrielle de la production réelle par habitant a été supérieure à 2,9 % trois fois, donc *gnews* = 3. De plus, en utilisant le déflateur du PIB du tableau B-4 du REP de 1997, nous pouvons calculer le taux moyen annuel d'inflation (en utilisant la formule de Fair)

entre le 4e trimestre de 1991 et le 3e trimestre de 1996, qui a été égal à 3,019 %. En utilisant ces données dans (10.23), nous avons alors :

$$\widehat{demvote} = 0{,}481 - 0{,}0435 + 0{,}0544 + 0{,}0108(3) - 0{,}0077(3{,}019) \approx 0{,}5011.$$

Par conséquent, en se basant sur les informations disponibles avant le résultat de l'élection de novembre 1996 et en utilisant le modèle ci-dessus, Bill Clinton aurait du obtenir une courte majorité des voix dans le duel bipartite, avec environ 50,1 % des voix. En réalité, Clinton a remporté les élections haut la main en recevant 54,65 % des voix dans son duel contre Bob Dole.

10.5 TENDANCE ET SAISONNALITÉ

Caractérisation des tendances des séries temporelles

Beaucoup de séries temporelles économiques ont tendance à croître au cours du temps. Nous devons reconnaître que certaines séries contiennent une **tendance** dans le temps, afin de tirer une inférence causale lors de l'utilisation de données de séries temporelles. Ignorer le fait que deux séquences suivent une tendance commune ou opposée peut nous amener à conclure faussement que les variations d'une variable sont causées par les variations d'une autre variable. Dans de nombreux cas, deux processus de séries temporelles semblent être corrélées simplement parce qu'ils suivent une tendance dans le temps, et ce pour des raisons liées en réalité à d'autres facteurs non observés.

La figure 10.2 présente la productivité du travail (production par heure de travail) aux États-Unis de 1947 à 1987. Cette série présente une tendance clairement à la hausse, ce qui reflète le fait que les travailleurs sont devenus plus productifs au fil du temps. D'autres séries, tout du moins sur certaines périodes de temps, peuvent présenter une tendance à la baisse. Parce que les tendances à la hausse sont plus fréquentes, nous allons nous concentrer sur celles-ci lors de notre discussion.

Quel genre de modèle statistique peut permettre de capturer correctement ces phénomènes ? Une formule simple est d'écrire la série $\{y_t\}$ comme :

$$y_t = \alpha_0 + \alpha_1 \times t + e_t, \; t = 1, 2, \ldots, \qquad \textbf{[10.24]}$$

où, dans le cas le plus simple, $\{e_t\}$ est indépendant et identiquement distribué (i.i.d.) avec $E(e_t) = 0$ et $Var(e_t) = \sigma_e^2$. Le paramètre α_1 multiplie le temps t, ce qui permet de créer une tendance linéaire dans le temps. L'interprétation de α_1 dans (10.24) est simple : en fixant l'ensemble des autres paramètres (ici en fixant e_t), α_1 mesure le changement de y_t d'une période à l'autre en raison de l'écoulement du temps. Mathématiquement, en définissant le changement de y_t de la période $t-1$ à t de l'équation (10.24), nous obtenons :

$$\Delta y_t = y_t - y_{t-1} = \alpha_1.$$

Une autre façon de penser à un processus ayant une tendance linéaire est de voir que sa valeur moyenne est une fonction linéaire du temps :

$$E(y_t) = \alpha_0 + \alpha_1 t. \qquad \textbf{[10.25]}$$

Si $\alpha_1 > 0$ alors, en moyenne, y_t augmente dans le temps et a une tendance à la hausse. Si $\alpha_1 < 0$ alors y_t a une tendance à la baisse. Les valeurs de y_t ne tombent pas exactement sur la droite de l'équation (10.25) à chaque période étant donné la présence d'une composante aléatoire, mais l'espérance se situe sur la droite. Contrairement à la moyenne, la variance de y_t est constante dans le temps : $Var(y_t) = Var(e_t) = \sigma_e^2$.

Si $\{e_t\}$ est une suite i.i.d., alors $\{y_t\}$ est indépendant, bien que n'étant pas identiquement distribué. Une représentation plus réaliste des séries temporelles avec tendance permet à $\{e_t\}$ d'être corrélé dans le temps, mais cela ne change pas le rôle de la tendance linéaire. En réalité, ce qui est important dans le cadre d'une régression sous les hypothèses du modèle linéaire classique est que $E(y_t)$ soit linéaire en t. Lorsque nous couvrirons les propriétés asymptotiques des estimateurs des MCO, nous discuterons du niveau acceptable de corrélation temporelle de $\{e_t\}$.

Figure 10.2 *Production par heure travaillée aux États-Unis pour les années 1947 à 1987 ; 1977 = 100.*

Complément de 10.4

Dans l'exemple 10.4, nous avons utilisé le taux de fécondité en tant que variable dépendante d'un modèle à retard répartis finis. De 1950 au milieu des années 1980s, *gfr* a connu une nette tendance négative. L'utilisation d'une tendance linéaire négative avec $\alpha_1 < 0$ est-elle réaliste pour l'ensemble des futures périodes de temps ? Justifiez.

Beaucoup de séries temporelles économiques sont mieux estimées par une **tendance exponentielle**, ce qui a lieu lorsqu'une série a le même taux de croissance moyen de période en période. La figure 10.3 représente les importations nominales annuelles aux États-Unis pendant les années 1948 à 1995 (REP de 1997, tableau B-101).

Durant les premières années, nous voyons que la variation en valeur des importations d'années en années est relativement faible, tandis que les variations augmentent à mesure que le temps passe. Ceci est en accord avec un *taux de croissance moyen constant* : la variation en pourcentage est à peu près la même à chaque période.

Dans la pratique, une tendance exponentielle d'une série temporelle est capturée par la modélisation du logarithme naturel de la série en tant que tendance linéaire (en supposant que $y_t > 0$) :

$$\log(y_t) = \beta_0 + \beta_1 t + e_t, \; t = 1, 2, \ldots \qquad \textbf{[10.26]}$$

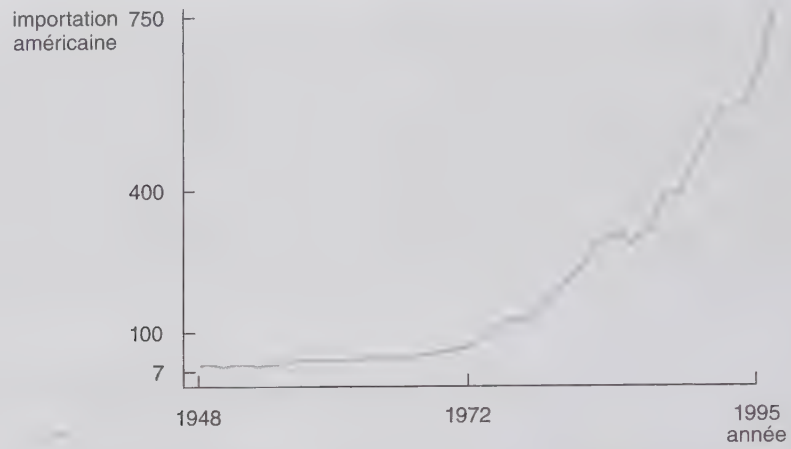


Figure 10.3 *Importation américaine en valeur (nominale) pour les années 1948 à 1995 (en milliards de dollars américains).*

Le passage à l'exponentiel montre que y_t lui-même contient une tendance exponentielle : $y_t = \exp(\beta_0 + \beta_1 t + e_t)$. Parce que nous allons vouloir utiliser des tendance exponentielle dans les modèles de régressions des séries temporelles, l'équation (10.26) s'avère être le moyen le plus simple de représenter ces séries.

Comment interpréter β_1 dans (10.26) ? Souvenez-vous que, pour de petites variations, $\Delta\log(y_t) = \log(y_t) - \log(y_{t-1})$ est approximativement égal au pourcentage de variation de y_t :

$$\Delta\log(y_t) \approx (y_t - y_{t-1})/y_{t-1}. \qquad \textbf{[10.27]}$$

Le côté droit de l'équation (10.27) est aussi appelé **taux de croissance** de y entre la période $t - 1$ et la période t. Pour passer du taux de croissance à un pourcentage, il suffit de multiplier le taux de croissance par 100. Si y_t suit (10.26), alors, en prenant en compte ces changements et avec $\Delta e_t = 0$:

$$\Delta\log(y_t) = \beta_1, \text{ pour tout } t. \qquad \textbf{[10.28]}$$

Pour le dire autrement, β_1 est approximativement égal au taux moyen de croissance par période de y_t. Par exemple, si t est une variable annuelle et $\beta_1 = 0{,}027$ alors y_t augmente en moyenne d'environ 2,7 % par an.

Bien que les tendances linéaires et exponentielles soient les plus courantes, les tendances temporelles peuvent prendre des formes plus complexes. Par exemple, au lieu du modèle de tendance linéaire (10.24), nous pourrions avoir une tendance quadratique en fonction du temps :

$$y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t. \qquad \textbf{[10.29]}$$

Si α_1 et α_2 sont positifs, alors la tendance est positive, comme on peut le voir facilement en estimant la pente (avec e_t fixe) :

$$\frac{\Delta y_t}{\Delta t} \approx \alpha_1 + 2\alpha_2 t. \qquad \textbf{[10.30]}$$

[Si vous êtes familier avec ce type de calcul, vous pouvez voir que le côté droit de l'équation (10.30) correspond à la dérivée de $\alpha_0 + \alpha_1 t + \alpha_2 t^2$ en fonction de t.] Si $\alpha_1 > 0$ et $\alpha_2 < 0$, la tendance a alors une forme de bosse. Cela n'est peut-être pas une très bonne description de certaines tendances, car cela requiert une tendance à la hausse suivie par une tendance à la baisse. Néanmoins, sur une période de temps donnée, cela peut être un moyen flexible de modélisation des séries temporelles qui ont des tendances plus complexes que celles des équations (10.24) et (10.26).

Utiliser les variables de tendance dans les régressions

La prise en compte de variables dépendantes ou explicatives ayant une tendance est assez simple. Tout d'abord, la spécificité des variables de tendance ne viole pas nécessairement les hypothèses du modèle linéaire classique TS.1 à TS.6. Cependant, nous devons être prudents en prenant en compte le fait que les facteurs inobservés de tendances qui affectent y_t pourraient également être corrélés avec les variables explicatives. Si nous ignorons cette possibilité, nous pouvons trouver une relation fallacieuse entre y_t et une ou plusieurs variables explicatives. Le phénomène consistant à détecter une relation entre deux ou plusieurs des variables avec tendance tout simplement parce que chaque série est croissante dans le temps est ce que l'on appelle un problème de régression fallacieuse. Heureusement, l'ajout d'une tendance temporelle élimine ce problème.

Pour être plus concret, considérons un modèle avec deux facteurs observés, x_{t1} et x_{t2}, affectant y_t. En outre, il existe des facteurs non observés qui sont systématiquement croissants ou décroissants dans le temps. Un modèle capturant cette relation peut s'écrire :

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 t + u_t. \qquad \textbf{[10.31]}$$

Cette équation rentre dans le cadre d'une régression linéaire classique avec $x_{t3} = t$. La prise en compte d'une tendance dans cette équation explicite le fait que y_t puisse croître ($\beta_3 > 0$) ou décroître ($\beta_3 < 0$) dans le temps pour des raisons indépendantes de x_{t1} et x_{t2}. Si (10.31) vérifie les hypothèses TS.1, TS.2, et TS.3, alors l'omission de t dans la régression et la simple régression de y_t par x_{t1} et x_{t2} va généralement biaiser les estimateurs β_1 et β_2. Ceci est spécialement vrai si x_{t1} et x_{t2} ont eux-mêmes une tendance, car ils peuvent être fortement corrélés avec t. L'exemple suivant montre pourquoi l'omission d'une tendance peut entrainer une régression fallacieuse.

EXEMPLE 10.7
Investissement résidentiel et prix de l'immobilier

Le fichier HSEINV présente les données annuelles concernant les investissements résidentiels et le prix de l'immobilier de résidence aux États-Unis de 1947 à 1988. Notons *invpc* l'investissement résidentiel réel par habitant (en milliers de dollars) et *price* l'indice des prix de l'immobilier résidentiel (égal à 1 en 1982). Une régression simple d'un modèle à élasticité constante, qui peut être considéré comme une équation de l'offre de logements, nous donne :

$$\widehat{\log(invpc)} = -0{,}550 + 1{,}241 \log(price)$$
$$(0{,}043)\ \ (0{,}382)$$
$$n = 42,\ R^2 = 0{,}208,\ \bar{R}^2 = 0{,}189. \qquad \textbf{[10.32]}$$

L'élasticité par habitant en fonction du prix est très élevée et significative ; statistiquement, cette élasticité n'est d'ailleurs pas différente de 1. Nous devons cependant être prudents ici. Les variables *invpc* et *price* ont toutes deux une tendance haussière. En particulier, si l'on régresse log(*invpc*) sur t, on obtient un coefficient de tendance égal à 0,0081 (écart-type = 0,0018) ; de la même façon, la régression de log(*price*) sur t donne un coefficient de tendance égal à 0,0044 (écart-type = 0,0004). Bien que les écarts-types des coefficients de tendance ne soient pas nécessairement fiables – ce type de régression ayant tendance à contenir une autocorrélation substantielle – les coefficients estimés révèlent une tendance à la hausse dans les deux cas.

Pour prendre en compte la tendance de nos variables, nous ajoutons donc au modèle une variable de tendance :

$$\widehat{\log(invpc)} = -0{,}913 - 0{,}381 \log(price) + 0{,}0098\, t$$
$$(0{,}136)\ (0{,}679)\qquad\qquad (0{,}0035)$$
$$n = 42,\ R^2 = 0{,}341,\ \bar{R}^2 = 0{,}307. \qquad \textbf{[10.33]}$$

L'histoire est désormais totalement différente : l'élasticité-prix estimée est maintenant négative et non-statistiquement différente de zéro. La variable de tendance est significative, et la valeur de son coefficient implique une hausse moyenne de 1 % par an de *invpc*. À partir de cette analyse, nous ne pouvons donc pas conclure que l'investissement résidentiel réel par habitant est influencé par les prix de l'immobilier. Il existe d'autres facteurs, capturés par la tendance, qui peuvent avoir un impact sur *invpc*, mais nous ne les avons pas modélisés ici. Le résultat de (10.32) montre une relation fallacieuse entre *invpc* et *price* due au fait que la variable de prix contient aussi une tendance haussière dans le temps.

Dans certains cas, l'ajout d'une tendance peut augmenter la significativité d'une variable-clé du modèle. Cela peut se produire si les variables dépendantes et indépendantes ont des tendances différentes (disons une haussière et une baissière), mais que le mouvement de la variable indépendante près de sa droite de tendance provoque un mouvement de la variable dépendante loin de sa droite de tendance.

EXEMPLE 10.8
Équation du taux de fécondité

Si nous ajoutons une tendance linéaire à l'équation (10.18), nous obtenons alors :

$$\widehat{gfr}_t = 111{,}77 + 0{,}279\, pe_t - 35{,}59\, ww2_t + 0{,}997\, pill_t - 1{,}15\, t$$
$$(3{,}36)\ (0{,}040)\qquad (6{,}30)\qquad (6{,}626)\qquad (0{,}19)$$
$$n = 72,\ R^2 = 0{,}662,\ \bar{R}^2 = 0{,}642 \qquad \textbf{[10.34]}$$

Le coefficient de *pe* est ici trois fois plus élevé que le coefficient estimé (10.18), et est nettement plus significatif. Il est intéressant de voir que la variable *pill* n'est plus significative une fois que l'on tient compte de la tendance linéaire. Comme nous pouvons le voir avec le coefficient de la tendance, *gfr* a diminué en moyenne sur la période, toutes choses égales par ailleurs.

Comme le taux de fécondité a connu à la fois une tendance à la hausse puis une tendance à la baisse au cours de la période de 1913 à 1984, nous pouvons voir la robustesse de l'effet estimé de *pe* en utilisant une tendance quadratique :

$$\widehat{gfr}_t = 124{,}09 + 0{,}348\, pe_t - 35{,}88\, ww2_t - 10{,}12\, pill_t$$
$$(4{,}36)\ (0{,}040)\qquad (5{,}71)\qquad (6{,}34)$$
$$- 2{,}53\, t + 0{,}0196\, t^2$$
$$(0{,}39)\quad (0{,}0050)$$
$$n = 72,\ R^2 = 0{,}727,\ \bar{R}^2 = 0{,}706. \qquad \textbf{[10.35]}$$

Le coefficient de *pe* est encore plus grand et encore plus significatif. Maintenant, la variable *pill* est de signe négatif comme attendu, et son effet est marginalement significatif. Les deux tendances sont significatives. La forme quadratique de la tendance est donc un moyen flexible de prendre en compte la tendance inhabituelle de *gfr*.

Vous pouvez vous demander pourquoi, dans l'exemple 10.8, nous nous sommes arrêtés à la forme quadratique de la tendance. En effet, rien ne nous empêchait d'ajouter par exemple t^3 comme variable indépendante, et d'ailleurs, cela pourrait même être justifié (voir exercice pratique sur ordinateur C6). Mais il est important de ne pas s'emporter en ajoutant un trop grand nombre de tendances dans un modèle. En effet, les tendances permettent de capturer les mouvements importants qui ne sont pas expliqués par les variables indépendantes du modèle. Si l'on inclut un grand nombre de termes polynomiaux en t, alors il est souvent possible d'expliquer la variable dépendante avec une bonne précision. Cependant,

cela ne permet pas de trouver les variables explicatives affectant y_t, ce qui est pourtant l'objectif de notre modélisation.

Supprimer la tendance d'une série temporelle avec une variable de tendance

Inclure une variable de tendance dans un modèle de régression permet d'obtenir une interprétation intéressante à propos de la série d'origine. Pour être plus concret, nous allons nous concentrer sur le modèle (10.31), mais nos conclusions peuvent ensuite être généralisées.

Lorsque l'on régresse y_t sur x_{t1}, x_{t2}, et t, nous obtenons l'équation ajustée :

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \hat{\beta}_2 x_{t2} + \hat{\beta}_3 t. \qquad \textbf{[10.36]}$$

Nous pouvons étendre les résultats de Frisch-Waugh sur l'interprétation partielle des MCO, que nous avons abordés dans la section 3-2, pour montrer que $\hat{\beta}_0$ et $\hat{\beta}_1$ peuvent être obtenus comme suit.

i. Régresser chacune des variables y_t, x_{t1}, et x_{t2} sur une constante et une variable de tendance t, et sauvegarder les résultats, en définissant, $\ddot{y}_t$, $\ddot{x}_{t1}$, $\ddot{x}_{t2}$, $t = 1, 2, \ldots, n$. Par exemple :

$$\ddot{y}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 t.$$

Nous pouvons donc voir cette variable comme étant linéairement sans tendance. En supprimant la tendance de y_t, nous avons estimé le modèle :

$$y_t = \alpha_0 + \alpha_1 t + e_t$$

par la méthode des MCO ; les résidus de cette régression, $\hat{e}_t = \ddot{y}_t$, n'ont alors plus de tendance (tout du moins dans l'échantillon). Une interprétation similaire est valable pour x_{t1} et x_{t2}

ii. Lancer la régression

$$\ddot{y}_t, \text{ sur } \ddot{x}_{t1}, \ddot{x}_{t2} \qquad \textbf{[10.37]}$$

(L'ajout d'une constante n'est pas nécessaire ; mais ajouter une constante ne change rien car son estimation sera égale à zéro). Cette régression nous donne alors exactement les mêmes résultats de $\hat{\beta}_1$ et $\hat{\beta}_2$ que dans (10.36).

Cela signifie donc que les estimations initiales de $\hat{\beta}_1$ et $\hat{\beta}_2$ peuvent être interprétées comme provenant d'une régression sans variable de tendance, mais où nous avons préalablement supprimé les tendances de la variable dépendante et des variables explicatives. Ces conclusions sont valides pour n'importe quelle nombre de variables indépendantes, ou bien si la tendance est de forme quadratique ou de tout autre forme polynomiale.

Si t est omis de (10.36), alors il n'y a pas de suppression de tendance, et y_t pourrait sembler être relié à un ou plusieurs x_{tj} simplement car les variables indépendantes contiennent aussi une tendance, comme nous l'avons vu dans l'exemple 10.7. Si le terme de tendance est significatif et que les résultats de la régression se trouvent fortement modifiés à la suite de l'ajout d'une tendance dans la régression, alors le résultat de la première régression sans tendance doit être traité avec suspicion.

L'interprétation de $\hat{\beta}_1$ et $\hat{\beta}_2$ montre l'intérêt d'ajouter une variable de tendance dans la régression si une ou plusieurs variables explicatives semblent avoir une tendance, et ce même si ce n'est pas le cas de la variable indépendante y_t. Si y_t n'a pas de tendance, mais que x_{t1} augmente avec le temps, alors exclure la tendance de la régression peut faire penser que x_{t1} n'a pas d'effet sur y_t, alors que les mouvements de x_{t1} autour de sa tendance peuvent en réalité affecter y_t. Cette situation est justement capturée si une tendance t est ajoutée dans la régression.

EXEMPLE 10.9
Taux d'emploi à Porto Rico

Si nous ajoutons une variable de tendance linéaire à l'équation (10.17), nous obtenons :

$$\widehat{\log(prepop_t)} = -8{,}70 - 0{,}169\log(mincov_t) + 1{,}06\log(usgnp_t) - 0{,}032\,t$$
$$(1{,}30)\ (0{,}044)\qquad\qquad (00{,}18)\qquad\qquad (0{,}005)$$
$$n = 38,\ R^2 = 0{,}847,\ \bar{R}^2 = 0{,}834. \qquad \textbf{[10.38]}$$

Le coefficient de log(*usgnp*) a radicalement changé : de – 0,012 et non significatif précédemment à 1,06 et largement significatif. Le coefficient du salaire minimum a été légèrement modifié, mais le fait que l'écart-type soit nettement plus faible rend log(*mincov*) encore plus significatif que précédemment.

La variable $prepop_t$ n'affiche pas clairement une tendance à la hausse ni à la baisse, mais log(*usgnp*) contient une tendance à la hausse. [Une régression log(*usgnp*) sur *t* donne un coefficient d'environ 0,03, donc *usgnp* augmente environ de 3 % par an sur la période.] Nous pouvons voir le coefficient de log(*usgnp*) de la manière suivante : lorsque le *usgnp* augmente de 1 % *au-dessus* de sa tendance de long terme, *prepop* augmente d'environ 1,06 %.

Calcul du R-Carré lorsque la variable dépendante contient une tendance

Le *R*-carré des régressions de séries temporelles est souvent très élevé, surtout lorsque l'on le compare avec le *R*-carré obtenu dans les analyses en coupe transversale. Cela signifie-t-il que nous comprenons mieux les facteurs ayant un impact sur *y* dans le cas des séries temporelles ? Pas nécessairement. D'un côté, les séries temporelles sont souvent présentées sous une forme agrégée (comme par exemple le salaire horaire moyen aux États-Unis), et les variables agrégées sont souvent plus faciles à expliquer que les données en provenance de coupe transversale, qui sont souvent désagrégées et concernent par exemple des individus, des familles ou bien des entreprises. Mais le *R*-carré et le *R*-carré ajusté peuvent être artificiellement élevés dans le cas de régression de séries temporelles si la variable dépendante présente une tendance. Souvenez vous que le R^2 est une mesure de l'erreur de variance relativement à la variance *y*. La formule du *R*-carré ajusté montre cette relation :

$$\bar{R}^2 = 1 - (\hat{\sigma}_u^2/\hat{\sigma}_y^2),$$

où $\hat{\sigma}_u^2$ est l'estimateur sans biais de la variance de l'erreur, $\hat{\sigma}_y^2$ = SCT / (n – 1), et SCT = $\sum_{t=1}^{n}(y_t - \bar{y})^2$. Estimer la variance de l'erreur lorsque y_t contient une tendance n'est pas problématique, à partir du moment où une variable de tendance est incluse dans le modèle de régression. Cependant, lorsque $E(y_t)$ suit, par exemple, une tendance linéaire [voir (10.24)], SCT/(n – 1) n'est plus un estimateur sans biais ou bien un estimateur convergent de $Var(y_t)$. En réalité, SCT/(n – 1) peut substantiellement surestimer la variance y_t, car cela ne prend pas en compte la tendance présente dans y_t.

Lorsque la variable dépendante contient une tendance linéaire, quadratique, ou bien tout autre tendance polynomiale, il est facile de calculer la qualité de l'ajustement en supprimant dans un premier temps la tendance de y_t. La méthode la plus simple consiste à calculer le *R*-carré usuel dans une régression où la variable dépendante a déjà été modifiée afin de supprimer la tendance. Par exemple, en partant du modèle (10.31), nous régressons tout d'abord y_t sur *t* afin d'obtenir les résidus de y_t, et ensuite nous régressons :

$$y_t \textit{ sur } x_{t1},\ x_{t2}, \text{ et } t. \qquad \textbf{[10.39]}$$

Le *R*-carré de cette régression est :

$$1 - \frac{\text{SSR}}{\sum_{t=1}^{n} \ddot{y}_t^2}, \qquad \textbf{[10.40]}$$

où SSR est similaire à la somme des carrés des résidus de l'équation (10.36). De fait $\sum_{t=1}^{n} \ddot{y}_t^2 \leq \sum_{t=1}^{n} (y_t - \bar{y})^2$ (cette inégalité est habituellement une inégalité stricte), le *R*-carré de (10.40) est inférieur ou égal, et habituellement inférieur, au *R*-carré de (10.36). (La somme des carrés des résidus est la même dans les deux régressions). Lorsque y_t contient une forte tendance linéaire, le *R*-carré de (10.40) peut être beaucoup plus faible que le *R*-carré usuel.

Le *R*-carré (10.40) reflète d'une meilleure façon la manière dont x_{t1} et x_{t2} permettent d'expliquer y_t, car il supprime l'effet de la tendance temporelle. En effet, il est toujours possible d'expliquer une variable ayant une tendance avec une quelconque forme de tendance, mais cela ne signifie pas que nous ayons trouvé un quelconque facteur ayant un impact sur y_t. Un *R*-carré ajusté peut aussi être calculé à partir de (10.40) : il suffit de diviser SSR par le nombre de degrés de liberté de (10.36), c'est-à-dire par $(n - 4)$, et de diviser $\sum_{t=1}^{n} \ddot{y}_t^2$ par $(n - 2)$, étant donné que deux variables de tendance sont utilisées pour supprimer la tendance de y_t. En général, SSR est divisé par le nombre de degrés de liberté dans une régression classique (en incluant les variables de tendance) et $\sum_{t=1}^{n} \ddot{y}_t^2$ est divisé par $(n - p)$, où p correspond au nombre de paramètres de tendance estimés pour supprimer la tendance de y_t. Wooldridge (1991a) fournit une approche détaillée en ce qui concerne la correction du nombre de degrés de liberté, mais un calcul simple permet d'obtenir une bonne approximation en utilisant le *R*-carré ajusté de la régression de $\ddot{y}_t$ sur t, t^2, ..., t^p, x_{t1}, ..., x_{tk}. Cela requiert simplement de supprimer la tendance de y_t afin d'obtenir $\ddot{y}_t$, et ensuite nous pouvons utiliser $\ddot{y}_t$ pour calculer les mesures classiques de précision du modèle.

EXEMPLE 10.10
Investissement en immobilier résidentiel

Dans l'exemple 10.7, nous avons vu que l'inclusion d'une variable de tendance linéaire en plus de la variable log(*price*) dans l'équation sur l'investissement en immobilier résidentiel avait un effet substantiel sur la valeur de l'élasticité-prix. Mais le *R*-carré de la régression (10.33), si on le considère tel quel, indique que le modèle explique 34,1 % de la variation de log(*invpc*). Ce résultat est trompeur. Si l'on supprime dans un premier temps la tendance log(*invpc*) puis que l'on régresse la variable sans tendance sur log(*price*) et *t*, le *R*-carré devient égal à 0,008, et le *R*-carré ajusté devient même négatif. De fait, les variations de log(*price*) autour de sa tendance n'ont pratiquement aucun effet sur les mouvements de log(*invpc*) autour de sa tendance. Ceci est cohérent avec le fait que la *statistique t* de log(*price*) dans l'équation (10.33) soit très faible.

Avant de terminer ce paragraphe, nous devons voir un dernier point. Dans le calcul de statistique F pour tester des hypothèses jointes, nous utilisons le *R*-carré habituel, sans élimination de la tendance. Souvenez-vous que le *R*-carré utilisé dans le calcul de statistique F est un simple dispositif de calcul, et donc la formule usuelle est toujours valide.

Saisonnalité

Lorsqu'une série temporelle est observée mensuellement ou trimestriellement (voire même hebdomadairement ou quotidiennement), cette série peut présenter un phénomène de saisonnalité. Par exemple, le nombre mensuel de mise en chantier dans le Mid-West est fortement influencé par la météo. Bien que la météo suive un processus en partie aléatoire, nous pouvons être certains que la météo durant le mois de janvier aura tendance à être moins clémente que la météo durant le mois de juin ; le nombre de mises en chantier est généralement donc plus élevé en juin qu'en janvier. Une façon de modéliser ce phénomène est d'accepter que les valeurs attendues de la série y_t soient différentes chaque mois. Pour prendre un autre exemple, les ventes au détail du quatrième trimestre chaque année sont typiquement plus élevées que les ventes du troisième trimestre, et ce grâce aux vacances de Nöel. Encore une fois, il est possible de capturer ce phénomène en acceptant que la moyenne des ventes au détail varie au cours de l'année, et ce en complément éventuellement de la présence d'une tendance. Par exemple, les ventes au détail du 1er trimestre de l'année précédente sont supérieures en valeur aux ventes du 4e trimestre il y a 30 ans, simplement car les ventes au détail ont fortement augmenté en valeur avec le temps. Cependant, si l'on compare les ventes au détail pour une année donnée, la saisonnalité due aux vacances de Noël tend à faire en sorte que les ventes au dernier trimestre sont généralement plus élevées.

Bien que de nombreuses séries mensuelles ou trimestrielles affichent une certaine saisonnalité, certaines n'en ont pas. Par exemple, il n'existe pas d'effet de saisonnalité en ce qui concerne le taux d'intérêt ou le taux d'inflation mensuel. De plus, les séries ayant une saisonnalité sont généralement « ajustées des variations saisonnières » avant leur publication au grand public. Une variable ajustée des variations saisonnières est une variable qui, en principe, affichait une saisonnalité, et dont la saisonnalité a été supprimée. L'ajustement saisonnier peut être fait de différentes façons, mais une discussion détaillée est au-delà de la portée de cet ouvrage [Voir Harvey (1990) et Hylleberg (1992) pour des traitements détaillés.]

L'ajustement saisonnier est devenu tellement courant qu'il est parfois même impossible de trouver les données non ajustées. Le PIB américain trimestriel est un exemple parfait. Dans le REP, de nombreuses données macroéconomiques sont rapportées avec des fréquences mensuelles (tout du moins pour les années récentes) et celles affichant des tendances saisonnières sont toutes corrigées des variations saisonnières. Les principales sources de séries macroéconomiques, y compris Citibase, corrigent aussi leurs séries en prenant en compte les variations saisonnières. Ainsi, la possibilité d'utiliser notre propre ajustement saisonnier est souvent limitée.

Parfois, nous travaillons avec des données non désaisonnalisées, et il est utile de savoir que des méthodes simples existent pour faire face à la saisonnalité dans les modèles de régression. En général, cela se fait en incluant un ensemble de **variables indicatrices saisonnières** pour tenir compte de la saisonnalité de la variable dépendante, des variables explicatives, ou des deux.

Cette approche est simple. Supposons que nous ayons des données mensuelles, et que nous pensons que les tendances saisonnières de l'année sont à peu près constantes dans le temps. Par exemple, comme Noël arrive toujours à la même période de l'année, on peut s'attendre à ce que les ventes au détail soient, en moyenne, plus élevées au cours du mois de décembre qu'au cours des mois précédents. Ou bien, étant donné que les conditions météorologiques sont très semblables au fil des années, on peut supposer que les mises en chantier dans le Midwest soient plus élevées en moyenne durant les mois d'été que durant les mois d'hiver. Un modèle général permettant de capter ces phénomènes peut s'écrire (avec donc des données mensuelles) :

$$y_t = \beta_0 + \delta_1 feb_t + \delta_2 mar_t + \delta_3 apr_t + \ldots + \delta_{11} dec_t + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \qquad \textbf{[10.41]}$$

où feb_t, mar_t, …, dec_t sont des variables indicatrices indiquant si la période t appartient au mois correspondant. Dans cette formule, janvier est le mois de base et β_0 est l'ordonnée pour janvier. S'il n'existe pas de

saisonnalité dans y_t, une fois les x_{tj} contrôlés alors les coefficients δ_1 à δ_{11} sont tous égaux à 0. Cela peut se tester facilement avec un F-test.

Complément de 10.5

Dans l'équation (10.41), quel est le coefficient pour le mois de mars ? Expliquez pourquoi les variables indicatrices saisonnières vérifient l'hypothèse de stricte exogénéité.

EXEMPLE 10.11
Effets des plaintes antidumping

Dans l'exemple 10.5, nous avons utilisé les données mensuelles, non corrigées des variations saisonnières, du fichier BARIUM. De ce fait, nous devons ajouter des variables indicatrices saisonnières afin de vérifier si cela ne modifie pas significativement notre conclusion. Cela pourrait en effet être le cas si le mois juste avant le dépôt de la plainte était un mois durant lequel les importations sont en moyenne plus fortes ou plus faibles que durant les autres mois de l'année. Lorsque nous ajoutons 11 variables indicatrices comme dans (10.41) et que nous testons l'hypothèse de significativité jointe, nous obtenons une p-value = 0,59, ce qui signifie que les variables indicatrices saisonnières sont jointement non significatives. De plus, aucun changement important n'est causé par l'incorporation de ces variables, une fois le niveau de significativité pris en compte. Dans leur étude, Krupp et Pollard (1996) ont utilisé trois variables indicatrices des saisons (automne, été et printemps, avec hiver comme saison de base), plutôt que d'utiliser une variable pour chaque mois, mais le raisonnement est essentiellement le même.

Si les données sont trimestrielles, alors nous devons introduire des variables indicatrices pour trois des quatre trimestres ; le dernier trimestre étant le trimestre de base. Parfois, il peut être utile de combiner des variables indicatrices de saisonnalité avec certaines variables x_{tj}, afin d'autoriser le fait que l'effet de x_t sur y_t puisse être différent au cours de l'année.

Tout comme l'inclusion d'une variable de tendance dans une régression permet de supprimer la tendance des données, l'inclusion de variables de saisonnalité permet de désaisonnaliser les données. Pour être plus concret, considérons l'équation (10.41) avec $k = 2$. Les coefficients β_1 et β_2 de x_1 et x_2 sont obtenus de la manière suivante :

(i) Régresser chacune des variables y_t, x_{t1}, et x_{t2} sur une constante et les variables de saisonnalité feb_t, mar_t, ..., dec_t, puis sauvegarder les résidus, que l'on appellera, $\ddot{y}_t$, $\ddot{x}_{t1}$ et $\ddot{x}_{t2}$ pour tout $t = 1, 2 \ldots, n$. Par exemple :

$$\ddot{y}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 feb_t - \hat{\alpha}_2 mar_t - \ldots - \hat{\alpha}_{11} dec_t.$$

Cette méthode permet de désaisonnaliser les séries mensuelles.

(ii) Lancer la régression, sans les variables indicatrices mensuelles, de y_1 sur x_{t1} et x_{t2} [tout comme dans (10.37)]. Cela nous donne $\hat{\beta}_1$ et $\hat{\beta}_2$

Dans certains cas, si y_t a une forte tendance saisonnière, une meilleure mesure de la qualité d'ajustement du modèle est un R-carré basé sur la variable y_t désaisonnalisée. Cela permet de supprimer la saisonnalité non-expliquée par les x_{tj}. Wooldridge (1991a) suggère un ajustement spécifique en fonction du nombre de degrés de liberté, mais il est aussi possible d'utiliser simplement le R-carré ajusté lorsque la variable dépendante a été désaisonnalisée.

Les séries temporelles affichant une saisonnalité peuvent aussi avoir en même temps une tendance ; dans ce cas, nous devons estimer un modèle de régression avec une variable de tendance et des variables indicatrices de saisonnalité. Les régressions peuvent ensuite s'interpréter comme des régressions avec des séries désaisonnalisées et sans tendance. La qualité de l'estimation est discutée dans Wooldridge (1991a) : essentiellement, nous supprimons la tendance et la saisonnalité de y_t en régressant y_t sur une variable de tendance et des variables indicatrices de saisonnalité avant de calculer le R-carré et le R-carré ajusté.

RÉSUMÉ

Dans ce chapitre, nous avons couvert les bases de l'économétrie des séries temporelles. Avec des hypothèses semblables à celles vues lors de l'analyse en coupe transversale, les estimateurs des MCO sont sans biais (sous les hypothèses TS.1 à TS.3), l'estimateur MCO est le meilleur estimateur linéaire sans biais (aussi appelé BLUE en anglais pour « Best-Linear-Unbiased-Estimator » sous les hypothèses TS.1 à TS.5), et les écarts-types, les statistiques t et F des MCO peuvent être utilisés pour l'inférence statistique (sous les hypothèses TS.1 à TS.6). À cause de la corrélation temporelle de nombreuses séries temporelles, nous devons explicitement faire des hypothèses sur la façon dont les erreurs sont liées aux variables explicatives pour toutes les périodes de temps, ainsi que sur la corrélation temporelle des termes d'erreur entre eux. Les hypothèses du modèle classique linéaire peuvent être assez restrictives pour certaines applications et analyses de séries temporelles, elles constituent néanmoins un point de départ naturel. Nous avons appliqué cela à la fois aux modèles statiques de régression et aux modèles à retard échelonnés finis.

Les logarithmes et les variables indicatrices sont régulièrement utilisés dans les régressions de séries temporelles et dans les études d'événements. Nous avons aussi discuté des indices et de la différence entre les séries temporelles mesurées en termes nominaux et en termes réels.

Les problèmes dus à la présence de tendance et de saisonnalité peuvent être facilement résolus via un modèle de régression multiple incluant une variable de tendance et des variables indicatrices de saisonnalité. Nous avons présenté les problèmes que peuvent poser l'utilisation du R-carré classique comme mesure de la qualité de l'ajustement, et proposé une alternative simple basée sur la suppression des tendances et la désaisonnalisation des séries.

HYPOTHÈSES DU MODÈLE LINÉAIRE CLASSIQUE POUR LES SÉRIES TEMPORELLES

Nous allons ici résumer les six hypothèses du modèle linéaire classique pour l'économétrie des séries temporelles. Les hypothèses TS.1 à TS.5 correspondent aux hypothèses de Gauss-Markov dans le cadre des séries temporelles (qui impliquent que l'estimateur MCO soit le meilleur estimateur linéaire sans biais – BLUE – et ait les caractéristiques classiques de variance). Nous avons simplement besoin de TS.1, TS.2, et TS.3 pour établir le caractère sans biais des MCO. Tout comme dans le cas des régressions avec des données en coupe transversale, l'hypothèse de normalité TS.6, est utilisée afin de pouvoir réaliser les inférences statistiques pour toutes tailles d'échantillon.

Hypothèse TS.1 (Linéarité des paramètres)

Le processus stochastique $\{(x_{t1}, x_{t2}, \ldots, x_{tk}, y_t) : t = 1, 2, \ldots, n\}$ suit un modèle linéaire

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \qquad \textbf{[10.8]}$$

où $\{u_t : t = 1, 2, \ldots, n\}$ correspond à la série des erreurs ou bruits. Ici, n est le nombre d'observations (périodes de temps).

Hypothèse TS.2 (Absence de colinéarité parfaite)

Dans l'échantillon (et donc dans le processus temporel sous-jacent), aucune variable indépendante n'est constante, ni est une combinaison linéaire parfaite d'autres variables indépendantes.

Hypothèse TS.3 (Espérance conditionnelle nulle)

Pour chaque t, l'espérance mathématique du terme d'erreur u_t, compte tenu des variables explicatives pour toutes les périodes, est égal à zéro. Mathématiquement,

$$E(u_t|\mathbf{X}) = 0, t = 1, 2, \ldots, n. \quad \textbf{[10.9]}$$

L'hypothèse TS.3 remplace l'hypothèse MLR.4 vue dans le cas des données en coupe transversale, ce qui signifie aussi que nous n'avons pas à vérifier l'hypothèse MLR.2 d'échantillonnage aléatoire. Souvenez-vous, l'hypothèse TS.3 implique que l'erreur est, à chaque période de temps, non-corrélée avec l'ensemble des variables explicatives pour toutes les périodes (ce qui inclut bien évidemment la période t).

Hypothèse TS.4 (Homoscédasticité)

Conditionnellement à $\mathbf{X}$, la variance de u_t est la même pour tout t : $\text{Var}(u_t|\mathbf{X}) = \text{Var}(u_t) = \sigma^2$,

$$t = 1, 2, \ldots, n.$$

Hypothèse TS.5 (Absence d'autocorrélation)

Conditionnellement à $\mathbf{X}$, les erreurs à deux périodes de temps différentes ne sont pas corrélées entre elles : $\text{Corr}(u_t, u_s|\mathbf{X}) = 0$, pour tout $t \neq s$

Souvenez-vous que nous avions ajouté l'hypothèse d'absence d'autocorrélation, en plus de celle d'homoscédasticité, pour obtenir les mêmes formules de variance que celles obtenues dans le cas des régressions en coupe transversale avec échantillonnage aléatoire. Comme nous le verrons dans le chapitre 12, l'hypothèse TS.5 est souvent violée, ce qui peut rendre l'inférence statistique classique douteuse.

Hypothèse TS.6 (Normalité)

Les erreurs u_t sont indépendantes de $\mathbf{X}$ et dont indépendamment et identiquement distribuée (i.i.d) selon une loi normale $(0, \sigma^2)$

MOTS-CLÉS

Autocorrélation p. 421
Corrélation sérielle p. 421
Corrigée des variations saisonnières p. 440
Désaisonnaliser p. 441
Distribution des retards p. 415
Effet cumulé p. 416
Élasticité court terme p. 426
Élasticité de long terme p. 426
Étude d'événement p. 427
Exogène contemporainement p. 419
Impact de court terme p. 415
Impact de long terme p. 415

PROBLÈMES

1. Êtes-vous en accord ou en désaccord avec chacune des propositions suivantes ? Justifiez brièvement vos choix :

i. Tout comme dans le cas des observations en coupe transversale, nous pouvons supposer que les observations des séries temporelles sont indépendamment distribuées.

ii. Les estimateurs des MCO issus d'une régression de séries temporelles sont sans biais si les trois premières hypothèses de Gauss-Markov sont vérifiées.

iii. Une variable avec tendance ne peut être utilisée comme variable dépendante dans un modèle de régression multiple.

iv. La saisonnalité n'est pas un problème lorsque nous utilisons des données annuelles.

2. Soit $gGDP_t$ le pourcentage annuel de variation du PIB et int_t le taux d'intérêt à court terme. Supposons que $gGDP_t$ est relié au taux d'intérêt par la relation :

$$gGDP_t = \alpha_0 + \delta_0 int_t + \delta_1 int_{t-1} + u_t,$$

où u_t n'est pas corrélé avec int_t, int_{t-1}, ni avec toutes autres valeurs passées du taux d'intérêt. Supposons que la Réserve Fédérale suive une règle de politique monétaire telle que :

$$int_t = \gamma_0 + \gamma_1(gGDP_{t-1} - 3) + v_t,$$

où $\gamma_0 > 0$ (ce qui signifie que lorsque la croissance du PIB de l'année précédente est supérieure à 3 %, la FED augmente le taux d'intérêt afin de prévenir du risque de surchauffe de l'économie). Si v_t n'est pas corrélé avec les valeurs de int_t et de u_t, justifiez que int_t doit être corrélé avec u_{t-1}. (*Astuce* : Réécrivez la première équation avec un retard, puis substituez $gGDP_{t-1}$ dans la seconde équation.) Quelle hypothèse de Gauss-Markov est violée dans cette situation ?

3. Supposons que y_t suive un modèle à retards échelonnés d'ordre 2 :

$$y_t = \alpha_0 + \delta_0 z_1 + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t.$$

Soit z^* la valeur d'équilibre de z_t et y^* la valeur d'équilibre de y_t, de telle sorte que

$$y^* = \alpha_0 + \delta_0 z^* + \delta_1 z^* + \delta_2 z^*$$

Montrez qu'une variation de y^*, due à un changement de z^*, est égale au multiplicateur de long terme multiplié par la variation de z^* :

$$\Delta y^* = \text{LPR} \cdot \Delta_Z *$$

Ceci permet de donner une nouvelle interprétation du multiplicateur de long terme.

4. Lorsque les trois variables indicatrices d'événements *befile6*, *affile6*, et *afdec6* sont supprimées de l'équation (10.22), nous obtenons $R^2 = 0{,}281$ et $\bar{R}^2 = 0{,}264$. Les variables d'évènements sont-elles conjointement significatives avec un intervalle de confiance à 10 % ?

5. Supposons que vous disposiez de données trimestrielles concernant le nombre de mises en chantier, le taux d'intérêt et le revenu réel par habitant. Spécifiez un modèle pour expliquer le nombre de mises en chantier, en prenant en compte la possibilité de présence d'une tendance ou de saisonnalité dans les variables.

6. Dans l'exemple 10.4, nous avons vu que nos estimateurs pour chacun des retards dans un graphique de distribution des retards étaient très imprécis. Une manière de soulager le problème de multicolinéarité est de supposer que δ_j suit une forme relativement simple. Pour voir cela avec un exemple, considérons un modèle avec quatre retards, tel que :

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \delta_3 z_{t-3} + \delta_4 z_{t-4} + u_t.$$

Maintenant, supposons que δ_j suive une fonction quadratique des retards, j :

$$\delta_j = \gamma_0 + \gamma_1 j + \gamma_2 j^2,$$

pour les paramètres γ_0, γ_1, et γ_2. Ceci est un exemple de modèle polynomial à retards échelonnés.

i. Utilisez la formule pour chaque δ_j dans le modèle à retards échelonnés, et écrivez le modèle en fonction des paramètres γ_h pour $h = 0$, 1 et 2.

ii. Expliquez la régression que vous feriez pour estimer les γ_h.

iii. Le modèle polynomial à retards échelonnés est une version restreinte du modèle général. Combien de restrictions sont imposées ? Comment est-il possible de tester cela ? (*Astuce* : Réfléchissez au F-test.)

7. Dans l'exemple 10.4, nous avons écrit un modèle qui contient explicitement le multiplicateur de long terme θ_0, tel que :

$$gfr_t = \alpha_0 + \theta_0 pe_t + \delta_1(pe_{t-1} - pe_t) + \delta_2(pe_{t-2} - pe_t) + u_t,$$

Les autres variables explicatives sont omises ici par simplicité. Comme toujours dans l'analyse d'une régression multiple, θ_0 peut avoir une interprétation toutes choses égales par ailleurs. À savoir, si pe_t augmente de 1 (dollar) tout en gardant $(pe_{t-1} - pe_t)$ et $(pe_{t-2} - pe_t)$ fixes, gfr_t doit augmenter de θ_0.

i. Si $(pe_{t-1} - pe_t)$ et $(pe_{t-2} - pe_t)$ sont fixes mais que pe_t augmente, qu'est ce qui doit être vrai à propos des changements de pe_{t-1} et pe_{t-2} ?

ii. Comment votre réponse à la question (i) vous permet de mieux interpréter le multiplicateur de long terme θ_0 de l'équation ci-dessus ?

8. Dans le modèle linéaire de l'équation (10.8), les variables explicatives $\mathbf{x}_t = (x_{t1}, \ldots, x_{tk})$ sont dites séquentiellement exogènes (ou bien *faiblement exogènes*) si :

$$E(u_t|\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1) = 0, \; t = 1, 2, \ldots,$$

de telle sorte que les erreurs ne soient pas prévisibles étant donné la valeur actuelle et toutes les valeurs passées des variables explicatives.

i. Expliquez pourquoi l'exogénéité séquentielle est impliquée par l'exogénéité stricte.

ii. Expliquez pourquoi l'exogénéité contemporaine est impliquée par l'exogénéité séquentielle.

iii. Les estimateurs des MCO sont-ils généralement sans biais sous l'hypothèse d'exogénéité séquentielle ? Justifiez.

iv. Considérez un modèle ayant pour but d'expliquer le taux annuel d'infection par le Virus HIV (*HIVrate*) en fonction du nombre de préservatifs utilisés par habitant (*pccon*) dans un pays, une région ou une province :

$$E(HIVrate_t|pccon_t, pccont_{t-1}, \ldots,) = \alpha_0 + \delta_0 pccon_t + \delta_1 pccon_{t-1} + \delta_2 pccon_{t-2} + \delta_3 pccon_{t-3}.$$

Expliquez pourquoi ce modèle satisfait l'hypothèse d'exogénéité séquentielle. Est-ce que cela semble être le cas aussi en ce qui concerne l'hypothèse stricte d'exogénéité ?

EXERCICES SUR ORDINATEUR

C1. En octobre 1979, la Réserve Fédérale a changé sa politique monétaire d'ajustement du taux d'intérêt, pour commencer à cibler l'offre de monnaie. En utilisant les données du fichier INTDEF, définissez une variable indicatrice égale à 1 pour toutes les années après 1979. Utilisez ensuite cette variable dans l'équation (10.15) afin de voir s'il existe un changement dans l'équation après 1979. Que pouvez-vous conclure ?

C2. Utilisez les données du fichier BARIUM pour cet exercice.

i. Ajoutez une tendance temporelle linéaire à l'équation (10.22). En dehors de cette tendance, est-ce que d'autres variables sont significatives ?

ii. Dans l'équation estimée de la question (i), testez la significativité jointe de toutes les variables (en dehors de la tendance). Que pouvez-vous en conclure ?

iii. Ajoutez des variables indicatrices mensuelles à cette équation afin voir s'il existe un phénomène de saisonnalité. L'inclusion de ces variables change-t-elle les estimateurs ou les écarts-types des variables du modèle initial ?

C3. Ajoutez la variable log(*prgnp*) à l'équation du salaire minimum de (10.38). Cette variable est-elle significative ? Interprétez son coefficient. Quel est l'impact de l'ajout de la variable log(*prgnp*) sur l'estimation du salaire minimum ?

C4. Utilisez les données du fichier FERTIL3, afin de vérifier que l'écart-type du multiplicateur de long terme de l'équation (10.19) est environ égal à 0,030.

C5. Utilisez pour cet exercice les données du fichier EZANDERS, concernant le nombre mensuel de demande d'allocations chômage dans le canton d'Anderson dans l'Indiana, de janvier 1980 à novembre 1988. En 1984, une zone économique spéciale (ZES) a été créée à Anderson (tout comme dans d'autres villes de l'Indiana). [voir Papke (1994) pour plus de détails.]

i. Régressez log(*uclms*) sur une tendance linéaire et 11 variables indicatrices mensuelles. Quelle a été la tendance générale concernant le nombre de demande d'allocations chômage sur la période ? (Interprétez le coefficient de la tendance temporelle linéaire.) Il y a-t-il un phénomène de saisonnalité visible ?

ii. Ajoutez à la régression (i) la variable indicatrice *ez*, égale à 1 à partir du mois où une Zone Économique Spéciale a été créée à Anderson. La création d'une zone économique spéciale a-t-elle permis de diminuer le nombre de demande d'allocations chômage dans le canton ? Si oui, dans quelle mesure ? [Utilisez la formule (7.10) du chapitre 7.]

iii. Quelles hypothèses devez-vous faire pour attribuer l'effet de (ii) à la création de la Zone Économique Spéciale ?

C6. Utilisez les données du fichier FERTIL3 pour cet exercice.

i. Régressez gfr_t sur t et t^2 et sauvegardez les résidus afin de désaisonnaliser gfr_t. Nommez cette variable gf_t.

ii. Régressez gf_t sur les autres variables de l'équation (10.35), en incluant t et t^2. Comparez le R-carré de cette régression avec le R-carré de (10.35). Que pouvez-vous conclure ?

iii. En ajoutant t^3 à l'équation, estimez de nouveau l'équation (10.35). Est-ce que ce terme est statistiquement significatif ?

C7. Utilisez les données du fichier CONSUMP pour cet exercice.

i. Estimez un modèle de régression simple reliant la croissance de la consommation réelle par habitant (biens non-durables et services) à la croissance du revenu réel par habitant. Utilisez la variation du logarithme pour les deux variables. Après avoir estimé l'équation, interprétez et discutez de la significativité statistique du modèle.

ii. Ajoutez un retard de la croissance du revenu réel par habitant à l'équation (i). Que pouvez-vous conclure à propos du temps d'ajustement de la consommation suite à une hausse du revenu réel ?

iii. Ajoutez le taux d'intérêt réel à l'équation (i). Cela a t-il un impact sur la croissance de la consommation ?

C8. Utilisez les données du fichier FERTIL3 pour cet exercice.

i. Ajoutez pe_{t-3} et pe_{t-4} à l'équation (10.19), puis testez la significativité jointe de ces retards.

ii. Trouvez la valeur estimée du multiplicateur de long terme et calculez son écart-type à partir de la question (i). Comparez ces valeurs avec celles obtenus dans l'équation (10.19).

iii. Estimez un modèle à retards échelonnés polynomiaux à partir du problème 6. Trouvez le multiplicateur de long terme, et comparez le avec le résultat obtenu dans un modèle non-restreint.

C9. Utilisez les données du fichier VOLAT pour cet exercice. La variable *rsp500* correspond au rendement mensuel de l'indice Standard & Poor's 500 (taux annuel, en incluant les variations de prix ainsi que les dividendes.) *i3* correspond au taux d'intérêt des bons du Trésor 3-mois, et *pcip* correspond au pourcentage de variation de la production industrielle (à un taux annuel).

i. Considérez l'équation :

$$rsp500_t = \beta_0 + \beta_1 pcip_t + \beta_2 i3_t + u_t.$$

Quels sont les signes attendus de β_1 et β_2 ?

ii. Estimez l'équation précédente en utilisant la méthode des MCO, et reportez les résultats de la régression. Interprétez le signe et la magnitude des coefficients.

iii. Quelles variables sont statistiquement significatives ?

iv. Est-ce que les résultats de (iii) impliquent que les rendements du S&P 500 sont prévisibles ? Justifiez.

C10. Considérez le modèle estimé en (10.15) ; en utilisant les données du fichier INTDEF.

i. Définissez la corrélation entre *inf* et *def* sur la période de l'échantillon, et commentez.

ii. Ajoutez un retard de *inf* et un retard de *def* à l'équation et présentez les résultats sous la forme habituelle.

iii. Comparez le multiplicateur de long terme estimé avec celui de l'équation (10.15). Sont-ils très différents ?

iv. Les deux retards du modèle sont-ils conjointement significatifs avec un seuil de confiance de 5 % ?

C11. Le fichier TRAFFIC2 contient 108 observations mensuelles concernant le nombre d'accidents de la route, la réglementation routière (et d'autres variables) pour l'État de Californie de janvier 1981 à décembre 1989. Utilisez les données afin de répondre aux questions suivantes.

i. Définissez la date à partir de laquelle la ceinture de sécurité est devenue obligatoire en Californie, ainsi que la date à partir de laquelle la limitation de vitesse sur l'autoroute a augmenté à 65 miles par heure.

ii. Régressez la variable log(*totacc*) sur une tendance temporelle linéaire et 11 variables indicatrices mensuelles, en utilisant janvier comme mois de base. Interprétez le coefficient de la tendance. Pouvez-vous dire qu'il existe une saisonnalité en ce qui concerne le nombre d'accidents de la route ?

iii. Ajoutez à la régression (ii) les variables *wkends*, *unem*, *spdlaw*, et *beltlaw*. Discutez du coefficient de la variable *unem*. Est-ce que le signe et la grandeur du coefficient ont un sens pour vous ?

iv. Dans la régression (iii), interprétez les coefficients des variables *spdlaw* et *beltlaw*. Est-ce que les résultats correspondent à ceux attendus ? Justifiez.

v. La variable *prcfat* correspond au pourcentage d'accidents ayant entraîné au moins un décès. Notez que cette variable est un pourcentage, et non une proportion. Quelle est la moyenne de *prcfat* sur la période ? Est-ce que ce chiffre vous semble-t-il correct ?

vi. Lancez la régression (iii) mais en utilisant *prcfat* comme variable indépendante à la place de log(*totacc*). Discutez des différents effets estimés et de la significativité des variables *spdlaw* et *beltlaw*

C12. i. Estimez l'équation (10.2) en utilisant toutes les données du fichier PHILLIPS, et présentez les résultats. Combien d'observations avez-vous désormais ?

ii. Comparez les estimateurs de (i) avec ceux de l'équation (10.14). En particulier, est-ce que l'ajout d'une année supplémentaire permet d'améliorer le modèle ? Justifiez.

iii. Maintenant, lancez la régression en utilisant seulement les données de 1997 à 2003. Comment varient les estimateurs par rapport à ceux de l'équation (10.14) ? Est-ce que les estimateurs de cette régression utilisant sept années sont assez précis pour en tirer des conclusions définitives ? Justifiez.

iv. Considérons une régression simple sur un échantillon de n périodes, que nous divisons ensuite en deux : une « période de temps 1 » et une « période de temps 2 ». Dans la période de temps 1, nous avons n_1 observations, et dans la période 2, nous avons n_2 observations (avec donc $n_1 + n_2 = n$). En vous appuyant sur les trois premières questions de cet exercice, commentez la phrase suivante : « En général l'estimation de la pente en utilisant n observations est à peu près égale à la moyenne pondérée des estimations de la pente sur les sous-échantillons 1 et 2, où les poids des pondérations sont respectivement n_1/n et n_2/n,) »

C13. Utilisez les données du fichier MINWAGE pour cet exercice. En particulier, utilisez les séries sur l'emploi et les salaires dans le secteur 232 (Accessoires pour hommes et garçons). La variable *gwage232* correspond à la croissance du salaire moyen mensuel dans le secteur 232 (variation du log), *gemp232* représente la croissance de l'emploi dans le secteur 232, *gmwage* la croissance du salaire minimum et *gcpi* la croissance de l'indice des prix à la consommation.

i. Régressez *gwage232* sur *gmwage*, *gcpi*. Est-ce que le signe et l'ampleur des coefficients de la régression vous semblent logiques ? Justifiez. Est-ce que la variable *gmwage* est statistiquement significative ?

ii. Ajoutez les 12 premiers retards de la variable *gmwage* à l'équation de la question (i). Pensez-vous qu'il soit nécessaire d'inclure ces retards afin d'estimer le multiplicateur de long terme permettant de mesurer l'effet de la croissance du salaire minimum sur la croissance des salaires du secteur 232 ? Justifiez.

iii. Régressez *gemp232* sur *gmwage*, *gcpi*. Est-ce que la croissance du salaire minimum semble avoir un effet simultané sur *gemp232* ?

iv. Ajoutez les 12 premiers retards à l'équation de la croissance du taux d'emploi. Est-ce que la croissance du salaire minimum a un impact significatif sur le taux de croissance de l'emploi, que ce soit à court terme ou à long terme ? Justifiez.

C14. Utilisez les données du fichier APPROVAL pour répondre aux questions suivantes. Le fichier comprend des données concernant 78 mois de la présidence de George W. Bush. (Les données se terminent en juillet 2007, avant la fin du mandat de George W. Bush.) En plus des variables économiques et des variables indicatrices permettant d'identifier divers événements, la base de données inclut le taux d'approbation, *approve*, collecté par Gallup. (Il est aussi possible de regarder l'Exercice sur Ordinateur C14 du chapitre 11 pour mieux comprendre les problèmes économétriques relatifs à l'analyse de ces données.)

i. Quelles sont les valeurs possibles de la variable *approve* ? Quelle est la moyenne de cette variable ?

ii. Estimez le modèle

$$approve_t = \beta_0 + \beta_1\, cpifood_t + \beta_2\, lrgasprice_t + \beta_3\, unemploy_t + \mu_t$$

où les deux premières variables sont exprimées sous forme logarithmique, et reportez les résultats de la régression de la manière habituelle.

iii. Interprétez les coefficients de l'estimation de la question (ii). Commentez les signes et les valeurs des effets, ainsi que la significativité statistique.

iv. Ajoutez les variables indicatrices *sep11* et *iraqinvade* à l'équation de la question (ii). Interprétez les coefficients des variables indicatrices. Sont-ils statistiquement significatifs ?

v. Est-ce que l'ajout de variables indicatrices dans la question (iv) a un impact important sur l'estimation du modèle ? Il y a t-il des coefficients de la question (iv) difficile à rationaliser ?

iv. Ajoutez *lsp500* à la régression de la question (iv). En contrôlant des autres facteurs, peut-on dire que l'évolution des marchés financiers à un impact important sur le taux d'approbation du président ?

CHAPITRE 11

UTILISATION DES MCO POUR L'ANALYSE DES SÉRIES TEMPORELLES

Traduction de Alain Durré

Dans le chapitre 10, nous avons discuté des propriétés de l'estimateur des MCO pour l'analyse des séries temporelles en échantillon fini, en évoquant un nombre important d'hypothèses. Si toutes les hypothèses du modèle classique pour l'analyse des séries temporelles sont vérifiées (hypothèses TS.1 à TS.6), l'estimateur des MCO a exactement les mêmes propriétés que celles dérivées dans le cas de données en coupe transversale. Dans cette situation, l'inférence statistique est réalisée exactement de la même manière que ce que nous avons vu précédemment.

Dans le chapitre 5 à propos de l'analyse en coupe instantanée, nous avons vu qu'il existait de bonnes raisons d'étudier les propriétés asymptotiques des MCO. En effet, si les termes d'erreurs ne suivent pas une distribution normale, alors nous devons utiliser le théorème central limite (TCL) pour justifier la validité du *t*-stat et des intervalles de confiance issus des MCO.

L'analyse asymptotique est encore plus importante dans le contexte des séries temporelles. (Ce qui est d'autant plus ironique qu'il est souvent très difficile d'obtenir des échantillons de grande taille pour les séries temporelles ; mais nous n'avons souvent pas d'autres choix que de nous baser sur des approximations asymptotiques). Dans la section 10.3, nous avons vu pourquoi l'hypothèse d'exogénéité (TS.3) peut être violée dans les modèles statiques et les modèles à retards répartis. Comme nous le verrons dans la section 11.2, les modèles contenant des variables dépendantes avec retards violent nécessairement l'hypothèse TS.3.

Malheureusement, l'analyse asymptotique pour les séries temporelles est bien plus complexe et risquée que dans le cas de données en coupe transversale. Dans le chapitre 5, nous avons obtenus les propriétés asymptotiques des MCO dans un contexte d'échantillonnage aléatoire. Mais les choses deviennent plus compliquées lorsque les observations peuvent être corrélées dans le temps. Cependant, la majorité des théorèmes sont toujours valables, tout du moins pour certains processus temporels (mais pas pour tous). La question principale est de savoir si la corrélation entre les variables à différentes périodes de temps tend rapidement vers 0 ou non. Les séries temporelles qui ont une corrélation temporelle substantielle demandent une attention particulière dans l'analyse de régression. Dans ce chapitre, nous allons voir certains de ces problèmes.

11.1 STATIONNARITÉ ET SÉRIES TEMPORELLES FAIBLEMENT DÉPENDANTE

Dans cette section, nous allons présenter les concepts clés nécessaires à l'application des approximations asymptotiques lors de l'analyse de régression de séries temporelles. L'important ici est d'avoir une vision globale de ces problèmes, et non pas d'en connaître tous les détails.

Stationnarité et non-stationnarité des séries temporelles

Historiquement, la notion de processus stationnaire a joué un rôle important dans l'analyse des séries temporelles. Une série temporelle est stationnaire si ses lois de probabilité sont stables dans le temps, au sens suivant : si l'on considère une suite de variables aléatoires à une période donnée, puis que nous décalions cette suite de *h* périodes, alors la probabilité de distribution jointe doit rester la même. Une définition formelle de la stationnarité s'ensuit.

Processus Stochastique Stationnaire. Le processus stochastique $\{x_t : t = 1, 2, \ldots\}$ est *stationnaire* si pour chaque ensemble d'indice temps $1 \leq t_1 < t_2 < \ldots < t_m$, la distribution jointe de $(x_{t1}, x_{t2}, \ldots, x_{tm})$ est la même que la distribution jointe de $(x_{t1+h}, x_{t2+h}, \ldots, x_{tm+h})$ pour tout entier $h \geq 1$.

Cette définition peut sembler abstraite, mais son interprétation est en réalité assez simple. Une implication (en choisissant $m = 1$ et $t_1 = 1$) est que x_t suit la même distribution x_1 pour tout $t = 2, 3, \ldots$ En d'autres

termes, la suite $\{x_t : t = 1, 2, \ldots\}$ est identiquement distribuée. Mais la stationnarité nécessite davantage. Par exemple, la distribution jointe de (x_1, x_2) (les deux premiers termes de la séquence) doit être la même que la distribution jointe de (x_t, x_{t+1}) pour tout $t \geq 1$. De nouveau, cela ne place aucune restriction sur la manière dont x_t et x_{t+1} sont liés l'un à l'autre ; en effet, ils peuvent même être fortement corrélés. La stationnarité requiert simplement que la nature de n'importe quelle corrélation entre des termes adjacents soit la même pour toutes les périodes de temps.

Un processus stochastique qui n'est pas stationnaire est appelé processus non-stationnaire. La stationnarité étant une caractéristique d'un processus stochastique sous-jacent et non pas d'une observation simple, il peut être difficile de déterminer si des données collectées proviennent d'un processus stationnaire ou non. Il est cependant simple de voir que certaines séquences ne sont pas stationnaires. Par exemple, un processus avec une tendance temporelle (voir section 10.5) est clairement non-stationnaire ; au minimum car sa moyenne n'est pas constante dans le temps.

Parfois, une forme faible de stationnarité est suffisante. Si $\{x_t : t = 1, 2, \ldots\}$ est un processus d'ordre deux finis, tel que, $[\mathrm{E}(x_t^2) < \infty]$ pour tout t, alors la définition suivante s'applique.

Complément de 11.1

Supposons que $\{y_t : t = 1, 2, \ldots\}$ soit généré par $y_t = \delta_0 + \delta_1 t + e_t$, où $\delta_1 \neq 0$, et que $\{e_t : t = 1, 2, \ldots\}$ soit i.i.d. de moyenne et de variance σ_e^2. Dans ce cas (i) le processus $\{y_t\}$ est-il stationnaire en covariance ? (ii) Le processus $(y_t - \mathrm{E}(y_t))$ est-il stationnaire en covariance ?

Covariance d'un processus stationnaire. Un processus stochastique $\{x_t : t = 1, 2, \ldots\}$ d'ordre deux finis $[\mathrm{E}(x_t^2) < \infty]$ est stationnaire en covariance si (i) $\mathrm{E}(x_t)$ est constant ; (ii) $\mathrm{Var}(x_t)$ est constante ; et que (iii) pour tout t, $h \geq 1$, $\mathrm{Cov}(x_t, x_{t+h})$ dépend seulement de h (et non de t).

La stationnarité en covariance ne s'intéresse qu'au deux premiers moments d'un processus stochastique : la moyenne et la variance du processus sont constantes dans le temps, et la covariance entre x_t et x_{t+h} ne dépend que de la distance h entre deux termes, et ne dépend donc aucunement de la période de temps initiale. Il s'ensuit donc immédiatement que la corrélation entre x_t et x_{t+h} ne dépend donc aussi que de h.

Si un processus stationnaire a un second moment fini, alors il est stationnaire en covariance ; mais l'inverse n'est pas vrai. Parfois, pour mettre l'accent sur le fait que la stationnarité est une exigence plus forte que la stationnarité en covariance, la première notion est parfois appelée *stationnarité stricte*. Parce que la stationnarité stricte permet de simplifier l'énoncé de certaines hypothèses, le terme « stationnarité » signifiera toujours pour nous « stationnarité stricte ».

Comment la stationnarité est-elle utilisée pour l'analyse des séries temporelles en économétrie ? D'un point de vue technique, la stationnarité simplifie l'expression de la loi des grands nombres (LGN) et du théorème central limite (TCL), bien que nous ne nous formaliserons pas avec cela dans ce chapitre. D'un point de vue pratique, si nous voulons comprendre la relation entre deux variables (ou plus) via une régression, nous devons supposer une certaine stabilité de cette relation dans le temps. Si nous acceptons que la relation entre deux variables (par exemple entre y_t et x_t) change arbitrairement à chaque période de temps, nous ne pouvons pas espérer apprendre beaucoup de cette relation si nous ne disposons que d'une seule série temporelle pour chaque variable.

L'utilisation d'un modèle de régression multiple de séries temporelles implique donc une certaine forme de stationnarité. Par la suite, les hypothèses TS.4 et TS.5 impliqueront une variance des erreurs constante dans le temps, et une corrélation entre les erreurs de deux périodes de temps adjacentes égale à 0.

Série temporelle faiblement dépendante

La stationnarité a à voir avec la distribution jointe d'un processus, car elle change dans le temps. Un concept très différent est celui de faible dépendance, qui place des restrictions sur la force des relations entre les variables aléatoires x_t et x_{t+h} lorsque la distance de temps entre elles, h, devient grand. La notion de faible dépendance est plus facilement explicable dans le cas d'une série temporelle stationnaire. Pour simplifier, une série temporelle stationnaire $\{x_t : t = 1, 2, \ldots\}$ est faiblement dépendante si x_t et x_{t+h} sont « presque indépendants » lorsque h augmente. Un énoncé similaire est vrai si la séquence est non-stationnaire, mais dans ce cas nous devons supposer que le concept de « presque indépendance » ne dépend pas du point de départ, t.

La description de faible dépendance du précédent paragraphe est nécessairement vague. Il n'est pas possible en effet de formaliser une définition, car aucune définition ne peut couvrir tous les cas de figure. Il existe de nombreuses formes spécifiques de faible dépendance, mais cela sort du cadre de cet ouvrage. [Voir White (1984), Hamilton (1994), et Wooldridge (1994b) pour un traitement approfondi de ces concepts]

Dans notre cas, une notion intuitive de la signification de faible dépendance est suffisante. Les séquences stationnaires en covariance peuvent être caractérisées en terme de corrélation : une série temporelle stationnaire en covariance est faiblement dépendante si la corrélation entre x_t et x_{t+h} tend vers 0 « assez rapidement » lorsque $h \to \infty$ (à cause de la stationnarité en covariance, la corrélation ne dépend pas du point de départ, t.) En d'autres termes, lorsque les variables sont de plus en plus espacées dans le temps, la corrélation entre elles devient de plus en plus petite. Les séquences de stationnarité en covariance où $\text{Corr}(x_t, x_{t+h}) \to 0$ lorsque $h \to \infty$ sont dites asymptotiquement non-corrélées. Intuitivement, c'est de cette manière que nous caractérisons usuellement la faible dépendance. Techniquement, nous devons faire l'hypothèse que la corrélation converge vers 0 assez rapidement, mais nous passerons rapidement sur ce point.

Pourquoi la faible dépendance est-elle importante pour l'analyse de régression ? Principalement car cela permet de remplacer l'hypothèse d'échantillonnage aléatoire en supposant à la place que la loi des grands nombres et le théorème central limite sont vérifiés. Le théorème central limite le plus connu pour les séries temporelles requiert la stationnarité et une certaine forme de faible dépendance : dans cette situation, les séries temporelles stationnaires de faible dépendance sont donc idéales dans l'analyse de régression multiple. Dans la section 11.2, nous démontrerons pourquoi la méthode des MCO peut-être justifiée d'une manière assez générale en faisant appel à la loi des grands nombres et au théorème central limite. Les séries temporelles qui ne sont pas faiblement dépendantes – dont des exemples seront donnés en section 11.3 – ne vérifient pas généralement le théorème central limite, et c'est pourquoi leur utilisation dans un modèle de régression multiple peut-être délicate.

L'exemple le plus simple de faible dépendance d'une série temporelle est une séquence indépendante et identiquement distribuée : une séquence qui est indépendante étant trivialement de faible dépendance. Un exemple plus intéressant d'une séquence de faible dépendance est

$$x_t = e_t + \alpha_1 e_{t-1}, \; t = 1, 2, \ldots, \qquad \textbf{[11.1]}$$

où $\{e_t : t = 0, 1, \ldots\}$ est une suite i.i.d. de moyenne nulle et de variance σ_e^2. Le processus $\{x_t\}$ est appelé processus à moyenne mobile d'ordre 1 [MA(1)] : x_t est une moyenne pondérée de e_t et e_{t-1} ; la période suivante, nous enlevons e_{t-1}, et alors x_{t+1} dépend de e_{t+1} et e_t. En définissant le coefficient e_t comme étant égal à 1 dans (11.1), nous ne perdons pas le caractère générique du modèle. [Dans l'équation (11.1), nous utilisons x_t et e_t comme des notations génériques d'un processus temporel]

Pourquoi un processus MA(1) est de faible dépendance ? Les termes adjacents sont corrélés parce que $x_{t+1} = e_{t+1} + \alpha_1 e_t$, $\text{Cov}(x_t, x_{t+1}) = \alpha_1 \text{Var}(e_t) = \alpha_1 \sigma_e^2$ et $\text{Var}(x_t) = (1 + \alpha_1^2)\sigma_e^2$, $\text{Corr}(x_t, x_{t+1}) = \alpha_1 / (1 + \alpha_1^2)$. Cependant, lorsque nous regardons les variables qui sont séparées par deux périodes de temps ou plus, elles ne sont pas corrélées (car elles sont indépendantes). Par exemple, $x_{t+2} = e_{t+2} + \alpha_1 e_{t+1}$ est indépendant de x_t car

$\{e_t\}$ est indépendant pour différents t. Étant donné l'hypothèse de distribution identique de e_t, $\{x_t\}$ dans (11.1) est en fait stationnaire. Ainsi, un processus MA(1) est stationnaire, de faible dépendance, et la LGN et le TCL peuvent s'appliquer à $\{x_t\}$.

Un exemple plus célèbre est le processus

$$y_t = \rho_1 y_{t-1} + e_t,\ t = 1, 2, \ldots \qquad \textbf{[11.2]}$$

Le point de départ de la série est y_0 ($t = 0$), et $\{e_t : t = 1, 2, \ldots\}$ est une série i.i.d. de moyenne nulle et de variance σ_e^2. Nous supposons de plus que e_t sont indépendants de y_0 et que $E(y_0) = 0$. Ce processus est appelé processus autorégressif d'ordre 1 [AR(1)].

L'hypothèse cruciale de faible dépendance d'un processus AR(1) est la condition de stabilité $|\rho_1| < 1$. Nous disons alors que $\{y_t\}$ est un processus AR(1) stable.

Pour voir qu'un processus AR(1) stable est asymptotiquement non corrélé, il est utile de supposer que le processus est stationnaire en covariance (en réalité, il peut généralement être démontré que $\{y_t\}$ est strictement stationnaire, mais la preuve est quelque peu technique.) Dans ce cas, $E(y_t) = E(y_{t-1})$, et à partir de (11.2) avec $\rho_1 \neq 1$, ceci est vrai uniquement si $E(y_t) = 0$. En considérant la variance de (11.2) et en utilisant le fait que e_t et y_{t-1} sont indépendants (et donc non-corrélé) $\text{Var}(y_t) = \rho_1^2 \text{Var}(y_{t-1}) + \text{Var}(e_t)$, et sous la condition de stationnarité en covariance, nous devons avoir $\sigma_y^2 = \rho_1^2 \sigma_y^2 + \sigma_e^2$. Comme $\rho_1^2 < 1$, et en utilisant la condition de stabilité, nous pouvons résoudre cela simplement pour σ_y^2 :

$$\sigma_y^2 = \sigma_e^2 / (1 - \rho_1^2). \qquad \textbf{[11.3]}$$

À partir de là, nous pouvons trouver la covariance entre y_t et y_{t+h} pour $h \geq 1$. En utilisant des substitutions successives,

$$\begin{aligned} y_{t+h} &= \rho_1 y_{t+h-1} + e_{t+h} = \rho_1(\rho_1 y_{t+h-2} + e_{t+h-1}) + e_{t+h} \\ &= \rho_1^2 y_{t+h-2} + \rho_1 e_{t+h-1} + e_{t+h} = \ldots \\ &= \rho_1^2 y_t + \rho_1^{h-1} e_{t+1} + \ldots + \rho_1 e_{t+h-1} + e_{t+h}. \end{aligned}$$

Étant donné que $E(y_t) = 0$ pour tout t, nous pouvons multiplier cette dernière équation par y_t, puis considérer l'espérance pour obtenir $\text{Cov}(y_t, y_{t+h})$. En utilisant le fait que e_{t+j} n'est pas corrélé avec y_t pour tout $j \geq 1$, nous avons

$$\begin{aligned} \text{Cov}(y_t, y_{t+h}) &= E(y_t y_{t+h}) = \rho_1^h E(y_t^2) + \rho_1^{h-1} E(y_t e_{t+1}) + \ldots + E(y_t e_{t+h}) \\ &= \rho_1^h E(y_t^2) = \rho_1^h \sigma_y^2. \end{aligned}$$

Parce que σ_y est l'écart-type de y_t et de y_{t+h}, nous pouvons facilement trouver la corrélation entre y_t et y_{t+h} pour tout $h \geq 1$:

$$\text{Corr}(y_t, y_{t+h}) = \text{Cov}(y_t, y_{t+h}) / (\sigma_y \sigma_y) = \rho_1^h. \qquad \textbf{[11.4]}$$

En particulier, $\text{Corr}(y_t, y_{t+1}) = \rho_1$, donc ρ_1 est égal à la corrélation entre deux termes adjacents quelconque de la série.

L'équation (11.4) est importante car elle montre bien que y_t et y_{t+h} sont corrélés pour $h \geq 1$, et que cette corrélation devient très faible lorsque h augmente étant donné que $|\rho_1| < 1, \rho_1^h \to 0$ lorsque $h \to \infty$. Même lorsque ρ_1 est grand – par exemple égal à 0,9, ce qui implique une corrélation positive très forte entre les termes adjacents – la corrélation entre y_t et y_{t+h} tend vers 0 assez rapidement. Par exemple $\text{Corr}(y_t, y_{t+5}) = 0{,}591$, $\text{Corr}(y_t, y_{t+10}) = 0{,}349$, et $\text{Corr}(y_t, y_{t+20}) = 0{,}122$. Si les indices t représentent des années, cela veut donc dire que la corrélation entre deux « y » séparés de 20 ans est d'environ 0,122. Lorsque ρ_1 est plus faible, la corrélation disparaît beaucoup plus rapidement (Vous pouvez essayer de vérifier cela avec $\rho_1 = 0{,}5$).

Cette analyse heuristique montre qu'un processus AR(1) stable est de faible dépendance. Un modèle AR(1) est d'une importance toute particulière dans l'analyse de régression multiple des séries temporelles. Nous couvrirons cela plus en détail dans le chapitre 12, et verrons l'utilité de ces processus pour la prévision dans le chapitre 18.

Il existe de nombreux autres types de séries temporelles de faible dépendance, comme par exemple le modèle autorégressif hybride et les processus de moyenne mobile. Mais les exemples précédents sont parfaitement adaptés à nos besoins et seront suffisants dans le cadre de ce chapitre.

Avant de terminer cette section, nous devons mettre l'accent sur un point qui entraîne souvent des confusions en ce qui concerne l'économétrie des séries temporelles. Une série ayant une tendance, bien que très certainement non-stationnaire, peut être de faible dépendance. En réalité, dans le modèle simple à tendance linéaire du chapitre 10 [voir équation (10.24)], la série $\{y_t\}$ est indépendante. Une série qui est stationnaire autour de sa tendance, et de faible dépendance, est souvent appelée *processus à tendance stationnaire*. (Notez que le nom n'est pas totalement descriptif, car nous supposons la faible dépendance en plus de la stationnarité). Des processus de ce type peuvent être utilisés dans des régressions tout comme dans le chapitre 10, à partir du moment où une variable de tendance appropriée est incorporée au modèle.

11.2 PROPRIÉTÉS ASYMPTOTIQUES DES MCO

Dans le chapitre 10, nous avons vu quelques cas pour lesquels les hypothèses classiques du modèle linéaire n'étaient pas vérifiées. Dans ces situations, nous devons faire appel aux propriétés asymptotiques des MCO, tout comme pour l'analyse des données en coupe transversale. Dans cette section, nous exposons les hypothèses et les résultats principaux qui permettent de justifier l'utilisation des MCO. Les démonstrations des théorèmes de ce chapitre sont quelque peu difficiles, et donc omises. Voir Wooldridge (1994b) pour un complément d'information.

Hypothèse TS.1′
Linéarité et faible dépendance

Nous supposons que le modèle est exactement le même que pour l'hypothèse TS.1, mais nous ajoutons l'hypothèse que $\{(\mathbf{x}_t, y_t) : t = 1, 2, \ldots\}$ est stationnaire et de faible dépendance. En particulier, la LGN et le TCL s'appliquent aux moyennes de l'échantillon.

La linéarité des paramètres signifie à nouveau que nous pouvons écrire le modèle sous la forme :

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \qquad \textbf{[11.5]}$$

où β_j sont les paramètres à estimer. Contrairement au chapitre 10, les x_{tj} peuvent inclure des retards de la variable dépendante, en plus des retards des variables explicatives.

Nous avons inclus l'hypothèse de stationnarité de TS.1′ afin de permettre une plus grande facilité dans l'interprétation des différentes hypothèses. Si nous utilisions avec attention les propriétés asymptotiques des MCO, comme nous le faisons dans l'annexe E, la stationnarité permettrait aussi de simplifier ces dérivées. Mais la stationnarité n'est pas critique pour que les MCO aient les propriétés asymptotiques standard. (Comme mentionné dans la section 11.1, en supposant que β_j sont constants dans le temps, nous supposons déjà une forme de stabilité de la distribution dans le temps). La restriction additionnelle importante de l'hypothèse TS.1′ par rapport à l'hypothèse TS.1 est l'hypothèse de faible dépendance. Dans la section 11.1, nous avons discuté assez longuement de la faible dépendance car ce n'est en aucun cas une

hypothèse anodine. Techniquement, l'hypothèse TS.1′ nécessite une dépendance faible de multiples séries temporelles (y_t ainsi que les éléments de $\mathbf{x}_t$), qui entraine l'ajout de restrictions sur la distribution jointe dans le temps. Les détails ne sont pas particulièrement important et sortent en tout cas du cadre de cet ouvrage ; voir Wooldridge (1994). Il est plus important de comprendre quel type de processus temporels persistants viole l'hypothèse de faible dépendance, ce que nous développerons dans la section suivante. Nous étudierons également l'utilisation de processus de ce type dans les modèles de régression multiple.

Naturellement, nous excluons toujours la parfaite colinéarité.

Hypothèse TS.2′
Absence de parfaite colinéarité

Même hypothèse que TS.2.

Hypothèse TS.3′
Moyenne conditionnelle nulle

Les variables explicatives $\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tk})$ sont simultanément exogènes comme dans l'équation (10.10) : $\mathrm{E}(u_t|\mathbf{x}_t) = 0$.

Ceci est l'hypothèse la plus naturelle concernant la relation entre u_t et les variables explicatives. Cette hypothèse est bien plus faible que l'hypothèse TS.3, car elle ne pose pas de restriction sur la manière dont u_t est relié aux variables explicatives à différentes périodes de temps. Nous verrons des exemples vérifiant TS.3′ juste après. Par stationnarité, nous postulons que si l'exogénéité simultanée est vérifiée pour une période de temps, cela est vrai aussi pour toutes les périodes. Supprimer la stationnarité nous demanderait simplement de vérifier que la condition est vraie pour tout $t = 1, 2, \ldots$

Dans certains cas, il est utile de savoir que la consistance des résultats suivants demande simplement que u_t ait une moyenne non conditionnelle égale à 0 et ne soit pas corrélé avec chacun des x_{tj} :

$$\mathrm{E}(u_t) = 0, \ \mathrm{Cov}(x_{tj}, u_t) = 0, \ j = 1, \ldots, k. \qquad \textbf{[11.6]}$$

Nous travaillerons principalement avec l'hypothèse de moyenne conditionnelle nulle, car cela permet d'obtenir une analyse asymptotique simple.

Théorème 11.1
Consistance des MCO

Sous les hypothèses TS.1′, TS.2′, et TS.3′, les estimateurs des MCO sont consistants : $\operatorname{plim} \hat{\beta}_j = \beta_j, j = 0, 1, \ldots, k$.

Il existe des différences pratiques importantes entre les théorèmes 10.1 et 11.1. Premièrement dans le théorème 11.1, nous avons conclu que les estimateurs des MCO sont consistants, mais pas nécessairement sans biais. Deuxièmement, dans le théorème 11.1, nous avons affaibli le sens dans lequel les variables explicatives doivent être exogènes, mais une faible dépendance est requise à propos des séries temporelles sous-jacentes. La faible dépendance est aussi cruciale pour obtenir une approximation des résultats de la distribution, ce que nous verrons par la suite.

EXEMPLE 11.1
Le modèle statique

Considérons un modèle statique avec deux variables explicatives :

$$y_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + u_t. \qquad [11.7]$$

Sous l'hypothèse de faible dépendance, la condition suffisante pour que les MCO soient consistants est :

$$E(u_t|z_{t1}, z_{t2}) = 0. \qquad [11.8]$$

Cela exclut donc que les variables omises qui sont dans u_t soient corrélées soit avec z_{t1} ou avec z_{t2}. De plus, aucune fonction de z_{t1} ou z_{t2} ne peut être corrélée avec u_t, et donc l'hypothèse TS.3′ exclut le problème de mauvaise spécification de forme fonctionnelle, comme dans les cas de la coupe transversale. D'autres problèmes, comme les erreurs de mesures des variables z_{t1} ou z_{t2}, peuvent faire que (11.8) n'est pas vérifiée.
Il est important de voir que l'hypothèse TS.3′ *ne supprime pas* la corrélation entre, par exemple, u_{t-1} et z_{t1}. Ce type de corrélation peut être présente si z_{t1} est relié à la valeur passée y_{t-1}, de telle sorte que

$$z_{t1} = \delta_0 + \delta_1 y_{t-1} + v_t. \qquad [11.9]$$

Par exemple, z_{t1} peut être une variable de politique monétaire, telle que le changement mensuel de l'offre de monnaie, et peut dépendre par exemple du taux d'inflation du mois précédent (y_{t-1}). Ce type de mécanisme entraîne généralement une corrélation entre z_{t1} et u_{t-1}. Ce type de réaction est acceptée sous l'hypothèse TS.3′.

EXEMPLE 11.2
Modèle à retards échelonnés finis

Dans un modèle à retards échelonnés finis :

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t, \qquad [11.10]$$

une hypothèse naturelle est que l'espérance de u_t, étant donné les valeurs courantes et passées de z, est nulle :

$$E(u_t|z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots) = 0. \qquad [11.11]$$

Cela signifie que, une fois que z_t, z_{t-1}, et z_{t-2} ont été inclus, aucun autre retard de z n'affectera $E(y_t|z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots)$; si cela n'était pas vrai, nous devrions rajouter d'autres retards dans notre équation. Par exemple, y_t peut mesurer le taux de variation annuel de l'investissement et z_t peut être une mesure du taux d'intérêt durant l'année t. Lorsque nous fixons $x_t = (z_t, z_{t-1}, z_{t-2})$, l'hypothèse TS.3′ est alors vérifiée et les MCO sont donc consistants. Tout comme dans l'exemple précédent, TS.3′ ne supprime pas une potentielle réaction de y aux valeurs futurs z.

Les deux exemples précédents ne requièrent pas nécessairement la théorie asymptotique, car les variables explicatives peuvent être strictement exogènes. L'exemple suivant viole clairement l'hypothèse d'exogénéité : dans cette situation, nous devons utiliser les propriétés asymptotiques des MCO

EXEMPLE 11.3
Modèle AR(1)

Considérons le modèle AR(1),

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t, \qquad [11.12]$$

où u_t a une espérance nulle, compte tenu de toutes les valeurs passées de y :

$$E(u_t|y_{t-1}, y_{t-2}, \ldots) = 0. \quad \textbf{[11.13]}$$

En combinant ces deux équations, nous avons :

$$E(y_t|y_{t-1}, y_{t-2}, \ldots) = E(y_t|y_{t-1}) = \beta_0 + \beta_1 y_{t-1}. \quad \textbf{[11.14]}$$

Ce résultat est très important. Premièrement, cela signifie que, après avoir contrôlé la présence d'un retard de y, aucun autre retard ne doit affecter l'espérance de y_t. (d'où l'origine de « premier ordre »). Deuxièmement, cette relation est supposée linéaire.

Parce que $\mathbf{x}_t$ ne contient que y_{t-1}, dans l'équation (11.13) cela implique donc que l'hypothèse TS.3′ est vérifiée. À l'opposée, l'hypothèse de stricte exogénéité nécessaire aux estimateurs sans biais (hypothèse TS.3) n'est pas vérifiée. Étant donné que les variables aléatoires pour toutes les périodes de temps incluent toutes les valeurs de y excepté la dernière, $(y_0, y_1, \ldots, y_{n-1})$, l'hypothèse TS.3 requiert que, pour tout t, u_t ne soit pas corrélé avec chacun des $y_0, y_1, \ldots, y_{n-1}$. Ceci ne peut pas être vrai. En réalité, et car u_t n'est pas corrélé avec y_{t-1} selon l'équation (11.13), u_t et y_t doivent être corrélés. Il est d'ailleurs facile de voir $\text{Cov}(y_t, u_t) = \text{Var}(u_t) > 0$. Ainsi, un modèle avec une variable dépendante retardée ne peut pas satisfaire l'hypothèse de stricte exogénéité TS.3.

Pour que la condition de faible dépendance soit vérifiée, nous devons faire l'hypothèse que $|\beta_1| < 1$, comme discuté dans la section 11.1. Si cette condition est vérifiée, alors le théorème 11.1 implique que les estimateurs MCO issus de la régression de y_t sur y_{t-1} produisent des estimateurs consistants de β_0 et β_1. Malheureusement, $\hat{\beta}_1$ est biaisé, et ce biais peut-être large pour les échantillons de petite taille ou si β_1 est proche de 1. (Pour β_1 proche de 1, $\hat{\beta}_1$ peut avoir un biais à la baisse important). Pour les échantillons de taille moyenne à large, $\hat{\beta}_1$ devrait être un bon estimateur de β_1.

Lorsque nous utilisons la procédure d'inférence standard, nous devons imposer les hypothèses d'homoscédasticité et d'absence de corrélation. Ces hypothèses sont moins restrictives que les contreparties du modèle linéaire classique vues dans le chapitre 10.

Hypothèse TS.4′
Homoscédasticité

Les erreurs sont simultanément homoscédastiques, c'est-à-dire, $\text{Var}(u_t|\mathbf{x}_t) = \sigma^2$.

Hypothèse TS.5′
Absence d'autocorrélation

Pour tout $t \neq s$, $E(u_t u_s|\mathbf{x}_t, \mathbf{x}_s) = 0$.

Dans l'hypothèse TS.4′, il est important de noter que nous conditionnons les variables explicatives seulement à la période t (contrairement à TS.4). Dans l'hypothèse TS.5′, nous conditionnons les variables explicatives uniquement aux périodes de temps coïncidant avec u_t et u_s. Comme indiqué, cette hypothèse est un peu difficile à interpréter, mais est une condition pour étudier les propriétés asymptotiques des MCO dans une large gamme de régression de séries temporelles. Lorsque nous considérons TS.5′, nous ignorons souvent le conditionnement de $\mathbf{x}_t$ et $\mathbf{x}_s$, et nous réfléchissons plutôt au fait que u_t et u_s ne soient pas corrélé, pour tout $t \neq s$.

La corrélation sérielle est souvent un problème pour les modèles statiques et les modèles à retards répartis finis : rien ne garantit en effet que les u_t non-observables ne sont pas corrélés dans le temps. Il est important de noter que l'hypothèse TS.5′ est vérifiée dans un modèle AR(1), comme indiqué par les équations (11.12) et (11.13). La variable explicative à la période t étant y_{t-1}, nous devons alors montrer que $E(u_t u_s | y_{t-1}, y_{s-1}) = 0$ pour tout $t \neq s$. Pour vérifier cela, supposons $s < t$ (l'autre cas fonctionne aussi par symétrie). Alors, comme $u_s = y_s - \beta_0 - \beta_1 y_{s-1}$, u_s est une fonction y pour les périodes avant t. Mais selon (11.13), $E(u_t | u_s, y_{t-1}, y_{s-1}) = 0$, et donc $E(u_t u_s | u_s, y_{t-1}, y_{s-1}) = u_s E(u_t | y_{t-1}, y_{s-1}) = 0$. En utilisant la loi des espérances itérées (voir annexe B), $E(u_t u_s | y_{t-1}, y_{s-1}) = 0$. Ceci est très important : à partir du moment où un retard est inclus dans l'équation (11.12), les erreurs doivent être sériellement non-corrélées. Nous discuterons de cette caractéristique des modèles dynamiques plus en profondeur dans la section 11.4.

Nous obtenons maintenant un résultat asymptotique pratiquement identique à celui de la coupe transversale.

Théorème 11.2
Normalité asymptotique des MCO

Sous les hypothèses TS.1′ à TS.5′, les estimateurs MCO sont asymptotiquement normalement distribués. De plus, les écarts-types, t-stat, F-stat et LM-stat issus des MCO sont asymptotiquement valides.

Ce théorème apporte une justification additionnelle à certains exemples du chapitre 10 : bien que les hypothèses du modèle classique ne sont pas vérifiés, les MCO sont tout de même consistants, et les procédures d'inférences usuelles sont valides. Bien sûr, cela dépend du fait que les hypothèses TS.1′ à TS.5′ soient effectivement vérifiées. Dans la section suivante, nous allons discuter de cas où l'hypothèse de faible dépendance n'est pas vérifiée. Les problèmes de corrélation sérielle et d'hétéroscédasticité seront traités dans le chapitre 12.

EXEMPLE 11.4
L'hypothèse d'efficience des marchés

Nous pouvons utiliser l'analyse asymptotique pour tester une version de l'hypothèse d'efficience des marchés (HEM). Soit y_t le taux de rendement (du mercredi au mercredi) d'un indice composite du NYSE. Une forme stricte de la HEM indique que l'information disponible sur le marché à une date antérieure à la semaine t ne doit pas permettre de prévoir le rendement pour la semaine t. En se basant uniquement sur l'information passée, l'hypothèse d'efficience des marchés indique que :

$$E(y_t | y_{t-1}, y_{t-2}, \ldots) = E(y_t). \qquad \textbf{[11.15]}$$

Si (11.15) est faux, alors il est possible d'utiliser l'information sur les rendements passés pour prévoir les rendements présents. L'hypothèse d'efficience des marchés suppose que si un investissement de ce type existait, alors il disparaitrait instantanément et donc ne serait en réalité pas exploitable.

Une façon simple de tester (11.15) est de spécifier un modèle AR(1) (11.12) comme modèle alternatif. Alors, l'hypothèse nulle peut s'écrire $H_0 : \beta_1 = 0$. Sous l'hypothèse nulle, l'hypothèse TS.3′ est vérifiée en utilisant (11.15), et, comme discuté précédemment, la corrélation sérielle n'est pas alors un problème. L'hypothèse d'homoscédasticité est $\text{Var}(y_t | y_{t-1}) = \text{Var}(y_t) = \sigma^2$, que nous supposons vraie pour le moment. Sous l'hypothèse nulle, les rendements sont sériellement non-corrélés, et donc nous pouvons sans risque affirmer qu'ils sont faiblement dépendants. Ainsi, nous pouvons utiliser le théorème 11.2 pour justifier l'utilisation des estimateurs MCO classiques pour $\hat{\beta}_1$, afin de tester $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$.

Les rendements hebdomadaires du fichier NYSE sont calculés à partir de données allant de janvier 1976 à mars 1989. Dans les rares cas où le mercredi se trouve être un jour férié, le cours de fermeture du jour suivant est utilisé pour calculer le rendement hebdomadaire. Le rendement hebdomadaire moyen sur la période est de 0,196 %, allant d'un maximum hebdomadaire de 8,45 % à un plus bas de –15,32 % durant le crack boursier d'octobre 1987. L'estimation du modèle AR(1) nous donne :

$$\widehat{return}_t = 0{,}180 + 0{,}059\,return_{t-1}$$
$$(0{,}081) \quad (0{,}038) \qquad \textbf{[11.16]}$$
$$n = 689,\ R^2 = 0{,}0035,\ \overline{R}^2 = 0{,}0020.$$

Le *t*-stat du coefficient de la variable $return_{t-1}$ est d'environ 1,55, et donc $H_0 : \beta_1 = 0$ ne peut pas être rejeté pour un seuil de confiance de 10 %. Cette estimation suggère donc une faible corrélation positive entre les rendements du NYSE d'une semaine à l'autre, mais cette relation n'est pas assez forte pour rejeter la HEM.

Dans l'exemple précédent, l'utilisation d'un modèle AR(1) pour tester l'hypothèse d'efficience des marchés ne permet pas de détecter les corrélations entre les rendements hebdomadaires espacés de plus d'une semaine. Il est facile d'estimer le modèle avec davantage de retards. Par exemple, un modèle autorégressif d'ordre deux, où modèle AR(2) model, s'écrit

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + u_t$$
$$\mathrm{E}(u_t | y_{t-1}, y_{t-2}, \ldots) = 0. \qquad \textbf{[11.17]}$$

Il existe des conditions de stabilité concernant β_1 et β_2 qui sont nécessaires afin de s'assurer que le processus AR(2) est de faible dépendance, mais ce n'est pas un problème ici car l'hypothèse nulle indique que l'hypothèse d'efficience des marchés est vérifiée.

$$H_0 : \beta_1 = \beta_2 = 0. \qquad \textbf{[11.18]}$$

Si nous ajoutons l'hypothèse d'homoscédasticité $\mathrm{Var}(u_t | y_{t-1}, y_{t-2}) = \sigma^2$, nous pouvons utiliser un *F*-stat standard pour tester (11.18). Si nous estimons un modèle AR(2) pour les rendements $return_t$, nous obtenons

$$\widehat{return}_t = 0{,}186 + 0{,}060\,return_{t-1} - 0{,}038\,return_{t-2}$$
$$(0{,}081) \ (0{,}038) \qquad (0{,}038)$$
$$n = 688,\ R^2\ 0{,}0048,\ \overline{R}^2 = 0{,}0019$$

(nous perdons alors une observation supplémentaire car nous ajoutons un retard à l'équation). Les deux retards sont individuellement non significatifs à un seuil de confiance de 10 %. Ils sont aussi conjointement non significatifs : en utilisant $R^2 = 0{,}0048$, nous pouvons estimer que le *F*-stat est d'environ $F = 1{,}65$; la *p*-value du *F* stat (avec 2 et 685 degrés de liberté) étant d'environ 0,193. Ainsi, nous ne pouvons rejeter (11.18) et ce même avec un seuil de confiance de 15 %.

EXEMPLE 11.5
Courbe de Phillips augmentée des anticipations

Une version linéaire d'une courbe de Phillips augmentée des anticipations peut s'écrire sous la forme

$$inf_t - inf_t^e = \beta_1 (unem_t - \mu_0) + e_t,$$

où μ_0 est le taux naturel de chômage et inf_t^e est le taux anticipé d'inflation formée en $t - 1$. Ce modèle suppose que le taux naturel de chômage est constant, ce que certains macro-économistes remettent en cause. La différence entre le taux de chômage actuel et le taux de chômage naturel est appelée chômage cyclique, tandis que

la différence entre l'inflation actuelle et l'inflation anticipée est appelée inflation non-anticipée. Le terme d'erreur, e_t, est appelé par les économistes « choc d'offre ». S'il existe un arbitrage entre inflation non anticipée et taux de chômage cyclique, alors $\beta_1 < 0$. [Pour une discussion détaillée concernant la courbe de Phillips augmentée des anticipations, voir Mankiw (1994, section 11.2).]

Pour compléter ce modèle, nous devons émettre des hypothèses en ce qui concerne les anticipations d'inflation. Selon les anticipations adaptives, la valeur attendue de l'inflation ne dépend que de l'inflation récente observée. Une formulation simpliste consiste à écrire que l'inflation attendue pour cette année est égale à l'inflation de l'année précédente : $inf_t^e = inf_{t-1}$ (voir Section 18.1 pour une formulation alternative des anticipations adaptatives). Sous cette condition, nous pouvons écrire :

$$inf_t - inf_{t-1} = \beta_0 + \beta_1 unem_t + e_t$$
ou
$$\Delta inf_t = \beta_0 + \beta_1 unem_t + e_t$$

où ($\Delta inf_t = inf_t - inf_{t-1}$ et $\beta_0 = -\beta_1 \mu_0$ doit-être positif, étant donné que $\beta_1 < 0$ et $\mu_0 > 0$.) Ainsi, si les anticipations sont adaptives, alors la courbe de Phillips augmentée des anticipations relie la variation de l'inflation au taux de chômage et à un choc d'offre, e_t. Si e_t n'est pas corrélé avec $unem_t$, comme cela est en général supposé, alors nous avons des estimateurs consistants, et, en utilisant les MCO, nous pouvons estimé β_0 et β_1 (Nous n'avons pas besoin de supposer que, par exemple, le taux de chômage futur n'est pas affecté par le choc d'offre présent). Nous supposons que TS.1′ à TS.5′ sont vérifiées. En utilisant les données du fichier PHILLIPS jusqu'à l'année 1996, nous estimons alors :

$$\widehat{\Delta inf_t} = 3{,}03 - 0{,}543\, unem_t$$
$$(1{,}38)\quad (0{,}230) \qquad \textbf{[11.19]}$$
$$n = 48,\ R^2 = 0{,}108,\ \overline{R}^2 = 0{,}088.$$

L'arbitrage entre chômage cyclique et inflation non anticipée est reflété dans l'équation (11.19) : une augmentation d'un point de *unem* diminue l'inflation non-anticipée d'environ un demi-point. Cet effet est statistiquement significatif (p-value $\approx 0{,}023$). Ces résultats permettent de contraster les résultats dans le cas d'une courbe de Phillips statique (exemple 10.1), où nous avions trouvé une faible relation positive entre l'inflation et le taux de chômage.

Comme nous pouvons écrire le taux naturel de chômage sous la forme $\mu_0 = \beta_0/(-\beta_1)$, nous pouvons utiliser (11.19) pour obtenir notre propre estimation de ce taux de chômage naturel : $\hat{\mu}_0 = \hat{\beta}_0 / (-\hat{\beta}_1) = 3{,}03 / 0{,}543 \approx 5{,}58$. De cette façon, nous estimons dans ce cas que le taux de chômage naturel est d'environ 5,6 %, ce qui est cohérent avec le consensus des économistes qui estiment que, historiquement, le taux de chômage naturel se situe entre 5 % et 6 %. L'écart-type de cet estimateur est difficile à obtenir car nous avons une fonction non-linéaire des estimateurs MCO. Wooldridge (2010, chapitre 3) traite de la théorie en cas de fonction générale non-linéaire. Dans notre application, l'écart-type est de 0,657, ce qui, en considérant un intervalle de confiance de 95 % (et en se basant sur une distribution normale), entraîne un intervalle de 4,29 % à 6,87 % en ce qui concerne le taux de chômage naturel.

Sous les hypothèses TS.1′ à TS.5′, nous pouvons montrer que les estimateurs des MCO sont asymptotiquement efficients, de la même manière que démontré précédemment dans le théorème 5.3, mais en remplaçant les observations en coupe transversale i par un indice temporel t. Enfin, les modèle ayant des variables explicatives avec tendance peuvent effectivement satisfaire les hypothèses TS.1′ à TS.5′, à condition que les variables soient stationnaires en tendance. Aussi longtemps que les tendances sont intégrées dans l'équation si cela est nécessaire, la procédure d'inférence usuelle est asymptotiquement valide.

Suite 11.2

Supposons que les anticipations s'écrivent sous la forme $inf_t^e = (1/2)inf_{t-1} + (1/2)inf_{t-2}$. Quelle régression effectueriez-vous pour estimer la courbe de Phillips augmentée des anticipations ?

11.3 UTILISATION DE SÉRIES TEMPORELLES HAUTEMENT PERSISTANTES DANS L'ANALYSE DE RÉGRESSION

La section précédente a montré que, à partir du moment où les séries temporelles utilisées sont de faible dépendance, les procédures d'inférence usuelles des MCO sont valides sous des conditions plus faibles que sous les hypothèses du modèle linéaire classique. Malheureusement, de nombreuses séries temporelles économiques ne peuvent pas être caractérisée par une faible dépendance. L'utilisation de séries temporelles avec une forte dépendance ne pose pas de problème pour l'analyse de régression, si les conditions du modèle linéaire classique du chapitre 10 sont vérifiées. Cependant, les procédures usuelles d'inférence ont de fortes chances de violer ces conditions lorsque les données ne sont pas de faible dépendance, car nous ne pouvons pas faire appel à la LGN ou au TCL. Dans cette section, nous donnons des exemples de séries temporelles hautement persistantes (ou fortement dépendantes), et montrons comment il est possible de transformer ces séries pour les utiliser dans l'analyse de régression.

Séries temporelles hautement persistantes

Dans le modèle AR(1) simple de l'équation (11.2), une hypothèse cruciale des séries est d'être de faible dépendance $|\rho_1| < 1$. Il s'avère que de nombreuses séries temporelles économiques sont mieux caractérisées par un processus AR(1) avec $\rho_1 = 1$. Dans cette situation, nous pouvons écrire

$$y_t = y_{t-1} + e_t, \; t = 1, 2, \ldots, \qquad \textbf{[11.20]}$$

où, de nouveau, nous supposons que $\{e_t : t = 1, 2, \ldots\}$ est indépendant et identiquement distribué, de moyenne nulle et de variance σ_e^2. Nous supposons que la valeur initiale, y_0, est indépendant de e_t pour tout $t \geq 1$.

Le processus (11.20) est appelé marche aléatoire. Ce nom vient du fait que y à la période t est obtenu en partant de la valeur précédente, y_{t-1}, puis en ajoutant à cette valeur une variable aléatoire indépendante de y_{t-1} et de moyenne nulle. Parfois, une marche aléatoire est définie différemment, en supposant différentes propriétés au niveau de la variable e_t (comme par exemple l'absence de corrélation plutôt que l'indépendance), mais la définition ci-dessus est suffisante pour nos besoins.

Premièrement, nous devons calculer l'espérance de y_t. La méthode la plus simple consiste à utiliser des substitutions répétées, pour obtenir :

$$y_t = e_t + e_{t-1} + \ldots + e_1 + y_0.$$

En prenant l'espérance des deux côtés de l'équation, nous obtenons

$$\begin{aligned} \mathrm{E}(y_t) &= \mathrm{E}(e_t) + \mathrm{E}(e_{t-1}) + \ldots + \mathrm{E}(e_1) + \mathrm{E}(y_0) \\ &= \mathrm{E}(y_0), \text{ pour tout } t \geq 1. \end{aligned}$$

Ainsi, l'espérance d'une marche aléatoire ne dépend pas de t. Une hypothèse populaire est que $y_0 = 0$ – le processus commence à zéro en date zéro – et dans ce cas, $\mathrm{E}(y_t) = 0$ pour tout t.

Par contre, la variance d'une marche aléatoire varie avec t. Pour calculer la variance d'une marche aléatoire, nous supposons par simplicité que y_0 n'est pas aléatoire, et donc que $\text{Var}(y_0) = 0$ (cette hypothèse n'affecte aucune des conclusions importantes concernant la variance d'une marche aléatoire). Alors, comme $\{e_t\}$ est i.i.d.

$$\text{Var}(y_t) = \text{Var}(e_t) + \text{Var}(e_{t-1}) + \ldots + \text{Var}(e_1) = \sigma_e^2 t. \qquad \textbf{[11.21]}$$

En d'autres termes, la variance d'une marche aléatoire est une fonction croissante linéaire du temps. Cela montre que le processus ne peut être stationnaire.

Plus important encore, une marche aléatoire affiche une forte persistance, dans le sens où la valeur de y aujourd'hui est importante pour déterminer la valeur de y dans un futur proche. Pour voir cela, en considérant h périodes,

$$y_{t+h} = e_{t+h} + e_{t+h-1} + \ldots + e_{t+1} + y_t.$$

Maintenant supposons qu'à la période t, nous voulions calculer l'espérance de y_{t+h} étant donné la valeur courante de y_t. Comme l'espérance de e_{t+j}, sachant y_t, est égal à zéro pour tout $j \geq 1$, nous avons

$$\text{E}(y_{t+h}|y_t) = y_t, \text{ pour tout } h \geq 1. \qquad \textbf{[11.22]}$$

Cela signifie que, peu importe l'horizon, la meilleure prévision de y_{t+h} est la valeur actuelle, y_t. Nous pouvons comparer cela avec le cas d'un processus AR(1) stable, où un argument similaire peut être utilisé pour montrer que

$$\text{E}(y_{t+h}|y_t) = \rho_1^h y_t, \text{ pour tout } h \geq 1.$$

Sous les conditions de stabilité, $|\rho_1| < 1$, et donc $\text{E}(y_{t+h}|y_t)$ tend vers 0 quand $h \to \infty$: la valeur de y_t devient de plus en plus faible, et $\text{E}(y_{t+h}|y_t)$ se rapproche de plus en plus de son espérance non conditionnelle $\text{E}(y_t) = 0$.

Lorsque $h = 1$, l'équation (11.22) est la réminiscence de l'hypothèse d'anticipations adaptatives utilisée dans l'exemple 11.5 : si l'inflation suit une marche aléatoire, alors la valeur anticipée de inf_t, étant donné toutes les valeurs passées de l'inflation, est simplement inf_{t-1}. Ainsi, un modèle de marche aléatoire pour l'inflation justifie l'utilisation des anticipations adaptatives.

Nous pouvons aussi voir que la corrélation entre y_t et y_{t+h} est proche de 1 lorsque t est grand et que $\{y_t\}$ suit une marche aléatoire. Si $\text{Var}(y_0) = 0$, il peut être montré que

$$\text{Corr}(y_t, y_{t+h}) = \sqrt{t/(t+h)}.$$

Par conséquent, la corrélation dépend du point de départ, t (de telle sorte que $\{y_t\}$ ne soit pas stationnaire en covariance). De plus, pour une période t donnée, la corrélation tend vers zéro lorsque $h \to \infty$, mais elle ne tend pas vers zéro très rapidement. En réalité, plus t est grand, moins la corrélation tend vers 0 rapidement lorsque h augmente. Si nous choisissons un h important, par exemple, $h = 100$, nous pouvons toujours choisir un « t » assez grand de telle sorte que la corrélation entre y_t et y_{t+h} soit arbitrairement proche de 1. (Si $h = 100$ et que nous voulons que la corrélation soit supérieure à 0,95, alors choisir $t > 1\,000$ permet d'obtenir le résultat souhaité.) Par conséquence, une marche aléatoire ne satisfait pas la condition de séquence asymptotiquement non-corrélée.

Le graphique 11.1 montre deux réalisations d'une marche aléatoire, générée à partir d'un ordinateur, avec une valeur initiale $y_0 = 0$ et pour e_t suivant une loi normale (0,1). Généralement, il n'est pas facile de déterminer uniquement graphiquement si une série temporelle est effectivement une marche aléatoire. Par la suite, nous discuterons d'une méthode informelle pour faire la distinction entre des séries à faible dépendance et à forte dépendance. Nous couvrirons les tests statistiques formels permettant de vérifier cela dans le chapitre 18.

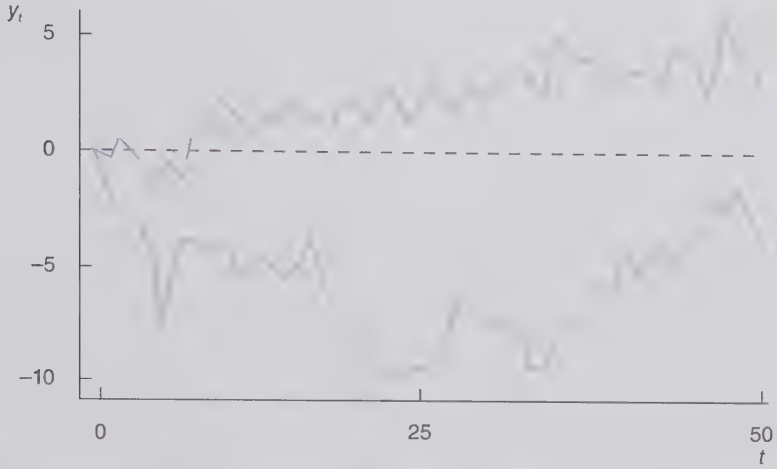


Figure 11.1 *Deux réalisations de marches aléatoires $y_t = y_{t-1} + e_t$, avec $y_0 = 0$, et y qui suit une loi normale (0,1), et n = 50.*

Une série qui est généralement vue comme un processus de marche aléatoire est le taux d'intérêt des bons du Trésor à 3 mois. Les données annuelles de cette variable sont tracées dans le graphique 11.2 pour les années 1948 à 1996.

Une marche aléatoire est un cas spécial de ce qui est connu sous le nom de processus à racine unitaire. Ce nom provient du fait que $\rho_1 = 1$ dans un modèle AR(1). Une classe plus générale de processus à racine unitaire est générée comme dans (11.20), mais $\{e_t\}$ peut maintenant être une série de faible dépendance (par exemple, $\{e_t\}$ pourrait lui-même suivre un processus MA(1) ou un processus AR(1) stable). Lorsque $\{e_t\}$ n'est pas une suite i.i.d., les propriétés de la marche aléatoires dérivées précédemment ne sont plus valables. Mais la caractéristique clé de $\{y_t\}$ est préservée : la valeur de y aujourd'hui est fortement corrélée avec y, même dans un futur lointain.

D'un point de vue de politique publique, il est souvent important de savoir si une série temporelle économique est fortement persistante ou non. Considérons par exemple le cas du Produit Intérieur Brut (PIB) aux États-Unis. Si le PIB est asymptotiquement non-corrélé, alors le niveau du PIB de l'année à venir est au mieux vaguement relié au PIB d'il y a par exemple 30 ans. Cela implique donc qu'une politique publique ayant affecté le PIB il y a de nombreuses années a un impact durable très faible. Par contre, si le PIB est fortement dépendant, le PIB de l'année prochaine est donc fortement corrélé au PIB d'il y a plusieurs années. Ainsi, une politique publique ayant entraîné un changement discret sur le PIB peut avoir des effets de long terme.

Figure 11.2 *Taux d'intérêt des bons du Trésor à 3 mois (1948-1996).*

Il est extrêmement important de ne pas confondre une tendance avec une situation de forte persistance. Une série peut contenir une tendance sans être fortement persistante, comme nous l'avons vu au chapitre 10. De plus, des facteurs tels que le taux d'intérêt, le taux d'inflation ou bien encore le taux de chômage sont considérés par beaucoup comme étant fortement persistant, sans qu'il n'existe une tendance nette, à la hausse ou à la baisse. Cependant, il existe souvent des séries fortement persistantes qui contiennent aussi une tendance nette. Un modèle permettant de répliquer ce comportement est la marche aléatoire avec dérive :

$$y_t = \alpha_0 + y_{t-1} + e_t, \; t = 1, 2, \ldots, \qquad \textbf{[11.23]}$$

où $\{e_t : t = 1, 2, \ldots\}$ et y_0 ont les mêmes propriétés que dans le cas d'un modèle de marche aléatoire. Ce qui est nouveau est le paramètre que nous avons appelé « terme de dérive ». Pour générer y_t, une constante α_0 est ajoutée au modèle en plus d'un terme aléatoire e_t et de la valeur précédente y_{t-1}. Nous pouvons voir que la valeur espérée de y_t suit une tendance temporelle linéaire du temps, en utilisant des substitutions successives :

$$y_t = \alpha_0 t + e_{t-1} + \ldots + e_1, y_0.$$

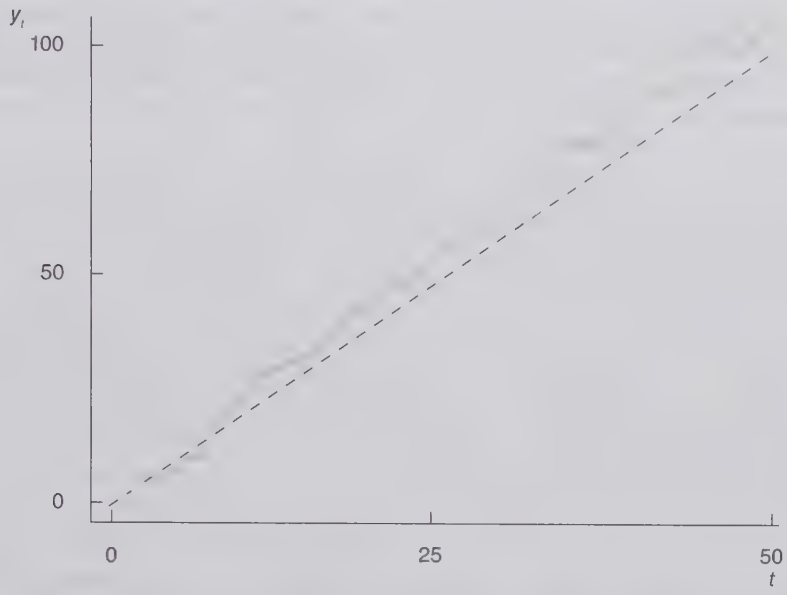


Figure 11.3 *Une réalisation d'une marche aléatoire avec dérive, $y_t = 2 + y_{t-1} + e_t$, avec $y_0 = 0$, e_t suit une loi normale (0, 9), et $n = 50$. La ligne pointillée représente l'espérance de y_t, $E(y_t) = 2t$.*

Par conséquent, si $y_0 = 0$, $E(y_t) = \alpha_0 t$: la valeur espérée de y_t augmente avec le temps si $\alpha_0 > 0$ et diminue avec le temps si $\alpha_0 < 0$. En raisonnant de la même manière que nous l'avions fait dans le cas d'une pure marche aléatoire, nous pouvons montrer que $E(y_{t+h}|y_t) = \alpha_0 h + y_t$, et donc que la meilleure prévision de y_{t+h} à la période t est y_t auquel on ajoute la dérive $\alpha_0 h$. La variance de y_t est alors la même que dans le cas d'une pure marche aléatoire.

Le graphique 11.3 représente une réalisation d'une marche aléatoire avec dérive, lorsque $n = 50$, $y_0 = 0$, $\alpha_0 = 2$, et où e_t est une variable aléatoire suivant une loi normale (0, 9). Comme nous pouvons le voir sur le graphique, y_t tend à croître avec le temps, mais la série ne retourne pas de façon régulière vers sa tendance.

Une marche aléatoire avec dérive est un autre exemple d'un processus à racine unitaire, car c'est un cas spécial $\rho_1 = 1$ d'un modèle AR(1) avec ordonnée à l'origine :

$$y_t = \alpha_0 + \rho_1 y_{t-1} + e_t.$$

Lorsque $\rho_1 = 1$ et $\{e_t\}$ est un processus quelconque de faible dépendance, nous obtenons alors un ensemble de classe de processus temporels fortement persistants, qui ont également une tendance linéaire moyenne.

Transformation des séries temporelles fortement persistantes

L'utilisation dans une équation de régression de séries temporelles ayant une forte persistance, caractérisée par un processus de racine unitaire, peut entraîner des résultats très trompeurs si les hypothèses du modèle classique linéaire ne sont pas respectées. Nous étudierons le problème de régressions fallacieuses dans le chapitre 18, mais pour le moment, nous devons garder en mémoire ces problèmes potentiels. Heureusement, des transformations simples sont possibles pour rendre un processus à racine unitaire faiblement dépendant.

Un processus faiblement dépendant est dit intégré d'ordre zéro, ou I(0). D'une manière pratique, cela signifie que rien ne doit être modifié sur les séries de ce type avant de faire une régression : les moyennes de

ce type de séries satisfaisant déjà le théorème central limite. Les processus à racine unitaire, tel que la marche aléatoire (avec ou sans dérive), sont dit intégrés d'ordre 1, ou I(1). Cela signifie que la première différence de ces processus est faiblement dépendante. Une série temporelle I(1) est souvent appelée processus stationnaire en différence, bien que ce nom soit quelque peu trompeur car mettant l'accent sur la stationnarité après la différenciation, plutôt que sur la faible dépendance des différences.

Le concept d'un processus I(1) est facile à voir en partant d'une marche aléatoire. Pour $\{y_t\}$ généré de la même manière que dans (11.20), pour $t = 1, 2, \ldots$,

$$\Delta y_t = y_t - y_{t-1} = e_t, \; t = 2, 3, \ldots ; \qquad \textbf{[11.24]}$$

la première différence de cette série $\{\Delta y_t = 2, 3, \ldots\}$ est une série i.i.d.. Globalement, si $\{y_t\}$ est généré par (11.24) dans lequel $\{e_t\}$ est un processus faiblement dépendant quelconque, alors $\{\Delta y_t\}$ est faiblement dépendant. De ce fait, lorsque nous supposons que des processus sont intégrés d'ordre 1, nous utilisons souvent la première différence afin d'utiliser ces variables par la suite dans une régression ; nous en verrons des exemples par la suite (accidentellement, le symbole « Δ » peut signifier « variation » ou « différence ». Lorsqu'une variable est appelée y, alors sa variation ou sa différence est souvent notée cy ou dy. Par exemple, la variation de prix peut être notée *cprice*.)

De nombreuses séries y_t strictement positives sont telles que $\log(y_t)$ est intégré d'ordre 1. Dans ce cas, nous utilisons la première différence des logarithmes, $\Delta\log(y_t) = \log(y_t) - \log(y_{t-1})$, dans les régressions. Alternativement, comme :

$$\Delta\log(y_t) \approx (y_t - y_{t-1})/y_{t-1}, \qquad \textbf{[11.25]}$$

nous pouvons utiliser la proportion ou le pourcentage de variation de y_t directement ; c'est ce que nous avons fait dans l'exemple 11.4, où, plutôt que d'exprimer l'hypothèse d'efficience des marchés en fonction du prix des actions, p_t, nous avons utilisé la variation en pourcentage hebdomadaire, $return_t = 100[(p_t - p_{t-1})/p_{t-1}]$. La quantité dans l'équation (11.25) est souvent appelée taux de croissance, mesuré en variation de proportion. Lorsque nous utilisons un ensemble particulier de données, il est important de savoir comment les taux de croissance sont mesurés – en proportion ou en pourcentage de variation. Parfois, pour une variable y, son taux de croissance est noté gy, de telle sorte que pour tout t, $gy_t = \log(y_t) - \log(y_{t-1})$ ou $gy_t = (y_t - y_{t-1})/y_{t-1}$. Souvent, ces quantités sont ensuite multipliées par 100 pour transformer une proportion en pourcentage de variation.

Différencier des séries temporelles avant de les utiliser a aussi un autre avantage : cela permet de supprimer les tendances linéaires. Ceci peut être démontré facilement, en écrivant une variable avec tendance linéaire sous la forme :

$$y_t = \gamma_0 + \gamma_1 t + v_t,$$

où v_t a une moyenne nulle. Alors, $\Delta y_t = \gamma_1 + \Delta v_t$, et donc $E(\Delta y_t) = \gamma_1 + E(\Delta v_t) = \gamma_1$. En d'autres termes, $E(\Delta y_t)$ est constant. Le même argument fonctionne pour $\Delta\log(y_t)$, lorsque $\log(y_t)$ suit une tendance. Ainsi, il est possible d'utiliser la première différence des variables présentant une tendance nette plutôt que d'inclure une variable de tendance dans la régression.

Déterminer si une série temporelle est I(1)

Déterminer si une série temporelle particulière est le résultat d'un processus I(1) ou d'un processus I(0) peut être quelque peu difficile. Certains tests statistiques peuvent être utilisés pour cela, mais ces techniques sont plus avancées. Une introduction au traitement de cette question sera faite au chapitre 18.

Il existe cependant des méthodes informelles qui peuvent fournir des indications utiles afin de voir si une série est approximativement caractérisée par une faible dépendance. Un outil simple est inspiré du modèle AR(1) : si $|\rho_1| < 1$, alors le processus est I(0) ; à l'opposé, le processus est I(1) si $\rho_1 = 1$. Précédemment, nous

avons montré que, lorsque le processus AR(1) est stable, $\rho_1 = \text{Corr}(y_t, y_{t-1})$. Ainsi, nous pouvons estimer ρ_1 à partir de la corrélation de l'échantillon entre y_t et y_{t-1}. Le coefficient de corrélation de l'échantillon est appelé autocorrélation de premier ordre de $\{y_t\}$; nous notons cela $\hat{\rho}_1$ En appliquant la loi des grands nombres, $\hat{\rho}_1$ est supposé proche de pour ρ_1, sous réserve que $|\rho_1| < 1$. (Cependant, $\hat{\rho}_1$ n'est pas un estimateur sans biais de ρ_1.)

EXEMPLE 11.6
Équation du taux de fécondité

Dans l'exemple 10.4, nous avions estimé le taux de fécondité *gfr*, en fonction de la valeur des allocations familiales, *pe*. L'autocorrélation de premier ordre de ces séries est très élevée : $\hat{\rho}_1 = 0{,}977$ pour *gfr* et $\hat{\rho}_1 = 0{,}964$ pour *pe*. Ces autocorrélations suggèrent la présence d'une racine unitaire, et soulèvent de nombreuses questions sur notre utilisation du *t*-stat issu des MCO. Souvenez-vous que les *t*-stats issus des MCO ne suivent une loi de Student que si toutes les hypothèses du modèle linéaire classique sont vérifiées. Pour relâcher ces hypothèses et appliquer les propriétés asymptotiques, il faut généralement que les séries temporelles sous-jacentes soient des processus I(0).

Nous estimons maintenant l'équation en utilisant les premières différences (et en supprimant la variable indicatrice pour plus de simplicité) :

$$\widehat{\Delta gfr} = -0{,}785 - 0{,}043\ \Delta pe$$
$$(0{,}502)\quad (0{,}028) \qquad \textbf{[11.26]}$$
$$n = 71,\ R^2 = 0{,}032,\ \overline{R}^2 = 0{,}018.$$

Maintenant, une augmentation de *pe* diminue *gfr* pour la même période, bien que les estimateurs ne soient pas statistiquement différents de zéro avec un seuil de confiance de 5 %. Cela nous donne donc des résultats très différents par rapport à l'analyse avec des variables en niveau, et cela jette un doute sur notre analyse précédente.

Si nous ajoutons deux retards de Δpe, les choses s'améliorent :

$$\widehat{\Delta gfr} = -0{,}964 - 0{,}036\ \Delta pe - 0{,}014\ \Delta pe_{-1} + 0{,}110\ \Delta pe_{-2}$$
$$(0{,}468)\ (0{,}027)\qquad (0{,}028)\qquad (0{,}027) \qquad \textbf{[11.27]}$$
$$n = 69,\ R^2 = 0{,}233,\ \overline{R}^2 = 0{,}197.$$

Bien que Δpe et Δpe_{-1} aient des coefficients négatifs, leurs coefficients sont petits et conjointement non significatifs (*p*-value = 0,28). Le second retard est fortement significatif, et indique une relation positive entre la variation de *pe* et les changements de *gfr* deux années plus tard. Cela a d'ailleurs beaucoup plus de sens qu'un effet simultané (Voir l'exercice pratique sur ordinateur C5 pour une analyse plus approfondie des équations en première différence.)

Nous pouvons utiliser la valeur de $\hat{\rho}_1$ pour nous aider à décider si un processus est I(1) ou I(0). Malheureusement, comme $\hat{\rho}_1$ est un estimateur, nous ne pouvons pas être sur que $\rho_1 < 1$. Idéalement, nous pourrions calculer un intervalle de confiance pour ρ_1, et voir si cet intervalle exclut la valeur $\rho_1 = 1$, mais cela est en réalité assez difficile. En effet les distributions des estimateurs sur l'échantillon de $\hat{\rho}_1$ sont très différentes lorsque ρ_1 est proche de 1 et lorsque ρ_1 est bien inférieur à 1. (En réalité, lorsque ρ_1 est proche de 1, $\hat{\rho}_1$ peut-être fortement biaisé à la baisse).

Dans le chapitre 18, nous montrerons comment il est possible de tester $H_0 : \rho_1 = 1$ contre $H_1 : \rho_1 < 1$. Pour le moment, nous pouvons utiliser $\hat{\rho}_1$ comme une bonne approximation de ρ_1 avant de décider si une série doit être différenciée. Il n'existe ni règle stricte, ni règle rapide pour faire ce choix. De nombreux économistes pensent que la la mise en différence de la série est justifiée si $\hat{\rho}_1 > 0{,}9$; mais d'autres peuvent prendre cette décision lorsque $\hat{\rho}_1 > 0{,}8$.

EXEMPLE 11.7
Salaires et productivité

La variable *hrwage* représente le salaire horaire moyen dans l'économie américaine, et *outphr* est la production par heure. Une façon d'estimer l'élasticité du salaire horaire en fonction de la production par heure est d'estimer l'équation :

$$\log(hrwage_t) = \beta_0 + \beta_1 \log(outphr_t) + \beta_2 t + u_t,$$

où on a inclus la tendance, car $\log(hrwage_t)$ et $\log(outphr_t)$ ont tous deux une nette tendance haussière linéaire. En utilisant les données du fichier EARNS pour les années 1947 à 1987, nous obtenons :

$$\widehat{\log(hrwage_t)} = -5{,}33 + 1{,}64 \log(outphr_t) - 0{,}018\, t \quad \textbf{[11.28]}$$
$$(0{,}37)\ (0{,}09) \qquad (0{,}002)$$
$$n = 41,\ R^2 = 0{,}971,\ \overline{R}^2 = 0{,}970.$$

(Nous avons reporté les mesures habituelles de précision du modèle ici ; il aurait été préférable de présenter les résultats basés sur une variable dépendante sans tendance, comme dans la section 10.5.). L'élasticité estimée semble trop élevée : une hausse de 1 % de la productivité entrainerait une hausse des salaires de 1,64 %. Comme l'écart-type est très faible, l'intervalle de confiance exclut clairement une élasticité unitaire. Les travailleurs américains auront sûrement du mal à croire que les salaires augmentent de plus de 1,5 % lorsque la productivité augmente de 1 %.

Les résultats de la régression (11.28) doivent être considérés avec attention. Même après avoir supprimer la tendance de $\log(hrwage)$, l'autocorrélation de premier-ordre est de 0,967, et pour $\log(outphr)$ sans tendance, $\hat{\rho}_1 = 0{,}945$. Cela suggère que les deux séries ont une racine unitaire ; nous devons donc ré-estimer l'équation en première différence (et donc nous n'avons plus besoin d'inclure une tendance) :

$$\widehat{\Delta \log(hrwage_t)} = -0{,}0036 + 0{,}809\, \Delta \log(outphr) \quad \textbf{[11.29]}$$
$$(0{,}0042)\ (0{,}173)$$
$$n = 40,\ R^2 = 0{,}364,\ \overline{R}^2 = 0{,}348.$$

Maintenant, une hausse de 1 % entraîne une augmentation du salaire réel d'environ 0,81 %, et l'estimateur n'est pas statistiquement différent de 1. Le R-carré ajusté montre que la croissance de la productivité explique environ 35 % de la croissance des salaires réels. (Voir Exercice C2 pour une version simple d'un modèle à retards répartis en première différence.)

Lorsque les séries étudiées présentent une nette tendance haussière ou baissière, il est plus logique d'obtenir l'autocorrélation de premier ordre après suppression de la tendance. Si la tendance n'est pas supprimée, la corrélation autorégressive a tendance à être surestimée, ce qui biaise à la hausse la possibilité de trouver une racine unitaire pour les processus avec tendance.

Dans les deux exemples précédents, la variable dépendante et les variables indépendantes semblaient avoir une racine unitaire. Dans d'autres cas, nous pouvons avoir un mélange de processus avec racine unitaire et de processus faiblement dépendant (avec ou sans tendance). Un exemple est donné dans l'exercice sur ordinateur C1.

11.4 MODÈLES DYNAMIQUE COMPLET ET ABSENCE DE CORRÉLATION SÉRIELLE

Dans le modèle AR(1) de l'équation (11.12), nous avons montré que, sous l'hypothèse (11.13), les erreurs $\{u_t\}$ sont nécessairement non-corrélées sériellement, au sens de l'hypothèse TS.5. Supposer qu'il n'y ait pas de corrélation sérielle est pratiquement la même chose que de considérer qu'un seul retard de y apparaît dans $E(y_t|y_{t-1}, y_{t-2}, \ldots)$.

Mais pouvons-nous faire le même raisonnement pour les autres modèles de régression ? La réponse est oui, bien que les conditions requises pour que les erreurs ne soient pas sériellement corrélées puissent être peu plausibles. Considérons par exemple un modèle de régression statique simple, à savoir :

$$y_t = \beta_0 + \beta_1 z_t + u_t, \qquad \textbf{[11.30]}$$

où y_t et z_t sont indexés par le temps de manière contemporaine. Pour que les estimateurs MCO soient consistants, nous avons seulement besoin que $E(u_t|z_t) = 0$. Généralement, les $\{u_t\}$ sont autocorrélés. Cependant, si nous supposons que :

$$E(u_t|z_t, y_{t-1}, z_{t-1}, \ldots) = 0, \qquad \textbf{[11.31]}$$

alors, (comme nous le montrerons dans un cas général ultérieurement) l'hypothèse TS.5′ est vérifiée. En particulier, les $\{u_t\}$ sont sériellement non-corrélées. Naturellement, l'hypothèse (11.31) implique que z_t est simultanément exogène, c'est-à-dire que $E(u_t|z_t) = 0$.

Pour mieux comprendre le sens de (11.31), nous pouvons écrire (11.30) et (11.31) comme :

$$E(y_t|z_t, y_{t-1}, z_{t-1}, \ldots) = E(y_t|z_t) = \beta_0 + \beta_1 z_t, \qquad \textbf{[11.32]}$$

où la première égalité est celle du taux intérêt présent. Cette équation montre que, une fois qu'on a tenu compte de z_t, aucun retard de y ou z ne permet d'expliquer la valeur présente de y. Ceci est une exigence forte, et est peu plausible lorsque les retards de la variable dépendante ont un pouvoir prédictif, ce qui est souvent le cas. Si cela est faux, nous pouvons alors nous attendre à ce que les erreurs soient auto-corrélées.

Ensuite, considérons un modèle à retards répartis finis intégrant deux retards :

$$y_t = \beta_0 + \beta_1 z_t + \beta_2 z_{t-1} + \beta_3 z_{t-2} + u_t. \qquad \textbf{[11.33]}$$

Étant donné que nous voulons capturer les effets avec retards de z sur y, nous supposons naturellement que (11.33) capture une dynamique de retards répartis :

$$E(y_t|z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots) = E(y_t|z_t, z_{t-1}, z_{t-2}) ; \qquad \textbf{[11.34]}$$

dans ce cas, au plus deux retards de z sont pris en compte. Si (11.31) est vérifiée, nous pouvons aller plus loin en indiquant qu'après avoir vérifié pour z et deux de ses retards, aucun retard de y et aucun retard additionnel z n'a d'impact sur y *actuel :*

$$E(y_t|z_t, y_{t-1}, z_{t-1}, \ldots) = E(y_t|z_t, z_{t-1}, z_{t-2}). \qquad \textbf{[11.35]}$$

L'équation (11.35) est plus probable que l'équation (11.32), mais cela exclut encore le fait que les retards de y puissent avoir un pouvoir prédictif additionnel sur le y contemporain.

Maintenant, considérons un modèle avec un retard de y et un retard de z :

$$y_t = \beta_0 + \beta_1 z_t + \beta_2 y_{t-1} + \beta_3 z_{t-1} + u_t.$$

Étant donné que ce modèle inclut un retard de la variable dépendante, l'équation (11.31) est une hypothèse naturelle, et implique que :

$$E(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots) = E(y_t|z_t, y_{t-1}, z_{t-1}).$$

En d'autres termes, après avoir vérifié pour z_t, y_{t-1}, et z_{t-1}, aucun autre retard de y ou de z n'affecte le y contemporain.

Dans le modèle général,

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \qquad \textbf{[11.36]}$$

où la variable explicative $\mathbf{x}_t = (x_{t1}, \ldots, x_{tk})$ peut contenir ou non des retards de y ou de z, l'équation (11.31) devient

$$\mathrm{E}(u_t|\mathbf{x}_t, y_{t-1}, \mathbf{x}_{t-1}, \ldots) = 0. \qquad \textbf{[11.37]}$$

En écrivant cela en termes de y_t,

$$\mathrm{E}(y_t|\mathbf{x}_t, y_{t-1}, \mathbf{x}_{t-1}, \ldots) = \mathrm{E}(y_t|\mathbf{x}_t) \qquad \textbf{[11.38]}.$$

En d'autres termes, peu importe ce qui est inclus dans $\mathbf{x}_t$, un nombre suffisant de retards ont déjà été inclus, de telle sorte que les retards supplémentaires de y ou des variables explicatives n'ont pas de pouvoir prédictif supplémentaire pour expliquer y_t. Lorsque cette condition est vérifiée, nous avons alors un modèle dynamique complet. Comme vu précédemment, le caractère complet de la dynamique peut être une hypothèse très forte pour les modèles statiques et les modèles à retards échelonnés.

Lorsque nous incluons des retards de y en tant que variables explicatives, nous considérons souvent que le modèle doit être dynamiquement complet. Nous couvrirons certaines exceptions par rapport à cette affirmation dans le chapitre 18.

Étant donné que (11.37) est équivalent à

$$\mathrm{E}(u_t|\mathbf{x}_t, u_{t-1}, \mathbf{x}_{t-1}, u_{t-2}, \ldots) = 0, \qquad \textbf{[11.39]}$$

Nous pouvons montrer qu'un modèle dynamiquement complet doit vérifier l'hypothèse TS.5′. (La dérivation n'est pas cruciale et peut être omise sans empêcher la compréhension). Pour être concret, prenons $s < t$. Alors, par la loi des espérances itérées (voir l'Annexe B),

$$\begin{aligned}\mathrm{E}(u_t u_s|\mathbf{x}_t, \mathbf{x}_s) &= \mathrm{E}[\mathrm{E}(u_t u_s|\mathbf{x}_t, \mathbf{x}_s, u_s)|\mathbf{x}_t, \mathbf{x}_s] \\ &= \mathrm{E}[u_s \mathrm{E}(u_t|\mathbf{x}_t, \mathbf{x}_s, u_s)|\mathbf{x}_t, \mathbf{x}_s],\end{aligned}$$

Où la seconde équation est obtenue car $\mathrm{E}(u_t u_s|\mathbf{x}_t, \mathbf{x}_s, u_s) = u_s \mathrm{E}(u_t|\mathbf{x}_t, \mathbf{x}_s, u_s)$. Maintenant, comme $s < t$, $(\mathbf{x}_t, \mathbf{x}_s, u_s)$ est un sous-ensemble de l'ensemble des variables conditionnelles à (11.39). Par conséquence, (11.39) implique que $\mathrm{E}(u_t|\mathbf{x}_t, \mathbf{x}_s, u_s) = 0$, et donc

$$\mathrm{E}(u_t u_s|\mathbf{x}_t, \mathbf{x}_s) = \mathrm{E}(u_s \cdot 0|\mathbf{x}_t, \mathbf{x}_s) = 0.$$

Ceci signifie donc que l'hypothèse TS.5′ est vérifiée.

Étant donné que la spécification d'un modèle dynamiquement complet signifie qu'il n'existe pas d'autocorrélation, cela implique t-il que tous les modèles doivent être dynamiquement complets ? Comme nous le verrons dans le chapitre 18, dans le cadre d'un modèle de prévision, la réponse est oui. Certains pensent que tous les modèles doivent être dynamiquement complets et que l'autocorrélation des erreurs d'un modèle est un signe de mauvaise spécification. Mais cette prise de position est trop rigide. Parfois, l'utilisation d'un modèle statique (comme la courbe de Phillips) ou d'un modèle à retards répartis finis (comme la mesure de l'ajustement du salaire de long-terme suite à une hausse de 1 % de la productivité) permet de répondre à nos besoins. Nous verrons dans le chapitre suivant comment détecter et corriger la corrélation sérielle dans des modèles de ce type.

EXEMPLE 11.8
Équation du taux de fécondité

Dans l'équation (11.27), nous avons estimé un modèle à retards répartis finis pour Δgfr en fonction de Δpe, en acceptant deux retards de Δpe. Pour que ce modèle soit dynamiquement complet au sens de (11.38), aucun retard de Δgfr ni retards supplémentaires de Δpe ne doivent apparaître dans l'équation. Nous pouvons facilement voir que cela est faux, en ajoutant Δgf_{t-1}, dont le coefficient est 0,300, et la statistique t est de 2,84. Ainsi, le modèle n'est pas dynamiquement complet au sens de (11.38).

Comment appréhender ce question ? Nous allons reporter l'interprétation des modèles généraux avec variables dépendantes retardées au chapitre 18. Mais le fait que (11.27) ne soit pas dynamiquement complet suggère qu'il existe peut-être une corrélation sérielle des termes d'erreurs. Nous verrons comment tester et corriger cela dans le chapitre 12.

Suite 11.3

Si (11.33) est vérifiée où $u_t = e_t + \alpha_1 e_{t-1}$ et où $\{e_t\}$ sont i.i.d. de moyenne zéro et de variance σ_e^2, l'équation (11.33) peut-elle être dynamiquement complète ?

La notion du caractère complet de la dynamique ne doit pas être confondue avec des hypothèses plus faibles concernant le nombre approprié de retards dans un modèle. Dans le modèle (11.36), les variables explicatives $\mathbf{x}_t$ sont dites séquentiellement exogènes si

$$\mathrm{E}(u_t|\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots) = \mathrm{E}(u_t) = 0, \; t = 1, 2, \ldots \qquad \textbf{[11.40]}$$

Comme discuté dans le problème 8 du chapitre 10, l'exogénéité séquentielle est impliqué par une stricte exogénéité, et l'exogénéité séquentielle conduit à une exogénéité contemporaine. De plus, étant donné que $(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots)$ est un sous-ensemble de $(\mathbf{x}_t, y_{t-1}, \mathbf{x}_{t-1}, \ldots)$, l'exogénéité séquentielle est impliquée par la complétude dynamique. Si $\mathbf{x}_t$ contient y_{t-1}, la complétude dynamique et l'exogénéité séquentielle constituent alors une seule condition. Le point clé est que, lorsque $\mathbf{x}_t$ ne contient pas y_{t-1}, l'exogénéité séquentielle ouvre la possibilité que la dynamique ne soit pas complète, dans le sens où cela ne capture par la relation entre y_t et toutes les valeurs passées de y et des autres variables explicatives. Mais dans un modèle à retards répartis finis – tel que celui estimé dans l'équation (11.27) – nous pouvons ne pas nous soucier de savoir si les valeurs passées de y ont un pouvoir prédictif sur y actuel. Notre intérêt principal est en effet de savoir si nous avons inclus un nombre suffisant de retards des variables explicatives pour capturer la dynamique des retards répartis. Par exemple, si nous supposons $\mathrm{E}(y_t|z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots) = \mathrm{E}(y_t|z_t, z_{t-1}, z_{t-2}) = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2}$, alors les variables explicatives $\mathbf{x}_t = (z_t, z_{t-1}, z_{t-2})$ sont séquentiellement exogènes parce nous supposons que deux retards du modèle étaient suffisants. Mais typiquement, le modèle n'est pas dynamiquement complet au sens de $\mathrm{E}(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, z_{t-2}, \ldots) = \mathrm{E}(y_t|z_t, z_{t-1}, z_{t-2})$, mais cela ne nous importe pas nécessairement. De plus, les variables explicatives d'un modèle à retards répartis finis peuvent être ou non strictement exogènes.

11.5 L'HYPOTHÈSE D'HOMOSCÉDASTICITÉ POUR LES SÉRIES TEMPORELLES

L'hypothèse d'homoscédasticité dans le cadre de régression de séries temporelles, particulièrement l'hypothèse TS.4′, est très proche de celle vue dans le cadre de régression en coupes transversales. Cependant, comme $\mathbf{x}_t$ peut contenir des retards de y ainsi que des retards des variables explicatives, nous allons brièvement discuter de la signification de l'hypothèse d'homoscédasticité dans le cadre de régression de séries temporelles.

Pour un modèle statique simple, par exemple :

$$y_t = \beta_0 + \beta_1 z_t + u_t, \qquad \textbf{[11.41]}$$

L'hypothèse TS.4′ requiert que $\text{Var}(u_t|z_t) = \sigma^2$.

Dans ce cas, bien que $\text{E}(y_t|z_t)$ soit une fonction linéaire de z_t, $\text{Var}(y_t|z_t)$ doit être constante.

Dans l'exemple 11.4, nous avons vu que, pour un modèle AR(1) comme celui de l'équation (11.12) l'hypothèse d'homoscédasticité est :

$$\text{Var}(u_t|y_{t-1}) = \text{Var}(y_t|y_{t-1}) = \sigma^2 ;$$

Bien que $\text{E}(y_t|y_{t-1})$ dépende de y_{t-1}, ce n'est pas le cas de $\text{Var}(y_t|y_{t-1})$.

Ainsi, l'écart dans la distribution de y_t ne peut pas dépendre de y_{t-1}.

Heureusement, tout cela est désormais plus clair. Si nous avons un modèle du type :

$$y_t = \beta_0 + \beta_1 z_t + \beta_2 y_{t-1} + \beta_3 z_{t-1} + u_t,$$

L'hypothèse d'homoscédasticité est

$$\text{Var}(u_t|z_t, y_{t-1}, z_{t-1}) = \text{Var}(y_t|z_t, y_{t-1}, z_{t-1}) = \sigma^2,$$

De telle sorte que la variance de u_t ne dépende pas de z_t, y_{t-1}, ou de z_{t-1} (ou de toute autre fonction du temps). En général, peu importe les variables du modèle, nous devons supposer que la variance de y_t, sachant toutes les variables explicatives, est constante. Si le modèle contient des retards de y ou des retards des variables explicatives, alors nous excluons explicitement une forme dynamique d'hétéroscédasticité (nous verrons cela plus en détails dans le chapitre 12). Mais dans un modèle statique, nous ne sommes intéressés que par $\text{Var}(y_t|z_t)$. Dans l'équation (11.41), aucune restriction directe n'est placée, par exemple, sur $\text{Var}(y_t|y_{t-1})$.

RÉSUMÉ

Dans ce chapitre, nous avons montré que l'utilisation des MCO peut être justifiée en utilisant l'analyse asymptotique, à condition que certaines conditions soient respectées. Idéalement, les processus de séries temporelles sont stationnaires et de faible dépendance, bien que la stationnarité ne soit pas cruciale. La faible dépendance est nécessaire pour appliquer les résultats standard, particulièrement le théorème central limite.

Les processus à tendance déterministe de faible dépendance peuvent être utilisés directement dans les régressions, à condition qu'une variable de tendance soit incluse dans le modèle (comme dans la section 10.5). La même chose est vraie pour les processus affichant une saisonnalité.

Lorsque les séries temporelles sont fortement persistantes (avec une racine unitaire), nous devons être extrêmement prudents avant d'utiliser ces séries directement dans un modèle de régression (sauf si nous sommes convaincus que les hypothèses du modèle linéaire classique du chapitre 10 sont vérifiées). Une alternative à cela consiste à utiliser la première différence de ces variables. Pour la plupart des séries temporelles économiques à forte persistance, la première différence est de faible dépendance. L'utilisation de la première différence change la nature du modèle, mais cette méthode fournit souvent autant d'information que l'utilisation d'un modèle en niveau. Lorsque les données sont fortement persistantes, les résultats en utilisant les différences premières sont généralement plus fiables qu'avec des données en niveau. Dans le chapitre 18, nous couvrirons certaines méthodes récentes plus avancées permettant d'utiliser des variables I(1) pour une analyse de régression multiples.

Lorsque les modèles sont dynamiquement complets, au sens qu'aucun retard supplémentaire (pour aucune variable) n'est nécessaire à l'équation, nous avons vu que les erreurs sont sériellement non-corrélés. Ceci est utile car certains modèles, comme par exemple les modèles autorégressifs, sont supposés être dynamiquement complets. Dans un modèle statique ou à retards répartis, l'hypothèse de complétude dynamique est généralement fausse, ce qui signifie que les erreurs sont autocorrélées. Nous verrons comment corriger cela dans le chapitre 12.

LES HYPOTHÈSES ASYMPTOTIQUES DE GAUSS-MARKOV POUR LES RÉGRESSIONS DE SÉRIES TEMPORELLES

Nous proposons ici un résumé des cinq hypothèses que nous avons utilisé dans ce chapitre afin de réaliser les inférences sur les échantillons de grande taille pour les régressions de séries temporelles. Souvenez-vous que nous avons introduit un nouvel ensemble d'hypothèses car les hypothèses du modèle linéaire classique sont souvent violées pour les séries temporelles, spécialement les hypothèses d'exogénéité stricte, d'absence d'autocorrélation et de normalité. Un point clé de ce chapitre est qu'une sorte de faible dépendance est requise pour s'assurer que le théorème central limite s'applique. Nous avons seulement utilisé les hypothèses TS.1′ à TS.3′ pour la consistance (et non pour le caractère sans biais) des MCO. Lorsque nous ajoutons TS.4′ et TS.5′, nous pouvons utiliser les intervalles de confiance, la t-stat et la F-stat comme étant approximativement valides dans le cas d'échantillons de grande taille. Contrairement aux hypothèses de Gauss-Markov et aux hypothèses du modèle classique linéaire, il n'existe pas de label attaché aux hypothèses TS.1′ à TS.5′. Cependant, ces hypothèses sont analogues à celles de Gauss-Markov, ce qui nous permet d'utiliser les inférences standard. Comme vu précédemment pour l'analyse d'échantillons de grande taille, nous nous dispensons de l'hypothèse de normalité.

Hypothèse TS.1′ (Linéarité et faible dépendance)

Un processus stochastique $\{(x_{t1}, x_{t2}, \ldots, x_{tk}, y_t) : t = 1, 2, \ldots, n\}$ est stationnaire et de faible dépendance, et suit un modèle linéaire

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t,$$

où $\{u_t : t = 1, 2, \ldots, n\}$ est la suite des erreurs. Ici, n est le nombre d'observations (nombre de période de temps).

Hypothèse TS.2′ (Absence de parfaite colinéarité)

Dans l'échantillon (et donc pour le processus temporel sous-jacent), aucune variable indépendante n'est constante ni une parfaite combinaison linéaire des autres.

Hypothèse TS.3′ (Moyenne conditionnelle nulle)

Les variables explicatives sont *simultanément exogènes*, c'est-à-dire que, $E(u_t|x_{t1}, \ldots, x_{tk}) = 0$. Souvenez-vous, TS.3′ est nettement plus faible que l'hypothèse de stricte exogénéité.

Hypothèse TS.4′ (Homoscédasticité)

Les erreurs sont simultanément homoscédastiques, c'est-à-dire que $\text{Var}(u_t|\mathbf{x}_t) = \sigma^2$, où $\mathbf{x}_t$ est une abréviation de $(x_{t1}, x_{t2}, \ldots, x_{tk})$.

Hypothèse TS.5′ (Absence d'autocorrélation)

Pour tout $t \neq s$, $E(u_t\, u_s|\mathbf{x}_t, \mathbf{x}_s) = 0$.

MOTS-CLÉS

Asymptotiquement non-corrélés p. 454
Autocorrélation de premier ordre p. 469
Faible dépendance p. 454
Forte dépendance p. 463
Forte persistance p. 464
Marche aléatoire p. 463
Marche aléatoire avec dérive p. 466
Modèle dynamiquement complet p. 472
Première différence p. 468
Processus autorégressif d'ordre 1 [ar(1)] p. 455
Processus intégré d'ordre 0 [i(0)] p. 467
Processus intégré d'ordre 1 [i(1)] p. 467
Processus à moyenne mobile d'ordre 1 [MA(1)] p. 454
Processus non-stationnaire p. 453
Processus à racine unitaire p. 465
Processus stable AR(1) p. 455
Processus stationnaire en différence p. 468
Processus à tendance stationnaire p. 456
Sériellement non corrélés p. 460, 475
Simultanément exogène p. 457
Simultanément homoscédastique p. 459
Stationnaire en covariance p. 453
Taux de croissance p. 468

PROBLÈMES

1. Soit $\{x_t : t = 1, 2, \ldots\}$ un processus stationnaire en covariance, défini par $\gamma_h = \text{Cov}(x_t, x_{t+h})$ pour tout $h \geq 0$. [Donc, $\gamma_0 = \text{Var}(x_t)$] Montrez que : $\text{Corr}(x_t, x_{t+h}) = \gamma_h/\gamma_0$.

2. Soit $\{e_t : t = -1, 0, 1, \ldots\}$ une suite de variables aléatoires indépendantes, identiquement distribuées, de moyenne zéro et de variance 1. Soit le processus stochastique défini par :

$$x_t = e_t - (1/2)e_{t-1} + (1/2)e_{t-2}, \; t = 1, 2, \ldots$$

i. Calculez $\text{E}(x_t)$ et $\text{Var}(x_t)$. Ces variables dépendent-elles de t ?

ii. Montrez que $\text{Corr}(x_t, x_{t+1}) = -1/2$ et $\text{Corr}(x_t, x_{t+2}) = 1/3$. (*Astuce* : Il est plus facile d'utiliser la formule du Problème 1.)

iii. A combien est égal $\text{Corr}(x_t, x_{t+h})$ pour $h > 2$?

iv. Est-ce que le processus $\{x_t\}$ est asymptotiquement non corrélé ?

3. Soit un processus temporel $\{y_t\}$ généré par $y_t = z + e_t$, pour tout $t = 1, 2, \ldots$, où $\{e_t\}$ est une suite i.i.d. de moyenne 0 et de variance σ_e^2. La variable aléatoire z ne dépend pas du temps ; sa moyenne est zéro et sa variance σ_z^2. En supposant que e_t n'est pas corrélé avec z.

i. Calculez l'espérance et la variance de y_t. Ces variables dépendent-telles de t ?

ii. Calculez $\text{Cov}(y_t, y_{t+h})$ pour n'importe quelle t et h. Est-ce que $\{y_t\}$ est stationnaire en covariance ?

iii. Utilisez (i) et (ii) pour montrer que pour tout t et h $\text{Corr}(y_t, y_{t+h}) = \sigma_z^2 / (\sigma_z^2 + \sigma_e^2)$.

iv. Est-ce que y_t satisfait les conditions intuitives pour être asymptotiquement non corrélé ? Justifiez.

4. Soit $\{y_t : t = 1, 2, \ldots\}$ un processus de marche aléatoire comme défini dans (11.20), avec $y_0 = 0$. Montrez que $\text{Corr}(y_t, y_{t+h}) = \sqrt{t/(t+h)}$ pour $t \geq 1, h > 0$.

5. Notons *gprice* le taux de croissance mensuel de l'indice général des prix et *gwage* le taux de croissance mensuel du salaire horaire [Les deux variables sont obtenues en considérant la première différence du logarithme *gprice* = $\Delta\log(price)$ et *gwage* = $\Delta\log(wage)$.]. En utilisant les données mensuelles du fichier, WAGEPRC, nous estimons le modèle à retards répartis suivant :

$$\begin{aligned} gprice = &- 0{,}00093 + 0{,}119\ gwage + 0{,}097\ gwage_{-1} + 0{,}040\ gwage_{-2} \\ &\quad (0{,}00057)\ (0{,}052)\ \ (0{,}039)\ \ (0{,}039) \\ &+ 0{,}038\ gwage_{-3} + 0{,}081\ gwage_{-4} + 0{,}107\ gwage_{-5} + 0{,}095\ gwage_{-6} \\ &\quad (0{,}039)\ \ (0{,}039)\ \ (0{,}039)\ \ (0{,}039) \\ &+ 0{,}104\ gwage_{-7} + 0{,}103\ gwage_{-8} + 0{,}159\ gwage_{-9} + 0{,}110\ gwage_{-10} \\ &\quad (0{,}039)\ \ (0{,}039)\ \ (0{,}039)\ \ (0{,}039) \\ &+ 0{,}103\ gwage_{-11} + 0{,}016\ gwage_{-12} \\ &\quad (0{,}039)\ \ (0{,}052) \end{aligned}$$

$$n = 273,\ R^2 = 0{,}317,\ \overline{R}^2 = 0{,}283.$$

i. Tracez la structure de distribution des retards. Pour quel retard l'effet de *gwage* sur *gprice* est-il le plus fort ? Pour quel retard est-il le plus faible ?

ii. Pour quels retards les t-statistiques sont-elles inférieures à 2 ?

iii. Quelle est l'impact de long terme estimé ? Est-il très différent de 1 ? Expliquez ce que le multiplicateur de long terme nous montre dans cet exemple.

iv. Quelle régression feriez-vous pour obtenir directement l'écart-type du multiplicateur de long terme ?

v. Comment testeriez-vous la significativité conjointe de six retards ou plus de *gwage* ? Quels seraient les degrés de libertés de la loi de Fisher ? (Attention ici, l'ajout de retards diminue le nombre d'observations.)

6. Soit $hy6_t$ le rendement à 3 mois (en pourcentage) procuré par l'achat d'un bon du trésor à 6 mois à la période $(t-1)$ et revendu 3 mois plus tard à la période t. Soit $hy3_{t-1}$ le rendement à trois mois procuré par l'achat d'un bon du trésor à 3 mois à la période $(t-1)$. A la période $(t-1)$, $hy3_{t-1}$ est connu, tandis que $hy6_t$ n'est pas connu car $p3_t$ (le prix d'un bon du trésor à 3mois) n'est pas connu à la période $(t-1)$. L'hypothèse des anticipations rationnelles (RE) suppose que ces deux investissements à trois mois doivent, en moyenne, apporter le même rendement. Mathématiquement, nous pouvons écrire cela sous la forme d'une espérance conditionnelle :

$$\text{E}(hy6_t | I_{t-1}) = hy3_{t-1},$$

où I_{t-1} correspond à l'ensemble des informations disponibles jusqu'à la date $t-1$. Cela revient à estimer le modèle

$$hy6_t = \beta_0 + \beta_1 hy3_{t-1} + u_t,$$

et à tester $H_0 : \beta_1 = 1$. (nous pouvons aussi tester $H_0 : \beta_0 = 0$, mais nous acceptons souvent qu'il existe une prime de maturité pour l'achat d'actifs de plus longue maturité, de telle sorte que $\beta_0 \neq 0$.)

a. (i) L'estimation de l'équation précédente en utilisant la méthode des MCO, avec les données du fichier INTQRT (espacé par intervalle de 3-mois) nous donne :

$$\widehat{hy6}_t = -0{,}058 + 1{,}104\, hy3_{t-1}$$
$$(0{,}070)\ \ (0{,}039)$$
$$n = 123,\ R^2 = 0{,}866.$$

Rejetez-vous $H_0 : \beta_1 = 1$ contre $H_0 : \beta_1 \neq 1$ avec un seuil de confiance de 1 % ? Est-ce que l'estimateur semble différent de 1 ?

a. (ii) Une autre implication de l'hypothèse des anticipations rationnelles est qu'aucune autre variable en t-1 ou précédemment ne permet d'expliquer $hy6_t$, une fois que l'on tient compte de $hy3_{t-1}$. En incluant un retard de l'écart (spread) entre le taux à 6 mois et le taux à 3 mois on obtient :

$$\widehat{hy6}_t = -0{,}123 + 1{,}053\, hy3_{t-1} + 0{,}480(r6_{t-1} - r3_{t-1})$$
$$(0{,}067)\ \ (0{,}039)\qquad\quad (0{,}109)$$
$$n = 123,\ R^2 = 0{,}885.$$

Maintenant, est-ce que le coefficient de $hy3_{t-1}$ est statistiquement différent de 1 ? Le retard du spread est-il significatif ? Selon cette équation, si, à la période $t - 1$, *r6* est plus élevé que *r3*, investiriez vous plutôt dans le bon du trésor à 3 mois ou dans le bon du trésor à 6 mois ?

iii) La corrélation entre $hy3_t$ et $hy3_{t-1}$ est égale à 0,914. Pourquoi cela pourrait-il soulever quelques préoccupations en ce qui concerne l'analyse précédente ?

iv) Comment testeriez-vous la saisonnalité dans l'équation estimée dans la partie (ii) ?

7. Un ajustement partiel du modèle est :

$$y_t^* = \gamma_0 + \gamma_1 x_t + e_t$$
$$y_t - y_{t-1} = \lambda(y_t^* - y_{t-1}) + a_t,$$

où y_t^* est le niveau désiré ou optimal de y, et y_t est le niveau actuel (observé). Par exemple, y_t^* est le taux de croissance optimal des stocks d'une entreprise, et x_t le taux de croissance des ventes d'une entreprise. Le paramètre γ_1 mesure l'effet de x_t sur y_t^*. La seconde équation décrit la façon dont le *y actuel* s'ajuste en fonction de la relation entre le y optimal à la période t et le y actuel en $t - 1$. Le paramètre λ mesure la vitesse d'ajustement, et satisfait $0 < \lambda < 1$.

(i) En injectant la première équation dans la seconde, montrez que l'on peut écrire : $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + u_t$.

En particulier, exprimez β_j en termes de γ_j et de λ, et exprimez u_t en termes de e_t et a_t. Par conséquent, le modèle à ajustement partiel aboutit à un modèle avec une variable dépendante à retard et une variable x contemporaine.

ii) Si $E(e_t|x_t, y_{t-1}, x_{t-1}, \ldots) = E(a_t|x_t, y_{t-1}, x_{t-1}, \ldots) = 0$ et que toutes les séries sont de faible dépendance, comment estimeriez vous β_j ?

iii) Si $\hat{\beta}_1 = 0{,}7$ et $\hat{\beta}_2 = 0{,}2$, quels sont les estimateurs de γ_1 et de λ ?

8. Supposons que l'équation :

$$y_t = \alpha + \delta y + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t$$

satisfasse l'hypothèse d'exogénéité séquentielle de l'équation (11.40).

i) Supposons que vous utilisiez la différence de cette équation, pour obtenir :

$$\Delta y_t = \delta + \beta_1 \Delta x_{t1} + \ldots + \beta_k \Delta x_{tk} + \Delta u_t.$$

Pourquoi utiliser la méthode des MCO avec une équation différenciée n'implique pas forcément une consistance des estimateurs de β_j ?

ii) Quelles hypothèses en ce qui concerne les variables explicatives de l'équation de base permettrait de s'assurer que les estimateurs MCO en première différence permettent une estimation consistante de β_j ?

iii) Soit $z_{t1}, \ldots, z_{tk}$ un ensemble de variables explicatives datées de manière contemporaine avec y_t. Si nous spécifions un modèle classique de régression du type $y_t = \beta_0 + \beta_1 z_{t1} + \ldots + \beta_k z_{tk} + u_t$, décrivez ce dont nous avons besoin pour supposer que $\mathbf{x}_t = \mathbf{z}_t$ soit séquentiellement exogène. Pensez-vous que ces hypothèses puissent être satisfaites dans des applications économiques ?

EXERCICES SUR ORDINATEUR

C1. Utilisez les données du fichier HSEINV pour cet exercice :

(i) Trouvez l'autocorrélation de premier ordre de *log*(*invpc*). Maintenant, déterminez l'autocorrélation après avoir enlevé la tendance de *log*(*invpc*). Faites la même chose pour *log*(*price*). Laquelle des deux séries semble avoir une racine unitaire ?

(ii) En vous basant sur vos réponses de (i), estimez l'équation :

$$\log(invpc_t) = \beta_0 + \beta_1 \Delta\log(price_t) + \beta_2 t + u_t$$

et reportez les résultats sous la forme standard. Interprétez le coefficient $\hat{\beta}_1$ et déterminez si celui-ci est statistiquement significatif.

(iii) Supprimez la tendance de $\log(invpc_t)$ et utilisez la variable sans tendance en tant que variable dépendante de la régression de la question (ii) (voir 10.5). Que se passe t-il au niveau du R^2 ?

(iv) Maintenant, utilisez $\Delta\log(invpc_t)$ comme variable dépendante. Comment vos résultats changent-ils par rapport à la question (ii) ? Est-ce que la tendance linéaire est toujours significative ? Justifiez.

C2. Dans l'exemple 11.7, considérez le taux de croissance du salaire horaire et de la production par heure en utilisant le logarithme naturel : *ghrwage* = $\Delta\log(hrwage)$ et *goutphr* = $\Delta\log(outphr)$. Considérons une extension simple du modèle estimé en (11.29) :

$$ghrwage_t = \beta_0 + \beta_1 goutphr_t + \beta_2 goutphr_{t-1} + u_t.$$

Ce modèle permet qu'une hausse de la productivité ait un effet direct et avec retard sur la hausse des salaires.

(i) Estimez l'équation à partir des données du fichier EARNS, et écrivez les résultats sous la forme standard. Est-ce que le retard de *goutphr* est statistiquement significatif ?

(ii) Si $\beta_1 + \beta_2 = 1$, une hausse permanente de la productivité est-elle entièrement transmise dans les salaires après un an ? Testez l'hypothèse $H_0 : \beta_1 + \beta_2 = 1$. Souvenez-vous, une façon de répondre à cette question est d'écrire l'équation sous une forme spécifique pour faire apparaître $\theta = \beta_1 + \beta_2$ dans le modèle, comme dans l'exemple 10.4 du chapitre 10.

(iii) Est-il nécessaire d'ajouter $goutphr_{t-2}$ dans le modèle ? Justifiez.

C3. (i) Dans l'exemple 11.4, il est possible que l'espérance du rendement à la période *t*, sachant les rendements passés, soit une fonction quadratique de $return_{t-1}$. Pour vérifier cette possibilité, utilisez les données du fichier NYSE pour estimer :

$$return_t = \beta_0 + \beta_1 return_{t-1} + \beta_2 return_{t-1}^2 + u_t$$

et écrivez vos résultats sous la forme classique.

(ii) Indiquez et testez l'hypothèse nulle que $E(return_t|return_{t-1})$ ne dépende pas de $return_{t-1}$. (*Astuce* : Il y a deux restrictions à tester ici.). Qu'en concluez-vous ?

(iii) Supprimez $return_{t-1}^2$ du modèle, et ajoutez un terme d'interaction $return_{t-1} \cdot return_{t-2}$. En considérant cette nouvelle équation, testez l'hypothèse d'efficience des marchés.

(iv) Que pouvez-vous conclure à propos de la prévision des rendements hebdomadaires du marché d'actions à partir des rendements passés du marché ?

C4. Utilisez les données du fichier PHILLIPS pour cet exercice, mais seulement les données jusqu'à l'année 1996.

(i) Dans l'exemple 11.5, nous avons supposé que le taux naturel du chômage était constant. Une forme alternative de la courbe de Phillips augmentée des anticipations autorise que le taux naturel du chômage dépende des valeurs passées du taux de chômage. Dans le cas le plus souple, le taux naturel du chômage en période t est égal à $unem_{t-1}$. Si nous supposons que les anticipations sont adaptatives, nous obtenons une courbe de Phillips où l'inflation et le taux de chômage sont exprimés en première différence :

$$\Delta inf = \beta_0 + \beta_1 \Delta unem + u.$$

Estimez ce modèle, écrivez les résultats sous la forme habituelle, puis discutez du signe, de la taille, et de la significativité de $\hat{\beta}_1$.

(ii) Quel modèle correspond-il le mieux aux données : l'équation (11.19) ou le modèle de la partie (i) ? Justifiez.

C5. (i) Ajoutez une tendance linéaire à l'équation (11.27). Est-ce que l'ajout d'une tendance est nécessaire pour les équations en premières différences ?

(ii) Supprimez la tendance et ajoutez les variables *ww2* et *pill* à l'équation (11.27). Est-ce que les variables sont conjointement significatives avec un seuil de confiance de 5 % ?

(iii) Ajoutez une tendance linéaire, *ww2* et *pill* à l'équation (11.27). Que se passe t-il en ce qui concerne l'ampleur et la significativité de la tendance par rapport à la question (i) ? Qu'en est-il du coefficient de *pill* par rapport à celui de la question (ii) ?

(iv) En utilisant le modèle de la question (iii), estimez le multiplicateur de long terme et indiquez son écart-type. Comparez les résultats à ceux de (10.19), où *gfr* et *pe* étaient exprimés en niveau plutôt qu'en première différence. Diriez-vous que la relation entre le taux de fécondité et le niveau des allocations familiales est particulièrement robuste ?

C6. Soit $inven_t$ la valeur réelle des stocks aux États-Unis durant l'année t, GDP_t le Produit Intérieur Brut réel et $r3_t$ le taux intérêt réel (ex-post) d'un bon du trésor à 3 mois. Le taux d'intérêt réel ex-post est environ égal à $r3_t = i3_t - inf_t$, où $i3_t$ est le d'un bon du trésor à 3 mois et inf_t est le taux d'inflation annuel [voir Mankiw (1994, section 6.4)]. La variation des stocks, $cinven_t$, correspond aux investissements en stocks pour l'année. L'effet accélérateur de l'investissement en stock relie *cinven* à *cGDP*, le taux de croissance :

$$cinven_t = \beta_0 + \beta_1 cGDP_t + u_t, \text{ où } \beta_1 > 0.$$

[Voir, par exemple, Mankiw (1994), chapitre 17.]

(i) Nous utilisons les données du fichier INVEN pour estimer l'effet d'accélérateur. Reportez les résultats sous la forme classique, et interprétez l'équation. Est-ce que $\hat{\beta}_1$ est statistiquement différent de zéro ?

(ii) Si le taux d'intérêt réel augmente, alors le coût d'opportunité de conserver des stocks augmentent, et donc une hausse du taux d'intérêt réel devrait diminuer les stocks. Ajoutez le taux d'intérêt réel au modèle précédent, et discutez des résultats.

(iii) Est-il préférable d'utiliser le niveau du taux d'intérêt réel ou bien la première différence, $cr3_t$?

C7. Utilisez les données du fichier CONSUMP pour cet exercice. Une version de la théorie du revenu permanent indique que la croissance de la consommation est imprévisible [une autre version de cette théorie est que le changement de consommation en lui-même est imprévisible, voir Mankiw (1994, chapitre 15, pour une discussion de la théorie du revenu permanent.] Soit $gc_t = \log(c_t) - \log(c_{t-1})$, le taux de croissance réel de la consommation par habitant (de biens non-durables et de services). Ainsi, la théorie du revenu permanent implique que $E(gc_t|I_{t-1}) = E(gc_t)$, où I_{t-1} correspond à l'ensemble de l'information connue à la période $(t-1)$; et dans ce cas, t correspond à une année.

(i) Testez l'hypothèse du revenu permanent en estimant $gc_t = \beta_0 + \beta_1 gc_{t-1} + u_t$. Exprimez clairement l'hypothèse nulle et les hypothèses alternatives. Que pouvez-vous en conclure ?

(ii) Pour tester la régression de la question (i) ajoutez les variables gy_{t-1}, $i3_{t-1}$, et inf_{t-1}. Est-ce que ces variables sont individuellement, ou conjointement, significatives avec un seuil de confiance de 5 % ? (Attention à reporter les *p*-values appropriées).

(iii) Dans la régression de la question (ii), commentez la *p*-value et le *statistique t* de gc_{t-1} ? Est-ce que cela signifie que la théorie du revenu permanent est maintenant vérifiée par les données ?

(iv) Dans la régression de la question (ii), à combien est égal le *F*-stat et quelle est la valeur de la *p*-value associée ? Est-ce que vos conclusions à propos de la théorie du revenu permanent sont maintenant en accord avec les résultats de la question (i) ?

C8. Utilisez les données du fichier PHILLIPS pour cet exercice.

(i) Estimez un modèle AR(1) classique pour l'équation du taux de chômage. Utilisez cette équation pour prévoir le taux de chômage de l'année 2004. Comparez cette prévision avec le taux de chômage réel de 2004 (Vous pouvez trouver ce chiffre dans un rapport récent du *Economic Report of the President.*)

(ii) Ajoutez un retard d'inflation au modèle AR(1) de la partie (i). La variable inf_{t-1} est-elle statistiquement significative ?

(iii) Utilisez l'équation de la question (ii) pour prévoir le taux de chômage en 2004. Est-ce que cette prévision est meilleure ou moins bonne que celle de la question (i) ?

(iv) Utilisez la méthode de la section 6.4 pour construire un intervalle de confiance à 95 % concernant le taux de chômage de 2004. Le taux de chômage réel de 2004 est-il compris dans cet intervalle ?

C9. Utilisez les données du fichier TRAFFIC2 pour cet exercice (l'exercice C11 du chapitre 10 était basé sur ces données).

(i) Calculez le coefficient d'autocorrélation de premier ordre de la variable *prcfat*. Êtes-vous interloqué par le fait que *prcfat* contienne une racine unitaire ? Faites la même chose pour le taux de chômage.

(ii) Estimez un modèle de régression multiple reliant la première différence de *prcfat*, $\Delta prcfat$, aux mêmes variables que dans la question (vi) de l'exercice C11 du chapitre 10 (en différenciant aussi dans un premier temps le taux de chômage). Ensuite, insérez une tendance, des variables indicatrices mensuelles, une variable de week-end et deux variables de politique monétaire (ne pas différencier ces variables). Commentez les résultats de cette régression.

(iii) Commentez l'affirmation suivante : « Nous devons toujours différencier une série temporelle si nous pensons que celle-ci comporte une racine unitaire avant de faire une régression multiple, car cette stratégie est sûre et permet d'obtenir des résultats très similaires par rapport à une régression en niveau » [Pour répondre à cette question, vous pouvez faire la régression de la question (vi) de l'exercice C11 du chapitre 10, si vous ne l'avez pas déjà faite.]

C10. Utilisez toutes les données du fichier PHILLIPS pour répondre à cette question (56 années de données).

(i) Ré-estimez l'équation (11.19) et indiquez vos résultats sous la forme usuelle. Est-ce que les estimations de l'ordonnée à l'origine et de la pente sont passablement modifiées lorsque vous ajoutez les années récentes à l'échantillon ?

(ii) Obtenez un nouvel estimateur du taux de chômage naturel, et comparez le avec celui estimé dans l'exemple 11.5.

(iii) Calculez l'autocorrélation de premier ordre de *unem*. À votre avis, est-ce que la racine est proche de 1 ?

(iv) Utilisez *cunem* comme variable explicative à la place de *unem*. Quelle variable explicative donne le *R*-carré le plus élevé ?

C11. La loi d'Okun – voir, par exemple, Mankiw (1994, chapitre 2) – implique la relation suivante entre le pourcentage de variation annuel du PIB réel, *pcrgdp*, et le taux de chômage, *cunem* :

$$pcrgdp = 3 - 2 * cunem.$$

Si le taux de chômage est stable, le PIB réel augmente de 3 % par an. Pour chaque point de pourcentage de hausse du chômage, le PIB réel diminue de deux points de pourcentage (Attention, il ne faut pas voir ici un lien de causalité, mais simplement une description statistique)
Pour voir si les données de l'économie américaine confirment la loi d'Okun, nous spécifions un modèle qui permet une déviation via un terme d'erreur, $pcrgdp_t = \beta_0 + \beta_1 cunem_t + u_t$.

(i) Utilisez les données du fichier OKUN pour estimer l'équation. Obtenez-vous exactement 3 pour l'ordonnée à l'origine et -2 pour la pente ? Vous attendiez-vous à cela ?

(ii) Trouvez le *t*-stat pour tester $H_0 : \beta_1 = -2$. Est-ce qu'il est possible de rejeter H_0 contre l'hypothèse bilatérale alternative, avec un seuil de confiance raisonnable ?

(iii) Trouvez le *t*-stat pour tester $H_0 : \beta_0 = 3$. Est-ce que vous rejeter H_0 avec un seuil de confiance de 5 % contre l'hypothèse bilatérale alternative ? Est-ce un rejet « fort » ?

(iv) Trouvez le *F*-stat et la *p*-value pour tester $H_0 : \beta_0 = 3$, $\beta_1 = -2$ contre l'hypothèse alternative que H_0 soit faux. Est-il possible de rejeter le test avec un seuil de confiance de 10 % ? Au final, diriez-vous que les données tendent à supporter ou à rejeter la loi d'Okun ?

C12. Utilisez les données du fichier MINWAGE pour cet exercice, en vous concentrant sur les variables concernant le salaire et l'emploi dans le secteur 232 (vêtements pour hommes et garçons). La variable *gwage*232 correspond à la variation mensuelle (différence du log) du salaire moyen dans le secteur 232 ; *gemp*232 est la variation de l'emploi dans le secteur 232 ; *gmwage* est la variation du salaire minimum fédéral et *gcpi* est la variation de l'Indice des Prix à la Consommation.

(i) Trouvez l'autocorrélation de premier ordre de *gwage*232. Est-ce que cette série présente de la faible dépendance ?

(ii) Estimez le modèle dynamique :

$$gwage232_t = \beta_0 + \beta_1 gwage232_{t-1} + \beta_2 gmwage_t + \beta_3 gcpi_t + u_t$$

par la méthode des MCO. En gardant constant la variation du salaire du mois précédent et la variation de l'IPC, est-ce qu'une hausse du salaire minimum fédéral entraîne une hausse simultanée de $gwage232_t$? Justifiez.

(iii) Maintenant, ajoutez un retard de la croissance de l'emploi, $gemp232_{t-1}$, à l'équation de la question (ii). Cette variable est-elle significative ?

(iv) En comparant avec le modèle sans $gwage232_{t-1}$ et $gemp232_{t-1}$, est-ce que l'ajout de deux variables retardées a un impact important sur le coefficient *gmwage* ?

(v) Lancez la régression de $gmwage_t$ sur $gwage232_{t-1}$ et $gemp232_{t-1}$, et indiquez le *R*-carré. Comment la valeur du R-carré vous permet de justifier votre réponse à la question (iv).

C13. Utilisez les données du fichier BEVERIDGE pour répondre à ces questions. Ce fichier contient des données à propos du taux de vacance et du taux de chômage aux USA, entre décembre 2000 et février 2001.

(i) Trouvez la corrélation entre *urate* et *urate*_1. Diriez-vous que la corrélation ressemble davantage à un processus à racine unitaire ou à un processus de faible dépendance ?

(ii) Répétez (i) mais en utilisant le taux d'emplois vacants, *vrate*.

(iii) La courbe de Beveridge relie le taux de chômage au taux d'emplois vacants. Une version linéaire simple de cette relation peut s'écrire :

$$urate_t = \beta_0 + \beta_1 vrate_t + u_t$$

où $\beta_1 < 0$ est attendu. Estimez β_0 et β_1 en utilisant la méthode des MCO, et reportez vos résultats sous la forme usuelle. Trouvez-vous bien une relation négative ?

(iv) Expliquez pourquoi vous ne pouvez-pas faire confiance aux intervalles de confiance affichés pour β_1 dans la question (iii). [Les outils nécessaires à l'étude de ce type de régression seront présentés dans le chapitre 18]

(v) En calculant la première différence de *urate* et *vrate* avant de lancer la régression, comment la pente varie t-elle par rapport à la question (iii) ? Est-ce que le coefficient est statistiquement différent de zéro ? [Cet exemple montre que différencier les variables avant d'utiliser la méthode des MCO n'est pas toujours une bonne stratégie. Nous verrons cela plus en détail dans le chapitre 18.]

C14. Utilisez les données du fichier APPROVAL pour répondre aux questions suivantes (voir également l'exercice C14 du chapitre 10)

(i) Calculez les autocorrélations d'ordre 1 pour les variables *approve* et *lrgasprice*. Sont-elles assez proches de l'unité pour s'inquiéter de la présence de racines unitaires ?

(ii) Considérons le modèle

$$approve_t = \beta_0 + \beta_1 lcpifood_t + \beta_2 lrgasprice_t + \beta_3 unemploy_t + \beta_4 sep11_t + \beta_5 iraqinvade_t + u_t,$$

où les deux premières variables sont exprimées en logarithme. Compte tenu des résultats obtenus à la question (i), pourquoi hésiteriez-vous à estimer ce modèle par MCO ?

(iii) Estimez l'équation de la question (ii) en différenciant toutes les variables (même les variables indicatrices). Comment interprétez-vous votre estimation de β_2 ? Est-elle statistiquement significative ? (Rapportez la *p*-valeur).

(iv) Interprétez votre estimation de β_4 et discutez de la significativité statistique.

(v) Ajoutez *lsp*500 au modèle décrit à la question (ii) et estimez l'équation en différence première. Discutez vos résultats pour la variable de marché.

CHAPITRE

12

CORRÉLATION SÉRIELLE ET HÉTÉROSCÉDASTICITÉ DANS L'ANALYSE DES SÉRIES TEMPORELLES

Traduction de Mikael Petitjean

Dans ce chapitre consacré aux séries temporelles, nous abordons le problème majeur de la corrélation sérielle au sein du terme d'erreur. Nous avons vu dans le chapitre précédent que les erreurs ne sont pas autocorrélées lorsque la spécification du modèle tient correctement compte de sa dynamique temporelle. Tester la présence de corrélation sérielle peut donc servir à déterminer si la dynamique d'un modèle a été spécifiée comme il se doit. Notez cependant que les erreurs des modèles statiques sont souvent autocorrélées même lorsque la spécification du modèle est la bonne. C'est également le cas pour les modèles à retards échelonnés finis. Il est donc important d'identifier les conséquences indésirables de la corrélation sérielle pour être capable de bien y remédier par la suite, en particulier pour ce type de modèles dont l'utilité n'est plus à démontrer.

Dans la section 12.1, nous montrons que certaines propriétés des MCO sont affectées par la présence de corrélation sérielle dans les erreurs. Dans la section 12.2, nous en identifions les différentes méthodes de détection. Dans un premier temps, nous abordons les tests qui ne s'appliquent qu'aux modèles dont les régresseurs sont strictement exogènes. Dans un second temps, nous étudions les tests valides asymptotiquement, dans lesquels les régresseurs peuvent prendre des formes plus générales, dont celle de la variable dépendante retardée. Dans la section 12.3, nous introduisons les méthodes de correction des erreurs en présence de corrélation sérielle. La section 12.4 montre que l'utilisation de variables en différence permet souvent de faire disparaître l'autocorrélation présente dans les erreurs. La section 12.5 est consacrée aux développements plus récents en matière d'ajustement des écarts-types estimés lorsque l'autocorrélation prend une forme très générale.

Dans le chapitre 8, nous avons discuté des méthodes de détection et de correction de l'hétéroscédasticité en présence de données en coupe transversale. Dans la section 12.6, nous montrons que ces méthodes peuvent être adaptées au cas particulier des séries temporelles. La procédure est fondamentalement similaire mais il y a un certain nombre de subtilités propres à l'autocorrélation temporelle dont il faudra tenir compte. Enfin, nous discutons très brièvement des conséquences liées à la présence de formes dynamiques d'hétéroscédasticité.

12.1 PROPRIÉTÉS DES MCO EN PRÉSENCE D'ERREURS AUTOCORRÉLÉES

Absence de biais et convergence

Dans le chapitre 10, nous avons démontré l'absence de biais des estimateurs des MCO sous les trois premières hypothèses de Gauss-Markov propres aux séries temporelles (de ST.1 à ST.3). Nous pouvons constater que le théorème 10.1 ne pose aucune restriction quant à la corrélation sérielle des erreurs. Il s'en suit que les $\hat{\beta}_j$ sont sans biais quel que soit le degré de corrélation sérielle des erreurs, à condition néanmoins que les variables explicatives soient strictement exogènes. De manière équivalente, les $\hat{\beta}_j$ ne sont pas biaisés en présence d'erreurs hétéroscédastiques.

Dans le chapitre 11, nous avons relâché l'hypothèse d'exogénéité stricte pour utiliser plutôt $E(u_t|\mathbf{x}_t) = 0$; lorsque les données sont faiblement dépendantes, nous avons montré que les $\hat{\beta}_j$ sont convergents, sans être nécessairement sans biais. Ces conclusions ne dépendent en rien de la nature de la corrélation sérielle dans les erreurs.

Efficacité et inférence

Étant donné que le théorème de Gauss-Markov (théorème 10.4) exige un terme d'erreur homoscédastique et exempt de toute corrélation sérielle, les estimateurs des MCO ne sont plus *BLUE* en présence d'erreurs

autocorrélées. Plus important encore, les écarts-types estimés des MCO et les tests d'hypothèse ne sont plus valides, même asymptotiquement. Nous pouvons le vérifier aisément dans le cas d'un modèle de **corrélation sérielle AR(1)** pour les erreurs. Supposons que

$$u_t = \rho u_{t-1} + e_t, \qquad t = 1, 2, \ldots, n \tag{12.1}$$

$$|\rho| < 1, \tag{12.2}$$

où les e_t sont des variables aléatoires non corrélées dont l'espérance est nulle et la variance est égale à σ_e^2 ; dans le chapitre 11, nous avons vu que l'hypothèse (12.2) représente la condition de stabilité.

Sous les quatre premières hypothèses de Gauss-Markov, la variance de l'estimateur des MCO pour la pente est calculée à partir du modèle de régression linéaire simple suivant :

$$y_t = \beta_0 + \beta_1 x_t + u_t.$$

Sans perte de généralité, nous supposons que la moyenne des x_t dans l'échantillon est nulle ($\bar{x} = 0$). L'estimateur des MCO $\hat{\beta}_1$ est donc égal à

$$\hat{\beta}_1 = \beta_1 + SCT_x^{-1} \sum_{t=1}^{n} x_t u_t, \tag{12.3}$$

où $SCT_x = \sum_{t=1}^{n} x_t^2$. Dans le calcul de $\hat{\beta}_1$ (étant donné $\mathbf{X}$), nous devons maintenant tenir compte de la corrélation sérielle dans les u_t :

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= SCT_x^{-2} \text{Var}\left(\sum_{t=1}^{n} x_t u_t\right) \\ &= SCT_x^{-2} \left(\sum_{t=1}^{n} x_t^2 \text{Var}(u_t) + 2 \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} x_t x_{t+j} \, E(u_t u_{t+j}) \right) \\ &= \frac{\sigma^2}{SCT_x} + 2 \left(\frac{\sigma^2}{SCT_x^2} \right) \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} \rho^j x_t x_{t+j} \end{aligned} \tag{12.4}$$

où $\sigma^2 = \text{Var}(u_t)$. Notez que nous avons utilisé le fait que $\text{E}(u_t \, u_{t+j}) = \text{Cov}(u_t, \, u_{t+j}) = \rho^j \sigma^2$ [voir l'équation (11.4)]. Le premier terme de l'équation (12.4), σ^2/SCT_x, représente la variance de $\hat{\beta}_1$ lorsque $\rho = 0$, ce qui correspond à la variance classique des MCO sous les hypothèses de Gauss-Markov. Si nous calculons la variance sur cette base, son estimateur sera biaisé lorsque $\rho \neq 0$ car nous ne tenons pas compte du second terme de l'équation (12.4). Comme nous le verrons dans plusieurs exemples, il est fréquent d'obtenir $\rho > 0$, auquel cas $\rho^j > 0$ pour tout j. De plus, dans les modèles de séries temporelles, les variables indépendantes affichent souvent une corrélation positive au cours du temps de telle sorte que le produit $x_t x_{t+j}$ est positif pour la plupart des paires t et $t + j$. Par conséquent, dans la plupart des applications économiques, le terme $\sum_{t=1}^{n-1} \sum_{j=1}^{n-t} \rho^j x_t x_{t+j}$ est positif et l'emploi de la formule traditionnelle de la variance des MCO, σ^2/SCT_x, conduit à sous-évaluer la vraie variance de l'estimateur des MCO. Si ρ est élevé et que x_t affiche une forte corrélation sérielle positive (ce qui est classique), l'estimateur standard de la variance peut se révéler particulièrement biaisé. L'erreur est alors de surévaluer la précision de l'estimateur du coefficient de pente des MCO.

Si $\rho < 0$, ρ^j est soit négatif (lorsque j est impair), soit positif (lorsque j est pair). Dans ces conditions, il est difficile de déterminer le signe du terme $\sum_{t=1}^{n-1}\sum_{j=1}^{n-t}\rho^j x_t x_{t+j}$. Il est néanmoins possible que l'emploi de la formule classique de la variance de l'estimateur des MCO nous conduise à surévaluer la vraie variance de $\hat{\beta}_1$. Dans tous les cas de figure, en présence de corrélation sérielle, l'utilisation de σ^2/SCT_x est inappropriée car elle nous donne une valeur biaisée de $\text{Var}(\hat{\beta}_1)$.

Étant donné que l'écart-type estimé de $\hat{\beta}_1$ est une estimation de la racine carrée de la variance de $\hat{\beta}_1$, l'estimation classique des MCO, $\hat{\sigma}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{SCT_x}$, n'est pas non plus valide en présence de corrélation sérielle. Par conséquent, les statistiques t basées sur la distribution de Student sont fausses et ne peuvent pas servir à effectuer des tests d'hypothèse simple. Sachant que la sous-estimation de l'écart-type estimé revient à surestimer les statistiques t, nous aurons tendance à rejeter trop souvent l'hypothèse nulle lorsque $\rho > 0$. La conclusion est la même pour les tests d'hypothèses multiples basés sur les statistiques F et ML, par exemple.

Pour aller plus loin 12.1

Supposez qu'au lieu du modèle AR(1), u_t suive un processus MA(1) tel que $u_t = \alpha e_{t-1} + e_t$. Trouvez $\text{Var}(\hat{\beta}_1)$ et montrez qu'elle est différente de σ^2/SCT_x lorsque $\alpha \neq 0$.

Qualité d'ajustement

On peut parfois lire que la présence de corrélation sérielle dans les erreurs d'un modèle de séries temporelles fausse les mesures de qualité d'ajustement, telles que le R carré et le R carré ajusté. Ce n'est heureusement pas le cas si les données sont stationnaires et faiblement dépendantes. Pour le comprendre, il est nécessaire de se remémorer la formule du R carré que nous avons utilisée dans le cadre des données en coupe transversale. Pour la population, $R^2 = 1 - \sigma_u^2/\sigma_y^2$ (voir la section 6.3). Cette définition reste valide dans le cadre de séries temporelles stationnaires et faiblement dépendantes, car les variances de l'erreur et de la variable dépendante ne changent pas au cours du temps. Grâce à la loi des grands nombres, on peut obtenir des estimations convergentes du R carré de la population. L'explication est la même dans le contexte des données en coupe transversale en présence d'hétéroscédasticité (voir la section 8.1). Étant donné qu'il n'existe pas d'estimateur sans biais du R carré de la population, cela n'a pas de sens de parler d'un R^2 biaisé par la présence de corrélation sérielle. Nous pouvons simplement affirmer que nos mesures de qualité d'ajustement continuent de représenter des estimateurs convergents du paramètre de la population. Cette conclusion n'est pas vraie si $\{y_t\}$ suit un processus I(1) puisque $\text{Var}(y_t)$ croît en fonction de t ; les mesures de qualité d'ajustement n'ont d'ailleurs pas beaucoup de sens dans un tel cas. Comme nous l'avons également vu dans la section 10.5, lorsqu'il s'agit de calculer le R carré, il est nécessaire de tenir compte de la tendance ou de la saisonnalité que la moyenne de y_t peut afficher. En dehors de ces difficultés, l'interprétation du R^2 et de sa version ajustée est la même qu'auparavant.

Corrélation sérielle en présence d'une variable dépendante retardée

Les étudiants qui découvrent l'économétrie sont souvent avertis des dangers qu'implique l'utilisation d'une variable dépendante retardée en tant que régresseur. Les manuels d'économétrie contiennent très souvent un avertissement indiquant que « les estimateurs des MCO ne sont pas convergents lorsque la variable dépendante est retardée et que les erreurs sont autocorrélées ». En réalité, cette affirmation est très imprécise et n'est valable que dans des situations plutôt singulières.

Pour le démontrer, supposons que la valeur attendue de y_t étant donné y_{t-1} soit linéaire :

$$E(y_t|y_{t-1}) = \beta_0 + \beta_1 y_{t-1}, \quad [12.5]$$

en respectant la condition de stabilité, $|\beta_1| < 1$. Si nous incluons un terme d'erreur, nous pouvons écrire

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t, \quad [12.6]$$

$$E(u_t|y_{t-1}) = 0. \quad [12.7]$$

Par construction, ce modèle respecte l'hypothèse ST.3′ (nullité de l'espérance conditionnelle de l'erreur) requise pour démontrer la convergence des MCO. Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ des MCO sont donc convergents. Néanmoins, il est important de noter que les erreurs $\{u_t\}$ *peuvent* être autocorrélées. Sous la condition (12.7), u_{t-1} n'est pas corrélée avec y_{t-1} mais rien n'empêche u_t d'être corrélée avec y_{t-2}. En effet, puisque $u_{t-1} = y_{t-1} - \beta_0 - \beta_1 y_{t-2}$, la covariance entre u_t et u_{t-1} est égale au terme $-\beta_1 \text{Cov}(u_t, y_{t-2})$ dont la valeur n'est pas nécessairement nulle. En conclusion, le modèle contient une variable dépendante retardée ; les erreurs peuvent souffrir de corrélation sérielle ; et les MCO permettent, malgré tout, d'obtenir des estimations convergentes de β_0 et de β_1. Certes, la corrélation sérielle dans les erreurs invalide les tests d'hypothèse mais elle n'a pas d'impact sur la propriété de convergence des estimateurs.

Dans quelles circonstances exactes pouvons-nous dès lors affirmer que les estimateurs des MCO ne sont pas convergents lorsque les erreurs sont autocorrélées et que la variable dépendante est retardée ? Nous pouvons l'affirmer lorsque le même modèle (12.6) est accompagné de l'hypothèse selon laquelle $\{u_t\}$ suit un processus AR(1) stable, tel que décrit en (12.1) et (12.2). Cette hypothèse implique que

$$E(e_t|u_{t-1}, u_{t-2}, \ldots) = E(e_t|y_{t-1}, y_{t-2}, \ldots) = 0. \quad [12.8]$$

Étant donné que e_t n'est pas corrélée avec y_{t-1} (par hypothèse), $\text{Cov}(y_{t-1}, u_t) = \rho \text{Cov}(y_{t-1}, u_{t-1})$; cette covariance ne sera jamais nulle, sauf si $\rho = 0$. Or, $\{u_t\}$ suit un processus AR(1) stable. Cela implique que les estimateurs de β_0 et de β_1 provenant de la régression de y_t sur y_{t-1} ne sont pas convergents.

Lorsque les erreurs u_t suivent un processus AR(1), le modèle (12.6) ne peut disposer d'estimateurs convergents. L'exactitude de cette affirmation ne la rend pas moins insolite. Quel est l'intérêt d'estimer les paramètres de (12.6) tout en supposant que les erreurs suivent un processus AR(1) ? Il n'y en a guerre. En réalité, si nous combinons (12.6) et (12.1), nous pouvons montrer que y_t suit un modèle autorégressif d'ordre 2, soit un modèle AR(2). Utilisons l'expression $u_{t-1} = y_{t-1} - \beta_0 - \beta_1 y_{t-2}$ dans laquelle nous insérons (12.1), soit $u_t = \rho u_{t-1} + e_t$. (12.6) peut donc s'écrire

$$\begin{aligned} y_t &= \beta_0 + \beta_1 y_{t-1} + \rho(y_{t-1} - \beta_0 - \beta_1 y_{t-2}) + e_t \\ &= \beta_0(1 - \rho) + (\beta_1 + \rho) y_{t-1} - \rho\beta_1 y_{t-2} + e_t \\ &= \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + e_t, \end{aligned}$$

où $\alpha_0 = \beta_0(1 - \rho)$, $\alpha_1 = \beta_1 + \rho$ et $\alpha_2 = -\rho\beta_1$. Étant donné (12.8), il s'ensuit que

$$E(y_t|y_{t-1}, y_{t-2}, \ldots) = E(y_t|y_{t-1}, y_{t-2}) = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2}. \quad [12.9]$$

Cela signifie que l'espérance de y_t, étant donné toutes les valeurs de y dans le passé, ne dépend que des *deux* premiers retards de y. En pratique, c'est l'équation (12.9) qu'il faudrait utiliser, notamment pour générer des prévisions, ce que nous ferons au chapitre 18. Sous les hypothèses désirables de stabilité portant sur les paramètres α_j (que nous détaillons dans la section 12.3), le modèle AR(2) peut être estimé par les MCO et offrir aux paramètres α_j des estimateurs convergents, normalement distribués sur le plan asymptotique.

En résumé, il est difficile de justifier sur le plan pratique l'utilisation d'un modèle dont la variable dépendante est retardée et dont les erreurs doivent suivre un processus autorégressif bien spécifique, tel que le processus AR(1). Très souvent, la présence de corrélation sérielle dans les erreurs indique que la forme fonctionnelle du modèle dynamique n'a pas été correctement spécifiée. Dans le cas que nous venons d'étudier, où la variable dépendante *et* les erreurs suivent un modèle AR(1), il aurait fallu ajouter y_{t-2} dans l'équation (12.6) pour obtenir une forme fonctionnelle correctement spécifiée.

Dans le chapitre 18, nous verrons plusieurs modèles dont la variable dépendante est retardée et dont les erreurs souffrent de corrélation sérielle. Même dans ces cas, les erreurs ne sont pas censées suivre un processus autorégressif bien particulier.

12.2 LA DÉTECTION DE L'AUTOCORRÉLATION

Dans cette section, nous décrivons plusieurs tests permettant de détecter la présence de corrélation sérielle dans les erreurs. Nous utilisons le modèle de régression multiple classique

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t.$$

Considérons dans un premier temps le cas où les variables explicatives sont strictement exogènes. Pour qu'elles le soient, l'erreur, u_t, ne doit pas être corrélée avec les variables explicatives, quelle que soit la période de temps considérée (voir la section 10.3). Cela exclut, entre autres, les modèles dont la variable dépendante est retardée.

Test t de détection de l'autocorrélation d'ordre 1 en présence de régresseurs strictement exogènes

Dans un modèle de régression multiple, la corrélation sérielle dans les erreurs peut apparaître sous différentes formes. Parmi les modèles dont les erreurs sont autocorrélées, le modèle le plus populaire et le plus facile à manipuler est le modèle AR(1), décrit par les équations (12.1) et (12.2). Dans la section précédente, nous avons identifié les conséquences liées à l'utilisation de la méthode des MCO lorsque les erreurs sont autocorrélées ; nous avons également dérivé la formule de la variance pour l'estimateur de la pente des MCO dans le cadre d'un modèle de régression linéaire simple dont les erreurs suivent un processus AR(1). Nous décrivons ici une manière de détecter la présence de corrélation sérielle de type AR(1). S'il n'y a *pas* de corrélation sérielle, l'hypothèse nulle n'est pas rejetée. À l'instar des tests d'hétéroscédasticité, l'hypothèse nulle représente le cas idéal ; nous attendons que les données nous fournissent des indices suffisamment clairs avant de rejeter l'hypothèse nulle d'absence d'autocorrélation.

Le premier test AR(1) que nous envisageons est un test asymptotique dans lequel les variables explicatives sont strictement exogènes : l'espérance de u_t est nulle, quelles que soient les valeurs affichées par chacune des variables explicatives dans le temps. En fonction de (12.1), nous devons également supposer que

$$\mathrm{E}(e_t|u_{t-1}, u_{t-2}, \ldots) = 0 \qquad \textbf{[12.10]}$$

et

$$\mathrm{Var}(e_t|u_{t-1}) = \mathrm{Var}(e_t) = \sigma_e^2. \qquad \textbf{[12.11]}$$

Il s'agit des hypothèses classiques du modèle AR(1), qui découlent du fait que $\{e_t\}$ est une séquence indépendante et identiquement distribuée ; ces hypothèses nous permettent de recourir à la théorie asymptotique que nous avons appliquée aux régressions dynamiques dans le chapitre 11.

Comme dans les tests d'hétéroscédasticité, l'hypothèse nulle n'est valide que si les hypothèses de Gauss-Markov sont respectées. Dans le modèle AR(1), l'hypothèse nulle selon laquelle les erreurs ne sont pas autocorrélées est

$$H_0 : \rho = 0. \qquad \textbf{[12.12]}$$

Comment pouvons-nous tester cette hypothèse ? Sous les conditions (12.10) et (12.11), en supposant que les u_t soient observables, nous pourrions directement appliquer le théorème 11.2 selon lequel les estimateurs des MCO dans une régression dynamique suivent asymptotiquement une loi normale. Dans ce cas,

$$u_t = \rho u_{t-1} + e_t, \; t = 2, \ldots, n. \qquad \textbf{[12.13]}$$

(Sous l'hypothèse nulle $\rho = 0$, $\{u_t\}$ affiche évidemment une faible dépendance.) En d'autres termes, nous pourrions estimer ρ à partir de la régression de u_t sur u_{t-1}, pour tout $t = 2, \ldots, n$, sans constante, et utiliser la statistique t obtenue pour ρ. Comme vous le savez maintenant, cela est impossible pour la simple raison que les erreurs u_t ne peuvent pas être observées. Néanmoins, comme nous l'avons fait pour les tests d'hétéroscédasticité, nous pouvons remplacer u_t par le résidu lui correspondant, $\hat{u}_t$, obtenu par les MCO. Comme $\hat{u}_t$ dépend des estimateurs des MCO $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$, l'utilisation de $\hat{u}_t$ au lieu de u_t dans la régression peut avoir des conséquences sur la distribution de la statistique t. Fort heureusement, grâce à l'hypothèse de stricte exogénéité, il s'avère que la distribution asymptotique de la statistique t, en présence d'un grand échantillon, n'est pas affectée par l'utilisation des résidus au lieu des erreurs. Une démonstration de ce type est trop élaborée pour être incluse dans un ouvrage d'introduction à l'économétrie comme celui-ci ; elle découle de l'article de Wooldridge (1991b).

Nous pouvons résumer le test asymptotique AR(1) de corrélation sérielle comme suit.

Tester la présence de corrélation sérielle AR(1) en présence de régresseurs strictement exogènes

i. Effectuer une régression des MCO de y_t sur x_{t1}, ..., x_{tk} et obtenir les résidus, $\hat{u}_t$, pour tout $t = 1, 2, \ldots, n$.

ii. Effectuer une régression de

$$\hat{u}_{it} \text{ sur } \hat{u}_{t-1}, \text{ pour tout } t = 2, \ldots, n, \qquad \textbf{[12.14]}$$

dans le but d'estimer le coefficient $\hat{\rho}$ de la variable $\hat{u}_{t-1}$ ainsi que la statistique $t_{\hat{\rho}}$. (Notez qu'une constante dans cette régression n'est pas indispensable ; la statistique t pour ρ n'en dépend que légèrement et elle sera asymptotiquement valide dans les deux cas.)

iii. Utiliser $t_{\hat{\rho}}$ pour tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$ de la manière habituelle. (En réalité, comme $\rho > 0$ est souvent la valeur attendue a priori, l'hypothèse alternative peut s'écrire $H_1 : \rho > 0$.) En règle générale, nous concluons que la corrélation sérielle pose problème lorsque H_0 est rejeté à un seuil de 5 %. Comme d'habitude, il est préférable d'indiquer la p-valeur du test.

Pour déterminer s'il est nécessaire de corriger les procédures d'inférence statistique en présence de corrélation sérielle, il est important de ne pas confondre significativité statistique et importance pratique. Par exemple, si l'échantillon est de grande taille, on peut rejeter l'hypothèse d'absence d'autocorrélation tout en constatant que la valeur de $\hat{\rho}$ est marginale sur le plan pratique ; si $\hat{\rho}$ est proche de zéro, cela n'a pas d'impact tangible sur les tests d'inférence statistique [voir l'équation (12.4)]. Il s'agirait là

de l'exception plutôt que de la règle car les séries temporelles en sciences sociales sont souvent relativement courtes.

EXEMPLE 12.1
Détecter la présence de corrélation sérielle d'ordre 1 dans la courbe de Phillips

Dans le chapitre 10, nous avons estimé une version statique de la courbe de Phillips qui modélise le compromis entre inflation et chômage aux États-Unis (voir l'exemple 10.1). Dans le chapitre 11, nous avons également étudié une version de la courbe de Phillips caractérisée par des anticipations adaptatives (voir l'exemple 11.5). Nous vérifions maintenant si le terme d'erreur de ces deux modèles souffre de corrélation sérielle. Dans la version basée sur les anticipations, nous avons une observation en moins puisque l'équation du modèle est $\Delta inf_t = inf_t - inf_{t-1}$.

Pour le modèle statique (basé sur 48 observations, jusqu'en 1996), les résultats de la régression (12.14) donnent $\hat{\rho} = 0{,}573$ et $t = 4{,}93$, avec une p-valeur $< 0{,}001$. Il est donc plus que probable que les erreurs de ce modèle soient affectées par une corrélation sérielle positive d'ordre 1. Une conséquence fâcheuse est que les écarts-types estimés et les statistiques t du chapitre 10 ne sont pas valides. Par contre, dans le modèle basé sur les anticipations, le test AR(1) de corrélation sérielle donne $\hat{\rho} = -0{,}036$, $t = -0{,}287$, p-valeur $= 0{,}775$ (avec 47 observations). Rien n'indique que la courbe de Phillips caractérisée par des anticipations adaptatives souffre de corrélation sérielle d'ordre 1 dans les erreurs.

Bien que le test de (12.14) repose sur le modèle AR(1), il permet également de détecter d'autres formes de corrélation sérielle. Souvenez-vous que $\hat{\rho}$ est un estimateur convergent de la corrélation entre u_t et u_{t-1}. Par conséquent, toute forme de corrélation sérielle qui se traduit par une corrélation non nulle entre deux erreurs adjacentes peut être détectée par ce test. Par contre, si la corrélation sérielle prend une autre forme, le test sera incapable de la détecter. Par exemple, ce sera le cas lorsque u_t et u_{t-2} sont corrélées mais que $\text{Corr}(u_t, u_{t-1}) = 0$.

La statistique t de la régression (12.14) n'est fiable que si les erreurs de l'équation (12.13) respectent l'hypothèse d'homoscédasticité décrite en (12.11). En réalité, il est aisé de rendre ce test valide en présence d'hétéroscédasticité dans e_t. Nous devons simplement utiliser la statistique t robuste à la présence d'hétéroscédasticité, définie au chapitre 8. Pour la courbe de Phillips qui est estimée dans l'exemple 12.1, la version robuste de la statistique t est égale à 4,03. Cette valeur est plus petite que celle obtenue pour sa version non robuste mais toutes deux aboutissent néanmoins à des p-valeurs inférieures à 1 %. Dans la Section 12.6, nous approfondirons notre analyse de l'hétéroscédasticité dans les régressions sur séries temporelles, en y introduisant des formes dynamiques d'hétéroscédasticité.

Pour aller plus loin 12.2

Comment pouvons-nous construire un intervalle de confiance à 95 % pour ρ en utilisant la régression (12.4) ?

Le test de Durbin-Watson

Un autre test AR(1) de corrélation sérielle est celui développé par Durbin et Watson. La **statistique de Durbin-Watson (*DW*)** est également basée sur les résidus des MCO :

$$DW = \sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2 / \sum_{t=1}^{n}(\hat{u}_t)^2 \qquad \textbf{[12.15]}$$

Un rapide calcul permet de montrer que DW est fortement lié à $\hat{\rho}$ tiré de (12.14). En effet,

$$DW \approx 2(1 - \hat{\rho}). \qquad \textbf{[12.16]}$$

Une raison qui explique l'absence de relation exacte entre les deux est que le dénominateur de $\hat{\rho}$ est $\sum_{t=2}^{n}(\hat{u}_{t-1})^2$ alors que celui de DW correspond à la somme des carrés des résidus. L'approximation (12.16) reste néanmoins fiable même lorsque la taille de l'échantillon est modérée. Par conséquent, le test basé sur DW et le test t basé sur $\hat{\rho}$ sont conceptuellement identiques.

Durbin et Watson (1950) ont cherché à dériver la distribution exacte de DW (conditionnellement à $\mathbf{X}$), ce qui exige le respect de l'ensemble des hypothèses du modèle linéaire classique, y compris l'hypothèse de normalité des erreurs. Cette distribution est difficile à obtenir car elle dépend de manière compliquée des valeurs de $\mathbf{X}$ dans l'échantillon. Elle dépend également de la taille de l'échantillon, du nombre de régresseurs et de la présence d'une constante. Dans certains logiciels économétriques, il est possible d'obtenir les p-valeurs pour DW mais cela reste plutôt rare. En tout état de cause, ces p-valeurs dépendent de l'ensemble des hypothèses du MRLC.

Au lieu de calculer les p-valeurs, certains ouvrages de référence en économétrie indiquent plutôt les valeurs critiques. Contrairement aux test t ou F, il n'y a malheureusement pas de valeur critique unique dans le test de Durbin-Watson. En raison de la difficulté d'obtenir la distribution exacte de DW sous l'hypothèse nulle, Durbin et Watson ont dû dériver une limite inférieure et une limite supérieure. Ces deux limites sont traditionnellement représentées par d_I (pour borne *inférieure*) et d_S (pour borne *supérieure*). Elles dépendent de plusieurs facteurs : le niveau de significativité désiré, la spécification de l'hypothèse alternative, le nombre d'observations et le nombre de régresseurs. (Nous supposons que la constante est incluse dans le modèle.) Néanmoins, ces valeurs critiques ne dépendent pas des p-valeurs prises par les variables explicatives.

En règle générale, le test DW est calculé en fonction de l'hypothèse alternative

$$H_1 : \rho > 0. \qquad \textbf{[12.17]}$$

Sur base de l'approximation (12.16), $\hat{\rho} \approx 0$ implique que $DW \approx 2$ et, lorsque $\hat{\rho} > 0$, $DW < 2$. Le rejet de l'hypothèse nulle (12.12) en faveur de l'hypothèse alternative (12.17) aura donc lieu lorsque la valeur de DW est significativement inférieure à deux. Pour déterminer si tel est le cas, nous sommes contraints de comparer DW à d_I et d_S. Si $DW < d_I$, nous pouvons rejeter H_0 en faveur de (12.17) ; si $DW < d_S$, c'est le cas contraire : H_0 n'est pas rejetée. Si $d_I \leq DW \leq d_S$, le test n'est pas probant et il est impossible de conclure.

Par exemple, si nous choisissons un seuil de significativité égal à 5 % avec $n = 45$ et $k = 4$, nous obtenons $d_S = 1{,}720$ et $d_I = 1{,}336$ [voir Savin et White (1977)]. Si $DW < 1{,}336$, nous pouvons rejeter l'hypothèse nulle d'absence de corrélation sérielle à un niveau de 5 % ; par contre, si $DW > 1{,}72$, H_0 n'est pas rejetée ; enfin, si $1{,}336 \leq DW \leq 1{,}72$, nous ne pouvons tirer aucune conclusion quant à la validité de H_0.

Dans l'exemple 12.1 portant sur l'exemple de la courbe de Phillips statique, le calcul de DW donne une valeur égale à 0,80. À un seuil de significativité de 1 %, Savin et White (1977) nous informent que la valeur critique de la borne inférieure, d_I, est égale à 1,32 lorsque $k = 1$ et $n = 50$. Par conséquent, nous pouvons rejeter l'hypothèse d'absence de corrélation sérielle à un seuil de 1 % en faveur de l'hypothèse alternative de corrélation sérielle positive. (Sur base du test t précédent, nous pouvons également conclure que la p-valeur est inférieure à 0,001, soit 0,1 %.) Par contre, dans le cas de la courbe de Phillips basée sur les anticipations adaptatives, $DW = 1{,}77$, ce qui dépasse largement la valeur critique de la borne supérieure à 1 % ($d_S = 1.40$). L'hypothèse nulle n'est donc pas rejetée.

Par rapport au test t décrit en (12.14), la possibilité d'obtenir une distribution exacte pour DW est le seul avantage du test de Durbin-Watson. Étant donné que les valeurs critiques ne sont exactes que sous l'ensemble des hypothèses du MRLC et qu'il existe une large zone au sein de laquelle aucune conclusion ne peut être tirée, l'utilité de la statistique DW reste relativement limitée. En effet, la statistique t de (12.14) est facile à calculer ; elle est asymptotiquement valide lorsque les erreurs ne sont pas distribuées selon une loi normale ; elle est valide en présence d'hétéroscédasticité liée aux x_{tj} ; et il est facile de la rendre robuste à toute forme d'hétéroscédasticité.

Test t de détection de l'autocorrélation d'ordre 1 en l'absence de régresseurs strictement exogènes

Lorsque les variables explicatives ne sont pas strictement exogènes (de telle sorte qu'au moins une variable x_{tj} est corrélée avec u_{t-1}, par exemple), ni le test t de la régression (12.14) ni le test de Durbin-Watson ne sont valides, même en présence d'un échantillon de grande taille. Ce sera toujours le cas pour les modèles qui contiennent la variable dépendante retardée comme régresseur : y_{t-1} et u_{t-1} sont toujours corrélées. C'est la raison pour laquelle Durbin (1970) a présenté deux versions alternatives de la statistique DW. Elles s'appliquent aux modèles dont les régresseurs peuvent ne pas être strictement exogènes. La première alternative est la statistique h de Durbin. Comme il est parfois impossible de calculer cette statistique, nous ne l'aborderons pas dans cet ouvrage.

La seconde alternative introduite par Durbin est un simple test t, similaire à (12.14), qui reste néanmoins valide en présence d'un nombre quelconque de régresseurs non strictement exogènes. Notez que le test fonctionne également si les variables explicatives sont strictement exogènes.

Tester la présence de corrélation sérielle d'ordre 1 en présence de régresseurs dont l'exogénéité stricte n'est pas requise

i. Effectuer une régression des MCO de y_t sur $x_{t1}, \ldots, x_{tk}$ pour estimer les résidus, $\hat{u}_t$, pour tout $t = 1, 2, \ldots, n$.

ii. Effectuer une seconde régression de

$$\hat{u}_t \text{ sur } x_{t1}, x_{t2}, \ldots, x_{tk}, \hat{u}_{t-1}, \text{ pour tout } t = 2, \ldots, n \qquad \textbf{[12.18]}$$

dans le but d'estimer le coefficient $\hat{\rho}$ de $\hat{u}_{t-1}$ et sa statistique t, $t_{\hat{\rho}}$.

iii. Utiliser $t_{\hat{\rho}}$ pour tester $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$ (ou utiliser une hypothèse unilatérale si nécessaire).

Dans l'équation (12.18), nous régressons les résidus des MCO sur *toutes* les variables indépendantes, y compris la constante et le résidu retardé. La statistique $t_{\hat{\rho}}$ obtenue à partir de l'estimation du coefficient de $\hat{u}_{t-1}$, est un test valide de l'hypothèse (12.12) relative au modèle AR(1) de (12.13) [en précisant que $\text{Var}(u_t|\mathbf{x}_t, u_{t-1}) = \sigma^2$ sous H_0]. Les x_{tj} peuvent inclure plusieurs versions retardées de la variable dépendante ainsi que d'autres variables explicatives qui ne sont pas nécessairement strictement exogènes.

L'inclusion de $x_{t1}, \ldots, x_{tk}$ permet de tenir explicitement compte de la corrélation qui peut exister entre u_{t-1} et chaque x_{tj}. Cela garantit que $t_{\hat{\rho}}$ suit approximativement une distribution t lorsque l'échantillon est de grande taille. Par contre, la statistique t de (12.14) exclut toute forme de corrélation entre les x_{tj} et u_{t-1} ; elle n'est donc pas fiable dès qu'un régresseur n'est pas strictement exogène. Remarquez en passant que nous

aurions pu choisir y_t à la place de $\hat{u}_t$ dans (12.18) sans que cela ne modifie la statistique t relative à $\hat{u}_{t-1}$ puisque $\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \ldots - \hat{\beta}_k x_{tk}$.

Pour que la statistique t de $\hat{u}_{t-1}$ soit valide en présence d'hétéroscédasticité de forme inconnue [en particulier, lorsque $\text{Var}(u_t|\mathbf{x}_t, u_{t-1})$ n'est pas constante], il suffit d'utiliser la statistique t robuste à la présence d'hétéroscédasticité, comme nous l'avons déjà précisé pour le test AR(1) reposant sur des régresseurs strictement exogènes.

EXEMPLE 12.2
Détecter la présence de corrélation sérielle d'ordre 1 dans l'équation du salaire minimum

Au chapitre 10 (voir l'exemple 10.9), nous avons estimé l'effet du salaire minimum sur le taux d'emploi à Porto Rico. Nous vérifions ici si les erreurs souffrent de corrélation sérielle en recourant au test t de Durbin, qui n'exige pas la stricte exogénéité du salaire minimum ($mincov_t$) ou des variables du PNB ($prgnp_t$ et $usgnp_t$). [Par rapport à l'équation (10.38), nous ajoutons $prgnp_t$, le log de la valeur réelle du PNB du Porto Rico, comme nous l'avons fait dans l'exercice sur ordinateur C3 du chapitre 10.] Nous supposons que les processus stochastiques sous-jacents sont stationnaires et faiblement dépendants, tout en tenant compte d'une tendance temporelle linéaire caractérisée par la variable t dans la régression.

Sachant que $\hat{u}_t$ représente les résidus des MCO, nous estimons la régression de

$$\hat{u}_t \text{ sur } \log(mincov_t), \log(prgnp_t), \log(usgnp_t), t, \text{ et } \hat{u}_{t-1},$$

sur base de 37 observations. Le coefficient estimé de $\hat{u}_{t-1}$ est $\hat{\rho} = 0{,}48$, le $t = 2{,}89$ et la p-valeur = 0,007 pour un test bilatéral. On peut donc conclure à la présence de corrélation sérielle d'ordre 1 dans les erreurs ; cela signifie que les statistiques t pour les $\hat{\beta}_j$, que nous utilisons dans les tests d'inférence statistique, ne sont pas fiables. Néanmoins, gardez bien à l'esprit que les estimateurs des $\hat{\beta}_j$ restent convergents à condition que l'exogénéité contemporaine entre u_t et chaque variable explicative soit respectée [sans oublier l'hypothèse de stationnarité et de faible dépendance des processus stochastiques]. Enfin, si nous avions utilisé la régression (12.14) au lieu de (12.18), nous aurions obtenu $\hat{\rho} = 0{,}417$ et $t = 2{,}63$; nos conclusions auraient été similaires.

Test de détection de l'autocorrélation d'ordre supérieur à 1

Il est facile de généraliser (12.18) pour tester la présence de corrélation sérielle d'ordre supérieur à 1. Par exemple, supposons que nous désirions tester

$$H_0 : \rho_1 = 0, \rho_2 = 0 \qquad \textbf{[12.19]}$$

dans un modèle AR(2),

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t.$$

Ce modèle alternatif d'autocorrélation nous permet de tester la présence d'une corrélation sérielle d'ordre 2. Comme d'habitude, nous estimons le modèle par les MCO pour obtenir des résidus, $\hat{u}_t$. Ensuite, nous effectuons la régression de

$$\hat{u}_t \text{ sur } x_{t1}, x_{t2}, \ldots, x_{tk}, \hat{u}_{t-1} \text{ et } \hat{u}_{t-2}, \text{ pour tout } t = 3, \ldots, n,$$

dans le but d'effectuer un test F de significativité jointe de $\hat{u}_{t-1}$ et $\hat{u}_{t-2}$. Si au moins un de ces deux retards est significatif, disons à 5 %, alors l'hypothèse nulle (12.19) de ce test est rejetée et nous pouvons conclure que les erreurs souffrent de corrélation sérielle.

De manière plus générale, nous pouvons tester la présence de corrélation sérielle dans un modèle autorégressif d'ordre q :

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_q u_{t-q} + e_t. \qquad [12.20]$$

L'hypothèse nulle est

$$H_0 : \rho_1 = 0, \rho_2 = 0, \ldots, \rho_q = 0. \qquad [12.21]$$

Détecter la présence de corrélation sérielle jusqu'à l'ordre q

i. Effectuer la régression des MCO de y_t sur $x_{t1}, \ldots, x_{tk}$ et obtenir les résidus, $\hat{u}_t$, pour tout $t = 1, 2, \ldots, n$.

ii. Effectuer la régression de

$$\hat{u}_t \text{ sur } x_{t1}, x_{t2}, \ldots, x_{tk}, \hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-q}, \text{ pour tout } t = (q+1), \ldots, n. \qquad [12.22]$$

iii. Réaliser un test F de significativité jointe de $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-q}$ dans (12.22). [La statistique F peut être calculée en utilisant y_t comme variable dépendante dans (12.22) ; la réponse sera identique.]

Si les x_{tj} sont strictement exogènes (impliquant que chaque variable x_{tj} n'est pas corrélée avec $u_{t-1}, u_{t-2}, \ldots, u_{t-q}$), alors leur inclusion dans (12.22) est inutile. Le fait d'inclure les x_{tj} dans la régression permet de rendre le test valide même lorsque l'hypothèse de stricte exogénéité est violée. Ce test repose néanmoins sur l'hypothèse d'homoscédasticité,

$$\text{Var}(u_t|\mathbf{x}_t, u_{t-1}, \ldots, u_{t-q}) = \sigma^2. \qquad [12.23]$$

Si cette hypothèse est également violée, nous pouvons recourir à la version du test robuste à la présence d'hétéroscédasticité, décrite au chapitre 8.

Au lieu de calculer la statistique du test F, il est également possible d'utiliser la statistique du multiplicateur de Lagrange (*ML*). (Nous avons montré au chapitre 5, sur base de données en coupe transversale, que la statistique *ML* peut être utilisée pour tester si des variables explicatives pouvaient être supprimées du modèle.) Pour tester (12.21), la statistique *ML* est

$$ML = (n - q) R_{\hat{u}}^2, \qquad [12.24]$$

où $R_{\hat{u}}^2$est le R carré de la régression (12.22). Sous l'hypothèse nulle, $LM \overset{a}{\sim} \chi_q^2$. Ce test est appelé le **test de Breusch-Godfrey** de corrélation sérielle. Comme dans le cas de la statistique F, la statistique *ML* requiert (12.23) mais il existe une version du test robuste à la présence d'hétéroscédasticité [Pour plus de détails, voir Wooldridge (1991b).]

Lorsque l'influence de la saisonnalité dans des données trimestrielles ou mensuelles n'a pas été neutralisée, il est souvent judicieux d'en tenir compte dans les tests de corrélation sérielle. Par exemple, si nous utilisons des données trimestrielles brutes, nous pouvons spécifier le modèle suivant

$$u_t = \rho_4 u_{t-4} + e_t. \qquad [12.25]$$

Il nous suffit maintenant de recourir aux tests de corrélation sérielle d'ordre 1. Lorsque les régresseurs sont strictement exogènes, nous pouvons effectuer un test t sur $\hat{u}_{t-4}$ dans la régression de

$$\hat{u}_t \text{ sur } \hat{u}_{t-4}, \text{ pour tout } t = 5, \ldots, n.$$

Même si les résidus de cette régression ne sont pas adjacents, il est également possible d'utiliser une version modifiée de la statistique de Durbin-Watson [voir Wallis (1972)]. Enfin, si les x_{tj} ne sont pas strictement exogènes, nous pouvons utiliser la régression (12.18) dans laquelle nous remplaçons $\hat{u}_{t-1}$ par $\hat{u}_{t-4}$.

EXEMPLE 12.3
Détecter la présence de corrélation sérielle jusqu'à l'ordre 3

Dans l'étude d'événement relative à l'industrie du chlorure de barium (voir l'exemple 10.5), nous avons utilisé des données mensuelles ; il peut donc s'avérer prudent de vérifier si les erreurs souffrent d'une corrélation sérielle qui dépasse l'ordre 1. En guise d'illustration, nous utilisons un modèle AR(3) pour caractériser les erreurs de l'équation (10.22). Après l'estimation de la régression (12.22), nous pouvons réaliser un test F de significativité jointe sur $\hat{u}_{t-1}$, $\hat{u}_{t-2}$, et $\hat{u}_{t-3}$. La statistique F du test est égale à 5,12. Au départ, nous avons $n = 131$ mais nous perdons trois observations dans la régression auxiliaire (12.22). Par ailleurs, comme la régression (10.22) contient 7 paramètres et que nous en ajoutons 3 dans (12.22), nous devons estimer 10 paramètres au total. Par conséquent, les degrés de liberté du test F sont 3 et 118, ce qui donne une p-valeur égale à 0,0023. On peut donc conclure que les erreurs sont affectées par la présence d'une corrélation sérielle qui peut aller jusqu'à l'ordre 3.

Dans l'exemple 12.3, les données sont mensuelles et l'influence de la saisonnalité n'a pas été prise en compte. Il est donc prudent de vérifier si u_t et u_{t-12} sont corrélées, par exemple. La régression de $\hat{u}_t$ sur $\hat{u}_{t-12}$ donne les résultats suivants : $\hat{\rho}_{12} = -0{,}187$ et p-valeur = 0,028. La saisonnalité présente dans les données conduit à une autocorrélation *négative* et significative à 5 %. Si nous incluons les variables explicatives dans le test, les résultats changent quelque peu : $\hat{\rho}_{12} = -0{,}170$ et p-valeur = 0,052. Cette autocorrélation négative est un résultat assez inhabituel mais la significativité statistique du test reste équivoque lorsque nous tenons compte d'une absence éventuelle d'exogénéité stricte de la part des variables explicatives.

Pour aller plus loin 12.3

Vous désirez vérifier s'il y a de la corrélation sérielle d'ordre 1 ou d'ordre 4 dans des données trimestrielles. En supposant que les régresseurs soient strictement exogènes, comment procédez-vous ?

12.3 LA CORRECTION DE L'AUTOCORRÉLATION EN PRÉSENCE DE RÉGRESSEURS STRICTEMENT EXOGÈNES

Si la présence de corrélation sérielle est détectée grâce à l'utilisation d'un des tests de la section 12.2, nous devons être capables d'en trouver le remède. Si notre objectif est d'estimer un modèle dont la dynamique est exhaustive, une nouvelle spécification du modèle doit être proposée. Pour les applications dont l'objectif n'est pas d'estimer un modèle dynamique plus complet, nous devons trouver un moyen de recourir à des tests valides d'inférence statistique : comme nous l'avons vu dans la section 12.1, les statistiques habituelles des MCO ne sont plus fiables. Dans cette section, nous partons du cas le plus classique, celui de la corrélation sérielle d'ordre 1. L'approche traditionnelle, que nous allons utiliser dans un premier temps, repose sur l'hypothèse que les régresseurs sont fixes ; autrement dit, les régresseurs doivent être strictement exogènes. Cette méthode de correction n'est donc pas appropriée si les variables explicatives incluent des variables dépendantes retardées.

Calcul de l'estimateur BLUE *en présence d'erreurs suivant un processus AR(1) connu*

Nous partons des hypothèses de Gauss-Markov, ST.1 à ST.4, mais nous assouplissons l'hypothèse ST.5 en supposant que les erreurs suivent le processus AR(1) suivant :

$$u_t = \rho u_{t-1} + e_t, \text{ pour tout } t = 1, 2, \ldots \qquad \textbf{[12.26]}$$

Rappelez-vous que l'hypothèse ST.3 implique que l'espérance conditionnelle de u_t est égale à zéro, soit $E(u|\mathbf{X}) = 0$. Dans le développement qui suit, nous simplifions la notation en supposant que le calcul est conditionnel à $\mathbf{X}$. Nous pouvons donc écrire que la variance de u_t est

$$\text{Var}(u_t) = \sigma_e^2/(1 - \rho^2). \qquad \textbf{[12.27]}$$

Pour plus de simplicité, prenons le cas du modèle de régression linéaire simple :

$$y_t = \beta_0 + \beta_1 x_t + u_t, \text{ pour tout } t = 1, 2, \ldots, n.$$

Sachant que les erreurs de cette équation souffrent de corrélation sérielle d'ordre 1, nous allons chercher à nous en débarrasser en transformant l'équation. Pour $t \geq 2$, nous pouvons écrire

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + u_{t-1},$$
$$y_t = \beta_0 + \beta_1 x_t + u_t.$$

Si nous multiplions la première équation par ρ et que nous la soustrayons de la seconde, nous obtenons

$$y_t - \rho y_{t-1} = (1 - \rho)\beta_0 + \beta_1(x_t - \rho x_{t-1}) + e_t, \; t \geq 2,$$

où $e_t = u_t - \rho u_{t-1}$. Nous obtenons

$$\tilde{y}_t = (1 - \rho)\beta_0 + \beta_1 \tilde{x}_t + e_t, \; t \geq 2, \qquad \textbf{[12.29]}$$

en posant

$$\tilde{y}_t = y_t - \rho y_{t-1} \text{ et } \tilde{x}_t = x_t - \rho x_{t-1}. \qquad \textbf{[12.28]}$$

$\tilde{y}_t$ et $\tilde{x}_t$ représentent des **variables en (quasi-)différence**. (Si $\rho = 1$, il s'agit de données en différence première ; souvenez-vous néanmoins que la condition de stabilité exige que $|\rho| < 1$). Notez également que le terme d'erreur dans (12.28) est exempt de la corrélation sérielle qui affectait u_t ; en réalité, cette équation respecte toutes les hypothèses de Gauss-Markov. Cela signifie que, si nous connaissons la vraie valeur de ρ, nous pouvons estimer β_0 et β_1 en régressant $\tilde{y}_t$ sur $\tilde{x}_t$, en n'oubliant pas de diviser la constante par $(1 - \rho)$.

Notez bien que les estimateurs des MCO de (12.28) ne sont pas *BLUE* car ils ignorent la première période de temps, $t = 1$. Nous pouvons y remédier en écrivant l'équation pour $t = 1$ comme suit :

$$y_1 = \beta_0 + \beta_1 x_1 + u_1. \qquad \textbf{[12.30]}$$

Étant donné que chaque e_t n'est pas corrélée avec u_1, nous pouvons ajouter (12.30) à (12.28) sans remettre en cause l'absence d'autocorrélation. Cependant, sur base de (12.27), nous pouvons constater que $\text{Var}(u_1) = \sigma_e^2/(1 - \rho^2) > \sigma_e^2 = \text{Var}(e_t)$. [Notez que l'équation (12.27) est violée lorsque $|\rho| \geq 1$, ce qui explique l'hypothèse de stabilité à laquelle nous recourrons.] Pour obtenir des erreurs de même variance, nous devons multiplier (12.30) par $(1 - \rho^2)^{1/2}$, soit

$$(1 - \rho^2)^{1/2} y_1 = (1 - \rho^2)^{1/2}\beta_0 + \beta_1(1 - \rho^2)^{1/2} x_1 + (1 - \rho^2)^{1/2} u_1$$

ou

$$\tilde{y}_1 = (1 - \rho^2)^{1/2}\beta_0 + \beta_1\tilde{x}_1 + \tilde{u}_1, \qquad \textbf{[12.31]}$$

en posant $\tilde{u}_1 = (1 - \rho^2)^{1/2}u_1$, $\tilde{y}_t = (1 - \rho^2)^{1/2}y_1$ et $\tilde{x}_t = (1 - \rho^2)^{1/2}x_1$. La variance de l'erreur dans (12.31) est Var $(\tilde{u}_1) = (1 - \rho^2)\text{Var}(u_1) = \sigma_e^2$. Nous pouvons désormais ajouter (12.31) à (12.28) et obtenir des estimateurs *BLUE* de β_0 et β_1 sous les hypothèses ST.1 à ST.4, malgré un processus AR(1) pour u_t. En fait, il s'agit d'estimateurs des *moindres carrés généralisés* (MCG). Nous avons rencontré d'autres estimateurs des MCG au chapitre 8, dans le contexte de l'hétéroscédasticité.

L'ajout d'autres variables explicatives (strictement exogènes) ne change pas grand-chose. Pour $t \geq 2$, nous avons

$$\tilde{y}_t = (1 - \rho)\beta_0 + \beta_1\tilde{x}_{t1} + \ldots + \beta_k\tilde{x}_{tk} + e_t, \qquad \textbf{[12.32]}$$

en posant $\tilde{x}_{tj} = x_{tj} - \rho x_{t-1,j}$. Lorsque $t = 1$, nous avons $\tilde{y}_1 = (1 - \rho^2)^{1/2}y_1$, $\tilde{x}_{1j} = (1 - \rho^2)^{1/2}x_{1j}$, et la constante est $(1 - \rho^2)^{1/2}\beta_0$. Si nous connaissons la valeur de ρ, il est relativement facile de transformer les données et d'estimer l'équation par les MCO. Sauf si $\rho = 0$, l'estimateur des MCG, c'est-à-dire l'estimateur des MCO appliqué aux données transformées, sera différent de l'estimateur des MCO appliqué aux données brutes. L'estimateur des MCG est donc *BLUE* et les tests t et F effectués sur (12.32) sont valides puisque les erreurs e_t sont exemptes de corrélation sérielle et d'hétéroscédasticité. Dans le pire des cas, cette validité sera asymptotique si les erreurs e_t ne sont pas distribuées normalement.

Estimation par les MCQG en présence d'erreurs suivant un processus AR(1) inconnu

Le gros inconvénient de l'estimateur des MCG est qu'en pratique, il est très rare de connaître la valeur de ρ. Heureusement, nous pouvons obtenir un estimateur convergent de ρ en régressant les résidus des MCO sur leurs valeurs retardées adjacentes, comme dans l'équation (12.14). Ensuite, nous pouvons utiliser la valeur estimée $\hat{\rho}$, au lieu de ρ, pour calculer les variables en différence. Enfin, nous pouvons utiliser les MCO pour estimer l'équation

$$\tilde{y}_t = \beta_0\tilde{x}_{t0} + \beta_1\tilde{x}_{t1} + \ldots + \beta_k\tilde{x}_{tk} + erreur_t, \qquad \textbf{[12.33]}$$

où $\tilde{x}_{t0} = (1 - \hat{\rho})$ pour $t \geq 2$, et $\tilde{x}_{10} = (1 - \hat{\rho}^2)^{1/2}$. Cette procédure permet d'obtenir un estimateur des **moindres carrés quasi-généralisés** (**MCQG**) pour chaque β_j. Le terme d'erreur de (12.33) contient e_t mais est également pollué par l'erreur d'estimation liée à $\hat{\rho}$. La bonne nouvelle est que cette erreur d'estimation n'affecte pas la distribution asymptotique de l'estimateur des MCQG.

Estimation du modèle AR(1) par les MCQG

i. Effectuer une régression des MCO de y_t sur $x_{t1}, \ldots, x_{tk}$ pour obtenir les résidus, $\hat{u}_t$, pour tout $t = 1, 2, \ldots, n$.

ii. Estimer l'équation (12.14) par les MCO pour obtenir $\hat{\rho}$.

iii. Construire l'équation (12.33) en utilisant $\hat{\rho}$ et estimer $\beta_0, \beta_1, \ldots, \beta_k$ par les MCO. Les écarts-types estimés ainsi que les statistiques t et F sont asymptotiquement valides.

En utilisant $\hat{\rho}$ plutôt que ρ, nous devons utiliser l'estimateur des MCQG dont les propriétés en échantillon fini ne sont pas aussi désirables qu'auparavant. Etant biaisé en échantillon fini, l'estimateur des MCQG n'est pas *BLUE*, même s'il reste convergent lorsque les données sont stationnaires et faiblement dépendantes. Par ailleurs, même si l'erreur e_t dans (12.32) est normalement distribuée, les statistiques t et F ne sont qu'approximativement distribuées selon les lois de Student et de Fisher en raison de l'erreur d'estimation liée à $\hat{\rho}$ et présente dans e_t. Cela ne pose pas de problème pour les applications empiriques qui utilisent des échantillons de grande taille.

Biaisé, l'estimateur des MCQG reste néanmoins asymptotiquement plus efficace que son équivalent des MCO lorsque les erreurs suivent un processus AR(1) (et que les variables sont strictement exogènes). Ici aussi, nous supposons que les données sont stationnaires et faiblement dépendantes.

En réalité, l'estimation du processus AR(1) par les MCQG n'est pas toujours effectuée de la même manière : elle dépend de la manière dont ρ est estimé et du traitement réservé à la première observation, $t = 1$. Par exemple, **l'estimation de Cochrane-Orcutt (CO)** ne tient pas compte de la première observation et utilise le $\hat{\rho}$ de (12.14). Quant à **l'estimation de Prais-Winsten (PW)**, elle incorpore la première observation et suit la procédure que nous avons décrite précédemment. Sur le plan asymptotique, à partir du moment où les séries temporelles ne sont pas trop courtes, cette différence n'a aucune importance.

Dans la pratique, les méthodes de Cochrane-Orcutt et de Prais-Winsten suivent une approche itérative. Autrement dit, la procédure en trois étapes décrite ci-dessus est répétée jusqu'à ce que l'estimation de ρ ne varie plus que marginalement. À la fin de chaque itération, nous partons d'une nouvelle série de résidus, ce qui permet d'obtenir une nouvelle estimation de ρ à partir de (12.14), de transformer les données à l'aide du nouveau $\hat{\rho}$, et d'estimer (12.33) par les MCO (et ainsi de suite jusqu'à ce que la variation de $\hat{\rho}$ devienne négligeable). Dans la plupart des logiciels économétriques, ce processus itératif est lancé automatiquement et aucune manipulation particulière de la part de l'utilisateur n'est requise. Il est difficile de savoir si la multiplication des itérations aide réellement. Dans certains cas, elle semble utile mais, d'un point de vue théorique, les propriétés asymptotiques de l'estimateur itéré sont identiques à celles de l'estimateur qui ne l'est pas. Pour plus de détails sur ce sujet et sur les autres méthodes d'estimation, voir Davidson et MacKinnon (1993, chapitre 10).

EXEMPLE 12.4
Étude d'événement dans l'industrie chimique et estimation par la méthode de Prais-Winsten

Estimons à nouveau l'équation portant sur la base de données BARIUM, qui avait été analysée dans l'exemple 10.5. Dans le tableau 12.1, nous comparons les résultats obtenus par la méthode des MCO à ceux qui découlent de la méthode des MCQG suivie par Prais et Winsten.

Le choix de la méthode d'estimation importe peu si nous considérons les coefficients significatifs sur le plan statistique. On constate, en effet, que les estimations obtenues par les MCQG sont très proches de celles des MCO [en particulier, pour log(*chempi*), log(*rtwex*), et *afdec6*]. Par contre, les coefficients qui ne sont pas statistiquement significatifs peuvent varier de manière importante, ce qui n'est pas non plus surprenant.

Remarquez que les écarts-types estimés (entre parenthèses) sont systématiquement plus élevés dans la seconde colonne que dans la première. C'est assez classique puisque les écarts-types estimés de Prais-Winsten sont précisément ajustés pour tenir compte de la présence de corrélation sérielle (positive, dans ce cas précis). Comme nous l'avons vu dans la section 12.1, les écarts-types estimés des MCO ont souvent tendance à sous-estimer la variation d'échantillonnage réelle des estimations ; ils ne sont donc pas fiables lorsqu'il existe une corrélation importante dans les erreurs. En conclusion, l'impact de la décision de la commission du commerce international des États-Unis sur les importations venant de Chine est moins significatif sur le plan statistique que nous le pensions auparavant. En effet, le t_{afdec6} des MCQG est égal à $-1,69$ alors que celui des MCO est égal à $-1,98$.

Enfin, l'estimation de l'équation par la méthode de Prais-Winsten donne un R carré substantiellement plus bas que celui obtenu par les MCO. Gardez néanmoins bien à l'esprit que cette comparaison est fallacieuse. Dans le cas des MCO, le R carré est calculé sur base des variables brutes. Dans l'approche de Prais-Winsten, le R carré provient de la dernière régression effectuée à partir des variables *transformées* à l'aide de $\hat{\rho}$. Bien que le R^2 est traditionnellement calculé et repris dans les résultats de la régression des MCQG, il est difficile de savoir ce qu'il représente exactement dans ce cas.

Tableau 12.3 *Variable dépendante : log(chnimp)*

Coefficient	MCO	Prais-Winsten
log(*chempi*)	3,12 (0,48)	2,94 (0,63)
log(*gas*)	0,196 (0,907)	1,05 (0,98)
log(*rtwex*)	0,983 (0,400)	1,13 (0,51)
befile6	0,060 (0,261)	−0,016 (0,322)
affile6	−0,032 (0,264)	−0,033 (0,322)
afdec6	−0,565 (0,286)	−0,577 (0,342)
Constante	−17,80 (21,05)	−37,08 (22,78)
$\hat{\rho}$	—	0,293
Observations *R* carré	131 0,305	131 0,202

Comparaison des MCO et des MCQG

Dans certaines applications empiriques, les estimations obtenues à l'aide des méthodes de Cochrane-Orcutt et de Prais-Winsten peuvent être sensiblement différentes de celles obtenues par les MCO (même si ce n'était pas le cas dans l'exemple 12.4). Cette différence est souvent considérée comme une indication de la supériorité des MCQG sur les MCO. Les choses ne sont pourtant pas aussi simples. Considérons la régression

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

dans laquelle les séries temporelles sont stationnaires. En considérant que la loi des grands nombres tient, la convergence des MCO pour β_1 sera vérifiée à condition que

$$\text{Cov}(x_t, u_t) = 0. \qquad \textbf{[12.34]}$$

Or, nous avons affirmé que les MCQG étaient convergents à condition que l'hypothèse de stricte exogénéité ne soit pas violée, ce qui est plus contraignant que (12.34). En réalité, nous pourrions montrer

que l'hypothèse la moins restrictive sur laquelle repose la convergence des MCQG, *en plus de l'hypothèse (12.34)*, est que la somme de x_{t-1} et x_{t+1} ne soit pas corrélée avec u_t :

$$\text{Cov}[(x_{t-1} + x_{t+1}), u_t] = 0. \qquad \textbf{[12.35]}$$

Concrètement, la convergence des MCQG repose sur l'absence de corrélation entre u_t, d'une part, et x_t, x_{t-1}, et x_{t+1}, d'autre part.

Comment peut-on expliquer que la convergence des MCQG dépende de la condition (12.35) ? Le raisonnement est simple si nous supposons que ρ est connu et que nous laissons tomber la première observation temporelle, comme dans la méthode de Cochrane-Orcutt. L'explication devient plus compliquée si nous utilisons $\hat{\rho}$ à la place ; elle n'apporte pas non plus d'éclairage supplémentaire. Reconnaissons tout d'abord que les propriétés asymptotiques d'un estimateur ne dépendent pas de l'ajout ou de la suppression d'une seule observation. Ensuite, notons que l'estimateur des MCQG utilise $x_t - \rho x_{t-1}$ comme régresseur dans l'équation où $u_t - \rho u_{t-1}$ représente le terme d'erreur. Grâce au théorème 11.1, nous savons que la convergence des MCO n'est vérifiée qu'en l'absence de corrélation entre le terme d'erreur et le régresseur. Dans ce cas précis, cela implique que $\text{E}[(x_t - \rho x_{t-1})(u_t - \rho u_{t-1})] = 0$. Si nous développons le calcul de l'espérance, nous obtenons

$$\begin{aligned}\text{E}[(x_t - \rho x_{t-1})(u_t - \rho u_{t-1})] &= \text{E}(x_t u_t) - \rho\text{E}(x_{t-1}u_t) - \rho\text{E}(x_t u_{t-1}) + \rho^2\text{E}(x_{t-1}u_{t-1}) \\ &= -\rho[\text{E}(x_{t-1}u_t) + \text{E}(x_t u_{t-1})]\end{aligned}$$

car $\text{E}(x_t u_t) = \text{E}(x_{t-1}u_{t-1}) = 0$ sous l'hypothèse (12.34). En outre, l'hypothèse de stationnarité implique que $\text{E}(x_t u_{t-1}) = \text{E}(x_{t+1}u_t)$ puisque nous ne faisons qu'avancer l'indice de temps d'une période. Par conséquent,

$$\text{E}(x_{t-1}u_t) + \text{E}(x_t u_{t-1}) = \text{E}[(x_{t-1} + x_{t+1})u_t].$$

Comme $\text{E}(u_t) = 0$, cette expression de l'espérance correspond précisément à la covariance de l'équation (12.35). Nous avons donc réussi à démontrer que la convergence de l'estimateur des MCG pour β_1 requiert les *deux* conditions (12.34) et (12.35) [Naturellement, si $\rho = 0$, (12.35) est inutile puisque nous retombons sur les MCO.]

Il en ressort que les MCQG et les MCO ne donneront pas nécessairement les mêmes estimations si la condition (12.35) n'est pas respectée. Dans un tel cas de figure, les MCO gardent leur convergence, qui ne dépend que de (12.34), alors que les MCQG la perdent. Si x a un effet retardé sur y, ou que x_{t+1} réagit aux variations de u_t, alors l'application des MCQG peut aboutir à des estimations erronées.

Comme les MCO et les MCQG sont des méthodes d'estimation différentes, il ne faut jamais s'attendre à obtenir des estimations identiques. Si les erreurs souffrent de corrélation sérielle et que les deux techniques donnent des estimations équivalentes des β_j, alors il est préférable d'utiliser les estimateurs des MCQG qui jouissent d'une plus grande efficacité et sont, dans le pire des cas, asymptotiquement valides. Si la divergence des estimations entre les MCO et les MCQG est plus grande, il faut encore pouvoir en déterminer la significativité sur le plan statistique. On pourrait recourir à la méthode de Hausman (1978) mais il nous est impossible de l'aborder dans cet ouvrage d'introduction à l'économétrie.

L'exemple 12.5 montre à quel point les estimations des MCO et des MCQG peuvent parfois diverger.

EXEMPLE 12.5
La version statique de la courbe de Phillips

Dans le tableau 12.2, sont indiquées les estimations de la courbe de Phillips statistique obtenues à la fois par la méthode des MCO et par la méthode de Prais-Winsten. Nous avons déjà estimé cette courbe en recourant aux MCO dans l'exemple 10.1.

Tableau 12.4 *Variable dépendante : inf*

Coefficient	MCO	Prais-Winsten
Unem	0,468 (0,289)	−0,716 (0,313)
Constante	1,424 (1,719)	8,296 (2,231)
$\hat{\rho}$	——	0,781
Observations *R* carré	49 0,053	49 0,136

Le coefficient qui nous intéresse est celui de *unem*. Or, la différence entre l'estimation de PW diffère très sensiblement de celle des MCO. Dans ce cas précis, nous aurions tendance à nous fier aux estimations de PW car elles collent aux indications théoriques quant à la nature du compromis entre inflation et chômage. En fait, les estimations de PW sont assez proches de celles que nous obtenons lorsque la courbe de Phillips est estimée en différence première (voir l'exercice sur ordinateur C4 du chapitre 11). Cette similitude des résultats n'est pas très surprenante puisque la méthode de PW implique une équation en différence basée sur $\hat{\rho}$, égal à 0,781 dans notre exemple. En conclusion, les liens entre *inf* et *unem* semblent ténus lorsque ces variables sont en niveau ; par contre, il semble bien exister une relation négative entre ces deux variables lorsqu'elles sont en différence.

Des cas similaires à celui de la courbe statique de Phillips posent d'épineux problèmes sur le plan empirique. Si nous sommes véritablement intéressés par la relation statique qui pourrait exister entre ces deux variables, alors les MCO sont préférables car les estimateurs seront convergents ; la seule condition est que le chômage et l'inflation soient des processus I(0) faiblement dépendants. Par contre, les MCO n'auront pas cette propriété désirable si une ou plusieurs de ces variables affichent une racine unitaire. Nous examinerons ce problème plus en détails dans le chapitre 18. Dans l'exemple 12.5, les MCQG aboutissent à des estimations plus proches de celles attendues par la théorie économique. Par ailleurs, la méthode des MCQG permet d'éliminer (approximativement) la présence de racines unitaires puisqu'elle consiste à estimer des équations similaires aux équations en différence première.

Correction par les MCQG d'une corrélation sérielle d'ordre supérieur à 1

Il est également possible de recourir aux MCQG pour se débarrasser d'une corrélation sérielle d'un ordre supérieur à 1. Harvey (1990) en présente une analyse exhaustive. Dans cet ouvrage, nous allons illustrer cette approche en recourant à un modèle autorégressif d'ordre 2 pour les erreurs. Le modèle AR(2) est

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t,$$

où $\{e_t\}$ satisfait les mêmes hypothèses que le modèle AR(1) précédent. Les conditions de stabilité du modèle AR(2) sont néanmoins plus compliquées à établir. Selon Harvey (1990), elles correspondent à

$$\rho_2 > -1,\ \rho_2 - \rho_1 < 1, \text{ et } \rho_1 + \rho_2 < 1.$$

Par exemple, le modèle est stable si $\rho_1 = 0{,}8$ et $\rho_2 = -0{,}3$; le modèle ne l'est pas lorsque $\rho_1 = 0{,}7$ et $\rho_2 = 0{,}4$.

En supposant que ces conditions soient respectées, nous pouvons transformer les variables du modèle dans le but de supprimer l'autocorrélation d'ordre 2 présente dans les erreurs. Dans un modèle de régression simple, si $t > 2$, nous avons

$$y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} = \beta_0(1 - \rho_1 - \rho_2) + \beta_1(x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}) + e_t$$

ou

$$\tilde{y}_t = \beta_0(1 - \rho_1 - \rho_2) + \beta_1 \tilde{x}_t + e_t, \; t = 3, 4, \ldots, n. \qquad \textbf{[12.36]}$$

Si ρ_1 et ρ_2 sont connus, nous appliquons la méthode des MCG qui consiste à estimer l'équation des variables transformées par les MCO. Comme ρ_1 et ρ_2 sont rarement connus, il faut souvent les estimer. Comme d'habitude, nous utilisons les résidus des MCO, $\hat{u}_t$, pour obtenir $\hat{\rho}_1$ et $\hat{\rho}_2$ à partir de la régression de

$$\hat{u}_t \text{ sur } \hat{u}_{t-1}, \hat{u}_{t-2}, \; t = 3, \ldots, n.$$

[Il s'agit de la même régression que celle utilisée pour tester la corrélation sérielle d'ordre 2 en présence de régresseurs strictement exogènes.] Nous devons ensuite utiliser $\hat{\rho}_1$ et $\hat{\rho}_2$, au lieu de ρ_1 et ρ_2, pour calculer les variables en différence. Cette procédure permet d'obtenir une version de l'estimateur des MCQG. Si nous avons plusieurs variables explicatives, chacune sera transformée de la manière suivante : $\tilde{x}_{tj} = \tilde{x}_{tj} - \hat{\rho}_1 x_{t-1,j} - \hat{\rho}_2 x_{t-2,j}$, lorsque $t > 2$.

Le traitement des deux premières observations est un peu particulier. Pour toutes les variables du modèle (y compris la constante), on doit appliquer les transformations suivantes :

$$\tilde{z}_1 = \{(1 + \rho_2)[(1 - \rho_2)^2 - \rho_1^2]/(1 - \rho_2)\}^{1/2} z_1 \text{ et}$$

$$\tilde{z}_2 = \{(1 - \rho_2^2)^{1/2} z_2 - [\rho_1(1 - \rho_1^2)^{1/2}/(1 - \rho_2)]z_1,$$

où z_1 et z_2 représentent n'importe quelle variable du modèle lorsque $t = 1$ et $t = 2$. Dans un souci de concision, nous n'allons pas dériver ces transformations. Sachez simplement qu'elles permettent d'éliminer l'autocorrélation entre les deux premières observations et d'obtenir une variance des erreurs égale à σ_e^2.

En règle générale, nous n'avons pas besoin de calculer nous-mêmes les variables transformées : les logiciels économétriques spécialisés dans le traitement des séries temporelles sont suffisamment élaborés pour estimer, sans le moindre problème, les modèles dont les erreurs suivent un processus AR(q).

12.4 CORRÉLATION SÉRIELLE ET VARIABLES EN DIFFÉRENCE PREMIÈRE

Dans le chapitre 11, nous avons expliqué que l'utilisation de variables en différence première permettait de transformer un processus intégré en un processus faiblement dépendant. Dans cette section, nous allons montrer qu'elle permet également de se débarrasser de la corrélation sérielle lorsque la persistance dans les données est élevée. Considérons un modèle de régression linéaire simple :

$$y_t = \beta_0 + \beta_1 x_t + u_t, \; t = 1, 2, \ldots, \qquad \textbf{[12.37]}$$

dans lequel u_t suit le processus AR(1) décrit en (12.26). Comme nous l'avons mentionné dans la section 11.3, l'utilisation des procédures d'inférence statistique basées sur les MCO est particulièrement inadéquate lorsque les variables y_t et x_t sont I(1), c'est-à-dire intégrées d'ordre 1. Nous en discuterons plus en détails dans le chapitre 18. Dans le cas où les erreurs $\{u_t\}$ de (12.37) suivent une marche aléatoire, l'équation perd toute signification car, notamment, la variance de u_t s'accroît avec le temps. Il est alors plus logique d'utiliser les variables en différence première :

$$\Delta y_t = \beta_1 \Delta x_t + \Delta u_t, \; t = 2, \ldots, n. \qquad \textbf{[12.38]}$$

Si u_t suit une marche aléatoire, alors $e_t \equiv \Delta u_t$. L'erreur e_t n'est donc pas autocorrélée ; elle affiche une espérance nulle et une variance constante. À condition que e_t ne soit pas non plus corrélée avec Δx_t, l'estimation de (12.38) par les MCO est valide, en notant que nous perdons la première observation naturellement.

Dans le cas où u_t ne suit pas une marche aléatoire mais que la valeur de ρ reste positive et élevée, il est souvent judicieux d'utiliser, malgré tout, des variables en différence première dans l'optique de se débarrasser de la corrélation sérielle présente dans les erreurs. L'inconvénient est que l'équation (12.38) n'est manifestement plus identique à l'équation de départ, (12.37). Nous n'avons néanmoins pas vraiment le choix car seuls les écarts-types estimés des MCO de (12.38) sont fiables. Inclure plusieurs variables explicatives ne change rien au raisonnement.

EXEMPLE 12.6
Équation du taux d'intérêt en différence première

Dans l'exemple 10.2, nous avons estimé une équation cherchant à expliquer le taux d'intérêt à trois mois du Trésor américain en fonction de l'inflation et du déficit budgétaire de l'état fédéral [voir l'équation (10.15)]. Si nous régressons les résidus de (10.15) sur un seul retard, nous obtenons $\hat{\rho} = 0{,}623$; la valeur est élevée et significative sur le plan statistique. On peut donc raisonnablement conclure que la corrélation sérielle pose problème dans cette équation.

Si nous utilisons les variables en différence première, les estimations de la régression donnent

$$\Delta i3_t = 0{,}042 + 0{,}149\ \Delta inf_t - 0{,}181\ \Delta def_t + e_t$$
$$(0{,}171)\ \ (0{,}092)\qquad (0{,}148) \qquad \textbf{[12.39]}$$
$$n = 55,\ R^2 = 0{,}176,\ \overline{R}^2 = 0{,}145 \qquad \textbf{[12.39]}$$

Les coefficients de cette régression en différence première sont très différents de l'équation de départ (10.15). Il est donc probable que les variables explicatives en niveau ne sont pas strictement exogènes ou qu'elles ont une racine unitaire. La corrélation entre $i3_t$ et $i3_{t-1}$ est d'ailleurs très élevée : elle est proche de 0,89. L'interprétation de (10.15) est donc sujette à caution. Par contre, la régression en différence première est, pour ainsi dire, exempte de toute corrélation sérielle puisque la régression de $\hat{e}_t$ sur $\hat{e}_{t-1}$ donne $\hat{\rho} = 0{,}072$ avec $\hat{\sigma}_{\hat{\rho}} = 0{,}134$. Étant donné que l'utilisation de variables en différence première permet d'éliminer à la fois les racines unitaires et la corrélation sérielle, la fiabilité des écarts-types estimés de (12.39) est vraisemblablement plus grande que celle des écarts-types estimés de (10.15). L'équation en différence première montre que les variations annuelles du taux à court-terme ne sont que marginalement influencées par les variations annuelles de l'inflation ou du déficit budgétaire. Le coefficient estimé de Δdef_t est même négatif ; il n'est pas pour autant différent de zéro sur le plan statistique puisque la p-valeur du test bilatéral n'est même pas inférieure à 20 %.

Comme nous l'avons expliqué au chapitre 11, la décision d'utiliser des variables en différence première n'est pas facile à prendre. Le raisonnement que nous venons de tenir a montré néanmoins que la différence première peut permettre de régler le problème de la corrélation sérielle dans le terme d'erreur. Nous y reviendrons au chapitre 18.

Pour aller plus loin 12.4

Vous avez estimé un modèle quelconque par les MCO. En utilisant la régression (12.14), vous obtenez $\hat{\rho} = 0{,}92$. Que faites-vous ?

12.5 CORRECTION DES ÉCARTS-TYPES ESTIMÉS APRÈS ESTIMATION PAR LES MCO

Depuis plusieurs années, la technique la plus populaire pour remédier aux problèmes d'autocorrélation dans les erreurs est de corriger les écarts-types estimés après l'estimation du modèle par les MCO. Bien que l'estimateur des MCO ne soit pas efficace, il existe de bonnes raisons pour procéder de la sorte. En premier lieu, il faut reconnaître que l'hypothèse de stricte exogénéité des variables explicatives est fréquemment violée. Or, dans un tel cas de figure, les MCQG ne sont ni efficaces, ni même convergents. En second lieu, les MCQG sont principalement utilisés lorsque les erreurs sont susceptibles de suivre un processus AR(1). Or, il est possible de rendre les écarts-types estimés des MCO robustes à des formes d'autocorrélation (et d'hétéroscédasticité) beaucoup plus générales que le processus AR(1).

Pour y voir plus clair, considérons l'équation (12.4) qui correspond à la variance de l'estimateur des MCO pour la pente d'un modèle de régression linéaire simple dont les erreurs suivent un processus AR(1). Il nous est possible d'estimer cette variance en utilisant les estimateurs classiques de ρ et σ^2. Pour que cette procédure soit valide, nous devons néanmoins supposer que les erreurs sont effectivement homoscédastiques et qu'elles suivent bien un modèle AR(1).

Nous allons maintenant montrer qu'il est possible de relâcher ses deux hypothèses. Davidson et MacKinnon (1993) ont montré qu'il était en effet possible de corriger les écarts-types estimés des MCO en présence de formes très générales d'hétéroscédasticité et de corrélation sérielle. Dans cet ouvrage, ces écarts-types estimés robustes des MCO sont calculés à l'aide d'une méthode simple qui est décrite dans Wooldridge (1989). Envisageons un modèle de régression linéaire multiple

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t,\ t = 1, 2, \ldots, n, \qquad \textbf{[12.40]}$$

que nous estimons par les MCO. Pour aller à l'essentiel, nous allons nous focaliser sur le calcul de l'écart-type estimé robuste de β_1. La méthode de calcul est assez simple. Commençons par considérer que x_{t1} est une fonction linéaire des autres variables explicatives et d'un terme d'erreur, soit

$$x_{t1} = \delta_0 + \delta_2 x_{t2} + \ldots + \delta_k x_{tk} + r_t,$$

dans laquelle l'espérance de l'erreur r_t et sa corrélation avec $x_{t2}, x_{t3}, \ldots, x_{tk}$ sont toutes deux égales à zéro.

Ensuite, il est possible de démontrer que la variance asymptotique de l'estimateur $\hat{\beta}_1$ des MCO est

$$\text{VarA}\,(\hat{\beta}_1) = (\sum_{t=1}^{n} \text{E}(r_t^2))^{-2}\ \text{Var}(\sum_{t=1}^{n} r_t u_t)$$

Sous l'hypothèse ST.5′ d'absence de corrélation sérielle, le terme $\{r_t u_t \equiv a_t\}$ n'est pas autocorrélé si bien que les écarts-types estimés des MCO seront valides sous l'hypothèse d'homoscédasticité ; si cette dernière hypothèse n'est pas vérifiée, la validité ne portera que sur les écarts-types estimés robustes à la présence l'hétéroscédasticité. Dans le cas où ST.5′ est violée, notre formule de $\text{VarA}(\hat{\beta}_1)$ doit être ajustée pour tenir compte de la corrélation entre a_t et a_s, lorsque $t \neq s$. Dans la pratique, il est fréquent de considérer que la corrélation est nulle lorsque les termes a_t et a_s sont éloignés de plusieurs périodes. Ce raccourci est justifié par le fait que la faible dépendance d'un processus implique que la corrélation doit précisément tendre vers zéro.

En partant du cadre général défini par Newey et West (1987), Wooldridge (1989) montre qu'il est possible d'estimer VarA $(\hat{\beta}_1)$ en procédant de la manière suivante. Définissons, tout d'abord, plusieurs termes. Soit "$\hat{\sigma}(\hat{\beta}_1)$", l'écart-type estimé des MCO pour $\hat{\beta}_1$; cet écart-type estimé est biaisé lorsque ST.5′

est violée. Soit $\hat{\sigma}$, l'écart-type de la régression (12.40) estimée par les MCO. Soit $\hat{r}_t$, les résidus de la régression auxiliaire de

$$x_{t1} \text{ sur } x_{t2}, x_{t3}, \ldots, x_{tk}, \qquad \textbf{[12.41]}$$

en incluant une constante, comme d'habitude. En considérant un nombre entier $g > 0$, définissons

$$\hat{v} = \sum_{t=1}^{n} \hat{a}_t^2 + 2\sum_{h=1}^{g}[1 - h/(g+1)]\left(\sum_{t=h+1}^{n} \hat{a}_t \hat{a}_{t-h}\right), \qquad \textbf{[12.42]}$$

où

$$\hat{a}_t = \hat{r}_t \hat{u}_t, \; t = 1, 2, \ldots, n.$$

L'estimation semble plus compliquée qu'elle ne l'est en réalité. Notez que le nombre entier g permet d'introduire la dose requise d'autocorrélation dans les écarts-types estimés. Après avoir obtenu $\hat{v}$, on peut directement estimer **l'écart-type estimé robuste à la présence de corrélation sérielle** pour $\hat{\beta}_1$, soit

$$\hat{\sigma}(\hat{\beta}_1) = [\text{"}\hat{\sigma}(\hat{\beta}_1)\text{"}/\hat{\sigma}]^2\sqrt{\hat{v}}. \qquad \textbf{[12.43]}$$

Autrement dit, nous estimons d'abord l'écart-type estimé de $\hat{\beta}_1$ par les MCO (comme si l'hypothèse ST.5′ était respectée), ce qui donne "$\hat{\sigma}(\hat{\beta}_1)$". Nous la divisions ensuite par $\hat{\sigma}$, calculons le carré du ratio, et multiplions le tout par la racine carrée de $\hat{v}$. Grâce à (12.43), nous pouvons obtenir des intervalles de confiance et des tests statistiques fiables.

Il est utile d'appréhender le calcul de $\hat{v}$ dans des cas plus simples. Lorsque $g = 1$,

$$\hat{v} = \sum_{t=1}^{n} \hat{a}_t^2 + \sum_{t=h+1}^{n} \hat{a}_t \hat{a}_{t-h}, \qquad \textbf{[12.44]}$$

et lorsque $g = 2$,

$$\hat{v} = \sum_{t=1}^{n} \hat{a}_t^2 + (4/3)\left(\sum_{t=2}^{n} \hat{a}_t \hat{a}_{t-1}\right) + (2/3)\left(\sum_{t=3}^{n} \hat{a}_t \hat{a}_{t-2}\right). \qquad \textbf{[12.45]}$$

Le nombre entier g correspond au nombre de termes que l'on doit ajouter pour tenir compte du degré d'autocorrélation dans les erreurs. Le facteur $[1 - h/(g+1)]$ dans (12.42) permet de garantir la non-négativité de $\hat{v}$, soit $\hat{v} \geq 0$ [Newey et West (1987) le démontrent]. Ce facteur est requis puisque $\hat{v}$ est l'estimation d'une variance dont la racine carrée intervient dans le calcul de (12.43).

L'écart-type estimé de (12.43) est également valide en présence d'hétéroscédasticité dans les erreurs. Dans la littérature scientifique, cet écart-type estimé et corrigé est identifié par le sigle « CHA » ; on parle d'écart-type estimé *« cohérent avec l'hétéroscédasticité et l'autocorrélation »*. Remarquez que si nous ignorons le second terme dans (12.42), alors (12.43) correspond à la version de l'écart-type estimé valide en présence d'hétéroscédasticité uniquement, que nous avons rencontrée au chapitre 8 (sans l'ajustement lié aux degrés de liberté).

La théorie qui sous-tend les écarts-types estimés et corrigés est technique et subtile. Rappelez-vous que nous sommes partis d'une situation dans laquelle la nature de la corrélation sérielle n'était pas connue. Dans un tel cas de figure, comment choisir une valeur appropriée pour g ? La théorie indique que (12.43) permet de tenir compte de formes très variées de corrélation sérielle à condition que g augmente en fonction de la taille de l'échantillon n. Plus grande sera la taille de l'échantillon, plus grande sera la flexibilité avec laquelle nous pouvons doser l'autocorrélation dans l'équation (12.42). La relation entre n et g a d'ailleurs

fait l'objet d'études relativement récentes que nous n'aborderons pas dans cet ouvrage. Pour des données annuelles, une valeur comme $g = 1$ ou $g = 2$ suffit généralement à incorporer la dynamique d'autocorrélation présente dans les erreurs. Pour des données à plus haute fréquence, g est ajusté à la hausse, en supposant naturellement que nous ayons suffisamment de données ($g = 4$ ou 8 pour des données trimestrielles et $g = 12$ ou 24 pour des données mensuelles). Newey et West (1987) suggèrent de calculer g en fonction de $4(n/100)^{2/9}$, tout en arrondissant à la valeur entière inférieure ; d'autres chercheurs ont proposé de choisir la partie entière de $n^{1/4}$. La recommandation de Newey-West est suivie par le logiciel économétrique Eviews®. Par exemple, si $n = 70$ (ce qui correspond plus ou moins à un échantillon de données annuelles depuis la seconde guerre mondiale), $g = 3$. (La partie entière de $n^{1/4}$ donne $g = 2$.)

Résumons maintenant la procédure à suivre pour rendre l'écart-type estimé de $\hat{\beta}_1$ robuste à la présence d'autocorrélation. Bien sûr, cette méthode est valable pour calculer l'écart-type estimé du coefficient de n'importe quelle autre variable explicative.

Écart-type estimé de $\hat{\beta}_1$ robuste à la présence de corrélation sérielle

i. Estimer (12.40) par les MCO dans le but d'obtenir "$\hat{\sigma}(\hat{\beta}_1)$", $\hat{\sigma}$, et les résidus $\{\hat{u}_t : t = 1, \ldots, n\}$.

ii. Obtenir les résidus $\{\hat{r}_t : t = 1, \ldots, n\}$ de la régression auxiliaire (12.41). Ensuite, calculer $\hat{a}_t = \hat{r}_t\hat{u}_t$ (pour chaque t).

iii. Calculer $\hat{v}$ à partir (12.42) en utilisant une valeur pour g.

iv. Calculer $\hat{\sigma}(\hat{\beta}_1)$ en fonction de (12.43).

En règle générale, sur le plan empirique, l'écart-type estimé robuste à l'autocorrélation, $\hat{\sigma}(\hat{\beta}_1)$, est plus grand que l'écart-type estimé traditionnel des MCO, "$\hat{\sigma}(\hat{\beta}_1)$", car les erreurs affichent très souvent une corrélation sérielle positive. Notez également qu'il est possible que les deux écarts-types estimés soient similaires en dépit d'une forte corrélation sérielle dans $\{u_t\}$. C'est en effet l'autocorrélation d'échantillonnage au sein de $\hat{a}_t = \hat{r}_t\hat{u}_t$ qui détermine les écarts-types estimés robustes de $\hat{\beta}_1$.

En sciences humaines, l'utilisation des écarts-types estimés robustes à l'autocorrélation est, en général, moins fréquente que le recours aux écarts-types estimés robustes à l'hétéroscédasticité. Plusieurs raisons peuvent l'expliquer. Tout d'abord, le calcul des écarts-types estimés robustes à l'autocorrélation n'est pas automatique puisqu'il requiert la détermination de g en fonction de (12.42). Certes, des logiciels économétriques ont automatisé la procédure mais vous devez l'utiliser telle quelle dans ce cas. Ensuite, les longues séries temporelles sont plus rares que les grandes bases de données en coupe transversale. (La finance est sans doute une exception). Or, lorsque les séries temporelles sont courtes (jusqu'à moins de 100 observations) et que les erreurs sont fortement corrélées, le calcul des écarts-types estimés robustes à l'autocorrélation devient beaucoup moins précis. Autrement dit, l'estimateur des MCO peut être particulièrement inefficace en présence d'un petit échantillon et d'une forte autocorrélation. Il est alors fréquent que la correction des écarts-types estimés diminue, voire élimine, la significativité statistique des coefficients qui existait au départ.

Si nous pensons que les variables explicatives sont strictement exogènes et qu'un processus AR(1) peut éventuellement caractériser la dynamique présente dans les erreurs, il est préférable d'utiliser la méthode des MCQG de PW ou de CO, car les estimateurs seront plus efficaces que ceux obtenus pas les MCO. Lorsque la corrélation sérielle du processus AR(1) est substantielle, l'utilisation de variables en différence est, en effet, susceptible de donner de meilleurs résultats que l'utilisation des MCO sur les données brutes. Par contre, si les erreurs suivent un processus plus complexe, les écarts-types estimés de PW ou de CO ne seront pas fiables. Dans un tel cas de figure, la solution consiste à transformer les variables en différence sur base de l'estimation

de ρ puis à appliquer les MCO sur les variables en différence en recourant aux écarts-types estimés robustes à la présence de corrélation sérielle. L'utilisation des écarts-types estimés robustes après la transformation en différence permet de tenir compte de la corrélation sérielle résiduelle qui peut invalider les tests d'inférence statistique. Notez que les écarts-types estimés robustes sont plus efficaces lorsque l'essentiel de l'autocorrélation a déjà été éliminé par la transformation des variables en différence [que ce soit dans le cadre d'un processus AR(1) ou d'un processus d'ordre supérieur]. Dans la section 8.4, en présence d'hétéroscédasticité, nous avons suivi une approche similaire en utilisant les MCQG avant de recourir aux écarts-types estimés robustes pour tenir compte de l'hétéroscédasticité résiduelle liée à une fonction de la variance mal spécifiée.

Les écarts-types estimés robustes à la présence de corrélation sérielle sont donc utiles lorsque nous pensons que les variables explicatives ne sont pas strictement exogènes car les estimateurs obtenus par les méthodes de PW et de CO ne sont même pas convergents dans ce cas. On peut également utiliser les écarts-types estimés robustes dans des modèles dont la variable dépendante retardée joue le rôle de variable explicative, en considérant naturellement qu'il existe de bonnes raisons pour ne pas éliminer l'autocorrélation résiduelle d'une autre façon.

Kiefer et Vogelsang (2005) propose une autre manière d'obtenir des tests d'inférence statique valides en présence de corrélation sérielle de forme générale. Contrairement à Newey et West, Kiefer et Vogelsang ne cherchent pas à déterminer le taux auquel g doit croître (en fonction de n) dans le but d'obtenir une statistique t asymptotiquement distribuée selon une loi normale. Ils cherchent plutôt à dériver la distribution en grand échantillon de t lorsque $b = (g + 1)/n$ prend une valeur non nulle. [Dans le cadre d'analyse de Newey-West, $(g + 1)/n$ converge toujours vers zéro.] Par exemple, lorsque $b = 1$, $g = n - 1$; ce résultat implique que *chaque* terme de covariance doit être présent dans l'équation (12.42). La statistique t qui en résulte suit une distribution asymptotique qui n'est pas celle de la loi normale. Pour un test bilatéral à un seuil de 5 %, Kiefer et Vogelsang obtiennent une valeur critique égale à 4,77 ; pour un test bilatéral à 10 %, la valeur critique est 3,764. Ces valeurs critiques sont plus grandes que celles issues de la distribution normale. L'avantage est que nous n'avons plus à nous soucier du nombre de covariances à inclure dans (12.42).

EXEMPLE 12.7
Le salaire minimum à Porto Rico

Nous allons calculer l'écart-type estimé robuste à la présence d'autocorrélation afin de mieux évaluer la significativité statistique de l'instauration du salaire minimum sur le taux d'emploi à Porto Rico. Dans l'exemple 12.2, nous avons conclu à la présence de corrélation sérielle d'ordre 1 dans les erreurs. Comme précédemment, nous ajoutons une tendance temporelle linéaire et deux variables de contrôle, log(*usgnp*) et log(*prgnp*).

L'estimation par les MCO de l'élasticité du taux d'emploi par rapport au salaire minimum est donnée par $\hat{\beta}_1 = -0{,}2123$. L'écart-type estimé traditionnel des MCO est "$\hat{\sigma}(\hat{\beta}_1)$" = 0,0402 et l'écart-type de la régression est $\hat{\sigma} = 0{,}0328$. En utilisant (12.45) pour laquelle $g = 2$, nous obtenons $\hat{v} = 0{,}000805$. Nous pouvons maintenant calculer l'écart-type estimé *CHA* : $\hat{\sigma}(\hat{\beta}_1) = [(0{,}0402/0{,}0328)^2]\sqrt{0{,}000805} \approx 0{,}0426$. Notez que l'écart-type estimé robuste est légèrement plus grand que l'écart-type estimé habituel des MCO, "$\hat{\sigma}(\hat{\beta}_1)$". La statistique t robuste est environ égale à $-4{,}98$, confirmant la forte significativité statistique de l'élasticité.

En guise de comparaison, l'estimation de PW donne $\hat{\beta}_1 = -0{,}1477$ et son écart-type estimé est égal à 0,0458. La différence entre les estimations obtenues par les MCO et les MCQG peut s'expliquer par le fait que l'hypothèse de stricte exogénéité des régresseurs est violée ; si tel est le cas, les MCQG ne sont pas convergents. Cette différence peut également provenir d'une simple erreur d'échantillonnage et ne pas être significative sur le plan statistique ; si tel est le cas, l'estimateur des MCQG est préférable car il est plus efficace et, au pire, asymptotiquement valide. Malheureusement, il est difficile de déterminer laquelle des deux explications est la bonne.

Avant de passer à la section suivante, gardez à l'esprit qu'il est également possible de construire des statistiques F robustes à la présence de corrélation sérielle. Ce sujet est néanmoins trop technique pour être abordé dans un ouvrage d'introduction à l'économétrie comme celui-ci. [Voir Wooldridge (1991b, 1995) et Davidson et MacKinnon (1993).]

12.6 HÉTÉROSCÉDASTICITÉ DANS LES RÉGRESSIONS SUR SÉRIES TEMPORELLES

Dans le chapitre 8, nous avons étudié les tests et les méthodes de correction de l'hétéroscédasticité dans le cadre de régressions sur données en coupe transversale. L'hétéroscédasticité peut également polluer les erreurs d'une régression sur séries temporelles. Comme dans l'analyse en coupe transversale, elle fausse les tests d'hypothèse, sans pour autant biaiser l'estimateur $\hat{\beta}_j$ ou lui faire perdre sa convergence.

Dans les études empiriques sur séries temporelles, l'hétéroscédasticité est souvent considérée comme un problème secondaire et bien moins important que celui lié la corrélation sérielle. Une bonne compréhension de ses particularités n'en reste pas moins utile.

Rappelez-vous que les tests d'inférence statistique sont asymptotiquement valides sous les hypothèses ST.1′ à ST.5′. Dans cette section, nous cherchons à identifier les conséquences de la violation de l'hypothèse d'homoscédasticité, ST.4′. L'hypothèse ST.3′ exclut la possibilité d'une mauvaise spécification, comme l'oubli d'une variable importante ou la présence d'erreur de mesure. L'hypothèse ST.5′ exclut la présence d'autocorrélation dans les erreurs. Il est important de souligner que la prise en compte de l'hétéroscédasticité ne règle pas pour autant les problèmes d'autocorrélation qui pourraient également polluer les erreurs si l'hypothèse ST.5′ n'était pas respectée.

La construction de statistiques robustes à la présence d'hétéroscédasticité

Lorsque nous avons étudié l'hétéroscédasticité dans le cadre des régressions en coupe transversale, nous avons noté qu'elle n'avait aucun impact ni sur l'absence de biais ni sur la convergence des estimateurs des MCO. Ces conclusions sont toujours valables lorsque nous travaillons sur des séries temporelles. Ni le théorème 10.1 sur l'absence de biais, ni le théorème 11.1 sur la convergence ne dépende de l'hypothèse d'homoscédasticité, ST.4′.

Dans la section 8.2, nous avons montré que les écarts-types estimés des MCO ainsi que les statistiques t et F pouvaient tenir compte de la présence d'une hétéroscédasticité de forme inconnue. Les ajustements nécessaires sont identiques pour les régressions sur séries temporelles dans lesquelles seule l'hypothèse ST.4′ n'est pas respectée. La plupart des logiciels économétriques offrent d'ailleurs des procédures d'inférence statistique valides en présence d'hétéroscédasticité.

Tester la présence d'hétéroscédasticité dans les erreurs

Dans certaines circonstances, il est important de vérifier si les erreurs d'une régression sur séries temporelles sont polluées par l'hétéroscédasticité : par exemple, lorsqu'il s'agit d'effectuer des tests d'inférence statistique en présence d'un échantillon de petite taille. Les tests que nous avons étudiés au chapitre 8 sont directement applicables à condition que l'on tienne compte des remarques suivantes. Notons tout d'abord que les erreurs u_t *ne* doivent *pas* être autocorrélées, car un test d'hétéroscédasticité est, en règle générale, faussé par la présence de corrélation sérielle. La première étape doit donc consister à utiliser un test d'autocorrélation

robuste à la présence d'hétéroscédasticité si nécessaire. Si l'hypothèse nulle du test d'autocorrélation est rejetée, il faut alors y remédier. Le test d'hétéroscédasticité ne doit intervenir qu'ensuite.

Considérons maintenant le test d'hétéroscédasticité de Breusch-Pagan :

$$u_t^2 = \delta_0 + \delta_1 x_{t1} + \ldots + \delta_k x_{tk} + v_t, \qquad \textbf{[12.46]}$$

dans lequel l'hypothèse nulle est $H_0 : \delta_1 = \delta_2 = \ldots = \delta_k = 0$. Sachant que u_t^2 est remplacé par $\hat{u}_t^2$, la statistique F n'est valide qu'en l'absence d'hétéroscédasticité *et* de corrélation sérielle dans $\{v_t\}$. Tous les tests standards d'hétéroscédasticité reposent implicitement sur ces deux conditions ; c'est le cas du test de White, que nous avons vu dans la section 8.3. L'hypothèse d'absence de corrélation sérielle dans $\{v_t\}$ exclut d'ailleurs certaines formes d'hétéroscédasticité dynamique que nous allons aborder dans la section suivante.

Si l'hypothèse nulle du test d'hétéroscédasticité est rejetée (et qu'il n'y a pas de trace d'autocorrélation dans u_t), alors il convient d'utiliser les écarts-types estimés robustes à la présence d'hétéroscédasticité. Une alternative est de recourir aux **moindres carrés pondérés** (voir la section 8.4). Quelle que soit la nature des données, l'application des MCP reste la même.

EXEMPLE 12.8
Héteroscedasticité et hypothèse d'efficience des marchés financiers (HEM)

Dans l'exemple 11.4, nous avons estimé le modèle de régression linéaire simple

$$return_t = \beta_0 + \beta_1 return_{t-1} + u_t. \qquad \textbf{[12.47]}$$

Sous l'HEM, $\beta_1 = 0$. Si nous testons cette hypothèse en utilisant la base de données NYSE, nous obtenons $t_{\beta 1} = 1{,}55$ avec $n = 689$. Sur le plan statistique, l'HEM n'est donc pas rejetée. L'HEM implique que le rendement attendu est constant car l'information disponible a déjà été complètement incorporée dans le rendement actuel. Elle n'impose pas de contrainte sur la variance conditionnelle. En fait, si nous effectuons le test d'hétéroscédasticité de Breusch-Pagan en régressant $\hat{u}_t^2$ sur $return_{t-1}$, nous obtenons

$$\hat{u}_t^2 = 4{,}66 - 1{,}104\ return_{t-1} + résidus_t$$
$$(0{,}43)\ \ (0{,}201) \qquad \textbf{[12.48]}$$
$$n = 689,\ R^2 = 0{,}042.$$

La statistique t de la variable $return_{t-1}$ est égale à environ $-5{,}5$, ce qui nous amène à rejeter avec confiance l'hypothèse d'absence d'hétéroscédasticité. Comme le coefficient de $return_{t-1}$ est négatif, nous pouvons en conclure que la volatilité des rendements est plus faible lorsque le rendement précédent est élevé, et vice versa. Nous avons mis à jour deux caractéristiques que l'on retrouve fréquemment en finance empirique : contrairement à la variance de ces rendements, la valeur attendue des rendements ne dépend pas des rendements passés.

Pour aller plus loin 12.5

Vous devez effectuer le test d'hétéroscédasticité de White dans le cadre de la régression sur séries temporelles (12.47). Comment procédez-vous ?

Hétéroscédasticité conditionnelle autorégressive

Depuis quelque temps, les économistes se sont particulièrement attachés à étudier les différentes formes dynamiques que pouvaient prendre l'hétéroscédasticité. Dans (12.46), l'hétéroscédasticité est dynamique si $\mathbf{x}_t$ contient la variable dépendante sous une forme retardée. Notez que des formes dynamiques d'hétéroscédasticité peuvent également apparaître même lorsque la régression ne contient pas de variable dépendante retardée.

Pour s'en rendre compte, nous considérons une régression simple statique,

$$y_t = \beta_0 + \beta_1 z_t + u_t,$$

et supposons que les hypothèses de Gauss-Markov sont vérifiées. Les estimateurs de MCO sont donc les estimateurs linéaires sans biais les plus efficaces. L'hypothèse d'homoscédasticité implique que $\text{Var}(u_t|\mathbf{Z})$ est constante, avec $\mathbf{Z}$ représentant tous les n résultats possibles de z_t. Même si la variance de u_t étant donné $\mathbf{Z}$ est constante, l'hétéroscédasticité peut malgré tout polluer le terme d'erreur u_t. En effet, Engle (1982) montre que la variance conditionnelle de u_t, étant donné les erreurs passées, n'est pas nécessairement nulle. C'est la raison pour laquelle Engle propose le **modèle d'hétéroscédasticité conditionnelle autorégressive**, appelé plus couramment le modèle « **ARCH** ». (Ce sigle signifie « autoregressive conditional heteroskedasticity » en anglais). Le modèle ARCH d'ordre 1 est caractérisé par

$$E(u_t^2|u_{t-1}, u_{t-2}, \ldots) = E(u_t^2|u_{t-1}) = \alpha_0 + \alpha_1 u_{t-1}^2. \qquad \textbf{[12.49]}$$

Dans ce modèle, le calcul conditionnel par rapport à $\mathbf{Z}$ est implicite. Notez que cette équation représente la variance conditionnelle de u_t étant donné toutes ses valeurs passées. Elle n'est donc valide que si $E(u_t|u_{t-1}, u_{t-2}, \ldots) = 0$, ce qui signifie que les erreurs ne peuvent pas être autocorrélées. Vu que la variance conditionnelle doit être positive, ce modèle n'a de sens que si $\alpha_0 > 0$ et $\alpha_1 \geq 0$; si $\alpha_1 = 0$, l'équation de la variance n'affiche aucune dynamique particulière.

Il est également intéressant d'écrire (12.49) sous la forme suivante :

$$u_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + v_t, \qquad \textbf{[12.50]}$$

où la valeur attendue de v_t (étant donné $u_{t-1}, u_{t-2}, \ldots$) est nulle par définition. (Notez que les v_t ne sont pas indépendantes des valeurs passées de u_t en raison de la contrainte $v_t \geq -\alpha_0 - \alpha_1 u_{t-1}^2$). L'équation (12.50) ressemble bien à un modèle autorégressif de u_t^2 (d'où son appellation ARCH). La condition de stabilité est $\alpha_1 < 1$, comme dans un modèle AR(1) traditionnel. Lorsque $\alpha_1 > 0$, ce sont les erreurs *au carré* qui contiennent de la corrélation sérielle (positive).

Quelle est la validité des MCO lorsque les erreurs suivent un processus ARCH comme celui de (12.50) ? Comme nous avons supposé que les hypothèses de Gauss-Markov étaient vérifiées, les MCO sont *BLUE*. Même si les erreurs ne sont pas normalement distribuées, les tests d'hypothèses restent asymptotiquement valides sous les hypothèses ST.1′ à ST.5′. Or, ces hypothèses sont vérifiées dans les modèles statiques ou à retard échelonnés dont les erreurs suivent un processus ARCH.

Si les MCO conservent toutes ses propriétés désirables lorsque les erreurs suivent un processus ARCH, pourquoi devrions-nous y attacher de l'importance ? Deux raisons peuvent l'expliquer. Tout d'abord, en présence de cette forme d'hétéroscédasticité dynamique, il est possible d'obtenir des estimateurs convergents des β_j qui soient *asymptotiquement* plus efficaces que les estimateurs des MCO. Par exemple, une procédure d'estimation de (12.50) basée sur les moindres carrés pondérés peut y parvenir. Une procédure basée sur le maximum de vraisemblance fonctionne également sous l'hypothèse que les erreurs u_t suivent une distribution normale conditionnelle. Ensuite, les économistes portent un intérêt réel à l'égard de la dynamique présente au sein de la variance conditionnelle, que ce soit en finance ou en macroéconomie. L'exemple qu'Engle (1982) donne dans son étude concerne d'ailleurs la variance de l'inflation au

Royaume-Uni ; il montre qu'une plus grande variance de l'erreur à la période précédente (autrement dit, une valeur plus élevée pour u_{t-1}^2) s'accompagne d'une plus grande variance de l'erreur durant la période actuelle. Comme la variance sert à calculer la volatilité et que la volatilité est elle-même un élément central dans l'évaluation des actifs financiers, il n'est pas étonnant que la famille des modèles ARCH soit également devenue populaire en finance.

Les modèles ARCH interviennent aussi lorsque l'espérance conditionnelle de la variable dépendante est dynamique. Soit le modèle dynamique suivant dans lequel la variable dépendante est y_t et la variable exogène contemporaine est z_t :

$$\mathrm{E}(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots) = \beta_0 + \beta_1 z_t + \beta_2 y_{t-1} + \beta_3 z_{t-1}.$$

Ce modèle dynamique n'inclut qu'un seul retard pour y et z. L'hypothèse classique est de supposer que $\mathrm{Var}(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots)$ est constante, comme nous l'avons vu au chapitre 11. Néanmoins, cette variance peut suivre un processus ARCH :

$$\begin{aligned}\mathrm{Var}(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots) &= \mathrm{Var}(u_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots)\\ &= \alpha_0 + \alpha_1 u_{t-1}^2,\end{aligned}$$

où $u_t = y_t - \mathrm{E}(y_t|z_t, y_{t-1}, z_{t-1}, y_{t-2}, \ldots)$. Depuis le chapitre 11, nous savons que la présence d'un processus ARCH dans les erreurs u_t ne peut pas avoir d'incidence sur la convergence des MCO ou sur la validité des écarts-types estimés (et autres statistiques) robustes à la présence d'hétéroscédasticité (puisque le processus ARCH ne représente qu'une forme particulière d'hétéroscédasticité.)

Si les modèles ARCH et leurs extensions vous intéressent, consultez les revues de la littérature de Bollerslev, Chou, et Kroner (1992) et de Bollerslev, Engle, et Nelson (1994).

EXEMPLE 12.9
Processus ARCH et rendements boursiers

Dans l'exemple 12.8, nous avons vu qu'il y avait de l'hétéroscédasticité dans les rendements boursiers hebdomadaires. Le modèle ARCH décrit en (12.50) permet de mieux caractériser cette hétéroscédasticité. Si nous estimons les erreurs de (12.47) par les MCO et que nous utilisons le carré de ces résidus pour les régresser sur leur valeur retardée, nous obtenons :

$$\begin{gathered}\hat{u}_t^2 = 2{,}95 + 0{,}337\,\hat{u}_{t-1}^2 + résidus_t\\ (0{,}44)\ \ (0{,}036)\\ n = 688,\ R^2 = 0{,}114\end{gathered} \qquad \textbf{[12.51]}$$

La valeur de la statistique t de $\hat{u}_{t-1}^2$ est supérieure à 9 ; l'hypothèse nulle selon laquelle il n'y a pas d'effet ARCH est donc clairement rejetée. Comme nous en avons discuté auparavant, le signe positif indique qu'une erreur de plus grande ampleur au temps $t - 1$ se traduit par une variance des rendements boursiers plus élevée aujourd'hui, au temps t.

Alors que les carrés des résidus sont autocorrélés (ce qui revient à affirmer que les erreurs suivent un processus ARCH), il est important de noter que les résidus des MCO ne le sont pas, conformément à l'HEM. En effet, la régression de $\hat{u}_t$ sur $\hat{u}_{t-1}$ donne $\hat{\rho} = 0{,}0014$ avec $t_{\hat{\rho}} = 0{,}038$.

Hétéroscédasticité et corrélation sérielle dans les modèles de régression sur séries temporelles

Dans un modèle de régression linéaire, il est également possible que les erreurs contiennent à la fois de l'hétéroscédasticité et de la corrélation sérielle. Si tel est le cas, nous pouvons toujours recourir aux MCO en utilisant les écarts-types estimés *CHA*, qui sont robustes à la présence d'hétéroscédasticité *et* de corrélation sérielle (voir la section 12.5).

Dans la plupart des cas, la corrélation sérielle représente un problème plus sérieux que l'hétéroscédasticité. En effet, par rapport à l'hétéroscédasticité, la corrélation sérielle a généralement un impact plus conséquent sur les écarts-types estimés et sur l'efficacité. Dans la section 12.2, nous avons souligné qu'il était relativement facile de réaliser un test d'autocorrélation robuste à la présence d'une hétéroscédasticité de forme générale ; c'est le cas du test de Breusch-Godfrey, par exemple. Si nous détectons la présence de corrélation sérielle dans les erreurs, nous pouvons ensuite recourir à la transformation de Cochrane-Orcutt (ou de Prais-Winsten) [voir l'équation (12.32)] et utiliser, dans la régression en différence, les écarts-types estimés robustes à la présence d'hétéroscédasticité. Une alternative consiste à utiliser le test de Breusch-Pagan ou de White dans l'équation (12.32).

Une autre méthode consiste à modéliser l'hétéroscédasticité et la corrélation sérielle pour appliquer ensuite une procédure de correction qui combine les MCP (pour s'attaquer à l'hétéroscédasticité) et la méthode de CO ou de PW (pour s'attaquer à l'autocorrélation). Plus concrètement, considérons le modèle

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \qquad \textbf{[12.52]}$$
$$u_t = \sqrt{h_t}\, v_t,$$
$$v_t = \rho v_{t-1} + e_t, \qquad |\rho| < 1.$$

Les variables $\mathbf{X}$ son indépendantes de e_t, pour tout t, mais h_t est une fonction des x_{tj}. Le processus $\{e_t\}$ a une espérance nulle, une variance constante, égale à σ_e^2, et n'est pas autocorrélé. Il s'en suit que $\{v_t\}$ correspond à un processus AR(1) stable. Or, l'erreur u_t est hétéroscédastique et autocorrélée, soit :

$$\text{Var}(u_t|\mathbf{x}_t) = \sigma_v^2 h_t,$$

où $\sigma_v^2 = \sigma_e^2/(1-\rho^2)$. Dès lors, comme $v_t = u_t/\sqrt{h_t}$, $\{v_t\}$ sera homoscédastique. En résumé, l'équation transformée

$$y_t/\sqrt{h_t} = \beta_0(1/\sqrt{h_t}) + \beta_1(x_{t1}/\sqrt{h_t}) + \ldots + \beta_k(x_{tk}/\sqrt{h_t}) + v_t \qquad \textbf{[12.53]}$$

dispose d'erreurs AR(1). Si nous connaissons la forme particulière que prend l'hétéroscédasticité (autrement dit, si nous connaissons h_t), il nous suffit de prendre en compte l'autocorrélation de v_t en estimant (12.53) à l'aide de la méthode de CO ou PW. Dans la plupart des cas, la fonction h_t n'est pas connue. Nous pouvons néanmoins l'estimer en utilisant la procédure d'estimation de h_t décrite dans la section 8.4. Nous pouvons résumer maintenant toute la procédure comme suit.

Estimation par les MCQG en présence d'hétéroscédasticité et de corrélation sérielle d'ordre 1 dans les erreurs

i. Estimer (12.52) par les MCO pour obtenir les résidus, $\hat{u}_t$.

ii. Régresser $g_t = log(\hat{u}_t^2)$ sur $x_{t1}, \ldots, x_{tk}$ (ou sur $\hat{y}_t, \hat{y}_t^2$) pour obtenir les valeurs ajustées, $\hat{g}_t$.

iii. Calculer $\hat{h}_t = \exp(\hat{g}_t)$.

iv. Estimer la régression transformée

$$\hat{h}_t^{-1/2} y_t = \beta_0 \hat{h}_t^{-1/2} + \beta_1 \hat{h}_t^{-1/2} x_{t1} + \ldots + \beta_k \hat{h}_t^{-1/2} x_{tk} + error_t \qquad \textbf{[12.54]}$$

par la procédure des MCQG de Cochrane-Orcutt ou de Prais-Winsten.

Les estimateurs des MCQG qui découlent de cette procédure sont asymptotiquement efficaces, à condition que les hypothèses du modèle (12.52) soient vérifiées naturellement. Les écarts-types estimés et les tests d'hypothèse sont valides sur le plan asymptotique. Par contre, si nous pensons que la fonction de la variance n'est pas correctement spécifiée ou que les erreurs suivent un autre processus que le processus AR(1), il est préférable d'utiliser les variables en (quasi-)différence dans (12.54), puis d'estimer cette nouvelle régression par les MCO en recourant à la correction de Newey-West. Même en présence d'une hétéroscédasticité ou d'une corrélation sérielle mal spécifiée, cette procédure nous permet d'obtenir des estimateurs asymptotiquement efficaces et des tests d'inférence statistique (asymptotiquement) valides.

RÉSUMÉ

Nous avons abordé l'épineux problème de la corrélation sérielle dans les modèles de régression linéaire multiple. Il est fréquent d'observer une corrélation positive entre erreurs adjacentes dans les modèles statiques et les modèles à retards échelonnés finis. En présence d'autocorrélation, les écarts-types estimés des MCO et les tests d'inférence statistique ne sont pas fiables (bien que les $\hat{\beta}_j$ restent sans biais ou, à tout le moins, convergents). En règle générale, les écarts-types estimés des MCO sous-estiment la véritable incertitude qui existe dans l'estimation des paramètres.

Le modèle d'autocorrélation le plus courant est le modèle AR(1). Il s'agit d'un processus intéressant car le test t de corrélation sérielle, basé sur les résidus des MCO, est simple à réaliser. Pour obtenir une statistique t valide sur le plan asymptotique, il suffit de régresser les résidus des MCO sur leurs valeurs retardées adjacentes, en supposant naturellement que les hypothèses de stricte exogénéité et d'homoscédasticité soient respectées. Rendre ce test robuste à la présence d'hétéroscédasticité ne pose pas de problème. Une alternative au test t a été proposée par Durbin et Watson dont l'utilité du test reste néanmoins limitée. Uniquement valide sous les hypothèses du MRLC, leur test présente également une large zone au sein de laquelle aucune conclusion ne peut être tirée quant à la validité de l'hypothèse nulle d'absence de corrélation sérielle.

En l'absence de régresseurs strictement exogènes, le test t de $\hat{u}_t$ sur $\hat{u}_{t-1}$ reste valide, à condition néanmoins que la régression auxiliaire contienne également tous les régresseurs. Nous pouvons aussi utiliser les tests F et ML pour vérifier la présence d'autocorrélation d'ordre supérieure à 1.

En présence d'une corrélation d'ordre 1 et de régresseurs strictement exogènes, la correction des écarts-types estimés peut s'effectuer à l'aide d'une procédure basée sur les MCQG ; nous avons étudié celle de Cochrane-Orcutt et de Prais-Winsten. Notez bien que les estimations des MCQG sont obtenues par les MCO à partir de variables *en (quasi) différence* et que tous les tests statistiques sont asymptotiquement valides. La quasi-totalité des logiciels économétriques disposent de fonctions intégrées permettant l'estimation de modèles dont les erreurs suivent un processus AR(1).

Lorsque l'hypothèse de stricte exogénéité des régresseurs n'est pas respectée, une autre manière de lutter contre les méfaits de la corrélation sérielle est d'utiliser les MCO en recourant aux calculs des écarts-types estimés robustes à la présence d'autocorrélation (et d'hétéroscédasticité). La plupart des logiciels économétriques appliquent la méthode de Newey et West (1987).

Enfin, nous avons identifié les caractéristiques de l'hétéroscédasticité présente dans les séries temporelles. Comme dans l'analyse en coupe transversale, l'hétéroscédasticité fausse les tests d'hypothèse, sans pour autant violer les propriétés d'absence de biais ou de convergence des estimateurs des MCO. Dans les séries temporelles néanmoins, l'hétéroscédasticité reste moins préoccupante que la corrélation sérielle. Les tests de Breusch-Pagan et de White, que nous avons étudiés au chapitre 8, peuvent être directement appliqués aux séries temporelles, à condition que les erreurs ne soient pas autocorrélées. Depuis plusieurs années maintenant, les économistes, en particulier ceux qui étudient les marchés financiers, se sont intéressés aux formes dynamiques d'hétéroscédasticité. Le modèle ARCH, dans lequel le carré des résidus est autocorrélé, en représente l'exemple le plus connu.

MOTS-CLÉS

Corrélation sérielle AR(1) p. 487
Données (quasi-)différenciées p. 498
Écart-type robuste à la présence d'autocorrélation p. 507, 508
Estimation de Cochrane-Orcutt (CO) p. 500
Estimation de Prais-Winsten (PW) p. 500
Moindres carrés pondérés p. 511
Moindres carrés quasi-généralisés (MCQG) p. 499
Hétéroscédasticité conditionnelle autorégressive (ARCH) p. 512
Statistique de Durbin-Watson (*DW*) p. 492
Test de Breusch-Godfrey p. 496

EXERCICES

1. Lorsque les erreurs d'un modèle de régression suivent un processus AR(1), comment pouvez-vous expliquer que les écarts-types estimés des MCO sous-estiment la variation d'échantillonnage dans les $\hat{\beta}_j$? Est-ce toujours le cas ?

2. « Les méthodes de Cochrane-Orcutt et de Prais-Winsten sont toutes deux utilisées pour obtenir des écarts-types estimés des MCO valides en présence de corrélation sérielle dans les erreurs. » Qu'en pensez-vous ?

3. Dans l'exemple 10.6, nous avons estimé une variante du modèle de Fair dont l'objectif était de prédire les résultats de l'élection présidentielle aux États-Unis.

i. Quel argument peut-on utiliser pour défendre l'hypothèse selon laquelle le terme d'erreur n'est pas autocorrélé ? (*Astuce* : À quelle fréquence ont lieu les élections présidentielles ?)

ii. Si nous régressons les résidus des MCO de (10.23) sur leurs valeurs retardées adjacentes, nous obtenons $\hat{\rho} = -0{,}068$ et $\hat{\sigma}(\hat{\rho}) = 0{,}240$. Quelle conclusion tirez-vous quant à la présence d'autocorrélation dans les u_t ?

iii. Dans le cadre de ce test de corrélation sérielle, faut-il s'inquiéter de la petite taille de l'échantillon ?

4. Vrai ou faux : « si les erreurs d'un modèle de régression suivent un processus ARCH, elles doivent être autocorrélées. »

5. i. Dans l'étude d'événement sur les « zones entreprises » (abordée dans l'exercice sur ordinateur C5 du chapitre 10), la régression des résidus des MCO sur leurs valeurs retardées adjacentes donne les résultats suivants : $\hat{\rho} = 0{,}841$ et $\hat{\sigma}(\hat{\rho}) = 0{,}053$. Quelle en est la conséquence sur les estimations obtenues par les MCO ?

ii. Si vous désirez utiliser les MCO tout en bénéficiant d'un écart-type estimé valide pour le coefficient de EZ, que devez-vous faire ?

6. Dans l'exemple 12.8, nous avons identifié la présence d'hétéroscédasticité dans le terme d'erreur de l'équation (12.47). Une solution est alors de calculer les écarts-types estimés robustes. Dans les résultats ci-dessous, ils sont indiqués entre crochets alors que les écarts-types estimés traditionnels le sont entre parenthèses :

$$\widehat{return}_t = 0{,}180 + 0{,}059\ return_{t-1}$$
$$(0{,}081)\ (0{,}038)$$
$$[0{,}085]\ [0{,}069]$$
$$n = 689,\ R^2 = 0{,}0035,\ \overline{R}^2 = 0{,}0020.$$

En se basant sur le test t robuste à la présence d'hétéroscédasticité, quelle est la significativité statistique de $return_{t-1}$?

7. Considérons le modèle de régression multiple suivant :

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t.$$

Nous supposons que les hypothèses ST.1, ST.2, ST.3, et ST.4 sont vérifiées.

i. Vous appliquez la méthode de Prais-Winsten car vous pensez que les erreurs $\{u_t\}$ suivent un processus AR(1), dont le paramètre autorégressif est ρ. Si, en réalité, les erreurs suivent un modèle AR(2) ou MA(1), comment pouvez-vous expliquer que les écarts-types estimés de Prais-Winsten ne sont pas valides ?

ii. Comment pourriez-vous combiner les procédures de Newey-West et de Prais-Winsten pour obtenir des écarts-types estimés valides ? Précisez bien chacune des étapes requises. [*Astuce* : l'équation (12.32) peut aider ; si $\{u_t\}$ ne suit pas un processus AR(1), e_t peut être remplacée par $u_t - \rho u_{t-1}$, où ρ est la probabilité limite de l'estimateur de $\hat{\rho}$. L'erreur $\{u_t - \rho u_{t-1}\}$ contient-elle encore de la corrélation sérielle ? Que pouvez-vous faire si c'est encore le cas ?]

iii. Expliquez la raison pour laquelle votre réponse au point (ii) ne change pas lorsque l'hypothèse TS.4 n'est pas remplie.

EXERCICES SUR ORDINATEUR

C1. Dans l'exemple 11.6, nous avons estimé un modèle à retards échelonnés finis en différence première :

$$\Delta gfr_t = \gamma_0 + \Delta pe_t + \Delta pe_{t-1} + \Delta pe_{t-2} + u_t.$$

Utilisez la base de données FERTIL3 pour tester la présence d'une corrélation sérielle AR(1) dans les erreurs.

C2. i. En recourant à la base de données WAGEPRC, estimez le modèle à retards échelonnés que nous avons rencontré dans le problème 5 du chapitre 11. Effectuez le test de corrélation sérielle AR(1) en utilisant la régression (12.14).

ii. Estimez à nouveau le modèle en utilisant la méthode d'itération de Cochrane-Orcutt. Quelle est votre estimation du multiplicateur (ou propension) de long terme ?

iii. Calculez l'écart-type estimé de ce multiplicateur de long terme à l'aide de la méthode de CO. (Vous devez donc estimer une équation en différence.) Déterminez si votre estimation du multiplicateur est statistiquement différente de 1 à un seuil de 5 %.

C3. i. Au point (i) de l'exercice sur ordinateur C6 du chapitre 11, nous avons estimé le modèle de l'accélérateur basé sur les investissements en stocks : $\Delta inven_t = \beta_0 + \beta_1\ \Delta GDP_t + u_t$. Testez maintenant si le terme d'erreur de cette équation contient une corrélation sérielle AR(1).

ii. Si vous identifiez la présence de corrélation sérielle, estimez à nouveau le modèle en utilisant la méthode de CO, puis comparez les résultats.

C4. i. Utilisez la base de données NYSE pour estimer l'équation (12.48). Soit $\hat{h}_t$, les valeurs ajustées de cette équation, qui correspondent aux estimations de la variance conditionnelle. Combien de $\hat{h}_t$ sont négatives ?

ii. Ajoutez $return_{t-1}^2$ dans (12.48) et calculez à nouveau les valeurs ajustées, $\hat{h}_t$. Reste-t-il des $\hat{h}_t$ négatives ?

iii. Utilisez les MCP pour estimer (12.47) en utilisant les $\hat{h}_t$ du point (ii). Comparez votre estimation de β_1 avec celle de l'équation (11.16). Testez l'hypothèse H_0 : $\beta_1 = 0$ et comparez le résultat à celui obtenu à l'aide des MCO.

iv. Estimez maintenant (12.47) par les MCP après avoir obtenu les $\hat{h}_t$ en recourant au modèle ARCH décrit en (12.51). Vos conclusions sont-elles différentes de celles obtenues au point (iii) ?

C5. Considérons la version du modèle de Fair auquel nous avons eu recours dans l'exemple 10.6. Au lieu de prédire la proportion des votes que capte le candidat démocrate, estimez un modèle de probabilité linéaire qui explique la probabilité de remporter l'élection pour les démocrates.

i. Utilisez la variable binaire *demwins* au lieu de *demvote* dans (10.23), puis présentez les résultats de la régression sous leur forme habituelle. Quelles variables déterminent la probabilité de gagner l'élection ? Utilisez les données jusqu'en 1992.

ii. Combien de valeurs ajustées sont inférieures à zéro ? Combien d'entre elles sont supérieures à un ?

iii. Utilisez la règle suivante : « si $\widehat{demwins} > 0{,}5$, les démocrates gagnent l'élection ». Sur base des 20 élections reprises dans l'échantillon, combien de fois cette prévision correspond-elle à la réalité ?

iv. Utilisez les valeurs de 1996 pour les variables explicatives et prédisez la probabilité que Clinton remporte l'élection. Votre prévision correspond-elle à la réalité ?

v. Utilisez un test *t* robuste à l'hétéroscédasticité pour tester la présence de corrélation sérielle AR(1) dans les erreurs. Quelle est votre conclusion ?

vi. Calculez et utilisez les écarts-types estimés robustes à l'hétéroscédasticité dans (i). La significativité statistique des variables explicatives est-elle différente ?

C6. i. Dans l'exercice sur ordinateur C7 du chapitre 10, nous avons estimé une relation simple entre le taux de croissance de la consommation et celui du revenu disponible. Vérifiez si le terme d'erreur de cette équation contient une corrélation sérielle AR(1). La base de données est CONSUMP.

ii. Dans l'exercice sur ordinateur C7 du chapitre 11, nous avons testé l'hypothèse du revenu permanent en régressant le taux de croissance de la consommation sur sa valeur retardée adjacente. Après avoir estimé cette régression, testez la présence d'hétéroscédasticité dans les erreurs en régressant le carré des résidus sur gc_{t-1} et gc_{t-1}^2. Qu'en concluez-vous ?

C7. i. Obtenez les estimations de Cochrane-Orcutt pour l'exemple 12.4 en utilisant la base de données BARIUM.

ii. Les estimations de CO sont-elles similaires à celles de Prais-Winsten ? Pouvait-on l'anticiper ?

C8. Dans cet exercice, vous devez utiliser la base de données TRAFFIC2.

i. Estimez par les MCO la régression de *prcfat* sur une tendance temporelle linéaire, des variables binaires mensuelles, et les variables *wkends*, *unem*, *spdlaw*, et *beltlaw*. Testez la présence de corrélation

sérielle AR(1) dans les erreurs en utilisant la régression (12.14). Est-il logique d'utiliser un test qui requiert la stricte exogénéité des régresseurs ?

ii. Estimez les écarts-types estimés robustes à la corrélation sérielle et à l'hétéroscédasticité pour les coefficients *spdlaw* et *beltlaw*, en considérant quatre retards dans la formule de Newey-West. Que devient la significativité statistique de ces deux variables ?

iii. Estimez le modèle par la méthode itérative de Prais-Winsten et comparez les estimations à celles des MCO. Observez-vous d'importantes modifications sur le plan des coefficients estimés et de la significativité statistique ?

C9. Le fichier FISH contient 97 observations journalières sur les prix et quantités de poisson échangées sur le marché de Fulton à New York. Utilisez log(*avgprc*) comme variable dépendante.

i. Régressez log(*avgprc*) sur quatre variables binaires relatives aux jours de la semaine, en considérant le vendredi comme jour de référence. Ajoutez une tendance temporelle linéaire. Pouvez-vous conclure qu'il existe une variation significative du prix entre les jours de la semaine ?

ii. Ajoutez maintenant les variables *wave2* et *wave3*, qui mesurent les hauteurs de vague en mer au cours des journées précédentes. Ces variables sont-elles significatives sur le plan statistique ? Expliquez le mécanisme par lequel une mer agitée peut conduire à une augmentation du prix du poisson.

iii. Que devient la tendance temporelle lorsque les variables *wave2* et *wave3* sont présentes dans la régression ? Comment expliquez-vous ce résultat ?

iv. Expliquez la raison pour laquelle il est raisonnable de penser que toutes les variables explicatives dans la régression sont strictement exogènes.

v. Tester la présence de corrélation sérielle d'ordre 1.

vi. Estimez les écarts-types de Newey-West en utilisant quatre retards. Comment évoluent les statistiques *t* de *wave2* et *wave3* ? Fallait-il s'attendre à une plus faible ou à une plus grande variation par rapport aux MCO ?

vii. Estimez à nouveau le modèle du point (ii) en utilisant la procédure itérative de Prais-Winsten. Les variables *wave2* et *wave3* sont-elles conjointement significatives sur le plan statistique ?

C10. Utilisez toutes les données disponibles dans PHILLIPS pour répondre aux questions suivantes.

i. Estimez par les MCO la version statique de la courbe de Phillips, $inf_t = \beta_0 + \beta_1 unem_t + u_t$, puis présentez les résultats de manière habituelle.

ii. En se servant des résidus des MCO du point (i), $\hat{u}_t$, estimez ρ en régressant $\hat{u}_t$ sur $\hat{u}_{t-1}$. (Vous pouvez utiliser une constante dans cette régression.) Peut-on conclure à la présence d'autocorrélation ?

iii. Estimez maintenant la courbe de Phillips en utilisant la méthode itérative de Prais-Winsten. Comparez l'estimation de β_1 à celle indiquée dans le tableau 12.2 ? Constatez-vous une forte différence entre les deux estimations (sachant que nous avons pris en compte les dernières années dans cet exercice) ?

iv. Au lieu de recourir à Prais-Winsten, utilisez Cochrane-Orcutt. Les estimations de ρ sont-elles similaires ? Qu'en est-il des estimations de β_1 ?

C11. Utilisez la base de données NYSE pour répondre aux questions suivantes.

i. Estimez l'équation (12.47) par les MCO pour en extraire les carrés des résidus, $\hat{u}_t^2$. Calculez ensuite la moyenne, le minimum et le maximum de cette série.

ii. Utilisez les résidus au carré du point (i) pour estimer le modèle d'hétéroscédasticité suivant :

$$\text{Var}(u_t|return_{t-1}, return_{t-2}, \ldots) = \text{Var}(u_t|return_{t-1}) = \delta_0 + \delta_1 return_{t-1} + \delta_2 return_{t-1}^2.$$

Quelles sont les estimations des coefficients, des écarts-types estimés, du R carré, et du R carré ajusté ?

iii. En fonction des résultats obtenus en (ii), quelle est la valeur de $return_{t-1}$ qui minimise la variance ? Quelle est la valeur de cette variance ?

iv. Obtenez-vous des estimations négatives de la variance à partir du modèle décrit au point (ii) ?

v. Le modèle du point (ii) est-il supérieur ou inférieur au modèle ARCH(1) que nous avons utilisé dans l'exemple 12.9 ? Expliquez.

vi. Dans le modèle ARCH(1) de l'équation (12.51), ajoutez un second retard, $\hat{u}_{t-2}^2$. Ce retard joue-t-il un rôle important ? Ce modèle ARCH(2) est-il préférable au modèle utilisé au point (ii) ?

C12. Utilisez la base de données INVEN pour cet exercice, comme nous l'avons fait dans l'exercice C6 du chapitre 11.

i. Obtenez les résidus des MCO pour le modèle de l'accélérateur $\Delta inven_t = \beta_0 + \beta_1 \Delta GDP_t + u_t$ et régresser $\hat{u}_t$ sur $\hat{u}_{t-1}$ pour tester la présence de corrélation sérielle dans u_t. Quelle est l'estimation de ρ ? L'autocorrélation représente-t-elle un sérieux problème ?

ii. Estimez le modèle de l'accélérateur par la méthode de PW et comparez l'estimation de β_1 à celle des MCO. Pourquoi faut-il s'attendre à des estimations similaires ?

C13. Utilisez la base de données OKUN pour répondre aux questions suivantes. Nous avons utilisé ces données dans l'exercice C11 du chapitre 11.

i. Estimez l'équation $pcrgdp_t = \beta_0 + \beta_1 cunem_t + u_t$ et tester la présence de corrélation sérielle AR(1) dans les erreurs *sans* recourir à l'hypothèse que $\{cunem_t: t = 1, 2, \ldots\}$ est strictement exogène. Quelle est votre conclusion ?

ii. Régressez les résidus au carré, $\hat{u}_t^2$, sur $cunem_t$, ce qui correspond au test de Breusch-Pagan dans le cas d'une régression simple. Quelle est votre conclusion ?

iii. Utilisez les écarts-types estimés robustes à la présence d'hétéroscédasticité pour l'estimation $\hat{\beta}_1$ des MCO. Ces écarts-types estimés diffèrent-ils sensiblement des écarts-types estimés classiques des MCO ?

C14. Utilisez la base de données MINWAGE pour cet exercice en se concentrant sur le secteur 232.

i. Estimez l'équation $gwage232_t = \beta_0 + \beta_1 gmwage_t + \beta_2 gcpi_t + u_t$ par les MCO et testez la présence d'une corrélation sérielle AR(1) dans les erreurs. Est-il important de vérifier si $gmwage_t$ et $gcpi_t$ sont strictement exogènes ? Quelle conclusion tirez-vous de ce test ?

ii. Estimez les écarts-types de Newey-West pour le modèle du point (i) en utilisant un seul retard égal à 12. Les écarts-types estimés de Newey-West sont-ils sensiblement différents des écarts-types estimés standards des MCO ?

iii. Estimez les écarts-types des MCO robustes à l'hétéroscédasticité et comparez-les aux écarts-types estimés standards ainsi qu'à ceux de Newey-West. Identifiez-vous un problème lié à la corrélation sérielle et à l'hétéroscédasticité ? Si oui, lequel des deux problèmes est le plus préoccupant ?

iv. Effectuez le test de Breusch-Pagan dans l'équation de départ et vérifiez s'il existe une hétéroscédasticité importante dans les erreurs.

v. Ajoutez les retards 1 à 12 de *gmwage* dans l'équation du point (i). Obtenez la *p*-valeur du test *F* de significativité jointe des retards 1 à 12. Effectuez à nouveau le test en le rendant robuste à la présence d'hétéroscédasticité. Quel en est l'impact sur la significativité jointe des retards ?

vi. Obtenez la *p*-valeur du test de significativité jointe du point (v) en utilisant cette fois-ci la correction de Newey-West. Quel enseignement en tirez-vous ?

vii. Si vous laissez tomber les retards de *gmwage*, aboutissez-vous à une estimation de la propension de long terme fort différente ?

C15. Utilisez la base de données BARIUM pour répondre aux questions suivantes.

i. Dans la tableau 12.1, les écarts-types estimés des MCO sont systématiquement inférieurs à ceux des MCQG (Prais-Winsten). Expliquez la raison pour laquelle cette comparaison est fallacieuse.

ii. Estimez à nouveau l'équation qui caractérise la colonne « MCO » du tableau 12.1 en utilisant cette fois-ci les écarts-types estimés de Newey-West basés sur une fenêtre de 4 mois ($g = 4$). Pour la variable *lchempi*, comparez les écarts-types estimés de Newey-West à ceux des MCO et de Prais-Winsten. Que devient cette comparaison pour la variable *afdec*6 ?

iii. Suivez les mêmes instructions qu'au point (ii) en fixant $g = 12$. Que deviennent les écarts-types estimés de *lchempi* et de *afdec*6 lorsque la fenêtre passe de 4 à 12 ?

C16. Utilisez les données du fichier APPROVAL pour répondre aux questions suivantes. Consultez également l'exercice sur ordinateur C14 du chapitre 11.

i. Estimez l'équation

$$approve_t = \beta_0 + \beta_1 lcpifood_t + \beta_2 lrgasprice_t + \beta_3 unemploy_t + \beta_4 sep11_t + \beta_5 iraqinvade_t + u_t$$

en différences premières (DP). Vérifiez si les erreurs de cette équation en DP souffrent de corrélation sérielle d'ordre 1. Testez-le en régressant $\hat{e}_t$ sur $\hat{e}_{t-1}$ sachant que $\hat{e}_t$ correspond aux résidus de cette équation en DP estimée par MCO.

ii. Estimez cette équation en DP par le méthode de Prais-Winsten. Quelle est l'estimation de β_2 par rapport à celle obtenue au point (i) par MCO ? Qu'en est-il de la significativité sur le plan statistique ?

iii. Dans l'équation en DP du point (i) estimée par MCO, utilisez les écarts-types de Newey-West en utilisant un retard, quatre retards, et huit retards. Qu'en est-il de la significativité sur le plan statistique de l'estimation de β_2 dans chacun de ses trois cas de figure ?

PARTIE 3

THÈMES AVANCÉS

Nous nous focalisons maintenant sur des sujets plus spécialisés qui ne sont habituellement pas traités dans un cours d'introduction d'un semestre. Certains de ces sujets nécessitent plus d'outils mathématiques que l'analyse par régression multiple des parties 1 et 2. Dans le chapitre 13, nous montrons comment appliquer la régression multiple aux coupes transversales indépendantes empilées. Les problèmes soulevés sont très semblables à ceux que l'on rencontre dans l'analyse standard des coupes transversales. La principale différence est qu'il est à présent possible d'étudier la façon dont les relations varient au cours du temps en incluant des variables indicatrices temporelles. Nous illustrons également la façon dont un jeu de données de panel peut être analysé dans un cadre de régression. Le chapitre 14 fait la lumière sur un ensemble de méthodes plus approfondies de traitement de données de panel. Ces méthodes, pour être plus complexes, n'en sont pas moins couramment utilisées dans la plupart des travaux appliqués.

Les chapitres 15 et 16 examinent le problème des variables explicatives endogènes. Le chapitre 15 présente la méthode des variables instrumentales et montre qu'elle permet de résoudre les problèmes de variables omises ainsi que les problèmes d'erreurs de mesure. Le chapitre 16 aborde la méthode des doubles moindres carrés. Souvent utilisée dans les travaux d'économie appliquée, elle est indispensable pour estimer les modèles à équations simultanées.

Le chapitre 17 traite de sujets relativement complexes utilisés dans les analyses en coupe transversale, comme les modèles pour les variables dépendantes qualitatives et les méthodes de correction des biais de sélection des échantillons. Le chapitre 18 est orienté dans une direction différente puisqu'il traite de certaines avancées récentes de l'économétrie des séries temporelles particulièrement utiles pour estimer des relations dynamiques.

Le chapitre 19 sera utile pour les étudiants qui doivent rédiger un mémoire ou d'autre types de travaux dans le domaine des sciences sociales appliquées. Ce chapitre propose des conseils pour choisir une question de recherche, collecter et analyser les données, et pour écrire un article.

CHAPITRE

13

EMPILER DES DONNÉES EN COUPES TRANSVERSALES DE PÉRIODES DIFFÉRENTES : MÉTHODES DE DONNÉES DE PANEL SIMPLE

Traduction de Maëlys de la Rupelle

Jusque là, notre étude de la régression multiple nous a permis d'analyser deux types de données bien distincts : des données en coupe transversale, ou des séries temporelles. Ces données sont fréquemment utilisées ; néanmoins, un nombre croissant de travaux empiriques s'appuient sur des données qui combinent les deux dimensions, et possèdent les caractéristiques des coupes transversales et des séries temporelles. Les méthodes de régression multiple s'appliquent également à ces jeux de données. Ces données permettent souvent de résoudre d'importantes questions politiques. Nous en donnerons plusieurs illustrations dans ce chapitre.

Dans ce chapitre, nous allons étudier deux sortes de jeux données : les coupes transversales indépendantes empilées, et les données de panel. **Un jeu de coupes transversales indépendantes empilées** s'obtient en échantillonnant de façon aléatoire une population de grande taille à différentes dates (souvent différentes années).

Par exemple, nous pourrions, chaque année, constituer un échantillon aléatoire parmi la population des salariés aux États-Unis, et collecter des informations sur le salaire horaire, le niveau d'étude, l'expérience, etc. Nous pourrions, tous les deux ans, sélectionner un échantillon aléatoire parmi les maisons vendues dans une zone urbaine particulière, et enregistrer le prix de vente, la surface en mètre carrés, le nombre de salles de bain, etc. D'un point de vue statistique, ces jeux de données ont une caractéristique particulière : les observations qu'ils contiennent ont été échantillonnées indépendamment. Cette manière de procéder constitue aussi un élément clé dans notre analyse de données en coupes transversales : elle supprime la corrélation entre les termes d'erreur de différentes observations.

Un jeu de coupes transversales indépendantes empilées et une coupe transversale n'ont pas les mêmes propriétés. La coupe transversale provient d'un échantillon aléatoire unique. Quand on répète l'échantillonnage à des dates différentes, on obtient des observations qui ne sont pas distribuées de manière identique. Par exemple, dans la plupart des pays, la distribution des salaires et des années d'études a changé au cours du temps. Comme nous le verrons, en pratique il est facile de traiter cette particularité dans un modèle de régression multiple : il suffit de permettre à la constante, et dans certains cas aux coefficients de pente, de varier au cours du temps. Nous examinerons de tels modèles dans la section 13.1. Dans la section 13.2, nous verrons utiliser des coupes transversales empilées pour analyser les conséquences d'une réforme politique.

Un jeu de **données de panel** conjugue lui aussi des dimensions individuelles et temporelles. Néanmoins, il ne se constitue pas de la même manière qu'un jeu de coupes transversales indépendantes empilées. Afin de collecter un panel de données – parfois appelées **données longitudinales** – nous suivons (ou essayons de suivre) les *mêmes* individus, familles, entreprises, villes, États, ou autres, au cours du temps. Par exemple, pour constituer un jeu de données de panel donnant des informations sur les salaires individuels, le nombre d'heures de travail, les années d'études, etc., on choisira de façon aléatoire des individus dans une population à un moment donné. Ensuite, ces mêmes personnes seront réinterrogées à plusieurs dates. Nous obtiendrons alors des données concernant les salaires, le nombre d'heures de travail, le nombre d'années d'études, etc. d'un même groupe d'individus pour différentes années.

Les jeux de données de panel sont assez faciles à collecter au niveau des écoles, des villes, des départements, des régions ou des pays. L'utilisation de données de panel améliore fortement l'analyse des politiques publiques ; nous allons aborder quelques exemples ci-après. Dans l'analyse économétrique de données de panel, nous ne pouvons pas supposer que les observations faites à des périodes différentes sont indépendantes. Par exemple, des facteurs non observés qui affectent le salaire d'un individu en 1990 affecteront également le salaire de cette même personne en 1991 ; c'est le cas, par exemple, des capacités individuelles. Les facteurs non observés qui affectent le taux de criminalité d'une ville en 1985 affecteront aussi le taux de criminalité de cette ville en 1990. C'est pourquoi des méthodes et des modèles spécifiques ont été développés pour analyser les données de panel. Dans les sections 13.3, 13.4 et 13.5, nous décrivons la méthode des différences premières, une méthode simple permettant de retirer les facteurs non observés et constants au cours du temps qui affectent les entités étudiées. Puisque les méthodes de données de panel sont légèrement plus complexes, nous nous baserons principalement sur notre intuition pour décrire les propriétés statistiques des procédures d'estimation, laissant les hypothèses plus détaillées pour la partie annexe de ce chapitre.

13.1 EMPILER DES COUPES TRANSVERSALES INDÉPENDANTES DE PÉRIODES DIFFÉRENTES

De nombreuses enquêtes sur des individus, des familles et des entreprises se répètent à intervalles réguliers, souvent chaque année. C'est le cas par exemple du recensement de la population active aux États-Unis, le « Current Population Survey » (ou CPS), qui échantillonne de façon aléatoire les foyers chaque année. (Voir par exemple CPS78_85, qui contient les données des CPS de 1978 et 1985). Si, à chaque nouvelle période, on sélectionne un nouvel échantillon aléatoire, et qu'on empile les différents échantillons aléatoires, nous obtenons ce qu'on appelle des coupes transversales indépendantes empilées.

Une des raisons qui justifie l'utilisation de coupes transversales indépendantes empilées est qu'elles permettent d'augmenter la taille de l'échantillon. En empilant les échantillons aléatoires collectés pour une même population, mais à différents moments, nous pouvons obtenir des estimateurs plus précis et des tests statistiques plus puissants. Dans ce cas précis, l'empilement n'est utile que dans la mesure où la relation entre la variable dépendante et au moins quelques-unes des variables indépendantes reste constante au cours du temps.

Comme nous l'avons indiqué dans l'introduction, l'utilisation de coupes transversales empilées entraîne seulement des complications statistiques mineures. Par exemple, pour prendre en compte le fait que la population peut avoir des distributions différentes à différentes périodes, nous permettons à la constante du modèle de varier selon les périodes, habituellement des années. Cela se fait facilement en incluant des variables indicatrices pour toutes les années sauf une, que l'on prend comme année de référence – souvent la première année de l'échantillon. Il est aussi possible que la variance de l'erreur change au cours du temps, nous en discuterons ultérieurement.

Parfois, l'évolution des coefficients des variables indicatrices annuelles est elle-même intéressante. Par exemple, un démographe peut s'intéresser à la question suivante : *après* avoir pris en compte le nombre d'années d'études, les caractéristiques relatives à la fécondité chez les femmes de plus de 35 ans ont-elles changé entre 1972 et 1984 ? L'exemple suivant illustre comment cette question peut être traitée simplement en utilisant une analyse par régression multiple avec des variables indicatrices annuelles.

EXEMPLE 13.1
La fertilité des femmes et son évolution

Le jeu de données dans le fichier FERTIL1, qui est semblable à celui utilisé par Sander (1992), provient de l'Enquête Générale Sociale (General Society Survey) du « National Opinion Research Center »[1] pour les années 1972 à 1984. Nous utilisons ces données pour estimer un modèle qui expliquerait le nombre total d'enfants par femme (*kids*).

Nous nous posons la question suivante : que peut-on dire de l'évolution du taux de fertilité, une fois qu'on tient compte de l'évolution de l'éducation, ou d'autres caractéristiques observables ? Les caractéristiques que nous voulons prendre en compte sont le nombre d'années d'études, l'âge, l'origine ethnique, la région du domicile à l'âge de 16 ans et le cadre de vie à l'âge de 16 ans. Les estimations sont reportées dans le tableau 13.1.

L'année de référence est 1972. Les coefficients des variables indicatrices annuelles montrent qu'il y a eu une forte baisse de la fertilité dans les années 1980. Par exemple, le coefficient de *y82* implique que, à éducation, âge et autres facteurs égaux, une femme avait en moyenne 0,52 moins d'enfants en 1982 qu'en 1972. C'est une baisse très importante : en considérant *educ*, *age* et les autres facteurs fixés, 100 femmes en 1982 sont censées avoir environ 52 enfants de moins que 100 femmes comparables en 1972. Puisque nous tenons compte de l'éducation, cette baisse ne doit rien à l'augmentation des niveaux moyens d'éducation, elle aussi responsable d'un déclin de la fertilité. (Le nombre moyen d'années d'études était de 12,2 ans en 1972 et de 13,3 ans en 1984). Les coefficients de *y82* et *y84* représentent la baisse de la fertilité due à d'autres facteurs qui ne sont pas compris dans les variables explicatives.

1 NDT : Une des plus grandes organisations indépendantes de recherche en sciences sociales aux États-Unis.

Étant donné que les variables indicatrices des années 1982 et 1984 sont individuellement assez significatives, il n'est pas surprenant qu'en tant que groupe, les variables indicatrices annuelles soient conjointement très significatives : le R carré de la régression sans ces dernières est de 0,1019, et cela conduit à $F_{6,1111} = 5,87$ associé à une p-valeur ≈ 0.

Les femmes ayant fait plus d'études ont moins d'enfants, et le coefficient est très significatif. Toutes choses égales par ailleurs, 100 femmes qui ont fait des études universitaires auront environ 51 enfants de moins que 100 femmes qui n'ont qu'un niveau d'études secondaires : 0,128(4) = 0,512. L'âge a un effet décroissant sur la fertilité. (Le point d'inflexion dans la fonction quadratique se situe à environ *age* = 46, âge auquel beaucoup de femmes ne peuvent plus avoir d'enfant).

Le modèle estimé dans le tableau 13.1 suppose que l'effet de chaque variable explicative, et notamment du niveau d'études, est resté constant. Cela n'est pas forcément vrai ; il vous sera demandé de tester cette question dans l'exercice sur ordinateur C1.

En outre, il peut y avoir de l'hétéroscédasticité dans le terme d'erreur sous-jacent à l'équation estimée. On peut la traiter en utilisant les méthodes du chapitre 8. Il convient de noter une différence intéressante ici : désormais, la variance de l'erreur peut varier au cours du temps même si elle ne change pas avec les valeurs de *educ*, *age*, *black*, etc. Néanmoins, les écarts-types estimés robustes à l'hétéroscédasticité et les tests statistiques restent valides. Le test de Breusch-Pagan peut s'obtenir en régressant les résidus des MCO au carré sur *toutes* les variables indépendantes du tableau 13.1, y compris les variables indicatrices annuelles. (Pour le cas particulier de la statistique de White, les valeurs ajustées $\widehat{kids}$ et les valeurs ajustées au carré sont utilisées comme variables indépendantes, comme toujours). La procédure des moindres carrés pondérés permettrait de rendre compte des variances susceptibles de changer au cours du temps. Dans la procédure discutée dans la section 8.4, les variables binaires annuelles seraient incluses dans l'équation (8.32).

Tableau 13.1 *Déterminants de la fertilité des femmes*

Variable dépendante : *kids*		
Variables indépendantes	Coefficients	Écarts-types estimés
educ	−0,128	0,018
age	0,532	0,138
age2	−0,0058	0,0016
Black	1,076	0,174
east	0,217	0,133
northcen	0,363	0,121
west	0,198	0,167
farm	−0,053	0,147
othrural	−0,163	0,175
town	0,084	0,124
smcity	0,212	0,160
y74	0,268	0,173
y76	−0,097	0,179

Variable dépendante : *kids*		
Variables indépendantes	Coefficients	Écarts-types estimés
y78	−0,069	0,182
y80	−0,071	0,183
y82	−0,522	0,172
y84	−0,545	0,175
constante	−7,742	3,052
$n = 1,129$		
$R^2 = 0,1295$		
$\bar{R}^2 = 0,1162$		

Nous pouvons aussi faire interagir une variable indicatrice annuelle avec les variables explicatives clés pour voir si les effets de cette variable ont changé d'une période à l'autre. L'exemple suivant étudie la façon dont le rendement de l'éducation et l'écart de salaire entre les sexes ont changé entre 1978 et 1985.

Pour aller plus loin 13.1

À partir des résultats du tableau 13.1, peut-on dire que, toutes choses égales par ailleurs, une femme d'origine afro-américaine est supposée avoir un enfant de plus qu'une femme d'une autre origine ?

EXEMPLE 13.2
Changements dans le rendement des années d'études et dans l'écart de salaire entre les sexes

On s'intéresse à un modèle de régression de log(*wage*) (où *wage* est le salaire horaire) sur données empilées pour les années 1978 (l'année de référence) et 1985 :

$$\log(wage) = \beta_0 + \delta_0\, y85 + \beta_1 educ + \delta_1\, y85.educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 union + \beta_5\, female + \delta_5\, y85.female + u \qquad [13.1]$$

La plupart des variables explicatives sont considérées comme connues. La variable *union* est une variable indicatrice égale à un si la personne est syndiquée, zéro sinon. La variable *y85* est une variable indicatrice égale à un si l'observation a été faite en 1985 et zéro si elle date de 1978. Il y a 550 personnes dans l'échantillon de 1978 et 534 dans celui de 1985.

La constante pour 1978 est β_0, celle de 1985 est $\beta_0 + \delta_0$, et le rendement des études en 1985 est $\beta_1 + \delta_1$. Par conséquent, δ_1 permet de mesurer à quel point le rendement d'une année d'études supplémentaire a changé entre 1978 et 1985. En 1978, la différence entre le log(*wage*) des hommes et des femmes est β_5 ; en 1985, la différence est $\beta_5 + \delta_5$. Ainsi, nous pouvons tester l'hypothèse nulle selon laquelle l'écart de salaire entre les sexes n'a pas changé au cours de ces sept ans en testant $H_0 : \delta_1 = 0$. L'alternative, à savoir que le différentiel de salaire entre les sexes a *diminué*, est $H_1 : \delta_1 > 0$. Pour simplifier, nous supposons que l'expérience et l'appartenance à un syndicat ont les mêmes effets sur les salaires quelle que soit l'année considérée.

Avant de pouvoir estimer ce modèle, nous devons résoudre un autre problème – concernant le salaire horaire, qui est ici en valeur nominale (ou courante) en dollars. Puisque les salaires nominaux augmentent mécaniquement avec l'inflation, il est très important de considérer l'effet de chaque variable explicative sur les salaires réels. Cela nécessite d'exprimer les salaires de 1985 en dollars de 1978 et pour cela de les déflater. (En utilisant l'indice des prix à la consommation (IPC) donné par l'*Economic Report of the President* de 1997, l'indice de déflation est de 107,6/65,2 ≈ 1,65.)

Bien que nous puissions facilement diviser chaque salaire de 1985 par 1,65, cela ne sera pas nécessaire si nous incluons une variable indicatrice pour l'année 1985 dans la régression *et* si nous utilisons log(*wage*) (et non le salaire) comme variable dépendante. L'utilisation d'un salaire réel ou d'un salaire nominal dans une forme fonctionnelle logarithmique n'affecte que le coefficient de la variable indicatrice annuelle, *y85*. Pour l'observer, notons *P85* le facteur de déflation des salaires de 1985 (1,65 si nous utilisons l'IPC). Ensuite, le log des salaires réels pour chaque personne i dans l'échantillon de 1985 est donné par :

$$\log(wage_i/P85) = \log(wage_i) - \log(P85).$$

Alors que $wage_i$ change selon les individus, *P85* ne change pas. Par conséquent, log(*P85*) sera intégré dans la constante associée à l'année 1985. (Cette conclusion serait différente si, par exemple, nous avions utilisé un autre indice des prix pour les personnes vivant dans différentes zones du pays). En d'autres termes, pour étudier comment le rendement de l'éducation ou l'écart de salaire entre les sexes a changé, il n'est pas nécessaire de transformer les salaires nominaux en salaires réels dans l'équation (13.1). L'exercice sur ordinateur C2 vous permet de le vérifier.

Si nous oublions d'introduire des constantes différentes pour 1978 et 1985, l'utilisation des salaires nominaux peut donner des résultats erronés. Si nous utilisons *wage* plutôt que log(*wage*) comme variable dépendante, il est important d'utiliser le salaire réel et d'inclure une variable indicatrice annuelle.

La discussion précédente revient souvent lorsque nous utilisons des valeurs en dollars, que ce soit pour la variable dépendante ou pour les variables indépendantes. Si les montants en dollars apparaissent sous forme logarithmique et qu'on inclut des variables indicatrices pour toutes les périodes (à l'exception bien sûr de la période de référence), l'utilisation de déflateurs de prix agrégés n'aura un impact que sur les constantes ; aucune estimation des coefficients de pente ne changera.

Nous utilisons maintenant les données de CPS78_85 pour estimer l'équation :

$$\begin{aligned}\log(wage) = {} & 0{,}459 + 0{,}118\ y85 + 0{,}0747\ educ + 0{,}0185\ y85.educ \\ & (0{,}093)\ (0{,}124)\quad (0{,}0067)\quad (0{,}0094) \\ & + 0{,}0296\ exper - 0{,}00040\ exper^2 + 0{,}202\ union \\ & (0{,}0036)\quad 0{,}00008)\quad (0{,}030) \\ & - 0{,}317\ female + 0{,}085\ y85.female \\ & (0{,}037)\quad (0{,}051) \\ & n = 1\ 084,\ R^2 = 0{,}426,\ \bar{R}^2 = 0{,}422. \end{aligned} \qquad \textbf{[13.2]}$$

Le rendement de l'éducation en 1978 est estimé à environ 7,5 % ; le rendement de l'éducation en 1985 est d'environ 1,85 points de pourcentage *plus élevé*, soit environ 9,35 % (7,5 + 1,85). Puisque la statistique t du terme d'interaction est de 0,0185/0,0094 ≈ 1,97, la différence entre 1978 et 1985 en matière de rendement de l'éducation est statistiquement significative au seuil de 5 % par rapport à une hypothèse alternative bilatérale.

Qu'en est-il de l'écart entre les sexes ? En 1978, toutes choses égales par ailleurs, une femme gagnait environ 31,7 % de moins qu'un homme (l'estimation précise est 27,2 %[2]). En 1985, l'écart en log(*wage*) est −0,317 + 0,085 = −0,232. Donc l'écart entre les sexes semble avoir chuté entre 1978 et 1985 d'environ 8,5 points de pourcentage. Le statistique t relative au terme d'interaction est d'environ 1,67, ce qui signifie qu'il est significatif au niveau de 5 % par rapport à l'hypothèse alternative unilatérale positive.

Que se passe-t-il si nous faisons interagir toutes les variables indépendantes avec *y85* dans l'équation (13.2) ? Cela revient à considérer deux équations distinctes, une pour 1978 et une autre pour 1985. C'est parfois souhaitable. Par exemple, dans le chapitre 7, nous discutons d'une étude de Krueger (1993) dans laquelle il estime le rendement de l'utilisation d'un ordinateur au travail. Krueger considère deux équations séparées, une utilisant le recensement de la population active aux États-Unis (CPS, pour « Current Population Survey ») de 1984 et l'autre utilisant le CPS de 1989. En comparant la manière dont le rendement des années d'études change au cours du temps et selon que l'utilisation de l'ordinateur soit prise en compte ou non dans la régression, il estime qu'entre le tiers et la moitié de la hausse du rendement de l'éducation au cours de ces cinq années peut être imputée à l'augmentation de l'utilisation de l'ordinateur. [Voir les tableaux VIII et IX dans Krueger (1993).]

Le test de Chow : une étude du changement structurel dans le temps

Dans le chapitre 7, nous avons discuté de la façon dont le test de Chow – qui est simplement un test F – peut être utilisé pour déterminer si une fonction de régression multiple varie entre deux groupes. Nous pouvons aussi appliquer ce test à deux périodes de temps différentes. On peut calculer la somme des carrés des résidus à partir de l'estimation sur données empilées ; on l'appelle la « SCR contrainte ». La SCR non contrainte correspond quant à elle à la SCR pour les deux périodes de temps considérées séparément. Le calcul de ces statistiques est identique à celui de la section 7.4. Une version robuste à l'hétéroscédasticité est également possible (voir la section 8.2).

L'exemple 13.2 présente une autre manière de calculer le test de Chow pour deux périodes de temps en faisant interagir chaque variable avec la variable indicatrice annuelle de l'une des deux années, et en testant la significativité jointe de la variable indicatrice annuelle et de tous les termes d'interaction. La constante d'un modèle de régression change souvent au cours du temps (en raison par exemple de l'inflation du prix du logement), et le test de Chow permet de détecter ces changements. En général, il est plus intéressant d'avoir des constantes différentes, et de tester ensuite si certains coefficients de pente changent au cours du temps (comme nous l'avons fait dans l'exemple 13.2).

Un test de Chow peut aussi se calculer pour plus de deux périodes. Tout comme dans le cas de deux périodes, il est habituellement plus intéressant de laisser les constantes varier au cours du temps puis de tester si les coefficients de pente ont changé au cours du temps. Nous pouvons généralement tester la constance des coefficients de pente en interagissant toutes les variables indicatrices temporelles (exceptée celle définissant le groupe de référence) avec une, plusieurs ou toutes les variables explicatives, puis en testant la significativité jointe des termes d'interaction. Les exercices sur ordinateur C1 et C2 en fournissent des illustrations. S'il y a un grand nombre de périodes et de variables explicatives, la construction d'un jeu entier d'interactions peut être fastidieux. De manière alternative, il est possible d'adapter l'approche décrite dans la question (vi) de l'exercice sur ordinateur C11 du chapitre 7. Tout d'abord, il faut estimer le modèle contraint en faisant une régression sur données empilées, avec une constante différente pour chaque période ; cela nous donne la SCR constrainte, SCR_c. Ensuite, il faut faire une régression pour chaque période de temps T, et on obtient la somme des carrés des résidus pour chaque période. La somme des carrés des résidus non contrainte est obtenue ainsi : $SCR_{nc} = SCR_1 + SCR_2 + \ldots + SCR_T$. S'il y a k variables explicatives (qui n'incluent pas les constantes ni les variables indicatrices temporelles) avec T périodes temporelles, alors nous testons les $(T - 1)k$ restrictions, et il y a $T + Tk$ paramètres dans le modèle non contraint. Donc, si $n = n_1 + n_2 + \ldots + n_T$ est le nombre total

2 NDT : L'estimation précise est $exp(-0{,}317) - 1 = 27{,}2$ %, car la variable dépendante est en log. Lorsqu'on a un modèle où la variable dépendante est $log(Y)$, et qu'on veut calculer l'effet d'une variable binaire sur Y, on doit prendre l'exponentielle de son coefficient et soustraire un. Lorsque le coefficient de la variable binaire est proche de zéro, la valeur du coefficient n'est pas très éloignée de la valeur de exp(coefficient) – 1, c'est pourquoi l'auteur dit de l'effet qu'il est d'environ 31,7 %.

d'observations, alors les *ddl* du test F sont $(T - 1)k$ et $n - T - Tk$. Nous calculons la statistique F comme suit : $[(SCR_c - SCR_{nc})/SCR_{nc}][(n - T - Tk)/(T - 1)k]$. Malheureusement, comme tout test F basé sur la somme des carrés des résidus ou sur les R carrés, ce test n'est pas robuste à l'hétéroscédasticité (et notamment à des variances changeant au cours du temps). Pour obtenir un test robuste à l'hétéroscédasticité, nous devons construire les termes d'interaction et faire une régression sur données empilées.

13.2 ANALYSER DES POLITIQUES PUBLIQUES À PARTIR DE COUPES TRANSVERSALES EMPILÉES

Les coupes transversales empilées peuvent s'avérer très utiles pour évaluer l'impact d'un événement ou d'une politique spécifique. L'exemple suivant montre comment deux jeux de données en coupe transversale collectés avant et après l'arrivée d'un événement peuvent être utilisés pour déterminer ses effets en matière économique.

EXEMPLE 13.3
Effet de la localisation d'un incinérateur de déchets sur les prix de l'immobilier

Kiel et McClain (1995) ont étudié les effets de la présence d'un nouvel incinérateur de déchets sur le prix des logements à North Andover, Massachusetts, aux États-Unis. Ils ont utilisé des données sur plusieurs années et une analyse économétrique assez complexe. Nous utilisons ici des données sur deux ans et des modèles simplifiés, mais notre analyse est similaire.

La rumeur concernant la construction probable d'un incinérateur à North Andover naquit après 1978. Sa construction ne débuta qu'en 1981. L'incinérateur était supposé être fonctionnel rapidement après le début de la construction, mais ne commença à fonctionner qu'en 1985. Dans cet exemple, nous utilisons des données relatives au prix des maisons vendues en 1978 et en 1982 ; les échantillons de 1978 et de 1982 sont indépendants. L'hypothèse testée est que le prix des maisons louées à proximité de l'incinérateur est inférieur à celui des maisons plus éloignées.

Pour illustrer cela, nous estimons qu'une maison est proche de l'incinérateur si elle se situe à moins de 3 miles de l'incinérateur. [Dans l'exercice sur ordinateur C3, il vous est plutôt demandé d'utiliser la distance réelle entre la maison et l'incinérateur, comme dans Kiel et McCLain (1995)]. Nous nous intéressons à l'impact de l'incinérateur sur la valeur réelle des maisons. Cela suppose que nous considérions le prix en dollars constants. Nous exprimons le prix des maisons en dollars de 1978, en utilisant l'indice des prix de l'immobilier de Boston. Considérons que *rprice* désigne le prix des maisons en termes réels. Un analyste naïf utiliserait seulement les données de 1981 et considèrerait un modèle très simple :

$$rprice = \gamma_0 + \gamma_1 nearinc + u, \qquad \textbf{[13.3]}$$

où *nearinc* est une variable binaire égale à un si la maison est proche de l'incinérateur, et à zéro sinon. L'estimation de cette équation en utilisant les données du fichier KIELMC donne :

$$\widehat{rprice} = 101\,307{,}5 - 30\,688{,}27\, nearinc$$
$$(3\,093{,}0) \quad (5\,827{,}71)$$
$$n = 142,\ R^2 = 0{,}165. \qquad \textbf{[13.4]}$$

Puisqu'il s'agit d'une régression simple avec une variable binaire, la constante du modèle correspond au prix de vente moyen pour les maisons éloignées de l'incinérateur. Le coefficient de *nearinc* correspond, lui, à la différence entre le prix de vente moyen des maisons proches de l'incinérateur et celui des maisons éloignées. L'estimation montre que le prix moyen de vente pour le premier groupe était de 30 688,27 $ inférieur à celui du second groupe. La statistique t est supérieure à cinq en valeur absolue, nous pouvons donc rejeter l'hypothèse selon laquelle la valeur moyenne des maisons proches de l'incinérateur est identique à la valeur des maisons qui en sont éloignées.

Hélas, l'équation (13.4) n'implique *pas* que la localisation à proximité de l'incinérateur entraîne la baisse de la valeur des logements. En effet, si nous estimons le même modèle de régression pour l'année 1978 (soit avant même que la construction d'un nouvel incinérateur n'ait été évoquée), nous obtenons :

$$\widehat{rprice} = 82\,517{,}23 - 18\,824{,}37\,nearinc$$
$$(2\,653{,}79)\quad(4\,744{,}59)$$
$$n = 179,\ R^2 = 0{,}082. \qquad \textbf{[13.5]}$$

Par conséquent, même avant qu'il n'y ait de rumeurs évoquant la construction d'un incinérateur, la valeur moyenne d'une maison proche du site était inférieure de 18 824,37 $ à la valeur moyenne d'une maison éloignée du site (82 517,23 $) ; la différence est aussi statistiquement significative. Cela semble indiquer que l'incinérateur a été construit dans une zone où le prix des logements était plus bas qu'ailleurs.

Comment est-il possible alors de savoir si la construction d'un nouvel incinérateur fait chuter les valeurs immobilières ? Il suffit d'observer comment le coefficient de *nearinc* a changé entre 1978 et 1981. La différence de prix entre les zones proches et éloignées de l'incinérateur était plus importante en 1981 qu'en 1978 (30 688,27 $ *versus* 18 824,37 $), même si on l'exprime en pourcentage du prix moyen des maisons éloignées de l'incinérateur. La différence entre les deux coefficients de *nearinc* est :

$$\hat{\delta}_1 = -30\,688{,}27 - (-18\,824{,}37) = -11\,863{,}9$$

$\hat{\delta}_1$ estime l'effet de l'incinérateur sur la valeur de l'immobilier à proximité. On appelle $\hat{\delta}_1$ **l'estimateur de la différence des différences** puisqu'on peut l'exprimer ainsi :

$$\hat{\delta}_1 = (\overline{rprice}_{81,sp} - \overline{rprice}_{81,se}) - (\overline{rprice}_{78,sp} - \overline{rprice}_{78,se}), \qquad \textbf{[13.6]}$$

où *sp* correspond à un « site à proximité de l'incinérateur » et *se* correspond à un « site éloigné ». En d'autres termes, $\hat{\delta}_1$ est le changement entre 1978 et 1981 de la différence des prix de l'immobilier entre les deux sites.

Pour tester si $\hat{\delta}_1$ est statistiquement différent de zéro, nous devons obtenir son écart-type estimé. En fait, $\hat{\delta}_1$ peut s'obtenir en estimant l'équation suivante :

$$\widehat{rprice} = \beta_0 + \delta_0\, y81 + \beta_1 nearinc + \delta_1 y81.nearinc + u, \qquad \textbf{[13.7]}$$

à l'aide des données empilées sur les deux ans. La constante, β_0, est le prix moyen d'une maison éloignée de l'incinérateur en 1978. Le paramètre δ_0 représente le changement des prix de l'immobilier à North Andover de 1978 à 1981.

[Une comparaison des équations (13.4) et (13.5) montre que les prix de l'immobilier à North Andover ont fortement augmenté sur cette période par rapport à l'indice des prix de l'immobilier à Boston]. Le coefficient de *nearinc*, β_1, mesure l'effet de la localisation indépendamment de la présence de l'incinérateur : comme nous l'avons vu dans l'équation (13.5), même en 1978, les maisons proches du site qui allait héberger l'incinérateur valaient moins cher que celles qui en étaient plus éloignées.

Le coefficient associé au terme d'interaction *y81·nearinc* est particulièrement intéressant : δ_1 mesure le déclin de la valeur des logements dû à la proximité de l'incinérateur, si l'on suppose que les valeurs des maisons proches et éloignées du site n'ont pas évolué de manière différente pour d'autres raisons.

Les estimations de l'équation (13.7) sont reportées dans la colonne (1) du tableau 13.2. Le seul nombre que nous ne pouvons pas obtenir à partir des équations (13.4) et (13.5) est l'écart-type de $\hat{\delta}_1$. La statistique *t* associée à $\hat{\delta}_1$ est d'environ −1,59, ce qui est peu significatif en comparaison avec l'hypothèse alternative unilatérale (*p*-valeur ≈ 0,057).

Kiel et McClain (1995) incluent diverses caractéristiques du marché de l'immobilier dans leur analyse de l'emplacement de l'incinérateur. Il y a deux raisons qui justifient cela. Tout d'abord, le type de maisons vendues près de l'incinérateur en 1981 a du être différent de celui des maisons vendues près de l'incinérateur en 1978 ; dès lors, il est important de prendre en compte de telles caractéristiques. Ensuite, même si les caractéristiques des maisons concernées ne changent pas, le fait de les prendre en compte peut fortement diminuer la variance de l'erreur, ce qui peut ensuite réduire l'écart-type estimé de $\hat{\delta}_1$. (Voir la section 6.3 pour la discussion). Dans la colonne (2) nous prenons en compte l'âge des maisons en utilisant une spécification quadratique. Cela augmente significativement le R carré (en réduisant la variance résiduelle). Le coefficient de *y81·nearinc* est alors plus grand en valeur absolue et son écart-type estimé est inférieur.

En plus des variables d'âge de la colonne (2), la colonne (3) prend en compte la distance de la maison à la frontière de l'État mesurée en pieds (*intst*), la surface du terrain en pieds carrés (*land*), la surface de la maison en pieds carrés (*area*), le nombre de pièces (*rooms*), et le nombre de pièces d'eau (*baths*). Cela produit une estimation de *y81·nearinc* plus proche de celle qui ne tient pas compte de ces facteurs, par contre, cela conduit à un écart-type estimé plus petit : la statistique t pour $\hat{\delta}_1$ est d'environ $-2,84$. Par conséquent, nous trouvons un effet plus significatif dans la colonne (3) que dans la colonne (1). Les estimations de la colonne (3) sont préférables puisqu'elles prennent en compte plus de facteurs et que leurs écarts-types estimés sont plus petits (sauf pour la constante, mais c'est sans importance ici). Le fait que *nearinc* ait un coefficient plus petit et qu'il soit non significatif dans la colonne (3) indique que parmi les caractéristiques prises en compte dans le modèle de régression (3) figurent des facteurs jouant un rôle important dans la détermination des prix des logements.

Afin de présenter notre méthode, nous avons utilisé le prix réel des logements en niveau dans les régressions reportées dans le tableau 13.2. En réalité, il est préférable d'utiliser log(*price*) [ou log(*rprice*)] pour obtenir un effet en pourcentage. Le modèle de base devient :

$$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u. \qquad \textbf{[13.8]}$$

À présent, $100 \cdot \delta_1$ donne une approximation, en pourcentage, de la baisse de la valeur des logements imputable à l'incinérateur. [Comme dans l'exercice 13.2, l'utilisation de log(*price*) *versus* log(*rprice*) n'a d'impact que sur le coefficient de *y81*.] En utilisant les mêmes 321 observations empilées, nous obtenons :

$$\widehat{\log(price)} = \underset{(0,31)}{11,29} + \underset{(0,045)}{0,457}\, y81 - \underset{(0,055)}{0,340}\, nearinc - \underset{(0,083)}{0,063}\, y81 \cdot nearinc$$

$$n = 321,\ R^2 = 0,409. \qquad \textbf{13.9]}$$

Le coefficient du terme d'interaction indique que suite à la construction du nouvel incinérateur, les maisons qui en sont proches ont perdu environ 6,3 % de leur valeur. Cependant, cette estimation n'est statistiquement pas différente de zéro. Mais lorsque nous incluons toutes les variables de contrôle, comme dans la colonne (3) du tableau 13.2 (mais avec *intst*, *land*, et *area* sous la forme logarithmique), le coefficient de *y81· nearinc* devient $-0,132$ avec une statistique t d'environ $-2,53$. De nouveau, la prise en compte des autres facteurs s'avère d'une importance cruciale. L'utilisation de la forme logarithmique nous permet d'estimer que les maisons proches de l'incinérateur ont été dévaluées d'environ 13,2 %.

Tableau 13.2 *Des effets de la localisation de l'incinérateur sur les prix des logements*

	Variable dépendante *rprice*		
Variable indépendante	**(1)**	**(2)**	**(3)**
constante	82 517,23 (2 726,91)	89 116,54 (2 406,05)	13 807,67 (11 166,59)
y81	18 790,29 (4 050,07)	21 321,04 (3 443,63)	13 928,48 (2 798,75)
nearinc	– 18 824,37 (4 875,32)	9 397,94 (4 812,22)	3 780,34 (4 453,42)
y81·nearinc	– 11 863,90 (7 456,65)	– 21 920,27 (6 359,75)	– 14 177,93 (4 987,27)
Autres facteurs	Non	*age, age²*	Ensemble complet
Observations *R* carré	321 0,174	321 0,414	321 0,660

La méthodologie utilisée dans l'exemple précédent a de nombreuses applications, surtout lorsque les données proviennent d'une **expérience naturelle** (ou d'une **quasi-expérience**). On parle d'expérience naturelle lorsqu'un événement exogène – comme un changement de politique gouvernementale – modifie l'environnement dans lequel évoluent les individus, les familles, les entreprises ou les villes. Une expérience naturelle implique toujours l'existence d'un groupe contrôle, qui ne subit pas l'influence du changement de politique, et un groupe de traitement, qui est supposé être affecté par le changement de politique. Contrairement à une expérience véritable, dans laquelle les groupes de traitement et de contrôle sont choisis de façon explicite et aléatoire, dans les expériences naturelles, les groupes de contrôle et de traitement sont déterminés par le changement de politique. Pour tenir compte de l'influence des différences systématiques caractérisant les groupes de traitement et de contrôle, nous avons besoin de données observées à deux périodes différentes, une qui précède le changement de politique et l'autre qui lui succède. Ainsi, notre échantillon est réparti en quatre groupes : le groupe de contrôle avant le changement, le groupe de contrôle après le changement, le groupe de traitement avant le changement et le groupe de traitement après le changement.

Appelons C le groupe de contrôle et T le groupe de traitement. Soit dT une variable indicatrice égale à un pour le groupe de traitement T, et à zéro sinon, et $d2$ une variable indicatrice associée à la deuxième période (celle qui suit le changement de politique). Nous nous intéressons au modèle suivant :

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + \textit{autres facteurs}, \qquad \textbf{[13.10]}$$

où y est la variable qui nous intéresse. Comme dans l'exemple 13.3, δ_1 mesure les effets de la politique. En l'absence d'autres facteurs dans la régression, $\hat{\delta}_1$ sera l'estimateur de la différence des différences, aussi appelé estimateur des doubles différences.

$$\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{2,C}) - (\bar{y}_{1,T} - \bar{y}_{1,C}), \qquad \textbf{[13.11]}$$

où la barre symbolise la moyenne, le premier indice indique l'année et le second indice identifie le groupe (contrôle ou traitement).

La présentation habituelle des résultats d'estimation d'un modèle de différence de différences est illustrée dans le tableau 13.3. Le tableau 13.3 montre que le paramètre δ_1, souvent appelé **l'effet moyen du traitement** (puisqu'il mesure l'effet du « traitement » ou de la politique sur le résultat moyen de y) peut être estimé de deux façons : (1) En calculant d'abord la différence moyenne entre les groupes de traitement et de contrôle pour chaque période, puis en faisant la différence des résultats obtenus, de façon similaire à l'équation (13.11) ; (2) En calculant le changement de moyenne au cours du temps pour chacun des deux groupes, de traitement et de contrôle ; puis en faisant la différence entre ces changements. Cela signifie simplement que l'on écrit $\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{1,T}) - (\bar{y}_{2,C} - \bar{y}_{1,C})$. Évidemment, le $\hat{\delta}_1$ estimé ne dépend pas de la façon dont on a calculé la différence, comme on peut le voir en réarrangeant l'expression.

Tableau 13.3 *L'estimateur par différence de différences (aussi appelé estimateur des doubles différences)*

	Avant	Après	Après – avant
Groupe de contrôle	β_0	$\beta_0 + \delta_0$	δ_0
Groupe de traitement	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Traitement – contrôle	β_1	$\beta_1 + \delta_1$	δ_1

Lorsque les variables explicatives sont ajoutées dans l'équation (13.10) (pour prendre en compte le fait que la population échantillonnée n'est pas la même aux deux périodes), l'estimation de δ_1 par les MCO n'a plus la forme simple de (13.11), mais son interprétation est similaire.

EXEMPLE 13.4
De l'effet des lois d'indemnisation des employés victimes d'accidents du travail sur les semaines d'inactivité

Meyer, Viscusi, et Durbin (1995) (MVD ci-après) ont étudié la durée (en semaines) durant laquelle un salarié victime d'un accident du travail reçoit une indemnisation salariale. Le 15 juillet 1980, l'État du Kentucky a augmenté le plafond maximum des revenus hebdomadaires couverts par le système d'indemnisation des accidents du travail. L'augmentation de ce plafond n'a pas d'effet sur l'indemnisation des employés à faibles revenus. Par contre, suite à la réforme, les employés à hauts revenus reçoivent une indemnisation plus élevée. Pour eux, il devient moins coûteux de rester bénéficiaire du système d'indemnisation, et de ne pas reprendre le travail. Le groupe de contrôle est constitué par les employés à faibles revenus ; le groupe de traitement est formé par les employés à revenus élevés ; on définit les employés à revenus élevés comme ceux qui étaient limités par le plafond en vigueur avant la réforme. En utilisant des échantillons aléatoires enquêtés avant et après la réforme, MVD regardent si des indemnités plus élevées en cas d'accident du travail entraînent un arrêt maladie plus long (tout le reste étant fixé). Ils ont estimé un premier modèle en différence de différences, en utilisant log(*durat*) comme variable dépendante. Notons *afchnge* la variable indicatrice relative aux observations recueillies après le changement de politique et *highearn* la variable indicatrice du groupe à hauts revenus, notre groupe de traitement. En utilisant les données du fichier INJURY, on obtient les résultats suivants (l'écart-type estimé est reporté entre parenthèses) :

$$\widehat{\log(durat)} = 1{,}126 + 0{,}0077\,afchnge + 0{,}256\,highearn$$
$$(0{,}031)\ (0{,}0447)\qquad (0{,}047)$$
$$+ 0{,}191\ afchnge.highearn$$
$$(0{,}069)$$
$$n = 5\,626,\ R^2 = 0{,}021. \qquad \textbf{[13.12]}$$

Par conséquent, $\hat{\delta}_1 = 0{,}191$ (t = 20,77), ce qui implique que la durée moyenne d'indemnisation des employés à hauts revenus a augmenté de 19 % du fait de la réforme. Le coefficient de *afchnge* est faible et non statistiquement significatif : comme attendu, l'augmentation du plafond des revenus n'a pas eu d'effet sur la durée d'indemnisation des employés à faibles revenus.

C'est un bon exemple de la façon dont on peut obtenir une estimation assez précise des effets d'un changement de politique même si nous ne pouvons pas expliquer la plupart des variations de la variable dépendante. Les variables explicatives de (13.12) expliquent seulement 2,1 % des variations de log(*durat*). Cela est logique : il y a évidemment beaucoup d'autres facteurs affectant la durée pendant laquelle une personne sera indemnisée, tels que la gravité de l'accident ou la sévérité des blessures. Heureusement, nous avons un échantillon de grande taille, ce qui nous permet d'obtenir une statistique t significative.

MVD ont aussi pris en compte d'autres facteurs comme le genre, le statut matrimonial, l'âge, l'industrie et le type de blessure. Ils tiennent compte ainsi du fait que les types de personnes et de blessures étaient peut-être systématiquement différents entre les groupes de revenus différents au cours de ces deux ans. Le fait de prendre en compte l'influence de ces facteurs n'a qu'un petit effet sur l'estimation de δ_1. (Voir l'exercice sur ordinateur C4).

Parfois, les deux groupes sont formés de personnes vivant dans deux États voisins aux États-Unis. Par exemple, pour mesurer l'effet d'un changement de la taxation des cigarettes sur leur consommation, nous pouvons échantillonner aléatoirement des individus dans deux États voisins pendant deux années. Dans l'État A, le groupe de contrôle, il n'y a pas eu de changement de taxation sur les cigarettes. Dans l'État B, le groupe de traitement, la taxe a augmenté (ou diminué) entre les deux années. La variable dépendante est la mesure de la consommation de cigarettes, et on peut estimer l'équation (13.10) pour déterminer les effets de la taxe sur la consommation de cigarettes.

Pour une enquête intéressante sur la méthodologie des expériences naturelles avec des exemples supplémentaires voir l'étude de Meyer (1995).

Pour aller plus loin 13.2

Comment interprétez-vous le coefficient et la statistique t associés à *highearn* dans l'équation (13.12) ?

13.3 ANALYSER DES DONNÉES DE PANEL SUR DEUX PÉRIODES

Nous analysons maintenant le type de données de panel le plus simple : pour une coupe transversale sur des individus, des écoles, des entreprises, des villes ou autres, nous disposons de données sur deux ans ; notons les t = 1 et t = 2, avec t = 1 correspondant à la plus ancienne. Le fichier CRIME2 contient des données sur la criminalité et le taux de chômage de 46 villes en 1982 et en 1987. Ici, t = 1 correspond à 1982 et t = 2 à 1987.

Que se passe-t-il si nous utilisons la coupe transversale de 1987 et que nous faisons une régression simple de *crmrte* sur *unem* ? Nous obtenons :

$$\widehat{crmrte} = 128{,}38 - 4{,}16\,unem$$
$$(20{,}76)\ \ (3{,}42)$$

$$n = 46,\ R^2 = 0{,}033.$$

Si l'on interprète de façon causale l'équation estimée, les résultats impliquent qu'une augmentation du taux de chômage fait baisser le taux de criminalité. Ce n'est peut-être pas ce à quoi nous pouvions nous attendre. Le coefficient de *unem* n'est pas statistiquement significatif : nous n'avons pas identifié de lien entre les taux de criminalité et de chômage.

De nombreuses variables sont omises de cette équation de régression simple, ce qui risque de biaiser nos estimations, comme nous l'avons vu précédemment. Une solution, pour résoudre le problème de variables omises, serait de prendre en compte un plus grand nombre de facteurs, comme la répartition de la population par âge et par sexe, son niveau d'études, les efforts de mise en application de la loi, etc. Néanmoins, cela peut s'avérer compliqué si les informations pertinentes sont difficiles à collecter ou à mesurer. Il y a une autre manière de traiter le problème : on peut considérer que les antécédents d'une ville en matière de criminalité donnent une information utile. Nous avions vus dans le chapitre 9 qu'en incluant le taux de criminalité, *crmrte*, de l'année antérieure – dans ce cas, 1982 – nous pouvions prendre en compte des tendances historiques, expliquant notamment que des villes différentes aient des taux de criminalité différents. Des données de panel peuvent pallier le problème de variable omise qui biaiserait une estimation en coupe transversale, et permettre d'évaluer un lien de causalité.

Les données de panel s'utilisent également d'une autre manière. Nous pouvons regrouper les facteurs non observés influençant les variables dépendantes en deux catégories : ceux qui sont constants et ceux qui varient au cours du temps. Si nous posons i l'unité d'observation de la coupe transversale et t la période, nous pouvons écrire un modèle avec une variable explicative observée comme suit :

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + \alpha_i + u_{it}, \; t = 1,2. \qquad \textbf{[13.13]}$$

Dans la notation y_{it}, i indique la personne, l'entreprise, la ville, etc. et t la période. La variable $d2_t$ est une variable indicatrice qui vaut zéro quand $t = 1$ et qui vaut un quand $t = 2$; elle ne change pas selon les individus c'est pourquoi elle n'est pas indicée par i. Par conséquent, la constante est β_0 quand $t = 1$ et $\beta_0 + \delta_0$ quand $t = 2$. Comme dans l'utilisation des coupes transversales indépendantes empilées, il est souvent nécessaire que la constante puisse varier au cours du temps. Dans l'exemple de la criminalité, les taux de criminalité des villes américaines dépendent probablement fortement des tendances de long terme communes à l'ensemble du pays ; or ces tendances ont varié au cours des cinq années considérées.

La variable a_i englobe tous les facteurs non observés constants au cours du temps qui influencent y_{it}. (Le fait que a_i n'ait pas d'indice t nous indique qu'a_i ne change pas au cours du temps). De façon générale, a_i est appelé **effet non observé**, **effet inobservé**, ou encore **effet fixe**. La dernière dénomination nous aide à nous rappeler que a_i est fixe au cours du temps. Le modèle en (13.13) est donc appelé **modèle à effets non observés**, **modèle à effets inobservés**, ou encore **modèle à effets fixes**. Enfin, a_i peut également être désigné par le terme **hétérogénéité non observée** (*ou hétérogénéité inobservée*, *hétérogénéité individuelle*, *hétérogénéité des entreprises*, *hétérogénéité des villes*, *etc.*).

L'erreur u_{it} est souvent appelée **l'erreur idiosyncratique** ou erreur variable dans le temps, puisqu'elle représente les facteurs non observés qui changent au cours du temps et qui affectent y_{it}. Ces erreurs sont très semblables à celles des modèles de séries temporelles.

Pour expliquer le taux de criminalité en 1982 et en 1987, nous pouvons écrire un modèle à effets fixes de la manière suivante :

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + \alpha_i + u_{it}, \qquad \textbf{[13.14]}$$

où *d87* est la variable indicatrice pour 1987. Puisque i indique les différentes villes, on appelle a_i l'effet non observé de la ville ou l'effet fixe de la ville : il représente tous les facteurs qui influent sur le taux de criminalité de la ville et ne changent pas au cours du temps. Les caractéristiques géographiques, comme la localisation de la ville aux États-Unis, sont incluses dans a_i. Beaucoup d'autres

facteurs ne sont pas exactement constants, mais ils peuvent être considérés comme constants sur une période de cinq ans, comme par exemple les caractéristiques démographiques de la population (âge, origine ethnique, et niveau d'études). Des villes différentes peuvent aussi avoir leurs propres méthodes pour répertorier les crimes et délits, et les habitants des villes peuvent avoir différentes attitudes face à la criminalité ; ce sont généralement des changements lents. Pour des raisons historiques, les villes peuvent avoir des taux de criminalité différents. Les facteurs historiques seront englobés dans l'effet non observé a_i.

Comment pourrions-nous estimer le paramètre d'intérêt, β_1, avec des données de panel sur deux ans ? Nous pourrions simplement regrouper les deux ans et utiliser les MCO, à l'instar de ce qui a été fait dans la section 13.1. Cette méthode a deux inconvénients. Le plus problématique est qu'afin de regrouper les MCO pour produire un estimateur convergent[3] de β_1, nous devons supposer que l'effet non observé, a_i, n'est pas corrélé avec x_{it}. En effet, nous pouvons écrire (13.13) de la manière suivante :

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + v_{it}, \; t = 1,2, \qquad \textbf{[13.15]}$$

où $v_{it} = a_i + u_{it}$ est souvent appelé **l'erreur composée**. D'après ce que nous avons vu avec les MCO, nous devons faire l'hypothèse que v_{it} n'est pas corrélé avec x_{it}, où $t = 1$ ou 2, pour pouvoir estimer β_1 (et les autres paramètres du modèle). Cela est vrai qu'on utilise une coupe transversale simple ou qu'on empile les deux coupes transversales. Par conséquent, même si on suppose que l'erreur idiosyncratique u_{it} n'est pas corrélée avec x_{it}, les MCO sur données empilées seront biaisés et non convergents si a_i et x_{it} sont corrélés. Le biais affectant alors les MCO sur données empilées est parfois appelé **biais d'hétérogénéité**. En réalité, il ne s'agit que d'un biais induit par l'omission d'une variable constante au cours du temps.

Pour aller plus loin 13.3

Supposons que a_i, u_{i1}, et u_{i2} soient de moyenne nulle et non corrélés deux à deux. Montrez que $\text{Cov}(v_{i1}, v_{i2}) = \text{Var}(a_i)$, de sorte que les erreurs composées sont positivement corrélées dans le temps, excepté pour $a_i = 0$. Qu'est-ce que cela implique à propos des écarts-types estimés habituels issus de l'estimation par MCO sur données empilées ?

Pour illustrer ce qui précède, nous estimons (13.14) par MCO sur les données empilées contenues dans le fichier CRIME2. Puisqu'il y a 46 villes et pour chaque ville, des observations pendant deux années, il y a un total de 92 observations :

$$\widehat{crmrte} = 93,42 + 7,94\, d87 + 0,427\, unem$$
$$(12,74)\ (7,98)\qquad (1,188)$$
$$n = 92,\ R^2 = 0,012. \qquad \textbf{[13.16]}$$

(Quand on reporte l'équation considérée, on enlève généralement les indices i et t). Le coefficient de *unem* a une statistique t très petite et est positif. L'utilisation de MCO sur les données empilées réunissant les deux années n'a pas entraîné de modification considérable des résultats par rapport à l'utilisation d'une coupe transversale simple. Ce n'est pas surprenant puisque l'utilisation de MCO sur données empilées ne résout pas le problème des variables omises. (Les écarts-types estimés sont incorrects dans cette équation en raison de la corrélation sérielle décrite à la question 13.3, mais nous n'en tenons pas compte ici puisque nous ne nous focalisons pas sur les MCO sur données empilées).

3 NDT : Rappel : un estimateur « convergent » est un estimateur qui converge vers la vraie valeur du paramètre.

Dans la plupart des applications, le principal intérêt de la collecte de données de panel est qu'il peut y avoir de la corrélation entre l'effet non observé, a_i, et les variables explicatives. Par exemple, dans l'équation sur la criminalité, nous voulons tenir compte du fait que le taux de chômage peut être corrélé avec les caractéristiques urbaines non mesurées qui affectent le taux de criminalité – ces caractéristiques sont incluses dans a_i. Cela est en fait assez simple : puisque a_i est constant au cours du temps, nous pouvons l'éliminer en prenant la différence entre les deux années. Plus exactement, pour l'observation d'une coupe transversale i, nous pouvons écrire les deux équations pour chaque année comme suit :

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + \alpha_i + u_{i2} \ (t = 2)$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + \alpha_i + u_{i1} \ (t = 1).$$

Si nous soustrayons la seconde équation à la première, nous obtenons :

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1}),$$

ou bien

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i, \qquad \textbf{[13.17]}$$

où Δ indique le changement de $t = 1$ à $t = 2$. L'effet non observé, a_i, n'apparaît pas dans (13.17) : il a été éliminé par la différence entre les deux années. Par ailleurs, la constante de (13.17) est en fait le changement de constante quand on passe de $t = 1$ à $t = 2$.

L'équation (13.17), que nous appelons **l'équation en différences premières**, est très simple. C'est simplement une équation unique en coupe transversale, où chaque variable est différenciée au cours du temps. Nous pouvons analyser (13.17) en utilisant les méthodes que nous avons développées dans la partie 1, à condition que les hypothèses clés soient satisfaites. La plus importante parmi celles-ci est que Δu_i n'est pas corrélé avec Δx_i. Cette hypothèse est valide si l'erreur idiosyncratique pour chaque période t, u_{it}, n'est pas corrélée avec la variable explicative pour les deux périodes. Il s'agit d'une autre version de l'hypothèse d'exogénéité stricte que nous avons vue dans le chapitre 10 pour les modèles de séries temporelles. En particulier, cette hypothèse exclut le cas dans lequel x_{it} est la variable dépendante retardée, $y_{i,t-1}$. Contrairement au chapitre 10, la variable x_{it} peut être corrélée aux variables non observables qui sont constantes au cours du temps. L'estimateur des MCO de β_1 obtenu à partir de (13.17) s'appelle **l'estimateur en différences premières**.

Dans l'exemple de la criminalité, considérer que Δu_i et $\Delta unem_i$ ne sont pas corrélés peut être correct, mais cela peut aussi être faux. Par exemple, supposons que l'effort des forces de police (qui est compris dans l'erreur idiosyncratique) augmente plus dans les villes où le taux de chômage diminue. Cela peut causer une corrélation négative entre Δu_i et $\Delta unem_i$, ce qui entraînerait alors un biais dans l'estimateur des MCO. Naturellement, ce problème peut être partiellement surmonté en introduisant plus de facteurs dans l'équation. Nous en parlerons ultérieurement. Comme précédemment, il est possible que nous n'ayons pas tenu compte de suffisamment de facteurs variant dans le temps.

De plus, Δx_i doit varier d'un i à l'autre. Cette condition cruciale n'est plus respectée si la variable explicative ne change pas au cours du temps, et ce quelle que soit la coupe transversale observée, ou encore si elle varie de manière identique pour chaque observation. Cela n'est pas un problème dans le cas du taux de criminalité puisque le taux de chômage change au cours du temps pour quasiment toutes les villes. Mais si i représente un individu et si x_{it} est une variable indicatrice de genre, $\Delta x_i = 0$ pour tout i ; on ne peut alors clairement pas estimer (13.17) par les MCO. C'est tout à fait logique : puisqu'on permet que a_i soit corrélé avec x_{it}, on ne peut pas espérer séparer l'effet de a_i sur y_{it} de l'effet des autres variables qui ne changent pas avec le temps.

Enfin, la dernière hypothèse nécessaire est l'hypothèse d'homoscédasticité. Cette hypothèse est raisonnable dans beaucoup de cas. Lorsqu'elle ne l'est pas, nous savons comment tester et corriger

l'hétéroscédasticité en utilisant les méthodes du chapitre 8. Parfois, on peut supposer à raison que (13.17) remplit toutes les conditions du modèle linéaire classique. Les estimateurs des MCO ne seront pas biaisés et dans ce cas, l'inférence statistique sera correcte.

Si nous estimons (13.17) à partir de données sur le taux de criminalité, nous obtenons :

$$\widehat{\Delta crmrte} = 15{,}40 + 2{,}22\,\Delta unem$$
$$(4{,}70)\ \ (0{,}88)$$
$$n = 46,\ R^2 = 0{,}127 \qquad \textbf{[13.18]}$$

ce qui donne une relation statistiquement significative et positive entre les taux de criminalité et de chômage. Ainsi, prendre la différence entre les deux périodes de temps afin d'éliminer les effets constants au cours du temps se révèle crucial. La constante de (13.18) donne aussi une information précieuse. Quand $\Delta unem = 0$, nous prédisons que le taux de criminalité augmente de 15,40 crimes pour 1000 personnes. Cela montre que les taux de criminalité ont globalement augmenté aux États-Unis entre 1982 et 1987.

Bien que nous ne soyons pas partis du modèle à effets inobservés décrit en (13.13), nous pouvons saisir intuitivement ce qu'implique l'utilisation des différences au cours du temps. Plutôt que de modéliser une relation en coupe transversale – qu'il serait difficile d'interpréter « *ceteris paribus* » en raison du problème des variables omises, l'équation (13.17) montre explicitement comment les changements de la variable explicative au cours d'une période affectent le changement de y au cours de la même période. Néanmoins, il reste très utile de garder l'équation (13.13) à l'esprit : elle nous rappelle explicitement qu'il est possible d'évaluer l'effet de x_{it} sur y_{it} en gardant a_i fixe.

Prendre la différence entre deux années de données de panel est un moyen puissant de tenir compte des effets inobservés. Mais cela a un coût. Tout d'abord, les jeux de données de panel sont plus difficiles à collecter que les coupes transversales, et plus particulièrement lorsqu'il s'agit de données individuelles. Il faut recourir à des enquêtes de suivi, de façon à garder la trace de chacun des individus interrogés, et à pouvoir les retrouver lors de la deuxième vague d'enquête.

Il est souvent difficile de localiser des personnes lors d'une seconde enquête. Il est compliqué également de retrouver des entreprises, qui peuvent faire faillite ou fusionner avec d'autres entreprises. Les données de panel sont beaucoup plus faciles à collecter pour les écoles, les villes, les départements, les régions et les États.

Même si nous avons collecté un jeu de données de panel, la méthode de différenciation utilisée pour éliminer a_i présente un autre inconvénient : elle peut réduire considérablement la variation des variables explicatives. Même si x_{it} varie souvent de façon importante pour chaque t, il est possible que Δx_i ne varie pas beaucoup, ce qui rend l'estimation de (13.17) par MCO peu précise. Comme nous le savons depuis le chapitre 3, plus la variation de Δx_i est faible, plus l'écart-type estimé de $\hat{\beta}_1$ risque d'être élevé. Il est parfois possible de contrer cet effet en utilisant une grande coupe transversale. Par ailleurs, il est souvent préférable d'utiliser des changements sur une longue durée plutôt que de considérer les différences d'une année à l'autre. À titre d'exemple, considérons la question de l'estimation du rendement des années d'études, en utilisant cette fois des données de panel individuelles sur deux ans. Le modèle pour un individu i est :

$$\log(wage_{it}) = \beta_0 + \delta_0 d_{2t} + \beta_1 educ_{it} + \alpha_i + u_{it},\ t = 1, 2,$$

où a_i représente l'habileté inobservée (ou les capacités individuelles innées) – qui est probablement corrélée à $educ_{it}$. De nouveau, nous choisissons un modèle où la constante change au cours du temps pour rendre compte de l'évolution des gains de productivité (et de l'inflation si le salaire $wage_{it}$ est exprimé

en termes nominaux). Puisque, par définition, les capacités innées ne changent pas au cours du temps, les méthodes de données de panel semblent être appropriées pour évaluer le rendement de l'éducation. L'équation en différences premières est :

$$\Delta \log(wage_i) = \delta_0 + \beta_1 \Delta educ_i + \Delta u_i, \qquad [13.19]$$

et nous pouvons l'évaluer par la méthode des MCO. Le problème est que nous nous intéressons aux adultes salariés. Or, pour la plupart des personnes employées, le niveau d'études ne varie pas au cours du temps. Si $\Delta educ_i$ n'a une valeur différente de zéro que pour une faible proportion de notre échantillon, il sera difficile d'obtenir un estimateur précis de β_1 à partir de (13.19), à moins que l'on ait un échantillon de très grande taille. En théorie, l'utilisation d'une équation en différences premières pour évaluer le rendement du niveau d'études est une bonne idée, mais cela ne fonctionne pas très bien avec les jeux de données de panel habituellement disponibles.

L'ajout de plusieurs variables explicatives ne pose pas de difficultés. Nous commençons par le modèle à effets non observés suivant :

$$y_{it} = \beta_0 + \delta_0 d_{2t} + \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + b_k x_{itk} + \alpha_i + u_{it}, \qquad [13.20]$$

avec $t = 1$ et 2. Cette équation est plus simple qu'elle ne semble. Chaque variable explicative comporte trois indices. Le premier correspond au nombre d'observations de la coupe transversale, le second à la période alors que le troisième renvoie simplement au numéro de variable.

EXEMPLE 13.5
De l'arbitrage entre temps de sommeil et travail

Nous utilisons les deux années des données de panel contenues dans le fichier SLP75_81, à partir de Biddle et Hamermesh (1990), pour analyser la relation entre le sommeil et le travail. Dans l'exercice 3 du chapitre 3, nous avons seulement utilisé la coupe transversale de 1975. Les données de panel pour 1975 et 1981 concernent 239 individus, ce qui constitue un échantillon de plus petite taille que la coupe transversale de 1975 qui en comprend plus de 700. Pour analyser le temps de sommeil total (en minutes par semaines), on peut écrire le modèle à effets fixes suivant :

$$slpnap_{it} = \beta_0 + \delta_0 d81_t + \beta_1 totwrk_{it} + \beta_2 educ_{it} + b_3 marrit + \beta_4\, yngkid_{it} + \beta_5\, gdhlth_{it} + a_i + u_{it},\ t = 1, 2.$$

L'effet inobservé, a_i, est un *effet individuel non observé* ou un *effet fixe individuel*. Il est important de prendre en compte la corrélation de a_i avec $totwrk_{it}$ car les facteurs (notamment biologiques) qui font que les personnes dorment plus ou moins (inclus dans a_i) peuvent être corrélés au temps que les gens passent à travailler.

Certaines personnes ont simplement plus d'énergie et cela leur permet de dormir moins et travailler plus. La variable *educ* correspond au nombre d'années d'études, *marr* est une variable indicatrice qui vaut un si l'individu est marié, *yngkid* est une variable indicatrice indiquant la présence d'un enfant en bas âge, et *gdhlth* une variable indicatrice égale à un si la personne est en bonne santé. Notez que nous n'incluons ni le sexe ni l'origine ethnique (comme nous l'avions fait dans l'analyse en coupe transversale) : ceux-ci ne changent pas au cours du temps et sont donc compris dans le terme a_i. Nous nous intéressons avant tout à l'évaluation du paramètre β_1.

Considérer la différence entre les deux années revient à estimer l'équation suivante :

$$\Delta slpnap_i = \delta_0 + \beta_1 \Delta totwrk_i + \beta_2 \Delta educ_i + \beta_3 \Delta marr_i + \beta_4 \Delta yngkid_i + \beta_5 \Delta gdhlth_i + \Delta u_i.$$

En supposant que le changement dans l'erreur idiosyncratique, Δu_i, n'est corrélé à aucun des changements des variables explicatives, nous pouvons obtenir des estimateurs convergents à l'aide de la méthode des MCO. Cela nous donne :

$$\widehat{\Delta slpnap} = -92{,}63 - 0{,}227\,\Delta totwrk - 0{,}024\,\Delta educ$$
$$(45{,}87)\ (0{,}036)\qquad (48{,}759)$$
$$+ 104{,}21\,\Delta marr + 94{,}67\,\Delta yngkid + 87{,}58\,\Delta gdhlth$$
$$(92{,}86)\qquad (87{,}65)\qquad (76{,}60)$$
$$n = 239,\ R^2 = 0{,}150. \qquad \textbf{[13.21]}$$

Le coefficient de $\Delta totwrk$ nous indique qu'il existe un arbitrage entre le temps consacré au sommeil et au travail : en gardant les autres facteurs fixés, une heure de travail supplémentaire est associée à 0,227(60) = 13,62 minutes de sommeil en moins. La statistique $t(-6{,}31)$ est très significative. Aucune autre estimation, mis à part la constante, n'est statistiquement différente de zéro. Le test F pour la significativité jointe de toutes les variables sauf $\Delta totwrk$ nous donne une p-valeur = 0,49 ce qui signifie qu'elles sont conjointement non significatives quel que soit le niveau de significativité retenu, et qu'elles peuvent être retirées de l'équation.

L'écart-type estimé de $\Delta educ$ est particulièrement grand en comparaison avec la valeur du paramètre estimé. Il s'agit du phénomène décrit précédemment pour l'équation de salaire. Dans l'échantillon de 239 personnes, 183 (76,6 %) ne changent pas de niveau d'études sur l'ensemble de la période considérée ; pour 90 % des personnes, le nombre d'années d'études change d'une année au plus. Comme le montre l'écart-type estimé extrêmement grand de $\hat{\beta}_2$, il y a trop peu de variation dans le niveau d'études pour évaluer β_2 avec suffisamment de précision. En tout état de cause, $\hat{\beta}_2$ apparaît très petit.

Les données de panel peuvent aussi servir à estimer des modèles à retards échelonnés finis (c'est-à-dire des modèles avec des variables explicatives retardées). Si nous estimons une équation où la variable dépendante est observée pendant deux années, s'il y a des variables explicatives retardées, nous devrons collecter des données pour les années qui précèdent. L'exemple suivant permet de mieux le comprendre.

EXEMPLE 13.6
Retards échelonnés du taux de criminalité et taux de résolution des crimes

Eide (1994) utilise des données de panel venant des districts de police de Norvège afin d'évaluer un modèle à retards échelonnés du taux de criminalité. La variable explicative est le « pourcentage de résolution » (*clrprc*) – le pourcentage de crimes et délits ayant été élucidés. Le taux de criminalité est mesuré en 1972 et en 1978. D'après Eide, il semble que le taux de résolution des crimes passés a un effet dissuasif sur les crimes actuels. Nous considérons les valeurs de *clrprc* au cours des deux années qui précèdent. Cela nous donne le modèle à effets non observés suivant pour les deux années :

$$\log(crime_{it}) = \beta_0 + \delta_0 d78_t + \beta_1 clrprc_{i,t-1} + \beta_2 clrprc_{i,t-2} + a_i + u_{it}.$$

Lorsque nous considérons l'équation en différences et que nous l'estimons en utilisant les données de CRIME3, nous obtenons :

$$\Delta\widehat{\log(crime)} = 0{,}086 - 0{,}0040\,\Delta clrprc_{-1} - 0{,}0132\,\Delta clrprc_{-2}$$
$$(0{,}064)\ (0{,}0047)\qquad (0{,}0052)$$
$$n = 53,\ R^2 = 0{,}193,\ \bar{R}^2 = 0{,}161. \qquad \textbf{[13.22]}$$

Le second retard est négatif et statistiquement significatif, ce qui implique qu'un pourcentage plus grand de résolution de crimes deux ans plus tôt aurait un effet dissuasif sur les crimes actuels. En particulier, une augmentation de *clrprc* de 10 points de pourcentage supplémentaires conduirait à une baisse du taux de criminalité estimé de 13,2 % deux ans plus tard. Cela suggère que le fait d'utiliser plus de ressources pour résoudre les crimes et parvenir à condamner les coupables peut réduire la criminalité dans le futur.

Organisation des données de panel

Recourir aux données de panel dans les études économétriques requiert de savoir de quelle manière les données doivent être organisées. Nous devons faire attention à ce que les observations associées à une même entité (personne, entreprise, ville, etc.) puissent être facilement utilisées. Plus concrètement, supposons que le jeu de données concerne les villes sur deux années. Dans la plupart des cas, le meilleur moyen d'accéder aux données est d'avoir *deux* entrées pour chaque ville, une pour chaque année : la première entrée pour chaque ville correspond à la première année et la seconde entrée concerne l'année suivante. Ces deux entrées devraient être adjacentes. Par conséquent, un jeu de données pour 100 villes et deux années contiendra 200 entrées. Les deux premières entrées concernent la première ville dans l'échantillon, les deux entrées suivantes correspondent à la seconde ville, etc. (Le tableau 1.5 du chapitre 1 en donne un exemple.) Le fait d'organiser les données relatives aux villes sur deux entrées facilite le calcul des différences premières, qu'il est pratique d'enregistrer dans la deuxième entrée correspondant à la ville. De plus, les données peuvent faire l'objet d'une analyse de coupes transversales empilées, qui peut alors être comparée à l'estimation en différences premières.

La plupart des jeux de données de panel sur deux périodes qui accompagnent ce texte sont organisées de cette façon (par exemple, CRIME2, CRIME3, GPA3, LOWBRTH, et RENTAL). Nous utilisons une extension directe de ce schéma pour les jeux de données de panel de plus de deux périodes.

La deuxième manière d'organiser des données de panel sur deux périodes est d'avoir une entrée unique par unité d'observation de la coupe transversale. Cela nécessite d'entrer chaque variable deux fois, pour chacune des deux périodes. Les données de panel du fichier SLP75_81 sont organisées de cette façon. À chaque individu correspond une entrée pour les variables *slpnap75*, *slpnap81*, *totwrk75*, *totwrk81*, etc. On trouve donc sur la même ligne la valeur des variables mesurées en 1975 et la valeur des variables mesurées en 1981. Il est facile de créer les différences de 1975 à 1981. Les jeux de données de panel TRAFFIC1 et VOTE2 sont également structurés de cette façon.

Cependant, le fait de mettre les données dans une entrée ne nous permet pas de réaliser une analyse de MCO sur données empilées en utilisant les deux périodes de données originelles. De plus, cette méthode d'organisation ne fonctionne pas pour les jeux de données de panel qui contiennent plus de deux périodes, cas que nous étudierons dans la section 13.5.

13.4 ÉVALUER DES POLITIQUES PUBLIQUES À PARTIR DE DONNÉES DE PANEL SUR DEUX PÉRIODES

Les jeux de données de panel sont très utiles pour analyser une politique ou une réforme, et en particulier évaluer une mesure ou un programme particulier. Dans l'organisation la plus simple d'une évaluation de programme, on obtient un échantillon d'individus, d'entreprises, de villes ou autres dans un premier temps. Certaines de ces entités, celles qui sont dans le groupe de traitement, bénéficient d'un programme

spécifique dans une deuxième période ; les autres forment le groupe de contrôle. Cette approche est similaire à celle décrite dans la littérature relative aux expériences naturelles, discutée précédemment, avec une différence importante : les *mêmes* entités de la coupe transversale sont observées à chaque période.

Par exemple, supposons que l'on veuille estimer les effets d'un programme de formation professionnelle dans le Michigan sur la productivité des travailleurs d'entreprises manufacturières (voir aussi l'exercice sur ordinateur C3 dans le chapitre 9). Soit $scrap_{it}$ le taux de rebut d'une entreprise i pendant l'année t (le taux de rebut est le pourcentage d'objets qui doivent être jetés en raison de la présence de défauts). Soit $grant_{it}$ une variable binaire égale à un si l'entreprise i reçoit une subvention pour une formation durant l'année t. Pour les années 1987 et 1988, le modèle est

$$scrap_{it} = \beta_0 + \delta_0 y88_t + \beta_1 grant_{it} + a_i + u_{it}, \; t = 1, 2, \qquad \textbf{[13.23]}$$

où $y88_t$ est la variable indicatrice de l'année 1988 et a_i est l'effet inobservé de l'entreprise ou l'effet fixe de l'entreprise. L'effet inobservé comprend des facteurs tels que la capacité du salarié moyen, le capital et la compétence managériale ; ce sont des facteurs qui sont *grosso modo* constants au cours du temps sur une période de deux ans. Nous pouvons craindre que l'effet inobservé, a_i, soit corrélé au fait que l'entreprise reçoive une subvention. En effet, les administrateurs du programme doivent donner la priorité aux entreprises dont les employés ont des compétences moindres. Le problème opposé peut aussi apparaître : afin d'améliorer les résultats du programme de formation professionnelle (du moins en apparence), les administrateurs peuvent donner les subventions aux employeurs dont les salariés sont les plus productifs. Heureusement, dans ce programme particulier, les subventions étaient attribuées selon le principe du « premier arrivé-premier servi » : ce n'étaient pas les administrateurs qui choisissaient les entreprises bénéficiaires. Cela étant, le fait qu'une entreprise postule plus tôt qu'une autre peut être corrélé avec la productivité de ses travailleurs. Dans ce cas, une analyse se basant sur une coupe transversale simple ou sur des coupes transversales empilées produirait des estimateurs biaisés et non convergents.

Si nous transformons l'équation pour l'écrire en différences premières et enlever a_i nous obtenons :

$$\Delta scrap_i = \delta_0 + \beta_1 \Delta grant_i + \Delta u_i. \qquad \textbf{[13.24]}$$

Ainsi, nous faisons simplement une régression de la variation du taux de rebut sur la variation de l'indicateur de subvention. Puisqu'aucune entreprise n'a reçu de subvention en 1987, pour tout i, $grant_{i1} = 0$, et donc, $\Delta grant_i = grant_{i2} - grant_{i1} = grant_{i2}$, ce qui indique simplement si l'entreprise a reçu une subvention en 1988. Cependant, en règle générale il est important de différencier toutes les variables (y compris les variables indicatrices) puisque cela sert à retirer a_i du modèle à effets inobservés (13.23).

L'estimation de l'équation en différences premières à partir des données du fichier JTRAIN donne :

$$\widehat{\Delta scrap} = -0{,}564 - 0{,}739\,\Delta grant$$
$$(0{,}405)\;\; (0{,}683)$$

$$n = 54, \; R^2 = 0{,}022.$$

Par conséquent, nous estimons que le fait d'avoir obtenu une subvention pour une formation professionnelle fait baisser le taux de rebut en moyenne de −0,739. Mais cette estimation n'est pas statistiquement différente de zéro.

Nous obtenons des résultats plus solides en utilisant log(*scrap*) et en estimant l'effet en pourcentage :

$$\Delta\widehat{\log(scrap)} = -0{,}057 - 0{,}317\,\Delta grant$$
$$(0{,}097)\;\; (0{,}164)$$

$$n = 54, \; R^2 = 0{,}067.$$

D'après cette estimation, le fait de recevoir une subvention pour une formation professionnelle fait baisser le taux de rebut d'environ 27,2 %. [exp(−0,317) − 1 ≈ −0,272]. La statistique t est d'environ −1,93, ce qui est faiblement significatif. En utilisant les données empilées, la régression de log(*scrap*) sur *y88* et *grant* donne $\hat{\beta}_1 = 0,057$ (écart-type = 0,431). Ainsi, nous ne trouvons pas de relation significative entre les taux de rebut et les subventions obtenues pour la formation professionnelle. Puisque ces résultats apparaissent très différents de ceux issus des estimations faites à partir du modèle en différences premières, cela suggère que les entreprises qui ont les employés aux capacités les plus faibles sont plus susceptibles de recevoir une subvention.

Nous allons à présent étudier le modèle d'évaluation de programme de manière plus générale. Soit y_{it} la variable dépendante, et $prog_{it}$ la variable explicative relative à la participation au programme.

Le modèle à effets non observés le plus simple est donné par :

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 prog_{it} + a_i + u_{it}. \qquad \textbf{[13.25]}$$

Si la participation au programme n'a eu lieu qu'à la seconde période, alors l'estimateur MCO de β_1 dans l'équation en différences a une représentation très simple :

$$\hat{\beta}_1 = \overline{\Delta y}_{traitement} - \overline{\Delta y}_{contrôle}. \qquad \textbf{[13.26]}$$

En d'autres termes, nous calculons le changement moyen de y entre les deux périodes pour les groupes de traitement et de contrôle. $\hat{\beta}_1$ correspond à la différence entre les deux. Il s'agit de la version pour données de panel de l'estimateur de différence des différences qui avait déjà été évoqué dans l'équation (13.11) pour deux coupes transversales empilées. Avec les données de panel, nous avons un avantage important : nous pouvons faire la différence de y au cours du temps pour les *mêmes* entités de la coupe transversale. Cela nous permet de prendre en compte les effets spécifiques de la personne, de l'entreprise ou de la ville, comme le montre le modèle présenté à l'équation (13.25).

Si la participation au programme a lieu durant les deux périodes, $\hat{\beta}_1$ ne peut pas s'écrire comme dans (13.26), mais on l'interprètera de la même manière : c'est le changement de la valeur moyenne de y lié à la participation au programme.

Le fait de tenir compte des facteurs variables au cours du temps ne change rien à la significativité. Nous considérons simplement leur différence entre une période et la suivante et nous les incluons avec $\Delta prog$. Cela nous permet de tenir compte des variables variant au cours du temps qui pourraient être corrélées avec les caractéristiques du programme.

Cette méthode en différences premières fonctionne aussi pour analyser les effets d'une politique qui varie en fonction de la ville ou de l'État. Nous présentons ci-dessous un exemple simple de ce cas.

EXEMPLE 13.7
De l'effet des lois sur la conduite en état d'ébriété sur les accidents de la route

De nombreux États des États-Unis pratiquent différentes politiques pour lutter contre la conduite en état d'ivresse. Nous allons étudier deux types de lois : les lois concernant les récipients ouverts – qui interdisent aux passagers la possession de récipients ouverts contenant des boissons alcoolisées – et les lois de suspension immédiate du permis – qui permettent aux tribunaux de suspendre le permis de conduire d'un conducteur en état d'ébriété avant qu'il ne soit jugé. Pour analyser leur impact, on peut utiliser une seule coupe transversale observée au niveau des États, et régresser le nombre d'accidents de la route (ou de conduites en états d'ivresse) sur des variables binaires indiquant les lois existant dans chaque État. Néanmoins, cela risque ne pas fonctionner puisque chaque État décide, au travers de processus législatifs, s'il a besoin de telles lois ou non. Par conséquent, la présence de ces lois dépend probablement du nombre moyen d'accidents de la route liés à l'alcool au cours des dernières années.

L'analyse sera plus convaincante si elle s'appuie sur des données de panel couvrant une période durant laquelle certains États ont adopté de nouvelles lois (tandis que d'autres ont abrogé des lois existantes). Le fichier TRAFFIC1 contient des données de 1985 et 1990 pour les 50 États américains et pour le district de Columbia. La variable dépendante est le nombre de morts par accident de la route pour 100 millions de miles parcourus en voiture (*dthrte*). En 1985, 19 États disposaient de lois sur les récipients ouverts contre 22 en 1990. En 1985, 21 États disposaient de lois de suspension du permis contre 29 en 1990.

L'estimation par les MCO du modèle en différences premières nous donne :

$$\widehat{\Delta dthrte} = -0,497 - 0,420\,\Delta open - 0,151\,\Delta admn$$
$$(0,052)\ (0,206)\qquad (0,117)$$
$$n = 51,\ R^2 = 0,119. \qquad \textbf{[13.27]}$$

Les résultats suggèrent ici que l'adoption d'une loi sur les récipients ouverts a fait diminuer le taux d'accidents de la route de 0,42. C'est un effet non négligeable, sachant que le taux moyen de décès sur la route en 1985 était de 2,7 avec un écart-type estimé d'environ 0,6.

L'estimation est statistiquement significative pour un niveau de 5 % contre une alternative bilatérale. La loi de suspension du permis a un effet moindre et sa statistique t est de seulement $-1,29$; mais le signe de cette estimation est celui que nous attendions.

La constante de cette équation montre que les accidents de la route diminuent considérablement dans tous les États sur une période de cinq ans, qu'il y ait eu ou non un changement de loi. Les États qui ont adopté une loi sur les récipients ouverts au cours de cette période ont en moyenne connu une baisse encore plus importante des taux d'accidents de la route.

D'autres lois ont aussi pu avoir un impact sur le taux d'accidents de la route, comme des lois sur le port de la ceinture de sécurité, sur le port des casques de moto, et sur les limitations de vitesses. De plus, nous pourrions aussi prendre en compte la distribution de certaines caractéristiques démographiques (comme la pyramide des âges ou la répartition hommes/femmes), ou mesurer l'influence d'une organisation telle que *Mothers Against Drunk Driving* (« les Mères Contre la Conduite en État d'Ivresse ») dans chaque État.

Pour aller plus loin 13.4

Dans l'exemple 13.7, $\Delta admn = -1$ pour l'État de Washington. Expliquez ce que cela signifie.

13.5 DIFFÉRENCIER LES VARIABLES SUR PLUS DE DEUX PÉRIODES

Il est également possible de différencier les variables sur plus de deux périodes. Pour l'illustrer, nous prenons N individus et $T = 3$ périodes pour chaque personne. Le modèle à effets fixes s'écrit de la manière suivante :

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \qquad \textbf{[13.28]}$$

pour $t = 1$, 2, et 3. (Le nombre total d'observations est alors $3N$). Vous noterez que nous incluons désormais deux variables indicatrices pour deux de nos périodes en plus de la constante. C'est une bonne idée d'introduire une constante distincte pour chaque période, plus particulièrement quand nous en avons peu. La période de référence est comme d'habitude $t = 1$. La constante pour la deuxième période est

$\delta_1 + \delta_2$, etc. Nous nous intéressons avant tout à β_1, β_2, ..., β_k. Si l'effet non observé a_i est corrélé avec une des variables explicatives, alors l'utilisation des MCO sur les trois ans de données empilées donnera des estimations biaisées et non convergentes.

L'hypothèse clé est que les erreurs idiosyncratiques ne sont pas corrélées avec la variable explicative aux différentes dates, soit :

$$\text{Cov}(x_{itj}, u_{is}) = 0, \text{ pour tout } t, s, \text{ et } j. \qquad \textbf{[13.29]}$$

Cela signifie que les variables explicatives sont strictement exogènes une fois que l'on a retiré l'effet inobservé, a_i. (Vous trouverez dans l'annexe de ce chapitre la définition de l'hypothèse d'exogénéité stricte en termes d'espérance conditionnelle nulle.)

L'hypothèse (13.29) écarte les cas pour lesquels les futures variables explicatives réagissent aux changements des erreurs idiosyncratiques, comme ce doit être le cas si x_{itj} est une variable dépendante retardée. Si nous omettons une variable importante variant au cours du temps, alors (13.29) ne sera généralement pas vérifiée. Si une (ou plusieurs) variable(s) explicative(s) sont mal mesurées, (13.29) peut ne pas être vérifiée, comme nous l'avons vu dans le chapitre 9. Dans les chapitres 15 et 16 nous discuterons de ce que l'on peut entreprendre dans de tels cas.

Si a_i est corrélé avec x_{itj}, alors x_{itj} sera corrélé avec l'erreur composée, $v_{it} = a_i + u_{it}$, dans (13.29). Nous pouvons éliminer a_i en considérant la différence entre les périodes adjacentes. Dans le cas de $T = 3$, nous soustrayons la première période de temps à la seconde et la seconde à la troisième. Cela nous donne :

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk} + \Delta u_{it}, \qquad \textbf{[13.30]}$$

pour $t = 2$ et 3. Nous n'avons pas d'équation en différences pour $t = 1$ puisqu'il n'y a rien à soustraire de l'équation quand $t = 1$. Désormais, (13.30) représente deux périodes pour chaque individu de l'échantillon. Si cette équation satisfait les hypothèses du modèle linéaire classique, alors les MCO sur données empilées donnent des estimateurs non biaisés, et les statistiques usuelles t et F sont valides. On peut aussi avoir recours aux résultats asymptotiques. La condition importante pour que les MCO soient convergents est que Δu_{it} ne soit pas corrélé avec Δx_{itj}, pour tout j et pour $t = 2$ ou 3.

Remarquez que l'équation (13.30) comprend les variables indicatrices temporelles en différences, $d2_t$ et $d3_t$, et qu'elle ne comporte pas de constante. Cela peut poser quelques problèmes, notamment pour le calcul du R carré. À moins que les constantes associées à des périodes différentes dans le modèle original (13.28) n'aient un intérêt direct – ce qui est rarement le cas – il est préférable d'estimer l'équation en différences premières avec une constante et une variable indicatrice pour une seule période (habituellement pour la troisième). L'équation devient alors :

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk}$$

Les estimations de β_j sont identiques dans les deux formulations.

L'approche est similaire lorsqu'il y a plus de trois périodes. Si nous disposons du même nombre T de périodes pour chaque entité parmi les N unités d'observations de la coupe transversale, nous pouvons dire que le jeu de données est un **panel cylindré** : nous disposons du même nombre d'observations à chaque période pour tous les individus, entreprises, villes, etc.

Quand T est petit par rapport à N, nous devons inclure une variable indicatrice pour chaque période de temps afin de rendre compte des changements qui n'ont pas été modélisés. Par conséquent, après une première différence, l'équation ressemble à :

$$\begin{aligned}\Delta y_{it} = {} & \alpha_0 + \alpha_3 d3_t + \alpha_4 d4_t + \ldots + \alpha_T dT_t + \beta_1 \Delta x_{it1} + \ldots \\ & + \beta_k \Delta x_{itk} + \Delta u_{it}, \; t = 2, 3, \ldots, T, \end{aligned} \qquad \textbf{[13.31]}$$

Nous avons alors $T - 1$ périodes pour chaque entité i dans cette équation en différences premières. Le nombre total d'observations est donné par $N(T - 1)$.

Il est facile d'estimer (13.31) par MCO sur données empilées, à condition que les observations aient été organisées proprement et que la différence ait été réalisée avec soin. Pour faciliter le calcul des différences premières, le fichier de données doit être composé de NT entrées. Les T premières entrées correspondent aux observations de la première unité (de la coupe transversale) rangées dans l'ordre chronologique ; les T entrées suivantes correspondent aux observations de la deuxième unité rangées dans l'ordre chronologique ; etc. Ensuite, on calcule les différences, et on encode le changement de $t - 1$ à t dans l'entrée correspondant à la t-ième période. Par conséquent, pour $t = 1$, la variable en différence devrait comporter des valeurs manquantes pour les N observations de la coupe transversale. Si vous ne faites pas cela, vous prenez le risque d'utiliser des observations fausses dans l'analyse par régression. On crée une observation non valide lorsque la dernière observation d'un individu, disons $i - 1$, est soustraite de la première observation pour l'individu i. Si vous avez fait la régression sur ces données en différences, et que NT ou $NT - 1$ observations sont reportées dans vos résultats de régression, alors vous avez probablement oublié d'assigner aux observations en $t = 1$ des valeurs manquantes.

Quand on utilise plus de deux périodes de temps, on doit supposer que Δu_{it} est non corrélé au cours du temps de sorte que les écarts-types estimés et les tests statistiques habituels soient valides. Cette hypothèse est parfois raisonnable, mais elle ne l'est plus si on suppose que les erreurs idiosyncratiques originelles, u_{it}, ne sont pas corrélées au cours du temps (nous utiliserons cette hypothèse dans le chapitre 14).

En effet, si nous supposons que les u_{it} ne sont pas autocorrélés dans le temps et sont de variance constante, alors on peut montrer que la corrélation entre Δu_{it} et $\Delta u_{i,t+1}$ est $-0,5$. Plus généralement, si u_{it} suit un modèle AR(1) stable, alors Δu_{it} est autocorrélé. C'est seulement lorsqu'u_{it} est supposé suivre un processus de marche aléatoire que Δu_{it} ne sera pas autocorrélé.

Il est facile de tester l'autocorrélation dans l'équation en différences premières. Soit $r_{it} = \Delta u_{it}$ l'erreur originelle en différences premières. Si r_{it} suit le modèle AR(1) $r_{it} = \rho r_{i,t-1} + e_{it}$, alors on peut facilement tester H_0 : $\rho = 0$. Nous estimons (13.31) par la méthode des MCO sur données empilées et nous obtenons les résidus $\hat{r}_{it}$.

Ensuite, nous faisons une régression simple par MCO sur données empilées de $\hat{r}_{it}$ sur $\hat{r}_{i,t-1}$, $t = 3, \ldots, T$, $i = 1, \ldots, N$, et nous calculons un test t standard pour le coefficient de $\hat{r}_{i,t-1}$. (Nous pouvons également calculer la statistique t robuste à l'hétéroscédasticité). Le coefficient $\hat{\rho}$ associé à $\hat{r}_{i,t-1}$ est un estimateur convergent de ρ. Puisque nous utilisons les résidus retardés, nous perdons une période. Par exemple, si nous avons $T = 3$, nous disposerons de deux périodes pour l'équation en différences, et le test d'autocorrélation sera simplement une régression en coupe transversale des résidus de la troisième période sur les résidus de la seconde période. Nous en donnons un exemple plus loin.

Nous pouvons apporter une correction à la présence de corrélation sérielle AR(1) dans r_{it} en utilisant la méthode des Moindres Carrés Quasi-Généralisés (MCQG). Idéalement, dans chaque observation en coupe transversale, nous devrions utiliser la transformation de Prais-Winsten basée sur $\hat{\rho}$ et décrite dans la section précédente. (Ici nous préférons clairement la transformation de Prais-Winsten à celle de Cochrane-Orcutt, puisque l'élimination de la première période signifierait la perte de N observations en coupe transversale). Malheureusement, les modules standard qui permettent d'apporter des corrections de type AR(1) pour les régressions de séries temporelles ne fonctionnent pas. Les méthodes standard de Prais-Winsten traitent les observations comme si elles suivaient un processus AR(1) en i et en t, ce qui n'a pas de sens, puisque nous supposons que les observations sont indépendantes d'un individu i à l'autre. Les corrections apportées aux écarts-types des MCO dans le cas d'un modèle autorisant la présence de corrélation sérielle de forme arbitraire (et d'hétéroscédasticité) peuvent être calculées lorsque N est grand (et N devrait être très supérieur à T). Traiter de manière détaillée des écarts-types et des tests statistiques qui seraient robustes à toutes les formes de corrélation sérielle et d'hétéroscédasticité dépasse le cadre de

cet ouvrage ; le lecteur intéressé pourra consulter à profit Wooldridge (2010, chapitre 10). Néanmoins, de telles statistiques sont faciles à calculer dans de nombreux modules de logiciels d'économétrie, et l'annexe vous en donnera une présentation intuitive.

Pour aller plus loin 13.5

De la corrélation sérielle en Δu_{it} conduirait-elle l'estimateur en différences premières à être biaisé et non convergent ? En quoi la corrélation sérielle est-elle un problème ?

En l'absence de corrélation sérielle dans les erreurs, les méthodes habituelles pour traiter l'hétéroscédasticité sont valides. Nous pouvons utiliser les tests d'hétéroscédasticité de Breusch-Pagan et de White du chapitre 8, et nous pouvons aussi calculer les écarts-types estimés robustes.

Dans l'exemple suivant, nous voyons pourquoi considérer les différences premières de données de panel disponibles pour plus de deux ans est particulièrement utile dans le cadre de l'évaluation des politiques publiques.

EXEMPLE 13.8
Des effets des zones d'activités sur le chômage

Papke (1994) a étudié les effets d'un programme de création de « zone d'entreprises[4] » (EZ, pour « enterprise zone ») dans l'État de l'Indiana aux États-Unis sur le nombre de demandes d'allocations chômage (appelées « unemployement claims » aux États-Unis). Les « zones d'entreprises » sont des endroits où le gouvernement apporte des subventions aux entreprises qui s'installent). Elle a analysé 22 villes de l'Indiana sur la période allant de 1980 à 1988. Six zones d'entreprises ont été désignées en 1984, et quatre de plus ont été assignées en 1985. Dans l'échantillon, douze villes n'ont pas accueilli de zone d'entreprises sur cette période ; elles constituent le groupe de contrôle.

On peut considérer le modèle simple d'évaluation de politique suivant :

$$\log(uclms_{it}) = \theta_t + \beta_1 ez_{it} + a_i + u_{it},$$

où $uclms_{it}$ est le nombre de demandes d'allocations chômage faites pendant l'année t dans la ville i. Le paramètre θ_t représente *simplement* une constante différente pour chaque période. Généralement, les demandes d'allocations chômage ont chuté dans tout l'État pendant cette période, et cela devrait être observé dans les constantes des différentes années. La variable binaire ez_{it} est égale à un si la ville i au temps t a une zone d'entreprises ; nous nous intéressons à β_1. L'effet non observé a_i représente les facteurs fixes ayant un impact sur le climat économique de la ville i. Puisque la désignation des zones d'entreprises n'a pas été déterminée de façon aléatoire – les zones d'entreprises sont habituellement des zones déprimées économiquement – il est probable que ez_{it} et a_i soient corrélés positivement (un a_i élevé signifiant plus de demandes d'allocations chômage, ce qui augmente les chances d'une ville de se voir attribuer une EZ). Ainsi, pour éliminer a_i, nous devrions écrire l'équation en différences suivante :

$$\Delta\log(uclms_{it}) = \alpha_0 + \alpha_1 d82_t + \ldots + \alpha_7 d88_t + \beta_1 \Delta ez_{it} + \Delta u_{it}. \qquad [13.32]$$

La variable dépendante dans cette équation, la variation de $\log(uclms_{it})$ correspond au taux de croissance annuel approximatif des demandes d'allocations chômage pour la période allant de l'année $t-1$ à t. Nous pouvons estimer cette équation pour les années allant de 1981 à 1988 en utilisant les données du fichier EZUNEM ; la taille totale de l'échantillon est de $22 \cdot 8 = 176$. L'estimation de β_1 est $\hat{\beta}_1 = -0{,}182$ (écart-type estimé = 0,078). Par conséquent, il semble que la présence d'une zone d'entreprise fasse baisser d'environ 16,6 % [$\exp(-0{,}182) - 1 \approx -0{,}166$] les demandes d'allocations chômage. Cet effet est important économiquement parlant et statistiquement significatif.

4 NDT : On peut considérer les « zones franches urbaines » en France comme un équivalent des « zones d'entreprises » américaines (elles en sont d'ailleurs inspirées).

Il n'y a pas de signe d'hétéroscédasticité dans l'équation : le test F de Breusch-Pagan donne $F = 0,85$, p-valeur = 0,557. Cependant, lorsqu'on ajoute les résidus des MCO retardés à l'équation en différences (et que l'on perd l'année 1981), on obtient $\hat{\rho} = -0,197$ ($t = -2,44$), ce qui témoigne d'une autocorrélation négative minime dans les erreurs en différences premières. Lorsque les erreurs sont corrélées négativement, les écarts-types estimés des MCO habituels ne tendent pas à minimiser outre mesure les écarts-types estimés, contrairement à ce qui se passe en présence d'autocorrélation positive (voir la section 12.1). Ainsi, la significativité d'une variable indicatrice de zone d'entreprises ne sera probablement pas affectée.

EXEMPLE 13.9
Les taux de criminalité par département (county) en Caroline du Nord

Cornwell et Trumbull (1994) ont utilisé les données de 90 départements (ou « county » aux États-Unis) de l'État de Caroline du Nord, pour les années allant de 1981 à 1987, pour estimer un modèle à effets inobservés de la criminalité ; les données sont contenues dans le fichier CRIME4. Ici, nous estimons une version plus simple de leur modèle, et nous considérons une équation en différences pour exclure l'effet non observé a_i. (Cornwell et Trumbull utilisent une transformation différente, que nous étudierons dans le chapitre 14). Plusieurs facteurs comprenant la localisation géographique, les attitudes face à la criminalité, les tendances historiques, et les conventions en matière de report[5] et de compilation des statistiques peuvent être englobés dans a_i. Le taux de criminalité correspond au nombre de crimes par personne, *prbarr* indique la probabilité estimée d'arrestation, *prbconv* la probabilité estimée de condamnation (pour une arrestation donnée), *prbpris* la probabilité de purger une peine de prison (pour une condamnation donnée), *avgsen* est la durée moyenne des peines prononcées et *polpc* le nombre d'officiers de police par habitants. Comme la plupart des études économétriques de la criminalité, nous utilisons les logs de toutes les variables pour estimer l'élasticité. Nous incluons également un jeu entier de variables indicatrices annuelles pour tenir compte de la tendance globale du taux de criminalité de l'État. Nous pouvons utiliser les années allant de 1982 à 1987 pour estimer l'équation en différences. Les nombres entre parenthèses sont les écarts-types estimés des MCO habituels ; les nombres entre crochets les écarts-types estimés robustes à l'hétéroscédasticité et à la corrélation sérielle :

$$\begin{aligned}\Delta\log(\widehat{crmrte}) &= 0,008 - 0,100\,d83 - 0,048\,d84 - 0,005\,d85\\ &\quad (0,017)\ (0,024)\ (0,024)\ (0,023)\\ &\quad [0,014]\ [0,022]\ [0,020]\ [0,025]\\ &\quad + 0,028\ d86 + 0,041\ d87 - 0,327\ \Delta\log(prbarr)\\ &\quad (0,024)\ (0,024)\ (0,030)\\ &\quad [0,021]\ [0,024]\ [0,056]\\ &\quad - 0,238\ \Delta\log(prbconv) - 0,165\ \Delta\log(prbpris)\\ &\quad (0,018)\ (0,026)\\ &\quad [0,040]\ [0,046]\\ &\quad - 0,022\ \Delta\log(avgsen) + 0,398\ \Delta\log(polpc)\\ &\quad (0,022)\ (0,027)\\ &\quad [0,026]\ [0,103]\end{aligned} \qquad \textbf{[13.33]}$$

$$n = 540,\ R^2 = 0,433,\ \overline{R}^2 = 0,422. \qquad \textbf{[13.33]}$$

Les trois variables de probabilité – d'arrestation, de condamnation et de purger une peine de prison – ont le signe attendu et sont statistiquement significatives. Par exemple, on prédit qu'une augmentation de 1 % de la probabilité d'être arrêté va faire diminuer le taux de criminalité d'environ 0,33 %. La variable associée à la sentence moyenne a un effet dissuasif léger, mais cet effet n'apparaît pas statistiquement significatif.

5 NDT : le taux de criminalité est sensible à la manière dont elle est reportée et mesurée.

Le coefficient de la variable du nombre de policier(s) par habitant est assez surprenant et est caractéristique de la plupart des études qui cherchent à expliquer les taux de criminalité. En l'interprétant de façon causale, cela signifie qu'une augmentation de 1 % du nombre de policier par habitant augmente le taux de criminalité d'environ 0,4 %. (La statistique t est très grande et vaut presque 15). Il est difficile de croire que la présence de plus de policiers fait augmenter la criminalité. Que se passe-t-il dans ce cas ? Il y a au moins deux possibilités. Tout d'abord, la variable du taux de criminalité est calculée à partir de crimes *signalés*. Donc il se peut que lorsqu'il y ait plus de policiers, plus de crimes soient signalés. Ensuite, la variable relative aux effectifs de police peut être endogène dans l'équation pour d'autres raisons : les départements[6] peuvent augmenter la taille des forces de police lorsqu'ils craignent que le taux de criminalité n'augmente. Dans ce cas, (13.33) ne peut pas être interprétée d'une façon causale. Dans les chapitres 15 et 16, nous étudierons les modèles et les méthodes d'estimation qui peuvent prendre en compte cette nouvelle forme d'endogénéité.

Le cas particulier du test de White pour l'hétéroscédasticité dans la section 8.3 donne $F = 75,48$ et une p-valeur = 0,0000, ce qui constitue donc une preuve solide d'hétéroscédasticité. (Techniquement, ce test n'est pas valide s'il y a de la corrélation sérielle, mais il est très indicatif. Le test d'autocorrélation AR(1) nous donne $\hat{\rho} = 2,233$, $t = 24,77$, donc il existe une autocorrélation négative. Le calcul des écarts-types estimés tient compte de l'autocorrélation et de l'hétéroscédasticité. (Nous n'en donnerons pas les détails ; les calculs sont semblables à ceux décrits dans la section 12.5 et sont réalisés par de nombreux logiciels d'économétrie. Voir Wooldridge (2010, chapitre 10) pour une discussion plus poussée). Aucune variable ne perd sa significativité statistique, mais les statistiques t associées aux variables dissuasives significatives deviennent beaucoup plus petites. Par exemple, la statistique t associée à la probabilité qu'il y ait condamnation (*prbconv*) s'échelonne de $-13,22$ (quand on utilise les écarts-types estimés) à $-6,10$ (quand on utilise les écarts-types robustes). De façon équivalente, les intervalles de confiance construits en utilisant les écarts-types estimés robustes devraient être plus larges que ceux basés sur les écarts-types estimés des MCO habituels.

On peut bien sûr appliquer le test de Chow sur les modèles de données de panel estimés par la méthode des différences premières. Comme dans le cas des coupes transversales empilées, nous voulons rarement tester si les constantes sont constantes au cours du temps ; pour plusieurs raisons, nous nous attendons à ce que les constantes soient différentes. Il est beaucoup plus intéressant de tester si les coefficients de pente ont varié au cours du temps, et nous pouvons facilement effectuer de tels tests en faisant interagir les variables explicatives qui nous intéressent avec les variables indicatrices temporelles. Il est intéressant de noter qu'alors que nous ne pouvons pas estimer les pentes des variables qui ne changent pas au cours du temps, nous pouvons tester si les effets marginaux des variables constantes au cours du temps ont changé au cours du temps. Pour illustrer cela, supposons que nous observons un échantillon aléatoire de personnes salariées pendant trois années – en 2000, 2002, et 2004. Nous spécifions le modèle suivant (avec, pour le log du salaire, *lwage*) :

$$lwage_{it} = \beta_0 + \delta_1 d02_t + \delta_2 d04_t + \beta_1 female_i + \gamma_1 d02_t female_i + \gamma_2 d04_t female_i + \mathbf{z}_{it}\lambda + a_i + u_{it},$$

où $\mathbf{z}_{it}\lambda$ désigne les autres variables explicatives incluses dans le modèle et leurs coefficients. Lorsque nous différencions le modèle, nous éliminons la constante pour l'année 2000, β_0, et l'écart des salaires entre les sexes en 2000, β_1. Cependant, le changement de $d02_t female_i$ est $(\Delta d02_t) female_i$, et n'est pas éliminé. Par conséquent, nous pouvons estimer comment l'écart des salaires a changé entre 2002 et 2004 relativement à 2000, et nous pouvons tester si $\gamma_1 = 0$, ou si $\gamma_2 = 0$, ou les deux.

Nous pourrions aussi nous demander si les avantages salariaux liés aux syndicats ont changé au cours du temps, auquel cas nous pourrions inclure dans le modèle $union_{it}$, $d02_t union_{it}$, et $d04_t union_{it}$. Les coefficients de toutes ces variables explicatives peuvent être estimés car on présume que $union_{it}$ varie au cours du temps.

6 NDT : « *counties* » en anglais.

Si on essaie d'estimer un modèle contenant des variables d'interaction en différenciant à la main, l'opération peut s'avérer délicate. Par exemple, dans l'équation précédente incluant le statut syndical, on doit simplement faire la différence des termes d'interaction, $d02_t union_{it}$ et $d04_t union_{it}$. On *ne peut pas* calculer la différence exacte sous la forme de $d02_t \Delta union_{it}$ et $d04_t \Delta union_{it}$, ni même remplacer $d02_t$ et $d04_t$ par leurs différences premières.

De façon générale, il est important de revenir au modèle de base et de se rappeler que la différenciation du modèle est utilisée pour éliminer a_i. Le plus facile est d'utiliser une commande qui permet d'analyser des données de panel en différences premières – cette commande est généralement présente dans les logiciels d'économétrie. (Nous verrons quelques unes des options possibles dans le chapitre 14.)

Les écueils potentiels des différences premières sur des données de panel

Dans cette section et dans les précédentes, nous avons postulé que la différenciation de nos données qui visait à éliminer l'effet inobservé constant dans le temps était une méthode valable pour identifier des effets causaux. Pour autant, le calcul des différences n'est pas exempt de difficultés. Nous avons déjà discuté des problèmes potentiels de cette méthode lorsque les variables explicatives clés ne varient que très peu au cours du temps (la méthode est inopérante lorsque les variables explicatives ne varient jamais au cours du temps). Malheureusement, même quand x_{itj} varie suffisamment dans le temps, l'estimation en différences premières (DP) peut présenter des biais importants. Nous avons déjà mentionné que l'exogénétié stricte des régresseurs était une hypothèse cruciale.

Malheureusement, comme montré dans Wooldridge (2010, section 11.1), le fait de disposer d'un plus grand nombre d'observations temporelles ne résout en général pas la non convergence de l'estimateur en DP lorsque les régresseurs ne sont pas strictement exogènes (par exemple, si $y_{i,t-1}$ est inclus dans x_{itj}).

Un autre inconvénient important de l'estimateur en DP est qu'il peut faire pire que les MCO sur données empilées si une des variables explicatives ou plus sont sujettes à des erreurs de mesure, comme dans le modèle classique des variables avec erreur de mesure discuté dans la section 9.3.

Différencier des régresseurs mal mesurés réduit leurs variations, ce qui est d'autant plus problématique que, du fait de l'erreur de mesure, ils seront corrélés avec le terme d'erreur en différences. Cela conduit à un biais numériquement significatif. Il peut être très difficile de résoudre ce problème. Voir la section 15.8 dans Wooldridge (2010, chapitre 11).

RÉSUMÉ

Nous avons étudié les méthodes permettant d'analyser des coupes transversales indépendantes empilées et des jeux de données de panel. On obtient des coupes transversales indépendantes lorsque différents échantillons aléatoires sont obtenus pour différentes périodes (habituellement des années). La méthode des MCO sur données empilées est la principale méthode d'estimation, et les procédures habituelles d'inférence statistique sont valables, y compris en présence d'hétéroscédasticité. (L'autocorrélation n'est pas un problème puisque les échantillons sont indépendants au cours du temps). En raison de la nature des séries temporelles, nous introduisons souvent plusieurs constantes. Nous pouvons aussi faire interagir les variables indicatrices temporelles avec certaines variables clés pour voir comment elles ont changé au cours du temps. Cela est particulièrement important pour les expériences naturelles dans la littérature relative à l'évaluation de politiques publiques.

Les jeux de données de panel sont de plus en plus utilisés dans les études appliquées, et particulièrement dans l'analyse des politiques. Ce sont des jeux de données dans lesquels les unités individuelles sont suivies au cours du temps. Les jeux de données en panel sont plus utiles pour prendre en compte des caractéristiques non observées invariantes dans le temps – relatives aux individus, entreprises, villes, etc. – que nous supposons être corrélées avec les variables explicatives dans notre modèle. Une manière d'éliminer l'effet inobservé est de différencier des données sur des périodes de temps conjointes. Ensuite, on peut analyser les différences par MCO. L'utilisation de données sur deux périodes aboutit donc à une régression en coupe transversale des données en différences. Les procédures d'inférence statistique habituelles sont asymptotiquement valides sous l'hypothèse d'homoscédasticité ; l'inférence exacte est possible sous l'hypothèse de normalité.

Pour plus de deux périodes, nous pouvons utiliser les MCO sur données empilées différenciées ; nous éliminons alors la première période en différenciant. En plus de l'homoscédasticité, nous devons faire l'hypothèse que les erreurs en différences ne sont pas autocorrélées afin de pouvoir utiliser les statistiques habituelles t et F. (L'annexe du chapitre contient une liste d'hypothèses auxquelles il convient de faire attention). Évidemment, toute variable constante au cours du temps est éliminée de l'analyse.

MOTS-CLÉS

Biais d'hétérogénéité p. 539
Coupes transversales indépendantes empilées p. 526
Données de panel p. 526
Données longitudinales p. 526
Effet fixe p. 538
Effet moyen du traitement p. 536
Effet non observé ou inobservé p. 538
Équation en différences premières p. 540
Erreur composée p. 539
Erreur idiosyncratique p. 538
Estimateur en différences premières p. 540
Estimateur en double-différences p. 533
Exogénéité stricte p. 540
Expérience naturelle p. 535
Hétérogénéité inobservée p. 538
Hétérogénéité non observée p. 538
Modèle à effets fixes p. 538
Modèle à effets non observés p. 538
Panel cylindré p. 548
Quasi-expérience p. 535
Regroupement (en grappes) p. 564
Variables indicatrices temporelles p. 531

EXERCICES

1. Dans l'exemple 13.1, nous supposons que les moyennes de tous les facteurs autres que *educ* sont restées constantes au cours du temps et que le niveau moyen d'années d'études est de 12,2 années pour l'échantillon de 1972 et de 13,3 années pour celui de 1984. En utilisant les estimations du tableau 13.1, calculez le changement estimé de la fertilité moyenne entre 1972 et 1984. (Assurez-vous de rendre compte du changement de la constante et de la variation du nombre moyen d'années d'études).

2. En utilisant les données du fichier KIELMC, les équations suivantes ont été estimées pour les années 1978 et 1981 :

$$\widehat{log(price)} = 11{,}9 - 0{,}547\,nearinc + 0{,}394\,y81\cdot nearinc$$
$$(0{,}26)\ (0{,}058)\qquad (0{,}080)$$
$$n = 321,\ R^2 = 0{,}220$$

et

$$\widehat{log(price)} = 11{,}18 + 0{,}563\,y81 - 0{,}403\,y81\cdot nearinc$$
$$(0{,}27)\ (0{,}044)\qquad (0{,}067)$$
$$n = 321,\ R^2 = 0{,}337.$$

Comparez les estimations du coefficient du terme d'interaction *y81·nearinc* avec celles de l'équation (13.9). Pourquoi ces estimations sont-elles si différentes ?

3. Pourquoi ne pouvons-nous pas utiliser les différences premières lorsque nous avons deux années de coupes transversales indépendantes (par opposition à des données de panel) ?

4. Si nous pensons que β_1 est positif dans (13.14) et que Δu_i et $\Delta unem_i$ sont négativement corrélés, quel est le biais de l'estimateur des MCO de β_1 dans l'équation du modèle en différences premières ? [*Indice* : revoyez l'équation (5.4)].

5. Supposons que nous voulions estimer l'effet de plusieurs variables sur l'épargne annuelle et que nous disposions d'un jeu de données individuelles en panel collectées le 31 janvier 1990 et le 31 janvier 1992. Si nous incluons une variable indicatrice pour 1992 et que nous calculons la différence première, est-il possible d'inclure également l'âge dans le modèle original ? Expliquez.

6. En 1985, ni la Floride ni la Géorgie ne disposaient de loi réprimant la possession de boissons alcoolisées ouvertes dans l'habitacle d'un véhicule automobile. En 1990, la Floride a voté une loi de ce type mais pas la Géorgie.

i. Supposons que vous puissiez collecter des échantillons aléatoires pour la population en âge de conduire dans les deux États, pour 1985 et 1990. Définissons *arrest* comme étant une variable binaire égale à un si une personne a été arrêtée pour conduite en état d'ébriété au cours de l'année. Sans prendre en compte aucun autre facteur, écrivez un modèle de probabilité linéaire permettant de tester si le vote de la loi de répression des boissons alcoolisées ouvertes a réduit la probabilité d'être arrêté pour conduite en état d'ébriété. Quel coefficient de votre modèle permet de mesurer l'effet de cette loi ?

ii. Pourquoi voudriez-vous prendre en compte d'autres facteurs dans le modèle ? Quels devraient être ces autres facteurs ?

iii. Maintenant, supposons que vous ne puissiez collecter que des données de 1985 et 1990 au niveau départemental dans les deux États. La variable dépendante serait la proportion de conducteurs ayant leur permis de conduire arrêtés pour conduite en état d'ivresse durant l'année. De quelle façon cette structure de données est-elle différente des données individuelles décrites dans la question (i) ? Quelle méthode économétrique utiliseriez-vous ?

7. i. En utilisant les données du fichier INJURY pour l'État du Kentucky, nous reprenons (13.12), en retirons *afchnge*, et nous obtenons l'équation estimée suivante :

$$\widehat{\log(durat)} = 1{,}129 + 0{,}253\,highearn + 0{,}198\,afchnge\cdot highearn$$
$$(0{,}022)\ (0{,}042)\qquad (0{,}052)$$
$$n = 5\,626,\ R^2 = 0{,}021.$$

Est-il surprenant que l'estimation du coefficient du terme d'interaction soit très proche de celle obtenue en (13.12) ? Expliquez.

ii. Lorsque *afchnge* est inclus dans l'équation mais que *highearn* en est retiré, le résultat est

$$\widehat{\log(durat)} = 1{,}233 - 0{,}100\,afchnge + 0{,}447\,afchnge\cdot highearn$$
$$(0{,}023)\ \ (0{,}040)\qquad\quad (0{,}050)$$

$$n = 5\,626,\ R^2 = 0{,}016.$$

Pourquoi le coefficient du terme d'interaction est-il maintenant beaucoup plus grand que celui dans (13.12) ? [*Indice* : dans l'équation (13.10), quelle est l'hypothèse faite sur les groupes de traitement et de contrôle lorsque $\beta_1 = 0$?]

EXERCICES SUR ORDINATEUR

C1. Utilisez les données du fichier FERTIL1 pour cet exercice.

i. Dans l'équation estimée dans le cadre de l'exemple 13.1, testez si le cadre de vie à l'âge de 16 ans a un effet sur la fertilité. (Le groupe de référence est la grande ville). Reportez la valeur de la statistique F et sa p-valeur.

ii. Testez si la région du pays dans laquelle on vit à l'âge de 16 ans (le groupe de référence est celui du Sud) a un effet sur la fertilité.

iii. Soit u le terme d'erreur de l'équation de la population. Supposez que la variance de u change au cours du temps (mais pas avec *educ*, *age*, etc.). Un modèle tenant compte de cela peut s'écrire ainsi :

$$u^2 = \gamma_0 + \gamma_1 y74 + \gamma_2 y76 + \ldots + \gamma_6 y84 + v.$$

En utilisant ce modèle, testez l'hétéroscédasticité de u. (*Indice* : votre test F doit avoir 6 000 et 1 122 degrés de liberté).

iv. Ajoutez les termes d'interaction *y74·educ*, *y76·educ*, ..., *y84·educ* au modèle estimé dans le tableau 13.1. Expliquez ce que représentent ces termes. Sont-ils conjointement significatifs ?

C2. Pour cet exercice, utilisez les données du fichier CPS78_85.

i. Comment interprétez-vous le coefficient de *y85* dans l'équation (13.2) ? A-t-il une interprétation intéressante ? (Soyez prudent ici ; vous devez expliquer les termes d'interactions *y85·educ* et *y85·female*.)

ii. En maintenant fixés les autres facteurs, quel est le pourcentage estimé d'augmentation du salaire nominal d'un homme ayant fait 12 ans d'études ? Proposez une régression permettant d'obtenir un intervalle de confiance pour cette estimation. [*Indice* : pour obtenir l'intervalle de confiance, remplacez *y85·educ* par *y85·*(*educ* – 12) ; référez-vous à l'exemple 6.3].

iii. Faites une nouvelle estimation de l'équation (13.2) en exprimant tous les salaires en dollars de 1978. En particulier, définissez le salaire réel comme *rwage* = *wage* pour 1978 et comme *rwage* = *wage*/1,65 pour 1985. Utilisez maintenant log(*rwage*) à la place de log(*wage*) dans l'estimation de (13.2). Quels sont les coefficients qui changent par rapport à ceux de l'équation (13.2) ?

iv. Expliquez pourquoi le R carré de la régression de la question (iii) n'est pas le même que celui de l'équation (13.2). (*Indice* : les résidus et donc la somme des carrés des résidus des deux régressions sont identiques)

v. Décrivez comment la participation syndicale a changé de 1978 à 1985.

vi. En commençant par l'équation (13.2), testez si les écarts de salaires entre syndiqués et non syndiqués ont changé au cours du temps. (Ce devrait être un simple test t).

vii. Est-ce que vos résultats trouvés dans les parties (v) et (vi) se contredisent ? Expliquez.

C3. Pour cet exercice, utilisez les données du fichier KIELMC

i. La variable *dist* est la distance entre chaque maison et le site de l'incinérateur, en pieds. Examinez le modèle suivant :

$$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 \log(dist) + \delta_1 y81 \cdot \log(dist) + u.$$

Si la construction de l'incinérateur réduit la valeur des maisons proches du site, quel est le signe attendu de δ_1 ? Qu'est-ce que cela signifie si $\beta_1 > 0$?

ii. Estimez le modèle de la question (i) et reportez les résultats sous la forme habituelle. Interprétez le coefficient de $y81 \cdot \log(dist)$. Que pouvez-vous en conclure ?

iii. Ajoutez *age*, age^2, *rooms*, *baths*, log(*intst*), log(*land*), et log(*area*) à l'équation. Maintenant, que pouvez-vous conclure concernant l'effet de la présence de l'incinérateur sur la valeur des logements ?

iv. Pourquoi le coefficient de log(*dist*) est-il positif et statistiquement significatif à la question (ii) mais pas à la question (iii) ? Qu'est-ce que cela nous apprend sur les variables de contrôle introduites à la question (iii) ?

C4. Utilisez les données du fichier INJURY dans cet exercice.

i. En utilisant les données du Kentucky, estimez de nouveau l'équation (13.12), en ajoutant les variables explicatives *male*, *married*, et un jeu entier de variables indicatrices pour les industries et les types d'accidents du travail. Comment l'estimation de *afchnge.highearn* varie lorsque ces autres facteurs sont pris en compte ? Est-ce que cette estimation est statistiquement significative ?

ii. Comment interprétez-vous les faibles valeurs de R carré de la question (i) ? Cela signifie-t-il que l'équation est inutile ?

iii. Estimez l'équation (13.12) en utilisant les données pour le Michigan. Comparez les estimations sur le terme d'interaction pour le Michigan et le Kentucky. Est-ce que l'estimation pour le Michigan est statistiquement significative ? Justifiez.

C5. Utilisez les données du fichier RENTAL pour cet exercice. Les données pour les années 1980 et 1990 incluent le prix des loyers ainsi que d'autres variables caractérisant les villes dans lesquelles sont établies des universités. L'idée est de voir si la présence d'un nombre élevé d'étudiants fait augmenter les loyers. Le modèle à effets non observés est :

$$\log(rent_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 \log(pop_{it}) + \beta_2 \log(avginc_{it}) + \beta_3 pctstu_{it} + a_i + u_{it},$$

où *pop* est la population de la ville, *avginc* est le revenu moyen, et *pctstu* est le nombre d'étudiants exprimé en pourcentage de la population de la ville (au cours de l'année scolaire).

i. Estimez l'équation par la méthode des MCO sur données empilées et reportez les résultats sous la forme standard. Comment interprétez-vous l'estimation de la variable indicatrice annuelle pour 1990 ? Que pouvez-vous conclure à partir de $\hat{\beta}_3$?

ii. Les écarts-types estimés que vous avez reportés dans la question (i) sont-ils valides ? Justifiez.

iii. Maintenant, exprimez l'équation en différences et estimez-la par les MCO. Comparez votre estimation de β_3 avec celle de la question (ii). Est-ce que la taille relative de la population d'étudiants semble affecter les prix des loyers ?

iv. Obtenez les écarts-types estimés robustes à l'hétéroscédasticité pour l'équation en différences premières de la question (iii). Cela modifie-t-il vos conclusions ?

C6. Pour cet exercice, utilisez CRIME3

i. Dans le modèle de l'exemple 13.6, testez l'hypothèse H_0 : $\beta_1 = \beta_2$. (*Indice* : définissez $\theta_1 = \beta_1 - \beta_2$ et écrivez β_1 en fonction de θ_1 et β_2. Substituez cette expression dans l'équation et réarrangez les termes. Faites ensuite un test t sur θ_1).

ii. Si $\beta_1 = \beta_2$, montrez que l'équation en différences peut aussi s'écrire

$$\Delta\log(crime_i) = \delta_0 + \delta_1 \Delta avgclr_i + \Delta u_i,$$

dans laquelle $\delta_1 = 2\beta_1$ et $avgclr_i = (clrprc_{i,-1} + clrprc_{i,-2})/2$ est la moyenne du pourcentage de résolution des crimes sur les deux années précédentes.

iii. Estimez l'équation de la question (ii). Comparez le R carré ajusté avec celui de (13.22). Quel modèle utiliseriez-vous finalement ?

C7. Utilisez GPA3 dans cet exercice. Le jeu de données concerne 366 étudiants athlètes d'une grande université pendant les semestres d'automne et de printemps. [Une analyse similaire est évoquée dans Maloney et McComick (1993), mais nous utilisons ici un vrai jeu de données de panel]. Puisque vous disposez de deux semestres de données pour chaque étudiant, l'utilisation d'un modèle à effets non observés est appropriée. La première question qui nous intéresse est la suivante : est-ce que les athlètes ont de moins bons résultats universitaires durant le semestre qui correspond à la saison des compétitions de leur sport ?

i. Utilisez les MCO sur données empilées pour estimer un modèle avec la moyenne générale du semestre[7] (*trmgpa*) comme variable dépendante. Les variables explicatives sont *spring*, *sat*, *hsperc*, *female*, *black*, *white*, *frstsem*, *tothrs*, *crsgpa*, et *season*. Interprétez le coefficient de *season*. Est-il statistiquement significatif ?

ii. La plupart des athlètes qui pratiquent leur sport durant l'automne sont les footballeurs. Supposez que le niveau de compétence intellectuelle des footballeurs diffère systématiquement de celui des autres athlètes. Si leurs compétences intellectuelles ne sont pas bien représentées par leurs résultats au test d'admission à l'université (le SAT, pour « Standard Aptitude Test ») et par le centile où ils se trouvaient classés au lycée, expliquez pourquoi les estimateurs des MCO sur données empilées seront biaisés.

iii. Maintenant, utilisez les données en différences, en considérant les différences entre les deux semestres. Quelles variables sont éliminées ? Testez l'existence d'un effet lié à la saison.

iv. Pensez-vous à une ou plusieurs variables variant au cours du temps, potentiellement importantes, qui auraient été omises dans l'analyse ?

C8. Le fichier VOTE2 inclut des données de panel des élections à la Chambre des Représentants en 1988 et 1990. Seuls les vainqueurs de 1988 qui se présentent à nouveau en 1990 apparaissent dans l'échantillon ; ce sont les titulaires. Le modèle à effets non observés qui explique le pourcentage des votes pour les titulaires en fonction des dépenses de campagne des deux candidats est :

$$vote_{it} = \beta_0 + \delta_0 d90_t + \beta_1 \log(inexp_{it}) + \beta_2 \log(chexp_{it}) + \beta_3 incshr_{it} + a_i + u_{it},$$

où $incshr_{it}$ est la part des dépenses de campagne des titulaires dans le total des dépenses de campagne (sous forme de pourcentage). L'effet non observé a_i contient les caractéristiques des titulaires – comme

7 NDT : « term Grade Point Average » en anglais, d'où l'abréviation *trmgpa*.

la « qualité » – de même que des informations sur le district constantes au cours du temps. Le genre et le parti du titulaire sont constants au cours du temps, donc ils sont englobés dans a_i. Nous nous intéressons à l'effet des dépenses faites pour la campagne sur les résultats des élections.

i. Écrivez l'équation donnée en différences, en faisant la différence entre les deux années. Estimez l'équation en différences par les MCO. Quelles variables sont individuellement significatives à un niveau de 5 % par rapport à l'hypothèse alternative bilatérale ?

ii. Dans l'équation de la question (i), testez la significativité jointe de $\Delta\log(inexp)$ et $\Delta\log(chexp)$. Reportez la *p*-valeur.

iii. Estimez de nouveau l'équation de la question (i) en utilisant $\Delta incshr$ comme unique variable indépendante. Interprétez le coefficient de $\Delta incshr$. Par exemple, si la répartition des dépenses des titulaires augmente de 10 %, comment cela devrait-il affecter la répartition des votes des titulaires ?

iv. Refaites la question (iii), mais utilisez maintenant seulement les élections où une même paire de candidats était en compétition en 1988 et en 1990. [Cela nous permet de prendre en compte également les caractéristiques des candidats concurrents qui seraient dans a_i. Levitt (1994) a mené une analyse plus poussée].

C9. Pour cet exercice utilisez le fichier CRIME4

i. Reprenez l'équation de l'exemple 13.9. Ajoutez les logs de chaque variable de salaire dans le jeu de données et estimez le modèle en faisant les différences premières. Comment le fait d'inclure ces variables affecte-t-il les coefficients des variables liées à la justice pénale ?

ii. Est-ce que toutes les variables de salaire dans (i) ont le signe attendu ? Sont-elles conjointement significatives ? Expliquez.

C10. Pour cet exercice, nous utilisons JTRAIN pour déterminer l'effet des subventions pour la formation professionnelle sur le nombre d'heures de formation professionnelle par employé. Le modèle de base pour les trois années est

$$hrsemp_{it} + \beta_0 + \delta_1 d88_t + \delta_2 d89_t + \beta_1 grant_{it} + \beta_2 grant_{i,t-1} + \beta_3 \log(employ_{it}) + a_i + u_{it}.$$

i. Estimez l'équation en différences premières. Combien d'entreprises sont utilisées dans l'estimation ? Combien d'observations totales seraient utilisées si chaque entreprise disposait des données pour toutes les variables (en particulier, *hrsemp*) sur les trois périodes considérées ?

ii. Interprétez le coefficient de *grant* et expliquez ce qu'il signifie.

iii. Est-il surprenant que $grant_{21}$ ne soit pas significatif ? Expliquez.

iv. Les plus grosses entreprises forment-elles plus ou moins leurs employés, en moyenne ? Quelle est l'ampleur de la différence de formation ?

C11. Le fichier MATHPNL contient des données de panel concernant les districts scolaires du Michigan pour les années allant de 1992 à 1998. Ces données sont l'équivalent au niveau des districts des données utilisées par Papke (2005) au niveau des écoles. Nous nous intéressons à *math4*, le pourcentage d'élèves en quatrième année d'école primaire[8] dans le district qui ont obtenu la moyenne à un test standard de mathématiques. La variable explicative clé est *rexpp*, ce qui correspond aux dépenses réelles par élève dans le district. Les montants sont en dollars de 1997. La variable de dépenses apparaîtra sous forme logarithmique.

8 NDT : Soit l'équivalent du CM1 en France.

i. Considérez le modèle statique à effets non observés suivant :

$$math4_{it} = \delta_1 y93_t + \ldots + \delta_6 y98_t + \beta_1 \log(rexpp_{it}) + \beta_2 \log(enrol_{it}) + \beta_3 lunch_{it} + a_i + u_{it},$$

$enrol_{it}$ est le nombre total d'inscriptions dans le district. $lunch_{it}$ est le pourcentage d'élèves éligibles au programme de repas scolaires subventionnés ; c'est une assez bonne mesure du taux de pauvreté du district. Expliquez pourquoi $\beta_1/10$ est la variation en point de pourcentage de $math4_{it}$ lorsque les dépenses réelles par élève augmentent d'environ 10 %.

ii. Utilisez les différences premières pour estimer le modèle de la question (i). L'approche la plus simple est d'introduire une constante dans l'équation en différences premières ainsi que des variables indicatrices pour les années allant de 1994 à 1998. Interprétez le coefficient de la variable des dépenses.

iii. Maintenant, ajoutez la variable des dépenses retardée au modèle et réestimez-le en utilisant les différences premières. Notez que vous perdez une autre année de données, donc vous n'utilisez les variations qu'à partir de 1994. Discutez des coefficients des variables de dépenses (variable courante et variable retardée) et de leur significativité.

iv. Donnez les écarts-types estimés robustes à l'hétéroscédasticité pour la régression en différences premières de la question (iii). Faites l'analyse comparative de ces écarts-types estimés avec ceux de la question (iii) pour les variables des dépenses.

v. Maintenant, donnez les écarts-types estimés robustes à la présence d'hétéroscédasticité et de corrélation sérielle. Qu'est-ce que cela implique pour la significativité de la variable de dépenses retardée ?

vi. Vérifiez que les erreurs sur les différences $r_{it} = \Delta u_{it}$ ont une autocorrélation négative en testant la présence d'autocorrélation AR(1).

vii. En vous basant sur un test joint entièrement robuste, vous semble-t-il nécessaire d'inclure les variables liées aux inscriptions et aux repas subventionnés dans le modèle ?

C12. Utilisez les données de MURDER dans cet exercice.

i. En utilisant les années 1990 et 1993, estimez l'équation suivante :

$$mrdrte_{it} = \delta_0 + \delta_1 d93_t + \beta_1 exec_{it} + \beta_2 unem_{it} + a_i + u_{it}, \; t = 1, 2$$

avec la méthode des MCO et reportez le résultat sous la forme habituelle. Ne tenez pas compte du fait que les écarts-types estimés des MCO habituels sont inappropriés à cause de la présence de a_i. Estimez-vous que la peine capitale a un effet dissuasif ?

ii. Calculez les estimations en DP (utilisez seulement les différences de 1990 à 1993 ; vous devriez obtenir 51 observations dans la régression en DP). Maintenant, que concluez-vous sur l'effet dissuasif ?

iii. Dans la régression en DP[9] de la question (ii), obtenez les résidus, soit $\hat{e}_i$. Faites tourner la régression de Breusch-Pagan de $\hat{e}_1^2$ sur $\Delta exec_i$, $\Delta unem_i$ et calculez le test F pour l'hétéroscédasticité. Faites la même chose pour le cas particulier du test de White [c'est-à-dire, faites la régression $\hat{e}_1^2$ *sur* $\hat{y}_i$, $\hat{y}_1^2$, où les valeurs ajustées viennent de la question (ii)]. Que pouvez-vous en conclure sur l'hétéroscédasticité dans l'équation en DP ?

iv. Faites la même régression à partir de la question (ii) mais cette fois, obtenez les statistiques t robustes à l'hétéroscédasticité. Que se passe-t-il ?

9 Rappel : en DP est l'abréviation pour « en différences premières ».

v. Sur quelle statistique t associée à $\Delta exec_i$ vous semble-t-il plus facile de vous appuyer : la statistique habituelle ou celle qui est robuste à l'hétéroscédasticité ? Pourquoi ?

C13. Utilisez les données de WAGEPAN dans cet exercice.

i. Considérez le modèle à effets non observés suivant :

$$lwage_{it} = \beta_0 + \delta_1 d81_t + \ldots + \delta_7 d87_t + \beta_1 educ_i + \gamma_1 d81_t educ_i + \ldots + \delta_7 d87_t educ_i + \beta_2 union_{it} + a_i + u_{it},$$

où a_i peut être corrélé avec $educ_i$ et $union_{it}$. Quels paramètres pouvez-vous estimer avec la méthode des différences premières ?

ii. Estimez l'équation de la question (i) en DP, et testez l'hypothèse nulle selon laquelle le rendement des années d'études n'a pas varié au cours du temps.

iii. Testez l'hypothèse de la question (ii) en utilisant un test complètement robuste, c'est-à-dire un test qui permet aux erreurs en DP Δu_{it} d'avoir une hétéroscédasticité arbitraire et d'être autocorrélées. Vos conclusions changent-elles ?

iv. Permettez maintenant à la différence de salaire entre syndiqués et non syndiqués de varier au cours du temps (avec les années d'études) et estimez l'équation par DP. Quelles sont les différences estimées en 1980 ? Et en 1987 ? Cette différence est-elle statistiquement significative ?

v. Testez l'hypothèse nulle selon laquelle la prime salariale des syndiqués n'a pas changé au cours du temps et discutez de vos résultats à la lumière de votre réponse à la question (iv).

C14. Utilisez les données de JTRAIN3 pour cette question

i. Estimez le modèle de régression simple $re78 = \beta_0 + \beta_1 train + u$, et reportez les résultats sous la forme habituelle. En vous basant sur cette régression, vous semble-t-il que la formation professionnelle au travail qui a été suivie en 1976 et 1977 a eu un effet positif sur les gains réels des employés en 1978 ?

ii. Maintenant utilisez la variation des gains réels, $cre = re78 - re75$, en tant que variable dépendante. (Il n'est pas nécessaire de faire la différence pour *train* puisque nous supposons qu'il n'y avait pas de programme de formation professionnelle au travail avant 1975. Ainsi, si on définit $ctrain = train78 - train75$ alors $ctrain = train78$ puisque $train75 = 0$). Désormais quel est l'effet estimé de la formation ? Discutez en quoi il est comparable avec l'estimation de la question (i).

iii. Calculez l'intervalle de confiance à 95 % pour l'effet du programme de formation professionnelle en vous servant des écarts-types estimés par les MCO habituels et des écarts-types estimés robustes à l'hétéroscédasticité, et décrivez vos résultats.

C15. Le jeu de données de HAPPINESS contient des coupes transversales indépendantes empilées pour les années paires allant de 1994 à 1998, obtenues par *l'Enquête Sociale Générale* (« General Social Survey »). La variable dépendante dans ce problème est la mesure du « bonheur », *vhappy*, qui est une variable binaire égale à un si la personne se dit « très heureuse » (au contraire de « heureuse » et de « pas très heureuse »).

i. Quelle année contient le plus grand nombre d'observations ? Laquelle en possède le moins ? Quel est le pourcentage de personnes dans l'échantillon se disant « très heureuses » ?

ii. Faites la régression de *vhappy* sur toutes les variables indicatrices annuelles, en laissant de côté *y*94 puisque 1994 est l'année de référence. Calculez une statistique robuste à l'hétéroscédasticité pour tester l'hypothèse nulle suivante : la proportion de personnes très heureuses n'a pas varié au cours du temps. Quelle est la *p*-valeur du test ?

iii. Ajoutez les variables indicatrices *occattend* et *regattend* à la régression de la question (ii). (Souvenez-vous que les coefficients sont interprétés relativement à un groupe de référence). Comment résumeriez-vous les effets de la fréquentation des églises sur le bonheur ?

iv. Définissez une variable, *highinc*, égale à un si les revenus de la famille sont supérieurs à 25 000 $. (Malheureusement, le même seuil est utilisé pour toutes les années, cette variable ne tient donc pas compte de l'inflation. Par ailleurs, 25 000 $ ne peut pas vraiment être considéré comme un revenu très élevé.) Incluez *highinc*, *unem10*, *educ* et *teens* dans la régression de la question (iii). Le coefficient de *regattend* est-il très affecté ? Que diriez-vous de sa significativité statistique ?

v. Discutez le signe, l'amplitude et la significativité statistique des quatre nouvelles variables de la question (iv). Ces estimations vous paraissent-elles avoir un sens ?

vi. En prenant en compte les facteurs de la question (iv), trouvez-vous qu'il y ait des différences de bonheur selon le genre ou l'origine ethnique ? Justifiez votre réponse.

C16. Utilisez les données du fichier COUNTYMURDERS pour répondre à cet exercice. La base de données liste les meurtres et les exécutions (peine capitale) dans les 2197 comtés des États-Unis.

i. Trouvez la valeur moyenne de *murdrate* dans tous les comtés sur l'ensemble de la période. Quel est l'écart type ? Pour quel pourcentage de l'échantillon la variable *murdrate* est-elle égale à zéro ?

ii. Pour combien d'observations la variable *execs* est-elle égale à zéro ? Quelle est la valeur maximale *d'execs* ? Pourquoi la moyenne *d'execs* est-elle si petite ?

iii. Considérez le modèle suivant

$$murdrate_{it} = \theta_t + \beta_1 execs_{it} + \beta_2 execs_{it-1} + \beta_3 percblack_{it} + \beta_4 percmale_i + \beta_5 perc1019 + \beta_6 perc2029 + a_i + u_{it}$$

où θ_t représente une constante différente pour chaque période, a_i est l'effet fixe comté, et u_{it} l'erreur idiosyncratique. Quelle hypothèse devons-nous faire sur les variables a_i et *execution* pour que les MCO groupés estiment de manière convergente les paramètres du modèle, en particulier β_1 et β_2 ?

iv. Estimez le modèle de la question (iii) par MCO groupés et reportez les estimateurs de β_1 et β_2 et leurs écarts types estimés. À votre avis, que se passe-t-il ?

v. Même si les estimateurs des MCO groupés sont convergents, pensez-vous que les écarts types estimés obtenus en (iv) sont corrects ? Expliquez.

vi. À présent, estimez l'équation de la question (iii) en utilisant les différences premières et éliminez ainsi a_i. Quels sont les nouveaux estimateurs de β_1 et β_2 ? Sont-ils très différents de ceux qui ont été obtenus à la question (iv) ?

vii. En utilisant les résultats obtenus à la question (vi), pouvez-vous dire qu'il y a un effet statistiquement significatif de la peine capitale sur le taux d'homicides ? Si possible, en plus des écarts types estimés habituels des MCO, utilisez les écarts types estimés robustes à la corrélation sérielle et à l'hétéroscédasticité.

ANNEXE 13A

13A.1 Hypothèses pour les MCO sur données empilées utilisant les différences premières

Dans cette annexe, nous présentons de manière rigoureuse les hypothèses concernant les estimateurs en différences premières. Leur démonstration est parfois évoquée ; vous la trouverez en détail dans Wooldridge (2010, chapitre 10).

Hypothèse DP.1

Pour chaque i le modèle est donné par :

$$y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, t = 1, \ldots, T,$$

où les β_j sont les paramètres à estimer et a_i est l'effet non observé.

Hypothèse DP.2

Nous disposons d'un échantillon aléatoire à partir de la coupe transversale.

Hypothèse DP.3

Chaque variable explicative varie au cours du temps (pour au moins un i), et aucune variable explicative n'est une combinaison linéaire parfaite des autres.

Pour faciliter l'expression de l'hypothèse suivante, notons $\mathbf{X}_i$ l'ensemble des variables explicatives pour toutes les périodes pour l'entité i ; ainsi, $\mathbf{X}_i$ désigne l'ensemble des x_{itj}, avec $t = 1, \ldots, T, j = 1, \ldots, k$.

Hypothèse DP.4

Pour chaque t, la valeur attendue de l'erreur idiosyncratique, sachant les valeurs des variables explicatives pour *toutes* les périodes et l'effet non observé, est zéro : $\mathbf{E}(u_{it}|\mathbf{X}_i, a_i) = \mathbf{0}$. Lorsque l'hypothèse DP.4 est valide, nous disons parfois que $\boldsymbol{x}_{itj}$ *est strictement exogène conditionnellement à l'effet non observé*. L'idée est qu'une fois que l'on tient compte de a_i, pour tout s et t, il n'y a pas de corrélation entre x_{isj} et l'erreur idiosyncratique restante u_{it}.

Comme nous l'avons précisé, l'hypothèse DP.4 est plus forte que nécessaire. Nous utilisons cette forme d'hypothèse puisqu'elle souligne que notre intérêt se porte sur l'équation suivante :

$$\mathrm{E}(y_{it}|\mathbf{X}_i, a_i) = \mathrm{E}(y_{it}|\mathbf{x}_{it}, a_i) = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i,$$

et donc que les β_j mesurent les effets marginaux des variables explicatives observées en gardant fixé ou « en prenant compte » l'effet inobservé a_i. Néanmoins, l'implication la plus importante de DP.4 est $\mathrm{E}(\Delta u_{it}|\mathbf{X}_i) = 0$, $t = 2, \ldots, T$. Cette implication est aussi suffisante pour assurer le caractère non biaisé de l'estimateur de DP.4. Par souci de cohérence, on peut simplement supposer que Δx_{itj} est non corrélé à Δu_{it} pour tout $t = 2, \ldots, T$ et $j = 1, \ldots, k$. Voir Wooldridge (2010, chapitre 10) pour une discussion approfondie.

Sous ces quatre premières hypothèses, les estimateurs en différences premières ne sont pas biaisés. L'hypothèse clé est la DP.4, puisqu'elle énonce la stricte exogénéité des variables explicatives. À l'aide de ces mêmes hypothèses, on peut aussi montrer que l'estimateur DP est convergent avec un T fixé et quand $N \to \infty$ (et peut-être aussi de façon plus générale).

Les deux hypothèses suivantes permettent de s'assurer que les écarts-types estimés et les tests statistiques issus des MCO sur données empilées en différences premières sont (asymptotiquement) valides.

Hypothèse DP.5

La variance des erreurs en différences, conditionnellement à toutes les variables explicatives, est constante : $\mathrm{Var}(\Delta u_{it}|\mathbf{X}_i) = \sigma^2$, $t = 2, \ldots, T$.

Hypothèse DP.6

Pour tout $t \neq s$, les erreurs idiosyncratiques en différences ne sont pas corrélées (conditionnellement à toutes les variables explicatives)

$$\text{Cov}(\Delta u_{it}, \Delta u_{is} | \mathbf{X}_i) = 0, \; t \neq s.$$

L'hypothèse DP.5 garantit que les erreurs en différences, Δu_{it}, sont homoscédastiques. L'hypothèse DP.6 établit que les erreurs différenciées ne sont pas autocorrélées, ce qui signifie que u_{it} suit une marche aléatoire (voir le chapitre 11). Dans Sous les hypothèses DP.1 à DP.6, l'estimateur en DP de β_j est le meilleur estimateur linéaire sans biais (conditionnellement aux variables explicatives).

Hypothèse DP.7

Conditionnellement à $\mathbf{X}_i$, les Δu_{it} sont des variables aléatoires normales indépendantes et identiquement distribuées. Lorsqu'on ajoute l'hypothèse DP.7, les estimateurs en DP sont distribués normalement, et les statistiques t et F obtenues par les MCO sur données empilées en différences ont des distributions exactes de t et de F. Sans DP.7, on peut revenir aux approximations asymptotiques habituelles.

13A.2 Calcul des écarts-types estimés robustes à la corrélation sérielle et à l'hétéroscédasticité quand elles sont de forme inconnue.

Puisque l'estimateur DP est convergent pour $N \to \infty$ selon les hypothèses DP.1 à DP.4, il serait très pratique de disposer d'une méthode simple pour obtenir des écarts-types estimés et des tests statistiques permettant n'importe quelle forme d'autocorrélation ou d'hétéroscédasticité dans les erreurs en DP, $e_{it} = \Delta u_{it}$.

Heureusement, étant donné que N est assez grand, et que T n'est pas « trop grand », il est facile d'obtenir des écarts-types estimés robustes et des tests statistiques appropriés. Comme nous l'avons mentionné dans le texte, un traitement plus détaillé dépasserait le cadre de cet ouvrage. Les arguments techniques regroupent les observations décrites dans les chapitres 8 et 12, dans lesquels les statistiques robustes à l'hétéroscédasticité et l'autocorrélation sont discutées. En fait, les données de panel présentent un avantage important : puisqu'on dispose d'une grande coupe transversale, il est possible de permettre une autocorrélation non contrainte dans les erreurs $\{e_{it}\}$ à condition que T ne soit pas trop grand. Nous pouvons comparer cette situation avec l'approche de Newey-West de la section 12.5, où les covariances estimées doivent être sous-pondérées quand les observations sont plus éloignées dans le temps.

L'approche générale qui permet d'obtenir des écarts-types estimés entièrement robustes et des tests statistiques avec des données de panel s'appelle le **regroupement** (*clustering*), et nous avons emprunté de nombreuses idées à la littérature sur l'échantillonnage en groupe, aussi appelé échantillonnage par grappes. L'idée est que chaque unité d'une coupe transversale est définie comme étant un groupe d'observations au cours du temps, et il est permis d'avoir une corrélation – autocorrélation – arbitraire et des variances qui changent au sein de chaque groupe. En raison de l'importance de l'échantillonnage par grappes, de nombreux logiciels économétriques disposent d'options adaptées, et tiennent compte de l'existence de grappes dans le calcul des écarts-types estimés et des tests statistiques. La plupart des commandes ressemblent à :

regress cy cx1 cx2 ... cxk, cluster(id)

où « id » est la variable qui contient les identifiants uniques pour chaque unité d'une coupe transversale (et le « c » avant chaque variable indique « changement »). L'option « *cluster(id)* » à la fin de la commande « *regress* » indique au logiciel de reporter tous les écarts-types estimés et les tests statistiques – y compris les statistiques t et de type F – de sorte qu'ils soient valides pour de grandes coupes transversales, avec n'importe quelle forme d'autocorrélation ou d'hétéroscédasticité. Le report de telles statistiques est très courant dans les travaux empiriques modernes sur données de panel. Souvent, les écarts-types corrigés seront beaucoup plus grands que les écarts-types estimés habituels ou que ceux qui ne corrigent que l'hétéroscédasticité. De plus grands écarts-types estimés reflètent mieux l'erreur d'échantillonnage dans les coefficients des MCO sur données empilées.

CHAPITRE

14

MÉTHODES AVANCÉES EN ÉCONOMÉTRIE DES DONNÉES DE PANEL

Traduction de Sophie Béreau

Dans ce chapitre, nous nous intéressons à deux méthodes permettant d'estimer des effets inobservés dans le cadre de modèles à données de panel qui sont tout aussi communes que la technique de différenciation première. Bien que ces méthodes s'avèrent plus difficiles à décrire et mettre en œuvre, un certain nombre de logiciels standard d'économétrie permettent de les utiliser facilement.

Dans la section 14.1, nous présentons l'estimateur du modèle à effets fixes. À l'instar de l'estimateur en différences premières, il repose sur une transformation permettant d'éliminer l'effet inobservé a_i préalablement à la phase d'estimation. Toutes les variables explicatives invariantes dans le temps sont ainsi retirées du modèle avec a_i.

Dans la section 14.2, nous présentons l'estimateur du modèle à effets aléatoires. Celui-ci est pertinent lorsqu'il est raisonnable de penser que l'effet inobservé n'est corrélé à aucune des variables explicatives du modèle. Le raisonnement est le suivant. Si nous introduisons suffisamment de variables de contrôle pertinentes dans notre équation, nous sommes alors en droit de penser que quelle que soit la source d'hétérogénéité résiduelle potentielle, elle ne pourra qu'engendrer de la corrélation sérielle dans le terme d'erreur composé et non de la corrélation entre le terme d'erreur et les variables explicatives du modèle. L'estimation de modèles à effets aléatoires par la méthode des moindres carrés généralisés est relativement aisée et peut être prise en charge par les logiciels économétriques usuels.

La section 14.3 introduit une approche relativement récente, celle du modèle à effets aléatoires corrélés. À cette occasion, nous proposons une revue synthétique des méthodes à effets fixes et aléatoires, méthodes qui s'avèrent en pratique bien utiles.

Dans la section 14.4, nous montrons comment les méthodes de données de panel peuvent être appliquées à d'autres structures de données, telles que les données d'appariement ou issues de techniques d'échantillonnage en grappes.

14.1 ESTIMATION DU MODÈLE À EFFETS FIXES

La différenciation première est une technique parmi d'autres pour éliminer les effets fixes a_i. Une méthode alternative, qui fonctionne mieux sous certaines hypothèses, est appelée **transformation *within***. Pour voir ce que cette méthode implique, considérons un modèle avec une unique variable explicative. Pour chaque unité individuelle i, on a :

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, \; t = 1, 2, \ldots, T. \qquad \textbf{[14.1]}$$

Pour chaque i, exprimons la moyenne de cette équation au cours du temps. Nous obtenons :

$$\bar{y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i, \qquad \textbf{[14.2]}$$

avec $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$ et ainsi de suite. Du fait que le paramètre a_i est constant au cours du temps, il apparaît à la fois dans (14.1) et (14.2). En soustrayant (14.2) de (14.1) pour chaque t, nous obtenons :

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i, \; t = 1, 2, \ldots, T,$$

ou

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it}, \; t = 1, 2, \ldots, T, \qquad \textbf{[14.3]}$$

avec $\ddot{y}_{it} = y_{it} - \bar{y}_i$ la **valeur de *y* centrée sur sa moyenne**, que l'on définit de façon identique pour $\ddot{x}_{it}$ et $\ddot{u}_{it}$. L'élément important relatif à l'équation (14.3) tient à ce que l'effet inobservé, a_i, a disparu. Ceci implique que nous sommes en mesure d'estimer (14.3) par la méthode des MCO sur les données empilées.

L'estimateur des MCO sur les données empilées[1] reposant sur des variables en écart à leur moyenne temporelle est appelé **estimateur à effets fixes** ou **estimateur *within***. Cette dernière appellation dérive de ce que les MCO appliqués à l'équation (14.3) reposent sur la variation temporelle de y et x au sein (*within*) de la dimension individuelle de chaque observation.

L'*estimateur between* correspond quant à lui, à l'estimateur des MCO sur l'équation en coupe transversale (14.2) (dans laquelle nous introduisons un constante β_0) : nous calculons les moyennes temporelles pour y et x et réalisons ensuite une régression en coupe. Nous n'étudierons pas ici l'estimateur *between* en détails car il est biaisé lorsque a_i est corrélé avec $\bar{x}_i$ (voir à ce titre l'exercice 2). Lorsque nous avons de bonnes raisons de penser qu'a_i n'est pas corrélé avec x_{it}, il est préférable d'avoir recours à l'estimateur à effets aléatoires que nous abordons dans la section 14.2. En effet, l'estimateur *between* ignore une partie importante de l'information relative à la façon dont les variables évoluent au cours du temps.

Ajouter des variables explicatives supplémentaires à l'équation n'engendre que peu de changements. Le **modèle à effets inobservés** dans sa forme originale est donné par :

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \ldots + \beta_k x_{itk} + a_i + u_{it},\ t = 1, 2, \ldots, T. \qquad \textbf{[14.4]}$$

Nous exprimons alors simplement chacune des variables en écart à sa moyenne temporelle, en incluant des éléments tels que des variables indicatrices temporelles, et effectuons une régression par la méthode des MCO sur les données empilées. À l'issue de ces transformations, l'équation pour chaque unité individuelle i est donnée par :

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \beta_2 \ddot{x}_{it2} + \ldots + \beta_k \ddot{x}_{itk} + \ddot{u}_{it},\ t = 1, 2, \ldots, T, \qquad \textbf{[14.5]}$$

que nous estimons par la méthode des MCO sur les données empilées.

Sous l'hypothèse d'exogénéité stricte des variables explicatives, l'estimateur à effets fixes est sans biais : grosso modo, cela implique que le terme d'erreur idiosyncratique u_{it} est décorrélé de chacune des variables explicatives pour chaque période. (Voir l'annexe du chapitre pour l'établissement rigoureux de ces hypothèses.) L'estimateur à effets fixes autorise de la corrélation arbitraire entre a_i et les variables explicatives à n'importe quelle période, à l'instar de la différenciation première. De ce fait, toute variable explicative invariante au cours du temps pour tout i se trouve éliminée par la transformation du modèle menant à l'expression de l'estimateur à effets fixes (la transformation *within*) puisque $\ddot{x}_{it} = 0$ pour tout i et t, si x_{it} est invariant au cours du temps. De ce fait, nous ne pouvons inclure aucune variable telle que le genre ou la distance d'une ville à la rivière la plus proche dans notre modèle de régression.

Pour aller plus loin 14.1

Supposons que dans le cadre d'un modèle expliquant le niveau d'épargne d'une famille pour les années 1990, 1991, et 1992, nous posions $kids_{it}$ comme le nombre d'enfants par famille i pour l'année t. Si le nombre d'enfants est constant au cours de cette période de trois années pour la plupart des familles de l'échantillon, quels problèmes cela pose-t-il pour l'estimation de l'effet du nombre d'enfants sur le niveau d'épargne ?

L'autre hypothèse requise pour la bonne pratique des MCO a trait à l'homoscédasticité et à l'absence de corrélation sérielle des erreurs u_{it} (au cours du temps) : voir l'annexe de ce chapitre pour plus de détails.

Il existe une subtilité dans la détermination du nombre de degrés de liberté de l'estimateur à effets fixes. Lorsque nous estimons l'équation (14.5) par la méthode des MCO sur les données empilées, nous

1 NDT : estimateur « *pool* » en anglais

considérons un total de NT observations et k variables indépendantes. [Notez qu'il n'y a pas de constante dans (14.5) ; celle-ci est éliminée du fait de la transformation menant au modèle à effets fixes.] Dès lors, nous devrions avoir $NT - k$ degrés de liberté. Ce calcul est cependant incorrect. Pour chaque observation en coupe i, nous perdons un degré de liberté (*ddl*) car nous avons écrit le modèle en écart à la moyenne. En d'autres termes, pour chaque i, les erreurs en écart à la moyenne $\ddot{u}_{it}$ somment à zéro lorsqu'elles sont sommées sur t, ce qui nous fait perdre un degré de liberté supplémentaire. (Notez qu'une telle contrainte n'existe pas dans l'expression originale du terme d'erreur idiosyncratique u_{it}.) De ce fait, le nombre de degrés de liberté approprié est $ddl = NT - N - k = N(T - 1) - k$. Heureusement pour nous, les logiciels modernes contenant des routines pour l'estimation des modèles à effets fixes calculent correctement le nombre de degrés de liberté *ddl*. Mais dans l'hypothèse où nous aurions à calculer à la main nos séries en écart à leur moyenne puis à procéder à l'estimation par les MCO sur les données empilées, nous devrions alors corriger les écarts-types estimés ainsi que les statistiques de test associées.

EXEMPLE 14.1
Des effets de la formation professionnelle sur le taux de rebut des entreprises

Nos données couvrent ici trois années, 1987, 1988, et 1989, pour 54 entreprises ayant reporté des taux de rebut chaque année. Aucune des firmes considérées n'a reçu de subvention avant 1988 ; en 1988, 19 entreprises ont reçu des subventions ; en 1989, 10 entreprises distinctes ont reçu des subventions. De ce fait, nous devons considérer la possibilité que la formation professionnelle reçue en 1988 a permis aux travailleurs d'être plus productifs en 1989. Ceci peut facilement se tester en incluant la valeur retardée de l'indicateur *grant*. Nous introduisons en outre des variables indicatrices pour 1988 et 1989. Les résultats d'estimation sont détaillés dans le tableau 14.1.

Nous avons reporté les résultats de façon à mettre en évidence l'interprétation des estimations du modèle à effets inobservés (14.4). Nous tenons compte ici de façon explicite de l'influence des effets inobservés constants au cours du temps dans a_i. L'expression du modèle en écart à la moyenne nous permet d'estimer les β_j, mais (14.5) n'apparaît pas alors comme la meilleure équation pour interpréter les résultats d'estimation.

Selon notre estimation, l'effet retardé de la subvention à la formation apparaît substantiellement plus important que l'effet contemporain : la formation professionnelle présente un effet au moins un an plus tard. Du fait que la variable expliquée est exprimée sous forme logarithmique, le modèle prédit que l'obtention d'une subvention en 1988 réduira le taux de rebut de l'entreprise en 1989 d'environ 34,4 % [$\exp(-0{,}422) - 1 \approx -0{,}344$] ; le coefficient associé à $grant_{-1}$ est significatif au seuil de 5 % contre l'alternative bilatérale. Le coefficient associé à *grant* est significatif au seuil de 10 %, et l'amplitude du coefficient est loin d'être triviale. Notons que les *ddl* sont obtenus par le calcul suivant : $N(T - 1) - k = 54(3 - 1) - 4 = 104$.

Le coefficient associé à *d89* indique que le taux de rebut était substantiellement moins élevé en 1989 qu'en 1987, même en l'absence de subvention à la formation professionnelle cette année-là. De ce fait, il est important de permettre des effets agrégés. Si nous avions omis les variables indicatrices annuelles, l'accroissement séculaire de la productivité du travail aurait pu être perçu comme provenant des subventions à la formation professionnelle. Le tableau 14.1 nous montre que même après avoir tenu compte des tendances agrégées de la productivité, les subventions à la formation professionnelle ont eu un effet estimé important.

Enfin, il est essentiel d'autoriser l'introduction d'effets retardés dans le modèle. Si nous omettons la variable $grant_{-1}$, alors nous faisons l'hypothèse que l'effet de la formation professionnelle ne dépasse pas l'année en cours. L'estimation du coefficient associé à *grant* lorsque nous retirons $grant_{-1}$ de l'équation est de $-0{,}082$ ($t = -0{,}65$) ; ce qui n'apparaît pas statistiquement significatif.

Pour aller plus loin 14.2

Selon le programme mis en place dans le Michigan, si une entreprise a bénéficié d'une subvention une année, elle n'est pas éligible pour une nouvelle subvention l'année suivante. Que cela implique-t-il pour la valeur de la corrélation entre les variables *grant* et $grant_{-1}$?

Tableau 14.1 *Estimation des effets fixes de l'équation de taux de rebut*

Variable dépendante : log(*scrap*)	
Variables indépendantes	**Coefficients (Écarts-types estimés)**
d88	−0,080 (0,109)
d89	−0,247 (0,133)
grant	−0,252 (0,151)
$grant_{-1}$	−0,422 (0,210)
Observations Degrés de liberté *R*-carré	162 104 0,201

Lorsque nous estimons un modèle à effets inobservés au moyen d'effets fixes, calculer une bonne mesure d'ajustement du modèle n'est pas chose aisée. Le *R*-carré donné dans le tableau 14.1 est construit sur base de la transformation *within* : c'est la valeur du *R*-carré obtenu à l'issue de l'estimation de l'équation (14.5). De ce fait, il peut être interprété comme la variation temporelle de y_{it} pouvant être expliquée par la variation temporelle des variables explicatives. D'autres manières de calculer le *R*-carré sont possibles, l'une d'elles sera présentée plus loin.

Bien que les variables invariantes dans le temps ne puissent pas être introduites directement dans le modèle à effets fixes, elles *peuvent* être prises en compte sous forme d'interactions avec d'autres variables changeant au cours du temps et en particulier avec des variables indicatrices annuelles. Par exemple, dans l'équation de salaire où le niveau d'éducation est constant dans le temps pour chacun des individus de l'échantillon, nous avons la possibilité de faire interagir le niveau d'éducation avec chacune des variables indicatrices annuelles pour mesurer de combien les rendements de l'éducation ont changé au cours du temps. Nous ne pouvons en revanche, pas avoir recours à des effets fixes pour estimer le niveau des rendements de l'éducation durant la période de référence. Cela implique que nous ne pouvons, pour aucune période, estimer les rendements de l'éducation, tout au plus le différentiel de rendement d'une année donnée par rapport à l'année de référence. La section 14.3 décrit une approche permettant aux coefficients associés aux variables invariantes dans le temps d'être estimés tout en préservant la nature du modèle à effets fixes.

Lorsque nous introduisons un ensemble de variables indicatrices annuelles – soient des variables indicatrices pour chacune des années sauf la première – nous ne pouvons estimer l'effet d'aucune variable dont les *variations* au cours du temps seraient identiques. Un exemple est celui des années d'expérience dans le cadre de données de panel où chaque individu travaille chaque année, de sorte que l'expérience augmente

chaque année d'une année pour chaque individu de l'échantillon. La présence de l'effet fixe a_i tient compte des différences entre les individus en matière d'années d'expérience à la date initiale. Mais ensuite, l'effet d'une année supplémentaire d'expérience professionnelle ne peut être distingué des effets du temps agrégés (puisque l'expérience augmente du même nombre d'année(s) pour chaque individu). Cela serait également vrai si, en lieu et place de variables indicatrices annuelles, nous avions eu recours à un modèle à tendance linéaire : pour chaque individu, l'expérience n'aurait pu être distinguée de la tendance linéaire.

EXEMPLE 14.2
Les rendements de l'éducation ont-t-ils évolué avec le temps ?

Les données contenues dans WAGEPAN sont reprises de Vella et Verbeek (1998). Chacun des 545 hommes de l'échantillon a travaillé chaque année de 1980 à 1987. Un certain nombre de variables de la base de données ont changé au cours du temps et parmi les plus importantes l'expérience, le statut marital, et le statut syndical. D'autres variables sont quant à elles restées inchangées telles que les origines ethniques ou le niveau d'éducation. Si nous avons recours à des effets fixes (ou différencions le modèle), nous ne pouvons inclure les variables relatives aux origines ethniques, au niveau d'éducation ou à l'expérience dans l'équation. Cependant, nous pouvons introduire la variable *educ* en interaction avec les variables indicatrices annuelles de 1981 à 1987 pour tester la constance des rendements de l'éducation sur la période. Notre variable dépendante est log(*wage*) ; quant aux variables indépendantes, nous considérons des variables indicatrices pour les statuts marital et syndical, un ensemble de variables indicatrices pour les différentes années ainsi que les variables d'interaction *d81•educ*, *d82•educ*, ..., *d87•educ*.

Les coefficients estimés relatifs à ces variables d'interaction sont tous positifs, et en général plus élevés pour les années récentes. Le coefficient le plus élevé est de 0,030 pour *d87•educ*, avec $t = 2{,}48$. En d'autres termes, les rendements de l'éducation sont estimés à environ 3 points de pourcentage de plus en 1987 qu'en 1980, l'année de référence. (Nous ne disposons pas d'estimation du niveau des rendements de l'éducation pour l'année de base pour les raisons invoquées précédemment.) L'autre terme d'interaction significatif est celui associé à *d86•educ* (coefficient = 0,027, $t = 2{,}23$). Les estimations relatives aux autres années n'apparaissent pas significatives au seuil de 5 % dans le cadre d'un test bilatéral. Si nous effectuons un test de significativité jointe de Fisher pour tous les sept termes d'interaction nous obtenons une p-valeur de 0,28 : cela illustre la possibilité d'obtenir un ensemble de variables non significatives alors que certaines d'entre elles se sont révélées l'être prises individuellement. [Les degrés de liberté (*ddl*) du test F sont de 7 et 3 799 ; le second terme provenant du calcul suivant : $N(T-1) - k = 545(8-1) - 16 = 3\,799$.] En général, les résultats sont cohérents avec l'hypothèse d'un accroissement des rendements de l'éducation sur la période considérée.

La régression sur variables indicatrices

Une approche traditionnelle du modèle à effets fixes consiste à faire l'hypothèse que l'effet inobservé, a_i, est un paramètre à estimer pour chaque unité individuelle i. Dès lors, dans l'équation (14.4), chaque a_i correspond à la constante spécifique pour l'individu i (ou l'entreprise i, la ville i, etc.). (Nous ne pouvons évidemment pas en faire de même dans le cadre d'une régression en coupe transversale : il y aurait alors $N + k$ paramètres à estimer avec seulement N observations. Nous avons besoin d'au moins deux périodes.) Pour estimer une constante individuelle par individu i nous devons introduire une constante pour chaque observation en coupe, en sus des variables explicatives (et probablement également des variables indicatrices pour chacune des dates). Cette méthode est appelée **régression sur variables indicatrices**. Même lorsque la dimension N est de taille modeste (mettons, $N = 54$ comme dans l'exemple 14.1), cela se traduit par de nombreuses variables explicatives – dans la plupart des cas, trop nombreuses pour permettre l'estimation du modèle de régression. De ce fait, cette méthode n'est pas très pratique pour les données de panel présentant une dimension en coupe importante.

Pour autant, les régressions sur variables indicatrices présentent certains aspects intéressants. Plus fondamentalement, elles nous donnent de façon *exacte* les estimations des coefficients β_j que nous aurions obtenues à partir de la régression sur les données en écart à leurs moyennes temporelles, les écarts-types et autres statistiques standard demeurant inchangées. De ce fait l'estimateur à effets fixes peut être obtenu à partir de la régression sur variables indicatrices. Un des avantages de cette méthode est qu'elle permet le calcul direct du nombre de degrés de liberté. Notons qu'il s'agit là d'un avantage mineur compte-tenu du fait que la plupart des logiciels d'économétrie disposent désormais des options nécessaires.

Le *R*-carré de la régression sur variables indicatrices est en général plus élevé. Cela provient du fait que nous introduisons autant de variables indicatrices que d'unités individuelles, expliquant ainsi une grande partie de la variabilité des données. Par exemple, si nous estimons le modèle à effets inobservés de l'exemple 13.8 au moyen de la régression sur variables indicatrices (possible avec $N = 22$), alors $R^2 = 0{,}933$. Cette valeur de *R*-carré ne devrait pas trop nous surprendre : il était attendu que nous expliquerions une grande part de la variation du nombre de déclarations de chômage en ayant recours à des variables indicatrices temporelles et géographiques. Tout comme dans l'exemple 13.8, l'estimation du coefficient de la variable indicatrice *EZ* est plus importante que la valeur du R^2.

Le *R*-carré issu de la régression sur variables indicatrices peut être utilisé pour calculer de façon standard la statistique de Fisher *F* en supposant, bien évidemment, que les hypothèses classiques du modèle de régression linéaire sont vérifiées (voir l'annexe de ce chapitre). Plus précisément, nous sommes en mesure de tester la significativité jointe de toutes les variables indicatrices individuelles ($N - 1$, puisque l'une d'entre elles a été choisie comme groupe de référence). Le *R*-carré non contraint est obtenu à partir de la régression réalisée sur l'ensemble des variables indicatrices individuelles ; alors que le *R*-carré contraint les ignore. Dans la grande majorité des cas, les variables indicatrices s'avèreront conjointement significatives.

De façon occasionnelle, les constantes estimées, soient les $\hat{\alpha}_i$, peuvent être intéressantes. C'est le cas lorsque nous souhaitons étudier la distribution des $\hat{\alpha}_i$ pour les différents individus *i*, ou à entreprises ou villes données, ou pour comparer son coefficient estimé à la valeur moyenne pour l'échantillon. Ces estimations sont directement accessibles à partir de la régression sur variables indicatrices mais sont rarement reportées lors de la mise en œuvre des routines des logiciels économétriques usuels (pour des raisons pratiques car les $\hat{\alpha}_i$ sont nombreux). Après avoir estimé le modèle à effets fixes, et quelle que soit la taille de *N* il est aisé de calculer les $\hat{\alpha}_i$ comme suit :

$$\hat{\alpha}_i = \bar{y}_i - \hat{\beta}_1\bar{x}_{i1} - \ldots - \hat{\beta}_k\bar{x}_{ik}, i = 1, \ldots, N, \qquad \textbf{[14.6]}$$

où les barres surmontant les variables désignent leurs moyennes temporelles et où les $\hat{\beta}_j$ sont les coefficients estimés du modèle à effets fixes. Par exemple, une fois estimé le modèle de criminalité, tout en ayant pris en compte l'influence d'un certain nombre de facteurs variant dans le temps, nous sommes en mesure d'obtenir une valeur $\hat{\alpha}_i$ pour une ville en particulier. Ces $\hat{\alpha}_i$ nous permettent de voir si les effets fixes inobservés contribuant à la criminalité sont supérieurs ou inférieurs à la moyenne.

Un certain nombre de logiciels économétriques proposant des routines pour estimer les modèles à effets fixes rapportent les résultats relativement à une « constante ». Cela peut prêter à confusion compte-tenu de notre remarque précédente : l'expression du modèle en écart à la moyenne empêche toute introduction de variable invariante dans le temps, dont la constante. [Voir équation (14.5).] Reporter les résultats relativement à une constante pour l'estimation du modèle à effets fixes (EF) dérive de la perception des éléments a_i comme autant de paramètres à estimer. En particulier, la constante reportée correspond à la moyenne des $\hat{\alpha}_i$ calculée pour les individus *i*. En d'autres termes, la constante globale du modèle correspond à la moyenne des effets fixes par individu et constitue un estimateur sans biais et convergent de $\alpha_i = \mathrm{E}(a_i)$.

Dans la plupart des études, les $\hat{\beta}_j$ sont dignes d'intérêt, et sont en général obtenus à partir de la régression sur les données exprimées en écart à leurs moyennes. De plus, il est courant de voir les a_i comme

des variables omises dont nous tenons compte au travers de la transformation *within*. Il est difficile de statuer quant à la meilleure approche pour estimer les a_i. En effet, même si $\hat{\alpha}_i$ est sans biais (sous les hypothèses EF.1 à EF.4 dans l'annexe de ce chapitre), il n'est pas convergent pour T fixé à mesure que $N \rightarrow \infty$. La raison en est qu'à mesure que nous ajoutons des observations individuelles, nous ajoutons de nouveaux a_i. Aucune information complémentaire ne vient étayer la connaissance de a_i lorsque T est fixé. Avec des T de plus grande dimension, nous pouvons obtenir de meilleures estimations des a_i, mais la plupart des données de panel présentent des structures caractérisées par N grand et T petit.

Effets fixes ou différences premières ?

Jusqu'à maintenant, si l'on met de côté la méthode des MCO sur les données empilées, nous avons passé en revue deux méthodes alternatives pour estimer les effets inobservés d'un modèle. L'une implique de différencier les données et l'autre, d'exprimer les variables en écart à leurs moyennes temporelles. Comment déterminer laquelle des deux méthodes est la plus appropriée ?

Nous pouvons éliminer un cas d'emblée, lorsque $T = 2$, les estimations par EF et DP, de même que les statistiques de tests, sont *identiques*, de ce fait, il importe peu de choisir une approche plutôt qu'une autre. Bien évidemment, l'équivalence entre les estimations à EF et en DP requiert que nous estimions le même modèle dans les deux cas. En particulier, comme discuté dans le cadre du chapitre 13, il est naturel d'inclure une constante dans le modèle à EF ; cette constante représente la deuxième période dans le modèle original lorsqu'il est écrit pour deux périodes. De ce fait, l'estimation EF doit inclure une variable indicatrice pour la deuxième période de façon à être rigoureusement identique à l'estimation du modèle en DP qui inclut une constante.

Avec $T = 2$, l'estimation à EF présente l'avantage d'être facile à mettre en œuvre, quel que soit le logiciel économétrique utilisé. Il est par ailleurs aisé de calculer des statistiques robustes à la présence d'hétéroscédasticité à l'issue de l'estimation en DP (puisque lorsque $T = 2$, l'estimation en DP s'assimile à une simple régression en coupe).

Lorsque $T \geq 3$, les estimateurs à EF et en DP diffèrent. Dans la mesure où les deux sont sans biais sous les hypothèses EF.1 à EF.4, nous ne pouvons utiliser l'absence de biais comme critère discriminant. De plus, les deux estimateurs sont convergents (pour T fixé et $N \rightarrow \infty$) sous EF.1 à EF.4. Pour N grand et T petit, le choix entre les deux estimateurs repose sur leur efficacité relative, déterminée par la corrélation sérielle du terme d'erreur idiosyncratique u_{it}. (Nous faisons l'hypothèse ici d'homoscédasticité des u_{it}, puisque les comparaisons d'efficacité requièrent cette propriété.)

Lorsque les u_{it} ne sont pas corrélées au cours du temps, l'estimateur à effets fixes est plus efficace que celui en différences premières (et les écarts-types estimés sont valides). Puisque le modèle à effets inobservés est spécifié avec des erreurs non corrélées au cours du temps (parfois seulement de façon implicite), l'estimateur à EF est plus utilisé que l'estimateur en DP. Nous devons toutefois garder à l'esprit que cela peut être faux. Dans de nombreuses applications, nous pouvons nous attendre à ce que les facteurs inobservés changent au cours du temps et soient de plus, sériellement corrélés. Si u_{it} suit une marche aléatoire – ce qui implique l'existence de niveaux de corrélation sérielle très forts et positifs – alors les différences premières Δu_{it} ne seront pas corrélées, et estimer le modèle en différences premières peut s'avérer plus judicieux. Dans de nombreux cas, le terme u_{it} exhibe une corrélation sérielle positive, mais sans doute plus faible que dans le cas d'un processus de marche aléatoire. Dans ce cas, nous ne pouvons aisément comparer l'efficacité des estimateurs par les méthodes à EF et en DP.

Il est difficile de tester si les u_{it} sont sériellement non corrélés à l'issue d'une estimation par EF : nous pouvons estimer les erreurs en écart à leur moyenne, $\ddot{u}_{it}$, mais pas les u_{it}. Néanmoins, dans la section 13.3, nous avons montré comment tester si les erreurs prises en différences premières étaient

non corrélées. Si cela semble être le cas, alors l'estimateur en DP peut être utilisé. Si l'on observe en revanche de la corrélation sérielle négative dans les Δu_{it}, l'estimateur à EF est probablement un meilleur estimateur. Il est souvent conseillé d'essayer les deux méthodes : si les résultats n'y sont pas sensibles, c'est encore mieux.

Lorsque T est grand, et plus particulièrement lorsque cette condition est couplée à N relativement petit (par exemple, $N = 20$ et $T = 30$), nous devons recourir à l'estimateur à EF avec prudence. Bien que les résultats relatifs aux distributions statistiques exactes des estimateurs et des statistiques de test tiennent quelles que soient les dimensions de N et T sous les hypothèses classiques du modèle à effets fixes, lorsque N est petit et T grand, l'inférence peut devenir très sensible à la violation de ces hypothèses. Plus spécifiquement, si nous recourons à des processus contenant une racine unitaire – voir chapitre 11 – le problème de régression fallacieuse peut surgir. Passer en différences premières présente l'avantage de stationnariser les séries temporelles en des processus à faible dépendance dans le temps. Dès lors, si nous avons recours au modèle en différences premières, nous pouvons avoir recours au théorème central limite même lorsque T est plus grand que N. La normalité des erreurs n'est pas requise, l'hétéroscédasticité et la corrélation sérielle peuvent être traitées comme dans le cadre du chapitre 13. L'inférence en présence d'un modèle à effets fixes est potentiellement plus sensible à la non normalité, l'hétéroscédasticité et la corrélation sérielle des erreurs idiosyncratiques.

À l'instar de l'estimateur en différences premières, l'estimateur à effets fixes peut être très sensible aux erreurs de mesure d'une ou de plusieurs variables explicatives. Néanmoins, si chacune des variables x_{itj} s'avère non corrélée avec u_{it}, mais que l'hypothèse stricte d'exogénéité est par ailleurs violée – par exemple, si la variable dépendante retardée est incluse parmi les variables explicatives du modèle ou s'il existe un effet de retour entre u_{it} et les futures réalisations de la variable explicative – alors l'estimateur à effets fixes présentera un biais bien moindre que l'estimateur en différences premières (sauf dans le cas où $T = 2$). Un résultat théorique d'importance tient au fait que le biais de l'estimateur en DP ne dépend pas de T, alors que celui à EF converge vers 0 à la vitesse $1/T$. Voir Wooldridge (2010, section 10.7) pour plus de détails.

En général, il est difficile de choisir entre les estimateurs à EF et en DP lorsqu'ils mènent à des résultats similaires. Il est alors courant de reporter les résultats issus des deux méthodes et de justifier pourquoi ceux-ci diffèrent.

Effets fixes sur des panels non cylindrés

Dans certains cas, et en particulier lorsque l'on étudie le comportement d'individus ou d'entreprises, il est courant que des données temporelles manquent pour certaines des unités individuelles considérées. Dans ce cas, nous qualifions la base de données de **panel non cylindré**. La mécanique d'estimation du modèle à effets fixes sur un panel non cylindré n'apparaît pas plus problématique que dans le cas d'un panel cylindré. Si T_i correspond au nombre de périodes pour l'unité individuelle i, nous utilisons simplement ces T_i observations lors de l'expression du modèle en écart aux valeurs moyennes. Le nombre total d'observations est alors de $T_1 + T_2 + \ldots + T_N$. Comme précédemment, un degré de liberté est perdu pour chaque observation individuelle, du fait de l'expression en écart à la moyenne. Tout logiciel d'économétrie proposant des routines pour l'estimation des effets fixes réalise les ajustements idoines. La régression sur variables indicatrices procède elle aussi à l'identique, et les degrés de liberté *ddl* sont obtenus de façon appropriée.

Il est aisé de voir que les unités pour lesquelles nous disposons d'information sur une seule période ne jouent aucun rôle dans l'analyse d'un modèle à effets fixes. Le passage en écart à la moyenne pour ces observations renvoie à zéro, elles ne sont alors pas utilisées dans l'estimation. (Si T_i est d'au plus deux pour tout i, nous pouvons avoir recours aux différences premières ; si $T_i = 1$ pour tout i, nous ne disposons pas du nombre minimal de périodes pour calculer des différences.)

La question la plus délicate dans le cadre des panels non cylindrés est de comprendre à quoi l'absence de données est due. Si l'on reprend l'exemple des villes et des États, pour certaines années, des données clés manquent. Si l'on fait l'hypothèse que les données manquantes constatées pour certaines unités individuelles i de notre panel ne sont pas corrélées avec le terme d'erreur idiosyncratique, u_{it}, le caractère non cylindré du panel ne pose pas de problème. Lorsque nous étudions des données relatives à des personnes, des familles ou des entreprises, les choses deviennent plus ardues. Imaginez, à titre d'exemple, que nous obtenions un échantillon aléatoire d'entreprises manufacturières en 1990, et que nous souhaitions tester l'incidence de la structure syndicale sur la rentabilité de la firme. Dans un monde idéal, nous pourrions avoir recours à une analyse en données de panel de façon à tenir compte de l'influence des caractéristiques inobservées relatives aux travailleurs ou au mode de gestion, celles-ci pouvant également être corrélées avec la proportion des employés syndiqués. Si nous collectons à nouveau des données pour les années ultérieures, il est possible qu'entre temps, certaines des entreprises aient disparu, soit parce qu'elles ont fait faillite soit en raison de fusions avec d'autres entreprises. Si c'est le cas, nous aurons alors affaire à un échantillon non aléatoire observé durant la période ultérieure. La question est alors la suivante : Si nous appliquons la méthode des effets fixes à un panel non cylindré, les estimateurs seront-ils bien sans biais (ou même convergents) ?

Si ce qui explique la disparition d'une entreprise de l'échantillon initial (un phénomène que l'on qualifie d'*attrition*) se trouve corrélé avec le terme d'erreur idiosyncratique – soient les facteurs inobservés variant dans le temps qui influencent le profit – alors nous faisons face à un problème de sélection d'échantillonnage qui a pour conséquence de biaiser les estimateurs (voir chapitre 9). Ce problème est un problème sérieux dans notre exemple. Néanmoins, une des caractéristiques utiles du modèle à effets fixes tient au fait qu'il *permet* à l'attrition d'être corrélée avec a_i, l'effet inobservé. L'idée est que compte-tenu de l'échantillon initial, certaines unités ont une plus grande probabilité que d'autres de disparaître de l'échantillon ; cela est capturé par le paramètre a_i.

EXEMPLE 14.3
De l'effet de la formation professionnelle sur le taux de rebut des entreprises

Nous ajoutons maintenant deux variables à l'analyse du tableau 14.1, soient $\log(sales_{it})$ et $\log(employ_{it})$, où *sales* correspond au chiffre d'affaire annuel des entreprises et *employ* au nombre total d'employés. Trois des 54 entreprises disparaissent complètement de l'analyse car ni leur chiffre d'affaire ni le nombre de leurs employés ne figurent dans les données. Cinq observations supplémentaires sont perdues du fait de données manquantes supplémentaires pour l'une ou l'autre de ces variables nous laissant avec $n = 148$. Recourir à l'estimation par effets fixes sur un panel non cylindré n'aura pas d'impact sur l'analyse d'ensemble, bien que l'effet estimé de la subvention soit plus élevé : $\hat{\beta}_{grant} = -0{,}297$, $t_{grant} = -1{,}89$; $\hat{\beta}_{grant-1} = -0{,}536$, $t_{grant-1} = -2{,}389$.

Résoudre le problème de l'attrition dans les données de panel s'avère en général complexe et dépasse le cadre de cet ouvrage. [Voir, par exemple, Wooldridge (2010, chapitre 19).]

14.2 MODÈLES À EFFETS ALÉATOIRES

Nous partons du même modèle à effets inobservés :

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \qquad \textbf{[14.7]}$$

dans lequel nous introduisons explicitement une constante, faisant ainsi l'hypothèse que l'effet inobservé, a_i, est de moyenne nulle (sans restreindre la portée générale du résultat). En temps normal, nous introduirions également des variables indicatrices temporelles parmi les variables explicatives. Le recours aux effets fixes ou aux différences premières a pour objet d'éliminer les a_i supposés être corrélés avec une ou plusieurs

des x_{itj}. Mais supposons que les a_i soient *décorrelés* de chacune des variables explicatives sur toute la période. Alors, recourir à cette transformation pour éliminer les a_i aura pour conséquence de générer des estimateurs inefficaces.

L'équation (14.7) devient un **modèle à effets aléatoires** lorsque l'on fait l'hypothèse que les effets inobservés a_i ne sont corrélés avec aucune des variables explicatives du modèle :

$$Cov(x_{itj}, a_i) = 0, t = 1, 2, \ldots, T ; j = 1, 2, \ldots, k. \quad \textbf{[14.8]}$$

Dans l'idéal, les hypothèses du modèle à effets aléatoires incluent toutes les hypothèses relatives au modèle à effets fixes plus la condition supplémentaire que les a_i sont indépendants de toutes les variables explicatives du modèle en tout temps. (Voir l'annexe de ce chapitre pour le détail des hypothèses retenues.) Si nous avons de bonnes raisons de penser que les effets inobservés a_i sont corrélés avec l'une ou plusieurs des variables explicatives, nous devrions alors plutôt recourir à la transformation en différences premières ou au modèle à effets fixes.

À partir du modèle (14.8) et suivant l'hypothèse des effets aléatoires, comment devrions nous nous y prendre pour estimer les paramètres β_j ? Il est important d'étudier si, dans la mesure où nous tablons sur l'hypothèse que a_i n'est pas corrélé avec les variables explicatives du modèle, les β_j sont estimés de façon convergente en ayant recours à une simple analyse en coupe : il n'est alors pas utile de recourir à des données de panel. Ceci étant, ne pas considérer les données de panel revient à ignorer une source d'information importante relativement aux autres périodes. Nous pourrions également utiliser l'intégralité de l'information dans le cadre d'une estimation par MCO sur les données empilées c'est-à-dire régresser y_{it} sur les variables explicatives et les variables indicatrices temporelles. Cette approche permet elle aussi d'obtenir des estimateurs convergents de β_j sous l'hypothèse d'effets aléatoires. Ce faisant, elle ignore certains des aspects essentiels de notre modèle. À supposer que nous définissions le **terme d'erreur composé** comme $v_{it} = a_i + u_{it}$ alors (14.7) peut être réécrit comme suit :

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + v_{it}. \quad \textbf{[14.9]}$$

Du fait que a_i est contenu dans le terme d'erreur composé à chaque période, les v_{it} sont sériellement corrélées au cours du temps. De fait, sous l'hypothèse d'erreurs aléatoires,

$$\text{Corr}(v_{it}, v_{is}) = \sigma_a^2 / (\sigma_a^2 + \sigma_u^2), t \neq s,$$

avec $\sigma_a^2 = \text{Var}(a_i)$ et $\sigma_u^2 = \text{Var}(u_{it})$. Cette corrélation (nécessairement) positive du terme d'erreur peut être potentiellement substantielle et, du fait que les écarts-types estimés issus des MCO sur données empilées ignorent cette corrélation, ils seront incorrects, de même que les statistiques de tests traditionnelles. Dans le chapitre 12, nous avions montré comment la méthode des moindres carrés généralisés (MCG) pouvait être utilisée en présence d'autocorrélation. Nous pouvons également avoir recours à cette technique dans le cas présent. Pour s'assurer des bonnes propriétés de la procédure, il convient de considérer des panels caractérisés par N grand et T relativement petit. Nous faisons l'hypothèse d'un panel cylindré, bien que cette méthode puisse être étendue au cas des panels non cylindrés.

Calculer la transformation des MCG qui permet d'éliminer la corrélation sérielle dans le terme d'erreur requiert des notions avancées d'algèbre linéaire [voir, à titre d'exemple, Wooldridge (2010, chapitre 10)]. Mais la transformation en tant que telle est plutôt simple. Soit

$$\theta = 1 - [\sigma_u^2 / (\sigma_u^2 + T\sigma_a^2)]^{1/2}, \quad \textbf{[14.10]}$$

qui est un paramètre borné entre zéro et un. L'expression du modèle transformé devient :

$$\begin{aligned} y_{it} - \theta\bar{y}_i = \beta_0(1 - \theta) + \beta_1(x_{it1} - \theta\bar{x}_{i1}) + \ldots \\ + \beta_k(x_{itk} - \theta\bar{x}_{ik}) + (v_{it} - \theta\bar{v}_i), \end{aligned} \quad \textbf{[14.11]}$$

où les éléments surmontés de barres désignent à nouveau les moyennes temporelles des variables correspondantes. Cette équation est intéressante car elle implique des variables exprimées **en quasi-écart à leur moyenne** pour chacune d'entre elles. L'estimateur à effets fixes soustrait aux variables leurs moyennes temporelles. La transformation du modèle menant à l'expression de l'estimateur à effets aléatoires repose sur la soustraction d'une fraction de ces moyennes temporelles, ces fractions dépendant de σ_u^2, σ_a^2 et du nombre de périodes T. L'estimateur des MCG correspond simplement à l'estimateur des MCO sur les données empilées de l'équation (14.11). Contre-toute attente, on peut montrer que les erreurs de l'équation (14.11) ne sont pas sériellement corrélées. (Voir exercice 3.)

La transformation présentée à l'équation (14.11) autorise la présence de variables explicatives constantes au cours du temps, et c'est là l'un des avantages du modèle à effets aléatoires (EA) sur celui à effets fixes ou en différences premières. Cette caractéristique du modèle vient du fait que le modèle à EA fait l'hypothèse que les effets inobservés ne sont pas corrélés avec les variables explicatives, qu'elles soient invariantes dans le temps ou non. Dès lors, dans une équation de salaire, nous pouvons inclure une variable telle que le niveau d'éducation même si celui-ci n'évolue pas au cours du temps. A noter que nous faisons l'hypothèse ici que le niveau d'éducation n'est pas corrélé avec a_i, qui contient à la fois l'habileté et le contexte familial des individus. Dans de nombreuses applications, la seule raison poussant à recourir aux données de panel est de permettre aux effets inobservés d'être corrélés avec les variables explicatives.

Le paramètre θ n'est pas connu en pratique, mais il peut toujours être estimé. Pour ce faire, il existe différentes approches telles que l'estimation par les MCO sur données empilées ou du modèle à effets fixes. En général, $\hat{\theta}$ prend la forme suivante $\hat{\theta} = 1 - \{1/[1 + T(\hat{\sigma}_a^2/\hat{\sigma}_u^2]\}^{1/2}$ avec $\hat{\sigma}_a^2$ un estimateur convergent de σ_a^2 et $\hat{\sigma}_u^2$ un estimateur convergent de σ_u^2. Ces estimateurs peuvent reposer sur l'évaluation des résidus issus de l'estimation par MCO sur données empilées ou de l'estimation du modèle à effets fixes. Une possibilité serait que $\hat{\sigma}_a^2 = \left[NT(T-1)/2 - (k+1)\right]^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it}\hat{v}_{is}$, avec $\hat{v}$ les résidus issus de l'estimation du modèle (14.9) par les MCO sur les données empilées. À partir de ce résultat, il est possible d'estimer σ_u^2 à partir de $\hat{\sigma}_u^2 = \hat{\sigma}_v^2 - \hat{\sigma}_a^2$, avec $\hat{\sigma}_v^2$, le carré de l'écart-type résiduel issu de la régression par les MCO sur les données empilées. [Voir Wooldridge (2010, chapitre 10) pour des discussions complémentaires sur le calcul de ces estimateurs.]

De nombreux logiciels d'économétrie permettent de réaliser l'estimation du modèle à effets aléatoires et le calcul automatique de certaines spécifications de $\hat{\theta}$. L'estimateur des moindres carrés quasi-généralisés (MCQG) qui utilise $\hat{\theta}$ à la place de θ est appelé **l'estimateur à effets aléatoires**. Sous les hypothèses associées au modèle à effets aléatoires nous montrons dans l'annexe de ce chapitre, que l'estimateur est convergent (mais pas sans biais) et asymptotiquement normal lorsque N devient grand pour T fixé. Les propriétés de l'estimateur à effets aléatoires (EA) avec N petit et T grand sont très largement inconnues bien que l'estimateur ait été très certainement utilisé dans ce contexte.

L'équation (14.11) nous permet de relier l'estimateur des EA à l'estimateur des MCO sur les données empilées (si $\theta = 0$) ainsi qu'à celui du modèle à effets fixes (si $\theta = 1$). En pratique, $\hat{\theta}$ n'atteint jamais les valeurs zéro ou un. Mais si $\hat{\theta}$ est proche de zéro, les estimations du modèle à effets aléatoires seront proches des résultats obtenus dans le cadre d'une estimation par les MCO sur les données empilées. C'est le cas quand l'effet inobservé, a_i, est relativement peu important (du fait de sa faible variance comparativement à σ_u^2). Il est plus courant pour σ_a^2 d'être relativement élevé comparativement à σ_u^2, impliquant que $\hat{\theta}$ sera plus proche de l'unité. À mesure que T devient grand, $\hat{\theta}$ tend vers un, rendant les estimations des modèles à EA et à EF très similaires.

Nous pouvons approfondir notre compréhension des mérites relatifs de ces deux estimateurs en écrivant l'erreur de l'équation (14.11) exprimée en quasi-écart à sa valeur moyenne comme suit :

$v_{it} - \theta\bar{v}_i = (1 - \theta)a_i + u_{it} - \theta\bar{u}_i$. Cette simple expression met en lumière le fait que les erreurs issues de l'équation transformée utilisée pour l'estimation du modèle à effets aléatoires pondèrent les effets inobservés par un facteur $(1 - \theta)$. Bien que la corrélation entre a_i et une ou plusieurs variables explicatives x_{itj} entraîne la non convergence de l'estimateur à effets aléatoires, nous voyons que la corrélation est atténuée par un facteur $(1 - \theta)$. À mesure que $\theta \rightarrow 1$, le biais tend vers zéro, ce qui est cohérent avec l'observation que l'estimateur à EA tend vers celui du modèle à EF. Si θ est proche de zéro, nous abandonnons une proportion plus grande de l'effet inobservé dans le terme d'erreur et par conséquent, le biais asymptotique de l'estimateur à EA sera plus important.

Dans les applications des modèles à EF et à EA, il est d'usage de faire apparaître les résultats d'estimation par les MCO sur les données empilées. Comparer ces trois jeux d'estimations permet de mieux comprendre la nature du biais engendré par l'écriture des effets inobservés, a_i, comme composante pleine et entière du terme d'erreur du modèle (à l'instar de l'estimation MCO sur les données empilées) ou partielle (comme dans le cadre de la transformation du modèle à EA). Mais nous devons garder à l'esprit que même lorsque a_i n'est corrélé avec aucune des variables explicatives du modèle en tout point du temps, les écarts-types estimés et tests statistiques issus de l'estimation par les MCO sur les données empilées ne sont en général pas valides puisqu'ils ignorent la corrélation sérielle potentiellement substantielle entre les erreurs composées, $v_{it} = a_i + u_{it}$. Comme mentionné dans le chapitre 13 (voir exemple 13.9), il est possible de calculer des écarts-types estimés ainsi que des valeurs de statistiques qui soient robustes à la corrélation sérielle (et l'hétéroscédasticité) de v_{it}. Un certain nombre de logiciels d'économétrie standard proposent ce type d'option. [Voir, par exemple, Wooldridge (2010, chapitre 10).]

EXEMPLE 14.4
Une équation de salaire estimée sur données de panel

Nous utilisons à nouveau les données contenues dans le fichier WAGEPAN pour estimer une équation de salaire pour les hommes. Nous avons recours à trois méthodes : les MCO sur les données empilées, le modèle à effets aléatoires, et le modèle à effets fixes. Pour les deux premières méthodes, nous incorporons la variable *educ*, ainsi que les variables indicatrices relatives aux origines ethniques *black* et *hispan* ; elles sont en revanche retirées du modèle à effets fixes. Les variables variantes dans le temps sont *exper*, $exper^2$, *union*, et *married*. Comme étudié dans la section 14.1, *exper* est retiré de l'analyse du modèle à effets fixes (bien que $exper^2$ soit maintenu). Chacune des régressions contient un ensemble de variables indicatrices temporelles. Les résultats d'estimation sont reportés dans le tableau 14.2.

Les coefficients associés aux variables *educ*, *black*, et *hispan* sont similaires pour les estimations par les MCO sur les données empilées et le modèle à effets aléatoire. Les écarts-types estimés par les MCO sur les données empilées correspondent à la formulation standard, ils sous-estiment la vraie valeur des écarts-types puisqu'ils ignorent la corrélation sérielle positive ; nous les reportons donc ici uniquement à titre de comparaison. Le profil de la variable expérience est quelque peu différent ; les coefficients des primes relatives au mariage et à l'appartenance à un syndicat chutent tous deux de façon drastique dans les estimations du modèle à effets aléatoires. Lorsque nous éliminons complètement l'effet inobservé en introduisant des effets fixes, la prime relative au mariage chute d'environ 4,7 %, bien qu'elle demeure statistiquement significative. Cette chute de la prime relative au mariage est cohérente avec l'idée que les hommes les plus susceptibles de recevoir de meilleurs salaires – comme cela est capturé par des effets inobservés plus élevés, a_i – sont aussi les plus susceptible d'être mariés. De ce fait, les résultats d'estimation du modèle par les MCO sur les données empilées illustrent qu'une grande partie de la prime de mariage reflète le fait que les hommes mariés gagneraient plus même s'ils n'étaient pas mariés. Les 4,7 % restants peuvent être interprétés de deux façons différentes : (1) le mariage rend réellement les hommes plus productifs (2) les employeurs paient aux hommes mariés une prime car le mariage est un gage de stabilité. Nous ne sommes pas en mesure de trancher entre ces deux hypothèses.

L'estimation de θ dans le cadre du modèle à effets aléatoires est de $\hat{\theta} = 0{,}643$, ce qui permet d'expliquer pourquoi, pour les variables présentant de la variabilité dans le temps, les estimations du modèle à effets aléatoires demeurent plus proches de celles du modèle à effets fixes que celles des MCO sur les données empilées.

Pour aller plus loin 14.3

La prime liée à l'appartenance syndicale estimée dans le modèle à effets fixes est d'environ 10 points de pourcentage plus basse que la prime estimée par les MCO sur les données empilées. Que cela suggère-t-il sur la corrélation entre *union* et l'effet inobservé ?

Tableau 14.2 *Trois estimateurs différents de l'équation de salaire*

Variable dépendante : log(*wage*)			
Variables indépendantes	**MCO sur les données empilées**	**Effets aléatoires**	**Effets fixes**
Educ	0,091 (0,005)	0,092 (0,011)	— — —
Black	−0,139 (0,024)	−0,139 (0,048)	— — —
Hispan	0,016 (0,021)	0,022 (0,043)	— — —
Exper	0,067 (0,014)	0,106 (0,015)	— — —
$Exper^2$	−0,0024 (0,0008)	−0,0047 (0,0007)	−0,0052 (0,0007)
Married	0,108 (0,016)	0,064 (0,017)	0,047 (0,018)
Union (appartenance à un syndicat)	0,182 (0,017)	0,106 (0,018)	0,080 (0,019)

Effets aléatoires ou effets fixes ?

Contrairement au modèle à effets aléatoires, le modèle à effets fixes permet un lien de corrélation non nul entre les a_i et les x_{itj}. De ce fait, le modèle à EF est considéré comme étant le cadre d'analyse le plus performant pour mesurer l'influence de variables explicatives, en raisonnant toutes choses égales par ailleurs. Pour autant, le modèle à effets aléatoires peut être utile dans certaines circonstances. Il l'est de façon très claire lorsque les variables explicatives clés du modèle sont invariantes dans le temps. Nous ne pouvons plus alors avoir recours au modèle à EF pour estimer leurs effets sur y. Par exemple, dans le tableau 14.2, nous devons recourir au modèle à EA (ou à l'estimation par les MCO sur les données empilées) pour estimer les rendements de l'éducation. Bien évidemment, dans ce cas précis, nous devons nous tourner vers

le modèle à EA et faisons alors l'hypothèse que l'effet inobservé n'est pas corrélé avec l'ensemble des variables explicatives. Ce faisant, il est recommandé d'introduire dans le modèle de régression autant de variables de contrôle invariantes dans le temps que possible. (Avec une analyse en effets fixes, cela n'est pas nécessaire.) L'estimation du modèle à EA est préférée aux MCO sur les données empilées car l'estimateur à EA est en général plus efficace.

Considérons maintenant le cas de variables explicatives variantes dans le temps. Y a-t-il des situations où le recours au modèle à effets aléatoires plutôt qu'aux effets fixes se justifie ? Oui : dans des situations où $\text{Cov}(x_{itj}, a_i) = 0$; mais celles-ci sont l'exception plutôt que la règle. À supposer que les réalisations de la variable explicative clé sont issues d'un protocole expérimental – par exemple, chaque année, les enfants sont assignés aléatoirement dans des classes de tailles différentes – alors le modèle à effets aléatoires paraît approprié pour estimer les effets de la taille de la classe sur la performance scolaire. Malheureusement, dans la plupart des cas, les régresseurs sont issus d'un processus vraisemblablement corrélé avec les préférences individuelles et avec le degré d'habileté des individus, tous deux capturés par la composante inobservée a_i.

Il est assez courant d'utiliser les deux méthodes d'estimation et de tester formellement la différence de significativité statistique des coefficients estimés des variables variantes dans le temps. (Dans le tableau 14.2, nous nous concentrerions alors sur les coefficients de *exper*^{2}, *married*, et *union*.) Hausman (1978) a été le premier à proposer un tel test. De nombreux logiciels d'économétrie proposent des routines permettant de calculer la statistique de test sous l'hypothèse nulle de validité des hypothèses du modèle à effets aléatoires – hypothèses listées dans l'annexe de ce chapitre. Le test repose sur l'idée que le modélisateur a recours au modèle à effets aléatoires sauf si le test d'Hausman rejette la condition (14.8). En pratique, l'absence de rejet de l'hypothèse nulle indique que les estimations des modèles EA et EF sont suffisamment proches pour être utilisés indifféremment, ou bien que l'incertitude entourant l'estimation des paramètres du modèle à EF est si grande que l'on ne peut conclure en pratique à des différences statistiques significatives entre les deux approches. Dans ce dernier cas, il semble légitime de se demander si l'on bénéficie de suffisamment d'informations pour fournir des estimations précises des coefficients. Le rejet du test d'Hausman est dans la pratique utilisé pour justifier l'invalidité de l'hypothèse centrale du modèle à EA, (14.8), et par suite, justifier le recours aux EF. (Naturellement comme dans toutes applications de statistique inférentielle, il convient de faire le distinguo entre significativité pratique et significativité statistique.) Wooldridge (2010, chapitre 10) élabore une discussion approfondie sur le sujet. Dans les sections qui suivent, nous discutons d'une modélisation alternative permettant de choisir entre les modèles à EA et à EF et d'utilisation plus simple.

En guise de dernière mise en garde, notons que dans les travaux empiriques, il n'est pas rare que les auteurs justifient le recours à l'un des deux modèles, EF plutôt que EA, suivant que les a_i sont perçus comme des paramètres à estimer ou des variables aléatoires. De telles considérations me paraissent résulter d'un problème mal posé. Dans ce chapitre, nous avons traité les a_i comme des variables aléatoires (14.7), indépendamment de la manière dont nous proposons d'estimer les β_j. Comme souligné précédemment, l'élément clé qui va déterminer le choix de la méthode d'estimation tient à la validité de l'hypothèse selon laquelle les a_i ne sont pas corrélés avec l'ensemble des x_{itj}. Néanmoins, dans certaines applications en données de panel, nous ne pouvons considérer notre échantillon comme un tirage purement aléatoire d'une population plus large, en particulier lorsque l'unité d'observation correspond à une entité géographique large (par exemple des États ou des provinces). Dans ce contexte, il est raisonnable de considérer les a_i comme des constantes distinctes à estimer pour chacune des unités individuelles. Dans ce cas, nous avons recours au modèle à EF : rappelez-vous que l'usage du modèle à EF implique automatiquement de considérer une constante spécifique par unité individuelle. Heureusement, sans s'engager dans un débat philosophique sur la nature exacte des a_i, le modèle à EF s'avère la plupart du temps plus convaincant que celui à EA pour l'évaluation de politiques économiques sur données agrégées.

14.3 LE MODÈLE À EFFETS ALÉATOIRES CORRÉLÉS

Dans certaines applications, il apparaît légitime d'assimiler les a_i (les effets inobservés) à des variables aléatoires à l'instar des autres variables du modèle. Ainsi, il existe une alternative aux effets fixes permettant aux a_i d'être corrélés aux variables explicatives. Pour présenter cette approche, considérons à nouveau le modèle simple présenté dans l'équation (14.1), où intervient une unique variable explicative x_{it}. Plutôt que de faire l'hypothèse que a_i est décorrélé des $\{x_{it} : t = 1, 2, \ldots, T\}$ – qui correspond à l'hypothèse du modèle à effets aléatoires – ou de prendre les moyennes temporelles pour éliminer a_i – approche par effets fixes – nous pourrions modéliser le lien de corrélation entre a_i et $\{x_{it} : t = 1, 2, \ldots, T\}$. Comme a_i est, par définition, constant au cours du temps, il peut être corrélé avec la valeur moyenne de x_{it} au cours du temps. Plus précisément, soit $\bar{x}_i = T^{-1}\sum_{t=1}^{T} x_{it}$ la moyenne des x_{it} au cours du temps. Faisons l'hypothèse d'un lien linéaire simple du type :

$$a_i = \alpha + \gamma\bar{x}_i + r_i, \qquad \textbf{[14.12]}$$

r_i n'étant corrélé à aucune des variables x_{it}. Du fait que $\bar{x}_i$ est une fonction linéaire des x_{it},

$$Cov(\bar{x}_i, r_i) = 0. \qquad \textbf{[14.13]}$$

Les équations (14.12) et (14.13) impliquent que a_i et $\bar{x}_i$ sont corrélés pour tout $\gamma \neq 0$.

Le **modèle à effets aléatoires corrélés** (EAC) combine les équations (14.12) et (14.1). En substituant la première à la deuxième on obtient :

$$y_{it} = \beta x_{it} + \alpha + \gamma\bar{x}_i + r_i + u_{it} = \alpha + \beta x_{it} + \gamma\bar{x}_i + r_i + u_{it}. \qquad \textbf{[14.14]}$$

L'équation (14.14) est intéressante car elle comprend toujours un terme d'erreur composé, $r_i + u_{it}$, qui consiste en une composante invariante dans le temps r_i ainsi qu'un choc idiosyncratique u_{it}. Plus fondamentalement, l'hypothèse décrite dans (14.8) tient lorsque nous remplaçons a_i par r_i. De plus, du fait que u_{it} est supposé être décorrélé de x_{is}, pour tout s et tout t, u_{it} est également décorrélé de $\bar{x}_i$. Toutes ces hypothèses se combinent et permettent l'estimation du modèle suivant :

$$y_{it} = \alpha + \beta x_{it} + \gamma\bar{x}_i + r_i + u_{it}, \qquad \textbf{[14.15]}$$

qui ressemble en tous points à l'équation du modèle à effets aléatoires, à ceci près qu'apparaît la moyenne temporelle de la variable explicative, $\bar{x}_i$. C'est par son introduction dans le modèle que l'on tient compte de l'influence de la corrélation entre a_i et la séquence des $\{x_{it} : t = 1, 2, \ldots, T\}$. Ce qui reste est capturé par le terme r_i qui est décorrélé des x_{it}.

Dans la plupart des logiciels d'économétrie, il est aisé de calculer les moyennes temporelles par unité individuelle pour chacune des variables, $\bar{x}_i$. Si l'on fait l'hypothèse que l'on dispose de tels calculs pour chaque unité i, que s'attend-on à obtenir en estimant le modèle à effets aléatoires transformé tel qu'écrit dans (14.15) ? Notons que l'estimation de (14.15) nous donne les valeurs de $\hat{\alpha}_{EAC}$, $\hat{\beta}_{EAC}$, et $\hat{\gamma}_{EAC}$ – soient les estimateurs à EAC. Le résultat est quelque peu décevant puisqu'il est possible de montrer – voir par exemple, Wooldridge (2010, chapitre 10) – que

$$\hat{\beta}_{EAC} = \hat{\beta}_{EF} \qquad \textbf{[14.16]}$$

où $\hat{\beta}_{EF}$ désigne l'estimateur à EF issu de l'équation (14.3). En d'autres termes, ajouter la moyenne temporelle $\bar{x}_i$ et utiliser l'estimateur à EA revient à retirer les moyennes temporelles et estimer le modèle par les MCO sur les données empilées.

Même si l'équation (14.15) n'est pas requise pour dériver l'expression de $\hat{\beta}_{EF}$, l'équivalence entre les estimations à EAC et à EF lui assure une interprétation élégante : $\hat{\beta}_{EF}$ tient compte de l'influence du niveau

moyen $\bar{x}_i$, tout en mesurant l'impact marginal de x_{it} sur y_{it}. En guise d'exemple, supposons que x_{it} représente le taux d'imposition applicable aux profits des entreprises d'un pays i à l'année t, et y_{it} une certaine mesure de l'activité économique dudit pays. En introduisant $\bar{x}_i$, la valeur moyenne du taux d'imposition en vigueur dans le pays durant les T années, nous prenons en compte les différences systématiques entre les pays pratiquant historiquement des taux de taxation très élevés et les pays plus accommodants vis-à-vis des entreprises – différences qui peuvent également avoir un impact sur les performances économiques.

Nous pouvons également reprendre l'équation (14.15) pour voir si les estimateurs à EF sont moins précis que les estimateurs à EA. Si nous fixons $\gamma = 0$ dans l'équation (14.15) nous obtenons alors l'estimateur à EA traditionnel de β soit : $\hat{\beta}_{EA}$. Cela implique que la corrélation entre x_{it} et $\bar{x}_i$ n'a aucune incidence sur l'estimateur à EA. À l'inverse, nous savons, depuis le chapitre 3 consacré à l'analyse de la régression multiple, que la corrélation entre x_{it} et $\bar{x}_i$ – c'est-à-dire la multicolinéarité – peut engendrer une variance plus élevée pour $\hat{\beta}_{EF}$. Parfois, la variance peut être bien plus élevée, en particulier lorsque x_{it} fluctue peu dans le temps, auquel cas x_{it} et $\bar{x}_i$ tendent à être très fortement corrélées. Dans le cas limite d'absence de variation au cours du temps pour tout i, la corrélation est parfaite – et l'approche par EF ne peut fournir un bon estimateur de β.

Au-delà de l'éclairage que cette approche peut apporter dans notre compréhension des estimateurs à EF et à EA, y a-t-il d'autres motifs de considérer l'approche à EAC comme pertinente, alors même qu'elle délivre la même information qu'une estimation à EF de β ? Oui, et il en existe au moins deux. Le premier tient au fait que l'estimateur à EAC repose sur une approche simple et formelle permettant de choisir entre les deux estimateurs standard que sont les estimateurs à EF et à EA. Comme discuté précédemment, l'estimateur à EA repose sur la condition que $\gamma = 0$ alors que le modèle à EF estime la valeur du paramètre γ. À partir de $\hat{\gamma}_{EAC}$ et de son écart-type estimé [obtenu à partir de l'estimation du modèle à EA de (14.15)], nous pouvons construire un test t sous l'hypothèse $H_0 : \gamma = 0$ contre $H_1 : \gamma \neq 0$. [L'annexe explique comment faire en sorte que ce test soit robuste à l'hétéroscédasticité et à la corrélation sérielle de $\{u_{it}\}$.] Si nous rejetons H_0 à un niveau de confiance suffisamment fort, alors nous rejetons également la spécification à EA en faveur de celle à EF. Comme toujours, notamment lorsque la taille de la dimension individuelle de l'échantillon est très élevée, il est important de ne pas confondre significativité statistique et intérêt économique.

L'intérêt de ce modèle tient également au fait qu'il introduit des variables invariantes dans le temps dans ce qui ressemble à un modèle à effets fixes. Par exemple, soit z_i la variable invariante dans le temps – cela pourrait être le genre ou bien encore un résultat obtenu à un test de QI dans l'enfance. Nous pourrions aisément trouver des arguments pour inclure z_i à l'équation (14.15) :

$$y_{it} = \alpha + \beta x_{it} + \gamma \bar{x}_i + \delta z_i + r_i + u_{it}, \qquad \textbf{[14.17]}$$

la notation relative au terme d'erreur demeure inchangée alors même que celle-ci n'inclut plus la composante z_i. Si nous estimons ce modèle comme un modèle à EF, il peut être montré que les propriétés de β sont les mêmes que celles de l'estimateur à EF de l'équation (14.1). En fait, une fois que nous incluons $\bar{x}_i$, nous pouvons inclure n'importe quelle variable invariante dans le temps, l'estimer par EA et obtenir $\hat{\beta}_{EF}$, le coefficient associé à x_{it}. De plus, nous obtenons une estimation de σ, bien qu'elle doive être interprétée avec prudence puisqu'elle ne traduit pas nécessairement l'effet causal de z_i sur y_{it}.

Une stratégie similaire au modèle à EAC peut être appliquée avec un grand nombre de variables explicatives variantes dans le temps (ou non variantes dans le temps). Lorsque l'équation augmentée des valeurs moyennes temporelles est estimée par EA, les coefficients des variables variantes dans le temps sont identiques à ceux d'une estimation par EF. À noter que lorsque le panel est cylindré il n'est pas utile d'inclure des moyennes temporelles des variables variantes dans le temps – et en particulier les variables indicatrices temporelles. (Avec T périodes, la moyenne temporelle d'une période donnée est de $1/T$, elle est donc constante pour tout i et tout t. Cela n'a donc aucun sens d'ajouter un ensemble d'éléments constants à une équation qui dispose d'ores et déjà d'une constante). Si la structure du panel est non cylindrée, alors les

moyennes des variables telles que les variables indicatrices temporelles peuvent changer selon la dimension individuelle considérée i – leur valeur dépendra du nombre d'observations temporelles dont nous disposons par unité individuelle i. Dans ce cas, les moyennes temporelles des variables changeant au cours du temps doivent être incluses.

L'exercice sur ordinateur 14 de ce chapitre illustre de quelle manière l'approche à EAC peut être appliquée au cas d'un panel cylindré comme celui décrit dans la base de données AIRFARE, et comment il est possible de tester la pertinence du modèle à EF par rapport au modèle à EAC.

Panels non cylindrés

Le modèle à EAC peut être estimé sur un panel non cylindré, mais requiert un traitement minutieux. Pour obtenir un estimateur qui reproduise les estimations à effets fixes sur les variables explicatives variantes dans le temps, il convient d'être très prudent lors du calcul des moyennes temporelles. En particulier, pour y ou tout x_j, une période temporelle particulière contribue à la moyenne, $\bar{y}_i$ ou $\bar{x}_{ij}$, si et seulement si des données sont observées pour l'ensemble des variables $(y_{it}, x_{it1}, \ldots, x_{itk})$ à cette date. Une façon de caractériser la situation consiste à définir une variable catégorielle, s_{it}, qui vaut 1 si des données sont bien observées pour l'ensemble des variables en t. Si un élément est manquant pour l'une des variables du modèle (ce qui inclut bien sûr le cas où aucune donnée n'est disponible en t), alors $s_{it} = 0$. (La notion d'indicateur de sélection est discutée plus avant dans le Chapitre 17.) Munis de cette définition, la formulation exacte de la moyenne temporelle de $\{y_{it}\}$ devient :

$$\bar{y}_i = T_i^{-1} \sum_{t=1}^{T} s_{it} y_{it}$$

avec T_i le nombre total de périodes temporelles pour lesquelles nous disposons de l'information complète pour l'individu i. En d'autres termes, nous ne calculons ici la moyenne temporelle que sur les périodes pour lesquelles nous disposons de données complètes.

Un autre point de subtilité concerne l'inclusion de variables muettes (*dummies*) temporelles, ou toute autre variable variant dans le temps, mais pas selon la dimension inter-individuelle, et le fait que nous devions également inclure leur moyenne temporelle (contrairement au cas des panels cylindrés pour lesquels ces moyennes étaient des constantes). À titre d'exemple, si $\{w_t : t = 1, \ldots, T\}$ est une série temporelle agrégée, au même titre qu'une dummy temporelle ou une tendance linéaire temporelle, alors :

$$\bar{w}_i = T_i^{-1} \sum_{t=1}^{T} s_{it} w_t$$

Du fait de la nature non cylindrée du panel, la valeur de $\bar{w}_i$ varie la plupart du temps selon la dimension inter-individuelle (à moins que les périodes temporelles non renseignées ne soient exactement les mêmes pour tous les individus du panel). De la même façon que pour les variables changeant selon les deux dimensions i et t, les moyennes temporelles des effets temporels agrégés sont facilement calculables à partir des routines pré-programmées disponibles couramment dans la plupart des logiciels d'économétrie.

Dans ce contexte, la dérivation formelle de l'estimateur est modifiée dans le cas d'un panel cylindré, que nous utilisions l'estimateur traditionnel à EA ou la version amendée à EAC. Plus précisément, le paramètre θ défini à l'équation (14.10) et que l'on utilise ensuite dans l'équation (14.11) pour obtenir les données exprimées en quasi-écart à leur moyenne, dépend de i au travers du nombre de période temporelles observées par unité individuelle i. Il suffit alors de remplacer T dans l'équation (14.10) par T_i. Les logiciels économétriques qui permettent l'estimation de modèles à effets aléatoires tiennent compte de cette différence lors de l'implémentation de la routine sur un panel non cylindré, de sorte qu'il n'est pas nécessaire d'opérer un quelconque ajustement du point de vue de l'utilisateur.

De façon générale, il convient de retenir qu'une fois les moyennes temporelles correctement calculées, utiliser une équation telle que celle décrite en (14.17) revient au même dans les cas cylindré et non cylindré. Nous pouvons toujours mettre en œuvre un test de significativité sur l'ensemble des moyennes temporelles de façon à choisir la spécification à EF ou EA « stricts », l'approche à EAC nous laissant la possibilité d'inclure des variables non variantes dans le temps.

Comme dans le cadre de l'estimation du modèle à effets fixes, une question clé tient à la compréhension des origines de la dimension non cylindrée du panel. Dans le cas d'un modèle à effets aléatoires « stricts », l'indicateur de sélection, s_{it}, ne peut être corrélé avec le terme d'erreur composé du modèle décrit dans l'équation (14.7), $a_i + u_{it}$, quelque soit la période t considérée. Sans cela, comme discuté dans Wooldridge (2010, Chapitre 19), l'estimateur à effets aléatoires n'est plus convergent. Comme évoqué en Section 14-1, l'estimateur du modèle à effets fixes permet, quant à lui, une corrélation arbitraire entre l'indicateur de sélection, s_{it}, et l'effet fixe, a_i. Dès lors, l'estimateur du modèle à effets fixes est plus robuste dans le contexte des panels non cylindrés. Par ailleurs, il est bon de rappeler que l'estimateur du modèle à effets fixes autorise une corrélation arbitraire entre les variables variantes dans le temps et a_i.

14.4 APPLIQUER LES TECHNIQUES DE DONNÉES DE PANEL À D'AUTRES STRUCTURES DE DONNÉES

Les différentes méthodes de données de panel peuvent être appliquées dans des contextes n'impliquant pas nécessairement l'observation de comportements au cours du temps. Par exemple, il est courant en démographie de considérer des fratries (parfois des jumeaux) pour tenir compte des caractéristiques inobservées liées à la structure familiale ou aux antécédents familiaux. En général, l'objectif est de mettre en lumière un « effet famille », qui serait commun aux membres d'une même fratrie et corrélé aux autres variables explicatives du modèle. Si les valeurs prises par les variables explicatives peuvent varier selon les membres d'une même fratrie, on préférera alors considérer toutes les différences prises deux à deux entre les membres de la fratrie – ou plus généralement, recourir à la transformation *within* au sein d'une famille pour estimer le modèle. En éliminant l'effet inobservé, on s'affranchit du biais potentiel causé par les caractéristiques communes aux membres d'une même famille. La plupart des logiciels économétriques estiment des modèles à effets fixes à partir de données semblables sans difficultés majeures.

À titre d'exemple, Geronimus et Korenman (1992) ont utilisé des données relatives à des couples de sœurs pour étudier dans quelle mesure une grossesse précoce affectait les revenus futurs. Si l'on considère le revenu dont un ménage a besoin, – qui dépend du nombre d'enfants – le modèle est donné par :

$$\log(incneeds_{fs}) = \beta_0 + \delta_0 sister2_s + \beta_1 teenbrth_{fs} + \beta_2 age_{fs} + other\ factors + a_f + u_{fs}, \quad \textbf{[14.18]}$$

avec f indiquant la famille et s la sœur considérée. La constante pour la première sœur est donnée par β_0, et celle pour la seconde par $\beta_0 + \delta_0$. La variable d'intérêt est $teenbrth_{fs}$, une variable binaire égale à un si la sœur s de la famille f a eu un enfant à l'adolescence. La variable age_{fs} correspond à l'âge de la sœur s dans la famille f ; Geronimus et Korenman ont également recours à d'autres variables de contrôle. L'élément non observé a_f, qui change seulement selon la famille, est un *effet familial inobservé* ou *effet fixe « famille »*. La question principale est de savoir si la variable *teenbrth* est corrélée avec l'effet « famille ». Si oui, alors les MCO sur les données empilées donneront un estimateur biaisé de l'impact d'une grossesse précoce sur les revenus futurs. Résoudre ce problème est relativement simple : au sein de chaque famille, on considère l'équation (14.18) pour chaque sœur et on en fait la différence. On obtient alors :

$$\Delta\log(incneeds) = \delta_0 + \beta_1 \Delta teenbrth + \beta_2 \Delta age + \ldots + \Delta u\ ; \quad \textbf{[14.19]}$$

ce qui permet d'éliminer l'effet « famille », a_f. L'équation obtenue peut alors être estimée par la méthode des MCO. Notons qu'il n'y a aucun élément temporel ici : la différenciation s'effectue entre sœurs. De même, il est possible que la constante de l'équation (14.18) ne soit pas la même pour les deux sœurs, ce qui a pour conséquence une constante non nulle dans l'équation issue de la différenciation (14.19). Si lors de l'encodage des données les sœurs sont classées de manière aléatoire, la constante estimée devrait être proche de zéro. Mais même dans un tel cas, il n'est pas dommageable d'inclure une constante dans l'équation (14.19). Cela peut permettre par exemple à la première sœur listée dans la base de données d'être celle qui se trouve dans la situation la plus défavorable.

Pour aller plus loin 14.4

Lorsque l'on a recours à la méthode des différences (décrite plus haut), cela a-t-il du sens d'inclure des variables indicatrices relatives aux origines ethniques du père et de la mère dans l'équation (14.18) ? Justifiez.

À partir des données relatives aux 129 couples de sœurs tirées de l'enquête datant de 1982 intitulée « National Longitudinal Survey of Young Women », Geronimus et Korenman ont d'abord estimé β_1 par la méthode des MCO sur les données empilées. Ils ont obtenu les valeurs de $-0{,}33$ et de $-0{,}26$; la seconde estimation vient d'une régression où a été prise en compte l'influence du contexte familial (avec des variables de contrôle comme le niveau d'éducation des parents) ; ces deux estimations sont toutes deux significatives [voir tableau 3 dans Geronimus et Korenman (1992)]. Il ressort de cette analyse que la maternité précoce a un impact important sur le niveau des revenus futurs d'une famille. Cependant, lorsque l'équation en différences premières est estimée, le coefficient de la variable *teenbrth* n'est plus que de $-0{,}08$, et n'apparaît pas statistiquement significatif. Cela suggère que c'est avant tout le contexte familial des femmes qui influence leur niveau de revenus futurs, et non la grossesse précoce en elle-même.

Geronimus et Korenman ont étudié d'autres dimensions pouvant être affectées par une grossesse précoce ; ils ont aussi utilisé deux autres bases de données ; dans certains cas, les estimations *within* au sein de la famille étaient économiquement importante et statistiquement significatives. Ils ont également montré que ces effets disparaissaient complètement lorsque les niveaux d'éducation des sœurs étaient pris en compte dans l'analyse.

Ashenfelter et Krueger (1994) ont estimé les rendements de l'éducation en utilisant un modèle en différences premières. Ils ont préalablement obtenu des données relatives à 149 sœurs jumelles[2] et collecté des informations sur leurs revenus, leurs niveaux d'éducation ainsi que d'autres variables. S'ils se concentrent sur de vrais jumelles, c'est qu'ils font l'hypothèse que les vrais jumeaux disposent des même niveaux d'habileté initiaux, et qu'on peut donc les éliminer en considérant la différence entre deux sœurs jumelles plutôt qu'en recourant à une estimation par les MCO sur les données empilées. Des sœurs jumelles ont le même âge, le même sexe, les mêmes origines ethniques, ces facteurs disparaissent donc de l'équation lorsque le modèle est exprimé en différences premières. De ce fait, Ashenfelter et Krueger régressent la différence de la variable log(*earnings*) sur la différence des niveaux d'éducation. Ils estiment le rendement de l'éducation à environ 9,2 % (t = 3,83). Il est à noter que ce chiffre est *plus important* que celui issu de l'estimation par les MCO sur les données empilées (qui était pour mémoire de 8,4 %, ce en tenant compte du genre, de l'âge et des origines ethniques). Ashenfelter et Krueger ont de plus estimé l'équation du modèle à effets aléatoires et ont obtenu le chiffre de 8,7 % pour les rendements de l'éducation. (Voir tableau 5 de leur article.) L'analyse du modèle à effets aléatoires est ici similaire au cas d'un panel à deux périodes.

2 L'échantillon est composé de vraies jumelles (homozygotes)

Les échantillons utilisés par Geronimus et Korenman (1992) puis Ashenfelter et Krueger (1994) sont des exemples d'**échantillons appariés**. En général, les méthodes des effets fixes et effets aléatoires peuvent être appliquées à des **échantillons en grappes**. Ceux-ci correspondent à des données en coupe transversale où chaque observation appartient à une « grappe » bien définie – ou encore à un même « groupe », ou *cluster* en anglais. Dans les exemples précédents, chaque famille peut être considérée comme une grappe. Prenons un autre exemple et supposons que nous observions des données relatives à la participation à un plan d'épargne retraite, où les entreprises offrent plusieurs contrats possibles. Nous pouvons considérer chaque entreprise comme une grappe. Dans ce contexte, il est clair que les effets inobservés spécifiques à chaque entreprise seront un facteur important pour expliquer les taux de participation dans les différents plans d'épargne proposés.

Considérons un autre exemple, et supposons que nous nous intéressions maintenant à la modélisation des décisions relatives au placement de l'épargne retraite. Il serait possible d'obtenir un échantillon aléatoire d'individus actifs – par ex. tous installés aux États-Unis – mais il est également courant d'échantillonner des entreprises à partir d'une population d'entreprises. Une fois ces entreprises échantillonnées, il serait aisé de collecter l'information relative à l'ensemble de leurs salariés ou à un sous-ensemble d'entre eux au sein de chacune des firmes. Dans chacun des cas, les données résultantes pourraient être considérées comme issues d'un échantillonnage par grappe du fait que le processus de sélection au hasard a d'abord été réalisé au niveau des firmes avant de concerner les entités individuelles. Il est dès lors attendu que des caractéristiques non observées au niveau de la firme (de même que les caratéristiques observées) jouent un rôle dans la prise de décision des individus et génèrent ainsi de la corrélation intra-firme qu'il convient de prendre en considération dans notre modélisation. Une estimation par un modèle à effets fixes est préférée lorsque nous nous attendons à ce que l'effet de cluster non observé – tel que capturé par a_i dans l'équation (14.12) – soit corrélé avec l'une ou plusieurs des variables explicatives du modèle. Par ailleurs, on peut aussi n'inclure que les variables explicatives qui varient, au moins partiellement, au sein des grappes. Les tailles des clusters étant rarement identiques, nous sommes donc amenés à recourir au modèle à effets fixes pour des panels non cylindrés.

Des données permettant d'étudier le lien entre la formation des étudiants et un ensemble de variables d'intérêt peuvent aussi revêtir la forme de données d'échantillon par grappe, avec un échantillon d'écoles obtenu à partir d'une population d'établissements, l'information relative aux étudiants étant ensuite dérivée à partir des données collectées au niveau de chacune des écoles. Chaque école s'assimile alors à un cluster ou une grappe, et il devient essentiel pour la validité de notre modélisation d'autoriser un « effet école » à être corrélé avec un ensemble de variables explicatives « clé » comme par exemple le fait que les étudiants participent à des programmes d'éducation prioritaire subsidiés par le gouvernement. En effet, dans la mesure où il est vraisemblable que le taux de participation des élèves à de tels programmes varie selon les écoles, il est bienvenu d'estimer dans ce contexte un modèle à effets fixes. Ainsi, il n'est pas rare de lire dans des travaux appliqués, des auteurs préciser de façon raccourcie « J'ai inclus des effets fixes "école" dans mon analyse ».

L'approche visant à estimer un modèle à EAC peut quant à elle être directement appliquée à des échantillons par grappe, car, du point de vue de l'estimation, un échantillon par grappe peut s'assimiler à une structure de panel non cylindré. Dès lors, les moyennes ajoutées à l'équation sont calculées comme des moyennes par grappe, au sein des écoles. La seule différence notable avec les données de panel tient à l'absence de pertinence de la notion de corrélation sérielle dans la composante idiosyncrasique des erreurs. Néanmoins, comme discuté dans Wooldridge (2010, Chapitre 20), il demeure pertinent de recourir à des écarts-types estimés robustes à l'échantillonnage par grappe, que l'on ait recours à l'estimation d'un modèle à effets fixes ou aléatoires.

Dans certains cas, les variables explicatives clés – souvent des variables relatives à des choix de politiques économiques – changent au niveau de la grappe, mais pas en son sein. Dans de tels cas, l'approche par effets fixes n'est pas applicable. Par exemple, nous pourrions être intéressés par l'impact de la qualité de

l'enseignement dispensé par un enseignant sur les performances scolaires des élèves, chacune des classes de l'école élémentaire correspondant à une grappe. Comme tous les élèves au sein d'une grappe sont suivis par le même enseignant, un effet grappe élimine *de facto* « l'effet classe », et par conséquent, toutes les variables observées relatives à la qualité de l'enseignement. Si nous introduisons de bonnes variables de contrôle dans notre équation, nous pouvons alors utiliser la méthode des effets aléatoires sur un panel en grappes non cylindré. La condition clé pour que l'estimateur des EA produise des estimations convaincantes est que les variables explicatives ne soient pas corrélées avec l'effet de grappe inobservé. La plupart des logiciels d'économétrie permettent d'estimer des modèles à effets aléatoires sur des panels en grappes non cylindrés sans beaucoup d'effort.

L'estimation par les MCO sur les données empilées est également une technique très employée dans le cas des échantillons en grappes lorsque l'élimination de l'effet de grappe par effets fixes est impossible ou non souhaitable. Cependant, comme pour les estimations sur données de panel, les écarts-types estimés par les MCO sont incorrects en présence d'effet de grappe. Il faut alors privilégier des méthodes d'estimation robustes à la présence de corrélation au sein de la grappe (et d'hétéroscédasticité). Un certain nombre de logiciels d'économétrie proposent des commandes simples pour corriger les écarts-types estimés ainsi que les statistiques de tests usuelles en présence de corrélation standard au sein de la grappe (de même que pour l'hétéroscédasticité). Ces corrections sont identiques à celles que l'on fait pour les MCO sur données de panel empilées, comme nous l'avons montré dans l'exemple 13.9. En guise d'exemple, Papke (1999) estime des modèles à probabilité linéaire pour la participation à des plans d'épargne suivant que les entreprises ont adopté ce type de plans. En raison de la forte probabilité d'un « effet entreprise », impliquant des corrélations entre les différents plans proposés au sein d'une même entreprise, Papke corrige les écarts-types estimés issus de l'estimation par les MCO pour tenir compte de l'effet de l'échantillonnage en grappes et de l'hétéroscédasticité dérivant du modèle à probabilité linéaire.

Avant de clôturer cette section, arrêtons-nous sur un ensemble de commentaires donnés ici par ordre d'importance. Compte tenu de la grande variété des outils prêts à l'emploi disponibles, qu'il s'agisse de l'estimation de modèles à effets fixes, à effets aléatoires ou d'inférence robuste à l'échantillonnage par grappe, il peut être tentant de chercher des raisons pour justifier l'emploi de méthodes de cluster là où cela n'est pas légitime. Par exemple, si un jeu de données est obtenu à partir d'un échantillonnage aléatoire sur la population sous-jacente, alors il n'y a en principe aucune raison de tenir compte d'éventuels effets de cluster dans le calcul des écarts-types estimés après estimation par les MCO. Le fait que les unités puissent être regroupées en cluster ex-post, c'est-à-dire à l'issue de l'échantillonnage aléatoire, ne justifie pas de procéder à de l'inférence robuste à de l'éventuelle corrélation au sein des clusters.

Pour illustrer ce propos, supposons que, sur une population d'élèves de fin de cycle primaire aux États-Unis, on échantillonne au hasard 50 000 individus, et que ces données sont ensuite étudiées à l'aide de techniques standards. Il pourrait être tentant de regrouper les élèves selon par exemple, les 50 États augmenté du District de Columbia – en faisant l'hypothèse qu'une variable muette permette d'identifier l'État dans la base – puis de traiter ces données comme des données de cluster. Ceci serait pourtant incorrect, et regrouper les écarts-types au niveau des États pourrait générer des écarts-types soit systématiquement trop élevés soit trop faibles en raison de la théorie asymptotique qui repose sur l'hypothèse que nous disposons de nombreux clusters, chacun d'entre eux étant de taille relativement faible. Dans tous les cas, un simple questionnement de bon sens permet de se rendre compte du caractère erroné de l'analyse. Par exemple, si nous connaissons le comté de résidence de chacun des élèves, pourquoi ne pas réaliser un cluster au niveau des comtés ? Ou à un autre niveau, nous pouvons également diviser les États-Unis en quatre zones de recensement et les considérer comme nos clusters – ce qui mènerait à différentes évaluations pour les écarts-types (qui n'ont aucune base théorique). Pour pousser l'argument à l'extrême, on pourrait considérer que nous avons réalisé un échantillonnage sur un unique cluster : les États-Unis dans leur ensemble, ce qui reviendrait à considérer que les écarts-types « cluster » ne peuvent être définis et que l'inférence serait impossible. La confusion provient

ici du fait que les clusters sont identifiés ex post – c'est-à-dire à l'issue du processus d'échantillonnage. Dans une véritable structure d'échantillonnage par grappe, les clusters sont définis en premier lieu puisqu'ils sont tirés au hasard au sein de la population sous-jacente, puis donnent lieu à un échantillonnage individuel à partir des clusters identifiés en première étape. Il serait possible de recourir aux méthodes d'analyse par grappe, si par exemple, une variable au niveau du district était créée à l'issue de l'échantillonnage aléatoire, puis utilisée dans l'estimation du modèle estimé au niveau individuel (i.e lorsque les entités individuelles sont les élèves). Procéder de la sorte peut générer de la corrélation dans les données observées au sein de chaque district. Rappelons ici que l'estimateur du modèle à effets fixes (dans ce cas, au niveau du district) correspond à l'estimation réalisée sur des moyennes par district. Ainsi, il est possible de tenir compte de la corrélation au sein des clusters au niveau du district, en sus des traditionnels effets fixes. Comme développé par Stock et Watson (2008) (dans le contexte de données de panel), en présence de tailles de clusters importantes, le niveau de corrélation au sein des clusters est généralement de faible importance, en revanche, lorsque la taille des clusters, il convient d'utiliser des écarts-types robustes à l'effet cluster.

RÉSUMÉ

Dans ce chapitre, nous avons poursuivi notre présentation des méthodes relatives aux données de panel, en étudiant les estimateurs des modèles à effets fixes et à effets aléatoires, puis le cas du modèle à effets aléatoires corrélés comme cadre général. Si on le compare à l'approche en différences premières, l'estimateur du modèle à effets fixes est efficace lorsque les erreurs idiosyncratiques ne sont pas autocorrélées (et homoscédastiques), en l'absence d'hypothèses sur la structure de corrélation entre l'effet inobservé a_i et les variables explicatives. Comme dans le cas de l'approche en différences premières, toute variable explicative invariante dans le temps disparaît de l'analyse. Les méthodes à effets fixes peuvent être étendues directement aux panels non cylindrés, sous réserve que les raisons pour lesquelles certaines données manquent ne soient pas systématiquement liées aux termes d'erreur idiosyncratiques.

L'estimateur du modèle à effets aléatoires est approprié lorsqu'on suppose que les effets non observés sont décorrélés de l'ensemble des variables explicatives du modèle. Dès lors, le paramètre a_i peut être considéré comme constitutif du terme d'erreur, et la méthode des moindres carrés généralisés (MCG) permet de traiter la corrélation sérielle découlant du modèle. En pratique, on peut employer les moindres carrés quasi-généralisés (MCQG) sur les données empilées exprimées en quasi-écart à leurs valeurs moyennes. La valeur estimée du paramètre de transformation, $\hat{\theta}$, indique si les estimations sont plus proches des estimations obtenues par les MCO sur les données empilées ou par le modèle à effets fixes. Si l'ensemble des hypothèses relatives aux effets aléatoires sont vérifiées, l'estimateur du modèle à effets aléatoires est asymptotiquement plus efficace – quand N devient grand pour T fixé – que ceux des MCO sur les données empilées, le modèle en différences premières, et celui à effets fixes (tous étant sans biais, convergents et asymptotiquement normalement distribués).

Le modèle de données de panel à effets aléatoires corrélés est devenu très populaire ces dernières années. Il propose un cadre simple pour tester la spécification la plus adaptée entre effets fixes et effets aléatoires, et permet en outre d'incorporer des variables invariantes dans le temps dans une équation qui permet d'estimer les paramètres associés aux variables variantes au cours du temps à l'instar du modèle à effets fixes. Enfin, il est à noter que les méthodes étudiées dans les chapitres 13 et 14 peuvent être étendues au cadre de données appariées ou d'échantillons en grappes (ou clusters). Les transformations en différences premières et « *within* » permettent d'éliminer l'effet des grappes. Si celui-ci n'est pas corrélé avec les variables explicatives du modèle, on peut utiliser les MCO sur les données empilées mais les écarts-types estimés et les statistiques de tests standard doivent être ajustés pour tenir compte de la corrélation au sein des grappes. De manière alternative, on peut recourir à l'estimation du modèle à effets aléatoires.

MOTS-CLÉS

Données en écart à leur moyenne p. 571
Données en quasi-écart à leur moyenne p. 576
Échantillon par grappe ou échantillon par *cluster* p. 585
Échantillons appariés p. 585
Effet de grappe ou effet de *cluster* p. 585
Effets aléatoires corrélés p. 580
Erreurs composées/termes d'erreurs composées p. 575
Estimateur [du modèle] à effets aléatoires p. 576
Estimateur [du modèle] à effets fixes p. 566
Estimateur *within* p. 566
Modèle à effets aléatoires p. 575
Modèle à effets inobservés p. 567
Panel non cylindré p. 573
Régression sur variables indicatrices p. 570
Transformation à effets fixes p. 567
Transformation *within* p. 566

EXERCICES

1. Supposons que les erreurs idiosyncratiques de l'équation (14.4), $\{u_{it} : t = 1, 2, \ldots, T\}$, ne soient pas corrélées sériellement et soient caractérisées par une variance constante, σ_u^2. Montrez que la corrélation entre les différences adjacentes, Δu_{it} et $\Delta u_{i,t+1}$, est de $-0{,}5$. Dès lors, sous les hypothèses du modèle à effets fixes, la différentiation première engendre de la corrélation sérielle d'une valeur connue.

2. En considérant une unique variable explicative, l'équation utilisée pour obtenir l'estimateur *between* est donnée par :

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + a_i + \bar{u}_i,$$

où les éléments surmontés d'une barre représentent les données moyennes au cours du temps. Nous faisons l'hypothèse que $E(a_i) = 0$ puisque nous avons inclus une constante dans le modèle. Supposons que $\bar{u}_i$ ne soit pas corrélé avec $\bar{x}_i$, mais que $\text{Cov}(x_{it}, a_i) = \sigma_{xa}$ pour tout t (et i du fait de l'échantillonnage aléatoire au niveau des entités individuelles).

i. Soit $\tilde{\beta}_1$ l'estimateur *between*, c'est-à-dire l'estimateur des MCO en utilisant les données moyennes, montrez que :

$$\text{plim } \tilde{\beta}_1 = \beta_1 + \sigma_{xa} / \text{Var}(\bar{x}_i),$$

où la limite en probabilité est définie pour $N \to \infty$. [*Astuce* : Voir les équations (5.5) et (5.6).]

ii. Supposons de plus que les x_{it}, pour tout $t = 1, 2, \ldots, T$, sont décorrélées et de variance constante σ_x^2. Montrez que la limite en probabilité est donnée par : $\text{plim } \tilde{\beta}_1 = \beta_1 + T(\sigma_{xa} / \sigma_x^2)$.

iii. Si les variables explicatives ne sont pas très fortement corrélées au cours du temps, que peut-on conclure à partir des éléments de la question (ii) : l'inconsistance de l'estimateur *between* est-elle plus faible lorsque la dimension temporelle s'accroît ?

3. Dans un modèle à effets aléatoires, définissez le terme d'erreur composé $v_{it} = a_i + u_{it}$, avec a_i une composante décorrélée de u_{it} et u_{it} une composante de variance constante σ_u^2 dont les réalisations sont décorrélées au cours du temps. Définissez $e_{it} = v_{it} - \theta \bar{v}_i$, avec θ tel que défini dans (14.10).

i. Montrez que $E(e_{it}) = 0$.

ii. Montrez que $Var(e_{it}) = \sigma_u^2$, $t = 1, \ldots, T$.

iii. Montrez que pour $t \neq s$, $Cov(e_{it}, e_{is}) = 0$.

4. Dans le but de déterminer les effets de la performance sportive des étudiants sur le nombre de candidatures, vous collectez des données pour un échantillon d'établissements d'enseignement supérieur de première catégorie pour les années 1985, 1990, et 1995.

i. Quelles mesures de succès sportif pourriez-vous inclure dans l'équation ? Quels sont les problèmes que cela pose d'un point de vue temporel ?

ii. De quels autres facteurs devriez-vous tenir compte dans l'équation ?

iii. Écrivez un modèle permettant d'estimer les effets de la performance sportive d'un établissement sur le taux de variation du nombre de candidatures. Comment vous y prendriez-vous pour estimer ce modèle ? Pourquoi choisiriez-vous cette méthode d'estimation ?

5. Supposez que vous ayez la possibilité de collecter les données suivantes pour un échantillon aléatoire d'étudiants en première et deuxième année de licence à l'université pour chacun des cours suivis durant le semestre : la note obtenue à l'examen final sanctionnant le cours, le taux de participation aux séances de cours, une variable indicatrice indiquant si le cours fait partie des matières principales du cursus de l'étudiant, la moyenne générale préalable au démarrage du semestre, la note obtenue à l'examen d'accès à l'université, le SAT.

i. Expliquez pourquoi vous pouvez assimiler cette base de données à une base de données en grappes. Combien d'observations vous attendez-vous à obtenir pour un étudiant type ?

ii. Écrivez un modèle similaire à celui décrit dans l'équation (14.18), expliquant les performances aux examens terminaux en fonction de l'assiduité en cours et d'autres caractéristiques. Utilisez la notation s en indice pour désigner l'étudiant et c pour la classe. Quelles sont les variables demeurant inchangées pour chaque étudiant ?

iii. Si vous compilez toutes les données et estimez le modèle par les MCO, quelles hypothèse faites-vous sur les caractéristiques inobservées des étudiants affectant leur performance et leur assiduité ? Quels rôles jouent les résultats obtenus au SAT et la valeur de la moyenne générale (GPA) au semestre antérieur ?

iv. Si vous pensez que le résultat au SAT et la moyenne générale (GPA) antérieure ne parviennent pas à rendre compte efficacement du degré d'habileté des étudiants, comment vous y prendriez-vous pour estimer l'effet de l'assiduité aux cours sur la performance aux examens finaux ?

6. Supposons que vous estimez le modèle correspondant au tableau 14.2 et que vous utilisez l'option du package « cluster » de Stata® 11 qui vous permet d'obtenir des écarts-types estimés robustes – c'est-à-dire robustes à l'hétéroscédasticité et à l'autocorrélation du terme d'erreur composé, $\{v_{it} : t = 1, \ldots, T\}$ – avec les résultats suivants : $\hat{\sigma}(\hat{\beta}_{educ}) = 0{,}11$, $\hat{\sigma}(\hat{\beta}_{black}) = 0{,}051$, $se(\hat{\beta}_{hispan}) = 0{,}09$, $\hat{\sigma}(\hat{\beta}_{exper}) = 0{,}020$, $\hat{\sigma}(\hat{\beta}_{exper2}) = 0{,}0010$, $\hat{\sigma}(\hat{\beta}_{married}) = 0{,}026$, et $\hat{\sigma}(\hat{\beta}_{union}) = 0{,}027$.

i. Comment ces écarts-types estimés se comportent-ils comparativement aux écarts-types estimés non robustes ? Pourquoi ?

ii. Comment les écarts-types estimés robustes issus de l'estimation par les MCO sur les données empilées se comportent-ils comparativement aux écarts-types estimés du modèle à effets fixes ? La présence de variables explicatives constantes ou variantes dans le temps peut-elle avoir une influence ?

iii. Lorsque les écarts-types estimés robustes des estimations à EA sont calculés, Stata® 11 reporte les chiffres suivants (nous nous intéressons seulement aux valeurs relatives aux variables variantes dans le temps ici) : $\hat{\sigma}(\hat{\beta}_{exper}) = 0{,}16$, $\hat{\sigma}(\hat{\beta}_{experq}) = 0{,}0008$, $\hat{\sigma}(\hat{\beta}_{married}) = 0{,}19$, et $\hat{\sigma}(\hat{\beta}_{union}) = 0{,}21$. [Ces valeurs sont robustes à tous types de corrélation sérielle et d'hétéroscédasticité des erreurs idiosyncratiques $\{u_{it} : t = 1, \ldots, T\}$ de même qu'à l'hétéroscédasticité de a_i.] Comment les écarts-types estimés robustes se comportent-ils en général comparativement aux écarts-types estimés usuels calculés pour le modèle à EA et reportés dans le tableau 2 ? Quelles conclusions pouvez-vous en tirer ?

iv. Comparez les quatre formes d'écarts-types estimés présentés en question (iii) avec leurs contreparties issues de l'estimation par les MCO sur les données empilées. Comment interprétez-vous le fait que tous les écarts-types estimés robustes soient inférieurs à leurs contreparties non robustes ?

7. Les données présentées dans la base CENSUS2000 consistent en un échantillon aléatoire d'individus installés aux États-Unis. Nous nous intéressons ici à l'estimation d'un modèle de régression simple reliant le log du revenu hebdomadaire *lweekinc*, aux variables explicatives *schooling*, *educ*. Nous disposons de 29 501 observations. Pour chaque individu, nous associons un indicateur mentionnant l'État d'origine (*state*) pour les 50 États augmenté du District de Columbia. Un identifiant plus subtil est donné par *puma*, qui reprend 610 valeurs différentes relatives à un ensemble de zones géographiques plus précises que les États.

L'estimation d'un modèle de régression simple de *lweekinc* sur *educ* donne un coefficient de pente égal à 0,1083 (à quatre décimales près). L'écart-type robuste à l'hétéroscédasticité est estimé à environ 0,0024. L'écart-type robuste à la présence de cluster au niveau de la mesure *puma* est estimé à environ 0,0027, la même mesure estimée au niveau de l'État est d'environ 0,0033. Pour construire un intervalle de confiance, laquelle de ces mesures d'écart-type vous paraît la plus fiable ? Justifiez.

EXERCICES SUR ORDINATEUR

C1. Nous utilisons les données contenues dans RENTAL pour cet exercice. Les données concernent le prix des loyers ainsi que d'autres variables relatives à des villes universitaires pour les années 1980 et 1990. L'idée est d'étudier si une présence accrue d'étudiants influence le loyer des biens immobiliers. Le modèle à effets inobservés est donné par :

$$\log(rent_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 log(pop_{it}) + \beta_2 log(avginc_{it}) + \beta_3 pctstu_{it} + a_i + u_{it},$$

avec *pop* la population dans la ville, *avginc* le revenu moyen, et *pctstu* la population estudiantine en pourcentage de la population totale (durant l'année scolaire).

i. Estimez l'équation par les MCO sur les données empilées et reportez les résultats sous la forme habituelle. Que pouvez-vous dire de l'estimation relative à la variable indicatrice pour l'année 1990 ? Que pouvez-vous conclure à partir de $\hat{\beta}_{pctstu}$?

ii. Les écarts-types estimés que vous reportez en question (i) sont-ils valides ? Justifiez.

iii. Différenciez maintenant l'équation et estimez-la par la méthode des MCO. Comparez votre estimation de $\hat{\beta}_{pctstu}$ avec celle discutée à la question (i). La proportion d'étudiants est-elle un facteur explicatif des loyers ?

iv. Estimez le modèle avec des effets fixes de façon à vérifier que vos résultats (paramètres et écarts-types estimés) sont identiques à ceux présentés en question (iii).

C2. Nous utilisons maintenant les données de CRIME4.

i. Ré-estimez le modèle à effets inobservés pour la criminalité comme dans l'exemple 13.9 en utilisant des effets fixes en lieu et place des différences premières. Observe-t-on des changements notables dans les signes ou la magnitude des coefficients estimés ? Qu'en est-il de la significativité statistique ?

ii. Passez en logarithmes chacune des variables de salaire de la base de données et estimez le modèle par effets fixes. Dans quelle mesure l'introduction de ces variables affecte les valeurs des coefficients estimés des variables relatives à la justice pénale décrites dans la question (i) ?

iii. Les variables de salaire discutées en question (ii) ont-elles toutes le signe attendu ? Justifiez. Sont-elles conjointement significatives ?

C3. Pour cet exercice, nous utilisons les données contenues dans la base JTRAIN pour étudier l'effet des subventions à la formation professionnelle sur la formation professionnelle par employé. Le modèle de base pour les trois années est donné par :

$$hrsemp_{it} = \beta_0 + \delta_1 d88_t + \delta_2 d89_t + \beta_1 grant_{it} + \beta_2 grant_{i,t-1} + \beta_3 log(employ_{it}) + a_i + u_{it}.$$

i. Estimez l'équation en utilisant des effets fixes. Combien de firmes sont utilisées dans cette estimation ? Combien d'observations seraient utilisées si chacune des entreprises disposait de données pour l'ensemble des variables explicatives du modèle (et en particulier, *hrsemp*) pour chacune des trois années ?

ii. Interprétez le coefficient relatif à *grant* et commentez sa significativité.

iii. Cela vous surprend-il que le coefficient associé à $grant_{-1}$ ne soit pas significatif ? Justifiez.

iv. Les grandes entreprises proposent-elles à leurs employés plus ou moins de temps de formation en moyenne ? De quel ordre sont les différences ? (Par exemple, si une entreprise a 10 % d'employés en plus, quel est le changement relatif dans le nombre d'heures allouées à la formation professionnelle ?)

C4. Dans l'exemple 13.8, nous avons utilisé les données de demandes d'allocation d'assurance chômage issues de Papke (1994) pour estimer les effets de la localisation des entreprises sur les demandes d'allocation chômage. Papke utilise également un modèle permettant de prendre en compte la dynamique propre à chaque ville comme suit :

$$\log(uclms_{it}) = a_i + c_i t + \beta_1 ez_{it} + u_{it},$$

avec a_i et c_i tous deux des effets inobservés. Cela permet de mieux prendre en compte l'hétérogénéité entre villes.

i. Montrez que, lorsque l'équation précédente est prise en différences premières nous obtenons :

$$\Delta\log(uclms_{it}) = c_i + \beta_1 \Delta ez_{it} + \Delta u_{it}, \; t = 2, \ldots, T.$$

Notez que l'équation en différences contient un effet fixe, c_i.

ii. Estimez l'équation en différences par la méthode des effets fixes. Quelle est la valeur estimée de β_1 ? Est-elle très différente de la valeur estimée obtenue dans l'exemple 13.8 ? L'effet de la localisation des entreprises est-il statistiquement significatif ?

iii. Ajoutez un ensemble de variables indicatrices temporelles pour estimer le modèle présenté en question (ii). Qu'arrive-t-il à la valeur estimée du paramètre β_1 ?

C5. i. Dans l'équation de salaire de l'exemple 14.4, expliquez pourquoi les variables indicatrices relatives à la profession pourraient être des variables omises d'importance pour l'estimation de la prime syndicale.

ii. Si chaque individu de l'échantillon avait conservé la même profession entre 1981 et 1987, auriez-vous besoin d'introduire des variables indicatrices relatives à la profession dans l'estimation du modèle à effets fixes ? Justifiez.

iii. À partir des données contenues dans la base WAGEPAN, introduisez huit variables indicatrices relatives aux professions dans l'équation et estimez-la en utilisant des effets fixes. Le coefficient associé à *union* change-t-il de beaucoup ? Qu'en est-il de sa significativité ?

C6. Ajoutez le terme d'interaction : $union_{it} \cdot t$ à l'équation estimée et dont les résultats figurent en tableau 14.2 pour voir si la croissance des salaires *growth* dépend du statut syndical. Estimez l'équation avec des effets aléatoires et des effets fixes et comparez les résultats.

C7. Nous utilisons les données étatiques relatives aux taux de criminalité et d'exécution issues du fichier MURDER pour l'exercice qui suit.

i. Considérons le modèle à effets inobservés suivant :

$$mrdrte_{it} = \eta_t + \beta_1 exec_{it} + \beta_2 unem_{it} + a_i + u_{it},$$

avec η_t les différentes variables indicatrices temporelles et a_t l'effet inobservé relatif à l'appartenance à un État donné. Si les exécutions passées de personnes jugées coupables de meurtres avaient eu un effet dissuasif, quel aurait dû être le signe de β_1 ? À votre avis, quel devrait être le signe de β_2 ? Justifiez.

ii. En restreignant l'étude aux années 1990 et 1993, estimez l'équation considérée à la question (i) au moyen d'une estimation par les MCO sur les données empilées. Ignorez le problème de corrélation sérielle dans le terme d'erreur composé. Les résultats d'estimation sont-ils en faveur de l'hypothèse d'un effet dissuasif de la peine capitale ?

iii. En vous focalisant toujours sur les années 1990 et 1993, estimez maintenant le modèle au moyen d'effets fixes. Vous pouvez recourir aux différences premières puisque vous n'utilisez que deux années de données. Vos résultats sont-ils maintenant en faveur de l'hypothèse d'un effet dissuasif de la peine capitale ? Ce résultat est-il robuste ?

iv. Calculez les écarts-types estimés robustes à l'hétéroscédasticité des paramètres estimés du modèle décrit en question (ii).

v. Identifiez l'État qui présente le nombre le plus important d'exécutions en 1993. (La variable *exec* correspond au nombre total d'exécutions réalisées en 1991, 1992, et 1993.) Identifiez le deuxième État en matière d'exécutions en 1993. Combien d'exécutions les séparent ?

vi. Estimez l'équation par la méthode des différences premières, en retirant l'État du Texas de votre analyse. Calculez les écarts-types standard et ceux robustes à l'hétéroscédasticité. Que trouvez-vous ? Que se passe-t-il ?

vii. Utilisez maintenant les trois années de données dont vous disposez et estimez le modèle à effets fixes. Introduisez l'État du Texas dans votre analyse. Discutez la taille et la significativité de l'effet dissuasif de la peine capitale comparativement aux effets mis en exergue dans les questions précédentes lorsque les seules années 1990 et 1993 étaient considérées.

C8. On considère les données issues de la base MATHPNL pour cet exercice. Notre objectif est de réaliser ici une estimation par effets fixes du modèle qui a été introduit en exercice C11 du chapitre 13 pour illustrer l'approche des différences premières. Nous nous intéressons au modèle suivant :

$$\begin{aligned} math4_{it} = {} & \delta_1 y94_t + \ldots + \delta_5 y98_t + \gamma_1 \log(rexpp_{it}) + \gamma_2 \log(rexpp_{i,t-1}) \\ & + \psi_1 \log(enrol_{it}) + \psi_2 lunch_{it} + a_i + u_{it}, \end{aligned}$$

La première année disponible (l'année de base) est l'année 1993, en raison de l'introduction de variables explicatives retardées d'une période dans le modèle.

i. Estimez le modèle par la méthode des MCO sur les données empilées et reportez les écarts-types estimés traditionnels. Veillez à bien introduire une constante dans le modèle ainsi que des variables indicatrices

annuelles de façon à ce que la composante a_t ait une valeur attendue non nulle. Quels sont les effets estimés des variables de dépenses ? Récupérez les résidus issus de l'estimation par les MCO, $\hat{v}_{it}$.

ii. Le signe du coefficient $lunch_{it}$ est-il conforme à vos attentes ? Interprétez la magnitude de ce coefficient. Diriez-vous que le taux de pauvreté dans le quartier a un effet important sur les taux de réussite ?

iii. Testez la présence de corrélation sérielle d'ordre 1 en estimant un processus AR(1) sur $\hat{v}_{it}$, c'est-à-dire en régressant $\hat{v}_{it}$ sur $\hat{v}_{it-1}$ pour les années 1994 à 1998. Vérifiez qu'il existe une forte corrélation sérielle positive et commentez ce résultat.

iv. Estimez maintenant l'équation au moyen d'effets fixes. Les variables retardées relatives aux niveaux de dépenses sont-elles toujours significatives ?

v. Pourquoi pensez-vous que dans le cadre de l'estimation par la méthode des effets fixes, les variables relatives à l'inscription et au programme de repas scolaires subventionnés apparaissent conjointement significatives ?

vi. On définit l'effet total – ou de long terme – des dépenses comme étant : $\theta_1 = \gamma_1 + \gamma_2$. Utilisez l'expression suivante $\gamma_1 = \theta_1 - \gamma_2$ pour obtenir une évaluation de l'écart-type estimé de $\hat{\theta}_1$. [*Astuce* : L'estimation standard par les effets fixes en utilisant $\log(rexpp_{it})$ et $z_{it} = \log(rexpp_{i,t-1}) - \log(rexpp_{it})$ comme variables explicatives devrait vous permettre de répondre à la question posée.]

C9. Le fichier PENSION contient des informations sur la participation à un plan d'épargne conçu pour les travailleurs américains. Une partie des observations fait référence à des couples issus d'une même famille, et à ce titre, constitue un exemple de données en grappes de petite taille (la taille de chaque grappe étant égale à deux).

i. En faisant fi de l'effet de grappe par famille, estimez le modèle par les MCO :

$$\begin{aligned} pctstck = {} & \beta_0 + \beta_1 choice + \beta_2 prftshr + \beta_3 female + \beta_4 age \\ & + \beta_5 educ + \beta_6 finc25 + \beta_7 finc35 + \beta_8 finc50 + \beta_9 finc75 \\ & + \beta_{10} finc100 + \beta_{11} finc101 + \beta_{12} wealth89 + \beta_{13} stckin89 \\ & + \beta_{14} irain89 + u, \end{aligned}$$

chacune des variables étant définie dans la base de données. La variable concentrant notre attention est la variable indicatrice *choice*, égale à un si le travailleur a le choix d'allouer ses fonds de pension entre différents types d'investissements. Quel est l'effet estimé de *choice* ? Cet effet est-il statistiquement significatif ?

ii. Les niveaux de revenu, de richesse, de détention d'actifs et de capitaux placés sur un compte épargne retraite s'avèrent-ils être des variables de contrôle importantes ? Justifiez.

iii. Déterminez combien de familles différentes sont présentes dans la base de données.

iv. Calculez maintenant les écarts-types estimés issus de la régression des MCO robustes aux corrélations existant au sein de chaque grappe (*i.e.* au sein de chaque famille). Diffèrent-ils de ceux calculés selon la méthode standard ? Êtes-vous surpris(e) ?

v. Estimez l'équation en prenant la différence entre les conjoints au sein de chaque de famille. Pourquoi les variables explicatives discutées dans la question (ii) disparaissent-elles lorsque l'on estime le modèle en différences premières ?

vi. Parmi les variables qui subsistent à la question (v), y en a-t-il qui demeurent significatives ? Êtes-vous surpris(e) ?

C10. Nous utilisons maintenant les données issues de la base AIRFARE pour cet exercice. Nous nous intéressons à l'estimation du modèle suivant :

$$\log(fare_{it}) = \eta_t + \beta_1 concen_{it} + \beta_2 \log(dist_i) + \beta_3 [\log(dist_i)]^2 + a_i + u_{it},\ t = 1, \ldots, 4,$$

où η_t symbolise les différentes variables indicatrices annuelles.

i. Estimez cette équation par la méthode des MCO sur les données empilées en prenant garde d'inclure les variables indicatrices annuelles. Si $\Delta concen = 0{,}10$, quel est le pourcentage estimé de hausse de la variable *fare* ?

ii. Évaluez l'intervalle de confiance au niveau de confiance de 95 % issu de l'estimation par les MCO du paramètre β_1. Quelle est la probabilité que cette évaluation ne soit pas fiable ? Si vous avez accès à un logiciel économétrique vous permettant de calculez les écarts-types estimés robustes, évaluez l'intervalle de confiance robuste associé au seuil de 95 % pour β_1. Comparez-le avec l'intervalle de confiance estimé selon la méthode usuelle et commentez vos résultats.

iii. Décrivez ce qui se passe lorsque l'on introduit comme déterminant la forme quadratique de log(*dist*). Plus spécifiquement, pour quelles valeurs de *dist* la relation entre log(*fare*) et *dist* devient-elle positive ? [*Astuce* : Identifiez d'abord la valeur seuil pour log(*dist*), puis calculez son exponentielle.] La valeur seuil est-elle en dehors de l'intervalle de définition de la variable en question ?

iv. Estimez maintenant l'équation au moyen des effets aléatoires. Dans quelle mesure l'estimation du paramètre β_1 change-t-elle ?

v. Estimez maintenant l'équation au moyen des effets fixes. Quelle estimation obtenez-vous pour le paramètre β_1 ? Pourquoi est-il similaire à l'estimation obtenue dans le cadre du modèle à effets aléatoires ? (*Astuce* : Quelle est la valeur de $\hat{\theta}$ dans le cadre de l'estimation à EA ?)

vi. Nommez deux caractéristiques de l'itinéraire (*route*) (autre que la distance entre deux arrêts) capturées par a_i. Est-il possible que ces effets soient corrélés avec $concen_{it}$?

vii. Êtes vous convaincu(e) qu'une plus forte concentration sur une route augmente les tarifs aériens ? Quelle est votre meilleure estimation ?

C11. Pour cette question, nous faisons l'hypothèse que vous avez accès à un logiciel économétrique permettant de calculer les écarts-types estimés robustes à la présence de corrélation sérielle arbitraire ainsi que d'hétéroscédasticité dans les données de panel.

i. Considérez les estimations par les MCO sur les données empilées du tableau 14.1, et obtenez les écarts-types estimés robustes à la présence de corrélation sérielle arbitraire (dans les erreurs composées, $v_{it} = a_i + u_{it}$) ainsi que d'hétéroscédasticité. Comment se comportent ces écarts-types robustes pour *educ*, *married*, et *union* comparativement aux calculs standards, non robustes ?

ii. Calculez maintenant les écarts-types estimés robustes pour les paramètres estimés du modèle à effets fixes autorisant la présence de corrélation sérielle arbitraire et d'hétéroscédasticité dans les erreurs idiosyncratiques, u_{it}. Comparez vos résultats aux écarts-types estimés standards, non robustes du modèle à effets fixes.

iii. Laquelle des deux méthodes, MCO sur les données empilées ou effets fixes, entraîne-t-elle l'ajustement le plus conséquent ? Pourquoi ?

C12. Rapportez-vous aux données contenues dans le fichier ELEM94_95 pour répondre à cette question. Les données font références à des écoles élémentaires dans l'État du Michigan. Dans cet exercice, nous abordons les données comme un échantillon en grappes, les grappes étant les quartiers où les écoles se situent.

i. Quels sont les nombres minimal et maximal d'écoles par quartier ? Quel est le nombre moyen d'écoles par quartier ?

ii. En usant de la méthode des MCO sur les données empilées (c'est-à-dire sur les 1848 écoles), estimez un modèle expliquant *lavgsal* à l'aide de *bs*, *lenrol*, *lstaff*, et *lunch* ; voir également l'exercice C11 du chapitre 9. Quels coefficient et écart-type estimé obtenez-vous pour la variable *bs* ?

iii. Calculez les écarts-types estimés robustes à la corrélation [par grappe] au sein d'un quartier (ainsi que l'hétéroscédasticité). Qu'advient-il à la statistique t de *bs* ?

iv. Toujours en vous fiant à l'estimation MCO sur les données empilées, retirez quatre observations pour lesquelles $bs > 0{,}5$ et calculez $\hat{\beta}_{bs}$ ainsi que son écart-type estimé robuste. Existe-t-il toujours un arbitrage entre salaires et bénéfices ?

v. Estimez l'équation par la méthode des effets fixes, en permettant un effet « quartier » commun pour les écoles issues d'un même quartier. À nouveau, retirez les observations pour lesquelles $bs > 0{,}5$. Que concluez-vous concernant l'arbitrage entre salaires et bénéfices ?

vi. A la lumière des estimations menées en questions (iv) et (v), et en réfléchissant à partir de l'effet fixe « quartier », discutez de la pertinence des compensations de salaire variables versées aux enseignants selon le quartier.

C13. Nous considérons la base de données DRIVING qui inclut des données de panel relatives à des États américains (pour les 48 États continentaux américains) de 1980 à 2004, soit un total de 25 ans. Différentes législations de sécurité routière sont considérées ici, concernant notamment les niveaux d'alcoolémie au delà desquels les chauffeurs sont considérés en état d'ivresse. Il est également fait mention de lois de suspension immédiate du permis – les permis pouvant alors être révoqués sans procès – ainsi que de régulations relatives au port de la ceinture. D'autres variables économiques et démographiques complètent la base.

i. Comment la variable *totfatrte* est-elle définie ? Quelle est la valeur moyenne de cette variable en 1980, 1992, et 2004 ? Régressez *totfatrte* sur les variables indicatrices annuelles de 1981 à 2004, et décrivez vos résultats. La conduite est-elle devenue plus sécurisée à la fin de la période ? Justifiez.

ii. Ajoutez les variables *bac*08, *bac*10, *perse*, *sbprim*, *sbsecon*, *sl*70*plus*, *gdl*, *perc*14_24, *unem*, et *vehicmilespc* à la régression mentionnée en question (i). Interprétez les coefficients relatifs à *bac*8 et *bac*10. La loi de suspension immédiate du permis de conduire (variable « *perse* ») a-t-elle un impact négatif sur le taux de mortalité ? Qu'en est-il de l'impact de la législation sur le port de la ceinture ? (Notez que si une législation a été mise en œuvre dans le courant de l'année, la variable sera égale à un nombre compris entre 0 et 1 permettant d'identifier le moment où la loi a été mise en œuvre.)

iii. Ré-estimez le modèle à partir de la question (ii) à l'aide d'un modèle à effets fixes (au niveau des États). Comment les coefficients estimés des variables *bac*08, *bac*10, *perse*, et *sbprim* se comportent-ils comparativement aux estimations issues de la régression MCO sur les données empilées ? Quel jeu d'estimations vous paraît le plus fiable ?

iv. Supposez que *vehicmilespc*, le nombre de miles conduit par tête, augmente de 1 000. À partir des estimations EF, quel est l'effet estimé de *totfatrte* ? Prenez soin d'expliquez vos résultats comme si vous vous adressiez à un non spécialiste.

v. En présence de corrélation sérielle et d'hétéroscédasticité dans les erreurs du modèle, les écarts-types estimés en question (iii) sont incorrects. Si possible, ayez recours à des écarts-types estimés robustes à l'effet de grappes (option « cluster »). Qu'advient-il de la significativité des vos variables d'intérêt de la question (iii) ?

C14. À partir des données issues de la base AIRFARE nous souhaitons répondre aux questions suivantes. Les estimations peuvent être comparées avec celles de l'exercice C10, de ce chapitre.

i. Calculez les moyennes au cours du temps de la variable *concen* ; et nommez-les *concenbar*. Combien de moyennes différentes pouvez-vous obtenir ? Reportez leurs valeurs minimale et maximale.

ii. Estimez l'équation suivante :

$lfare_{it} = \beta_0 + \delta_1 y98_t + \delta_2 y99_t + \delta_3 y00_t + \beta_1 concen_{it} + \beta_2 ldis_{it} + \beta_3 ldistsq_i + \gamma_1 concenbar_i + a_i + u_{it}$ par effets aléatoires. Vérifiez que $\hat{\beta}_1$ est bien identique à l'estimation par les effets fixes effectuée dans l'exercice C10.

iii. Si vous retirez *ldist* et *ldistsq* de l'estimation réalisée en (i) mais introduisez à la place $concenbar_i$, qu'advient-il du coefficient estimé $\hat{\beta}_1$? Qu'advient-il de l'estimation de γ_1 ?

iv. À partir de l'équation mentionnée en question (ii) et des écarts-types estimés standards tirés de l'estimation par EA, testez $H_0 : \gamma_1 = 0$ dans le cadre d'un test bilatéral. Reportez la *p*-valeur. Que pouvez-vous en conclure relativement aux estimations du modèle à EA par rapport à celles du modèle à EF pour le paramètre β_1 ?

v. Si possible, calculez la statistique *t* (et donc la *p*-valeur) associée au test qui soit robuste à la corrélation sérielle et à l'hétéroscédasticité. Cela change-t-il les conclusions obtenues à la question (iv) ?

C15. À partir des données issues de la base COUNTYMURDERS répondez aux questions suivantes. Les données couvrent les meurtres et exécutions (peines capitales) enregistrées au niveau de 2 197 comtés (« *counties* ») aux États-Unis. Voir également l'Exercice sur ordinateur C16 du Chapitre 13.

i. Soit le modèle :

$$murdrate_{it} = \theta_t + \delta_0 execs_{it} + \delta_1 execs_{i,t-1} + \delta_2 execs_{i,t-2} + \delta_3 execs_{i,t-3} + \beta_5 percblack_{it} + \beta_6 percmale_{it} + \beta_7 perc1019_{it} + \beta_8 perc2029_{it} + a_i + u_{it}$$

avec θ_t une constante différant pour chaque période temporelle *t*, a_i l'effet fixe pays, et u_{it} le terme d'erreur idiosyncrasique. Pourquoi est-il pertinent d'inclure des retards de la variable d'intérêt, *execs*, dans l'équation ?

ii. Estimez le modèle de régression de la question (i) par les MCO et reportez les valeurs estimées des paramètres δ_0, δ_1, δ_2 et δ_3 de même que les écarts-types estimés sur les données empilées. Trouvez-vous que les exécutions ont un effet dissuasif sur les meurtres ? Fournissez une explication impliquant a_i.

iii. Estimez maintenant l'équation de la question (i) en utilisant des effets fixes de façon à éliminer a_i. Quelles sont les nouvelles valeurs estimées des δ_j ? Sont-elles très différentes de celles estimées à la question (ii) ?

iv. Évaluez l'effet de long terme à partir des estimations de la question (iii). À partir des écarts-types estimés du modèle à effets fixes standard, l'impact de long-terme est-il statistiquement différent de 0 ?

v. Si possible, calculez des écarts-types pour les estimations issues du modèle à effets fixes qui soient robustes à des formes arbitraires d'hétéroscédasticité et/ou d'autocorrélation dans $\{u_{it}\}$. Qu'advient-il de la significativité $\hat{\delta}_j$? Quid de l'impact à long terme ?

ANNEXE 14A

14A.1 Hypothèses pour les effets fixes et effets aléatoires

Dans cette annexe, nous explicitons les hypothèses relatives à l'estimation des modèles à effets fixes et effets aléatoires. Nous proposons également une discussion des propriétés de ces estimateurs selon différents jeux d'hypothèses. Pour une vérification détaillée de ces assertions, on consultera avec profit Wooldridge (2010, chapitre 10).

Hypothèse EF.1

Pour chaque i, le modèle est donné par :

$$y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, \; t = 1, \ldots, T,$$

avec β_j les paramètres à estimer et a_i l'effet inobservé.

Hypothèse EF.2

Nous disposons d'un échantillon aléatoire pour les unités individuelles.

Hypothèse EF.3

Chacune des variables explicatives varie au cours du temps (au moins pour quelques uns des individus i), et aucune variable explicative n'est une parfaite combinaison linéaire des autres.

Hypothèse EF.4

La valeur espérée de l'erreur idiosyncratique conditionnellement aux variables explicatives à toutes dates t et à l'effet inobservé, est nulle pour tout t : $\mathrm{E}(u_{it}|\mathbf{X}_i, a_i) = 0$.

Sous ces quatre premières hypothèses – qui sont identiques à celles du modèle en différences premières – l'estimateur à effets fixes est sans biais. À nouveau, l'élément clé est l'hypothèse de stricte exogénéité, EF.4. Sous les mêmes hypothèses, l'estimateur EF est convergent lorsque T est fixé et $N \to \infty$.

Hypothèse EF.5

$$\mathrm{Var}(u_{it}|\mathbf{X}_i, a_i) = \mathrm{Var}(u_{it}) = \sigma_u^2, \text{ pour tout } t = 1, \ldots, T.$$

Hypothèse EF.6

Les erreurs idiosyncratiques ne sont pas corrélées pour tout $t \neq s$ (conditionnellement aux variables explicatives et à l'effet inobservé a_i) : $\mathrm{Cov}(u_{it}, u_{is}|\mathbf{X}_i, a_i) = 0$.

Sous les hypothèses EF.1 à EF.6, l'estimateur des effets fixes de β_j est le meilleur estimateur de la classe des estimateurs linéaires sans biais (BLUE). Puisque l'estimateur en différences premières est un estimateur linéaire sans biais, il est nécessairement moins bon que l'estimateur à EF. L'hypothèse responsable de ce résultat est l'hypothèse EF.6, qui implique que les erreurs idiosyncratiques ne sont pas corrélées sériellement.

Hypothèse EF.7

Conditionnellement à $\mathbf{X}_i$ et a_i, les erreurs u_{it} sont indépendantes et identiquement distribuées selon une loi de distribution $\mathrm{N}(0, \sigma_u^2)$.

L'hypothèse EF.7 implique EF.4, EF.5, et EF.6, mais est plus forte puisqu'elle fait l'hypothèse d'une distribution normale pour les erreurs idiosyncratiques. Si nous ajoutons EF.7, l'estimateur à effets fixes est distribué normalement et les statistiques t et F suivent des distributions exactes de Student et Fisher respectivement. En l'absence de EF.7, nous ne pouvons que recourir à des approximations asymptotiques. Sans hypothèses spécifiques, ces approximations requièrent N grand et T petit.

Les hypothèses de bonne pratique du modèle à effets aléatoires incluent EF.1, EF.2, EF.4, EF.5, et EF.6. (l'hypothèse EF.7 pourrait être ajoutée mais cela n'ajoute rien en pratique car nous devons estimer u.) Comme nous ne soustrayons qu'une fraction des moyennes temporelles, nous pouvons permettre l'introduction de variables constantes dans le temps. Dès lors, EF.3 peut être remplacée par :

Hypothèse EA.1

Aucune variable explicative n'est une parfaite combinaison linéaire des autres.

Autoriser l'introduction de régresseurs invariants dans le temps est coûteux : cela nécessite de faire une hypothèse concernant le lien entre l'effet inobservé, a_i, et les variables explicatives.

Hypothèse EA.2

L'hypothèse EF.4 est vérifiée, et la valeur espérée de a_i conditionnellement à l'ensemble des variables explicatives est constante, soit : $E(a_i|\mathbf{X}_i) = \beta_0$.

Il s'agit là de l'hypothèse qui évacue toute corrélation possible entre l'effet inobservé et les variables explicatives, élément clé permettant de distinguer les modèles à effets fixes et effets aléatoires. Comme nous supposons que a_i n'est corrélé avec aucun $\mathbf{x}_{it}$, nous pouvons inclure l'hypothèse de régresseurs invariants dans le temps. (Techniquement, l'expression en quasi-écarts à la moyenne ne retire qu'une fraction de la moyenne temporelle et non son intégralité.) La valeur espérée de a_i peut être non nulle selon l'hypothèse EA.4 et dès lors le modèle à effets aléatoires contient une constante, β_0, à l'instar de l'équation (14.7). Rappelons-nous qu'il est d'usage d'inclure également un ensemble de variables indicatrices temporelles, la première année étant celle de référence.

Nous devons en outre imposer l'homoscédasticité de la composante a_i comme suit :

Hypothèse EA.3

En plus de l'hypothèse EF.5, la variance de a_i conditionnellement aux variables explicatives est supposée constante : $\text{Var}(a_i|\mathbf{X}_i) = \sigma_a^2$.

Sous ces six hypothèses (EF.1, EF.2, EA.3, EA.4, EA.5, et EF.6), l'estimateur des effets aléatoires est convergent et asymptotiquement normal lorsque N devient grand pour T fixé. En fait, la convergence et la normalité asymptotique dérivent des quatre premières hypothèses, mais sans les deux dernières, les écarts-types et les statistiques de tests ne seraient pas correctes. De plus, selon ces six hypothèses, l'estimateur des EA est asymptotiquement efficace. Cela signifie que pour des grands échantillons, l'estimateur des EA aura de plus petits écarts-types estimés que l'estimateur des MCO sur les données empilées (lorsque les valeurs robustes sont considérées). Concernant les coefficients des variables variantes dans le temps (les seuls estimables par les EF), l'estimateur à EA est plus efficace que celui à EF – et souvient de façon substantielle. Pour autant, l'estimateur du modèle à EF n'a pas vocation à être efficace sous les hypothèses de bonne pratique des EA ; il a pour ambition d'être robuste à la présence de corrélation entre a_i et x_{itj}. Comme souvent en économétrie, il existe un arbitrage entre la robustesse et l'efficacité. Pour plus de détails sur la validité de ces hypothèses, on se rapportera avec profit à Wooldridge (2010, chapitre 10).

14A.2 Inférence robuste à la présence de corrélation sérielle et d'hétéroscédasticité pour les modèles à effets fixes et effets aléatoires

Une des hypothèses clés pour l'inférence à partir des modélisations à EF, à EA ou à EAC, est celle de l'absence de corrélation sérielle dans les erreurs idiosyncratiques, $\{u_{it} : t = 1, \ldots, T\}$ – voir l'hypothèse EF.6. Bien évidemment, l'hétéroscédasticité peut être un problème en tant que tel et exclut de fait l'inférence standard (voir hypothèse EF.5). Comme discuté en annexe du chapitre 13, les mêmes problèmes se posent lorsque l'on procède à l'estimation du modèle en différences premières lorsque $T \geq 3$ périodes.

Heureusement, comme nous l'avons vu pour l'estimation en différences premières, il existe maintenant des solutions simples pour permettre une inférence robuste – c'est-à-dire une inférence robuste à des violations arbitraires des hypothèses EF.5 et EF.6 et, dans le cadre des modèles à effets aléatoires (corrélés), de l'hypothèse EA.5. À l'instar de l'estimation en différences premières, l'approche générale visant à obtenir des écarts-types estimés ainsi que des statistiques de tests robustes consiste à appliquer une correction aux erreurs afin de tenir compte de la structure particulière des données – une méthode connue sous l'appellation de **clustering**. Ceci étant dit, la technique de *clustering* peut être appliquée à différents types d'équations. Par exemple, pour une estimation du modèle à EF, nous l'appliquerons à l'équation exprimée en écart aux valeurs moyennes (14.5). Pour l'estimation du modèle à EA, elle le sera à l'équation en quasi-écart aux valeurs moyennes (14.11) [un commentaire similaire tient pour le modèle à EAC, où les moyennes temporelles sont introduites comme variables explicatives additionnelles]. De plus amples développements sur cette question dépassent le cadre de cet ouvrage, pour plus de détails, on se rapportera avec profit à Wooldridge (2010, chapitre 10). La bonne compréhension des motivations derrière cette approche est en revanche essentielle. Si possible, nous devrions calculer les écarts-types estimés, intervalles de confiance, et statistiques de tests qui soient valides pour un ensemble d'unités individuelles de grande dimension sous les hypothèses les moins restrictives. L'estimateur des EF ne requiert que les hypothèses EF.1 à EF.4 pour être sans biais et convergent (lorsque $N \to \infty$ avec T fixé). De ce fait, un chercheur scrupuleux vérifiera au moins que les corrections réalisées dans le but de rendre les estimateurs robustes à la présence de corrélation sérielle et d'hétéroscédasticité dans les erreurs affectent l'inférence. L'expérience montre que c'est souvent le cas.

L'application des techniques de *clustering* pour tenir compte de la corrélation sérielle au sein de données de panel se justifie facilement lorsque N est significativement plus grand que T, mais pas lorsque N est petit et T plus grand. Calculer les statistiques robustes après des estimations à EF ou à EA est relativement simple avec la plupart des logiciels économétriques, et requiert un ajout du type de « cluster(id) » à la fin de la commande, « id » faisant ici référence à l'identifiant des unités individuelles.

CHAPITRE 15

ESTIMATION PAR VARIABLES INSTRUMENTALES ET DOUBLES MOINDRES CARRÉS

Traduction de Marion Leturcq

Dans ce chapitre, nous approfondissons le problème des **variables explicatives endogènes** dans les modèles de régression multiple. Dans le chapitre 3, nous avons discuté du biais des estimateurs par MCO quand une variable importante est omise. Dans le chapitre 5, nous avons vu qu'en général, les MCO ne convergent pas quand des variables sont omises. Dans le chapitre 9, nous avons montré que le biais de variables omises peut être éliminé (ou au moins atténué) quand une variable de substitution (*proxy*) pour une variable explicative non observée est disponible. Hélas, on ne dispose pas toujours de variables de substitution qui conviennent.

Dans les deux chapitres précédents, nous avons expliqué comment les estimations avec des effets fixes ou en différence première peuvent être utilisées sur données de panel pour estimer les effets de variables indépendantes qui varient au cours du temps en présence de variables omises qui sont fixes au cours du temps. Bien que ces méthodes soient très utiles, nous n'avons pas toujours accès à des données de panel. Quand bien même nous aurions accès à des données de panel, elles ne nous apporteraient pas grand-chose lorsque nous nous intéressons à l'effet d'une variable qui ne varie pas au cours du temps : l'estimation par différence première, de même que l'introduction d'effets fixes, élimine les variables explicatives qui restent constantes au cours du temps. De plus, les méthodes avec données de panel que nous avons étudiées jusqu'à présent ne résolvent pas le problème qui se pose quand les variables omises varient dans le temps et sont corrélées aux variables explicatives.

Dans ce chapitre, nous adoptons une autre approche face au problème de l'endogénéité. Nous verrons comment utiliser la méthode des variables instrumentales pour résoudre le problème de l'endogénéité d'une ou plusieurs variable(s) explicative(s). La méthode des doubles moindres carrés (DMC) est la deuxième méthode la plus populaire, après les moindres carrés ordinaires, d'estimation d'équations linéaires en économétrie appliquée.

Nous montrerons d'abord comment utiliser les méthodes par variables instrumentales (VI) pour obtenir des estimateurs convergents en présence de variables omises. Les VI peuvent aussi être utilisées pour résoudre les problèmes d'**erreur de mesure** sur les régresseurs, sous certaines conditions. Le chapitre suivant montrera comment estimer des modèles d'équations simultanées en utilisant des méthodes par VI.

Nous présentons l'estimation par variables instrumentales en suivant la même procédure que nous avons suivie pour les moindres carrés ordinaires dans la partie 1, dans laquelle nous faisions l'hypothèse que nous disposions d'un échantillon aléatoire tiré dans une population sous-jacente. C'est un point de départ commode car, en plus de simplifier les notations, cela permet d'insister sur le fait que les hypothèses importantes pour l'estimation par VI doivent être exprimées en termes de population sous-jacente (de la même façon que pour les MCO). Nous avons montré dans la partie 2 que les MCO peuvent être mis en œuvre sur des données de séries temporelles : c'est également possible dans le cadre des méthodes par variables instrumentales. La section 15.7 soulève quelques problèmes spécifiques que l'on rencontre quand les méthodes par VI sont appliquées sur des données de séries temporelles. Nous verrons dans la section 15.8 quelques applications sur données agrégées en coupe transversale et sur données de panel.

15.1 MOTIVATION : LES VARIABLES OMISES DANS UN MODÈLE DE RÉGRESSION SIMPLE

Nous avons vu jusqu'ici trois options pour répondre au risque d'un biais de variable omise (ou d'hétérogénéité inobservée) : (1) on peut ignorer le problème et assumer les conséquences d'un estimateur biaisé et non convergent, (2) on peut essayer de trouver et d'utiliser une variable de substitution (*proxy*) adéquate, ou (3) on peut faire l'hypothèse que la variable omise ne varie pas au cours du temps et utiliser les méthodes avec effets fixes ou par différences premières, vues dans les chapitres 13 et 14. La première réponse peut s'avérer satisfaisante quand les estimations vont dans le même sens que le biais pour les paramètres qui nous intéressent. Par

exemple, si on peut dire que l'estimateur d'un paramètre de valeur positive, par exemple l'effet de la formation professionnelle sur les salaires futurs, est biaisé vers zéro et que l'on estime que l'effet est positif et statistiquement significatif, on a appris quelque chose : la formation professionnelle a un effet positif sur les salaires et on a probablement sous-estimé son effet. Malheureusement, le cas contraire est fréquent : on aboutit fréquemment à une estimation d'une ampleur trop importante, ce qui rend difficile de tirer des conclusions utiles.

La solution qui consiste à utiliser une variable de substitution, discutée dans la section 9.2, peut aussi donner des résultats satisfaisants mais il n'est pas toujours possible de trouver une bonne variable de substitution. Cette approche tente de résoudre le problème de variable omise en remplaçant la variable inobservée par une variable de substitution.

Une autre approche consiste à laisser la variable inobservée dans le terme d'erreur, mais au lieu d'estimer le modèle par MCO, utiliser une méthode d'estimation qui prend en compte la présence d'une variable omise. C'est ce que fait la méthode des variables instrumentales.

À titre d'illustration, prenons le problème des capacités inobservées dans l'équation de salaire pour les adultes en emploi. Un modèle simple est :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + e,$$

où *wage* représente le salaire, *abil* les capacités inobservées et *e* est le terme d'erreur. Dans le chapitre 9, nous avons montré comment, sous certaines hypothèses, une variable de substitution comme le QI peut remplacer les capacités inobservées, et dans ce cas on obtient un estimateur de β_1 convergent grâce à la régression de

$$log(wage) \text{ sur } educ, QI.$$

Supposons, cependant, qu'on ne dispose pas de variable de substitution (ou que celle-ci n'a pas les propriétés nécessaires pour produire un estimateur convergent de β_1). Dans ce cas, nous mettons *abil* dans le terme d'erreur, et on se retrouve avec le modèle de régression simple

$$log(wage) = \beta_0 + \beta_1 educ + u, \qquad \textbf{[15.1]}$$

où *u* contient *abil*. Bien sûr, si l'équation (15.1) est estimée par MCO, on aboutira à un estimateur biaisé et non convergent de β_1 si *educ* et *abil* sont corrélées.

Il se trouve que l'équation (15.1) peut toujours servir de base à l'estimation, à condition que l'on puisse trouver une variable instrumentale pour *educ*. Afin de décrire cette approche, nous écrivons le modèle de régression simple :

$$y = \beta_0 + \beta_1 x + u, \qquad \textbf{[15.2]}$$

dans lequel on pense que *x* et *u* sont corrélées :

$$Cov(x,u) \neq 0. \qquad \textbf{[15.3]}$$

La méthode des variables instrumentales fonctionne, que *x* et *u* soient corrélées ou non, mais, pour des raisons que l'on verra plus tard, il vaut mieux utiliser les MCO si *x* n'est pas corrélée à *u*.

Afin d'obtenir des estimateurs convergents de β_0 et β_1 quand *x* et *u* sont corrélées, on a besoin d'informations supplémentaires. C'est une nouvelle variable, qui satisfait certaines propriétés, qui nous apporte ces informations. Imaginons que nous disposions d'une variable observable *z* qui satisfasse deux hypothèses : (1) *z* n'est pas corrélée à *u*, c'est-à-dire :

$$Cov(z,u) = 0 ; \qquad \textbf{[15.4]}$$

(2) *z* est corrélée à *x*, c'est-à-dire :

$$Cov(z,x) \neq 0. \qquad \textbf{[15.5]}$$

Alors nous appelons z une **variable instrumentale** pour x, ou parfois simplement un **instrument** pour x.

Il est nécessaire que l'instrument z vérifie l'équation (15.4), ce que l'on résume parfois en disant « z est exogène dans l'équation (15.2) », et on appelle donc l'équation (15.4) équation d'**exogénéité de l'instrument**. Dans un contexte de variable omise, l'exogénéité de l'instrument signifie que z ne doit pas avoir d'effet marginal sur y (une fois qu'on a tenu compte de l'influence de x et des variables omises), et z ne doit pas être corrélée aux variables omises. L'équation (15.5) signifie que z doit avoir un lien, positif ou négatif, avec la variable explicative exogène x. On appelle parfois cette condition « **pertinence de l'instrument** » (comme dans « z est pertinent pour expliquer les variations en x »).

Ces deux conditions que doit vérifier une variable instrumentale présentent une différence importante. Puisque l'équation (15.4) concerne la covariance entre z et le terme inobservé u, on ne peut généralement pas espérer pouvoir tester cette hypothèse : dans une grande majorité des cas, il faut faire appel au comportement économique ou à l'introspection pour justifier que $\text{Cov}(z,u) = 0$. (Dans certains cas peu courants, on pourrait observer d'une variable de substitution pour certains facteurs contenus dans u, auquel cas on peut vérifier si z et la variable de substitution ne sont pas tout à fait corrélées. Bien sûr, si on dispose d'une bonne variable de substitution pour un élément important de u, on pourrait juste ajouter la variable de substitution en tant que variable explicative et estimer l'équation ainsi étendue par moindres carrés ordinaires. Voir section 9.2.)

À l'opposé, la condition selon laquelle z est corrélée à x (dans la population) peut être testée, étant donné un échantillon aléatoire tiré de la population. La façon la plus simple de faire cela est d'estimer une régression simple entre x et z. Dans la population, nous avons :

$$x = \pi_0 + \pi_1 z + v \qquad \textbf{[15.6]}$$

Par conséquent, puisque $\pi_1 = \text{Cov}(z,x)/\text{Var}(z)$, l'hypothèse (15.5) est vérifiée si et seulement si $\pi_1 \neq 0$. Ainsi, on doit être en mesure de *rejeter* l'hypothèse nulle selon laquelle

$$H_0 : \pi_1 = 0 \qquad \textbf{[15.7]}$$

contre l'hypothèse alternative $H_1 : \pi_1 \neq 0$, à un seuil de significativité suffisamment faible (c'est-à-dire 5 % ou 1 %). Si c'est le cas, on peut avoir confiance dans le fait que (15.5) est vérifiée.

En ce qui concerne l'équation *log(wage)* en (15.1), une variable instrumentale z pour *educ* : (1) ne doit pas être corrélée aux capacités inobservées (et aucun facteur inobservé pouvant affecter le salaire) et (2) doit être corrélée au niveau d'études. Quelque chose comme le dernier chiffre du numéro de sécurité sociale[1] d'un individu vérifie très certainement la première condition : il n'est pas corrélé aux capacités inobservées parce qu'il est déterminé de façon aléatoire. Néanmoins, c'est précisément parce que le dernier chiffre du numéro de sécurité social est déterminé de façon aléatoire qu'il n'est pas non plus corrélé avec l'éducation, il s'agit par conséquent d'une mauvaise variable instrumentale pour *educ*.

Ce qu'on appelle une *variable de substitution* (*proxy*) pour une variable omise est un mauvais instrument pour la raison opposée. Par exemple, dans le cas où on oublierait les capacités inobservées dans l'équation de *log(wage)*, une variable de substitution pour *abil* doit être le plus fortement possible corrélée à *abil*. Une variable instrumentale *ne doit pas être corrélée* à *abil*. Par conséquent, alors que *QI* est un bon candidat comme variable de substitution de *abil*, ce n'est pas une bonne variable instrumentale pour *educ*.

Il n'est pas facile de savoir si d'autres variables instrumentales potentielles vérifient la condition d'exogénéité requise en (15.4). Pour les équations de salaire, les économistes du travail utilisent des variables sur le milieu familial comme variables instrumentales pour l'éducation. Par exemple, le niveau d'études de la mère (*motheduc*) est corrélé positivement au niveau d'études de l'enfant, comme on peut le constater en

1 Identifiant individuel donné par le système d'assurance maladie français.

collectant un échantillon de données sur des personnes en emploi et en faisant une régression simple de *educ* sur *motheduc*. Par conséquent, *motheduc* satisfait l'équation (15.5). Le problème est que l'éducation de la mère peut aussi être corrélée aux capacités inobservées de l'enfant (via les capacités inobservées de la mère et peut être la qualité de l'éducation en bas âge), auquel cas la condition (15.4) n'est pas vérifiée.

On peut aussi choisir la taille de la fratrie pendant l'enfance (*sibs*) comme VI pour *educ* dans l'équation (15.1). En général, avoir plus de frères et sœurs est associé à un niveau d'éducation plus faible. Ainsi, si la taille de la fratrie n'est pas corrélée aux capacités inobservées, elle peut faire office de variable instrumentale pour *educ*.

En guise de second exemple, considérons le problème de l'estimation de l'effet causal du fait de sécher les cours sur le résultat à l'évaluation finale. Dans le cadre d'une régression simple, on a

$$score = \beta_0 + \beta_1 skipped + u, \qquad \textbf{[15.8]}$$

où *score* est le résultat à l'évaluation finale et *skipped* est le nombre de cours manqués au cours du semestre. On doit certainement se préoccuper du fait que le nombre de cours séchés (*skipped*) est corrélé à d'autres facteurs contenus dans u : les étudiants les plus talentueux et les plus motivés ratent certainement moins de cours que les autres. Ainsi, une régression simple de *score* sur *skipped* ne donnera sûrement pas une bonne estimation de l'effet causal des cours séchés.

Quelle variable pourrait constituer une bonne VI pour *skipped* ? Nous avons besoin de quelque chose qui n'ait pas d'effet direct sur *score* et qui ne soit pas corrélé avec les capacités ou la motivation de l'étudiant. En même temps, la VI doit être corrélée à *skipped*. Une option serait d'utiliser la distance entre le lieu de vie et le campus. Certains étudiants doivent utiliser un moyen de transport pour aller au campus, ce qui peut accroître la propension à rater des cours (en raison du mauvais temps, d'un réveil difficile, etc.). Ainsi, *skipped* est potentiellement corrélé à *distance*, ce qui peut être vérifié en régressant *skipped* sur *distance* et en procédant à un test de Student, comme nous l'avons décrit ci-dessus.

Peut-on penser que *distance* n'est pas corrélée à u ? Dans le modèle de régression simple (15.8), certains facteurs dans u pourraient être corrélés à *distance*. Par exemple, les étudiants issus d'une famille modeste vivent peut-être en dehors du campus. Si les revenus ont un impact sur les performances de l'étudiant, cela peut conduire *distance* à être corrélée à u. La section 15.2 montre comment utiliser les VI dans un contexte de régression multiple, de sorte que les autres facteurs qui influencent *score* peuvent être directement inclus dans le modèle. En conséquence, *distance* pourrait être un bon instrument pour *skipped*. Une approche par variable instrumentale peut ne pas être du tout nécessaire si une bonne variable de substitution pour les capacités inobservées de l'étudiant existe, comme par exemple les résultats de l'étudiant aux examens précédents à ce semestre.

Il y a une dernière chose qu'il est bon de souligner avant de s'intéresser au mécanisme de l'estimation par VI : au moment d'utiliser une régression simple dans l'équation (15.6) pour tester (15.7), il est important de prendre note du signe (et même de l'amplitude) de $\hat{\pi}_1$, et pas seulement de son seuil de significativité. Quand on justifie pourquoi z est un bon candidat pour être l'instrument d'une variable explicative endogène x, il faut inclure, au moment de discuter le choix des instruments, une discussion sur la nature du lien entre x et z. Par exemple, en raison de l'influence de la génétique et du milieu social, une corrélation positive entre le niveau d'étude de l'enfant (x) et celui de la mère (z) a du sens. Si vous trouvez dans votre échantillon de données une corrélation négative entre les niveaux d'études (à savoir, $\hat{\pi}_1 < 0$) alors utiliser le niveau d'études de la mère comme une VI pour le niveau d'étude de l'enfant risque de ne pas être convaincant. (Et ceci n'a rien à voir avec le fait que la condition (15.4) soit vérifiée.) Lorsqu'on cherche à mesurer si le fait de rater des cours a un effet sur les résultats aux examens, on devrait trouver une relation positive et statistiquement significative entre *séchés* et *distance* afin d'argumenter en faveur de l'utilisation de *distance* comme VI pour *séchés* : une relation négative serait difficile à justifier (et cela suggèrerait qu'il existe des variables omises

importantes qui conduisent à une corrélation négative – des variables qui devraient elles-mêmes être incluses dans le modèle (15.8)).

Nous allons maintenant démontrer que le fait de disposer d'une variable instrumentale permet d'estimer de manière convergente les paramètres de l'équation (15.2). En particulier, nous montrons que les hypothèses (15.4) et (15.5) servent à *identifier* le paramètre β_1. L'**identification** d'un paramètre signifie, dans ce contexte, que l'on peut écrire β_1 en termes de moments de la population qui peuvent être estimés en utilisant un échantillon de données. Afin d'écrire β_1 en termes de covariances de la population, on utilise l'équation (15.2) : la covariance entre z et y est

$$Cov(z,y) = \beta_1 Cov(z,x) + Cov(z,u).$$

À présent, sous l'hypothèse (15.4), $Cov(z,u) = 0$, et sous l'hypothèse (15.5), $Cov(z,x) \neq 0$. Par conséquent, il est possible de résoudre l'équation en β_1 de la façon suivante :

$$\beta_1 = \frac{Cov(z,y)}{Cov(z,x)} \qquad \textbf{[15.9]}$$

[Notons qu'il n'est pas possible de faire ce calcul simple si z et x ne sont pas corrélés, c'est-à-dire si $Cov(z,x) = 0$.] L'équation (15.9) montre que β_1 est la covariance de la population entre z et y divisée par la covariance de la population entre z et x, ce qui montre que β_1 est identifié. Étant donné un échantillon aléatoire, nous estimons les quantités de la population par leurs analogues de l'échantillon. Après avoir simplifié la taille de l'échantillon au numérateur et au dénominateur, on obtient l'estimateur des variables instrumentales (VI) de β_1 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \qquad \textbf{[15.10]}$$

Étant donné un échantillon de données sur x, y et z, il est facile d'obtenir un estimateur des VI en (15.10). L'estimateur des VI de β_0 est simplement $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, ce qui ressemble fort à l'estimateur des MCO de la constante, à ceci près que l'estimateur de la pente, $\hat{\beta}_1$, est maintenant l'estimateur des VI.

Ce n'est pas par hasard que l'on obtient l'estimateur des MCO de $\hat{\beta}_1$ quand $z = x$. En d'autres termes, lorsque x est exogène, il peut être utilisé comme son propre instrument, et l'estimateur des VI est alors identique à l'estimateur des MCO.

On montre, en appliquant simplement la loi des grands nombres, que l'estimateur VI converge vers $\hat{\beta}_1$: $plim(\hat{\beta}_1) = \beta_1$, à condition que les hypothèses (15.4) et (15.5) soient vérifiées. Si l'une des deux hypothèses n'est pas vérifiée, les estimateurs VI ne sont pas convergents (nous reviendrons sur cela plus loin). L'estimateur des VI, par essence, est toujours biaisé dès lors que x et u sont corrélés, c'est-à-dire quand l'estimation par variable instrumentale est nécessaire. Cela signifie que l'estimateur des VI peut avoir un biais substantiel avec de petits échantillons, c'est une des raisons pour lesquelles on préfère les grands échantillons.

Lorsque l'on discute de l'application des variables instrumentales, il est important de faire attention au vocabulaire. De même que les MCO, les VI sont une méthode d'*estimation*. Se référer à « un modèle de variables instrumentales » a peu de sens – de même que l'expression « modèle des MCO » a peu de sens. Comme on l'a vu, un modèle est une équation comme celle en (15.8), qui est un cas particulier du modèle générique de l'équation (15.2). Lorsqu'on fait face à un modèle comme celui en (15.2), on peut choisir d'estimer les paramètres de ce modèle de nombreuses façons différentes. Avant ce chapitre, nous nous sommes principalement concentrés sur les MCO, mais, par exemple, on sait aussi d'après le chapitre 8, qu'on peut

utiliser une méthode d'estimation alternative, les moindres carrés pondérés (et là, il y a encore en général un certain nombre de possibilités pour les poids). Si on a une variable instrumentale potentielle pour x, appelée z, on peut opter pour une estimation par variables instrumentales. Il est vrai que la méthode d'estimation choisie dépend du modèle et de ses hypothèses. Mais la définition et l'existence des estimateurs ne dépendent pas du modèle ou des hypothèses : souvenons-nous qu'un estimateur est seulement une règle pour combiner les données. On comprend ce que le chercheur veut dire en utilisant une phrase telle que « j'ai estimé un modèle de variables instrumentales » mais cette façon de parler trahit un manque de compréhension de la différence entre un modèle et une méthode d'estimation.

Inférence statistique avec l'estimateur des VI

Étant donné la similitude de la structure des estimateurs des VI et des MCO, il n'est pas surprenant que l'estimateur des VI ait approximativement une distribution normale pour de grands échantillons. Afin de pouvoir faire l'inférence de β_1, nous avons besoin de l'écart-type estimé du coefficient, que l'on pourrait utiliser pour calculer des statistiques de Student et des intervalles de confiance. L'approche classique est d'imposer une hypothèse d'homoscédasticité, exactement comme dans le cas des MCO. Dorénavant, l'hypothèse d'homoscédasticité est exprimée conditionnellement à la variable instrumentale, z, non pas à la variable explicative endogène, x. En plus des hypothèses précédentes sur u, x et z, on ajoute

$$E(u^2|z) = \sigma^2 = Var(u) \qquad \textbf{[15.11]}$$

On peut montrer que, sous les hypothèses (15.4), (15.5), et (15.11), la variance asymptotique de $\hat{\beta}_1$ est

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} \qquad \textbf{[15.12]}$$

où σ_x^2 est la variance de x dans la population, σ^2 la variance de u dans la population, et $\rho_{x,z}^2$ est le carré de la corrélation entre x et z dans la population. Ce dernier nous indique dans quelle mesure x et z sont corrélées dans la population. Comme pour l'estimateur des MCO, la variance asymptotique décroît vers 0 à la vitesse $1/n$, où n est la taille de l'échantillon.

L'équation (15.12) est intéressante pour deux raisons. D'abord, elle propose une façon d'obtenir l'écart-type pour un estimateur VI. Toutes les quantités de l'équation (15.12) peuvent être estimées de manière convergente étant donné un échantillon aléatoire. Pour estimer σ_x^2, on calcule simplement la variance de l'échantillon de x_i ; pour estimer $\rho_{x,z}^2$, on peut faire tourner une régression de x_i sur z_i afin d'obtenir le R-carré, à savoir $R_{x,z}^2$. Pour finir, pour estimer σ^2, on peut utiliser les résidus VI,

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,\ i = 1, 2, \ldots, n$$

où $\hat{\beta}_0$ et $\hat{\beta}_1$ sont les estimateurs VI. Un estimateur convergent de σ^2 ressemble très fortement à l'estimateur de σ^2 à partir d'une simple régression MCO :

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

où il est classique de prendre en compte le nombre de degrés de liberté (même si cela a peu d'effet au fur à mesure que la taille de l'échantillon s'accroît).

L'écart-type (asymptotique) de $\hat{\beta}_1$ est la racine carrée de la variance asymptotique estimée, qui est donnée par :

$$\frac{\hat{\sigma}^2}{SST_x \cdot R_{x,z}^2} \qquad \textbf{[15.13]}$$

où SST_x est la somme des carrés des x_i. (Il faut garder en tête que la variance de l'échantillon des x_i est SST_x/n, donc les tailles d'échantillon se simplifient et donnent (15.13).) L'écart-type qui en résulte peut être utilisé pour construire soit des statistiques de Student pour tester des hypothèses sur β_1, soit des intervalles de confiance pour β_1. $\hat{\beta}_0$ a aussi un écart-type que l'on ne présente pas ici. N'importe quel logiciel moderne d'économétrie calcule les écarts-types estimés des coefficients après une estimation par VI. Il y est rarement nécessaire de faire les calculs à la main.

La seconde raison pour laquelle (15.12) est intéressante est qu'elle nous permet de comparer les variances asymptotiques des estimateurs VI et MCO (lorsque x et u ne sont pas corrélées). Sous les hypothèses de Gauss-Markov, la variance de l'estimateur des MCO est σ^2/SST_x, alors que la formule comparable pour l'estimateur VI est $\dfrac{\sigma^2}{SST_x \cdot R^2_{x,z}}$: elles diffèrent seulement en ceci que $R^2_{x,z}$ apparaît au dénominateur de la variance VI. Du fait que le R-carré est toujours inférieur à 1, la variance VI est toujours plus grande que la variance MCO (quand les MCO sont valides). Si $R^2_{x,z}$ est petit, alors la variance VI peut être bien plus grande que la variance MCO. Pour mémoire, $R^2_{x,z}$ mesure la force de la relation linéaire entre x et z dans l'échantillon. Si x et z sont à peine corrélées, $R^2_{x,z}$ peut être faible, et ceci se traduit par une grande variance de l'estimateur VI dans l'échantillon. Plus z est corrélée à x, plus $R^2_{x,z}$ est proche de 1, et plus la variance de l'estimateur des VI est faible. Dans la cas où $z = x$, $R^2_{x,z} = 1$, et on retrouve la variance MCO, comme on s'y attendait.

La discussion précédente met en valeur le coût de mettre en place une estimation VI quand x et u ne sont pas corrélées : la variance asymptotique de l'estimateur VI est toujours plus grande, et parfois bien plus grande, que la variance asymptotique de l'estimateur des MCO.

EXEMPLE 15.1
Estimer les rendements de l'éducation chez les femmes mariées

Nous utilisons les données sur les femmes mariées en emploi de la base MROZ afin d'estimer les rendements de l'éducation au moyen du modèle de régression simple

$$\log(wage) = \beta_0 + \beta_1 educ + u. \qquad \textbf{[15.14]}$$

où *wage* correspond au salaire et *educ* au niveau d'études.

Afin de comparer, nous considérons d'abord l'estimation par les MCO :

$$\widehat{\log(wage)} = -0{,}185 + 0{,}109\, educ$$
$$(0{,}185)\ \ (0{,}014) \qquad \textbf{[15.15]}$$
$$n = 428,\ R^2 = 0{,}118.$$

L'estimation de β_1 implique un rendement de presque 11 % par année d'éducation supplémentaire.

Ensuite, nous utilisons l'éducation du père (*fatheduc*) comme variable instrumentale pour *educ*. Il nous faut justifier l'absence de corrélation entre *fatheduc* et u. La deuxième condition est qu'*educ* et *fatheduc* soient corrélées. On peut très aisément vérifier cela en utilisant une régression simple de *educ* sur *fatheduc* (en utilisant seulement les femmes en emploi dans l'échantillon) :

$$\widehat{educ} = 10{,}24 + 0{,}269\, fatheduc$$
$$(0{,}28)\ \ (0{,}029) \qquad \textbf{[15.16]}$$
$$n = 428,\ R^2 = 0{,}173.$$

La statistique de Student pour *fatheduc* est de 9,28, ce qui nous indique que la corrélation entre *educ* et *fatheduc* est positive et statistiquement significative. (En fait, *fatheduc* explique 17 % de la variation d'*educ* dans l'échantillon.) Utiliser *fatheduc* comme VI pour *educ* nous donne :

$$\widehat{\log(wage)} = 0{,}441 + 0{,}059\, educ$$
$$(0{,}446)\ (0{,}035)$$
$$n = 428,\ R^2 = 0{,}093. \qquad \textbf{[15.17]}$$

L'estimation VI du rendement de l'éducation est 5,9 %, ce qui est à peine plus de la moitié de l'estimation MCO. Cela suggère que l'estimation MCO est trop élevée, ce qui est cohérent avec un biais de variable omise. Mais nous devons garder en tête que ce sont des estimations pour seulement un seul échantillon : nous ne saurons jamais si 0,109 est au dessus du vrai rendement de l'éducation, ou si 0,059 est plus proche du vrai rendement de l'éducation. De plus, l'écart-type du coefficient de l'estimation par VI est deux fois et demie plus grand que l'écart-type du coefficient des MCO (ce à quoi on s'attendait, pour les raisons que nous avons indiquées avant). L'intervalle de confiance à 95 % de β_1 en utilisant les MCO est bien plus étroit que celui obtenu en utilisant les VI. En fait, l'intervalle de confiance des VI contient l'estimation des MCO. Par conséquent, même si les différences entre (15.15) et (15.17) sont en pratique importantes, on ne peut pas dire si la différence est *statistiquement* significative. Nous verrons comment tester cela dans la section 15.5.

Dans l'exemple précédent, les rendements de l'éducation estimés en utilisant les VI étaient moins élevés qu'en utilisant les MCO, ce qui correspond à nos attentes. Mais cela n'est pas forcément le cas, comme le montre l'exemple suivant.

EXEMPLE 15.2
Estimer les rendements de l'éducation chez les hommes

On utilise maintenant WAGE2 pour estimer les rendements de l'éducation chez les hommes. On utilise la variable *sibs* (nombre de frères et sœurs) comme instrument pour *educ*. Ces dernières sont corrélées négativement, ce qu'on peut vérifier grâce à la régression simple :

$$\widehat{educ} = 14{,}14 - 0{,}228\, sibs$$
$$(0{,}11)\ \ (0{,}030)$$
$$n = 935,\ R^2 = 0{,}057.$$

Cette équation implique que chaque frère (ou sœur) est associé, en moyenne, à 0,23 années d'éducation en moins. Si nous faisons l'hypothèse que *sibs* n'est pas corrélée au terme d'erreur dans (15.14), alors l'estimateur VI est convergent. Estimer l'équation (15.14) en utilisant *sibs* comme instrument d'*educ* donne

$$\widehat{\log(wage)} = 5{,}13 + 0{,}122\, educ$$
$$(0{,}36)\ (0{,}026)$$
$$n = 935.$$

(Nous trouvons ici *R*-carré négatif, donc nous ne le reportons pas. Nous discutons du *R*-carré dans le cas de l'estimation par VI plus loin.) Pour comparer, l'estimation de β_1 par les MCO est 0,059 avec un écart-type de 0,006. Contrairement à l'exemple précédent, l'estimation des VI est maintenant bien plus élevée que l'estimation par les MCO. Même si on ne sait pas si la différence est statistiquement significative, cela n'est pas cohérent avec un biais de variable omise dans l'estimation par MCO. Il se peut que *sibs* soit également corrélée aux

capacités : plus de frères et sœurs signifie, en moyenne, moins d'attention de la part des parents, ce qui peut diminuer les capacités. On peut aussi interpréter ce résultat en supposant que l'estimateur des MCO est biaisé vers zéro à cause des erreurs de mesure d'*educ*. Cela n'est pas entièrement satisfaisant car, comme nous l'avons évoqué dans la section 9.3, *educ* ne suit certainement pas le modèle classique d'erreur de mesure des régresseurs.

Dans les exemples précédents, la variable explicative endogène (*educ*) ainsi que les variables instrumentales (*fatheduc*, *sibs*) sont de nature quantitative. Mais rien n'empêche les variables explicatives ou instrumentales d'être des variables binaires. Angrist et Krueger (1991), dans leur analyse la plus simple, ont eu une idée intelligente en proposant une variable instrumentale binaire pour *educ* qui fait usage des données du recensement sur les hommes, aux États-Unis. On appelle *frstqrt* une variable qui vaut 1 si l'homme est né au cours du premier trimestre de l'année et 0 sinon. On peut penser que le terme d'erreur dans (15.14) – et, en particulier, les capacités inobservées – ne devrait pas avoir de lien avec le trimestre de naissance. Mais *frstqrt* doit aussi être corrélée à *educ*. Il se trouve que le nombre d'année d'études dépend *effectivement* de manière systématique du trimestre de naissance dans la population. Angrist et Krueger justifient de manière convaincante que cela est dû aux lois sur la scolarité obligatoire en vigueur dans les États américains. En bref, les étudiants nés plus tôt dans l'année entrent à l'école, en général, plus âgés. Par conséquent, ils atteignent l'âge de fin de scolarité obligatoire (16 ans dans la plupart des États), en ayant atteint un niveau d'éducation un petit peu plus faible que des étudiants qui sont entrés à l'école plus jeunes. Pour les étudiants qui finissent le lycée, par contre, Angrist et Krueger montrent qu'il n'y a pas de lien entre le nombre d'années d'éducation et le trimestre de naissance.

En raison du fait que le nombre d'années d'éducation varie seulement un tout petit peu avec le trimestre de naissance – ce qui signifie que $R^2_{x,z}$ dans (15.13) est très faible –, Angrist et Krueger avaient besoin d'un échantillon de très grande taille pour obtenir une estimation des VI suffisamment précise. En utilisant 247 199 hommes nés entre 1920 et 1929, l'estimation MCO indiquait un rendement de l'éducation de 0,0801 (avec un écart-type de 0,0004), et l'estimation des VI indiquait 0,0715 (0,0219) ; ces résultats sont reportés dans le tableau III de l'article de Angrist et Krueger. Il faut remarquer que la statistique de Student est très grande pour l'estimation par les MCO (à peu près 200), alors que la statistique de Student pour l'estimation des VI est seulement de 3,26. Ainsi, l'estimation des VI est statistiquement différente de zéro, mais son intervalle de confiance est plus large que celui basé sur l'estimation par les MCO.

Angrist et Krueger trouvent un résultat intéressant : l'estimation des VI n'est pas très différente de l'estimation par les MCO. En fait, en utilisant les hommes nés au cours de la décennie suivante, l'estimation des VI est un peu plus élevée que l'estimation par les MCO. On pourrait interpréter ce résultat comme une preuve qu'il n'y a pas de biais lié à l'omission des capacités dans l'équation de salaire quand celle-ci est estimée par les MCO. Cependant, l'article d'Angrist et Krueger a été critiqué d'un point de vue économétrique. Comme expliqué dans Bound, Jaeger et Baker (1995), il se pourrait que la période de naissance soit corrélée à des facteurs inobservés qui ont un effet sur le salaire. Comme nous l'expliquerons dans la prochaine sous-section, l'estimateur des VI peut poser de gros problèmes s'il existe une corrélation entre z et u, même très faible.

Dans le cadre de l'analyse de politiques publiques, la variable explicative endogène est souvent binaire. Par exemple, Angrist (1990) étudie l'effet d'être un ancien combattant de la guerre du Vietnam sur les revenus au cours de la vie. Un modèle simple est :

$$\log(wage) = \beta_0 + \beta_1 veteran + u \qquad \textbf{[15.18]}$$

où *veteran* est une variable binaire valant 1 si la personne est un ancien combattant. Le problème avec l'estimation par MCO est qu'il peut y avoir un effet d'*auto-sélection*, comme nous l'avons mentionné dans le

chapitre 7 : il se pourrait que les personnes qui parviennent à tirer le plus de l'armée décident de s'engager, ou que la décision de s'engager soit corrélée à d'autres caractéristiques qui affectent les revenus. *veteran* et u sont donc corrélées.

Angrist fait remarquer que le tirage au sort pour le Vietnam offre un cadre d'expérience naturelle (voir aussi chapitre 13) qui crée une variable instrumentale pour *veteran*. Les jeunes hommes recevaient un numéro de tirage qui déterminait s'ils allaient être appelés à servir dans l'armée au Vietnam. Les numéros attribués étaient (au final) assignés de façon aléatoire, il semble donc possible que le numéro de tirage au sort ne soit pas corrélé au terme d'erreur u. Mais ceux qui avaient un numéro suffisamment petit devaient aller servir au Vietnam, donc la probabilité d'être un ancien combattant est corrélée au numéro de tirage. Si ces deux assertions sont vraies, le numéro de tirage au sort est un bon candidat pour être une VI de *veteran*.

Pour aller plus loin 15.1

Si certains hommes à qui on avait assigné un petit numéro de tirage prolongent leur scolarité afin de réduire la probabilité d'être appelé, le numéro de tirage est-il un bon instrument de *veteran* dans (15.18) ?

Il est aussi possible d'avoir une variable explicative endogène binaire ainsi qu'une variable instrumentale binaire. Voir le problème 1 pour un exemple.

Propriétés des VI avec une variable instrumentale faible

Nous avons déjà vu que l'estimateur des VI est convergent dès lors que z et u ne sont pas corrélées et que z et x sont corrélées, que ce soit positivement ou négativement. Néanmoins, nous avons également vu que les estimations par VI peuvent présenter de grands écarts-types estimés des coefficients, surtout si z et x sont seulement un peu corrélées. Les conséquences d'une faible corrélation entre z et x peuvent être encore plus graves : l'estimateur des VI pourrait être très biaisé asymptotiquement dès que z et u sont corrélées, même si elles ne le sont que faiblement.

Nous pouvons voir cela en étudiant la probabilité limite de l'estimateur VI quand z et u sont corrélées. En notant $\hat{\beta}_{1,\text{VI}}$ l'estimateur VI, on peut écrire

$$plim\ \hat{\beta}_{1,\text{VI}} = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)}\frac{\sigma_u}{\sigma_x} \qquad [15.19]$$

où σ_u et σ_x sont respectivement les écarts-types de u et x dans la population. La partie intéressante de cette équation est celle qui contient les termes de corrélation. Elle montre que même si $Corr(z,u)$ est faible, l'estimateur VI est loin d'être convergent si $Corr(z,x)$ est également faible. Ainsi, même si seule la convergence nous préoccupe, utiliser les VI plutôt que les MCO n'apporte pas nécessairement d'amélioration si la corrélation entre z et u est plus petite que celle entre x et u. En utilisant le fait que $Corr(x,u) = Cov(x,u) / (\sigma_x\sigma_u)$ combiné à l'équation [5.3], on peut écrire la probabilité limite de l'estimateur des MCO – que l'on appellera $\hat{\beta}_{1,OLS}$ – ainsi :

$$plim\ \hat{\beta}_{1,OLS} = \beta_1 + Corr(x,u)\frac{\sigma_u}{\sigma_x} \qquad [15.20]$$

En comparant ces deux formules, on voit qu'il est possible que les directions des biais asymptotiques soient différentes pour les VI et pour les MCO. Par exemple, supposons que $Corr(x,u) > 0$, $Corr(z,x) > 0$ et $Corr(z,u) < 0$. Alors l'estimateur VI est biaisé négativement tandis que l'estimateur des MCO est biaisé positivement (asymptotiquement). En pratique, une telle situation est peu courante. Lorsque la direction du biais est la même et que la corrélation entre z et x est faible, le cas est plus problématique. Pour être concret,

supposons que x et z sont toutes les deux corrélées positivement à u et que $Corr(z,x) > 0$. Le biais asymptotique de l'estimateur des VI est alors plus faible que celui de l'estimateur des MCO seulement si $Corr(z,x) < Corr(x,u)$. Si $Corr(z,x)$ est faible, alors ce qui peut paraître une petite corrélation entre z et u est amplifiée et l'estimateur des VI est alors pire que l'estimateur des MCO, même quand on ne considère que la question du biais. Par exemple, si $Corr(z,x) = 0{,}2$, $Corr(z,u)$ doit être inférieur à un cinquième de $Corr(x,u)$ pour que le biais asymptotique de l'estimateur des VI soit inférieur à celui de l'estimateur des MCO. Dans un grand nombre d'applications, la corrélation entre l'instrument et x est inférieure à 0,2. Malheureusement, il est rare d'avoir une idée des ampleurs relatives de $Corr(z,u)$ et de $Corr(x,u)$, nous ne savons donc jamais avec certitude quel estimateur a le plus grand biais asymptotique (à moins, bien sûr, de faire l'hypothèse que $Corr(z,u) = 0$).

Dans l'exemple d'Angrist et Krueger (1991) mentionné plus haut, dans lequel x représente le nombre d'années d'études et z est une variable binaire indiquant le trimestre de naissance, la corrélation entre z et x est très faible. Bound, Jaeger et Baker (1995) proposent quelques raisons pour lesquelles le trimestre de naissance et u peuvent être un petit peu corrélés. À partir de l'équation (15.19), on voit que cela peut conduire l'estimateur des VI à avoir un biais substantiel.

Quand z et x ne sont pas corrélées du tout, les choses vont particulièrement mal, que z soit corrélée à u ou pas. L'exemple suivant illustre pourquoi on doit toujours vérifier si la variable explicative endogène est corrélée au candidat VI.

L'exemple précédent montre que l'estimation par les VI peut aboutir à des résultats étranges quand la condition de pertinence de l'instrument, $Corr(z,x) \neq 0$, n'est pas vérifiée. Ce problème, que l'on nomme problème d'**instrument faible**, présente un plus grand intérêt d'un point de vue pratique. On le définit de façon non formelle comme le problème d'avoir une corrélation « faible » (mais pas nulle) entre z et x. Il est difficile de définir à partir de quel moment un instrument faible est trop faible pour une application précise, mais la littérature théorique récente, complétée par des simulations, a apporté de nombreux éclaircissements sur ce point. Staiger et Stock (1997) ont formalisé le problème des instruments faibles en modélisant la corrélation entre z et x comme une fonction de la taille de l'échantillon. Pour être plus précis, on fait l'hypothèse que la corrélation se réduit et tend vers zéro au rythme $1/\sqrt{n}$. Évidemment, la distribution asymptotique de l'estimateur des variables instrumentales est différente des distributions asymptotiques usuelles, pour lesquelles la corrélation est supposée fixe et différente de zéro. Une des conséquences du travail de Stock-Staiger est que l'inférence statistique usuelle, qui s'appuie sur les statistiques de Student et la distribution normale standard, peut conduire à de graves erreurs. Nous discutons de ce point plus en détail dans le paragraphe 3 de ce chapitre.

EXEMPLE 15.3
Estimer l'effet du tabagisme sur le poids à la naissance

Dans le chapitre 6, nous avons estimé l'effet du tabagisme sur le poids à la naissance des enfants. Sans autres variables explicatives, le modèle est

$$\log(bwght) = \beta_0 + \beta_1 packs + u \qquad [15.21]$$

où *bwght* est le poids à la naissance et *packs* le nombre de paquets de cigarettes par jour fumées par la mère. La variable *packs* pourrait être corrélée avec d'autres facteurs de santé ou avec l'accès aux soins prénataux, *packs* et u seraient alors corrélées. Une variable instrumentale potentielle pour *packs* est le prix moyen des cigarettes dans l'État de résidence, que l'on nomme *cigprice*. Nous ferons l'hypothèse que *cigprice* et u ne sont pas corrélées (même si les subventions de l'État aux soins de santé peuvent être corrélées aux taxes sur les cigarettes).

Si les cigarettes sont un bien de consommation classique, la théorie économique de base nous explique que *packs* et *cigprice* doivent être négativement corrélées, de sorte que *cigprice* peut être utilisée comme VI de *packs*. Pour vérifier cela, on régresse *packs* sur *cigprice*, en utilisant les données de BWGHT :

$$\widehat{packs} = 0{,}067 + 0{,}0003\, cigprice$$
$$(0{,}103)\ \ (0{,}0008)$$
$$n = 1.388,\ R^2 = 0{,}0000,\ \bar{R}^2 = -0{,}0006.$$

Les résultats indiquent qu'il n'y a pas de relation entre le fait de fumer pendant la grossesse et le prix des cigarettes, ce qui n'est pas si surprenant au vu du caractère addictif du tabagisme.

Il ne faudrait pas utiliser *cigprice* comme VI de *packs* dans (15.21) car *packs* et *cigprice* ne sont pas corrélées. Mais que se passe-t-il si on le fait ? Les résultats des VI sont :

$$\widehat{\log(bwght)} = 4{,}45 + 2{,}99\, packs$$
$$(0{,}91)\ \ (8{,}70)$$
$$n = 1.388$$

(le *R*-carré correspondant est négatif). Le coefficient de *packs* est énorme et son signe n'est pas celui auquel on s'attendait. L'écart-type estimé du coefficient est également très grand, et *packs* n'est pas significatif. Ces estimations n'ont néanmoins aucun sens : *cigprice* ne vérifie pas le seul critère d'une VI que l'on peut toujours tester, l'hypothèse (15.5).

Calcul du R-carré après l'estimation VI

La plupart des programmes de régression calculent un *R*-carré après une estimation VI, en suivant la formule classique : $R^2 = 1 - SCR / SCT$, dans laquelle *SCR* est la somme des carrés des *résidus* des *VI*, et *SCT* est la somme des carrés de y. Contrairement aux MCO, le *R*-carré de l'estimation des VI peut être négatif en pratique, parce que la sommes des carrés des *résidus des VI* peut être plus grande que la somme des carrés de y. Bien que ce ne soit pas vraiment coûteux de donner le *R*-carré pour une estimation par VI, cela ne présente pas non plus grand intérêt. Quand x et u sont corrélées, nous ne pouvons pas décomposer la variance de y en $\beta_1^2 Var(x) + Var(u)$, donc le *R*-carré n'a pas d'interprétation naturelle. De plus, comme nous allons le voir dans la section 15.3, ces *R*-carrés *ne peuvent pas* être utilisés de la manière classique pour calculer les tests de Fisher de restrictions jointes.

Si notre but était de rendre le *R*-carré le plus grand possible, nous utiliserions toujours les MCO. La méthode des VI a pour but de donner de meilleures estimations pour l'effet *ceteris paribus* de x sur y quand x et u sont corrélées ; la qualité de l'ajustement n'est donc pas le but. Obtenir un *R*-carré élevé à partir d'une régression des MCO ne nous apporte pas grand-chose si on ne peut pas estimer de façon convergente β_1.

15.2 ESTIMATION DU MODÈLE DE RÉGRESSION MULTIPLE PAR VI

L'estimateur par VI pour le modèle de régression simple est facilement étendu au cas de la régression multiple. Commençons par le cas où seule une des variables explicatives est corrélée au terme d'erreur. Prenons l'exemple d'un modèle linéaire standard avec deux variables explicatives :

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \qquad \textbf{[15.22]}$$

Nous l'appelons équation structurelle pour insister sur le fait que nous nous intéressons aux β_j, ce qui veut simplement dire que l'équation a pour but de mesurer une relation causale. Nous utilisons ici une nouvelle notation pour distinguer les variables endogènes des **variables exogènes**. La variable dépendante y_1 est clairement endogène, puisqu'elle est corrélée à u_1. Les variables y_2 et z_1 sont les variables explicatives et u_1 est le terme d'erreur. Comme d'habitude, on fait l'hypothèse que la valeur espérée de u_1 est zéro : $E(u_1) = 0$. Nous utilisons z_1 pour indiquer que cette variable est exogène dans (15.22) (z_1 n'est pas corrélée à u_1). Nous utilisons y_2 pour indiquer que cette variable est suspectée d'être corrélée à u_1. Nous ne spécifions pas pourquoi y_2 et u_1 sont corrélées, mais pour le moment le plus simple est de penser que u_1 contient une variable omise corrélée à y_2. L'équation (15.22) doit ses origines aux modèles d'équations simultanées (que nous allons voir dans le chapitre 16), mais nous l'utilisons dans un cadre plus général pour distinguer facilement les variables explicatives exogènes des variables explicatives endogènes dans un modèle de régression multiple.

Pour illustrer (15.22), prenons par exemple :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u_1 \qquad \textbf{[15.23]}$$

Où $y_1 = \log(wage)$, $y_2 = educ$ et $z_1 = exper$. En d'autres termes, on fait l'hypothèse que *exper* est exogène dans (15.23), mais nous autorisons *educ* – pour les raisons habituelles – à être corrélée avec u_1.

Nous savons que si (15.22) est estimée par MCO, *tous* les estimateurs seront biaisés et non convergents. Nous suivons donc la stratégie suggérée dans la section précédente et nous partons à la recherche d'un instrument pour y_2. Puisque nous avons fait l'hypothèse que z_1 n'est pas corrélée à u_1, peut-on utiliser z_1 comme instrument pour y_2, en admettant que y_2 et z_1 soient corrélées ? La réponse est non. Puisque z_1 apparaît lui aussi comme variable explicative dans (15.22), il ne peut pas faire office de variable instrumentale pour y_2. Nous avons besoin d'une autre variable exogène – appelons-la z_2 – qui *n'intervient pas* dans (15.22). Par conséquent, l'hypothèse cruciale est que z_1 et z_2 ne sont pas corrélées à u_1, on suppose aussi que la valeur espérée de u_1 est zéro, ce qui est sans perte de généralité quand l'équation a une constante :

$$E(u_1) = 0,\ \mathrm{Cov}(z_1, u_1) = 0,\ \mathrm{Cov}(z_2, u_1) = 0 \qquad \textbf{[15.24]}$$

Puisqu'on a fait l'hypothèse que la valeur espérée de u_1 était égale à zéro, les deux dernières conditions sont équivalentes à $E(z_1u_1) = E(z_2u_1) = 0$, donc en s'inspirant de la méthode des moments, on obtient des estimateurs $\hat{\beta}_0$, $\hat{\beta}_1$, et $\hat{\beta}_2$ en résolvant les équations empiriques équivalentes à celles données en (15.24), c'est-à-dire :

$$\sum_{i=1}^{n}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0 \qquad \textbf{[15.25]}$$

Il s'agit d'un système de trois équations linéaires pour les trois inconnues $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\beta}_2$, que l'on peut facilement résoudre étant donnée l'information sur y_1, y_2, z_1 et z_2. Ces estimateurs sont appelés *estimateurs par variables instrumentales*. Si on pense que y_2 est exogène et qu'on choisit $z_2 = y_2$, les équations en (15.25) sont exactement les conditions de premier ordre qu'on a vues pour les estimateurs des MCO (voir équations en [3.13]).

Ici aussi, il est nécessaire que la variable instrumentale z_2 soit corrélée à y_2, mais la présence de z_1 dans l'équation (15.22) rend plus compliqué le type de corrélation qui doit exister entre ces deux variables. Il nous faut maintenant écrire l'hypothèse en termes de corrélation *partielle*. La façon la plus simple d'écrire

la condition est d'exprimer la variable explicative endogène comme une fonction linéaire des variables exogènes et du terme d'erreur :

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2 \qquad \textbf{[15.26]}$$

Où, par construction, $E(v_2) = 0$, $Cov(z_1, v_2) = 0$ et $Cov(z_2, v_2) = 0$ et les π_j sont des paramètres inconnus. La condition d'identification cruciale (avec celle en (15.24)) est :

$$\pi_2 \neq 0 \qquad \textbf{[15.27]}$$

En d'autres termes, y_2 et z_2 sont encore corrélées, après qu'on a pris en compte l'influence de z_1. Cette corrélation peut être positive ou négative, mais elle ne doit pas être égale à zéro. Il est facile de tester (15.27) : on estime (15.26) par MCO et on utilise un test de Student (éventuellement en prenant en compte l'hétéroscédasticité). Il faudrait toujours tester cette hypothèse. Malheureusement, nous ne pouvons pas tester si z_1 et z_2 ne sont pas corrélées à u_1 ; nous espérons pouvoir le justifier en se basant sur un raisonnement économique ou par introspection.

L'équation (15.26) est un exemple d'**équation en forme réduite**, ce qui signifie qu'on a écrit une variable endogène en termes de variables exogènes. Cette appellation vient des modèles d'équations simultanées (que nous étudierons dans le prochain chapitre), mais c'est un concept utile dès lors que nous avons affaire à une variable explicative endogène. Cette appellation nous permet de distinguer ce type d'équation de l'équation structurelle en (15.22).

Pour aller plus loin 15.2

Imaginons que nous voulions estimer l'effet de la consommation de cannabis sur la moyenne des notes à l'université, au niveau licence. Pour la population des étudiants en fin de licence, nous appelons *daysused* le nombre de jours du dernier mois au cours desquels un étudiant a fumé du cannabis et intéressons-nous à l'équation structurelle, où *colGPA* est la note moyenne de l'étudiant en licence et *SAT* est la note moyenne de l'étudiant obtenue au concours d'entrée à l'université :

$$colGPA = \beta_0 + \beta_1 daysused + \beta_2 SAT + u$$

i. On appelle *percHS* le pourcentage, pour un étudiant de lycée, d'élèves de sa classe de terminale qui déclaraient consommer régulièrement du cannabis. Si c'est un candidat de VI pour *daysused*, écrivez la forme réduite de *daysused*. Pensez-vous que (15.27) est vérifiée ?
ii. Pensez-vous que *percHS* est vraiment exogène dans l'équation structurelle ? Quels problèmes peuvent être rencontrés ?

Il est très facile d'ajouter plus de **variables explicatives exogènes** au modèle. On peut écrire le modèle structurel ainsi :

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1 \qquad \textbf{[15.28]}$$

dans lequel on pense que y_2 est corrélée à u_1. On appelle z_k une variable qui n'apparaît pas dans (15.28) mais qui est aussi exogène. Ainsi, nous faisons l'hypothèse que

$$E(u_1) = 0, \text{Cov}(z_j, u_1) = 0, \qquad j = 1, \ldots, k \qquad \textbf{[15.29]}$$

Sous l'hypothèse (15.29), $z_1, \ldots, z_{k-1}$ sont les variables exogènes qui apparaissent dans (15.28). En effet, celles-ci jouent le rôle de leurs propres instruments pour estimer les β_j dans (15.28). Le cas particulier où $k = 2$ est donné par les équations (15.25) : en plus de z_2, z_1 apparaît dans l'ensemble des conditions de moment utilisées pour obtenir les estimations par VI. En règle générale, $z_1, \ldots, z_{k-1}$ sont utilisées dans les conditions de moment en plus de la variable instrumentale pour y_2, z_k.

La forme réduite de y_2 est :

$$y_2 = \pi_0 + \pi_1 z_1 + \ldots + \pi_{k-1} z_{k-1} + \pi_k z_k + v_2 \qquad [15.30]$$

Et nous avons besoin de la corrélation partielle entre z_k et y_2 :

$$\pi_k \neq 0 \qquad [15.31]$$

Sous les hypothèses (15.29) et (15.31), z_k est un instrument valide pour y_2. (On ne se soucie pas des autres π_j dans (15.30), certains d'entre eux peuvent être égaux à zéro.) On fait également l'hypothèse mineure qu'il n'existe pas de relation linéaire parfaite entre les variables exogènes, ce qui est analogue à l'hypothèse d'absence de colinéarité parfaite dans le contexte des MCO.

Pour l'inférence statistique standard, nous avons besoin de faire l'hypothèse que u_1 est homoscédastique. Nous exprimons ces hypothèses de manière attentive dans un contexte plus général dans la section 15.3.

Tableau 15.1 *Variable dépendante : log(wage)*

Variables explicatives	MCO	VI
Educ	0,075 (0,003)	0,132 (0,055)
Exper	0,085 (0,007)	0,108 (0,024)
exper²	−0,0023 (0,0003)	−0,0023 (0,0003)
Noir	−0,199 (0,018)	−0,147 (0,054)
Aire Urbaine	0,136 (0,020)	0,112 (0,032)
Sud	−0,148 (0,026)	−0,145 (0,027)
Observations *R-carré*	3010 0,300	3010 0,238
Autres variables explicatives : *smsa66, reg662, …, reg669*		

EXEMPLE 15.4
Utiliser la proximité de l'université comme VI pour l'éducation

Card (1995) utilise les données pour un échantillon d'hommes en 1976 pour estimer les rendements de l'éducation. Il utilise une variable binaire indiquant si la personne a grandi à proximité d'une université (*nearc4*) comme variable instrumentale pour l'éducation. Il considère une équation de type *log(wage)*, dans laquelle il tient compte de l'influence d'autres variables classiques : expérience, une variable indiquant si la personne est noire, une indicatrice indiquant si la personne vit dans une aire urbaine et une autre indiquant si la personne vit dans le sud des États-Unis, un ensemble complet de variables binaires pour indiquer chacune des régions et une variable binaire indiquant si la personne vivait dans une aire urbaine en 1966. Pour que

nearc4 soit un instrument valide, elle ne doit pas être corrélée au terme d'erreur dans l'équation de salaire (ce que l'on suppose), et elle doit être corrélée de façon partielle à *educ*. Afin de vérifier cette dernière condition, nous régressons *educ* sur *nearc4* et l'ensemble des variables exogènes qui apparaissent dans l'équation. (Pour le dire autrement, nous estimons la forme réduite de *educ*.) En utilisant les données de CARD, on obtient, en forme condensée :

$$\widehat{educ} = 16{,}64 + 0{,}320\ nearc4 - 0{,}413\ exper + \ldots$$
$$(0{,}24)\ \ (0{,}088)\qquad\qquad (0{,}034)$$
$$n = 3{,}010,\ R^2 = 0{,}477.$$

Nous nous intéressons au coefficient et à la statistique de Student de *nearc4*. Le coefficient nous enseigne que, en 1976, toutes choses égales par ailleurs (expérience, origine ethnique, région, etc.), les gens qui vivaient près d'une université en 1966 avaient un tiers d'année d'éducation de plus que ceux qui n'ont pas grandi près d'une université. La statistique de Student pour *nearc4* est de 3,64, ce qui donne une *p-value* dont les trois premières décimales sont nulles. Par conséquent, si *nearc4* n'est pas corrélée aux facteurs inobservés dans le terme d'erreur par ailleurs, on peut utiliser *nearc4* comme VI pour *educ*.

Les estimations par VI et par MCO sont données dans le tableau 15.1. À l'instar des écarts-types estimés par MCO, il est nécessaire d'ajuster les écarts-types estimés par VI pour le nombre de degrés de liberté afin d'estimer la variance de l'erreur. Certains programmes statistiques font automatiquement l'ajustement pour le nombre de degrés de liberté, mais d'autres ne le font pas.

Il est intéressant de constater que l'estimation par VI des rendements de l'éducation est près de deux fois plus grande que l'estimation par MCO, mais l'écart-type estimé du coefficient de l'estimation par VI est plus de 18 fois plus grand que l'écart-type estimé du coefficient de l'estimation par MCO. L'intervalle de confiance à 95 % pour l'estimation par VI est entre 0,024 et 0,239, ce qui représente une ampleur très large. Quand on pense qu'*educ* est endogène, la présence d'intervalles de confiance très grands est le prix à payer pour obtenir une estimation convergente des rendements de l'éducation.

Comme nous l'avons discuté précédemment, nous ne devrions pas nous attarder sur le fait que le *R*-carré est plus petit dans l'estimation par VI : par définition, le *R*-carré obtenus par MCO sera toujours plus grand parce que les MCO minimisent la somme du carré des résidus.

Il est important de noter, surtout lorsqu'il s'agit d'étudier les effets d'une intervention de politique publique, qu'il existe également une forme réduite pour y_1. Dans le cadre de l'équation (15.28) avec z_k comme VI pour y_2, la forme réduite de y_1 prend toujours la forme de :

$$y_1 = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_k z_k + e_1 \qquad \textbf{[15.32]}$$

où $\gamma_j = \beta_{j+1} + \beta_1 \pi_j$ pour $j < k$, $\gamma_k = \beta_1 \pi_k$ et $e_1 = u_1 + \beta_1 v_2$, ce qu'on peut vérifier en remplaçant (15.30) dans (15.28) et en réécrivant les termes. Puisque les z_k sont exogènes dans (15.32), les γ_j peuvent être estimés de manière convergente par MCO. En d'autres termes, nous régressons y_1 sur toutes les variables exogènes, y compris z_k, la VI de y_2. Ce n'est que si nous voulons estimer β_1 dans (15.28) que l'on a besoin de mettre en application les VI.

Quand y_2 est une variable 0-1 qui indique la participation à un programme et que z_k est une variable 0-1 qui indique l'éligibilité pour participer à ce programme – qui est, nous l'espérons, soit distribué de manière aléatoire entre les individus ou, sinon, une fonction des autres variables exogènes $z_1, \ldots, z_{k-1}$ (comme le revenu) – l'interprétation du coefficient γ_k est intéressante. Au lieu d'être une estimation de l'effet du programme lui-même, il représente une estimation de l'effet d'*offrir* le programme. Contrairement à β_1 dans (15.28), qui mesure les effets du programme lui-même, γ_k prend en compte le fait que certaines unités éligibles choisissent de ne pas participer au programme. Dans la littérature de l'évaluation de politiques publiques,

γ_k est un exemple de paramètre représentant l'*intention-de-traiter* : il mesure l'effet d'être *éligible* et non l'effet d'avoir effectivement participé. Le coefficient d'intention-de-traiter, $\gamma_k = \beta_1 \pi_k$, dépend de l'effet de la participation, β_1, et de la variation (en général l'accroissement) de la probabilité de participation due à l'éligibilité, π_k. (Quand y_2 est binaire, l'équation (15.30) est un modèle de probabilité linéaire, et par conséquent π_k mesure la variation *ceteris paribus* de la probabilité que $y_2 = 1$ quand z_k passe de zéro à un.)

15.3 LES DOUBLES MOINDRES CARRÉS

Dans la section précédente, nous avons fait l'hypothèse que nous avions une seule variable explicative endogène (y_2), avec une variable instrumentale pour y_2. Il arrive fréquemment que nous disposions de plus d'une variable exogène exclue du modèle structurel et potentiellement corrélées à y_2, ce qui signifie que ce sont des instruments valides pour y_2. Dans cette section, nous présentons comment utiliser plusieurs variables instrumentales en même temps.

Une seule variable explicative endogène

Intéressons-nous au modèle structurel (15.22), qui a une variable explicative endogène et une variable explicative exogène. Supposons maintenant que l'on ait *deux* variables exogènes exclues de (15.22) : z_2 et z_3. Nos hypothèses selon lesquelles z_2 et z_3 n'apparaissent pas dans (15.22) et ne sont pas corrélées au terme d'erreur u_1 sont appelées **restrictions d'exclusion**.

Si z_2 et z_3 étaient toutes les deux corrélées à y_2, nous pourrions simplement utiliser chacune d'entre elles comme instrument, comme dans la section précédente. Mais alors nous aurions deux estimateurs VI et aucun des deux, en général, ne serait efficace. Puisqu'aucune des variables z_1, z_2 et z_3 n'est corrélée à u_1, aucune combinaison linéaire de ces trois variables n'est corrélée à u_1, et donc toute combinaison linéaire des variables exogènes est une VI valide. Pour avoir la meilleure VI, nous choisissons celle qui est la plus fortement corrélée à y_2. Il se trouve que celle-ci nous est donnée par l'équation de forme réduite de y_2. Nous écrivons :

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2 \qquad \textbf{[15.33]}$$

où

$$E(v_2) = 0, Cov(z_1, v_2) = 0 \; Cov(z_2, v_2) = 0, Cov(z_3, v_2) = 0$$

Ainsi, la meilleure VI pour y_2 (sous les hypothèses explicitées dans l'annexe du chapitre) est la combinaison linéaire des z_j dans (15.33), que l'on appelle y_2^* :

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 \qquad \textbf{[15.34]}$$

Pour que cette VI ne soit pas parfaitement corrélée à z_1, il faut qu'au moins l'un des π_2 ou π_3 soit différent de zéro :

$$\pi_2 \neq 0 \text{ ou } \pi_3 \neq 0 \qquad \textbf{[15.35]}$$

Cette hypothèse est celle qui permet l'identification, une fois que l'on a supposé que les z_j sont toutes exogènes. (La valeur de π_1 est sans importance.) L'équation structurelle (15.22) n'est pas identifiée si $\pi_2 = 0$ et $\pi_3 = 0$. On peut tester $H_0 : \pi_2 = 0$ et $\pi_3 = 0$ contre (15.35) en utilisant une statistique de Fisher.

Il peut être utile de regarder l'équation (15.33) sous cet angle : on sépare y_2 en deux parties. La première partie est y_2^*, la part de y_2 qui n'est pas corrélée au terme d'erreur u_1. La seconde partie est v_2, qui est potentiellement corrélée à u_1, raison pour laquelle y_2 est probablement endogène.

Avec l'information sur les z_j, nous pouvons calculer y_2^* pour chaque observation, à condition de connaître les paramètres π_j de la population. En pratique, ce n'est jamais le cas. Nous pouvons cependant toujours estimer la forme réduite par MCO, comme nous l'avons vu dans la section précédente. Ainsi, en utilisant l'échantillon à disposition, nous régressons y_2 sur z_1, z_2 et z_3 et nous obtenons les valeurs prédites :

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3 \qquad \textbf{[15.36]}$$

(c'est-à-dire qu'on obtient $\hat{y}_{i2}$ pour chaque i). À ce stade, il faudrait vérifier que z_2 et z_3 sont significatifs de manière jointe dans (15.33) à un niveau de significativité suffisamment petit (pas plus grand que 5 %). Si z_2 et z_3 ne sont pas significatifs de manière jointe dans (15.33) alors une estimation par VI nous fait juste perdre notre temps.

Une fois que l'on a $\hat{y}_2$, on peut l'utiliser comme VI pour y_2. On utilise, pour estimer β_0, β_1 et β_2 les trois équations suivantes : les deux premières équations de (15.25), et la troisième est remplacée par

$$\sum_{i=1}^{n} \hat{y}_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0 \qquad \textbf{[15.37]}$$

La résolution du système de trois équations pour les trois inconnues nous donne les estimateurs des variables instrumentales.

Quand on dispose de plusieurs instruments, l'estimateur des variables instrumentales qui utilise $\hat{y}_{i2}$ comme instrument est aussi appelé **estimateur des double moindres carrés (DMC)** ou estimateur des moindres carrés en deux étapes. La raison est simple. En utilisant l'algèbre des MCO, on peut montrer qu'en utilisant $\hat{y}_2$ comme variable instrumentale de y_2, les estimateurs par variable instrumentale $\hat{\beta}_0$, $\hat{\beta}_1$ et $\hat{\beta}_2$ sont *identiques* à l'estimateur des MCO obtenu par la régression de

$$y_1 \text{ sur } \hat{y}_2 \text{ et } z_1 \qquad \textbf{[15.38]}$$

En d'autres termes, nous pouvons obtenir l'estimateur des DMC en deux étapes. La première étape consiste à faire tourner la régression en (15.36), à partir de laquelle on obtient les valeurs ajustées $\hat{y}_{2i}$. La seconde étape est la régression par les MCO (15.38). Puisqu'on utilise $\hat{y}_2$ au lieu de y_2, les estimations par les DMC peuvent être très différentes des estimations par les MCO.

Certains économistes interprètent la régression en (15.38) de la façon suivante : la valeur ajustée $\hat{y}_2$ est la version estimée de y_2^* et y_2^* n'est pas corrélée à u_1. Par conséquent, dans un premier temps, les doubles moindres carrés « purgent » y_2 de sa corrélation à u_1 avant de procéder à une régression MCO en (15.38). On peut montrer cela en introduisant $y_2 = y_2^* + v_2$ dans (15.22) :

$$y_1 = \beta_0 + \beta_1 y_2^* + \beta_2 z_1 + u_1 + \beta_1 v_2 \qquad \textbf{[15.39]}$$

Le terme d'erreur composite $u_1 + \beta_1 v_2$ a maintenant une moyenne nulle et il n'est corrélé ni à y_2^* ni à z_1 : c'est pour cela que l'on peut utiliser les MCO.

La plupart des logiciels d'économétrie ont des commandes spéciales pour les doubles moindres carrés, il n'est donc pas nécessaire de réaliser explicitement les deux étapes. En fait, dans la plupart des cas, il vaut mieux éviter de faire soi-même manuellement la seconde étape, car les écarts-types estimés des coefficients et les statistiques de test obtenues de cette façon *ne sont pas* valides. (Cela vient du fait que le terme d'erreur dans (15.39) inclut v_2, mais seule la variance de u_1 intervient pour les écarts-types des coefficients.) Tous les programmes qui permettent de calculer une régression par les doubles moindres carrés demandent de spécifier la variable dépendante, la liste des variables explicatives (qu'elles soient endogènes ou exogènes), et la liste complète des variables instrumentales (c'est-à-dire toutes les variables exogènes). La présentation des résultats est en général assez similaire à celle des MCO.

Dans le modèle (15.28), avec une seule variable instrumentale pour y_2, l'estimateur des variables instrumentales de la section 15.2 est identique à l'estimateur par doubles moindres carrés. Par conséquent, lorsque l'on a une variable instrumentale pour chaque variable explicative endogène, on peut appeler la méthode d'estimation « par variable instrumentale » ou « par doubles moindres carrés ».

Ajouter plus de variables exogènes ne change pas grand-chose. Par exemple, supposons que l'équation de salaire (*wage*) soit :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1 \qquad \textbf{[15.40]}$$

où u_1 n'est corrélée ni à *exper*, ni à $exper^2$. Supposons qu'on pense également que ni l'éducation de la mère (*motheduc*) et ni celle du père (*fatheduc*) ne sont corrélées à u_1. Alors on peut les utiliser toutes les deux comme variable instrumentale pour *educ*. L'équation de forme réduite pour *educ* est :

$$educ = \pi_0 + \pi_1 exper + \pi_2 exper^2 + \pi_3 motheduc + \pi_4 fatheduc + v_2 \qquad \textbf{[15.41]}$$

et il est nécessaire que $\pi_3 \neq 0$ ou $\pi_4 \neq 0$ (ou les deux, bien entendu) pour l'identification.

EXEMPLE 15.5
Les rendements de l'éducation pour les femmes actives en emploi

Nous estimons l'équation (15.40) en utilisant les données de MROZ. Tout d'abord, nous testons $H_0 : \pi_3 = 0, \pi_4 = 0$ dans (15.41) en utilisant un test de Fisher (F-test). Le résultat nous indique : $F = 55{,}40$, et la *p-valeur* = 0,0000. Comme on s'y attendait, *educ* est partiellement corrélée à l'éducation des parents.

En estimant (15.40) par doubles moindres carrés, nous obtenons les résultats suivants, présentés sous forme d'équation :

$$\widehat{\log(wage)} = .048 + .061\, educ + .044\, exper - .0009\, exper^2$$
$$(.400)\ (.031)\qquad (.013)\qquad (.0004)$$
$$n = 428,\ R^2 = .136.$$

On estime que les rendements de l'éducation sont de 6,1 %, résultat qu'il faut comparer à l'estimation par les MCO, qui est d'à peu près 10,8 %. Les écarts-types estimés des coefficients sont relativement grands, donc l'estimation DMC est à peine statistiquement significative au niveau de 5 %, contre l'hypothèse alternative bilatérale.

Les hypothèses nécessaires pour que l'estimateur par DMC ait les propriétés asymptotiques désirées sont explicitées dans l'annexe du chapitre, mais il est utile de les résumer rapidement ici. Réécrivons l'équation structurelle déjà vue en (15.28) :

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1 \qquad \textbf{[15.42]}$$

Nous faisons l'hypothèse qu'aucun des z_j n'est corrélé à u_1. De plus, nous avons besoin d'au moins une variable exogène qui soit *absente* de (15.42) et qui soit partiellement corrélée à y_2. Ces hypothèses nous assurent la convergence. Pour que les écarts-types estimés usuels des estimateurs par doubles moindres carrés et les statistiques de Student soient asymptotiquement valides, il nous faut aussi une hypothèse d'homoscédasticité : la variance du terme d'erreur structurel u_1 ne doit dépendre d'aucune variable exogène. Pour les applications sur les séries temporelles, nous avons besoin de plus d'hypothèses, ce que nous verrons dans la section 15.7.

Multicolinéarité et DMC

Dans le chapitre 3, nous avons présenté le problème de la multicolinéarité et nous avons montré comment la corrélation entre les régresseurs peut conduire à de grands écarts-types estimés pour les estimations MCO. La multicolinéarité peut causer des problèmes encore plus graves dans le cadre des doubles moindres carrés. Afin de comprendre pourquoi, indiquons que la variance (asymptotique) de l'estimateur de β_1 par DMC peut être approximée par

$$\sigma^2 / [\widehat{SCT}_2(1 - \hat{R}_2^2)] \qquad \textbf{[15.43]}$$

Où $\sigma^2 = Var(u_1)$, $\widehat{SCT}_2$ est la variation totale de $\hat{y}_2$ et $\hat{R}_2^2$ est le R-carré d'une régression de $\hat{y}_2$ sur toutes les autres variables exogènes qui apparaissent dans l'équation structurelle. La variance de l'estimateur par DMC est plus grande que celle de l'estimateur par MCO pour deux raisons. D'abord, $\hat{y}_2$ a moins de variation que y_2 par construction. (En guise de rappel : Somme totale des carrés = somme des carrés de la partie expliquée + somme des carrés des résidus, la variation de y_2 est la somme totale des carrés, alors que la variation de $\hat{y}_2$ est la somme des carrés de la partie expliquée issue de la régression de première étape.) De plus, la corrélation entre $\hat{y}_2$ et les variables exogènes dans (15.42) est souvent beaucoup plus grande que la corrélation entre y_2 et ces mêmes variables. Le problème de la multicolinéarité dans les DMC vient de cette seconde raison.

Afin d'illustrer le problème, prenons l'exemple 15.4. Quand on régresse la variable *educ* sur les variables exogènes du tableau 15.1 (desquelles on a exclu la variable *nearc4*), le R-carré est de 0,475. C'est un niveau de multicolinéarité raisonnable, mais l'important est que l'écart-type estimé de $\hat{\beta}_{educ}$ obtenu par MCO est assez faible. Lorsque l'on construit les valeurs ajustées $\widehat{educ}$ à partir de la première étape et qu'on la régresse sur les variables exogènes qui apparaissent dans le tableau 15.1, on trouve un R-carré de 0,995, ce qui correspond à un très haut niveau de multicolinéarité entre $\widehat{educ}$ et les autres variables exogènes du tableau. (Le fait que le R-carré est grand n'est pas très surprenant puisque $\widehat{educ}$ est fonction de toutes les variables exogènes du tableau 15.1, en plus de *nearc4*.) L'équation (15.43) montre que si $\hat{R}_2^2$ est proche de un, alors l'écart-type estimé de l'estimateur par DMC peut être très grand. Mais comme pour les MCO, on peut compenser un grand $\hat{R}_2^2$ par un échantillon de très grande taille.

Détecter des instruments faibles

Dans la section 15.1 de ce chapitre, nous avons brièvement abordé le problème des instruments faibles. Nous nous sommes concentrés sur l'équation (15.19), qui montre comment une faible corrélation entre l'instrument et l'erreur peut conduire à un très grand écart (et donc à un biais) si l'instrument z est également peu corrélé avec la variable explicative x. Le même problème peut se poser dans le contexte du modèle à équations multiples présenté dans l'équation (15.42), que nous disposions d'un seul instrument pour y_2 ou de plus d'instruments que nécessaire.

Nous avons également mentionné les conclusions de Staiger et Stock (1997), et nous discutons maintenant plus en profondeur des implications pratiques de cette recherche. Il est important de noter que Staiger et Stock étudient le cas où toutes les variables instrumentales sont exogènes. L'hypothèse d'exogénéité des instruments étant satisfaite, ils se concentrent sur le cas où les instruments sont faiblement corrélés avec y_2 et ils étudient la validité des écarts types, des intervalles de confiance et des statistiques t portant sur le coefficient β_1 sur y_2. Le mécanisme qu'ils ont utilisé pour modéliser la faible corrélation a mené à une découverte importante : même avec des échantillons de très grandes tailles, l'estimateur par DMC peut être biaisé et présenter une distribution très différente de la loi normale.

En s'appuyant sur Staiger et Stock (1997), Stock et Yogo (2005) (SY par la suite) ont proposé des méthodes pour détecter les situations où des instruments faibles pourraient entraîner un biais substantiel et une inférence statistique faussée. En pratique, Stock et Yogo ont établi des règles quant à la taille nécessaire

de la statistique t (avec un seul instrument) ou de la statistique F (avec plus d'un instrument) obtenues à partir de la régression de première étape. La théorie est beaucoup trop compliquée pour être expliquée en détail ici. Au lieu de cela, nous allons voir quelques règles empiriques simples proposées par Stock et Yogo qui sont faciles à mettre en œuvre.

L'implication principale du travail de Stock et Yogo est que rejeter statistiquement l'hypothèse nulle dans la régression de première étape à des niveaux de significativité habituels n'est pas suffisant. Par exemple, dans l'équation (15.6), il ne suffit pas de rejeter l'hypothèse nulle énoncée dans (15.7) au seuil de significativité de 5 %. En utilisant les calculs de biais pour l'estimateur des variables instrumentales, SY recommandent de procéder à l'inférence VI de façon habituelle seulement si la statistique de test t issue de la régression de première étape a une valeur absolue supérieure à $\sqrt{10} \approx 3,2$. Les lecteurs reconnaîtront que cette valeur est bien supérieure au 95e percentile de la distribution normale standard, soit 1,96, ce qui correspondrait à un seuil de significativité standard de 5 %. Cette même règle empirique s'applique au modèle de régression multiple avec une seule variable explicative endogène, y_2, et une seule variable instrumentale, z_k. En particulier, la statistique t dans l'hypothèse de test (15.31) doit être au moins égale à 3,2 en valeur absolue.

Stock et Yogo traitent aussi du cas de l'estimation par DMC. Dans ce cas, nous devons nous concentrer sur la statistique de la régression de première étape F pour exclure les variables instrumentales pour y_2, et la règle fixée par SY est $F > 10$. (Remarque : il s'agit en fait de la même règle que celle basée sur la statistique t lorsqu'il n'y a qu'un seul instrument, puisque $t^2 = F$.) Par exemple, considérons l'équation (15.34), où nous avons deux instruments pour y_2, z_2 et z_3. Alors la statistique de test F, du test de l'hypothèse nulle :

$$H_0 : \pi_2 = 0, \ \pi_3 = 0$$

devrait être telle que $F > 10$. Rappelez-vous que ce n'est pas la statistique F globale pour toutes les variables exogènes de (15.34). Nous testons uniquement les coefficients des VI proposés pour y_2, c'est-à-dire les variables exogènes qui n'apparaissent pas dans (15.22). Dans l'exemple 15.5, la statistique F pertinente est de 124,76, qui est donc bien supérieure à 10, ce qui signifie que nous n'avons pas à nous soucier des instruments faibles. (Par contre, nous doutons évidemment de l'exogénéité des variables relatives à l'éducation des parents.)

La règle empirique qui veut que la statistique F soit supérieure à 10 fonctionne bien dans la plupart des modèles et elle est facile à retenir. Cependant, comme toutes les règles empiriques impliquant une inférence statistique, cela n'a pas de sens d'utiliser 10 comme un seuil de façon tranchée. Par exemple, on peut probablement continuer son analyse si $F = 9,94$, car on est assez proche de 10. La règle empirique devrait être utilisée comme une ligne directrice. Stock et Yogo ont des suggestions plus détaillées pour les cas où il y a beaucoup d'instruments pour y_2, par exemple cinq ou plus. Le lecteur intéressé pourra consulter l'article de Stock et Yogo. La plupart des chercheurs en économie appliquée adoptent 10 comme valeur de référence.

Plusieurs variables explicatives endogènes

Les doubles moindres carrés peuvent également être utilisés pour estimer des modèles dans lesquels on compte plusieurs variables explicatives endogènes. Par exemple, on peut examiner le modèle

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1 \qquad \textbf{[15.44]}$$

Où $E(u_1) = 0$ et u_1 n'est corrélée à aucun des z_1, z_2 ou z_3. Les variables y_2 et y_3 sont des variables explicatives endogènes : chacune d'elles peut être corrélée à u_1.

Pour estimer (15.44) par DMC, il nous faut *au moins deux* variables exogènes qui n'apparaissent pas dans (15.44) mais qui sont corrélées à y_2 et y_3. Supposons d'abord que l'on dispose de deux variables exogènes exclues, appelées z_4 et z_5. Ensuite, en suivant notre analyse du cas avec une seule variable explicative endogène, il faut que l'une ou l'autre de z_4 et z_5 apparaisse dans la forme réduite de y_2 et dans la forme réduite

de y_3. (Comme dans le cas précédent, on peut tester cette condition au moyen d'un test de Fisher.) Même si cette condition est nécessaire pour l'identification, elle n'est malheureusement pas suffisante. Imaginons que z_4 apparaisse dans les deux formes réduites, mais que z_5 n'apparaisse dans aucune des deux. Alors, nous n'avons pas vraiment deux variables partiellement corrélées à y_2 et y_3. Les doubles moindres carrés ne nous permettront donc pas d'obtenir d'estimateurs convergents des β_j.

En règle générale, quand on a plus d'une variable explicative endogène dans un modèle de régression, pour plusieurs raisons compliquées, on peut très bien ne pas aboutir à l'identification. Mais on peut facilement exprimer une condition nécessaire pour obtenir l'identification, que l'on appelle la **condition d'ordre**.

Condition d'ordre pour l'identification d'une équation. Il nous faut au moins autant de variables exogènes exclues qu'il y a de variables explicatives endogènes incluses dans l'équation structurelle. La condition d'ordre est facile à vérifier, puisqu'il s'agit simplement de compter le nombre de variables endogènes et de variables exogènes. La condition suffisante pour l'identification est appelée **condition de rang**. Nous avons évoqué des cas particuliers de la condition de rang plus haut, par exemple lorsque nous avons discuté l'équation (15.35). Nous avons besoin, pour énoncer la condition de rang dans un cadre général, de faire appel à l'algèbre matriciel, ce qui va au-delà des objectifs de l'ouvrage. (Voir Wooldridge, 2010, chapitre 5).

Pour aller plus loin 15.3

Pour expliquer les taux d'agressions à main armée (*violent*), au niveau d'une ville, on fait appel au modèle suivant, dans lequel la variable *controlgun* est une variable binaire qui indique s'il existe des lois pour le contrôle des armes. D'autres variables permettent de tenir compte de l'influence du taux de chômage (*unem*), de la densité (*popul*), du pourcentage de noirs (*percblck*), du pourcentage de jeunes qui ont entre 18 et 21 ans parmi la population (*age18_21*), auxquelles on ajoute d'autres variables dont on veut prendre en compte l'influence.

$$violent = \beta_0 + \beta_1 controlgun + \beta_2 unem + \beta_3 popul + \beta_4 percblck + \beta_5 age18\_21 + \ldots$$

Des chercheurs ont estimé des équations similaires en utilisant comme instruments pour *controlgun* des variables comme : le nombre de membres de la *National Rifle Association*[2] dans la ville ou le nombre d'abonnés à des magazines sur les armes à feu (voir par exemple Kleck et Patterson, 1993). Ces instruments vous paraissent-ils convaincants ?

Test d'hypothèses multiples après une estimation par DMC

Il faut être prudent quand on teste des hypothèses multiples dans un modèle estimé par DMC. Il est tentant d'utiliser soit la somme des carrés des résidus soit la forme *R*-carré de la statistique *F*, comme nous l'avons vu dans le cadre des MCO dans le chapitre 4. Le *R*-carré dans le cadre des DMC peut être négatif, ce qui laisse à penser que la façon usuelle de calculer les statistiques *F* n'est peut-être pas appropriée : c'est effectivement le cas. En fait, si on utilise les résidus de l'estimation par DMC pour calculer la somme des carrés des résidus pour les modèles contraints et non contraints, il n'est pas garanti que la somme des carrés des résidus de l'estimation du modèle contraint soit supérieure à la somme des carrés des résidus de l'estimation du modèle non contraint. Si on se trouve dans le cas contraire, on pourrait avoir une statistique *F* négative.

Il est possible de combiner la somme des carrés des résidus de la régression de seconde étape (comme celle en (15.38)) avec la somme des carrés des résidus de l'estimation du modèle non contraint pour obtenir une statistique qui présente approximativement une distribution *F* pour les grands échantillons. Nous ne présentons pas les détails ici : en effet, la plupart des logiciels d'économétrie ont des programmes proposant

2 Association à but non lucratif dont le but est de promouvoir les armes à feu aux États-Unis (note de la traduction).

des commandes de test faciles d'utilisation qui peuvent être utilisées pour tester des hypothèses multiples après une estimation par DMC. Davidson et MacKinnon (1993) et Wooldridge (2010 ; chapitre 5) expliquent comment calculer une sorte de statistique F dans le cadre des DMC.

15.4 SOLUTION DES VI AUX PROBLÈMES D'ERREUR DE MESURE SUR LES RÉGRESSEURS

Dans les précédentes sections, nous avons vu comment utiliser les variables instrumentales pour résoudre le problème des variables omises, mais on peut aussi les utiliser pour traiter le problème d'erreur de mesure sur les régresseurs. Pour illustrer le problème, prenons le modèle

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u \qquad \textbf{[15.45]}$$

Où y et x_2 sont observées mais pas x_1^*. On appelle x_1 une mesure observée de x_1^* : $x_1 = x_1^* + e_1$, avec e_1 l'erreur de mesure. Nous avons vu, dans le chapitre 9, que les MCO sont biaisés et ne convergent pas lorsqu'on utilise x_1 à la place de x_1^* et que x_1 et e_1 sont corrélés. Il est possible de voir cela en écrivant :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1) \qquad \textbf{[15.46]}$$

Si les hypothèses classiques sur les erreurs de mesures sont vérifiées, l'estimateur par MCO de β_1 présente un biais d'atténuation (ou biais vers zéro, c'est-à-dire que l'ampleur du coefficient est sous-estimée). On ne peut rien faire pour régler ce problème sans hypothèse supplémentaire.

Il est possible, dans certains cas, d'utiliser la méthode des variables instrumentales pour résoudre le problème d'erreur de mesure. Dans (15.45), nous faisons l'hypothèse que u n'est pas corrélée à x_1^*, x_1 et x_2. Dans le cadre d'un modèle d'erreur de mesure classique, nous faisons l'hypothèse que e_1 n'est pas corrélée à x_1^* et x_2. Cela implique l'exogénéité de x_2 dans (15.46), mais la corrélation de x_1 et u. Nous avons besoin d'une VI pour x_1. Cette variable devra être corrélée à x_1, sans être corrélée à u (afin de pouvoir l'exclure de (15.45)), et elle ne devra pas être corrélée à l'erreur de mesure, e_1.

Une solution est d'obtenir une autre mesure de x_1^*, que l'on appelle z_1. Puisque c'est x_1^* qui affecte y, faire l'hypothèse que z_1 n'est pas corrélée à u va de soi. Si on écrit $z_1 = x_1^* + a_1$, où a_1 est l'erreur de mesure de z_1, alors il nous faut faire l'hypothèse que e_1 et a_1 ne sont pas corrélées. En d'autres termes, x_1 et z_1 mesurent x_1^* avec erreur, mais leurs erreurs de mesure ne sont pas corrélées. Évidemment, x_1 et z_1 sont corrélées car elles dépendent toutes les deux de x_1^*, donc on peut utiliser z_1 comme instrument pour x_1.

Comment peut-on obtenir deux mesures pour une seule variable ? Parfois, on demande à un groupe de salariés leur salaire annuel et leurs employeurs donnent une seconde mesure. On peut demander indépendamment à chaque conjoint, dans un couple marié, de déclarer le niveau d'épargne ou de revenu de la famille. Dans l'étude d'Ashenfelter et Krueger (1994) citée dans la section 14.3, on a demandé à chacun des jumeaux de renseigner le niveau d'étude de son frère ou de sa sœur, ce qui donne une seconde mesure que l'on peut utiliser comme instrument pour le niveau d'éducation renseigné par la personne elle-même dans une équation de salaire. (Ashenfelter et Krueger utilisent une différence première et la méthode des VI en même temps afin de prendre en compte également le problème de l'omission des capacités individuelles inobservées, ce qu'on verra dans la section 15.8.) Néanmoins, en général, il est rare d'avoir deux mesures d'une même variable explicative.

Une alternative serait d'utiliser d'autres variables exogènes comme instruments pour la variable qui est potentiellement mal mesurée. Par exemple, notre utilisation de *motheduc* (éducation de la mère) et de *fatheduc* (éducation du père) comme instruments pour *educ* (éducation de l'enfant) dans l'exemple 15.5 peut servir à cela. Si on pense que $educ = educ^* + e_1$, alors les estimations par VI de l'exemple 15.5 ne souffrent

pas d'erreur de mesure si *motheduc* et *fatheduc* ne sont pas corrélées à l'erreur de mesure, e_1. Cette hypothèse semble plus raisonnable que de penser que *motheduc* et *fatheduc* ne sont pas corrélées aux capacités inobservées qui sont contenues dans u dans l'équation (15.45).

On peut aussi adopter les méthodes par variables instrumentales quand on utilise des variables comme les résultats à un examen dans le but de tenir compte de l'influence de certaines caractéristiques inobservées. Dans la section 9.2, nous avons vu que, sous certaines hypothèses, des variables de substitution (ou *proxy*) peuvent être utilisées pour régler le problème de biais de variable omise. Dans l'exemple 9.3, on utilise le QI comme une variable de substitution pour les capacités inobservées. Cela conduit simplement à introduire la variable QI dans le modèle et l'estimer par MCO. Mais, il y a une alternative à cette méthode, qui fonctionne même si la variable QI ne vérifie pas complètement les conditions de variable de substitution. Pour illustrer cela, écrivons l'équation de salaire ainsi :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + abil + u \qquad \textbf{[15.47]}$$

Où, encore une fois, la variable *abil* est omise. Mais nous disposons de deux résultats à des examens qui sont des *indicateurs* des capacités inobservées. Nous supposons que les résultats peuvent s'écrire de la façon suivante :

$$test_1 = \gamma_1 abil + e_1$$

Et

$$test_2 = \delta_1 abil + e_2$$

Où $\gamma_1 > 0$ et $\delta_1 > 0$. Puisque ce sont les capacités inobservées qui affectent le salaire, nous pouvons faire l'hypothèse que $test_1$ et $test_2$ ne sont pas corrélées à u. En écrivant *abil* en fonction du résultat au premier examen et en introduisant cette écriture dans (15.47), on obtient :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \alpha_1 test_1 + (u - \alpha_1 e_1) \qquad \textbf{[15.48]}$$

Où $\alpha_1 = 1 / \gamma_1$. Maintenant, si on suppose que e_1 n'est corrélée à aucune des variables explicatives dans (15.47), y compris *abil*, alors e_1 est nécessairement corrélée à $test_1$. (Il faut remarquer que *educ n'est pas* endogène dans (15.48) ; toutefois, $test_1$ l'est.) Cela signifie que l'estimation de (15.48) nous donnera des estimateurs non convergents des β_j (ainsi que de α_1). Sous les hypothèses que nous avons faites, $test_1$ ne satisfait pas complètement les conditions pour être une variable de substitution.

En supposant que e_2 n'est corrélée à aucune des variables explicatives de (15.47) *et* que e_1 et e_2 ne sont pas corrélées, alors e_1 n'est pas corrélée au second résultat d'examen, $test_2$. Par conséquent, il est possible d'utiliser $test_2$ comme instrument pour $test_1$.

EXEMPLE 15.6
Utilisation de deux résultats d'examen comme indicateurs des capacités

Nous utilisons les données dans WAGE2 pour mettre en œuvre la procédure précédente, dans laquelle la variable *IQ* (QI) joue le rôle du premier résultat d'examen et la variable *KWW* (connaissance du monde du travail – *knowledge of the world of work*) celui du deuxième résultat d'examen. Les variables explicatives sont les mêmes que celles de l'exemple 9.3 : *educ* (éducation), *exper* (expérience), *tenure* (fonction), *married* (vaut 1 si la personne est mariée), *south* (vaut 1 pour la région sud), *urban* (vaut 1 si habite en zone urbaine) et *black* (vaut 1 si la personne est Noire). À la place d'ajouter la variable *IQ* et faire une estimation par MCO, comme nous l'avons fait pour la colonne (2) du tableau 9.2, nous ajoutons *IQ* et nous utilisons *KWW* comme instrument de *IQ*. Le coefficient pour *educ* est de 0,025 (écart-type = 0,017). Cette estimation est faible, et elle n'est pas statistiquement différente de zéro. Ce résultat est déconcertant, il suggère qu'une de nos hypothèses n'est pas vérifiée : peut-être que e_1 et e_2 sont en fait corrélées.

15.5 TEST D'ENDOGÉNÉITÉ ET TEST DE SURIDENTIFICATION

Dans cette section, nous décrivons deux tests importants pour l'estimation par variables instrumentales.

Test d'endogénéité

L'estimateur des DMC est moins efficace que l'estimateur par MCO quand les variables explicatives sont exogènes : comme nous l'avons vu, les écarts-types estimés des coefficients des DMC peuvent être très grands. Par conséquent, il est utile de disposer d'un test de l'endogénéité d'une variable explicative qui montre que les DMC sont vraiment nécessaires. Il est facile de mettre en place un tel test.

À titre d'illustration, supposons qu'il n'y ait qu'une seule variable que nous pensons être endogène dans le modèle suivant :

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1 \qquad \textbf{[15.49]}$$

dans lequel z_1 et z_2 sont exogènes. Nous disposons également de deux variables exogènes supplémentaires, z_3 et z_4, qui n'apparaissent pas dans (15.49). Si y_2 n'était pas corrélée à u_1, on pourrait estimer (15.49) par MCO. Comment peut-on tester la corrélation de y_2 et u_1 ? Hausman (1978) suggère de comparer directement les estimations par les MCO et les estimations par les DMC, afin de déterminer si elles sont statistiquement différentes. Après tout, les MCO comme les DMC sont convergents si toutes les variables sont exogènes. Si les estimations par DMC et par MCO sont significativement différentes, nous pouvons conclure que y_1 est certainement endogène (tout en maintenant les z_j exogènes).

C'est une bonne idée de calculer les estimations par MCO et par DMC pour voir si elles sont effectivement différentes. Afin de déterminer si les différences sont statistiquement significatives, le moyen le plus simple est d'utiliser un test de régression. Celui-ci est basé sur l'estimation de la forme réduite de y_2, qui est, dans ce cas :

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \qquad \textbf{[15.50]}$$

Comme aucun des z_j n'est corrélé à u_1, on voit dans cette équation que y_2 n'est pas corrélée à u_1 si et seulement si v_2 n'est pas corrélée à u_1 – ce que nous souhaitons vérifier. Nous pouvons écrire $u_1 = \delta_1 v_2 + e_1$, où e_1 n'est pas corrélée à v_2 et a une espérance nulle. Alors, u_1 et v_2 ne sont pas corrélées si et seulement si $\delta_1 = 0$. La façon la plus simple de tester cela est d'inclure v_2 comme régresseur supplémentaire dans (15.49) et faire un test de Student. Implémenter cette procédure pose un problème : v_2 n'est pas observée, puisque c'est le terme d'erreur de (15.50). On peut néanmoins obtenir les résidus de la forme réduite $\hat{v}_2$, puisqu'il est possible d'estimer la forme réduite de y_2 par MCO. Par conséquent, nous estimons par MCO :

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + erreur \qquad \textbf{[15.51]}$$

Pour terminer, nous testons $H_0 : \delta_1 = 0$ grâce à une statistique de Student. Si H_0 est rejetée à un niveau de significativité faible, on peut conclure que y_2 est endogène car v_2 et u_1 sont corrélées.

Pour tester l'endogénéité d'une seule variable explicative

i. Estimer la forme réduite de y_2 en la régressant sur *toutes* les variables exogènes (c'est-à-dire celles de l'équation structurelle ainsi que les variables instrumentales supplémentaires). Calculer les résidus $\hat{v}_2$.

ii. Ajouter $\hat{v}_2$ à l'équation structurelle (qui inclut y_2) et tester la significativité de $\hat{v}_2$ au moyen d'une régression MCO. Si le coefficient de $\hat{v}_2$ est statistiquement différent de zéro, en conclure que y_2 est effectivement endogène. On peut utiliser des tests de Student robustes à l'hétéroscédasticité.

EXEMPLE 15.7
Rendements de l'éducation des femmes en emploi

On peut tester l'endogénéité de *educ* dans (15.40) en calculant les résidus $\hat{v}_2$ issus de l'estimation de la forme réduite (15.41) (en utilisant seulement les femmes en emploi) et en les injectant dans (15.40). En suivant cette démarche, le coefficient de $\hat{v}_2$ est $\hat{\delta}_1 = 0{,}058$ avec $t = 1{,}67$. Ceci est une preuve assez fragile de l'existence d'une corrélation positive entre u_1 et v_2. C'est certainement une bonne idée d'indiquer les deux estimations parce que l'estimation des DMC des rendements de l'éducation (6,1 %) est nettement inférieure à l'estimation des MCO (10,8 %).

Il est intéressant de remarquer que dans la régression de l'étape (ii) pour le test d'endogénéité, les estimations des coefficients pour toutes les variables explicatives (sauf bien sûr, $\hat{v}_2$) sont identiques aux estimations DMC. Par exemple, estimer (15.51) par les MCO aboutit aux mêmes $\hat{\beta}_j$ qu'estimer (15.49) par les DMC. L'avantage de cette équivalence est qu'elle permet de vérifier facilement si on a fait la régression adéquate pour tester l'endogénéité. Elle propose également une interprétation différente et utile des DMC : l'ajout de la $\hat{v}_2$ à l'équation d'origine comme variable explicative, suivi de la mise en œuvre des MCO, permet de nettoyer de l'endogénéité de y_2. Donc, en commençant par estimer (15.49) par les MCO, il est possible de quantifier dans quelle mesure il est important de considérer que y_2 est endogène en regardant comment $\hat{\beta}_1$ varie quand $\hat{v}_2$ est ajouté à l'équation. Quel que soit le résultat des tests statistiques, on peut voir si la variation de $\hat{\beta}_1$ va dans la direction attendue et si cette variation a du sens dans la pratique.

Il est aussi possible de tester l'endogénéité de plusieurs variables explicatives. Pour chaque variable suspectée d'être endogène, nous obtenons les résidus de la forme réduite, comme dans la partie (i). Ensuite, nous testons la significativité jointe de ces résidus dans l'équation structurelle, en utilisant en test F. La significativité jointe nous informe qu'au moins l'une des variables explicatives suspectes est endogène. On teste autant de restrictions d'exclusion que le nombre de variables explicatives suspectées d'endogénéité.

Test de suridentification

Quand nous avons introduit l'estimateur par variable instrumentale simple dans la section 15.1, nous avons insisté sur le fait que l'instrument doit satisfaire deux conditions : il ne doit pas être corrélé au terme d'erreur (condition d'exogénéité) et il doit être corrélé à la variable explicative endogène (condition de pertinence). Nous avons également vu que, même pour des modèles avec d'autres variables explicatives, la deuxième condition peut être testée au moyen d'un test de Student (quand on dispose d'un seul instrument) ou d'un test de Fisher (quand on dispose de plusieurs instruments). Dans le cadre d'un estimateur des VI simple, nous avons remarqué que la condition d'exogénéité ne peut pas être testée. Cependant, si nous disposons de plus d'instruments que ce dont nous avons besoin, nous pouvons tester si certains d'entre eux sont corrélés au terme d'erreur structurel ou pas.

Pour donner un exemple précis, considérons une nouvelle fois l'équation (15.49) avec deux instruments pour y_2 : z_3 et z_4. Pour mémoire, z_1 et z_2 jouent le rôle de leurs propres instruments. Mais puisqu'on a deux instruments pour y_2, nous pouvons estimer (15.49) en utilisant, par exemple, seulement z_3 comme VI pour y_2. Nous appelons $\check{\beta}_1$ l'estimateur par VI de β_1 qui en découle. Ensuite, nous pouvons estimer (15.49) en utilisant seulement z_4 comme VI pour y_2 et nous appelons $\tilde{\beta}_1$ cet estimateur par VI. Si tous les z_j sont exogènes et si z_3 et z_4 sont toutes les deux partiellement corrélées à y_2, alors $\check{\beta}_1$ et $\tilde{\beta}_1$ sont deux estimateurs convergents pour β_1. Par conséquent, si notre choix des instruments a un sens, $\check{\beta}_1$ et $\tilde{\beta}_1$ ne devraient être différents qu'à cause des erreurs d'échantillonnage. Hausman (1978) propose de construire un test, pour tester si z_3 et z_4 sont tous deux exogènes, en s'appuyant sur la différence $\check{\beta}_1 - \tilde{\beta}_1$. Nous indiquerons un peu plus loin un moyen plus simple de construire un test valide, mais avant cela, nous devons comprendre comment interpréter le résultat du test.

Si nous arrivons à la conclusion que $\check{\beta}_1$ et $\tilde{\beta}_1$ sont statistiquement différents, alors nous devons en conclure que soit z_3, soit z_4, (soit les deux) ne satisfait pas la condition d'exogénéité. Malheureusement, nous ne sommes pas en mesure de savoir laquelle des deux variables ne satisfait pas cette condition (à moins de décréter immédiatement que z_3, par exemple, est exogène). Par exemple, si y_2 représente le nombre d'années d'études dans une équation de salaire, z_3 l'éducation de la mère et z_4 celle du père, une différence statistiquement significative entre les deux estimateurs des VI implique que l'une ou les deux variables représentant les niveaux d'études des parents est (sont) corrélée(s) à u_1 dans (15.49).

Évidemment, quand on rejette des instruments parce qu'ils ne sont pas exogènes, cela soulève un vrai problème et cela nécessite de changer d'approche d'estimation. Mais il y a un problème plus grave et plus subtil lorsqu'on compare des estimations par VI : elles peuvent paraître similaires même si aucune des deux variables instrumentales ne satisfait la condition d'exogénéité. Dans l'exemple précédent, il est probable que si le niveau d'étude de la mère est positivement corrélé à u_1, alors celui du père le sera aussi. Par conséquent, les deux estimations par VI peuvent paraître similaires même si aucune des deux n'est convergente. En effet, puisque les VI, dans cet exemple, sont choisis en utilisant des raisonnements similaires, les utiliser séparément dans des procédures VI peut tout à fait conduire à deux estimations similaires alors qu'aucune n'est convergente. Il ne faut donc pas se sentir trop rassuré par le fait la procédure VI réussit à passer le test d'Hausman.

Quand on compare deux estimations par VI, on a souvent affaire au problème qu'elles semblent, dans les faits, différentes et qu'on ne peut pourtant pas rejeter statistiquement l'hypothèse nulle selon laquelle elles sont convergentes vers le même paramètre de population. Par exemple, quand on estime (15.40) en utilisant *motheduc* (éducation de la mère) comme seule VI, on trouve un coefficient pour *educ* de 0,049 (0,037). Si on utilise *fatheduc* (éducation du père) comme seule VI, on trouve un coefficient de 0,070 (0,034). (Résultat peu surprenant, puisque l'estimation obtenue en utilisant les deux variables d'éducation des parents comme instruments est entre ces deux estimations, à savoir 0,061 (0,031).) Au moment d'avoir un discours en termes de politiques publiques, la différence entre 5 % et 7 % pour les rendements estimés de l'éducation est importante. Mais comme nous l'avons montré dans l'exemple 15.8, la différence n'est pas statistiquement significative.

La démarche qui consiste à comparer les estimations par VI du même paramètre est un exemple de test de **restrictions de suridentification**. L'idée générale est que nous disposons de plus d'instruments que ce dont nous avons besoin pour estimer les paramètres de manière convergente. Dans l'exemple précédent, nous avions un instrument de plus que nécessaire, ce qui nous donne une restriction de suridentification que l'on peut tester. Dans un cadre plus général, supposons que l'on dispose de q instruments de plus que ce dont on a besoin. Par exemple, avec une variable explicative endogène y_2, et trois instruments potentiels pour y_2, nous disposons de $q = 3 - 1 = 2$ restrictions de suridentification. Lorsque q est supérieur ou égal à 2, la comparaison de plusieurs estimations par les VI n'est pas aisée. Par contre, calculer une statistique de test à partir des résidus des DMC est aisé. L'idée est la suivante : si tous les instruments étaient exogènes, les résidus des DMC ne devraient pas être corrélés aux instruments ou seulement à cause des erreurs d'échantillonnage. S'il y a $k + 1$ paramètres et $k + 1 + q$ instruments, les résidus des DMC ont une moyenne de zéro et, de la même manière, ils ne sont pas corrélés avec k combinaisons linéaires des instruments. (Ce résultat mathématique implique, comme cas particulier, le fait que les résidus des MCO ont une moyenne de zéro et qu'ils ne sont pas corrélés aux k variables explicatives.) Par conséquent, le test vérifie si les résidus des DMC sont corrélés à q fonctions linéaires des instruments. Il n'est pas nécessaire de décider soi-même des fonctions en question, le test fait cela automatiquement.

Le test suivant, issu des régressions, est valide si l'hypothèse d'homoscédasticité présentée dans l'annexe du chapitre (hypothèse DMC.5) est vérifiée.

Test des restrictions de suridentification

i. Estimer l'équation structurelle par DMC et calculer les résidus DMC, $\hat{u}_1$

ii. Régresser $\hat{u}_1$ sur toutes les variables exogènes. Calculer le R-carré, que l'on appelle R_1^2

iii. Sous l'hypothèse nulle selon laquelle aucune des variables instrumentales n'est corrélée à u_1, $nR_1^2 \sim \chi_q^2$, où q est la différence entre le nombre de variables instrumentales qui n'apparaissent pas dans le modèle et le nombre total de variables explicatives endogènes. Si nR_1^2 est supérieur à la valeur critique au seuil de 5 % (par exemple) d'une distribution χ_q^2, rejeter H_0 et conclure qu'au moins quelques-unes des VI ne sont pas exogènes.

EXEMPLE 15.8
Rendements de l'éducation pour les femmes en emploi

En utilisant *motheduc* et *fatheduc* comme variables instrumentales pour *educ* dans (15.40), nous disposons d'une seule restriction de suridentification. En régressant les résidus des DMC, $\hat{u}_1$, sur *exper*, *exper*^{2}, *motheduc* et *fatheduc*, nous obtenons $R_1^2 = 0{,}0009$. Par conséquent, $nR_1^2 = 428 \times 0{,}0009 = 0{,}3852$, ce qui représente une très petite valeur pour une distribution χ_1^2 (p-value = 0,535). Par conséquent, les variables pour les niveaux d'études des parents ne sont pas rejetées par le test de suridentification. En ajoutant aussi le niveau d'étude du conjoint *(huseduc)*, nous obtenons deux restrictions de suridentification et nous avons $nR_1^2 = 1{,}11$ (p-value = 0,574). Sous réserve des conditions présentées précédemment, ajouter *huseduc* (niveau d'études du conjoint) à la liste des variables instrumentales semble raisonnable, puisque cela réduit l'écart-type de l'estimation obtenue par DMC : l'estimation par DMC du coefficient pour *educ* en utilisant les trois instruments est de 0,080 (écart-type = 0,022), ce qui rend *educ* beaucoup plus significatif que lorsque *huseduc* n'est pas utilisé comme VI ($\hat{\beta}_{educ} = 0{,}061$, écart-type = 0,031).

Lorsque $q = 1$, une question vient naturellement à l'esprit : qu'apporte le test basé sur la régression en plus du test basé sur la comparaison directe des estimations ? En fait, les deux procédures sont les mêmes asymptotiquement. Pour des raisons pratiques, il est très sensé de calculer les deux estimations par VI et de voir dans quelle mesure elles sont différentes. En règle générale, quand $q \geq 2$, on peut comparer les estimations par DMC en utilisant l'ensemble des instruments, aux estimations obtenues en utilisant chaque instrument un par un. En faisant cela, on peut voir si les différentes estimations par les VI sont proches les unes des autres, que l'hypothèse nulle soit rejetée par le test de suridentification ou pas.

Dans l'exemple précédent, nous avons fait allusion à une caractéristique générale des DMC : sous leurs hypothèses standards, ajouter des instruments à la liste améliore l'efficacité asymptotique des DMC. Il est toutefois nécessaire pour cela que chaque nouvel instrument soit réellement exogène (dans le cas contraire, les DMC ne sont même pas convergents) et ce résultat est seulement un résultat asymptotique. Avec des échantillons de taille classique, ajouter trop d'instruments (c'est-à-dire augmenter le nombre de restrictions de suridentification) peut conduire l'estimateur des DMC à être gravement biaisé. Une discussion détaillée nous conduirait trop loin. Une bonne illustration est donnée par Bound, Jaeger et Baker (1995) qui soutiennent que les estimations par DMC des rendements de l'éducation obtenus par Angrist et Krueger (1991), qui utilisent un grand nombre de variables instrumentales, sont potentiellement très biaisées (même en disposant de centaines de milliers d'observations !).

Le test de suridentification peut être utilisé dès que nous disposons de plus d'instruments que ce dont nous avons besoin. Si nous disposons tout juste d'assez d'instruments, le modèle est dit *juste identifié* et le R-carré de la partie (ii) est égal à 0. Comme nous l'avons mentionné plus haut, nous ne pouvons pas tester l'exogénéité des instruments dans un modèle juste identifié.

Le test peut être rendu robuste à n'importe quelle forme d'hétéroscédasticité ; voir Wooldridge (2010, chapitre 5) pour plus de détails.

15.6 DOUBLES MOINDRES CARRÉS ET HÉTÉROSCÉDASTICITÉ

L'hétéroscédasticité, dans le cadre des DMC, soulève les mêmes questions que dans le cadre des MCO. Ce qu'il faut savoir est qu'il est possible d'obtenir des écarts-types estimés des coefficients et des statistiques de test qui sont asymptotiquement robustes à l'hétéroscédasticité de n'importe quelle forme inconnue. En fait, l'expression [8.4] reste valide si les $\hat{r}_{ij}$ sont les résidus de la régression des $\hat{x}_{ij}$ sur les autres $\hat{x}_{ih}$, dans lesquelles « ^ » signifie qu'il s'agit des valeurs prédites des régressions de première étape (pour les variables explicatives endogènes). Wooldridge (2010, chapitre 5) apporte une explication plus détaillée. Certains logiciels d'économétrie font cela systématiquement.

Il est également possible de tester l'hétéroscédasticité, en utilisant un test analogue au test de Breusch-Pagan que nous avons vu dans le chapitre 8. On appelle $\hat{u}$ les résidus obtenus par DMC et on note $z_1, z_2, \ldots, z_m$ l'ensemble des variables exogènes (incluant celles qu'on utilise comme VI pour les variables explicatives endogènes). Ainsi, sous certaines hypothèses raisonnables (présentées par exemple dans Wooldridge, 2010, chapitre 5), une statistique asymptotique valide est la statistique F usuelle utilisée pour tester la significativité jointe des régresseurs dans une régression de $\hat{u}^2$ sur $z_1, z_2, \ldots, z_m$. L'hypothèse nulle d'homoscédasticité est rejetée si les z_j sont significatifs de manière jointe.

En mettant en œuvre ce test dans le cadre de l'exemple 15.8, en utilisant *motheduc*, *fatheduc* et *huseduc* comme instruments pour *educ*, on obtient $F_{5,422} = 2{,}53$ et la p-value est de 0,029. Ce résultat nous indique la présence d'hétéroscédasticité au seuil de 5 %. Il semblerait donc logique de construire des écarts-types estimés robustes à l'hétéroscédasticité afin de la prendre en compte.

Si on sait de quelle façon la variance des erreurs dépend des variables exogènes, nous pouvons utiliser une procédure de triple moindres carrés, fondamentalement la même que dans la section 8.4. Après avoir estimé un modèle estimant $Var(u|z_1, z_2, \ldots, z_m)$, il faut diviser la variable dépendante, les variables explicatives et toutes les variables instrumentales de l'observation i par $\sqrt{\hat{h}_i}$, $\hat{h}_i$ étant la variance estimée. (La constante, qui est à la fois une variable explicative et une VI, est divisée par $\sqrt{\hat{h}_i}$, voir section 8.4.) Ensuite, il faut utiliser les DMC sur l'équation transformée en utilisant les instruments transformés.

15.7 APPLICATION DES DMC SUR DES ÉQUATIONS DE SÉRIES TEMPORELLES

Quand on utilise les DMC sur des données de séries temporelles, on retrouve la plupart des problèmes soulevés dans les chapitres 10, 11 et 12 lors de l'application des MCO dans ce cadre. Écrivons l'équation structurelle pour chaque période temporelle de la façon suivante :

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t \qquad \textbf{[15.52]}$$

dans laquelle une ou plusieurs des variables explicatives x_{tj} peut être corrélée à u_t. On appelle $z_{t1}, \ldots, z_{tm}$, l'ensemble des variables exogènes telles que :

$$E(u_t) = 0, Cov(z_{tj}, u_t) = 0, \; j = 1, \ldots, m$$

Toute variable explicative exogène est aussi l'une des z_{tj}. Pour l'identification, il est nécessaire d'avoir $m \geq k$ (il nous faut au moins autant de variables exogènes qu'il y a de variables explicatives).

La façon de fonctionner des DMC est la même pour les séries temporelles que pour les données en coupe instantanée, mais pour les séries temporelles, les propriétés statistiques des DMC dépendent de la tendance temporelle et des propriétés de corrélation des suites sous-jacentes. En particulier, il faut prendre

soin d'inclure des tendances temporelles si on a une variable dépendante ou des variables explicatives qui ont une tendance temporelle. Une tendance temporelle est exogène, elle peut donc toujours être utilisée comme son propre instrument. Il en va de même pour les variables binaires prenant en compte la saisonnalité si on utilise des données mensuelles ou trimestrielles.

Les données qui présentent une forte persistance (qui ont une racine unitaire) doivent être utilisées avec précaution, comme pour les MCO. Souvent, il est requis de différencier l'équation avant l'estimation, ainsi que l'instrument.

Pour aller plus loin 15.4

Afin de tester l'effet d'une croissance des dépenses du gouvernement sur la croissance de la production, on peut écrire le modèle :

$$cPIB_t = \beta_0 + \beta_1 cGOV_t + \beta_2 INV_t + \beta_3 ACT_t + u_t$$

où c indique la croissance, *PIB* est la production intérieur brute réelle, *GOV* représente les dépenses réelles du gouvernement, *INV* est le rapport des investissements intérieurs bruts sur le *PIB* et *ACT* est la taille de la population active. (Voir équation (6) de Ram, 1986) Sous quelles hypothèses une variable binaire indiquant si le président en $t-1$ est Républicain est-elle une VI adaptée pour $cGOV_t$?

Sous des hypothèses semblables à celles du chapitre 11 pour les propriétés asymptotiques des MCO, les DMC sur données de séries temporelles sont convergents et suivent asymptotiquement une loi normale. En fait, il suffit d'écrire les hypothèses en question en remplaçant les variables explicatives par les variables instrumentales puis d'ajouter les hypothèses identificatrices des DMC. Par exemple, l'hypothèse d'homoscédasticité sera énoncée de la façon suivante :

$$E(u_t^2 | z_{t1}, \ldots, z_{tm}) = \sigma^2 \qquad \textbf{[15.53]}$$

et l'hypothèse d'absence de corrélation sérielle s'écrit :

$$E(u_t u_s | z_t, z_s) = 0, \text{ pour tout } t \neq s \qquad \textbf{[15.54]}$$

où z_t représente l'ensemble des variables exogènes à l'instant t. Une présentation complète des hypothèses est donnée dans l'annexe du chapitre. Nous proposerons des exemples d'exercices sur les DMC appliqués sur séries temporelles dans le chapitre 16, voir également l'exercice sur ordinateur C4.

Dans le cadre des MCO, l'hypothèse d'absence de corrélation sérielle est rarement vérifiée avec des données de séries temporelles. Par chance, il est très facile de tester la présence de corrélation sérielle de la forme AR(1). Si on écrit $u_t = \rho u_{t-1} + e_t$ et qu'on remplace cette écriture dans l'équation (15.52), on obtient :

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + \rho u_{t-1} + e_t, t \geq 2 \qquad \textbf{[15.55]}$$

Pour tester $H_0 : \rho_1 = 0$, il faut remplacer u_{t-1} par les résidus obtenus par DMC, $\hat{u}_{t-1}$. De plus, si x_{tj} est endogène dans (15.52), alors elle est aussi endogène dans (15.55), donc il nous faut encore utiliser une VI. Puisque e_t n'est corrélée à aucune des valeurs passées de u_t, $\hat{u}_{t-1}$ peut être utilisée comme son propre instrument.

Tester la corrélation sérielle AR(1) après les DMC

i. Estimer (15.52) par DMC et récupérer les résidus DMC, $\hat{u}_t$

ii. Estimer

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + \rho \hat{u}_{t-1} + error_t, t = 2, \ldots, n$$

par DMC, en utilisant les mêmes instruments que ceux utilisés en (i), en plus de $\hat{u}_{t-1}$. Utiliser la statistique de Student sur $\hat{\rho}$ pour tester $H_0 : \rho = 0$.

De la même façon que pour la version MCO de ce test, vue dans le chapitre 12, la statistique t a seulement une justification asymptotique, mais, en pratique, elle fonctionne bien. Une version robuste à l'hétéroscédasticité peut être utilisée pour se prémunir de l'hétéroscédasticité. En outre, il est possible d'utiliser plus de variables retardées pour les résidus afin de tester une corrélation sérielle de plus haut degré en utilisant un test de Fisher de significativité jointe.

Que doit-on faire si on détecte de la corrélation sérielle ? Certains logiciels informatiques disposent de programmes qui calculent des écarts-types estimés des coefficients robustes à des formes assez générales de corrélations sérielles et d'hétéroscédasticité. Ces calculs sont très similaires à ceux qu'on a vus dans la section 12.5 dans le cadre des MCO. (Voir Wooldridge (1995) pour des formules et d'autres méthodes de calcul.)

Une méthode alternative est d'utiliser un modèle AR(1) et de prendre en compte la corrélation sérielle. La procédure est similaire à celle des MCO et ajoute des restrictions aux variables instrumentales. L'équation quasi-différenciée est la même qu'en [12.32] :

$$\tilde{y}_t = \beta_0(1-\rho) + \beta_1\tilde{x}_{t1} + \ldots + \beta_k\tilde{x}_{tk} + e_t, \; t \geq 2 \qquad \textbf{[15.56]}$$

où $\tilde{x}_{tj} = x_{tj} - \rho x_{t-1,j}$. (On peut utiliser l'observation $t = 1$ de la même façon que dans la section 12.3, mais nous omettons ce cas ici pour des raisons de simplicité.) La question est alors : que peut-on utiliser comme variable instrumentale ? Il paraît naturel d'utiliser les instruments quasi-différenciés, $\tilde{z}_{tj} = z_{tj} - \rho z_{t-1,j}$. Cependant, cela fonctionne seulement si les erreurs originales u_t, dans (15.52), ne sont pas corrélées aux instruments aux dates t, $t - 1$ et $t + 1$. Ce qui revient à dire que les variables instrumentales doivent être strictement exogènes dans (15.52), ce qui, par exemple, exclut la possibilité d'utiliser des retards de la variable dépendante. Cela exclut également les cas dans lesquels les mouvements futurs des variables instrumentales réagissent aux variations courantes et passées du terme d'erreur u_t.

DMC avec des erreurs AR(1)

i. Estimer (15.52) par DMC et récupérer les résidus des DMC, $\hat{u}_t$, $t = 1, 2, \ldots, n$

ii. Récupérer $\hat{\rho}$ de la régression de $\hat{u}_t$ sur $\hat{u}_{t-1}$, $t = 2, \ldots,$ n, et construire les variables quasi-différenciée $\tilde{y}_t = y_t - \hat{\rho}y_{t-1}$, $\tilde{x}_{tj} = x_{tj} - \hat{\rho}x_{t-1,j}$ et $\tilde{z}_{tj} = z_{tj} - \hat{\rho}z_{t-1,j}$, pour $t \geq 2$. (Souvenez-vous, dans la plupart des cas, certaines des variables instrumentales sont aussi des variables explicatives.)

iii. Estimer (15.56) (dans laquelle on remplace ρ par $\hat{\rho}$) par DMC, en utilisant les $\tilde{z}_{tj}$ comme instruments. En faisant l'hypothèse que (15.56) satisfait les hypothèses des DMC de l'annexe du chapitre, les statistiques de test des DMC usuelles sont valides asymptotiquement.

Nous pouvons également utiliser l'observation de la première date, comme dans l'estimation du modèle avec des variables explicatives exogènes proposée par Prais-Winsten. Les variables transformées de la première date (la variable dépendante, les variables explicatives et les variables instrumentales) sont obtenues simplement en multipliant toutes les valeurs de première date par $(1 - \hat{\rho})^{1/2}$. (Voir aussi section 12.3.)

15.8 L'APPLICATION DES DMC AUX DONNÉES DE COUPES AGRÉGÉES ET AUX DONNÉES DE PANEL

Utiliser les méthodes par variables instrumentales sur des données de coupes indépendantes agrégées ne soulève pas de nouvelle difficulté. Comme pour les modèles estimés par MCO, il est souvent nécessaire d'introduire des variables binaires temporelles pour prendre en compte des effets temporels agrégés. Ces

variables binaires sont exogènes (parce que le temps qui passe est exogène), elles peuvent donc jouer le rôle de leurs propres instruments.

EXEMPLE 15.9
Effet de l'éducation sur la fécondité

Dans l'exemple 13.1, nous avons utilisé les données de coupes transversales agrégées de FERTIL1 pour estimer les effets de l'éducation sur la fécondité des femmes, en tenant compte de l'influence de divers autres facteurs. Comme dans Sanders (1992), nous laissons à *educ* la possibilité d'être endogène dans l'équation. Les variables instrumentales d'*educ* sont les niveaux d'études de la mère (*motheduc*) et du père (*fatheduc*). L'estimation DMC de β_{educ} est de $-0{,}153$ (écart-type = 0,039), qu'il faut comparer à l'estimation MCO de $-0{,}128$ (écart-type = 0,018). L'estimation par DMC indique que l'effet de l'éducation sur la fécondité est plus important que celui mesuré précédemment, mais les écarts-types estimés des coefficients estimés par les DMC sont plus de deux fois plus grands que les écarts-types estimés par les MCO. (En fait, l'intervalle de confiance à 95 % calculé des estimations DMC contient l'estimation MCO.) Les estimations des MCO et des DMC de β_{educ} ne sont pas statistiquement différentes, ce que l'on peut voir en testant l'endogénéité de *educ* en utilisant la méthode vue en 15.5 : si on inclut les résidus de la forme réduite $\hat{v}_2$, avec les autres régresseurs du tableau 13.1 (en incluant également *educ*), la statistique t de $\hat{v}_2$ est de 0,702, ce qui n'est significatif pour aucun niveau de significativité raisonnable. Par conséquent, dans ce cas, nous pouvons conclure que la différence entre les DMC et les MCO peut être entièrement attribuée à des erreurs d'échantillonnage.

L'estimation par variable instrumentale peut être combinée aux méthodes de données de panel, notamment la différence première, pour estimer les paramètres de manière convergente en présence d'effets inobservés et de l'endogénéité d'une ou plusieurs variables explicatives. L'exemple simple qui suit illustre cette combinaison de méthodes.

EXEMPLE 15.10
Formation professionnelle et productivité du salarié

Imaginons que l'on veuille estimer l'effet d'une heure supplémentaire de formation professionnelle sur la productivité. Nous considérons le modèle simple suivant, sur données de panel, pour les années 1987 et 1988 :

$$\log(scrap_{it}) = \beta_0 + \delta_0 d88_t + \beta_1 hrsemp_{it} + a_i + u_{it}, t = 1, 2$$

Dans lequel $scrap_{it}$ est le taux de rebut de l'entreprise i au cours de l'année t et $hrsemp_{it}$ est le nombre d'heures de formation professionnelle par employé. Comme d'habitude, nous admettons des effets fixes d'années et un effet fixe inobservé entreprise, a_i.

Pour des raisons évoquées au cours de la section 13.2, nous devrions nous préoccuper de la corrélation potentielle de $hrsemp_{it}$ à a_i, qui contient les capacités inobservées des salariés. Comme nous l'avons vu plus haut, nous nous débarrassons de a_i en utilisant une différence première :

$$\Delta\log(scrap_i) = \delta_0 + \beta_1 \Delta hresmp_i + \Delta u_i \qquad \textbf{[15.57]}$$

Normalement, nous pourrions estimer cette équation par MCO. Mais que se passe-t-il si Δu_i est corrélée à $\Delta hrsemp_i$? Par exemple, une entreprise peut se mettre à embaucher des salariés plus qualifiés et réduire le niveau de formation professionnelle en même temps. Dans ce cas, nous avons besoin d'une variable instrumentale pour $\Delta hrsemp_i$. Trouver une VI de ce type est souvent difficile, mais on peut utiliser le fait que certaines entreprises ont reçu des subventions à la formation en 1988. Si on fait l'hypothèse que l'attribution de subventions n'est pas corrélée à Δu_i (ce qui n'est pas incongru, car les subventions ont été attribuées en début d'année), alors

$\Delta grant_i$ est un instrument valide, à condition que $\Delta hrsemp$ et $\Delta grant$ soient corrélées. En utilisant les données de JTRAIN, sur lesquelles on calcule la différence entre 1987 et 1988, la régression de première étape nous donne :

$$\Delta hrsemp = 0{,}51 + 27{,}88\ \Delta grant$$
$$(1{,}56)\quad (3{,}13)$$
$$n = 45,\ R^2 = 0{,}392.$$

Celle-ci confirme que la variation du nombre d'heures de formation professionnelle est fortement et positivement corrélée au fait d'avoir reçu une subvention à la formation professionnelle en 1988. En fait, l'attribution d'une subvention à la formation professionnelle a augmenté la formation par employé de presque 28 heures, et l'octroi de la subvention explique près de 40 % de la variation de $\Delta hrsemp$. L'estimation par doubles moindres carrés de (15.57) nous donne :

$$\Delta\log(scrap) = -0{,}033 - 0{,}014\ \Delta hrsemp$$
$$(0{,}127)\quad (0{,}008)$$
$$n = 45,\ R^2 = 0{,}016.$$

Celle-ci signifie que l'on estime qu'une augmentation de 10 heures de formation professionnelle par salarié réduit le taux de rebut de 14 %. Le nombre moyen d'heures de formation des entreprises de l'échantillon s'élève à environ 17 heures par salarié, avec un minimum de zéro et un maximum de 88.

À titre de comparaison, l'estimation de (15.57) par MCO nous donne $\hat{\beta}_1 = 0{,}0076$ (écart-type = 0,0045), donc l'ampleur du coefficient β_1 estimé par DMC est presque deux fois plus grande que lorsqu'il est estimé par MCO et statistiquement, il est un petit peu plus significatif.

Lorsque $T \geq 3$, l'équation différenciée peut présenter une corrélation sérielle. Le même test et la même correction pour la corrélation sérielle de type AR(1) présenté en section 15.7 peuvent être utilisés, en agrégeant toutes les régressions en i et en t. La transformation de Prais-Winsten pour la première date pourrait être utilisée afin de ne pas perdre l'ensemble de l'information d'une période.

Il est également nécessaire de faire appel aux méthodes par variables instrumentales pour estimer les modèles à effets fixes inobservés qui contiennent des retards de la variable dépendante. Cela provient du fait que, une fois l'équation différenciée, $\Delta y_{i,t-1}$ est corrélée à Δu_{it} parce que $y_{i,t-1}$ et $u_{i,t-1}$ sont corrélées. Il est possible d'utiliser deux retards ou plus de y comme variables instrumentales pour $\Delta y_{i,t-1}$. (Voir Wooldridge, 2010, Chapitre 11, pour plus de détails.)

Les variables instrumentales différenciées peuvent également être utilisées sur des échantillons pour lesquels on a apparié les observations. Ashenfelter et Krueger (1994) différencie l'équation de salaire entre jumeaux afin d'éliminer les capacités inobservées :

$$\log(wage_2) - \log(wage_1) = \delta_0 + \beta_1(educ_{2,2} - educ_{1,1}) + (u_2 - u_1)$$

Où $wage_1$ représente le salaire du premier jumeau et $wage_2$ celui du second jumeau, $educ_{1,1}$ (resp. $educ_{2,2}$) est le nombre d'années d'étude du premier jumeau tel qu'il a lui-même reporté (resp. second jumeau). Afin de prendre en compte des erreurs de mesure potentielles dans les nombres d'années d'études auto-déclarés, Ashenfelter et Krueger ont utilisé $(educ_{2,1} - educ_{1,2})$ comme instrument pour $(educ_{2,2} - educ_{1,1})$, dans laquelle $educ_{2,1}$ (resp. $educ_{1,2}$) est le nombre d'années d'étude du second jumeau tel que l'a déclaré le premier jumeau (resp. du premier jumeau tel que l'a déclaré le second). L'estimation de β_1 par les VI est de 0,167 ($t = 3{,}88$), que l'on peut comparer à l'estimation par les MCO sur les différences premières qui nous donne 0,092 ($t = 3{,}83$). (Voir Ashenfelter et Krueger, 1994, tableau 3).

RÉSUMÉ

Dans le chapitre 15, nous avons introduit la méthode des variables instrumentales comme méthode permettant d'estimer les paramètres d'un modèle linéaire quand une ou plusieurs variables explicatives sont endogènes. Une variable instrumentale doit avoir deux propriétés : (1) elle doit être exogène, à savoir qu'elle ne doit pas être corrélée au terme d'erreur de l'équation structurelle, (2) elle doit être partiellement corrélée à la variable explicative endogène. Il est généralement difficile de trouver une variable qui présente ces deux propriétés.

La méthode des doubles moindres carrés permet d'utiliser plus de variables instrumentales que l'on a de variables explicatives. Elle est couramment utilisée dans les études empiriques en sciences sociales. Si elle est correctement mise en œuvre, elle permet d'estimer des effets *ceteris paribus* en présence de variables explicatives endogènes. Elle peut être appliquée sur des données de coupe transversale, de série temporelle ou des données longitudinales. Cependant, si les instruments sont faibles, c'est-à-dire s'ils sont corrélés au terme d'erreur ou seulement faiblement corrélés à la variable explicative endogène, ou les deux, alors les DMC sont pires que les MCO.

Si on dispose de variables instrumentales valides, alors on peut tester si la variable explicative est endogène, en utilisant le test présenté dans la section 15.5. De plus, même s'il est impossible de tester si toutes les variables instrumentales sont exogènes, on peut au moins tester si certaines d'entre elles le sont, à condition d'avoir plus d'instruments qu'il ne nous faut pour obtenir une estimation convergente, c'est-à-dire, si le modèle est suridentifié.

Dans ce chapitre, nous avons utilisé les problèmes de variables omises et d'erreurs de mesure pour illustrer la méthode des variables instrumentales. Les méthodes par variables instrumentales sont également indispensables pour les modèles d'équations simultanées, ce que nous verrons dans le chapitre 16.

MOTS-CLÉS

Condition d'ordre p. 623
Condition de rang p. 623
Équation en forme réduite p. 615
Équation structurelle p. 614
Erreur de mesure sur les régresseurs p. 602
Estimateur par doubles moindres carrés p. 620
Estimateur par variables instrumentales p. 602
Exogénéité d'un instrument p. 604
Expérience naturelle p. 611
Identification d'un paramètre p. 606
Instruments faibles p. 612
Pertinence d'un instrument p. 604
Restrictions d'exclusion p. 618
Restrictions de suridentification p. 628
Variable exogène p. 614
Variable explicative endogène p. 602
Variable explicative exogène p. 615
Variable instrumentale p. 604
Variables omises p. 602

EXERCICES

1. Nous examinons un modèle simple pour estimer l'effet de posséder un ordinateur (*PC*) sur la moyenne des notes d'un étudiant en fin de licence (*MOY*) dans une grande université publique.

$$MOY = \beta_0 + \beta_1 PC + u$$

PC est une variable binaire indiquant si l'étudiant possède un ordinateur personnel.

i. Pour quelles raisons peut-on penser que le fait de posséder un ordinateur est potentiellement corrélé à u ?

ii. Expliquez pour quelles raisons la variable *PC* pourrait être associée aux revenus annuels des parents. Est-ce que cela signifie que le revenu des parents constitue une bonne variable instrumentale pour *PC* ? Pourquoi ?

iii. Imaginez que, trois années auparavant, l'université a offert des subventions pour acheter des ordinateurs à la moitié des étudiants entrant à l'université environ. Les étudiants qui recevaient une subvention avaient été désignés aléatoirement. Expliquez précisément comment vous utiliseriez cette information pour construire une variable instrumentale pour *PC*.

2. Imaginez que vous vouliez estimer l'effet de l'assiduité sur les performances scolaires des étudiants, comme dans l'exemple 6.3. On considère le modèle de base suivant :

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + u$$

pour lequel les variables ont été définies dans le chapitre 6.

i. On appelle *dist* la distance entre la résidence étudiante où vivent les étudiants et l'amphithéâtre où le cours a lieu. Pensez-vous que *dist* et u sont corrélées ?

ii. Nous faisons l'hypothèse que *dist* et u ne sont pas corrélées. Quelles sont les autres hypothèses que *dist* doit satisfaire pour être un instrument valide pour *atndrte* ?

iii. Nous ajoutons, dans l'équation [6.18], un terme d'interaction $priGPA \times atndrte$:

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA \times atndrte + u$$

Si *atndrte* est corrélée à u alors, en règle générale, $priGPA \times atndrte$ le sera aussi. Quelle variable pourrait constituer un bon instrument pour $priGPA \times atndrte$? (Indice : si $E(u|priGPA, ACT, dist) = 0$, ce qui est le cas quand *priGPA*, *ACT* et *dist* sont toutes les trois exogènes, alors toute fonction de *priGPA* et *dist* n'est pas corrélée à u.)

3. Nous nous intéressons au modèle de régression simple

$$y = \beta_0 + \beta_1 x + u$$

Et on appelle z une variable binaire, instrument pour x. Utiliser (15.10) pour montrer que l'estimateur des VI $\hat{\beta}_1$ peut s'écrire de la façon suivante :

$$\hat{\beta}_1 = (\bar{y}_1 - \bar{y}_0) / (\bar{x}_1 - \bar{x}_0)$$

où $\bar{y}_0$ et $\bar{x}_0$ sont respectivement les moyennes empiriques des y_i et des x_i sur le sous-échantillon pour lequel $z_i = 0$, et où $\bar{y}_1$ et $\bar{x}_1$ sont respectivement les moyennes empiriques des y_i et des x_i sur le sous-échantillon pour lequel $z_i = 1$. Cet estimateur, connu sous le nom d'*estimateur par groupe (grouping estimator)*, a été proposé pour la première fois par Wald (1940).

4. Imaginez que vous vouliez, pour un État donné des États-Unis, utiliser des données de série temporelle annuelle pour estimer l'effet du salaire minimum au niveau de l'État sur le taux d'emploi des jeunes de 18 à 25 ans (*EMP*). On considère le modèle suivant :

$$gEMP_t = \beta_0 + \beta_1 gMIN_t + \beta_2 gPOP_t + \beta_3 gGSP_t + \beta_4 gGDP_t + u_t$$

Dans lequel MIN_t représente le salaire minimum, exprimé en dollar réels, POP_t la population de 18 à 25 ans, GSP_t la production brute dans l'État et GDP_t est le produit national brut des États-Unis. Le préfixe g indique que l'on s'intéresse au taux de croissance entre $t-1$ et t, pour lequel on calcule généralement une approximation grâce à la différence des logs.

i. Si on pense que l'État choisit le salaire minimum en se basant en partie sur des facteurs inobservés (de l'économètre) qui affectent également l'emploi des jeunes, quel problème présente l'estimation par MCO ?

ii. On appelle $USMIN_t$ le salaire minimum des États-Unis, également mesuré en termes réels. Pensez-vous que $gUSMIN_t$ soit corrélé à u_t ?

iii. La loi américaine impose que le salaire minimum de chaque État soit au moins aussi important que le salaire minimum des États-Unis. Expliquez pourquoi cela fait de $gUSMIN_t$ un candidat potentiel comme VI pour $gMIN_t$.

5. En se reportant aux équations (15.19) et (15.20), posons $\sigma_u = \sigma_x$, de sorte que la variation du terme d'erreur dans la population est la même que celle de x. Supposons que la variable instrumentale, z, est faiblement corrélée à u : $Corr(z,u) = 0{,}1$. Supposons également que z et x sont corrélées de façon un peu plus forte : $Corr(z,x) = 0{,}2$

i. Quel est le biais asymptotique de l'estimateur des VI ?

ii. De combien doit être la corrélation entre x et u pour que l'estimateur des MCO présente un biais plus grand que celui de l'estimateur des VI ?

6. i. Considérons le modèle avec une variable explicative endogène, une variable explicative exogène et une variable supplémentaire exogène : écrivez la forme réduite de y_2 comme en (15.26) et remplacez-la dans l'équation structurelle (15.22). Cela nous donne la forme réduite de y_1 :

$$y_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + v_1$$

Écrivez α_j en termes de β_j et de π_j.

ii. Écrivez le terme d'erreur de la forme réduite v_1, en fonction de u_1, v_2 et des paramètres.

iii. Comment feriez-vous pour estimer de façon convergente les α_j ?

7. Le modèle simple suivant a pour objectif de mesurer l'effet d'un programme pour encourager les parents à choisir l'école de leurs enfants (parmi les écoles faisant partie du réseau « *choice school* ») sur la performance à examen standardisé (voir Rouse (1998) pour la motivation et l'exercice sur ordinateur C11 pour l'analyse d'un sous-ensemble des données de Rouse) :

$$score = \beta_0 + \beta_1 choice + \beta_2 faminc + u_1$$

où *score* représente le résultat à un examen général au niveau de l'État, *choice* est une variable binaire indiquant si l'élève a été à une école du réseau « *choice school* » au cours de l'année précédente, et *faminc* représente le revenu de la famille. La variable instrumentale pour *choice* est *grant*, le montant de la bourse, en dollar, attribuée aux élèves pour payer les frais d'inscription aux écoles du réseau « *choice school* ». Nous tenons compte de l'influence des revenus de la famille car le montant attribué diffère en fonction de ceux-ci.

i. Même en intégrant *faminc* dans l'équation, pourquoi la variable *choice* risque-t-elle d'être corrélée à u_1 ?

ii. Si, au sein de chaque tranche de revenu, le montant de la bourse est assigné de façon aléatoire, la variable *grant* est-elle corrélée à u_1 ?

iii. Écrivez l'équation de forme réduite pour la variable *choice*. Que nécessite *grant* pour être partiellement corrélée à *choice* ?

iv. Écrivez l'équation de forme réduite pour *score*. Expliquez en quoi c'est utile. (Indice : comment interprétez-vous le coefficient de *grant* ?)

8. Votre objectif est de tester si les filles qui ont été scolarisées dans un lycée de filles ont de meilleurs résultats en mathématiques que celles qui ont été scolarisées dans un lycée mixte. Vous disposez d'un échantillon aléatoire de filles qui ont achevé le lycée dans un État des États-Unis, et *score* indique le résultat en mathématiques d'un examen de mathématique standardisé. Soit *girlshs* une variable binaire indiquant si l'élève a été scolarisé dans un lycée de filles.

i. De l'influence de quels autres facteurs auriez-vous envie de tenir compte dans l'équation ? (On fait l'hypothèse que vous pouvez collecter des données sur ces facteurs.)

ii. Écrivez une équation reliant *score* à *girlshs* et aux autres facteurs que vous avez cités dans la question (i).

iii. On suppose que le soutien des parents et la motivation sont des facteurs que l'on ne peut pas mesurer et qui sont donc dans le terme d'erreur de la question (ii). Peuvent-ils vraisemblablement être corrélés à *girlshs* ? Expliquez.

iv. Soit *numghs* une variable indiquant pour chaque fille le nombre de lycées de filles qui se trouvent à moins de 20 kilomètres de chez elle. Présentez les hypothèses sous lesquelles *numghs* s'avère un instrument valide pour *girlshs*.

v. Supposons que, au moment d'estimer la forme réduite pour *girlshs*, vous trouviez que le coefficient de *numghs* est négatif et statistiquement significatif. Pensez-vous que procéder à une estimation par VI, en utilisant *numghs* comme instrument pour *girlshs* est une bonne idée ? Expliquez.

9. Imaginez que, dans l'équation (15.8), vous ne disposiez pas d'un bon instrument pour *skipped*. Néanmoins, vous disposez de deux informations supplémentaires concernant les étudiants : vous connaissez la note obtenue à l'examen d'entrée à l'université ainsi que la note moyenne obtenue pendant la scolarité jusqu'au semestre précédent. Que feriez-vous à la place d'une estimation par VI ?

10. Dans un article, Evans et Schwab (1995) ont étudié les effets d'avoir été scolarisé dans un lycée catholique sur la probabilité de poursuivre ses études à l'université (*college* aux États-Unis). Pour être concret, on appelle *college* une variable binaire qui vaut un si l'étudiant va à l'université et zéro dans le cas contraire. On appelle *CathHS* une variable binaire qui vaut un si l'étudiant est allé dans un lycée catholique. Le modèle de probabilité linéaire s'écrit :

$$college = \beta_0 + \beta_1 CathHS + autres\ facteurs + u$$

« *autres facteurs* » inclut les variables suivantes : sexe, origines ethniques, revenu de la famille et éducation des parents.

i. Pour quelle raison *CathHS* pourrait être corrélée à u ?

ii. Evans et Schwab disposent de données sur les résultats à un examen standardisé que les étudiants ont passé lorsqu'ils étaient en deuxième année de licence. Que peut-on faire avec cette variable afin d'améliorer l'estimation *ceteris paribus* d'avoir été à un lycée catholique ?

iii. On note *CathRel* une variable binaire qui vaut un si l'étudiant est catholique. Discuter les conditions requises pour qu'elle puisse être considérée comme une VI valide pour *CathHS* dans l'équation précédente. Laquelle des deux peut être testée ?

iv. Comme on pouvait l'espérer, être catholique a un effet positif significatif sur le fait d'avoir été scolarisé dans un lycée catholique. Pensez-vous que *CathRel* est un instrument convaincant pour *CathHS* ?

11. Considérons un modèle simple de série temporelle dans lequel la variable explicative souffre d'une erreur de mesure classique :

$$y_t = \beta_0 + \beta_1 x_t^* + u_t \qquad \textbf{[15.58]}$$
$$x_t = x_t^* + e_t$$

dans lequel u_t est de moyenne zéro et n'est corrélé ni à x_t^*, ni à e_t. On observe seulement y_t et x_t. On fait l'hypothèse que e_t est de moyenne nulle et qu'elle n'est pas corrélée à x_t^*. Nous supposons également que x_t^* est de moyenne nulle (cette dernière hypothèse sert seulement à simplifier les calculs).

i. Écrivez $x_t^* = x_t - e_t$ et injectez-la dans (15.58). Montrez que le terme d'erreur dans la nouvelle équation, qu'on appelle v_t, est corrélé négativement à x_t si $\beta_1 > 0$. Qu'est-ce que cela implique pour l'estimateur par MCO de β_1 obtenu par la régression de y_t sur x_t ?

ii. En plus des hypothèses précédentes, nous supposons que u_t et e_t ne sont pas corrélées aux valeurs passées de x_t^* et de e_t, en particulier, elles ne sont corrélées ni à x_{t-1}^*, ni à e_{t-1}. Montrer que $E(x_{t-1}v_t) = 0$, où v_t est le terme d'erreur du modèle de la question (i).

iii. Vous semble-t-il possible que x_t et x_{t-1} soient corrélées ? Expliquez.

iv. Quelle stratégie vous est suggérée par les réponses aux questions (ii) et (iii) afin d'estimer de manière convergente β_0 et β_1 ?

EXERCICES SUR ORDINATEUR

C1. Utilisez les données de WAGE2 pour cet exercice.

i. Dans l'exemple 15.2, en utilisant *sibs* comme instrument pour *educ*, l'estimation par VI nous indique que le rendement de l'éducation est de 0,122. Afin de vous convaincre qu'utiliser *sibs* comme VI pour *educ* n'est *pas* la même chose que remplacer *educ* par *sibs* puis avoir recours à une régression MCO, régressez log(*wage*) sur *sibs* et expliquez vos résultats.

ii. La variable *brthord* nous renseigne sur le rang de naissance (*brthord* vaut 1 pour l'ainé, 2 pour le cadet, et ainsi de suite). Expliquez pourquoi il est probable que *educ* et *brthord* soient négativement corrélées. Régressez *educ* sur *brthord* pour déterminer s'il existe une corrélation négative et statistiquement significative.

iii. Utilisez *brthord* comme VI pour *educ* dans l'équation (15.1). Donnez les résultats et interprétez-les.

iv. Supposons maintenant que l'on veuille inclure le nombre de frères et sœurs dans l'équation de salaire, afin de tenir compte de l'influence, dans une certaine mesure, de l'environnement familial.

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 sibs + u$$

On veut utiliser *brthord* comme VI pour *educ*, en faisant l'hypothèse que *sibs* est exogène. La forme réduite de *educ* s'écrit :

$$educ = \pi_0 + \pi_1 sibs + \pi_2 brthord + v$$

Écrivez l'hypothèse d'identification et testez-la.

v. Estimez l'équation de la question (iv) en utilisant *brthord* comme VI pour *educ* (et *sibs* joue le rôle de sa propre VI). Commentez les écarts-types estimés de $\hat{\beta}_{educ}$ et de $\hat{\beta}_{sibs}$.

vi. En utilisant les valeurs ajustées obtenues dans la question (iv), $\widehat{educ}$, calculez la corrélation entre $\widehat{educ}$ et *sibs*. Utilisez ces résultats pour expliquer les résultats obtenus à la question (v).

C2. Les données de la base FERTIL2 contiennent, pour un échantillon de femmes au Botswana en 1988, des informations sur leur nombre d'enfants (*children*), leur nombre d'années d'éducation (*educ*), leur âge (*age*), ainsi que des variables sur le statut religieux et économique.

i. Estimez par MCO le modèle suivant :

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

Interprétez les résultats. En particulier, en gardant *age* constant, quels sont les effets estimés d'une année d'éducation supplémentaire sur la fécondité ? Si 100 femmes reçoivent une année d'éducation supplémentaire, combien d'enfants en moins pense-t-on qu'elles auront ?

ii. La variable *frsthalf* est une variable binaire qui vaut un si la femme est née au cours des 6 premiers mois de l'année. En postulant que *frsthalf* n'est pas corrélée au terme d'erreur de la question (i), montrez que *frsthalf* est un bon candidat pour être une VI pour *educ*. (Indice : vous devez faire une régression.)

iii. Estimez le modèle de la question (i) en utilisant *frsthalf* comme VI pour *educ*. Comparez l'effet de l'éducation estimé par VI à celui estimé par MCO dans la question (i).

iv. Ajoutez au modèle les variables binaires *electric*, *tv* et *bicycle* (qui valent un si le ménage est équipé respectivement de l'électricité, la télévision et un vélo) et supposons qu'elles sont exogènes. Estimez l'équation par MCO et par DMC et comparez les coefficients estimés pour *educ*. Interprétez le coefficient pour *tv* et expliquez pourquoi posséder un téléviseur a un effet négatif sur la fécondité.

C3. Utilisez les données de CARD pour cet exercice.

i. L'équation estimée pour l'exemple 15.4 peut s'écrire :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \ldots + u$$

où les autres variables explicatives sont détaillées dans le tableau 15.1. Afin d'obtenir un estimateur des VI convergent, l'instrument pour *educ*, c'est-à-dire *nearc4* (université à proximité du lieu de résidence pendant l'enfance), ne doit pas être corrélée à u. Est-il possible que *nearc4* soit corrélée à d'autres variables contenues dans le terme d'erreur, comme les capacités inobservées ? Expliquez.

ii. Pour un sous-échantillon d'hommes présents dans la base de données, on connaît les résultats à un test de QI (*IQ*). Régressez *IQ* sur *nearc4* pour vérifier si les résultats moyens au test de QI sont différents si l'homme a grandi à proximité d'une université. Qu'en concluez-vous ?

iii. Régressez à présent *IQ* sur *nearc4*, *smsa66*, et des variables binaires régionales en 1966 : *reg662*, ..., *reg669*. Existe-t-il un lien entre *IQ* et *nearc4* une fois qu'on a neutralisé l'effet des variables binaires régionales ? Réconciliez ce résultat aux résultats obtenus pour la question (ii).

iv. À partir des questions (ii) et (iii), que concluez-vous quant à l'importance de tenir compte de l'influence de *smsa66* et des variables binaires régionales en 1966 dans l'équation de log(*wage*) ?

C4. Pour cet exercice, utilisez les données de INTDEF.
On considère l'équation simple qui lie les taux des bons du trésor américain à trois mois (*i3*) au taux d'inflation (*infl*, variable construite à partir de l'indice des prix à la consommation – *Consumer Price Index*) suivante :

$$i3_t = \beta_0 + \beta_1 infl_t + u_t$$

i. Estimez cette équation par MCO, en ne prenant pas en compte la première date pour pouvoir comparer les résultats plus tard. Présentez les résultats sous la forme usuelle.

ii. Certains économistes pensent que l'indice des prix à la consommation n'est pas une bonne mesure du taux réel d'inflation, ce qui implique que l'estimation par MCO estimée en question (i) souffre d'un biais d'erreur de mesure. Réestimez l'équation de la question (i), en utilisant $infl_{t-1}$ comme VI pour $infl_t$. Dans quelle mesure l'estimation VI de β_1 est-elle comparable à l'estimation MCO ?

iii. Calculez à présent la différence première de l'équation :

$$\Delta i3_t = \beta_1 \Delta infl_t + \Delta u_t$$

Estimez cette équation par MCO et comparez l'estimation de β_1 obtenue avec les précédentes estimations.

iv. Est-il possible d'utiliser $\Delta infl_{t-1}$ comme instrument pour $\Delta infl_t$ dans l'équation différenciée calculée dans la question (iii) ? Expliquez. (Indice : $\Delta infl_{t-1}$ et $\Delta infl_t$ sont-elles suffisamment corrélées ?)

C5. Utilisez les données de CARD pour cet exercice.

i. Dans le tableau 15.1, la différence entre l'estimation du rendement de l'éducation obtenu par MCO et celle obtenue par VI est importante d'un point de vue économique. Calculez les résidus de la forme réduite, $\hat{v}_2$, issus de la régression de *educ* sur *nearc4*, *exper*, *exper*^{2}, *black*, *smsa*, *south*, *smsa66*, *reg662*, ..., *reg669* – voir tableau 15.1.

Utilisez les pour tester si *educ* est exogène, c'est-à-dire déterminez si les estimations MCO et VI sont *statistiquement* différentes.

ii. Estimez l'équation par DMC, en ajoutant *nearc2* comme instrument. Est-ce que le coefficient de *educ* change beaucoup ?

iii. Testez la seule restriction de suridentification issue de la question (ii).

C6. Utilisez les données de MURDER pour cet exercice.
La variable *mrdrte* représente le taux d'homicide, c'est-à-dire le nombre d'homicides pour 100.000 personnes. La variable *exec* représente le nombre total de prisonniers exécutés durant l'année en cours et les deux précédentes, *unem* est le taux de chômage de l'État.

i. Combien d'États ont exécuté au moins un prisonnier en 1991, 1992 ou 1993 ? Dans quel État les exécutions sont-elles les plus nombreuses ?

ii. En utilisant les deux années 1990 et 1993, faites tourner une régression sur données agrégées de *mrdrte* sur *d93*, *exec* et *unem*. Que pensez-vous du coefficient de *exec* ?

iii. En utilisant les variations de 1990 à 1993 seulement (pour un total de 51 observations), estimez par MCO l'équation

$$\Delta mrdrte = \delta_0 + \beta_1 \Delta exec + \beta_2 \Delta unem + \Delta u$$

Reportez les résultats sous la forme standard. Ces estimations-ci laissent-elles entendre que la peine capitale semble avoir un effet dissuasif ?

iv. Les variations du nombre d'exécutions peuvent être, au moins en partie, liées aux variations attendues du nombre d'homicides, ce qui aurait pour conséquence une corrélation entre $\Delta exec$ et Δu dans la question (iii). Il peut être raisonnable de penser que $\Delta exec_{-1}$ n'est pas corrélée à Δu. (Après tout, $\Delta exec_{-1}$ dépend du nombre d'exécutions qui ont eu lieu il y a trois ans ou plus.) Régressez $\Delta exec$ sur $\Delta exec_{-1}$ pour voir si elles sont suffisamment corrélées et interprétez le coefficient de $\Delta exec_{-1}$.

v. Réestimez l'équation de la question (iii), en utilisant $\Delta exec_{-1}$ comme instrument pour $\Delta exec$. Supposez que $\Delta unem$ est exogène. En quoi vos conclusions changent-elles par rapport à celles de la partie (iii) ?

C7. Pour cet exercice, utilisez les données de PHILLIPS.

i. Dans l'exemple 11.5, nous avons estimé une courbe de Phillips augmentée des anticipations (*expectations augmented Phillips curve*) de la forme :

$$\Delta infl_t = \beta_0 + \beta_1 unem_t + e_t$$

Dans laquelle $\Delta infl_t = infl_t - infl_{t-1}$, avec $infl_t$ l'inflation à la date t. En estimant cette équation par MCO, nous avons fait l'hypothèse que le choc d'offre e_t n'était pas corrélé à $unem_t$, le taux de chômage en date t. Si cette hypothèse est fausse, que peut-on dire de l'estimateur par MCO de β_1 ?

ii. Supposons que e_t n'est pas prévisible, étant donnée toute l'information passée : $E(e_t|infl_{t-1}, unem_{t-1}, \ldots) = 0$. Expliquez en quoi cette hypothèse fait de $unem_{t-1}$ un bon candidat de VI pour $unem_t$.

iii. Régressez $unem_t$ sur $unem_{t-1}$. Les variables $unem_{t-1}$ et $unem_t$ sont-elles significativement corrélées ?

iv. Estimez la courbe de Phillips augmentée des anticipations par VI. Reportez les résultats sous une forme usuelle et comparez-les aux estimations VI obtenues dans l'exemple 11.5.

C8. Pour cet exercice, utilisez les données de 401KSUBS.
L'équation d'intérêt est le modèle de probabilité linéaire suivant :

$$pira = \beta_0 + \beta_1 p401k + \beta_2 inc + \beta_3 inc^2 + \beta_4 age + \beta_5 age^2 + u$$

L'objectif est ici de tester si les individus font un choix entre la participation à un plan 401(k) (système d'épargne retraite par capitalisation aux États-Unis), représentée par la variable *p401k* et la participation à un compte individuel de retraite (autre système d'épargne retraite – *individual retirement account*), représentée par la variable *pira*. On cherche donc à estimer β_1.

i. Estimez l'équation par MCO et discutez l'effet estimé de *p401k*.

ii. Quel est le problème potentiel de l'estimation par moindres carrés ordinaires quand l'objectif est d'estimer le choix, toutes choses égales par ailleurs, entre différents types de plans d'épargne retraite ?

iii. La variable *e401k* est une variable binaire qui vaut un si le salarié est éligible à un plan 401(k). Expliquez à quelles conditions *e401k* peut être une VI valide pour *p401k*. Ces hypothèses vous semblent-elles raisonnables ?

iv. Estimez la forme réduite de *p401k* et vérifiez que *e401k* fait état d'une corrélation partielle significative avec *p401k*. Puisque la forme réduite est aussi un modèle de probabilité linéaire, calculez des écarts-types des coefficients robustes à l'hétéroscédasticité.

v. Estimez maintenant l'équation structurelle par VI et comparez l'estimation de β_1 à l'estimation MCO. Une fois de plus, il est nécessaire de calculer des écarts-types des coefficients robustes à l'hétéroscédasticité.

vi. Testez l'hypothèse nulle selon laquelle *p401k* est en fait exogène, en utilisant un test robuste à l'hétéroscédasticité.

C9. Le but de cet exercice est de comparer les estimations des coefficients et de leurs écarts-types obtenues en utilisant correctement les DMC et celles obtenues en utilisant des procédures inappropriées. Utilisez les données de WAGE2.

i. Utilisez une commande des DMC pour estimer l'équation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 black + u$$

en utilisant *sibs* comme VI pour *educ*. Reportez les résultats sous la forme habituelle.

ii. Nous allons maintenant mettre en œuvre manuellement les DMC. Dans un premier temps, régressez $educ_i$ sur $sibs_i$, $exper_i$, $tenure_i$ et $black_i$, et calculez les valeurs ajustées $\widehat{educ}_i$, $i = 1, \ldots, n$. Ensuite, régressez la régression de seconde étape de $\log(wage_i)$ sur $\widehat{educ}_i$, $exper_i$, $tenure_i$ et $black_i$, $i = 1, \ldots, n$. Vérifiez que les $\hat{\beta}_j$ sont identiques à ceux obtenus dans la question (i), mais que les écarts-types estimés des coefficients sont légèrement différents. Les écarts-types estimés obtenus au cours de la régression de seconde étape lorsque les DMC sont mis en œuvre à la main sont généralement faux.

iii. À présent, nous allons mettre en place la procédure en deux étapes suivante, qui conduit généralement à des estimations des paramètres non convergentes pour les β_j, et pas seulement à des écarts-types faux. Pour la première étape, régressez $educ_i$ sur $sibs_i$ seulement, et calculez les valeurs ajustées, que l'on appelle $\widehat{educ}_i$. (Remarquez que cette première étape n'est pas correcte.) Ensuite, en seconde étape, régressez $\log(wage_i)$ sur $\widehat{educ}_i$, $exper_i$, $tenure_i$ et $black_i$, $i = 1, \ldots, n$. Comparez l'estimation du rendement de l'éducation obtenue à partir de cette procédure en deux étapes incorrecte à celle obtenue à partir de la procédure des DMC correcte.

C10. Utilisez les données de HTV pour cet exercice.

i. Estimez par MCO une régression linéaire simple, en régressant log(*wage*) (logarithme du salaire) sur *educ* (nombre d'années d'étude). Sans autre variable explicative, quel est l'intervalle de confiance à 95 % du rendement d'une année d'étude supplémentaire ?

ii. La variable *ctuit* exprime, en milliers de dollars, la variation des frais de scolarité pour les étudiants de 17 à 18 ans. Montrez que *ctuit* et *educ* ne sont pas significativement corrélées. Qu'est-ce que cela nous indique quant à la possibilité d'utiliser *ctuit* comme VI pour *educ* dans le cadre d'une régression simple ?

iii. À présent, ajoutez au modèle de régression simple de la question (i) une fonction quadratique pour l'expérience ainsi qu'un ensemble complet de variables binaires régionales pour la résidence à la date d'observation ainsi que pour la résidence à l'âge de 18 ans. Ajoutez également des variables indicatrices indiquant si le lieu de résidence à la date d'observation et celui à l'âge de 18 ans sont dans une zone urbaine. Quel est le rendement estimé d'une année d'étude supplémentaire ?

iv. En utilisant à nouveau *ctuit* comme VI potentielle pour *educ*, estimez la forme réduite de *educ*. (Bien entendu, la forme réduite d'*educ* inclut à présent les variables explicatives de la question (iii).) Montrez que *ctuit* est maintenant statistiquement significative dans la forme réduite d'*educ*.

v. Estimez le modèle de la question (iii) par VI, en utilisant *ctuit* comme VI pour *educ*. Comparez l'intervalle de confiance du rendement de l'éducation obtenu ici à celui obtenu par MCO dans la question (iii).

vi. Pensez-vous que la procédure VI de la question (v) est convaincante ?

C11. La base de données VOUCHER.DTA est un sous-échantillon des données utilisées par Rouse (1998) et peut être utilisée pour estimer les effets du choix d'école (parmi les écoles faisant partie du réseau « *choice school* », voir également exercice 7) sur les performances académiques. La scolarisation dans une école du réseau « *choice school* » a été financée par un bon d'échange, dont les récipiendaires ont été déterminés par tirage aléatoire parmi ceux qui avaient postulé. Le sous-échantillon de données a été choisi de sorte que tous les étudiants observés aient réussi l'examen de mathématiques en 1994 (qui est la dernière année disponible dans les données de Rouse). Malheureusement, comme le faisait remarquer Rouse, le résultat à l'examen de mathématiques n'est pas observé pour de nombreux étudiants, certainement à cause de l'attrition (c'est-à-dire ceux qui ont quitté leur école publique de quartier à Milwaukee). On observe dans les données des étudiants qui ont postulé au programme de bons d'échange et qui ont été acceptés, des étudiants qui ont postulé au programme de bons d'échange et qui n'ont pas été acceptés, ainsi que des étudiants qui n'ont pas postulé au programme de bons d'échange. Par conséquent, même si les récipiendaires des bons d'échange ont été déterminés par tirage aléatoire parmi ceux qui ont postulé, nous ne disposons pas nécessairement d'un échantillon tiré dans une population pour laquelle se voir attribuer un bon d'achat est un évènement déterminé aléatoirement. (En particulier, il est important de remarquer que les étudiants qui n'ont jamais postulé au programme de bons d'échange sont potentiellement différents de ceux qui ont postulé en moyenne, et il est impossible de prévoir la nature de ces différences avec l'information dont nous disposons.)

Rouse (1998) utilise des méthodes d'analyse propres aux données de panel comme celles que nous avons vues dans le chapitre 14, afin de prendre en compte des effets fixes étudiants. Elle utilise également des méthodes

par variables instrumentales. Ce problème nécessite de mener une analyse sur coupe transversale dans laquelle le fait d'avoir été tiré et donc d'avoir gagné un bon d'échange joue le rôle de variable instrumentale du fait d'avoir été scolarisé dans une école du réseau *« choice school »*. En fait, nous construisons deux variables, car nous observons plusieurs années pour chaque étudiant. La première variable construite, *choiceyrs*, représente le nombre d'années, entre 1991 et 1994, au cours desquelles l'étudiant a été scolarisé dans une école du réseau *« choice school »*, cette variable prend des valeurs entre 0 et 4. La seconde variable, *selectyrs*, indique le nombre d'années au cours desquelles l'étudiant a été tiré pour recevoir un bon d'échange. Si l'étudiant a postulé pour participer au programme en 1990 et qu'il a reçu un bon d'échange, alors *selectyrs* = 4. S'il ou elle a postulé pour le programme en 1991 et qu'il/elle a reçu un bon d'échange alors *selectyrs* = 3, et ainsi de suite. La variable d'intérêt est *mnce*, le centile atteint par le résultat de l'étudiant à un examen de mathématique passé en 1994.

i. Parmi les 990 étudiants de l'échantillon, combien n'ont jamais reçu de bon d'échange ? Combien d'entre eux ont reçu un bon d'échange valable pour les 4 années ? Combien d'étudiants ont été finalement scolarisés dans une école du réseau *« school choice »* pendant les quatre années ?

ii. Estimez par MCO un modèle simple, en régressant *choiceyrs* sur *selectyrs*. Le lien entre ces variables va-t-il dans la direction à laquelle vous vous attendiez ? La relation entre ces variables est-elle forte ? Est-ce que la variable *selectyrs* vous semble un bon candidat de VI pour *choiceyrs* ?

iii. Estimez par MCO un modèle simple, en régressant *mnce* sur *choiceyrs*. Quel résultat obtenez-vous ? Que se passe-t-il si vous ajoutez les variables *black*, *hispanic* et *female* (variables binaires indiquant si la personne est noire, hispanique et une femme, respectivement) ?

iv. Pourquoi la variable *choiceyrs* risque d'être endogène dans une équation du type :

$$mnce = \beta_0 + \beta_1 choiceyrs + \beta_2 black + \beta_3 hispanic + \beta_4 female + u_1$$

v. Estimez l'équation de la question (iv) par VI, en utilisant *selectyrs* comme VI pour *choiceyrs*. Est-ce que l'utilisation d'une procédure des VI nous indique que la scolarisation dans une école du réseau *« choice school »* a un effet positif ? Que pensez-vous des coefficients obtenus pour les autres variables explicatives ?

vi. Afin de tenir compte du cas où les performances passées affectent la participation au tirage pour obtenir un bon d'échange (également pour l'attrition), ajoutez *mnce90* (le résultat à l'examen de mathématique en 1990) dans l'équation de la question (iv). Estimez l'équation par MCO et par VI, puis comparez les résultats obtenus pour β_1. D'après l'estimation par VI, de combien chaque année passée dans une école de réseau *« school choice »* améliore-t-elle le résultat en centile à l'examen de mathématique ? Peut-on considérer que cet effet est quantitativement important ?

vii. Pourquoi l'analyse proposée dans la question (vi) n'est pas entièrement convaincante ? (Indice : par rapport à la question (v), qu'est-ce que le nombre d'observations change et pourquoi ?)

viii. On appelle *choiceyrs1*, *choiceyrs2*, etc. des variables binaires qui indiquent différents nombres d'années qu'un étudiant peut avoir passées dans une école du réseau *« school choice »* (entre 1991 et 1994). On appelle *selectyrs1*, *selectyrs2*, etc. des variables binaires ayant une définition similaire, mais qui indiquent avoir été tiré au sort pour recevoir un bon d'échange. Estimez par VI l'équation suivante, en utilisant les quatre variables binaires *selectyrs* comme instruments (Comme avant, les variables *black*, *hispanic*, et *female* jouent le rôle de leur propre VI.) :

$$mnce = \beta_0 + \beta_1 choiceyrs1 + \beta_2 choiceyrs2 + \beta_3 choiceyrs3 + \beta_4 choiceyrs4 + \beta_5 black + \beta_6 hispanic + \beta_7 female + u_2$$

Décrivez vos résultats. Ont-ils du sens ?

C.12 Utilisez les données de la base CATHOLIC pour répondre à cet exercice. Le modèle d'intérêt est le suivant :

$$math12 = \beta_0 + \beta_1 cathhs + \beta_2 lfaminc + \beta_3 motheduc + \beta_4 fatheduc + u$$

où *cathhs* est une variable binaire qui indique si un élève fréquente une école secondaire catholique.

i. Combien y a-t-il d'élèves dans l'échantillon ? Quel pourcentage de ces élèves fréquente une école secondaire catholique ?

ii. Estimer l'équation ci-dessus par MCO. Quelle est la valeur estimée de β_1 ? Quel est son intervalle de confiance à 95 % ?

iii. En utilisant la variable *parcath* comme instrument pour la variable *cathhs*, estimer la forme réduite pour la variable *cathhs*. Quelle est la statistique de test *t* pour *parcath* ? Cette valeur indique-t-elle que l'on fait face à un problème d'instrument faible ?

iv. Estimer l'équation ci-dessus par VI, en utilisant *parcath* comme variable instrumentale pour la variable *cathhs*. Comparer la valeur estimée par VI ainsi que l'intervalle de confiance à 95 % aux quantités estimées par MCO.

v. Tester l'hypothèse nulle selon laquelle la variable *cathhs* serait exogène. Quelle est la *valeur-p* du test ?

vi. Supposons que vous ajoutiez l'interaction entre la variable *cathhs* et la variable *mothereduc* au modèle ci-dessus. Pourquoi cette interaction est-elle généralement endogène ? Pourquoi l'interaction entre la variable *parcath* et la variable *mothereduc* est-elle une bonne candidate pour être une variable instrumentale pour l'interaction entre la variable *cathhs* et la variable *mothereduc* ?

vii. Avant de créer les interactions décrites dans la question (vi), calculer d'abord la moyenne empirique de *mothereduc* et construire les variables $cathhs \cdot (mothereduc - \overline{mothereduc})$ et $parcath \cdot (mothereduc - \overline{mothereduc})$. Ajouter la première interaction au modèle à estimer et utiliser la seconde comme variable instrumentale. Bien entendu, *cathhs* est elle aussi instrumentée. Le terme d'interaction est-il statistiquement significatif ?

viii. Comparer le coefficient de la variable *cathhs* estimé en (vii) à celui estimé pour la question (iv). L'inclusion de l'interaction est-elle importante pour estimer l'effet partiel moyen ?

ANNEXE 15A

15A.1 Les hypothèses des doubles moindres carrés

Cette annexe décrit les hypothèses nécessaires sous lesquelles l'estimateur par DMC présente des propriétés asymptotiques désirables. Dans un premier temps, nous présentons les hypothèses pour des applications en coupe transversale avec un échantillon aléatoire. Ensuite, nous décrivons les hypothèses supplémentaires qu'il faut ajouter à celles-ci afin de pouvoir appliquer les DMC aux données de séries temporelles ou de panel.

15A.2 Hypothèse DMC.1 (Linéarité dans les paramètres)

Le modèle peut s'écrire de la forme suivante dans la population :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Dans lequel les $\beta_0, \beta_1, \ldots, \beta_k$ sont les paramètres inconnus (et constants) d'intérêt et u est le terme aléatoire inobservé ou le terme aléatoire de perturbation. Les variables instrumentales sont notées z_j.

Il est important de remarquer que l'hypothèse DMC.1 est quasiment identique à RLM.1 (à ceci près que DMC.1 mentionne la notation des variables instrumentales, z_j). En d'autres termes, le modèle auquel on s'intéresse est le même que celui présenté pour l'estimation par MCO des β_j. Il est parfois facile de perdre de vue que l'on peut mettre en œuvre différentes méthodes d'estimation pour le même modèle. Malheureusement, il est assez courant d'entendre des chercheurs dire « J'ai estimé un modèle des MCO » ou « J'ai utilisé un modèle des DMC ». Ce type d'affirmation n'a aucun sens. Les MCO et les DMC sont différentes méthodes d'*estimation* qui peuvent être appliquées sur le *même* modèle. Il est vrai qu'elles ont des propriétés statistiques désirables sous différentes hypothèses, mais elles permettent toutes les deux d'estimer la même relation, donnée par l'équation en DMC.1 (ou RLM.1). Cette remarque est similaire à celle que l'on a faite pour les modèles de données de panel avec effets inobservés que l'on a vus dans les chapitres 13 et 14 : MCO sur données agrégées, différence première, effets fixes et effets aléatoires sont différentes méthodes d'estimation pour le même modèle.

15A.3 Hypothèse DMC.2 (Échantillon aléatoire)

On dispose d'un échantillon aléatoire sur les variables y, x_j et z_j.

15A.4 Hypothèse DMC.3 (Condition de rang)

i. Il n'existe pas de relation linéaire parfaite entre les variables instrumentales.
ii. La condition de rang pour l'identification est vérifiée.

Avec une seule variable explicative endogène, comme dans l'équation (15.42), la condition de rang est facile à décrire. On appelle $z_1, \ldots, z_m$ les variables exogènes, parmi lesquels $z_k, \ldots, z_m$ n'apparaissent pas dans le modèle structurel (15.42). La forme réduite de y_2 s'écrit :

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_{k-1} z_{k-1} + \pi_k z_k + \ldots + \pi_m z_m + v_2$$

Nous savons qu'il est nécessaire qu'au moins l'un des $\pi_k, \ldots, \pi_m$ soit différent de zéro. Cela signifie qu'au moins une variable exogène n'apparaît pas dans (15.42) (condition d'ordre). Présenter la condition de rang pour 2 variables explicatives endogènes ou plus nécessite le recours au calcul matriciel (Voir Wooldridge, 2010, chapitre 5).

15A.5 Hypothèse DMC.4 (Variable instrumentale exogène)

Le terme d'erreur u est de moyenne nulle et aucune VI n'est corrélée à u.
Rappelez-vous que chaque x_j qui n'est pas corrélée à u joue le rôle de sa propre VI.

15A.6 Théorème 15A.1

Sous les hypothèses DMC.1 à DMC.4, l'estimateur par DMC est convergent.

15A.7 Hypothèse DMC.5 (Homoscédasticité)

Soit z l'ensemble des variables instrumentales. Alors, $E(u^2|z)=\sigma^2$

15A.8 Théorème 15A.2

Sous les hypothèses DMC.1 à DMC.5, les estimateurs par DMC ont asymptotiquement une distribution normale. Les estimateurs convergents de la variance asymptotique sont donnés par l'équation (15.43), dans laquelle on remplace σ^2 par $\hat{\sigma}^2 = (n-k-1)^{-1}\sum_{i=1}^{n}\hat{u}_i^2$, les $\hat{u}_i$ étant les résidus DMC.

L'estimateur par DMC est aussi le meilleur estimateur par VI sous les cinq hypothèses énoncées avant. Nous ne faisons qu'annoncer ce résultat ici, la preuve est donnée dans Wooldridge (2010, Chapitre 5).

15A.9 Théorème 15A.3

Sous les hypothèses DMC.1 à DMC.5, l'estimateur par DMC est asymptotiquement efficace dans la classe des estimateurs par VI qui utilisent des combinaisons linéaires des variables exogènes comme instruments.

Si l'hypothèse d'homoscédasticité n'est pas vérifiée, les estimateurs par DMC ont toujours une distribution asymptotiquement normale, mais les écarts-types estimés des coefficients (et donc les statistiques de test de Student et de Fisher) doivent être ajustés. De nombreux logiciels d'économétrie le font systématiquement. De plus, si cette hypothèse n'est pas vérifiée, l'estimateur par DMC n'est plus, en général, l'estimateur par VI asymptotiquement efficace. Nous n'étudions pas dans cet ouvrage d'autres estimateurs efficaces. (voir Wooldridge, 2010, chapitre 8).

Pour l'application aux données de séries temporelles, il nous faut ajouter des hypothèses. D'abord, comme pour les MCO, il nous faut faire l'hypothèse que toutes les séries (y compris les VI) sont faiblement indépendantes : cela nous assure le fait que la loi des grands nombres et le théorème central limite sont vérifiés. Il nous faut également ajouter une hypothèse d'absence de corrélation sérielle, afin d'assurer la validité des écarts-types estimés des coefficients classiques et des statistiques de test, ainsi que l'efficacité asymptotique.

15A.10 Hypothèse DMC.6 (Absence de corrélation sérielle)

L'équation (15.54) est vérifiée.

Pour les applications sur données de panel, nous avons besoin d'une hypothèse similaire d'absence de corrélation sérielle. Les tests et les corrections pour la corrélation sérielle sont discutés dans la section 15.7.

CHAPITRE

16

MODÈLES À ÉQUATIONS SIMULTANÉES

Traduction de Pierre André

Dans le chapitre précédent, nous avons montré comment la méthode des variables instrumentales pouvait résoudre deux types de problèmes d'endogénéité : l'endogénéité due aux variables omises et celle due à l'erreur de mesure. Conceptuellement, ces deux problèmes sont simples. Dans le cas des variables omises, en mesurant l'effet *ceteris paribus* d'une variable explicative (ou de plusieurs), nous souhaiterions prendre en compte une variable inobservée (ou plusieurs) dans le *ceteris paribus*. Dans le cas de l'erreur de mesure, nous voudrions estimer l'effet de variables explicatives sur y, mais certaines variables sont mal mesurées. Dans les deux cas, les paramètres d'intérêt auraient pu être estimés par MCO avec de meilleures données.

Une autre forme importante d'endogénéité des variables explicatives est la **simultanéité**. Cette forme apparaît quand plusieurs variables explicatives sont *déterminées en même temps* que la variable dépendante, par exemple par un mécanisme d'équilibre (comme nous le verrons plus tard). Dans ce chapitre, nous étudierons des méthodes pour estimer des modèles à équations simultanées (MES) simples. Bien que ce chapitre n'ait pas l'ambition de couvrir de manière exhaustive les MES, nous allons traiter les modèles les plus souvent utilisés.

La méthode la plus utilisée pour estimer des modèles à équations simultanées est la méthode des variables instrumentales. Dans ce cas, la solution au problème de simultanéité est donc similaire aux solutions proposées pour répondre aux problèmes de variables omises et d'erreur de mesure. Cependant, l'estimation pratique et l'interprétation des MES sont toujours des défis. Nous commencerons par discuter la nature et le domaine de pertinence des équations simultanées dans la section 16.1. Dans la section 16.2, nous montrerons pourquoi l'estimation d'un système d'équations simultanées par les MCO est biaisée et non convergente dans le cas général.

La section 16.3 décrit l'identification dans un système à deux équations et la section 16.4 aborde rapidement les modèles avec plus de deux équations. Les modèles à équations simultanées sont utilisés pour modéliser les séries temporelles agrégées et la section 16.5 discute des problèmes spécifiques à ce genre de modèles. La section 16.6 évoque l'estimation des modèles à équations simultanées sur des données de panel.

16.1 DESCRIPTION DES MODÈLES À ÉQUATIONS SIMULTANÉES

Quand on utilise un modèle à équations simultanées, le point le plus important est de toujours garder à l'esprit que chaque équation du système doit avoir une interprétation causale *ceteris paribus*. Comme nous n'observons que les résultats à l'équilibre, nous devons utiliser un raisonnement contrefactuel pour construire les équations du modèle à équations simultanées. Il faut penser à la fois aux caractéristiques potentielles d'un individu et à ses caractéristiques observées.

L'exemple classique de MES est composé des équations d'offre et de demande d'un bien ou d'un facteur de production (par exemple le travail). Pour être concret, appelons h_s le nombre annuel d'heures travaillées dans l'agriculture, mesuré au niveau du département français, et appelons w le salaire horaire moyen obtenu par les travailleurs. On peut écrire une fonction simple d'offre de travail

$$h_s = \alpha_1 w + \beta_1 z_1 + u_1, \qquad \textbf{[16.1]}$$

où z_1 est une caractéristique affectant l'offre de travail agricole, par exemple le salaire moyen dans le secteur manufacturier du département. Le terme d'erreur, u_1, contient les autres facteurs déterminant l'offre de travail. [Beaucoup de ces facteurs sont observés et pourraient en réalité être inclus dans l'équation (16.1). Pour illustrer simplement notre problème, nous n'incluons qu'un facteur, z_1.] L'équation (16.1) est un exemple d'**équation structurelle**. Cette appellation vient du fait que la fonction d'offre de travail est tirée de la théorie économique et a une interprétation causale. Le coefficient α_1 indique dans quelle mesure l'offre de travail est affectée par le salaire ; si h_s et w sont en forme logarithmique, α_1 est l'élasticité prix de l'offre de travail. En général, on s'attendrait à ce qu'α_1 soit positif (quoique la théorie économique n'exclut pas non plus $\alpha_1 \leq 0$).

Les élasticités prix de l'offre de travail sont importantes pour déterminer dans quelle mesure les travailleurs souhaiteront travailler plus si le taux d'imposition sur les salaires change. Si z_1 est le salaire dans l'industrie, on s'attend à ce que $\beta_1 \leq 0$: toutes choses égales par ailleurs, si les salaires augmentent dans l'industrie, plus de travailleurs vont se tourner vers l'industrie et moins vers l'agriculture.

Quand on représente graphiquement l'offre de travail agricole, on représente le temps de travail total en fonction du salaire moyen, en gardant z_1 et u_1 constants. Un changement de z_1 modifie l'offre de travail, de même qu'un changement de u_1. La différence est que z_1 est observée alors que u_1 ne l'est pas. z_1 est parfois appelée *facteur d'offre observé*, et u_1 est appelé *facteur d'offre inobservé*.

En quoi l'équation (16.1) diffère-t-elle de ce qu'on a observé jusque ici ? La différence est subtile. L'équation (16.1) est supposée vraie pour toutes les valeurs de salaire possible, mais on ne peut pas s'attendre à ce que le salaire varie de manière exogène entre les départements. S'il était possible de faire une expérience fixant les niveaux de salaire agricole et industriel de chaque département et de mesurer les temps de travail des salariés pour mesurer l'offre de travail dans chaque département h_s, il serait possible d'estimer (16.1) par les MCO. Malheureusement pour l'économètre, cette expérience n'est pas faisable. Il est possible de collecter des données sur les salaires moyens dans ces deux secteurs ainsi que le temps de travail total dans le secteur agricole. Pour analyser ces données, il faut bien comprendre que les données collectées sont au mieux déterminées par l'intersection de l'offre *et* de la demande de travail. En supposant que les marchés du travail soient à l'équilibre, on observe donc les *valeurs d'équilibre* des salaires et de la quantité de travail.

Pour décrire la manière dont les salaires et la quantité de travail sont déterminés, il faut une fonction de demande de travail, que nous supposons donnée par

$$h_d = \alpha_2 w + \beta_2 z_2 + u_2, \qquad \textbf{[16.2]}$$

où h_d est la demande de travail. Comme pour la fonction d'offre, la demande de travail dépend du salaire, w, une fois z_2 et u_2 fixés. La variable z_2 – disons la surface de terres agricoles – est un *facteur de demande observé*, alors qu'u_2 est un *facteur de demande inobservé*.

Comme pour l'équation d'offre de travail, l'équation de demande de travail est une équation structurelle : elle peut être obtenue à partir de la maximisation du profit des exploitants agricoles. Si h_d et w sont sous forme logarithmique, α_2 est l'élasticité prix de la demande de travail. La théorie économique nous apprend que $\alpha_2 < 0$. Comme le travail et la terre sont compléments dans la fonction de production agricole, nous nous attendons à trouver $\beta_2 > 0$.

Remarquez que les équations (16.1) et (16.2) décrivent des mécanismes totalement différents. L'offre de travail décrit le comportement des travailleurs et la demande de travail décrit le comportement des exploitants agricoles. Chaque équation a une interprétation *ceteris paribus* et pourrait se suffire à elle-même. Elles sont cependant liées dans l'analyse économétrique parce que les salaires et la quantité de travail *observés* sont déterminés par l'intersection entre l'offre et la demande. En d'autres termes, pour chaque département i, la quantité de travail observée h_i et le salaire observé w_i sont déterminés par la condition d'équilibre

$$h_{is} = h_{id}. \qquad \textbf{[16.3]}$$

Pour chaque département, nous observons uniquement les heures de travail à l'équilibre, qui seront donc appelées h_i.

Nous pouvons combiner la condition d'équilibre (16.3) avec des équations de demande et d'offre de travail, obtenant alors

$$h_i = \alpha_1 w_i + \beta_1 z_{i1} + u_{i1} \qquad \textbf{[16.4]}$$

et

$$h_i = \alpha_2 w_i + \beta_2 z_{i2} + u_{i2}, \qquad [16.5]$$

où nous inclurons explicitement l'indice i pour rappeler que h_i et w_i sont les valeurs obtenues à l'équilibre pour le département i. Ces deux équations constituent un **modèle à équations simultanées (MES)**, qui a plusieurs caractéristiques importantes. D'abord, une fois z_{i1}, z_{i2}, u_{i1}, et u_{i2} fixés, ces deux équations déterminent h_i and w_i. (Précisément, il faut aussi supposer $\alpha_1 \neq \alpha_2$, c'est-à-dire que les pentes de la demande et de l'offre ne sont pas les mêmes, voir problème 1.) Pour cette raison, h_i et w_i sont les variables endogènes de ce MES. Que dire de z_{i1} et z_{i2} ? Comme elles ne sont pas déterminées par le modèle, nous les considérons comme des **variables exogènes**. D'un point de vue statistique, l'hypothèse principale concernant z_{i1} et z_{i2} est qu'aucune d'entre elles n'est corrélée avec les erreurs des équations de demande et d'offre, respectivement u_{i1} et u_{i2}. Ces dernières sont deux exemples d'**erreurs structurelles**, parce qu'elles apparaissent dans des équations structurelles.

Il est important de remarquer également que, si z_1 et z_2 sont exclues du modèle, il est impossible de dire quelle équation est la fonction d'offre et quelle équation est la fonction de demande. Quand z_1 est le salaire industriel, le raisonnement économique nous dit que c'est un facteur affectant l'offre de travail agricole, parce que cela mesure le coût d'opportunité du travail agricole ; quand z_2 est la surface des terres agricoles, la théorie du producteur nous indique que cela doit apparaître dans l'équation de demande de travail. Si z_1 et z_2 sont les mêmes – par exemple, le niveau d'éducation moyen des adultes du département, qui peut à la fois affecter l'offre et la demande de travail agricole – alors les deux équations sont les mêmes, et il n'y a donc aucun espoir de les identifier séparément. Cela illustre le problème d'identification des modèles à équations simultanées, que nous discuterons plus généralement dans la section 16.3.

EXEMPLE 16.1
Nombre d'assassinats et de policiers

Les villes voudraient certainement savoir quelle serait la diminution du nombre d'assassinats si on augmentait la taille des forces de maintien de l'ordre. Un modèle simple pour répondre à cette question sur données transversales s'écrit

$$murdpc = \alpha_1 polpc + \beta_{10} + \beta_{11} incpc + u_1, \qquad [16.6]$$

où *murdpc* est le nombre de meurtres par habitant, *polpc* est le nombre de policiers par habitant, et *incpc* est le revenu moyen (comme on peut le voir, l'indice i n'est pas inclus). Nous supposerons que le revenu par habitant est exogène dans cette équation. En pratique, il faudrait certainement inclure d'autres facteurs, comme la structure par âge et sexe des habitants, éventuellement des variables géographiques et des variables qui mesurent la sévérité des peines. Pour simplifier les choses, focalisons-nous sur l'équation (16.6).

La question à laquelle nous souhaiterions répondre est : si une ville augmente ses forces de police de manière exogène, cela va-t-il, en moyenne, diminuer la criminalité ? Si nous pouvions choisir la taille des forces de police de manière exogène pour un échantillon représentatif des villes, alors l'équation (16.6) pourrait être estimée par les MCO. Il n'est bien entendu pas possible de faire ce genre d'expérience. Mais pouvons-nous considérer que la taille des forces de police est déterminée de manière exogène ? Probablement pas : le budget de la police municipale dépend probablement de la criminalité attendue dans la ville. Pour refléter ceci, nous pouvons supposer une seconde relation :

$$polpc = \alpha_2 murdpc + \beta_{20} + autres\ facteurs. \qquad [16.7]$$

Nous nous attendons à trouver $\alpha_2 > 0$: toutes choses égales par ailleurs, les villes dont le taux de criminalité (attendu) est plus élevé auront plus de policiers par habitant. Une fois que les autres facteurs sont spécifiés dans cette équation (16.7), nous avons un modèle à deux équations. Seule l'équation (16.6) nous intéresse vraiment, mais comme nous le verrons dans la section 16.3, il faut savoir précisément comment la seconde équation est spécifiée pour estimer la première.

Un autre point important est que (16.7) décrit le comportement des équipes municipales, alors que (16.6) décrit les actions des assassins potentiels. Cela donne à chacune de ces deux équations une interprétation claire toutes choses égales par ailleurs, ce qui fait des équations (16.6) et (16.7) un modèle à équation simultanées interprétable.

Les exemples les plus convaincants de MES ressemblent à l'exemple des équations d'offre et de demande. Chaque équation, prise séparément, doit avoir une interprétation comportementale, *ceteris paribus*. Puisqu'on observe seulement les valeurs à l'équilibre, l'écriture d'un MES demande de se poser des questions contrefactuelles comme : combien d'heures de travail les travailleurs *voudraient*-ils offrir si les salaires *étaient* différents de leur valeur à l'équilibre ? L'exemple 16.1 donne un autre exemple de MES où chaque équation a une interprétation *ceteris paribus*.

Donnons maintenant un exemple d'usage inadapté de MES.

EXEMPLE 16.2
Dépenses de logement et épargne

Supposons que nous ayons fait l'hypothèse que, pour un ménage tiré au sort dans la population, les dépenses annuelles de logement et d'épargne soient déterminées de manière jointe par :

$$housing = \alpha_1 saving + \beta_{10} + \beta_{11} inc + \beta_{12} educ + \beta_{13} age + u_1 \quad [16.8]$$

et

$$saving = \alpha_2 housing + \beta_{20} + \beta_{21} inc + \beta_{22} educ + \beta_{23} age + u_2, \quad [16.9]$$

où *inc* est le revenu annuel, et *educ* et *age* sont mesurés en années. Il semblerait que ces équations sont une manière pertinente de voir comment les dépenses de logement et d'épargne sont déterminées. Mais il faut se demander ce que serait une de ces équations sans l'autre. Aucune des deux n'a d'interprétation *ceteris paribus*, car *housing* et *saving* sont choisis par le même ménage. Par exemple, se poser la question suivante n'a aucun sens : si le revenu annuel augmentait de 10.000 €, de combien les dépenses de logement évolueraient, *pour un niveau d'épargne donné* ? Si le revenu familial augmente, un ménage changera généralement son choix de dépenses de logement et d'épargne. Mais l'équation (16.8) laisse penser que nous cherchons à connaître les effets de *inc*, *educ*, et *age* en gardant *saving* inchangé. Cette expérience de pensée n'est pas intéressante. Dans tout modèle basé sur des principes économiques, en particulier si les ménages maximisent une fonction d'utilité, les ménages changent à la fois *housing* et *saving* en fonction de *inc* et des prix relatifs du logement et de l'épargne. Les variables *educ* et *age* affecteraient les préférences pour la consommation, l'épargne et le risque. En conséquence, *housing* et *saving* pourraient toutes les deux être exprimées en fonction du revenu, de l'éducation, de l'âge, et d'autres variables affectant le problème du consommateur (comme les rendements de l'investissement immobilier et des autres investissements).

Même si nous décidions que le MES composé de (16.8) et (16.9) a un sens, il est impossible d'estimer ses paramètres (nous discuterons ce problème d'une manière plus générale en section 16.3). On ne peut pas distinguer les deux équations, sauf en supposant que le revenu, l'éducation ou l'âge apparaît dans une équation mais pas dans l'autre, ce qui n'aurait pas de sens.

Cet exemple est un mauvais exemple de MES, mais nous pourrions vouloir tester si, toutes choses égales par ailleurs, il existe un arbitrage entre dépenses de logement et épargne. Dans ce cas, il faudrait estimer (16.8) avec les MCO, sauf s'il y a un biais de variables omises ou un problème d'erreur de mesure.

Trop souvent, les MES sont utilisés dans des situations tout aussi problématiques que l'exemple 16.2. Le problème est que les deux variables endogènes sont choisies par le même agent. Par conséquent, aucune des équations ne s'interprète individuellement. Un autre exemple d'utilisation fallacieuse des MES serait de

modéliser le nombre d'heures hebdomadaires passées à étudier et à travailler. Chaque étudiant choisit ces deux variables de manière simultanée – probablement en considérant, entre autres, les salaires potentiels, les capacités scolaires, son goût pour les études. Exactement comme dans l'exemple 16.2, spécifier deux équations où chacune dépend de l'autre n'a pas de sens. Le message à retenir ici est celui-ci : quand deux variables sont déterminées de manière simultanée, cela ne veut *pas* toujours dire qu'un modèle à équations simultanées est pertinent. Pour qu'un MES ait un sens, il faut que chaque équation ait une interprétation *ceteris paribus*. Comme nous en avons discuté précédemment, l'exemple de l'offre et de la demande, ou l'exemple 16.1, correspondent à ce cas. En général, les raisonnements économiques de base, parfois aidés de modèles économiques simples, peuvent aider à utiliser les MES intelligemment (et à savoir quand il ne faut pas les utiliser).

Pour aller plus loin 16.1

Pindyck et Rubinfeld (1992, section 11.6) décrivent un modèle de publicité où des entreprises monopolistiques choisissent leurs prix et les dépenses de publicités pour maximiser leur profit. Cela a-t-il un sens d'utiliser un MES pour modéliser ces variables au niveau de l'entreprise ?

16.2 BIAIS DE SIMULTANÉITÉ DES MCO

On peut facilement comprendre qu'une variable explicative déterminée simultanément avec la variable dépendante est généralement corrélée avec le terme d'erreur, ce qui entraîne un biais et la non-convergence des MCO. Considérons le modèle structurel à deux équations

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \qquad [16.10]$$

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2 \qquad [16.11]$$

et intéressons-nous à l'estimation de la première équation. Les variables z_1 et z_2 sont exogènes, donc aucune n'est corrélée avec u_1 ou u_2. Pour simplifier, nous supprimons la constante dans chaque équation.

Pour montrer qu'y_2 est généralement corrélée avec u_1, nous résolvons le modèle à deux équations en écrivant y_2 en fonction des variables exogènes et des termes d'erreur. Si nous mettons le terme de droite de (16.10) à la place de y_1 dans (16.11), nous trouvons

$$y_2 = \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + u_1) + \beta_2 z_2 + u_2$$

ou

$$(1 - \alpha_2\alpha_1)y_2 = \alpha_2\beta_1 z_1 + \beta_2 z_2 + \alpha_2 u_1 + u_2. \qquad [16.12]$$

Nous devons donc faire une hypothèse sur les paramètres pour trouver y_2 :

$$\alpha_2\alpha_1 \neq 1. \qquad [16.13]$$

Cette hypothèse peut être restrictive ou non, en fonction des cas. Dans l'exemple 16.1, nous pensons que $\alpha_1 \leq 0$ et $\alpha_2 \geq 0$, ce qui implique $\alpha_1\alpha_2 \leq 0$; donc (16.13) est très raisonnable dans l'exemple 16.1.

Si la condition (16.13) est vraie, nous pouvons diviser (16.12) par $(1 - \alpha_2\alpha_1)$ et écrire y_2 comme

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + v_2, \qquad [16.14]$$

où $\pi_{21} = \alpha_2\beta_1/(1 - \alpha_2\alpha_1)$, $\pi_{22} = \beta_2/(1 - \alpha_2\alpha_1)$, et $v_2 = (\alpha_2 u_1 + u_2)/(1 - \alpha_2\alpha_1)$. L'équation (16.14), qui exprime y_2 en fonction des variables exogènes et des termes d'erreur, est **l'équation de forme réduite** prédisant y_2, un concept que nous avons introduit dans le chapitre 15 dans le contexte de l'estimation des variables instrumentales. Les paramètres π_{21} et π_{22} sont appelés **paramètres de forme réduite**. Nous pouvons constater

que ce sont des fonctions non-linéaires des **paramètres structurels** qui apparaissent dans les équations (16.10) et (16.11).

L'erreur de forme réduite, v_2, est une fonction linéaire des erreurs structurelles u_1 et u_2. Comme u_1 et u_2 ne sont pas corrélées à z_1 et z_2, v_2 n'est pas non plus corrélée à z_1 ou z_2. Nous pouvons donc estimer π_{21} et π_{22} par les MCO, ce qui est utilisé pour les doubles moindres carrés (nous y reviendrons dans la section suivante). De plus, les paramètres de forme réduite peuvent parfois être intéressants en eux-mêmes, bien que nous nous focalisions ici sur l'estimation de (16.10).

Une forme réduite prédisant y_1 existe aussi sous l'hypothèse (16.13) ; les calculs sont similaires à ceux pour obtenir (16.14). Elle a les mêmes propriétés que la forme réduite prédisant y_2.

Nous pouvons utiliser l'équation (16.14) pour montrer que, sauf dans des cas très particuliers, l'estimation de l'équation (16.10) par les MCO produira une estimation biaisée et non-convergente de α_1 et β_1. Comme z_1 et u_1 ne sont pas corrélés, la question est de savoir si y_2 et u_1 sont corrélés. Dans l'équation de forme réduite (16.14), nous voyons que y_2 et u_1 sont corrélés si et seulement si v_2 et u_1 sont corrélées (puisque z_1 et z_2 sont supposées exogènes). Mais v_2 est une fonction linéaire de u_1 et u_2, donc en général, v_2 est corrélée à u_1. En fait, si nous supposons que u_1 et u_2 ne sont pas corrélées, alors v_2 et u_1 sont *toujours* corrélées quand $\alpha_2 \neq 0$. Même si α_2 est nul – ce qui veut dire que y_1 n'apparaît pas dans l'équation (16.11) – v_2 et u_1 seront corrélées si u_1 et u_2 sont corrélées. Quand $\alpha_2 = 0$ et u_1 et u_2 ne sont pas corrélées, y_2 et u_1 ne sont pas non plus corrélées. Ces conditions sont très restrictives : si $\alpha_2 = 0$, y_2 n'est plus déterminée simultanément avec y_1. Si en plus, la corrélation entre u_1 et u_2 est nulle, cela règle le problème de variables omises ou d'erreur de mesure contenue dans u_1 qui serait corrélée avec y_2. Il n'est donc pas surprenant que l'estimation de l'équation (16.10) par les MCO fonctionne dans ce cas.

Quand y_2 est corrélé avec u_1 pour des raisons de simultanéité, on peut dire que les MCO ont un **biais de simultanéité**. Trouver le sens du biais est difficile dans le cas général, comme nous avons vu avec les biais de variables omises dans les chapitres 3 et 5. Mais dans les modèles simples, nous pouvons deviner le sens du biais. Par exemple, supposons que l'équation (16.10) peut être simplifiée en supprimant z_1, et supposons que u_1 et u_2 ne sont pas corrélées. La covariance entre y_2 et u_1 est alors

$$\begin{aligned}\text{Cov}(y_2,u_1) &= \text{Cov}(v_2,u_1) = [\alpha_2 / (1 - \alpha_2\alpha_1)]\text{E}(u_1^2) \\ &= [\alpha_2 / (1 - \alpha_2\alpha_1)]\sigma_1^2,\end{aligned}$$

où $\sigma_1^2 = \text{Var}(u_1) > 0$. Donc le biais asymptotique de l'estimateur des MCO de α_1 a le même signe que $\alpha_2/(1 - \alpha_2\alpha_1)$. Si $\alpha_2 > 0$ et $\alpha_2\alpha_1 < 1$, le biais asymptotique est positif. (Malheureusement, comme dans notre calcul de variables omises dans la section 3.3, les conclusions ne se généralisent pas à des modèles plus généraux. Mais elles peuvent donner des intuitions utiles.) Par exemple, dans l'exemple 16.1, nous pensons que $\alpha_2 > 0$ et $\alpha_2\alpha_1 \leq 0$, ce qui veut dire que l'estimateur de α_1 par les MCO serait biaisé vers le haut. Si $\alpha_1 = 0$, les MCO estimeraient, en moyenne, un effet *positif* du nombre de policiers sur le nombre d'assassinats. Puisque nous nous attendons à ce que le nombre de policiers réduise (ceteris paribus) le nombre d'assassinats, le biais vers le haut veut dire que les MCO auront tendance à sous-estimer l'efficacité des forces de police.

16.3 IDENTIFIER ET ESTIMER UNE ÉQUATION STRUCTURELLE

Comme nous l'avons vu dans la section précédente, les MCO sont biaisés et non-convergents quand on les applique à une équation structurelle d'un modèle à équations simultanées. Dans le chapitre 15, nous avions vu que la méthode des doubles moindres carrés peut résoudre le problème de variable explicative endogène. Nous allons maintenant voir comment les DMC peuvent être appliqués aux MES.

Le fonctionnement des DMC est similaire à celui du chapitre 15. La principale différence est que, comme une équation structurelle est spécifiée pour chaque variable endogène, nous pouvons voir immédiatement s'il y a suffisamment de variables instrumentales pour estimer chaque équation. Commençons par discuter du problème d'identification.

Identification d'un système à deux équations

Nous avions mentionné la notion d'identification au chapitre 15. Quand nous estimons un modèle par les MCO, la condition d'identification principale est qu'aucune des variables explicatives ne doit être corrélée au terme d'erreur. Comme nous l'avons vu dans la section 16.2, cette condition fondamentale ne tient généralement plus pour les MES. Cependant, il est toujours possible d'identifier (c'est-à-dire d'estimer de manière convergente) les paramètres des équations d'un MES avec des variables instrumentales, comme dans le cas de l'erreur de mesure ou des biais de variables omises.

Avant de considérer un MES général à deux équations, focalisons-nous sur un exemple simple d'offre et de demande pour comprendre quelques intuitions. Écrivons le système à l'équilibre (c'est-à-dire en imposant $q_s = q_d = q$) :

$$q = \alpha_1 p + \beta_1 z_1 + u_1 \qquad \textbf{[16.15]}$$

et

$$q = \alpha_2 p + u_2. \qquad \textbf{[16.16]}$$

Pour illustrer, supposons que q est la consommation de lait par habitant dans le département, que p est le prix moyen du litre de lait dans le département, et que z_1 est le prix du fourrage pour nourrir les animaux, que nous supposerons exogènes dans les équations d'offre et de demande de lait. Cela veut dire que (16.15) doit être l'équation d'offre, le prix du fourrage affectant l'offre de lait ($\beta_1 < 0$) et non sa demande. La fonction de demande n'inclut pas de variable affectant seulement la demande.

Étant donné un échantillon contenant (q, p, z_1) et tiré aléatoirement, quelle(s) équation(s) peut-on estimer ? En d'autres termes, laquelle (ou lesquelles) est une **équation identifiée** ? Nous allons montrer que l'équation de *demande*, (16.16), est identifiée, alors que l'équation d'offre ne l'est pas. Nous pouvons facilement le voir en utilisant les mêmes règles que pour l'estimation par variables instrumentales du chapitre 15 : z_1 peut être utilisé comme variable instrumentale pour la variable de prix dans l'équation (16.16). En revanche, comme z_1 apparaît dans l'équation (16.15), il n'y a pas de variable instrumentale pour le prix dans l'équation d'offre.

Intuitivement, l'équation de demande est identifiée parce qu'il existe une variable observée, z_1, qui change l'offre sans affecter la demande. La figure 16.1 montre que, quand z_1 varie (et en supposant qu'il n'y a pas de terme d'erreur), on peut tracer la courbe de demande. Comme d'habitude, la présence d'un facteur de demande inobservé u_2 nous fait estimer l'équation de demande avec une erreur, mais les estimateurs seront convergents à partir du moment où z_1 n'est pas corrélé avec u_2.

L'équation d'offre ne peut pas être tracée sur le graphique car il n'y a pas de facteur exogène et observé déplaçant la courbe de demande. Les facteurs inobservés déplaçant la courbe de demande ne peuvent pas nous aider, il nous faut des facteurs observés. Si, de la même façon que dans la fonction de demande de travail (16.2), on pouvait observer un facteur de demande de lait exogène – par exemple le revenu moyen, qui peut influencer la demande de lait – alors la fonction d'offre de lait serait aussi identifiée.

En résumé : *Dans le système composé par (16.15) et (16.16), c'est la présence d'une variable exogène dans l'équation d'offre qui nous permet d'estimer l'équation de demande.*

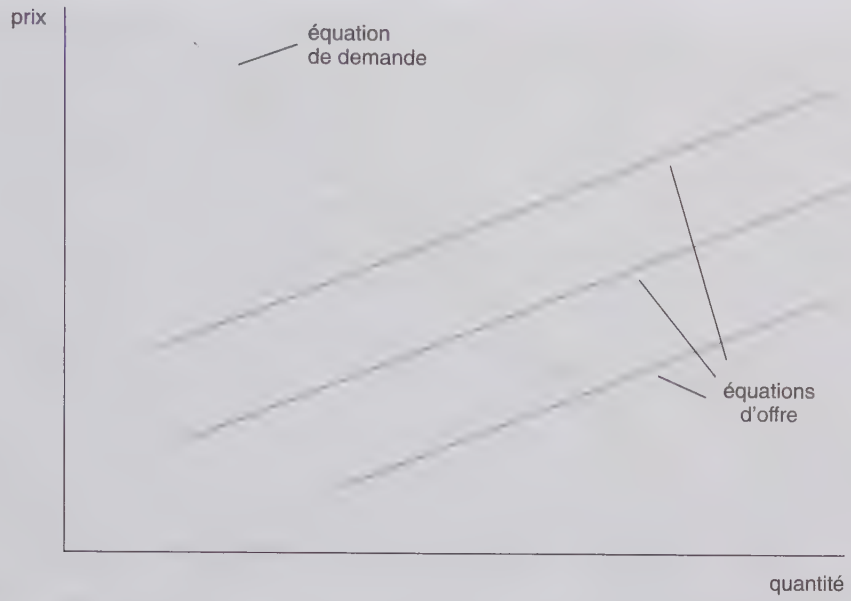


Figure 16.1 *Le déplacement des équations de courbe d'offre permet de tracer l'équation de la courbe de demande. Chaque équation de l'offre correspond à une valeur différente de la variable exogène z_1.*

Généraliser cette discussion sur l'identification à un modèle plus général à deux équations n'est pas difficile. Écrivons ces deux équations :

$$y_1 = \beta_{10} + \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\beta}_1 + u_1 \qquad \textbf{[16.17]}$$

et

$$y_2 = \beta_{20} + \alpha_2 y_1 + \mathbf{z}_2\boldsymbol{\beta}_2 + u_2, \qquad \textbf{[16.18]}$$

où y_1 et y_2 sont les variables endogènes, et u_1 et u_2 sont les termes d'erreurs structurels.

La constante de la première équation est β_{10}, et celle de la seconde équation est β_{20}. La variable z_1 contient un ensemble de k_1 variables exogènes apparaissant dans la première équation : $\mathbf{z}_1 = (z_{11}, z_{12}, \ldots, z_{1k_1})$. De la même manière, z_2 est l'ensemble de k_2 variables exogènes affectant la seconde équation : $\mathbf{z}_2 = (z_{21}, z_{22}, \ldots, z_{2k_2})$. Dans beaucoup de cas, il y aura des variables communes à z_1 et z_2. Pour alléger l'écriture, nous utiliserons la notation matricielle

$$\mathbf{z}_1\boldsymbol{\beta}_1 = \beta_{11}z_{11} + \beta_{12}z_{12} + \ldots + \beta_{1k_1}z_{1k_1}$$

et

$$\mathbf{z}_2\boldsymbol{\beta}_2 = \beta_{21}z_{21} + \beta_{22}z_{22} + \ldots + \beta_{2k_2}z_{2k_2}$$

c'est-à-dire : $\mathbf{z}_1\boldsymbol{\beta}_1$ regroupe toutes les variables exogènes de la première équation, chacune multipliée par son coefficient, de même pour $\mathbf{z}_2\boldsymbol{\beta}_2$. (Certains auteurs utilisent plutôt la notation $\mathbf{z}_1'\boldsymbol{\beta}_1$ et $\mathbf{z}_2'\boldsymbol{\beta}_2$. Si l'approche matricielle appliquée à l'économétrie vous intéresse, vous pouvez regarder l'Annexe E.)

Puisque z_1 et z_2 contiennent en général des variables exogènes différentes, cela veut donc dire que nous avons imposé des **restrictions d'exclusion** au modèle. En d'autres termes, nous *faisons l'hypothèse* que certaines variables exogènes n'apparaissent pas dans la première équation et que d'autres sont absentes

de la seconde équation. Comme nous l'avons vu dans l'exemple de l'offre et de la demande, cela permet de distinguer les deux équations structurelles.

Quand le système (16.17)-(16.18) peut-il être résolu pour trouver y_1 et y_2 (trouver y_1 et y_2 en fonction de toutes les variables exogènes et des erreurs structurelles u_1 et u_2) ? La condition est la même que dans (16.13), c'est-à-dire quand $\alpha_2\alpha_1 \neq 1$. La démonstration est quasiment la même que pour le modèle simple de la section 16.2. Sous cette hypothèse, une forme réduite existe pour y_1 et y_2.

La question clé est (par exemple) : sous quelles hypothèses peut-on estimer les paramètres de (16.17) ? C'est le problème de l'identification. Il est facile de trouver une **condition de rang** pour que l'équation (16.17) soit identifiée.

Condition de Rang pour l'Identification d'une Équation Structurelle. La première équation d'un modèle à deux équations simultanées est identifiée si et seulement si la *deuxième* équation contient au moins une variable exogène (avec un coefficient non-nul) exclue de la première équation.

C'est la condition nécessaire et suffisante pour que (16.17) soit identifiée. La **condition d'ordre**, que nous avons discutée au chapitre 15, est une condition nécessaire pour que la condition de rang soit vraie. La condition d'ordre, pour identifier la première équation, dit qu'au moins une variable exogène est exclue de cette équation. La vérifier est trivial, une fois que les deux équations ont été écrites. La condition de rang demande un peu plus : au moins une des variables exogènes exclues de la première équation doit avoir un coefficient estimé non-nul dans la seconde équation. Cela permet de s'assurer qu'au moins une des variables omises de la première équation apparaît dans l'équation de forme réduite prédisant y_2 et que nous pouvons donc utiliser cette variable comme instrument pour y_2. Cette condition peut être testée avec un test t ou un test F, comme dans le chapitre 15 ; et comme nous le montrons dans quelques exemples ci-dessous.

La condition d'identification de la seconde équation est, logiquement, strictement symétrique à la condition d'identification de la première équation. De plus, si nous écrivons les équations d'offre de travail et de demande de travail de la section 16.1 – de sorte que y_1 apparaisse dans le terme de gauche *des deux* équations, avec y_2 dans le terme de droite – la condition d'identification est identique.

EXEMPLE 16.3
Temps de travail des femmes mariées salariées

Pour illustrer le problème de l'identification, considérons l'offre de travail des femmes mariées salariées. En guise de fonction de demande, écrivons le salaire offert en fonction du nombre d'heures de travail et des variables de productivité habituelles. En imposant l'équilibre du marché, les équations structurelles sont

$$hours = \alpha_1 \log(wage) + \beta_{10} + \beta_{11} educ + \beta_{12} age + \beta_{13} kidslt6 + \beta_{14} nwifeinc + u_1 \qquad [16.19]$$

et

$$\log(wage) = \alpha_2 hours + \beta_{20} + \beta_{21} educ + \beta_{22} exper + \beta_{23} exper^2 + u_2 \qquad [16.20]$$

La variable *age* est l'âge de la femme, en années, *kidslt6* est le nombre d'enfants de moins de six ans, *nwifeinc* est le revenu ne provenant pas du salaire de la femme (ce qui inclut entre autres les revenus de son mari), et *educ* et *exper* sont respectivement le nombre d'années d'éducation et d'expérience. Toutes les variables, hormis *hours* et log(*wage*), sont supposées exogènes. (Il s'agit là d'une hypothèse discutable, car *educ* pourrait être corrélé à des capacités inobservées dans les deux équations. Mais pour se focaliser sur ce qui nous intéresse ici, nous ignorerons le problème des capacités inobservées.) La forme fonctionnelle de ce système – où *hours* est en niveau mais *wage* est en forme logarithmique – est populaire en économie du travail. Ce système pourrait très bien être décrit par les équations (16.17) et (16.18), en définissant $y_1 = hours$ et $y_2 = \log(wage)$.

La première équation est la fonction d'offre. Elle satisfait la condition d'ordre parce que deux variables exogènes, *exper* et $exper^2$, sont omises de l'équation d'offre de travail. Ces restrictions d'exclusion sont des hypothèses capitales : nous supposons que, à salaire, éducation, âge, nombre de jeunes enfants et autres sources de revenu donnés, l'expérience passée n'affecte pas l'offre de travail actuelle. C'est certainement discutable, mais tenons-nous en là dans cet exemple.

Étant donné les équations (16.19) et (16.20), la condition de rang pour identifier la première équation est : au moins un des coefficients d'*exper* et $exper^2$ doit être non-nul dans l'équation (16.20). Si $\beta_{22} = 0$ et $\beta_{23} = 0$, il n'y a plus de variable exogène apparaissant dans la seconde équation qui n'apparaisse pas dans la première (*educ* apparaît dans les deux équations). La condition de rang pour l'identification de (16.19) en termes de forme réduite prédisant log(*wage*) est équivalente, elle s'écrit

$$\log(wage) = \pi_{20} + \pi_{21}educ + \pi_{22}age + \pi_{23}kidslt6 + \pi_{24}nwifeinc + \pi_{25}exper + \pi_{26}exper^2 + v_2. \quad \textbf{[16.21]}$$

Pour l'identification, il faut que $\pi_{25} \neq 0$ ou $\pi_{26} \neq 0$, ce qui peut être testé en utilisant une statistique F standard, comme nous l'avons vu dans le chapitre 15.

L'équation de salaire, (16.20), est identifiée si au moins des coefficients de *age*, *kidslt6*, et *nwifeinc* est non-nul dans (16.19). Cela revient au même de supposer que la forme réduite prédisant *hours* – qui a le même terme de droite que (16.21) – dépend d'au moins une variable parmi *age*, *kidslt6*, et *nwifeinc*. En spécifiant l'équation de salaire, nous faisons l'*hypothèse* que *age*, *kidslt6*, et *nwifeinc* n'ont aucun effet sur le salaire obtenu, une fois que *hours*, *education*, et *experience* sont pris en compte. Cette hypothèse serait indésirable si une de ces variables affectait directement la productivité des femmes, ou si les femmes étaient discriminées sur le marché du travail en fonction de leur âge ou de leurs enfants.

Dans l'exemple 16.3, nous prenons comme population d'intérêt les femmes mariées salariées (donc le nombre d'heures travaillées est toujours strictement positif). On exclut donc ici les femmes mariées qui ne travaillent pas. Il est évident que nous ne pouvons pas observer le salaire d'une femme qui ne travaille pas. Nous aborderons ces sujets au chapitre 17 ; mais pour le moment, pensons aux équations (16.19) et (16.20) comme n'ayant un sens que pour les femmes pour lesquelles *hours* > 0.

EXEMPLE 16.4
Inflation et ouverture commerciale

Romer (1993) propose des modèles théoriques d'inflation qui prédisent que des pays plus « ouverts » commercialement auront des taux d'inflation plus faibles. Son analyse empirique prédit le taux d'inflation annuel moyen (depuis 1973) en fonction du rapport moyen entre importation et produit interieur (ou national) brut depuis 1973 – ce qui est sa mesure d'ouverture commerciale. En plus d'estimer son équation principale par les MCO, il utilise des variables instrumentales. Bien que Romer ne spécifie pas les deux équations dans un système à équations simultanées, il a en tête un système à deux équations :

$$inf = \beta_{10} + \alpha_1 open + \beta_{11}\log(pcinc) + u_1 \quad \textbf{[16.22]}$$

$$open = \beta_{20} + \alpha_2 inf + \beta_{21}\log(pcinc) + \beta_{22}\log(land) + u_2, \quad \textbf{[16.23]}$$

où *pcinc* est le revenu par habitant en 1980, en dollars américains (supposé exogène), et *land* est la superficie du pays, en miles carrés (également supposé exogène). L'équation (16.22) est l'équation d'intérêt, et on suppose que $\alpha_1 < 0$ (les économies plus ouvertes ont un taux d'inflation plus faible). La seconde équation vient du fait que l'ouverture commerciale peut également dépendre, entre autres, du taux d'inflation moyen. La variable log(*pcinc*) apparaît dans les deux équations, mais log(*land*) apparaît seulement dans la seconde équation. Il est en effet possible que, ceteris paribus, les petits pays soient plus ouverts commercialement (et donc $\beta_{22} < 0$).

En utilisant la règle d'identification donnée plus haut, l'équation (16.22) est identifiée si $\beta_{22} \neq 0$. L'équation (16.23) n'est *pas* identifiée car elle contient les deux variables exogène. Mais l'équation d'intérêt est (16.22).

Pour aller plus loin 16.2

Si nous observons la croissance de l'offre monétaire pour chaque pays depuis 1973, que nous supposerons exogène, cela peut-il aider à identifier l'équation (16.23) ?

Estimation par les DMC

Après avoir déterminé si une équation est identifiée, cette équation peut être estimée par les doubles moindres carrés. Les variables instrumentales sont les variables exogènes qui apparaissent dans une équation seulement.

EXEMPLE 16.5
Offre de travail des femmes mariées salariées

Nous utilisons les données sur les femmes mariées salariées contenues dans le fichier MROZ pour estimer l'équation d'offre de travail (16.19) par les DMC. La liste exhaustive des instruments est composée de : *educ*, *age*, *kidslt6*, *nwifeinc*, *exper*, et $exper^2$. La courbe d'offre de travail estimée est

$$\begin{aligned} \widehat{hours} = {} & 2\,225{,}66 + 1\,639{,}56 \log(wage) - 183{,}75\, educ \\ & \quad (574{,}56) \quad (470{,}58) \qquad\quad (59{,}10) \\ & - 7{,}81\, age - 198{,}15\, kidslt6 - 10{,}17\, nwifeinc \\ & \quad (9{,}38) \qquad (182{,}93) \qquad (6{,}61) \\ & \qquad\qquad n = 428, \end{aligned} \qquad \textbf{[16.24]}$$

ce qui montre que la courbe d'offre de travail est croissante. Le coefficient estimé de log(*wage*) s'interprète de cette manière : en gardant les autres variables constantes, $\widehat{hours} \approx 16{,}4(\%wage)$. Nous pouvons calculer l'élasticité de l'offre de travail en multipliant les deux parties de l'équation par 100/*hours* :

$$100.(\Delta\widehat{hours}/hours) \approx (1\,640/hours)(\%\Delta wage)$$

ou

$$\%\Delta\widehat{hours} \approx (1\,640/hours)(\%\Delta wage),$$

ce qui implique que l'élasticité de l'offre de travail (en fonction du salaire) est 1 640/*hours* [L'élasticité n'est pas constante dans ce modèle car la variable dépendante de (16.24) est *hours*, et non log(*hours*).]. Au temps de travail moyen de l'échantillon, 1 303 heures par an, l'élasticité estimée est 1 604/1 303 ≈ 1,26, ce qui veut dire que, pour une hausse de salaire de 1 %, l'augmentation du temps de travail est de plus de 1 %. Cette élasticité est donc relativement grande. Pour des temps de travail plus élevés, l'élasticité serait plus petite ; pour des temps de travail plus faibles, comme *hours* = 800, l'élasticité est supérieure à deux.

À titre de comparaison, quand (16.19) est estimée par les MCO, le coefficient de log(*wage*) est −2,05 (se = 54,88), ce qui veut dire que le salaire n'affecte pas le temps de travail. Pour confirmer l'endogénéité de log(*wage*) dans (16.19), on peut utiliser le test de la section 15.5. Si on ajoute les résidus de la forme réduite $\hat{v}_2$ à l'équation et qu'on estime l'équation par les MCO, la statistique *t* de $\hat{v}_2$ est −6,61, ce qui est très significatif. On en conclut donc que log(*wage*) est endogène.

L'équation de salaire (16.20) peut également être estimée par les DMC. Cela donne

$$\widehat{\log(wage)} = -0{,}656 + 0{,}00013\ hours + 0{,}110\ educ$$
$$(0{,}338)\ (0{,}00025)\qquad (0{,}016)$$
$$+ 0{,}035\ exper - 0{,}00071\ exper^2$$
$$(0{,}019)\qquad (0{,}00045)$$
$$n = 428. \qquad \textbf{[16.25]}$$

La différence avec les précédentes estimations d'équation de salaire est que *hours* est une variable explicative, et que les DMC permettent de prendre en compte l'endogénéité de *hours* (et nous supposons que *educ* et *exper* sont exogènes). Le coefficient de *hours* n'est pas statistiquement significatif, ce qui veut dire qu'on ne peut pas prouver que le salaire offert augmente la durée du temps de travail. Les autres coefficients sont similaires aux coefficients obtenus par les MCO en supprimant *hours* de l'équation.

Les variables instrumentales peuvent également être directement appliquées à l'estimation des effets de l'ouverture commerciale sur l'inflation.

EXEMPLE 16.6
Inflation et ouverture commerciale

Avant d'estimer (16.22) avec les données contenues dans OPENNESS, vérifions si *open* a une corrélation partielle suffisante avec l'instrument proposé, log(*land*). La régression de forme réduite est

$$\widehat{open} = 117{,}08 + 0{,}546\ \log(pcinc) - 7{,}57\ \log(land)$$
$$(15{,}85)\ (1{,}493)\qquad (0{,}81)$$
$$n = 114,\ R^2 = 0{,}449.$$

La statistique *t* de log(*land*) est supérieure à 9 en valeur absolue, ce qui vérifie l'hypothèse de Romer selon laquelle les petits pays sont plus ouverts au commerce. Il n'est pas pertinent d'interpréter le fait que log(*pcinc*) soit si peu significatif dans cette régression.

L'estimation de (16.22) en utilisant log(*land*) comme variable instrumentale pour *open* donne

$$\widehat{inf} = 26{,}90 - 0{,}337\ open + 0{,}376\ \log(pcinc)$$
$$(15{,}40)\ (0{,}144)\qquad (2{,}015)$$
$$n = 114. \qquad \textbf{[16.26]}$$

Le coefficient d'*open* est statistiquement significatif, approximativement au seuil de 1 % contre l'alternative unilatérale ($\alpha_1 < 0$). L'effet est également économiquement important : pour chaque point de pourcentage d'augmentation du rapport importations sur PIB, le taux d'inflation annuel moyen diminue en moyenne d'un tiers de point de pourcentage. Par comparaison, l'estimation par les MCO donne −0,215 (se = 0,095).

Pour aller plus loin 16.3

Comment pourrait-on tester si les estimations par MCO et par variables instrumentales de l'effet d'*open* sont statistiquement différentes ?

16.4 SYSTÈMES AVEC PLUS DE DEUX ÉQUATIONS

Les modèles à équations simultanées peuvent contenir plus de deux équations. L'étude de l'identification de ces modèles dans le cas général est difficile, et cela nécessite de l'algèbre matricielle. Cependant, une fois qu'on a prouvé qu'une équation d'un système est identifiée dans le cas général, on peut estimer cette équation par les DMC.

Identification dans les systèmes avec trois équations ou plus

Nous utiliserons un système à trois équations pour illustrer les problèmes qui se posent pour l'identification de MES compliqués. En supprimant les constantes, le modèle s'écrit

$$y_1 = \alpha_{12}y_2 + \alpha_{13}y_3 + \beta_{11}z_1 + u_1 \quad \textbf{[16.27]}$$

$$y_2 = \alpha_{21}y_1 + \beta_{21}z_1 + \beta_{22}z_2 + \beta_{23}z_3 + u_2 \quad \textbf{[16.28]}$$

$$y_3 = \alpha_{32}y_2 + \beta_{31}z_1 + \beta_{32}z_2 + \beta_{33}z_3 + \beta_{34}z_4 + u_3, \quad \textbf{[16.29]}$$

où les y_g sont les variables endogènes, et les z_j sont exogènes. Le premier indice sur les paramètres indique l'équation, le second indique la variable ; nous utiliserons α pour les coefficients des variables endogènes, et β pour les coefficients des variables exogènes.

Laquelle (ou lesquelles) de ces équations peut être estimée ? Il est difficile en général de montrer qu'une équation d'un MES avec plus de deux équations est identifiée, mais il est facile de voir que certaines équations ne sont *pas* identifiées. Dans le système (16.27) à (16.29), nous pouvons facilement voir que (16.29) entre dans cette catégorie. Comme toutes les variables exogènes entrent dans cette équation, il n'y a pas de variable instrumentale pour y_2. Nous ne pouvons donc pas estimer de manière convergente les paramètres de cette équation. Comme discuté en section 16.2, l'estimation par les MCO de cette équation ne sera généralement pas convergente.

Et l'équation (16.27) ? Cela semblerait fonctionner, puisque z_2, z_3, et z_4 sont toutes exclues de l'équation – ce qui est un autre exemple de *restriction d'exclusion*. Bien qu'il y ait deux variables endogènes dans cette équation, nous avons trois variables instrumentales potentielles pour y_2 et y_3. L'équation (16.27) satisfait donc la condition d'ordre. Nous allons définir de manière générale la condition d'ordre pour des MES quelconques.

Condition d'ordre pour l'identification. Une équation d'un MES quelconque satisfait la condition d'ordre si le nombre de variables exogènes *exclues* de l'équation est au moins aussi grand que le nombre de variables endogènes dans le terme de droite.

La seconde équation, (16.28), satisfait aussi la condition d'ordre puisqu'il y a une variable exogène exclue, z_4, et une variable endogène dans le terme de droite, y_1.

Comme nous l'avons discuté dans le chapitre 15 et dans la section précédente, la condition d'ordre est seulement nécessaire, et non suffisante, pour l'identification. Par exemple, si $\beta_{34} = 0$, z_4 n'apparaît pas dans le système, ce qui veut dire qu'elle n'est corrélée ni avec y_1, ni avec y_2, ni avec y_3. Si $\beta_{34} = 0$, la seconde équation n'est pas identifiée, puisque z_4 ne peut servir de variable instrumentale pour y_1. Cela illustre encore le fait que l'identification d'une équation dépend des valeurs des paramètres dans les autres équations (dont nous ne pouvons jamais être sûrs).

L'identification d'un MES peut échouer pour de nombreuses raisons souvent compliquées. Pour obtenir des conditions suffisantes, il faut étendre la condition de rang des systèmes à deux équations. C'est possible, mais cela demande de l'algèbre matricielle [voir, par exemple, Wooldridge (2010, chapitre 9)].

Dans beaucoup d'applications, on suppose que, sauf si une équation n'est manifestement pas identifiée, une équation qui satisfait la condition de rang est identifiée.

La terminologie des équations suridentifiées et juste identifiées présentée dans le chapitre 15 vient initialement des MES. Au vu de la condition d'ordre, (16.27) est une **équation suridentifiée**, puisque nous n'avons besoin que de deux variables instrumentales (pour y_2 et pour y_3) alors qu'il y en a trois (z_2, z_3, et z_4) ; il y a donc une restriction suridentifiante dans cette équation. D'une manière générale, le nombre de restrictions suridentifiantes est la différence entre le nombre variables exogènes dans le système et le nombre de variables dans l'équation. Ces restrictions suridentifiantes peuvent être testées en utilisant le test de suridentification de la section 15.5. L'équation (16.28) est une équation **juste identifiée**, et la troisième équation est une **équation non-identifiée**.

Estimation

Indépendamment du nombre d'équations d'un MES, chaque équation identifiée peut être estimée par les DMC. Les instruments d'une équation sont les variables exogènes qui apparaissent dans les autres équations du système. Les tests d'endogénéité, d'hétéroscédasticité, de corrélation sérielle, et de restriction suridentifiante sont donc les mêmes qu'au chapitre 15.

Il se trouve que, quand un système de deux équations ou plus est correctement spécifié et sous quelques hypothèses supplémentaires, les *méthodes d'estimation de système* sont généralement plus efficaces que l'estimation de chaque équation par les DMC. Les *triples moindres* carrés constituent la méthode d'estimation de système la plus utilisée dans le contexte des MES. Ces méthodes, avec ou sans variable explicative endogène, ne sont pas couvertes dans ce livre [Voir, par exemple, Wooldridge (2010, chapitres 7 et 8).]

16.5 MODÈLES À ÉQUATIONS SIMULTANÉES ET SÉRIES TEMPORELLES

Les gros systèmes d'équations simultanées utilisés pour décrire l'économie d'un pays font partie des applications les plus anciennes des MES. écrivons un modèle Keynésien simple de demande agrégée (qui ignore exportations et importations) :

$$C_t = \beta_0 + \beta_1(Y_t - T_t) + \beta_2 r_t + u_{t1} \quad [16.30]$$

$$I_t = \gamma_0 + \gamma_1 r_t + u_{t2} \quad [16.31]$$

$$Y_t \equiv C_t + I_t + G_t, \quad [16.32]$$

où

C_t = consommation

Y_t = revenu

T_t = recettes fiscales

r_t = taux d'intérêt

I_t = investissement

G_t = dépenses publiques

[Voir, par exemple, Mankiw (1994, chapitre 9).] Pour illustrer, supposons que t représente une année.

La première équation est une fonction de consommation agrégée, où la consommation dépend du revenu disponible, du taux d'intérêt, et d'une erreur structurelle inobservée u_{t1}. La seconde équation est une fonction d'investissement très simple. L'équation (16.32) est une *équation comptable*, c'est le résultat de la comptabilité nationale des revenus : elle tient par définition, sans erreur. Il n'y a donc aucun sens à estimer (16.32), mais nous avons besoin de cette équation pour boucler le modèle.

Puisque ce système a trois équations, il y a trois variables endogènes. Au vu des deux premières équations, il est évident que C_t et I_t seront endogènes. De plus, Y_t est endogène du fait de l'équation comptable. Nous supposerons, au moins dans ce modèle, que T_t, r_t, et G_t sont exogènes, c'est-à-dire qu'ils ne sont corrélés ni à u_{t1} ni à u_{t2} (nous discuterons des problèmes associés à ce type d'hypothèses plus loin).

Si r_t est exogène, il semble logique d'estimer l'équation (16.31) par les MCO. La fonction de consommation, en revanche, dépend du revenu disponible, qui est endogène, puisque Y_t l'est. Nous avons deux instruments disponibles sous les hypothèses d'exogénéité faites ici : T_t et G_t. Si nous suivons nos préconisations pour l'estimation d'équations sur données transversales, nous estimerions (16.30) par les DMC avec les instruments (T_t, G_t, r_t).

Les modèles semblables à celui des équations (16.30) à (16.32) sont rarement estimés aujourd'hui, pour plusieurs bonnes raisons. Premièrement, il est très difficile de justifier que, au niveau agrégé, les impôts, les taux d'intérêt, et les dépenses publiques sont exogènes. Il est clair que les impôts dépendent directement du revenu. Par exemple, avec un taux marginal d'imposition unique τ_t l'année t, $T_t = \tau_t Y_t$. Nous pouvons facilement remplacer $(Y_t - T_t)$ par $(1 - \tau_t)Y_t$ dans (16.30) et nous pourrons toujours estimer l'équation par les DMC en supposant que les dépenses publiques sont exogènes. Nous pourrions aussi ajouter le taux d'imposition à la liste des instruments, s'il est exogène. Mais les dépenses publiques et le taux d'imposition sont-ils exogènes ? Ils peuvent probablement l'être en théorie, si le gouvernement fixe le taux d'imposition et les dépenses indépendamment de ce qu'il se passe dans l'économie. Toutefois, il est difficile d'en être sûr : les dépenses publiques dépendent généralement du revenu national et pour de hauts niveaux de revenu national, il suffit de taux d'imposition plus faibles pour assurer les mêmes recettes fiscales totales. Par ailleurs, supposer que les taux d'intérêt sont exogènes est très discutable. Il serait possible d'écrire un modèle plus réaliste avec une offre et une demande de monnaie, et les taux d'intérêts pourraient êtres estimés de manière jointe avec C_t, I_t, et Y_t. Mais alors trouver suffisamment de variables exogènes pour identifier toutes les équations devient très difficile (et les problèmes de ces modèles mentionnés ci-dessous sont toujours pertinents).

Certains ont avancé que certaines dépenses publiques, par exemple les dépenses militaires, sont exogènes pour une variété d'applications à équations simultanées – voir, par exemple, Hall (1998) et Ramey (1991). D'autres ne croient pas à ces hypothèses et de toutes façons, la dépense militaire n'est pas toujours corrélée aux variables endogènes comme cela est nécessaire [voir Shea (1991) pour une discussion et l'exercice sur ordinateur C6 pour un exemple].

Un modèle comme celui présenté en (16.30) à (16.32) présente un second problème : il est complètement statique. C'est particulièrement vrai avec des données mensuelles ou trimestrielles, mais c'est aussi vrai avec des données annuelles, car certaines variables mettent du temps à s'ajuster. (Un argument en faveur des modèles statiques du type Keynésien est qu'ils visent à décrire le long terme sans se préoccuper des dynamiques de court terme.) Ajouter des éléments dynamiques dans le modèle n'est pas très difficile. Par exemple, nous pourrions ajouter le revenu retardé (c'est-à-dire le revenu de la période précédente) dans l'équation (16.31) :

$$I_t = \gamma_0 + \gamma_1 r_t + \gamma_2 Y_{t-1} + u_{t2}. \qquad \textbf{[16.33]}$$

En d'autres termes, nous ajoutons une **variable endogène** retardée (mais pas I_{t-1}) à l'équation d'investissement. Pouvons-nous considérer qu'Y_{t-1} est exogène dans cette équation ? Sous certaines hypothèses sur

u_{t2}, la réponse est oui. Une variable endogène retardée dans un MES est généralement appelée **variable prédéterminée**. Les retards de variables exogènes sont aussi prédéterminés. Si nous supposons que u_{t2} n'est corrélé ni avec les variables exogènes à date t (ce qui est standard), ni avec toutes les variables endogènes et exogènes *passées*, alors Y_{t-1} n'est pas corrélé avec u_{t2}. Puisque r_t est exogène, nous pouvons estimer (16.33) avec les MCO.

Si nous ajoutons la consommation retardée à (16.30), nous pouvons considérer que C_{t-1} est exogène dans cette équation, mais il faut pour cela faire les mêmes hypothèses sur u_{t1} que nous avons faites pour u_{t2} dans le paragraphe précédent. Le revenu disponible courant est toujours endogène dans

$$C_t = \beta_0 + \beta_1(Y_t - T_t) + \beta_2 r_t + \beta_3 C_{t-1} + u_{t1}, \qquad \textbf{[16.34]}$$

donc nous pourrions estimer cette équation par les DMC en utilisant les instruments (T_t, G_t, r_t, C_{t-1}). Si l'investissement est déterminé par (16.33), Y_{t-1} doit être inclus dans la liste des instruments. [Pour comprendre pourquoi, écrire la forme réduite de Y_t à partir des variables exogènes et prédéterminées – T_t, r_t, G_t, C_{t-1}, et Y_{t-1} – en utilisant (16.32), (16.33) et (16.34). Du fait que Y_{t-1} apparaît dans cette forme réduite, il doit être utilisé comme variable instrumentale.]

La présence d'une dynamique dans les MES agrégés est clairement une amélioration par rapport aux MES statiques, au moins dans des buts de prévision économiques. Mais il reste d'importants problèmes pour estimer des MES en utilisant des données de séries temporelles agrégées, certains d'entre eux ont été discutés dans les chapitres 11 et 15. Rappelons-nous que la validité des procédures d'inférence habituelles, MCO et DMC, tient à la notion de *dépendance faible*. Malheureusement, des séries comme la consommation agrégée, le revenu, l'investissement, et même les taux d'intérêts, semblent violer l'hypothèse de dépendance faible (Dans la terminologie du chapitre 11, elles ont des *racines unitaires*.). Ces séries ont aussi tendance à avoir une tendance exponentielle, même si cela peut être partiellement contourné en utilisant une transformation logarithmique et en changeant les formes fonctionnelles. En général, sans même parler de petits échantillons, les propriétés des MCO et DMC sur grands échantillons sont compliquées et dépendent de nombreuses hypothèses quand elles sont appliquées à des équations avec des variables I(1). Nous en parlerons brièvement au chapitre 18. Un traitement détaillé et général du problème se trouve dans Hamilton (1994).

Cette discussion veut-elle dire qu'il est vain d'appliquer les MES à des données de séries temporelles ? Pas du tout. Les problèmes de tendance et de persistance élevée peuvent être traités en spécifiant des systèmes en différences premières ou en taux de croissance. Mais il faut prendre en compte le fait que le MES en différences premières est différent de celui en niveaux [Par exemple, si la croissance de la consommation est spécifiée comme une fonction de la croissance du revenu disponible et de la variation des taux d'intérêts, c'est un modèle différent de (16.30)]. De plus, comme nous en avons discuté, ajouter une dynamique dans un MES n'est pas particulièrement difficile. Enfin, il est souvent un peu moins difficile de trouver des variables vraiment exogènes à mettre dans un MES avec des données désagrégées. Par exemple, pour les industries manufacturières, Shea (1993) montre que la production dans les autres industries (ou plus précisément la croissance de leur production) peut être utilisée comme instrument dans le but d'estimer des fonctions d'offre. Ramsey (1991) a également une analyse convaincante de la manière d'estimer des fonctions de coûts industrielles avec des variables instrumentales sur données de séries temporelles.

L'exemple suivant montre comment il est possible de mobiliser des données agrégées pour tester la théorie économique selon laquelle la consommation dépend uniquement du revenu permanent, généralement appelée *hypothèse du revenu permanent* (HRP). L'approche utilisée dans cet exemple n'est pas strictement celle des modèles à équations simultanées, mais il est néanmoins possible de penser que les croissances du revenu et de la consommation (ainsi que les taux d'intérêts) sont déterminées de manière jointe.

EXEMPLE 16.7
Test de l'hypothèse du revenu permanent

Campbell et Mankiw (1990) ont utilisé une méthode de variables instrumentales pour tester différentes formes de l'hypothèse du revenu permanent. Nous utiliserons des données annuelles de 1959 à 1995 dans CONSMUP pour répliquer une de leurs analyses. Campbell et Mankiw utilisent des données trimestrielles se terminant en 1985.

Une des équations estimées par Campbell et Mankiw est (en utilisant notre notation) :

$$gc_t = \beta_0 + \beta_1 gy_t + \beta_2 r3_t + u_t, \qquad \textbf{[16.35]}$$

où

$gc_t = \Delta\log(c_t)$ = croissance annuelle de la consommation par habitant en termes réels (en excluant les biens durables).
gy_t = croissance du revenu disponible réel.
$r3_t$ = Le taux d'intérêt réel (ex-post) mesuré par le rendement des bonds du trésor américain à trois mois : $r3_t = i3_t - inf_t$, le taux d'inflation étant basé sur l'Indice des Prix à la Consommation.

Les taux de croissance de la consommation et du revenu disponible n'ont pas de tendance, et ils sont faiblement dépendants ; nous supposerons que c'est également le cas de $r3_t$, de sorte que la théorie asymptotique standard puisse s'appliquer.

La caractéristique clé de l'équation (16.35) est que l'HRP implique que le terme d'erreur u_t a une moyenne nulle conditionnellement aux informations disponibles à la date $t - 1$ ou plus tôt : $E(u_t|I_{t-1}) = 0$. Cependant, u_t n'est *pas* nécessairement non-corrélé à gy_t ou $r3_t$. Une manière traditionnelle de voir les choses est que ces variables sont déterminées de manière jointe, mais nous n'écrirons pas le système à trois équations en entier.

Comme u_t n'est pas corrélé avec les variables de la date $t - 1$ ou antérieures, les valeurs retardées de gc, gy et $r3$ sont des instruments valides pour estimer (16.35) (c'est aussi le cas des autres variables observables retardées, mais nous ne les utiliserons pas ici). Quelles sont les hypothèses d'intérêt ? La forme pure de l'HRP indique que : $\beta_1 = \beta_2 = 0$. Campbell et Mankiw expliquent que β_1 est positif si une part de la population consomme son revenu courant au lieu de son revenu permanent. L'HRP avec un taux d'intérêt réel variable implique que $\beta_2 > 0$.

L'estimation de (16.35) par les DMC, en utilisant gc_{t-1}, gy_{t-1}, et $r3_{t-1}$ comme instruments pour les variables endogènes gy_t et $r3_t$, donne

$$\widehat{gc}_t = 0{,}0081 + 0{,}586\ gy_t - 0{,}00027\ r3_t$$
$$(0{,}0032)\ (0{,}135)\qquad (0{,}00076)$$
$$n = 35,\ R^2 = 0{,}678. \qquad \textbf{[16.36]}$$

En conséquence, la forme pure de l'HRP est fortement rejetée, puisque le coefficient de gy est grand d'un point de vue économique (une hausse du revenu disponible de 1 % augmenterait la consommation de plus de 0,5 %) et statistiquement significatif ($t = 4{,}34$). En revanche, le coefficient du taux d'intérêt réel est très petit et statistiquement non significatif. Ces résultats sont qualitativement similaires à ceux de Campbell et Mankiw.

L'HRP implique aussi que les erreurs $\{u_t\}$ n'ont pas de corrélation sérielle. Après l'estimation par les DMC, nous obtenons les résidus $\hat{u}_t$, et incluons $\hat{u}_{t-1}$ comme variable explicative dans (16.36) ; nous utilisons toujours les instruments gc_{t-1}, gy_{t-1}, $r3_{t-1}$, et $\hat{u}_{t-1}$ est son propre instrument (voir section 15.7). Le coefficient d'$\hat{u}_{t-1}$ est $\hat{\rho} = 0{,}187$ (écart-type estimé = 0,133), il y a donc des signes de corrélation sérielle, bien que ce ne soit pas significatif au seuil de 5 %. Campbell et Mankiw expliquent pourquoi, avec les données trimestrielles, une corrélation sérielle positive entre les termes d'erreur peut exister même si l'HRP est vraie. Certains de ces arguments tiennent également avec des données annuelles.

Pour aller plus loin 16.4

Supposons que vous disposiez, pour une ville en particulier, de données sur la consommation par habitant de poisson, le revenu total par habitant, le prix du poisson, les prix du poulet et du bœuf. Les revenus, les prix du poulet et du bœuf sont considérés comme exogènes. Supposons qu'il n'y ait pas de saisonnalité dans la demande de poisson, mais qu'il y en ait dans l'offre de poisson. Comment utiliser cette information pour mesurer une équation avec élasticité constante de la demande de poisson ? Écrivez l'équation et discutez de son identification. (*Indice* : Vous devez avoir 11 variables instrumentales pour le prix du poisson.)

Utiliser des taux de croissance ou des variables I(1) est relativement commun dans les applications sur séries temporelles. Par exemple, Shea (1993) estime des courbes d'offres de branches industrielles et sa spécification est basée sur des taux de croissance.

Si un modèle structurel contient une tendance – qui pourrait capturer des facteurs exogènes ayant une tendance mais qui ne sont pas directement modélisés – alors cette tendance est sa propre variable instrumentale.

16.6 MODÈLES À ÉQUATIONS SIMULTANÉES SUR DONNÉES DE PANEL

Les modèles à équations simultanées peuvent aussi être utilisés sur des données de panel. Par exemple, on peut vouloir estimer des équations d'offre de travail et de salaire, comme dans l'exemple 16.3, pour un groupe de personnes travaillant pendant un certain laps de temps. En plus de permettre la détermination simultanée des variables à chaque date, nous pouvons alors avoir des effets inobservés dans chaque équation. Dans l'équation d'offre de travail, il serait par exemple utile de pouvoir prendre en compte des différences de préférences pour le loisir entre individus supposées invariantes au cours du temps.

Pour estimer un MES avec des données de panel, l'approche se fait en deux étapes : (1) éliminer les effets inobservés des équations d'intérêt en utilisant des effets fixes ou des différences premières, puis (2) trouver des variables instrumentales pour les variables restant exogènes dans l'équation transformée. Cela peut être particulièrement difficile : pour une analyse convaincante, il faut trouver des instruments qui changent au cours du temps. Pour le comprendre, écrivons un MES sur données de panel :

$$y_{it1} = \alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\beta}_1 + a_{i1} + u_{it1} \qquad [16.37]$$

$$y_{it2} = \alpha_2 y_{it1} + \mathbf{z}_{it2}\boldsymbol{\beta}_2 + a_{i2} + u_{it2}, \qquad [16.38]$$

où i est l'identifiant individuel, t est la date, et $\mathbf{z}_{it1}\boldsymbol{\beta}_1$ ou $\mathbf{z}_{it2}\boldsymbol{\beta}_2$ sont des fonctions linéaires d'un ensemble de variables exogènes dans chaque équation. Dans le cas le plus général, il est possible de supposer que les effets inobservés a_{i1} et a_{i2} sont corrélés avec *toutes* les variables explicatives, même les éléments de z. Cependant, il faut supposer que les erreurs structurelles idiosyncratiques, u_{it1} et u_{it2}, ne sont corrélées avec $\mathbf{z}$ dans aucune des équations à aucune des dates, c'est dans ce sens que les $\mathbf{z}$ doivent être exogènes. Sauf dans des cas particuliers, y_{it2} est corrélé à u_{it1}, et y_{it1} est corrélé à u_{it2}.

Supposons que nous sommes intéressés par l'équation (16.37). Nous ne pouvons pas l'estimer par les MCO, puisque l'erreur composite $a_{i1} + u_{it1}$ est potentiellement corrélée à toutes les variables explicatives. Écrivons les équations en différence premières pour supprimer l'effet inobservé a_{i1} :

$$\Delta y_{it1} = \alpha_1 \Delta y_{it2} + \Delta \mathbf{z}_{it1}\boldsymbol{\beta}_1 + \Delta u_{it1}. \qquad [16.39]$$

(Comme d'habitude avec les premières différences ou les données dé-moyennées, il est seulement possible d'estimer l'effet des variables qui changent au cours du temps pour au moins une des unités de la coupe.) Le terme d'erreur de cette équation n'est plus corrélé avec Δz_{it1} par hypothèse. Mais Δy_{it2} et Δu_{it1} peuvent être corrélés. Il faut donc trouver un instrument pour Δy_{it2}.

Comme dans le cas des données uniquement en coupe ou des séries temporelles pures, les variables instrumentales potentielles sont dans l'*autre* équation : des éléments de $\mathbf{z}_{it2}$ qui ne sont pas dans $\mathbf{z}_{it1}$. En pratique, nous avons besoin d'éléments de $\mathbf{z}_{it2}$ qui *varient au cours du temps* et ne sont pas dans $\mathbf{z}_{it1}$. Nous avons en effet besoin d'un instrument pour Δy_{it2} et le changement d'une variable d'une date à l'autre a peu de chances d'être corrélé au *niveau* d'une variable exogène. En effet, si on calcule la première différence de (16.38), on voit que les instruments naturels pour Δy_{it2} sont les éléments de Δz_{it2} qui ne sont pas aussi dans Δz_{it1}.

Afin d'illustrer les problèmes qui peuvent se poser, considérons une version en panel de la fonction d'offre de travail de l'exemple 16.3. Supposons que l'on trouve l'équation suivante en différences premières :

$$\Delta hours_{it} = \beta_0 + \alpha_1 \Delta \log(wage_{it}) + \Delta(autres\ facteurs_{it}),$$

et nous aimerions utiliser $\Delta exper_{it}$ comme instrument pour $\Delta \log(wage_{it})$. Le problème est que, comme nous nous intéressons aux personnes qui travaillent à chaque date, $\Delta exper_{it} = 1$ quels que soient i et t (Chaque personne gagne une année d'expérience par année). Nous ne pouvons pas avoir d'instrument qui a la même valeur quels que soient i et t, donc nous devons trouver quelque chose d'autre.

La participation à un programme expérimental peut souvent être utilisée pour trouver des instruments dans un contexte de panel. Dans l'exemple 15.10, nous avions utilisé l'obtention de bourses à la formation professionnelle comme instrument pour l'évolution du nombre d'heures de formation professionnelle, dans l'objectif de mesurer les effets de la formation professionnelle sur la productivité des travailleurs. Nous pourrions utiliser ceci dans le contexte d'un MES : la formation professionnelle et la productivité des travailleurs sont déterminées de manière jointe, mais recevoir une bourse pour la formation professionnelle est exogène dans l'équation (15.57).

Il est parfois possible de trouver des stratégies de variables instrumentales astucieuses et convaincantes dans des applications sur données de panel, comme l'illustre l'exemple suivant.

EXEMPLE 16.8
Effet de la population carcérale sur les taux de criminalité

Afin d'estimer l'effet causal des hausses de la population carcérale au niveau des États américains, Levitt (1996) utilise les jugements sur la surpopulation carcérale pour instrumenter la croissance de la population des prisons. L'équation que Levitt estime est en différences premières, nous pouvons écrire le modèle à effets fixes sous-jacent :

$$\log(crime_{it}) = \theta_t + \alpha_1 \log(prison_{it}) + \mathbf{z}_{it1}\boldsymbol{\beta}_1 + a_{i1} + u_{it1}, \qquad [16.40]$$

où θ_t représente les constantes à chaque date, et *crime* et *prison* sont mesurés pour 100 000 habitants (la variable de population carcérale est mesurée au dernier jour de l'année précédente). Le vecteur contient le log du nombre de policiers par habitants, le log du revenu par habitant, le taux de chômage, la proportion de personnes noires et celle de personnes vivant dans une aire urbaine (*metropolitan area*), et les proportions de chaque groupe d'âge dans la population.

En différenciant (16.40), on obtient l'équation estimée par Levitt :

$$\Delta \log(crime_{it}) = \xi_t + \alpha_1 \Delta \log(prison_{it}) + \Delta \mathbf{z}_{it1}\boldsymbol{\beta}_1 + \Delta u_{it1}. \qquad [16.41]$$

La simultanéité entre le taux de criminalité et la population carcérale, ou plus précisément entre leurs taux de croissance, rend généralement l'estimation par les MCO de (16.41) non-convergente. En utilisant le taux de criminalité violente et un sous-échantillon des données de Levitt (dans PRISON, pour les années de 1980 à 1993, pour $51 \times 14 = 714$ observations au total), l'estimateur des MCO d'α_1 obtenu sur données empilées est $-0{,}171$ (écart-type estimé = 0,048). Nous pouvons également estimer (16.41) par les DMC, en utilisant deux variables binaires comme instrument pour $\Delta \log(prison)$: si une décision juridique sur la surpopulation carcérale a été rendue l'année précédente et si cela a été le cas dans les deux années précédentes. L'estimateur des DMC de α_1 sur données empilées est $-1{,}032$ (écart-type estimé = 0,370). L'effet estimé par les DMC est donc bien plus grand et sans surprise, il est aussi bien moins précis. Levitt trouve des résultats similaires en utilisant des données sur une durée plus longue (mais avec des observations manquantes pour certains États) et plus d'instruments.

Tester la corrélation sérielle dans $r_{it1} = \Delta u_{it1}$ est facile. Après l'estimation des DMC sur données empilées, on peut en calculer les résidus, $\hat{r}_{it1}$. Ensuite, il faut ajouter ces résidus retardés dans l'équation d'intérêt, estimer cette équation par les DMC, en utilisant $\hat{r}_{it1}$ comme instrument pour lui-même. La première année de données est perdue à cause de la variable retardée. Ensuite, la statistique t habituelle des DMC sur les résidus retardés est un test valide pour la corrélation sérielle. Dans l'exemple 16.8, le coefficient de $\hat{r}_{it1}$ est de seulement environ 0,076, avec $t = 1{,}67$. Avec un coefficient si petit et une statistique t si faible, on peut supposer sans risque qu'il n'y a pas de corrélation sérielle.

Une approche alternative pour estimer les MES avec des données de panel est d'utiliser la transformation à effets fixes et ensuite, d'utiliser une technique de variables instrumentales comme dans le cas des DMC sur données empilées. Une procédure simple est d'estimer l'équation dé-moyennée par DMC sur données empilées, ce qui ressemblerait à

$$\ddot{y}_{it1} = \alpha_1 \ddot{y}_{it2} + \ddot{\mathbf{z}}_{it1}\boldsymbol{\beta}_1 + \ddot{u}_{it1}, \; t = 1, 2, \ldots, T, \qquad \textbf{[16.42]}$$

où $\ddot{\mathbf{z}}_{it1}$ et $\ddot{\mathbf{z}}_{it2}$ sont des variables instrumentales. C'est équivalent à l'utilisation des DMC dans leur formulation avec des variables indicatrices pour l'unité d'observation, où les variables indicatrices pour l'unité d'observation seraient leurs propres instruments. Ayres et Levitt (1998) ont appliqué les DMC sur des données dé-moyennées pour estimer l'effet de l'outil de prévention électronique des vols « LoJack » sur les pourcentages de voitures volées dans les villes. Si (16.42) est estimée directement, le nombre de degrés de liberté doit être corrigé : c'est $N(T-1) - k_1$, où k_1 est le nombre d'éléments dans α_1 et $\boldsymbol{\beta}_1$. En incluant des variables indicatrices pour l'unité d'observation et en utilisant les DMC sur données empilées avec les données de départ, on trouve le bon nombre de degrés de liberté. Un traitement détaillé des DMC sur données de panel est donné dans Wooldridge (2010, chapitre 11).

RÉSUMÉ

L'utilisation de modèles à équations simultanées est utile quand chaque équation du système a une interprétation *ceteris paribus*. Les cas où les équations décrivent l'offre et la demande sur un marché sont de bons exemples, comme ceux où les équations décrivent le comportement économique de différents agents. Les cas les plus fréquents sont ceux de l'offre et de la demande, mais il y a beaucoup d'autres usages des MES en économie et dans les sciences sociales.

Une caractéristique importante des MES est que, puisque tout le système est spécifié, on voit clairement quelles variables sont supposées exogènes et quelles variables apparaissent dans chaque équation. En prenant un système complet, nous pouvons déterminer quelles équations sont identifiables (c'est-à-dire celles que nous pourrons estimer). Dans le cas important des systèmes à deux équations, l'identification de

la première équation (par exemple) est facile à déterminer : il faut qu'au moins une variable exclue de la première équation apparaisse avec un coefficient non-nul dans la seconde équation.

Comme nous le savions déjà à partir des chapitres précédents, l'estimation par les MCO d'une équation qui contient une variable explicative endogène mène généralement à des estimateurs biaisés et non-convergents. En revanche, les DMC peuvent être utilisés pour estimer toute équation identifiée d'un système. Il existe des méthodes plus sophistiquées pour estimer des systèmes, qui ne sont pas traitées ici.

Dans certaines applications, la nuance entre un problème de variable omise et un problème de la simultanéité est subtile. Les deux problèmes, sans parler des erreurs de mesure, peuvent apparaître dans la même équation. Un bon exemple est l'offre de travail des femmes mariées. La durée de la scolarité (*educ*) apparaît à la fois dans l'offre de travail et dans l'équation de salaire [voir les équations (16.19) et (16.20)]. Si des capacités inobservées apparaissent dans le terme d'erreur de l'offre de travail, alors le salaire et l'éducation sont tous deux endogènes. Mais il est important de remarquer qu'une équation estimée par les DMC doit pouvoir s'interpréter indépendamment du reste du système.

Les MES peuvent aussi s'appliquer à des données de séries temporelles. Comme pour l'estimation par les MCO, il faut faire attention, au moment d'utiliser les DMC, à l'existence de tendance ou de problèmes de cointégration. Les problèmes de corrélation sérielle peuvent être traités comme dans la section 15.7. Nous proposons également un exemple d'estimation de MES sur données de panel, où l'estimation d'une équation en différences premières supprime les effets inobservés. Il est donc possible d'estimer cette équation en différences premières par des DMC sur données empilées, exactement comme au chapitre 15. Sinon, dans certains cas, il est possible de dé-moyenner toutes les variables, y compris les variables instrumentales et ensuite d'utiliser les DMC sur données empilées. Cela revient à mettre une variable indicatrice pour chaque unité d'observation en coupe et d'utiliser les DMC, les variables indicatrices étant leurs propres instruments. Appliquer les MES sur données de panel est très utile, car cela permet de tenir compte, en même temps, de l'hétérogénéité inobservée et de la simultanéité. Ce type d'applications deviennent de plus en plus fréquentes et ne sont pas particulièrement difficiles à estimer.

MOTS-CLÉS

Biais de simultanéité p. 655
Condition de rang p. 658
Condition d'ordre p. 658, 662
Équation de forme réduite p. 654
Équation identifiée p. 656
Équation juste identifiée p. 663
Équation non identifiée p. 663
Équation structurelle p. 650
Équation suridentifiée p. 663
Erreur de forme réduite p. 654
Erreur structurelle p. 652
Modèles à équations simultanées (MES) p. 652
Paramètres de forme réduite p. 654
Restrictions d'exclusion p. 657
Simultanéité p. 650
Variable endogène p. 652
Variable endogène retardée p. 664
Variable exogène p. 652
Variable prédéterminée p. 665

EXERCICES

1. Écrivons un système à deux équations du type « offre et demande », c'est-à-dire, avec la même variable y_1 (par exemple la « quantité ») dans le terme de gauche :

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$
$$y_1 = \alpha_2 y_2 + \beta_2 z_2 + u_2.$$

i. Si $\alpha_1 = 0$ ou $\alpha_2 = 0$, expliquez pourquoi il existe une forme réduite pour y_1 (Souvenez-vous : la forme réduite exprime y_1 en une fonction linéaire des variables exogènes et de l'erreur structurelle). Si $\alpha_1 \neq 0$ et $\alpha_2 = 0$ trouvez la forme réduite d'y_2.

ii. Si $\alpha_1 \neq 0$ $\alpha_2 \neq 0$ et $\alpha_1 \neq \alpha_2$ trouvez la forme réduite d'y_1. y_2 a-t-il une forme réduite dans ce cas ?

iii. D'après vous, la condition risque-t-elle d'être vraie dans l'exemple de l'offre et de la demande ? Expliquez pourquoi.

2. Appelons *corn* la consommation de maïs au niveau du département, en kilogrammes par habitant, *price* le prix du kg de maïs, *income* le revenu par habitant du département, et *rainfall* les précipitations en cm durant la dernière saison de culture du maïs. Le modèle à équations simultanées suivant impose la condition d'équilibre que l'offre est égale à la demande :

$$corn = \alpha_1 price + \beta_1 income + u_1$$
$$corn = \alpha_2 price + \beta_2 rainfall + \gamma_2 rainfall^2 + u_2.$$

Quelle est l'équation d'offre et quelle est l'équation de demande ? Pourquoi ?

3. Dans l'exercice 3 du chapitre 3, nous avons estimé une équation pour tester si les gens font un arbitre entre le nombre de minutes par semaines passées à dormir (*sleep*) et le nombre de minutes par semaines passées à travailler (*totwrk*) pour un échantillon aléatoire d'individus. Nous avions également inclus dans l'équation l'éducation et l'âge. Comme *sleep* et *totwrk* sont choisies de manière jointe par les individus, peut-on critiquer l'estimation de l'alternative entre dormir et travailler en raison du biais de simultanéité ? Expliquez.

4. Supposons que le salaire annuel (*earnings*) et la consommation d'alcool (*alcohol*) soient déterminés par le MES suivant :

$$\log(earnings) = \beta_0 + \beta_1 alcohol + \beta_2 educ + u_1$$
$$alcohol = \gamma_0 + \gamma_1 \log(earnings) + \gamma_2 educ + \gamma_3 \log(price) + u_2,$$

où *price* est l'indice des prix locaux de l'alcool (en incluant les taxes). Supposons que *educ* et *price* soient exogènes. Si β_1, β_2, γ_1, γ_2, et γ_3 sont tous différents de zéro, quelle équation est identifiée ? Comment pourriez-vous estimer cette équation ?

5. Nous cherchons à mesurer l'efficacité de l'utilisation des préservatifs sur la transmission des maladies sexuellement transmissibles parmi les lycéens sexuellement actifs. Écrivons un modèle simple :

$$infrate = \beta_0 + \beta_1 conuse + \beta_2 percmale + \beta_3 avginc + \beta_4 city + u_1,$$

où

infrate = la proportion des lycéens sexuellement actifs affectés par une maladie vénérienne.

conuse = la part des garçons qui disent utiliser régulièrement des préservatifs.

avginc = le revenu familial moyen.

city = une variable binaire indiquant si l'école est en ville.

L'unité d'observation est l'école.

i. En supposant que l'équation peut s'interpréter de manière causale, toutes choses égales par ailleurs, quel devrait être le signe de β_1 ?

ii. Pourquoi *infrate* et *conuse* pourraient-ils être déterminés simultanément ?

iii. Si l'utilisation de préservatifs s'accroît avec le taux de maladies vénériennes, c'est-à-dire si $\gamma_1 > 0$ dans l'équation

$$conuse = \gamma_0 + \gamma_1 infrate + other\ factors,$$

quel est le biais probable si on estime β_1 par les MCO ?

iv. Appelons *condis* la variable binaire prenant valeur 1 si une école a un programme de distribution de préservatifs. Comment peut-on utiliser cette variable pour mesurer β_1 (et les autres betas) par la méthode des variables instrumentales ? Quelle est l'hypothèse à faire sur *condis* dans les deux équations ?

6. Étudions un modèle de probabilités linéaires, prédisant si les employeurs offrent une retraite complémentaire (*pension*), en fonction – entre autres – du taux de syndiqués (*percunion*) dans l'entreprise :[1]

$$pension = \beta_0 + \beta_1 percunion + \beta_2 avgage + \beta_3 avgeduc + \beta_4 percmale + \beta_5 percmarr + u_1.$$

i. Pourquoi *percunion* et *pension* risquent-ils d'être déterminés simultanément ?

ii. Supposons qu'il est possible d'obtenir des informations complémentaires sur les familles des employés. Comment pourrait-on utiliser cette information pour créer un instrument pour *percunion* ?

iii. Comment pourrait-on vérifier si cette variable est au moins un instrument raisonnable pour *percunion* ?

7. Une université américaine cherche à estimer la demande de tickets pour les matchs de basketball féminin de son université. Vous disposez des données de séries temporelles sur 10 saisons, soit environ 150 observations au total. Un modèle possible serait

$$lATTEND_t = \beta_0 + \beta_1 lPRICE_t + \beta_2 WINPERC_t + \beta_3 RIVAL_t + \beta_4 WEEKEND_t + \beta_5 t + u_t,$$

où

$PRICE_t$ = le prix de l'entrée en termes réels – c'est-à-dire corrigé de l'inflation.

$WINPERC_t$ = le pourcentage de victoires depuis le début de la saison.

$RIVAL_t$ = une variable binaire indiquant une rivalité préexistante avec cet adversaire.

$WEEKEND_t$ = une variable binaire indiquant si la partie se joue pendant un week-end.

l est le logarithme népérien, la fonction de demande a donc une élasticité-prix constante.

i. Expliquez pourquoi inclure une variable de tendance dans cette équation est une bonne idée.

ii. L'offre de tickets est fixée par la capacité maximale du stade, supposons que celle-ci n'a pas changé au cours des 10 années. Cela veut dire que l'offre ne dépend pas du prix. Peut-on dire pour autant que le prix est exogène dans l'équation de demande ? (Indice : la réponse est non)

iii. Supposons que le prix d'admission change progressivement – disons au début de chaque saison. Les prix sont choisis sur la base de l'affluence moyenne de la saison précédente, ainsi que de la réussite

1 Note de traduction : *avgage* est l'âge moyen des salariés de l'entreprise, *avgeduc* leur éducation moyenne, *percmale* le pourcentage d'hommes, et *percmarr* le pourcentage de personnes mariées.

sportive de l'année dernière. Sous quelles hypothèses le pourcentage de victoires de l'année précédente ($SEASPERC_{t-1}$) est-il un instrument valide pour $lPRICE_t$?

iv. Cela vous semble-t-il raisonnable d'ajouter le (log des) prix des matchs de basketball masculin dans l'équation ? Pourquoi ? Quel serait le signe du coefficient selon la théorie économique ? Une autre variable liée au basketball masculin entre-t-elle dans les déterminants de l'affluence au basketball féminin ?

v. Si on a peur que certaines des séries, notamment *lATTEND* et *lPRICE*, aient des racines unitaires, comment peut-on modifier l'équation estimée ?

vi. Si certaines parties se jouent à guichets fermés, quel problème cela pose-t-il pour l'estimation d'une fonction de demande ? (Indice : si la partie se joue à guichets fermés, peut-on observer la demande ?)

8. Nous cherchons à mesurer l'effet des dépenses scolaires sur le marché de l'immobilier local. Appelons *HPRICE* le prix médian du logement dans une commune et *EXPAND* les dépenses scolaires municipales par habitant de cette commune. Nous voulons estimer le modèle suivant sur des données de panel pour les années 1992, 1994 et 1996 :

$$lHPRICE_{it} = \theta t + \beta_1 lEXPEND_{it} + \beta_2 lPOLICE_{it} + \beta_3 lMEDINC_{it} + \beta_4 PROPTAX_{it} + a_{i1} + u_{it1},$$

où $POLICE_{it}$ est la dépense de police par habitant, $MEDINC_{it}$ est le revenu médian, et $PROPTAX_{it}$ est le taux d'imposition sur le logement[2]. *l* est le logarithme népérien. Les dépenses scolaires municipales et le prix des logements sont déterminés simultanément, car la valeur des logements affecte la base fiscale des communes et donc les ressources disponibles pour les écoles.

Supposons que le financement des écoles ait été radicalement modifié en 1994 : au lieu d'être financées par les impôts sur l'immobilier, le financement des écoles aurait été totalement pris en charge par l'État. Appelons $lSTATEALL_{it}$ le log des dépenses de l'État pour la commune *i* l'année *t*, qui est exogène dans l'équation précédente. Comment pourrait-on estimer β_j ?

EXERCICES SUR ORDINATEUR

C1. Utilisez le fichier de données SMOKE pour cet exercice.

i. Nous voulons estimer les effets de la consommation de tabac sur le revenu annuel (par exemple liés aux maladies, ou à une baisse de productivité). Écrivons le modèle :

$$\log(income) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u_1,$$

où *cigs* est le nombre moyen de cigarettes fumées par jour. Comment interprétez-vous β_1 ?

ii. Pour tenir compte du fait que la consommation de cigarettes peut être déterminée simultanément au revenu, écrivons l'équation de demande de cigarettes

$$cigs = \gamma_0 + \gamma_1 \log(income) + \gamma_2 educ + \gamma_3 age + \gamma_4 age^2 + \gamma_5 \log(cigpric) + \gamma_6 restaurn + u_2,$$

où *cigpric* est le prix d'un paquet de cigarettes (en centimes de dollars), et *restaurn* est une variable binaire prenant pour valeur 1 si des lois restreignent la consommation de cigarettes dans les restaurants dans l'État où la personne vit. En supposant ces variables exogènes, quels signes attendez-vous pour γ_5 et γ_6 ?

2 Note de traduction : impôts locaux en France.

iii. Sous quelle(s) hypothèse(s) l'équation de la question (i) est-elle identifiée ?

iv. Estimez l'équation de revenu par les MCO et discutez le coefficient de β_1.

v. Estimez la forme réduite prédisant *cigs* (rappelez-vous que cela revient à régresser *cigs* sur toutes les variables exogènes). log(*cigpric*) et *restaurn* sont-ils significatifs dans la forme réduite ?

vi. Estimons maintenant l'équation de revenu par les DMC. Discutez le coefficient estimé de β_1 et comparez-le avec le coefficient des MCO.

vii. Pensez-vous que le prix des cigarettes et les lois limitant la consommation de cigarettes sont exogènes dans l'équation de revenu ?

C2. Utilisez le fichier de données MROZ pour cet exercice.

i. Ré-estimez la fonction d'offre de travail de l'exemple 16.5, en utilisant log(*hours*) comme variable dépendante. Comparez l'élasticité estimée (qui est maintenant constante) avec l'élasticité tirée de l'équation (16.24) au nombre moyen d'heures travaillées.

ii. Dans l'équation d'offre de travail de la question (i), utilisez *motheduc* et *fatheduc* comme variables instrumentales pour prendre en compte une éventuelle endogénéité d'*educ*. Faites attention au fait qu'il y a maintenant deux variables endogènes dans l'équation.

iii. Testez les restrictions suridentifiantes dans l'estimation des DMC de la question (ii). Le test rejette-t-il les instruments ?

C3. Utilisez le fichier de données OPENNESS pour cet exercice.

i. Comme log(*pcinc*) n'est significatif ni dans (16.22) ni dans la forme réduite d'*open*, supprimez-le de l'analyse. Estimez (16.22) avec les MCO et les variables instrumentales, en omettant log(*pcinc*). Cela change-t-il les principaux résultats ?

ii. En continuant à omettre log(*pcinc*) de notre analyse, *land* ou log(*land*) sont-ils de meilleurs instruments pour *open* ? (Indice : régressez *open* sur chacune de ces variables prises séparément, puis ensemble).

iii. Revenons maintenant à (16.22). Ajoutons la variable indicatrice *oil* à l'équation, que nous supposerons exogène. Estimez l'équation par la méthode des variables instrumentales. Être un producteur de pétrole a-t-il un effet *ceteris paribus* sur l'inflation ?

C4. Utilisez le fichier de données CONSUMP pour cet exercice.

i. Dans l'exemple 16.7, testez les restrictions suridentifiantes en estimant (16.35) avec la méthode de la section 15.5. Que concluez-vous ?

ii. Campbell et Mankiw (1990) utilisent les *doubles* retards de toutes les variables comme instruments, à cause des risques d'erreur de mesure et de la durée de diffusion des informations. Ré-estimez (16.35) en utilisant seulement gc_{t-2}, gy_{t-2}, et $r3_{t-2}$ comme instruments. Quelles sont les différences avec l'estimation (16.36) ?

iii. Régressez gy_t sur les instruments de la question (ii) et testez si gy_t est suffisamment corrélée avec ces instruments. Pourquoi est-ce important ?

C5. Utilisez un « *Economic Report of the President* » (daté de 2005 au moins) pour mettre à jour les données de CONSUMP au moins jusqu'en 2003. Cela change-t-il les résultats principaux ?

C6. Utilisez le fichier de données CEMENT pour cet exercice.

i. Écrivons une fonction d'offre (inverse) où la croissance mensuelle des prix du ciment (*gprc*) dépend de la croissance de la quantité (*gcem*) :

$$gprc_t = \alpha_1 gcem_t + \beta_0 + \beta_1 gprcpet_t + \beta_2 feb_t + \ldots + \beta_{12} dec_t + u_t^s$$

où *gprcpet* (croissance du prix du pétrole) est supposé exogène, et *feb*, ..., *dec* sont des variables indicatrices mensuelles. Quels signes prédiriez-vous pour α_1 et β_1 ? Estimez cette équation par les MCO. La fonction d'offre est-elle croissante ?

ii. La variable *gdefs* représente la croissance mensuelle des dépenses réelles de défense aux États-Unis. Sous quelles hypothèses sur *gdefs* cette variable est-elle un bon instrument pour *gcem* ? Testez si *gcem* a une corrélation partielle avec gdefs (Ne vous préoccupez pas d'une corrélation sérielle possible en forme réduite). Est-il possible d'utiliser *gdefs* comme instrument en estimant la fonction d'offre ?

iii. Shea (1993) pense que la croissance de la construction résidentielle (*gres*) et non-résidentielle (*gnon*) sont des instruments valides pour *gcem*. Ces facteurs de demande pourraient ne pas être trop corrélés avec l'erreur de l'offre u_t^s. Testez si *gcem* a une corrélation partielle avec *gres* et *gnon* ; là encore, ne vous préoccupez pas des corrélations sérielles en forme réduite.

iv. Estimez la fonction d'offre, en utilisant *gres* et *gnon* comme instruments pour *gcem*. Que concluez-vous sur la fonction d'offre statique de ciment ? [La fonction d'offre dynamique est apparemment croissante ; voir Shea (1993)]

C7. Référez-vous à l'exemple 13.9 et au fichier de données CRIME4.

i. Supposons que vous pensiez qu'après avoir différencié pour supprimer l'effet inobservé, $\Delta\log(polpc)$ et $\Delta\log(crmrte)$ sont déterminés simultanément, notamment parce que les hausses de la criminalité sont associées à une hausse des effectifs de police. Comment cela peut-il expliquer le coefficient positif de l'équation (13.33) ?

ii. La variable *taxpc* est le montant de taxes collectées par personne dans le comté (aux États-Unis). Est-ce raisonnable d'exclure ce facteur de l'équation de criminalité ?

iii. Estimez la forme réduite prédisant $\Delta\log(polpc)$ en utilisant les MCO sur données empilées, en incluant l'instrument potentiel, $\log(taxpc)$. $\Delta\log(taxpc)$ semble-t-il être un bon instrument ? Pourquoi ?

iv. Supposons que l'État de Caroline du Nord ait donné des primes à certains comtés pour augmenter la taille des forces de police. Comment pourrait-on utiliser cette information pour mesurer l'effet de l'augmentation des forces de police sur le taux de criminalité ?

C8. Utilisez les données du fichier FISH, qui viennent de Graddy (1995) pour cet exercice. Ces données sont aussi utilisées dans l'exercice sur ordinateur C9 dans le chapitre 12. Maintenant, nous estimerons une fonction de demande de poisson.

i. Supposons que l'équation de demande – à l'équilibre à chaque période – s'écrive

$$\log(totqty_t) = \alpha_1 \log(avgprc_t) + \beta_{10} + \beta_{11} mon_t + \beta_{12} tues_t + \beta_{13} wed_t + \beta_{14} thurs_t + u_{t1},$$

la demande peut donc changer en fonction du jour de la semaine. Si la variable de prix est endogène, quel type d'information supplémentaire faut-il pour estimer de manière convergente les paramètres de l'équation de demande ?

ii. Les variables $wave2_t$ et $wave3_t$ sont des mesures de la hauteur des vagues dans l'océan sur les derniers jours. Quelles sont les deux hypothèses nécessaires pour se servir de $wave2_t$ et $wave3_t$ comme instruments de $\log(avgprc_t)$ en estimant l'équation de demande ?

iii. Régressez $\log(avgprc_t)$ sur des variables indicatrices de jour de semaine et les deux mesures de vagues. $wave2_t$ et $wave3_t$ sont-ils significatifs ? Quelle est la *p*-valeur du test ?

iv. Estimez maintenant l'équation de demande par les DMC. Quelle est l'intervalle de confiance à 95 % de l'élasticité-prix de la demande ? L'élasticité estimée est-elle raisonnable ?

v. Calculez les résidus des DMC, $\hat{u}_{t1}$. Ajouter un lag, $\hat{u}_{t-1,1}$, dans l'équation de demande estimée par les DMC. Souvenez-vous d'utiliser $\hat{u}_{t-1,1}$ comme son propre instrument. Y a-t-il un signe de corrélation AR(1) entre les erreurs de l'équation de demande ?

vi. Puisque l'équation d'offre dépend des variables de hauteur des vagues, quelles sont les deux hypothèses nécessaires pour estimer l'élasticité-prix de l'offre ?

vii. Dans l'équation de forme réduite prédisant $\log(avgprc_t)$, les variables indicatrices de jour de semaine sont-elles significatives de manière jointe ? Que concluriez-vous sur notre capacité à mesurer l'élasticité-prix de l'offre ?

C9. Dans cet exercice, utilisez le fichier AIRFARE, mais seulement l'année 1997.

i. Écrivons une fonction de demande de billets d'avion aux États-Unis :

$$\log(passen) = \beta_{10} + \alpha_1 \log(fare) + \beta_{11} \log(dist) + \beta_{12}[\log(dist)]^2 + u_1,$$

où

passen = nombre moyen de passagers par jour.

fare = prix moyen du billet d'avion.

dist = distance parcourue (en miles).

Si c'est vraiment une fonction de demande, quel est le signe d'α_1 ?

ii. Estimez l'équation de la question (i) avec les MCO. Quelle est l'élasticité-prix estimée ?

iii. *concen* est une mesure de concentration du marché (précisément, c'est la part de marché du transporteur le plus important). Expliquez avec des mots les hypothèses nécessaires pour considérer que *concen* est exogène dans l'équation de demande. Supposons maintenant que *concen* est exogène dans l'équation de demande. Estimez la forme réduite de log(*fare*) et vérifiez que *concen* a un effet positif sur log(*fare*).

iv. Estimez la fonction de demande avec les variables instrumentales. Quel est maintenant l'élasticité-prix de la demande estimée ? Est-elle différente du coefficient estimé par les MCO ?

v. En regardant l'estimateur des variables instrumentales, décrivez l'effet de la distance sur la demande de billets d'avion.

C10. Utilisez tout le panel contenu dans AIRFARE pour cet exercice. Écrivons l'équation de demande d'un modèle à équations simultanées à effets inobservés :

$$\log(passen_{it}) = \theta_{t1} + \alpha_1 \log(fare_{it}) + a_{i1} + u_{it1},$$

où les variables de distance sont inclues dans a_{i1}.

i. Estimez la fonction de demande en utilisant des effets fixes, pensez à inclure également des variables indicatrices d'année. Quelle est l'élasticité estimée ?

ii. Utilisez des effets fixes pour estimer la forme réduite

$$\log(fare_{it}) = \theta_{t2} + \pi_{21} concen_{it} + a_{i2} + v_{it2}.$$

Faites un test pour vous assurer que $concen_{it}$ peut être utilisé comme variable instrumentale pour $\log(fare_{it})$.

iii. Estimez maintenant la fonction de demande avec la transformation à effets fixes et des variables instrumentales, comme dans l'équation (16.42). Quelle est l'élasticité estimée ? Est-elle statistiquement significative ?

C11. Pour estimer des *Courbes d'Engel*, une méthode habituelle est de modéliser l'effet dépense totale, et éventuellement de variables démographiques, sur la part d'un bien dans les dépenses. Une spécification classique s'écrit

$$sgood = \beta_0 + \beta_1 ltotexpend + demographics + u,$$

où *sgood* est la part d'un bien donné dans les dépenses totales et *ltotexpend* est le log de la dépense totale. Le signe et le coefficient de β_1 pour les différents types de bien sont les variables d'intérêt.

Pour prendre en compte l'endogénéité potentielle de *ltotexpend* – qui peut être vue comme un problème de variables omises, d'équations simultanées, ou les deux – le log du revenu familial est souvent utilisé comme variable instrumentale. Appelons *lincome* le log du revenu familial. Pour le reste de cette question, utilisez les données contenues dans EXPENDSHARE, qui viennent de Blundell, Duncan et Pendakur (1998).

i. Utilisez *sfood*, la part des dépenses utilisées pour la nourriture, comme variable dépendante. Quelles valeurs prend *sfood* ? Est-ce surprenant qu'il n'y ait pas de zéros ?

ii. Estimez l'équation

$$sfood = \beta_0 + \beta_1 ltotexpend + \beta_2 age + \beta_3 kids + u \qquad \textbf{[16.43]}$$

avec les MCO, et notez le coefficient de *ltotexpend*, $\hat{\beta}_{OLS,1}$, ainsi que son écart-type estimé robuste à l'hétéroscédasticité. Interprétez le résultat.

iii. Estimez l'équation de forme réduite de *ltotexpend*, en utilisant *lincome* comme instrument, et en incluant *age* et *kids*. En supposant que *lincome* est exogène dans (16.43), *lincome* est-il un instrument valide pour *ltotexpend* ?

iv. Estimez maintenant (16.43) avec les variables instrumentales. Comparez $\hat{\beta}_{IV,1}$ avec $\hat{\beta}_{OLS,1}$. Quels sont les intervalles de confiance robustes à 95 % ?

v. Testez l'hypothèse nulle que *ltotexpend* est exogène dans (16.43) avec le test de la section 15.5. Assurez-vous de noter et d'interpréter la p-valeur. Y a-t-il des restrictions suridentificantes à tester ?

vi. Remplacez *sfood* par *salcohol* dans (16.43) et estimez l'équation par les MCO et les DMC. Quels sont maintenant les coefficients de *ltotexpend* ?

CHAPITRE 17

MODÈLES À VARIABLE DÉPENDANTE LIMITÉE ET CORRECTION POUR LA SÉLECTION DE L'ÉCHANTILLON

Traduction de Pierre André

Dans le chapitre 7, nous avons étudié le modèle à probabilités linéaires, qui n'est que l'application des régressions linéaires multiples aux variables binaires. Les variables binaires sont des exemples de **variable dépendante limitée (VDL)**. Une VDL est une variable dépendante ne pouvant pas prendre toutes les valeurs possibles. Une variable binaire ne peut prendre que deux valeurs, zéro et un. Dans la section 7.7, nous avions discuté l'interprétation de régressions multiples avec des variables prédites discrètes, notamment le cas où y prend un petit nombre de valeurs entières – par exemple, le nombre de fois où un homme est arrêté dans l'année, ou le nombre d'enfants d'une femme. Dans d'autres cas, nous avons trouvé d'autres variables dépendantes limitées, comme par exemple le pourcentage de personnes cotisant pour la retraite (qui doit être entre zéro et 100) ou la moyenne scolaire (généralement entre zéro et 20 en France).

Beaucoup de variables économiques sont limitées d'une certaine manière, souvent parce qu'elles sont forcément positives. Par exemple, le salaire horaire, les prix de l'immobilier, les taux d'intérêt nominaux sont toujours positifs. Mais toutes les variables n'ont pas nécessairement besoin d'un traitement particulier. Si les variables strictement positives prennent beaucoup de valeurs, un traitement spécifique est rarement nécessaire. Quand y est discret et prend un petit nombre de valeurs, cela n'a par contre pas de sens de le considérer comme une variable approximativement continue. Les modèles linéaires ne sont pas nécessairement inadaptés quand y est discret. Cependant, nous avons vu au chapitre 7 que dans le cas des variables dépendantes binaires, le modèle de probabilités linéaires a des inconvénients. Dans la section 17.1, nous discuterons des modèles logit et probit, qui répondent aux limites du MPL ; l'inconvénient est qu'ils sont plus difficiles à interpréter.

D'autres formes de variable dépendante limitée apparaissent en économétrie, notamment pour modéliser le comportement des individus, des familles ou des entreprises. Le comportement d'optimisation mène souvent à des **solutions en coin** pour une part non négligeable de la population. C'est-à-dire qu'un nombre important d'observations ont choisi une quantité ou une valeur nulle, par exemple. Chaque année, une proportion importante des familles donne un montant nul à des associations humanitaires. La distribution des montants donnés aux associations caritatives est répartie sur beaucoup de valeurs positives différentes, mais avec énormément d'observations dont la valeur est zéro. Un modèle linéaire pourrait prédire le montant donné mais il mènerait probablement à des montants prédits négatifs pour certaines familles. Il n'est pas possible de prédire le logarithme du montant donné, puisque ce montant prend souvent la valeur zéro. Le modèle Tobit, couvert en section 17.2, est conçu pour modéliser des variables dépendantes limitées avec des solutions en coin.

Les variables de comptage sont une autre forme importante de VDL : elles prennent des valeurs entières non négatives. La section 17.3 illustre le fait que les modèles de Poisson sont adaptés à ce cas.

Dans certains cas, les variables dépendantes limitées sont dues à une censure des données, ce que nous introduisons en section 17.4. Le problème général de la sélection de l'échantillon, où l'échantillon observé n'est pas tiré aléatoirement dans la population d'intérêt, est traité dans la section 17.5.

Les modèles à variable dépendante limitée peuvent être utilisés sur des données de séries temporelles ou de panel, mais ils sont le plus souvent appliqués aux données en coupe. Les problèmes de sélection de l'échantillon sont généralement limités aux données en coupe ou de panel. Dans ce chapitre, nous nous limiterons aux applications sur données en coupe. Wooldridge (2010) analyse le problème des données de panel et donne beaucoup d'autres détails sur les applications sur données tant en coupe qu'en panel.

17.1 LES MODÈLES LOGIT ET PROBIT POUR LES RÉPONSES BINAIRES

Le modèle à probabilités linéaires est simple à estimer et à utiliser, mais nous avions discuté de ses inconvénients dans la section 7.5. Les deux inconvénients les plus importants sont que les valeurs prédites peuvent être plus petites que zéro ou plus grandes que un, et que les effets marginaux des variables explicatives

(apparaissant en niveaux) sont constants. Ces limites des MPL peuvent être dépassées en utilisant des **modèles à réponse binaire** plus sophistiqués.

Les modèles à réponse binaire visent principalement à prédire la **probabilité de réponse**

$$P(y = 1|\mathbf{x}) = P(y = 1|x_1, x_2, \ldots, x_k), \qquad [17.1]$$

où **x** représente l'ensemble des variables explicatives. Par exemple, quand y est une variable indicatrice d'emploi, **x** pourrait contenir des caractéristiques individuelles comme l'éducation, l'âge, la situation familiale, et d'autres facteurs affectant l'emploi, comme une variable indicatrice binaire pour la participation à une formation récente à la recherche d'emploi.

Spécification des modèles logit et probit

Dans le MPL, nous supposons que la probabilité de réponse est une fonction linéaire d'un ensemble de paramètres, β_j ; voir l'équation (7.27). Pour éviter les limites des MPL, considérons la classe des modèles à réponse binaire de la forme

$$P(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}), \qquad [17.2]$$

où G est une fonction prenant des valeurs strictement comprises entre zéro et un : $0 < G(z) < 1$ pour tout nombre réel z. La probabilité de réponse estimée sera donc toujours strictement comprise entre zéro et un. Comme dans les chapitres précédents, nous écrirons $\mathbf{x}\boldsymbol{\beta} = \beta_1 x_1 + \ldots + \beta_k x_k$.

Différentes fonctions non linéaires ont été utilisées pour la fonction G, assurant que les probabilités soient entre zéro et un. Les deux fonctions couvertes ici sont utilisées dans l'immense majorité des applications (avec le MPL). Dans le **modèle logit**, G est la fonction logistique :

$$G(z) = \exp(z)/[1 + \exp(z)] = \Lambda(z), \qquad [17.3]$$

qui est comprise entre zéro et un pour tout nombre réel z. C'est la fonction de distribution cumulative d'une variable aléatoire logistique standard. Dans le **modèle probit**, G est la fonction de répartition (fr) de la loi normale standardisée, qu'on écrit avec l'intégrale

$$G(z) = \Phi(z) \equiv \int_{-\infty}^{z} \phi(v)\,dv, \qquad [17.4]$$

où $\phi(z)$ est la densité de la loi normale standardisée :

$$\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2). \qquad [17.5]$$

Ce choix de G permet encore de s'assurer que (17.2) est toujours strictement comprise entre zéro et un pour toutes les valeurs des x_j.

Les fonctions G en (17.3) et (17.4) sont toutes deux croissantes. Ces deux fonctions croissent le plus vite en $z = 0$, $G(z) \to 0$ quand $z \to -\infty$, et $G(z) \to 1$ quand $z \to \infty$. La fonction logistique est dessinée dans la figure 17.1. La fr de la loi normale ressemble beaucoup à la fr de la loi logistique.

Les modèles logit et probit peuvent être expliqués par un **modèle à variables latentes** sous-jacent. Soit y^* une variable inobservée, ou *latente*, et supposons que

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + e, \quad y = 1[y^* > 0], \qquad [17.6]$$

où nous introduisons la notation $1[\cdot]$ pour définir les variables dépendantes binaires. La fonction $1[\cdot]$ est appelée *fonction indicatrice*, elle prend la valeur un si l'événement entre crochets est vrai, et zéro sinon. De ce fait, y vaut un si $y^* > 0$, et y vaut zéro si $y^* \leq 0$. *Nous* supposerons que e est indépendant de x et que e suit soit une loi de distribution logistique standardisée, soit une loi de distribution normale standardisée. Dans les deux cas,

e est distribué de manière symétrique autour de zéro, ce qui veut dire que $1 - G(-z) = G(z)$ pour tout nombre z. Les économistes tendent à préférer l'hypothèse de normalité pour e, c'est pourquoi le modèle probit est plus populaire que le modèle logit en économétrie. De plus, plusieurs problèmes de spécification dont nous parlerons plus tard sont plus faciles à analyser avec les modèles probit grâce aux propriétés de la loi normale.

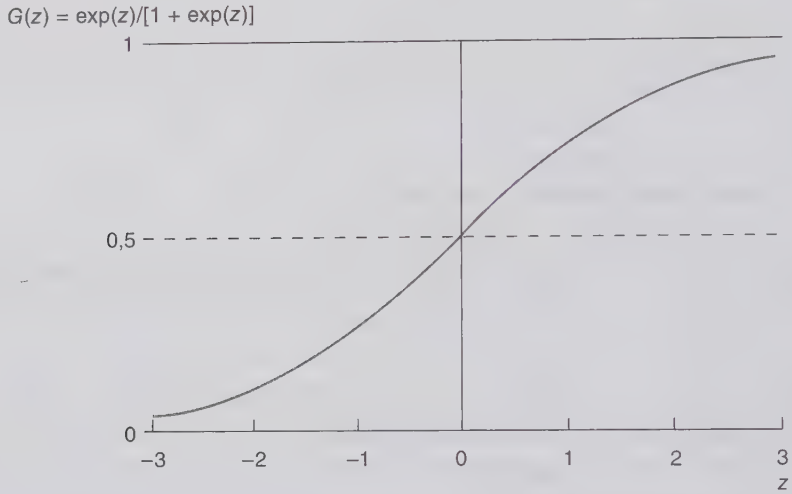


Figure 17.1 *Graphique de la fonction logistique* G(z) = exp(z)/[1 + exp(z)].

La probabilité de réponse d'y peut être tirée de (17.6) et des hypothèses mentionnées plus haut :

$$\begin{aligned} P(y = 1|x) &= P(y^* > 0|x) = P[e > -(\beta_0 + \mathbf{x}\boldsymbol{\beta})|\mathbf{x}] \\ &= 1 - G[-(\beta_0 + \mathbf{x}\boldsymbol{\beta})] = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}), \end{aligned}$$

ce qui est exactement la même chose que (17.2).

Dans la plupart des applications des modèles à réponse binaire, l'objectif est de mesurer l'effet des x_j sur la probabilité de réponse $P(y = 1|\mathbf{x})$. La formulation à variable latente pourrait donner l'impression que nous nous intéressons aux effets des x_j sur y^*. Comme nous le verrons, pour les logit et probit, le *signe* de l'effet de x_j sur $E(y^*|\mathbf{x}) = \beta_0 + \mathbf{x}\boldsymbol{\beta}$ et sur $E(y|\mathbf{x}) = P(y = 1|\mathbf{x}) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$ est toujours le même. Mais la variable latente y^* a rarement une unité correctement définie (Par exemple, y^* pourrait être la différence d'utilité entre deux actions). Les valeurs des β_j ne sont donc pas très utiles en tant que telles (au contraire des modèles à probabilités linéaires). Dans la plupart des cas, le but est d'estimer l'effet des x_j sur la probabilité de succès $P(y = 1|x)$, mais le fait que $G(\cdot)$ soit non linéaire le rend plus compliqué.

Pour trouver l'effet marginal de variables approximativement continues sur la probabilité de réponse, cela nécessite quelques calculs. Si x_j est approximativement continue, son effet marginal sur $p(\mathbf{x}) = P(y = 1|\mathbf{x})$ vient des dérivées partielles :

$$\frac{\partial p(x)}{\partial x_j} = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})\beta_j, \text{ où } g(z) \equiv \frac{dG}{dz}(z). \qquad \textbf{[17.7]}$$

Comme G est la fr d'une variable aléatoire continue, g est une fonction de densité de probabilités. Dans le cas des logit et probit, $G(\cdot)$ est une fr strictement croissante, donc $g(z) > 0$ pour tout z. L'effet marginal de x_j sur $p(\mathbf{x})$ dépend de $\mathbf{x}$ via le terme positif $g(\beta_0 + \mathbf{x}\boldsymbol{\beta})$ ce qui veut dire que l'effet marginal a toujours le même signe que β_j.

L'équation (17.7) montre que l'effet *relatif* de deux variables explicatives continues ne dépend pas de $\mathbf{x}$: le rapport des effets partiels de x_j et x_h est β_j/β_h. Dans le cas fréquent où g est une densité symétrique en zéro, avec un unique maximum en zéro, l'effet est le plus grand quand $\beta_0 + \mathbf{x}\boldsymbol{\beta} = 0$. Par exemple, dans le cas du probit avec $g(z) = \phi(z)$, $g(0) = \phi(0) = 1/\sqrt{2\pi} \approx 0{,}40$. Dans le cas logit, $g(z) = \exp(z)/[1 + \exp(z)]^2$, et donc $g(0) = 0{,}25$.

Si, par exemple, x_1 est une variable explicative binaire, l'effet d'un changement de x_1 de zéro à un, toutes choses égales par ailleurs, s'écrit

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \ldots + \beta_k x_k). \qquad \textbf{[17.8]}$$

Cela dépend encore des valeurs des autres x_j. Par exemple, si y est une variable indicatrice d'emploi et x_1 est une variable indicatrice de participation à une formation à la recherche d'emploi, cet effet dépend d'autres caractéristiques qui affectent l'employabilité, comme l'éducation et l'expérience. Remarquez que le signe de β_1 est suffisant pour savoir si le programme a un effet positif ou négatif. Mais pour connaître la *valeur* de cet effet, nous devons estimer toutes les quantités de (17.8).

Nous pouvons aussi utiliser la différence dans (17.8) pour d'autres variables discrètes (comme le nombre d'enfants). Si x_k est une de ces variables, l'effet d'une augmentation de x_k de c_k à $c_k + 1$ est

$$\begin{aligned} &G[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k(c_k + 1)] \\ &- G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k c_k). \end{aligned} \qquad \textbf{[17.9]}$$

On peut facilement ajouter les formes fonctionnelles habituelles dans les variables explicatives. Par exemple, dans le modèle

$$\mathrm{P}(y = 1|\mathbf{z}) = G(\beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3),$$

l'effet marginal de z_1 sur $\mathrm{P}(y = 1|\mathbf{z})$ est $\partial \mathrm{P}(y = 1|\mathbf{z})/\partial z_1 = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_1 + 2\beta_2 z_1)$ et l'effet marginal de z_2 sur cette probabilité est $\partial \mathrm{P}(y = 1|\mathbf{z})/\partial z_2 = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_3/z_2)$, où $\mathbf{x}\boldsymbol{\beta} = \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3$. En conséquence, $g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_3/100)(1/z_2)$ est une approximation de l'effet d'une augmentation de z_2 de 1 % sur la probabilité de réponse.

Il peut être utile de mesurer l'élasticité de la probabilité de réponse par rapport à des variables explicatives, mais il faut être attentif à l'interprétation des changements en pourcentages de probabilités. Par exemple, le changement d'une probabilité de 0,04 à 0,06 représente une augmentation de 2 *points de pourcentages*, mais une augmentation relative de 50 % par rapport à la valeur initiale. On peut calculer dans le modèle précédent l'élasticité de $\mathrm{P}(y = 1|\mathbf{z})$ par rapport à z_2, qui est $\beta_3[g(\beta_0 + \mathbf{x}\boldsymbol{\beta})/G(\beta_0 + \mathbf{x}\boldsymbol{\beta})]$. L'élasticité par rapport à z_3 est $(\beta_4 z_3)$ $[g(\beta_0 + \mathbf{x}\boldsymbol{\beta})/G(\beta_0 + \mathbf{x}\boldsymbol{\beta})]$. Dans le premier cas, l'élasticité a toujours le même signe que β_3, mais cela dépend de tous les paramètres et des valeurs des variables explicatives. Si $z_3 > 0$, la seconde élasticité a toujours le même signe que le paramètre β_4.

Les modèles avec interactions entre variables explicatives peuvent être un peu compliqués, mais il faut aussi calculer les dérivées partielles et les effets marginaux pour les valeurs d'intérêt. Quand on mesure l'effet des variables discrètes – quelle que soit la complexité du modèle – il faut utiliser (17.9). Ceci est discuté plus en détail dans la sous-section sur l'interprétation des estimations en page **686.**

Estimation des modèles logit et probit par maximum de vraisemblance

Comment peut-on estimer les modèles de réponse binaire non linéaires ? Le MPL peut être estimé par les moindres carrés ordinaires (voir la section 7.5) ou, dans certains cas, par les moindres carrés pondérés (voir la section 8.5). Or, quand $\mathrm{E}(y|\mathbf{x})$, est non linéaire, on ne peut pas utiliser les MCO ou les MCP. Il serait

possible d'utiliser des versions non linéaires de ces méthodes, mais ce n'est pas plus difficile d'utiliser une **estimation par maximum de vraisemblance (EMV)** (voir l'annexe 17A pour une courte discussion). Jusqu'à maintenant, nous avions peu besoin des EMV ; même si nous avions remarqué que l'estimateur des MCO est un estimateur du maximum de vraisemblance (conditionnelle aux variables explicatives). Pour estimer les modèles à variable dépendante limitée, les estimations par maximum de vraisemblance sont indispensables. Comme l'EMV est basée sur la distribution de y conditionnellement à $\mathbf{x}$, l'hétéroscédasticité de $\text{Var}(y|\mathbf{x})$ est prise en compte automatiquement.

Supposons que nous ayons un échantillon aléatoire avec n observations. Pour trouver l'estimateur du maximum de vraisemblance – conditionnellement aux variables explicatives – il faut écrire la distribution de probabilités de y_i conditionnellement à $\mathbf{x}_i$. Elle s'écrit

$$f(y|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^y[1 - G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y}, \; y = 0, 1, \qquad \textbf{[17.10]}$$

où, pour simplifier, nous incluons la constante dans le vecteur $\mathbf{x}_i$. On peut facilement voir que cette expression donne $G(\mathbf{x}_i\boldsymbol{\beta})$ quand $y = 1$, et $1 - G(\mathbf{x}_i\boldsymbol{\beta})$ quand $y = 0$. La fonction de log-vraisemblance de l'observation i est une fonction des paramètres et des données $(\mathbf{x}_i, y_i)$; elle s'obtient en calculant le log de (17.10) :

$$\ell_i(\boldsymbol{\beta}) = y_i\log[G(\mathbf{x}_i\boldsymbol{\beta})] + (1 - y_i)\log[1 - G(\mathbf{x}_i\boldsymbol{\beta})]. \qquad \textbf{[17.11]}$$

Comme $G(\cdot)$ est toujours strictement comprise entre zéro et un pour les logit et probit, $\ell_i(\boldsymbol{\beta})$ est toujours définie quel que soit $\boldsymbol{\beta}$.

La log-vraisemblance d'un échantillon de taille n se calcule en prenant la somme de (17.11) sur toutes les observations : $\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta})$. L'EMV de $\boldsymbol{\beta}$, noté $\hat{\boldsymbol{\beta}}$, maximise cette log-vraisemblance. Si $G(\cdot)$ est la fr logistique, $\hat{\boldsymbol{\beta}}$ est *l'estimateur logit* ; si $G(\cdot)$ est la fr normale centrée réduite, $\hat{\boldsymbol{\beta}}$ est l'*estimateur probit*.

Comme le problème de maximisation est non linéaire, il n'existe pas de formules donnant les estimateurs du maximum de vraisemblance d'un logit ou d'un probit. En plus de rendre ces estimateurs plus difficiles à trouver, cela rend la théorie statistique beaucoup plus compliquée pour les logit et probit que pour les MCO ou même les DMC. La théorie générale de l'EMV montre néanmoins que sous des conditions très générales, l'EMV est convergent, asymptotiquement normal et asymptotiquement efficace [Voir Wooldridge (2010, chapitre 13) pour une discussion générale]. Nous utiliserons ces résultats ici ; appliquer des modèles logit et probit est assez facile dès lors que l'on comprend leur fonctionnement statistique.

Chaque $\hat{\beta}_j$ a un écart-type estimé (asymptotique) qui a une formule compliquée donnée dans l'annexe du chapitre. Une fois que nous avons les écarts-types estimés – qui sont donnés par tout logiciel qui calcule des estimations de probit et logit –, il est possible de faire des tests t (asymptotiques) exactement comme dans le cas des MCO, DMC, et des autres estimateurs que nous avons rencontrés. Par exemple, pour tester $H_0 : \beta_j = 0$, il faut calculer la statistique $t = \hat{\beta}_j / \text{se}\hat{\beta}_j$ et faire le test comme d'habitude, une fois décidé si on veut une alternative unilatérale ou bilatérale.

Test d'hypothèses multiples

Il est aussi possible de tester des hypothèses multiples dans les modèles logit et probit. Dans la plupart des cas, ce seront des tests de plusieurs restrictions d'exclusion, comme dans la section 4.5. Nous nous focaliserons donc sur les restrictions d'exclusion ici.

Il y a trois manières de tester des restrictions d'exclusion dans les modèles logit et probit. Le test du multiplicateur de Lagrange, ou du score, ne nécessite d'estimer le modèle que sous l'hypothèse nulle, comme

dans le cas linéaire de la section 5.2 ; nous ne le couvrirons pas ici, parce qu'il est rarement nécessaire pour tester des restrictions d'exclusion [Voir Wooldridge (2010, chapitre 15) pour d'autres usages de ce test dans les modèles à réponse binaire].

Le test de Wald nécessite l'estimation du modèle non restreint. Dans le cas linéaire, la **statistique de Wald** est une transformation simple de la statistique F ; il n'est donc pas nécessaire de couvrir la statistique de Wald séparément. La formule de la statistique de Wald est donnée dans Wooldridge (2010, chapitre 15). Cette statistique est donnée par les logiciels d'économétrie qui permettent de tester des restrictions d'exclusion après que le modèle non restreint ait été estimé. Elle suit asymptotiquement une distribution du chi-deux, *df* étant le nombre de restrictions testées.

Si les modèles restreint et non restreint sont tous deux faciles à estimer – ce qui est généralement le cas des restrictions d'exclusions –, le *test du rapport de vraisemblance* (*RV*) devient très intéressant. Le test *RV* est basé sur la même idée que le test *F* des modèles linéaires. Le test *F* mesure l'accroissement de la somme des carrés des résidus quand on supprime des variables du modèle. Le test *RV* mesure la différence des fonctions de log-vraisemblance entre les modèles restreint et non restreint. Le principe est que comme l'EMV maximise une fonction de log-vraisemblance, supprimer des variables *diminue* généralement la log-vraisemblance – en tout cas, cela ne l'augmente pas (De la même manière, le R-carré n'augmente jamais quand on supprime des variables d'une régression). La question est de savoir si la baisse de la log-vraisemblance est suffisamment importante pour que l'on puisse conclure que les variables concernées sont importantes. Nous pourrons en décider une fois dotés d'une statistique de test et d'un ensemble de valeurs critiques.

La **statistique du rapport de vraisemblance** est *le double* de la différence des log-vraisemblances :

$$RV = 2(\mathcal{L}_{ur} - \mathcal{L}_{r}), \qquad \textbf{[17.12]}$$

où $\mathcal{L}_{ur}$ est la log-vraisemblance du modèle non restreint et $\mathcal{L}_{r}$ est la log-vraisemblance du modèle restreint. Comme $\mathcal{L}_{ur} \geq \mathcal{L}_{r}$, *RV* n'est jamais négatif, et généralement strictement positif. En calculant la statistique *RV* pour des modèles à réponse binaire, il est important de comprendre que la log-vraisemblance est toujours négative. Cela vient de l'équation (17.11) : comme y_i est soit zéro soit un, et les deux variables dans la fonction log sont strictement comprises entre zéro et un, cela veut dire que leurs logs naturels sont négatifs. Le fait que les deux fonctions de log-vraisemblance soient négatives ne change pas le calcul de la statistique *RV* ; il faut simplement garder les signes négatifs de l'équation (17.12).

La multiplication par deux dans (17.12) est nécessaire pour que *RV* ait une distribution du chi-deux sous H_0. Si nous testons q restrictions d'exclusion, $RV \overset{a}{\sim} \chi^2_q$. Cela veut dire que, pour tester H_0 au seuil de 5 %, nous utiliserons les valeurs critiques au 95^e percentile dans la distribution χ^2_q Le calcul des p-valeurs est facile dans la plupart des logiciels.

Pour aller plus loin 17.1

Un modèle probit prédit le fait qu'une entreprise se fasse racheter durant une année :

$$\begin{aligned} P(takeover = 1|\mathbf{x}) = \Phi(&\beta_0 + \beta_1 avgprof \\ &+ \beta_2 mktval \\ &+ \beta_3 debtearn \\ &+ \beta_4 ceoten \\ &+ \beta_5 ceosal \\ &+ \beta_6 ceoage), \end{aligned}$$

où *takeover* est une variable binaire, *avgprof* est la marge bénéficiaire des années précédentes, *mktval* est la valeur de marché de l'entreprise, *debtearn* est le ratio d'endettement sur les bénéfices, et *ceoten*, *ceosal*, et

ceoage sont la durée en poste, le salaire et l'âge du PDG. Formulons l'hypothèse nulle que, toutes choses égales par ailleurs, les variables liées au PDG ne changent pas la probabilité de rachat. Combien y a-t-il de *ddl* dans la distribution du chi-deux pour les tests du *RV* ou de Wald ?

Interpréter des estimations de logit et probit

Sur des ordinateurs récents, d'un point de vue pratique, les aspects les plus difficiles des modèles logit et probit sont la présentation et l'interprétation des résultats. Les coefficients estimés, leurs écarts-types estimés, ainsi que la valeur de la log-vraisemblance sont donnés par tous les logiciels qui estiment des logit et des probit, et devraient être donnés dans toute application. Les coefficients donnent les signes des effets marginaux de chaque x_j sur la probabilité de réponse, et x_j est statistiquement significatif quand on peut rejeter $H_0 : \beta_j = 0$ à un niveau de significativité suffisamment petit.

Comme nous en avons un peu discuté dans la section 7.5 pour le modèle à probabilité linéaire, le **pourcentage de prédictions correctes** permet de mesurer l'adéquation du modèle estimé. Comme précédemment, définissons une prédiction binaire sur y_i : un si la probabilité prédite est au moins 0,5, et zéro sinon. Mathématiquement, $\tilde{y}_i = 1$ si $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq 0{,}5$ et $\tilde{y}_i = 0$ si $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) < 0{,}5$. Étant donné les $\{\tilde{y}_i : i = 1, 2, \ldots, n\}$, nous pouvons voir si $\tilde{y}_i$ prédit correctement y_i pour toutes les observations. Il y a quatre possibilités pour chaque paire $(y_i, \tilde{y}_i)$; quand les deux sont zéro et quand les deux sont un, la prédiction est correcte. Dans les deux cas où l'une est zéro et l'autre est un, la prédiction est incorrecte. Le pourcentage de prédictions correctes est le pourcentage de fois où $\tilde{y}_i = y_i$.

Bien que le pourcentage de prédictions correctes soit une mesure d'adéquation du modèle estimé aux données utile, elle peut induire en erreur. En particulier, il est possible d'avoir un taux élevé de prédictions correctes même quand la variable prédite la plus rare est très mal prédite. Par exemple, supposons que $n = 200$, y_i est nul pour 160 observations, et sur ces 160 observations, $\tilde{y}_i$ est nul dans 140 cas (nous prédisons donc 87,5 % des $y_i = 0$). Même si *aucune* des prédictions n'est correcte quand $y_i = 1$, 70 % des prédictions sont toujours bonnes (140/200 = 0,7). Nous voulons souvent prédire correctement le cas le moins probable (comme quand quelqu'un est suspecté d'avoir commis un crime), et nous devrions donc montrer directement la qualité de la prédiction pour chaque valeur de y_i. Cela peut donc être justifié de calculer le pourcentage de prédictions correctes pour chaque y_i. Le Problème 1 demande de montrer que le pourcentage de prédictions correctes est une moyenne pondérée de $\hat{q}_0$ (le pourcentage de prédictions correctes quand $y_i = 0$) et $\hat{q}_1$ (le pourcentage de prédictions correctes quand $y_i = 1$), les poids étant respectivement les pourcentages de zéros et de uns dans l'échantillon.

Le seuil de 0,5 évoqué ci-dessus peut parfois être critiqué, notamment quand la variable dépendante est improbable. Supposons par exemple que $\bar{y} = 0{,}08$ (seulement 8 % de « positifs » dans l'échantillon), il se peut très bien que l'on ne prédise *jamais* $y_i = 1$ car la probabilité estimée ne dépasse jamais 0,5. Alternativement, on peut utiliser le pourcentage de positifs dans l'échantillon comme seuil $-0{,}08$ dans l'exemple précédent. En d'autres termes, on définit $\tilde{y}_i = 1$ quand $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq 0{,}08$ et zéro sinon. Cette règle va évidemment accroître le nombre de prédictions de positifs, mais cela a un coût : le nombre d'erreurs augmentera nécessairement dans la prédiction des zéros (des « négatifs »). En terme de pourcentage total de prédictions correctes, cela peut être bien moins efficace que d'utiliser le seuil à 0,5.

Une troisième option est de choisir un seuil tel que la part des $\tilde{y}_i = 1$ dans l'échantillon est $\bar{y}$. (Ou aussi proche que possible de $\bar{y}$.) En d'autres termes, cherchons le seuil $\tau, 0 < \tau < 1$, tel que si on définit $\tilde{y}_i = 1$ par $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq \tau$, alors $\sum_{i=1}^{n} \tilde{y}_i \approx \sum_{i=1}^{n} y_i$. (Le tâtonnement pour trouver la valeur désirée peut être fastidieux, mais reste faisable. Dans certains cas, il peut être impossible de faire en sorte que le nombre de positifs prédits soit exactement égal au nombre de positifs dans l'échantillon.) Maintenant, étant donné

l'ensemble des $\tilde{y}_i$, nous pouvons calculer le pourcentage de prédictions correctes pour chaque y_i ainsi que le pourcentage total de prédictions correctes.

Il existe également plusieurs mesures de pseudo *R*-carré pour les réponses binaires. McFadden (1974) suggère la mesure $1 - \mathcal{L}_{ur}/\mathcal{L}_o$, où $\mathcal{L}_{ur}$ est la log-vraisemblance du modèle estimé, et $\mathcal{L}_o$ est la log-vraisemblance du modèle avec seulement une constante. Pourquoi cette mesure a-t-elle un sens ? Souvenons-nous que les log-vraisemblances sont négatives, donc $\mathcal{L}_{ur}/\mathcal{L}_o = |\mathcal{L}_{ur}|/|\mathcal{L}_o|$. De plus, $|\mathcal{L}_{ur}| \leq |\mathcal{L}_o|$. Si les covariables n'ont aucun pouvoir explicatif, $\mathcal{L}_{ur}/\mathcal{L}_o = 1$, et le pseudo *R*-carré sera zéro, exactement comme dans les régressions linéaires où les covariables n'ont aucun pouvoir explicatif. Habituellement, $|\mathcal{L}_{ur}| < |\mathcal{L}_o|$, donc $1 - \mathcal{L}_{ur}/\mathcal{L}_o > 0$. Si $\mathcal{L}_{ur}$ était nul, le pseudo *R*-carré serait un. En réalité, $\mathcal{L}_{ur}$ ne peut être nul dans un modèle probit, puisque cela voudrait dire que la probabilité estimée est toujours un quand $\boldsymbol{y}_i = 1$ et toujours zéro quand $\boldsymbol{y}_i = 0$.

Les autres pseudo *R*-carrés pour les probit et logit sont plus directement liés au *R*-carré des estimations par MCO d'un modèle de probabilités linéaires. Pour un modèle logit ou probit, appelons probabilités prédites les $\hat{y}_i = G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$. Comme ces probabilités sont aussi une estimation de $\mathrm{E}(y_i|\mathbf{x}_i)$, nous pouvons baser un *R*-carré sur la proximité entre les $\hat{y}_i$ et les y_i. Une possibilité évidente au vu de l'analyse des régressions linéaires est de calculer le carré de la corrélation entre y_i et $\hat{y}_i$. Rappelez-vous : pour les régressions linéaires, c'est une autre manière d'obtenir le *R*-carré habituel ; voir l'équation (3.29). On peut donc calculer un pseudo *R*-carré pour les probit et logit qui soit directement comparable à celui des modèles à probabilités linéaires. Dans tous les cas, l'adéquation du modèle aux données est généralement moins importante que la mesure d'effets *ceteris paribus* convaincants des variables explicatives.

Il est souvent intéressant de mesurer l'effet de x_j sur la probabilité de réponse $\mathrm{P}(y = 1|\mathbf{x})$. Si x_j est (grossièrement) continue,

$$\Delta\hat{\mathrm{P}}(y = 1|\mathrm{x}) \approx [g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j]\Delta x_j, \qquad \textbf{[17.13]}$$

pour de « petites » variations de x_j. Donc, si $\Delta x_j = 1$, le changement dans la probabilité estimée d'avoir un positif est approximativement $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j$. En comparaison avec le modèle à probabilités linéaires, l'inconvénient de l'utilisation d'un modèle probit ou logit est que les effets marginaux dans l'équation (17.13) sont plus délicats à résumer à cause du facteur d'échelle $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$, qui dépend de x (c'est-à-dire de toutes les variables explicatives). Il est possible d'en donner les valeurs pour des valeurs intéressantes de x_j – comme les moyennes, médianes, maximums, quartiles inférieurs et supérieurs – et de voir comment $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$ évolue. Bien que cela semble intéressant, cela peut vite s'avérer fastidieux même avec un nombre modéré de variables explicatives.

Pour résumer rapidement l'ampleur des effets marginaux, le fait qu'un seul facteur d'échelle multiplie $\hat{\beta}_j$ pour obtenir les effets marginaux est pratique (au moins pour les variables approximativement continues). Une méthode fréquemment utilisée dans les logiciels d'économétrie qui estiment des modèles probit et logit est de remplacer les variables explicatives par leur moyenne. En d'autres termes, le facteur multiplicatif est

$$g(\hat{\beta}_0 + \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}) = g(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + \ldots + \hat{\beta}_k\bar{x}_k), \qquad \textbf{[17.14]}$$

où $g(\cdot)$ est la densité de la loi normale standardisée dans le cas du modèle probit et $g(z) = \exp(z)/[1 + \exp(z)]^2$ dans le cas du modèle logit. L'idée de départ est que (17.14) permet, après multiplication par les $\hat{\beta}_j$, d'obtenir l'effet marginal des x_j pour l'individu « moyen » de l'échantillon. En multipliant un coefficient par (17.14), on obtient donc en général un **effet marginal au point moyen (EMPM)**.

L'utilisation des EMPM pour résumer l'effet marginal des variables explicatives pose au moins deux problèmes. Tout d'abord, s'il y a des variables explicatives discrètes, leur moyenne ne représente personne dans

l'échantillon (ou dans la population, en l'occurrence). Par exemple, si $x_1 = femme$ et 47,5 % de l'échantillon sont des femmes, quel sens y a-t-il à faire des calculs en $\bar{x}_1 = 0{,}475$ pour représenter un individu « moyen » ? Ensuite, si une variable explicative continue apparaît de manière non linéaire – par exemple, avec un log ou un carré – il n'est pas clair que nous voulions la moyenne de la fonction non linéaire ou la moyenne de la variable de départ. Par exemple, devrions-nous utiliser $\log(\overline{sales})$ ou $\log(sales)$ pour la taille moyenne de l'entreprise ? Par défaut, les logiciels d'économétrie donnent la première forme pour calculer le facteur d'échelle de (17.14) : le logiciel calcule la moyenne des variables incluses dans l'équation probit ou logit.

Une autre approche contourne la question du choix des valeurs de covariables pour mesurer le facteur d'échelle. Ce second facteur d'échelle consiste à calculer la moyenne des effets marginaux sur l'échantillon, ce qui mène à un **effet marginal moyen** (**EMM**), aussi appelé **effet partiel moyen** (**EPM**). Pour une variable continue x_j, l'effet marginal moyen est $n^{-1}\sum_{i=1}^{n}[g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})\hat{\beta}_j] = \left[n^{-1}\sum_{i=1}^{n} g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})\right]\hat{\beta}_j$. Le terme multipliant $\hat{\beta}_j$ agit comme un facteur d'échelle :

$$n^{-1}\sum_{i=1}^{n} g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}). \qquad \textbf{[17.15]}$$

L'équation (17.15) se calcule facilement après une estimation logit ou probit, avec $g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) = \phi(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$ dans le cas probit et $g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) = \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})/[1 + \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})]^2$ dans le cas logit. Les deux facteurs d'échelle diffèrent – potentiellement de beaucoup – puisque (17.15) mesure la moyenne d'une fonction non linéaire au lieu de cette fonction non linéaire prise à la moyenne [comme dans (17.14)].

Comme ces deux facteurs d'échelle dépendent de l'approximation donnée en (17.13), aucun n'a réellement de sens pour une variable explicative discrète. Au lieu de cela, il est plus logique d'utiliser l'équation (17.9) pour estimer le changement de probabilités. En changeant x_k de c_k à $c_k + 1$, l'analogue discret d'un effet partiel basé sur (17.14) est

$$G[\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_{k-1}\bar{x}_{k-1} + \hat{\beta}_k(c_k + 1)]$$
$$- G(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_{k-1}\bar{x}_{k-1} + \hat{\beta}_k c_k), \qquad \textbf{[17.16]}$$

où G est la fr de la loi normale standardisée dans le cas probit et $G(z) = \exp(z)/[1 + \exp(z)]$ dans le cas logit. L'effet marginal moyen, qui est généralement plus comparable aux estimations de MPL est

$$n^{-1}\sum_{i=1}^{n}\{G[\hat{\beta}_0 + \hat{\beta}_1\bar{x}_{i1} + \ldots + \hat{\beta}_{k-1}\bar{x}_{ik-1} + \hat{\beta}_k(c_k + 1)]$$
$$- G(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_{i1} + \ldots + \hat{\beta}_{k-1}\bar{x}_{ik-1} + \hat{\beta}_k c_k)\}. \qquad \textbf{[17.17]}$$

La quantité de l'équation (17.17) est un effet « partiel » car toutes les variables explicatives à part x_k sont fixées à leurs valeurs observées. Ce n'est pas nécessairement un effet « marginal », puisqu'une augmentation de x_k de c_k à $c_k + 1$ n'est pas forcément « marginale » (ou « petite ») ; cela dépend de la définition de x_k. Retrouver l'expression de (17.17) pour un logit ou un probit est relativement simple. Pour chaque observation, il faut tout d'abord calculer la probabilité de succès pour les deux valeurs de x_k, en prenant les valeurs de l'échantillon pour les autres variables explicatives (Nous aurons ainsi n différences estimées). Ensuite, il faut calculer la moyenne sur tout l'échantillon des différences entre les probabilités estimées. Pour des x_k binaires, (17.16) et (17.17) se calculent facilement avec plusieurs logiciels d'économétrie, dont Stata.®

L'expression (17.17) a une interprétation particulièrement simple quand x_k est une variable binaire. Pour chaque observation i, nous estimons la différence entre la probabilité que $y_i = 1$ quand $x_k = 1$ et $x_k = 0$, soit :

$$G(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1} x_{i,k-1} + \hat{\beta}_k) - G(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1} x_{i,k-1}).$$

Pour tout i, cette différence est l'effet estimé de changer x_k de zéro à un, que l'observation i ait $x_{ik} = 1$ ou $x_{ik} = 0$. Par exemple, supposons qu'y est une variable indicatrice d'emploi (égale à un si la personne est employée) après participation à un programme de formation à la recherche d'emploi, indicatrice notée x_k. Nous estimons la différence de probabilité d'emploi pour chaque personne dans les deux états (programme de formation ou pas). Ce *raisonnement contrefactuel* est similaire à celui du chapitre 16, que nous avions utilisé pour introduire les modèles à équations simultanées. L'effet du programme de formation estimé sur la probabilité d'emploi est la moyenne des différences de probabilités estimées. Pour donner un autre exemple, supposons qu'y indique si un prêt a été accordé à une famille en fonction du revenu, de la richesse, des autres crédits, et ainsi de suite – ce seraient les éléments de $(x_{i1}, x_{i2}, \ldots, x_{i,k-1})$ –, et si le chef de ménage est blanc ou non-blanc. Nous espérons avoir suffisamment pris en compte les autres éléments pour que l'effet estimé soit l'effet de l'origine ethnique sur les chances d'obtenir un prêt.

Dans les applications avec un probit, logit, et un MPL, pour comparer des estimations, le plus logique est de décrire le facteur d'échelle ci-dessus. Il est néanmoins utile d'avoir une méthode plus rapide pour comparer la valeur des coefficients. Comme dit plus haut, pour le probit, $g(0) \approx 0{,}4$, alors que pour le logit, $g(0) = 0{,}25$. Pour rendre probit et logit grossièrement comparables, il faut multiplier les coefficients du probit par $0{,}4/0{,}25 = 1{,}6$, ou multiplier les coefficients du logit par 0,625. Dans le MPL, $g(0)$ est exactement un, donc les coefficients des logits peuvent être divisés par quatre pour être comparés à des MPL ; les pentes des probits peuvent être divisées par 2,5 pour être comparables aux MPL. Dans la plupart des cas, il est cependant nécessaire de faire les comparaisons exactes, que l'on peut trouver en utilisant les facteurs d'échelle donnés en (17.15) pour les logit et probit.

EXEMPLE 17.1
Participation des femmes mariées à la force de travail

Nous utilisons maintenant les données contenues dans MROZ pour estimer un modèle prédisant la participation à la force de travail tiré de l'exemple 8.8 – voir aussi la section 7.5 – avec un logit et un probit. Les estimations suivant un modèle de probabilités linéaires tirées de l'exemple 8.8 sont aussi données, en utilisant des écarts-types estimés robustes à l'hétéroscédasticité. Les résultats sont donnés dans le tableau 17.1, avec les écarts-types estimés entre parenthèses.

Les estimations des trois modèles racontent une histoire similaire. Les signes des coefficients sont les mêmes dans les trois modèles, et les variables statistiquement significatives sont les mêmes. Le pseudo R-carré rapporté pour le MPL est le R-carré habituel des MCO ; pour le logit et le probit, le pseudo R-carré est celui basé sur les log-vraisemblances détaillé plus haut.

Comme nous l'avons déjà mentionné, la *taille* des coefficients estimés n'est pas directement comparable entre les modèles. Au lieu de cela, calculons les facteurs d'échelle des équations (17.14) et (17.15). Si nous évaluons la densité de probabilité de la loi normale standardisée $\phi(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k)$ à la moyenne de l'échantillon pour les variables explicatives (en incluant les moyennes de *exper*^{2}, *kidslt6*, et *kidsge6*), le résultat est approximativement 0,391. Quand on calcule (17.14) pour le logit, on trouve environ 0,243. Le rapport entre les deux $0{,}391/0{,}243 \approx 1{,}61$ est très proche de la règle empirique pour comparer les estimations d'un probit et d'un logit : multiplier les estimations d'un probit par 1,6. En revanche, pour comparer les probit et logit à un MPL, il vaut mieux utiliser (17.15). Ces facteurs d'échelle sont approximativement 0,301 (probit) et 0,179 (logit). Par exemple, l'effet marginal moyen d'*educ* mesuré par le logit est environ $0{,}179(0{,}221) \approx 0{,}040$, et celui mesuré par le probit est presque $0{,}301(0{,}131) \approx 0{,}039$, les deux sont très proches de l'estimation par MPL, qui est 0,038. Même pour la variable discrète *kidslt6*, les effets marginaux moyens mesurés par les logit et probit sont proches des coefficients du MPL (–0,262). Ils sont $0{,}179(-1{,}443) \approx -0{,}258$ (logit) et $0{,}301(-0{,}868) \approx -0{,}261$ (probit).

Le tableau 17.2 donne l'effet partiel moyen de toutes les variables explicatives et pour les trois modèles estimés. Nous avons obtenu les estimations et les écarts-types avec le logiciel Stata®. Ces EMM traitent toutes les variables explicatives comme des variables continues, même le nombre d'enfants. Le calcul de l'EMM de *exper* demande d'être un peu attentif, car il faut prendre en compte la forme fonctionnelle quadratique d'*exper*. Dans le cas d'un modèle linéaire, il faut également calculer la dérivée et en calculer la moyenne. Dans la colonne du MPL, l'EMM d'*exper* est la moyenne de la dérivée par rapport à *exper*, soit 0,039 – 0,012 *exper*, sur toutes les observations *i*. (Les autres EMM de la colonne du MPL sont simplement les coefficients des MCO du tableau 17.1.) Les EMM d'*exper* pour les modèles logit et probit prennent également en compte la forme quadratique d'*exper*. Comme ce tableau le montre clairement, les EMM et leur significativité statistique sont très similaires pour toutes les variables explicatives pour ces trois modèles.

Pour aller plus loin 17.2

En utilisant le modèle probit et des approximations dans le calcul, quelle est le changement de la probabilité prédite quand *exper* augmente de 10 à 11 ?

La plus grande différence entre le modèle MPL et les modèles logit et probit est que le MPL fait l'hypothèse que les effets marginaux sont *constants*, pour *educ*, *kidslt6*, etc. alors que les modèles logit et probit impliquent que les effets marginaux ont des amplitudes décroissantes. Dans le MPL, un enfant en plus réduit la probabilité de participation à la force de travail d'environ 0,262, quel que soit le nombre d'enfants d'une femme (et quelles que soient les autres variables explicatives). Nous pouvons voir que les effets marginaux estimés avec un probit fonctionnent différemment. Pour être concrets, prenons une femme avec *nwifeinc* = 20,13, *educ* = 12,3, *exper* = 10,6, et *age* = 42,5 – qui sont proches de la moyenne de l'échantillon – et *kidsge6* = 1. Quelle est la diminution de la probabilité de travailler s'il y a un jeune enfant au lieu de zéro ? Évaluons la fr de la loi normale standardisée, $\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k)$, avec *kidslt6* = 1 et avec *kidslt6* = 0, et les autres variables explicatives données plus haut. Nous trouvons approximativement 0,373 – 0,707 = –0,334, ce qui veut dire que la participation à la force de travail est plus basse d'environ 0,334, quand une femme a un jeune enfant. Si une femme passe d'un à deux jeunes enfants, la probabilité diminue encore, mais l'effet marginal est plus faible : 0,117 – 0,373 = –0,256. Remarquez que l'effet marginal du modèle de probabilités linéaires, supposé mesurer l'effet aux alentours de la moyenne, se trouve entre ces deux estimations (Par ailleurs, les calculs donnés ici utilisent généralement des coefficients arrondis à trois chiffres après la virgule et différeront légèrement de ceux donnés par un logiciel de statistiques qui fait moins d'erreurs d'arrondi).

Tableau 17.1 *Prédiction la participation des femmes à la force de travail : coefficients des modèles MPL, logit, et probit*

Variable dépendante : *inlf*			
Variables Explicatives	**MPL (MCO)**	**Logit (EMV)**	**Probit (EMV)**
nwifeinc	–0,0034 (0,0015)	–0,021 (0,008)	–0,012 (0,005)
educ	0,038 (0,007)	0,221 (0,043)	0,131 (0,025)
exper	0,039 (0,006)	0,206 (0,032)	0,123 (0,019)
$exper^2$	–0,00060 (0,00018)	–0,0032 (0,0010)	–0,0019 (0,0006)

Variable dépendante : *inlf*			
Variables Explicatives	**MPL (MCO)**	**Logit (EMV)**	**Probit (EMV)**
age	−0,016 (0,002)	−0,088 (0,015)	−0,053 (0,008)
kidslt6	−0,262 (0,032)	−1,443 (0,204)	−0,868 (0,119)
kidsge6	0,013 (0,013)	0,060 (0,075)	0,036 (0,043)
constant	0,586 (0,151)	0,425 (0,860)	0,270 (0,509)
Pourcentage de Prédictions Correctes Log-Vraisemblance Pseudo *R*-carré	73,4 – 0,264	73,6 −401,77 0,220	73,4 −401,30 0,221

Tableau 17.2 *Effets marginaux moyens pour la participation à la force de travail*

Variables explicatives	**MPL**	**Logit**	**Probit**
nwifeinc	−0,0034 (0,0015)	−0,0038 (0,0015)	−0,0036 (0,0014)
educ	0,038 (0,007)	0,039 (0,007)	0,039 (0,007)
exper	0,027 (0,002)	0,025 (0,002)	0,026 (0,002)
age	−0,016 (0,002)	−0,016 (0,002)	−0,016 (0,002)
kidslt6	−0,262 (0,032)	−0,258 (0,032)	−0,261 (0,032)
kidsge6	0,013 (0,014)	0,011 (0,013)	0,011 (0,013)

La figure 17.2 illustre les différences de probabilité de réponse estimée entre un modèle de réponse binaire non linéaire et le modèle de probabilités linéaires. La figure donne la probabilité de participation à la force de travail en fonction de la durée de scolarisation pour les modèles de probabilités linéaires et probit (Le graphique pour le modèle logit est très proche de celui du modèle probit). Dans les deux cas, les variables explicatives autres qu'*educ* sont fixées à leurs moyennes de l'échantillon. Ici, les équations des courbes de la figure sont $\widehat{inlf} = 0{,}102 + 0{,}038\ educ$ pour le modèle linéaire et $\widehat{inlf} = \Phi(-1{,}403 + 0{,}131\ educ)$ pour le modèle probit. Aux niveaux d'instruction les plus faibles, le modèle de probabilités linéaires estime des chances de participation à la force de travail plus élevées que le probit. Par exemple, pour huit années d'instruction, le modèle de probabilités linéaires estime une probabilité de participation à la force de travail de 0,406 alors que le probit estime une probabilité d'environ 0,361. Les estimations sont les mêmes aux alentours de 11,33 années d'instruction. Aux niveaux d'instruction les plus élevés, le modèle probit donne des probabilités

de participation à la force de travail plus élevées. Dans l'échantillon, le nombre minimal d'années d'instruction est 5 et le plus grand est 17, on ne peut donc pas vraiment faire de comparaison hors de cet intervalle.

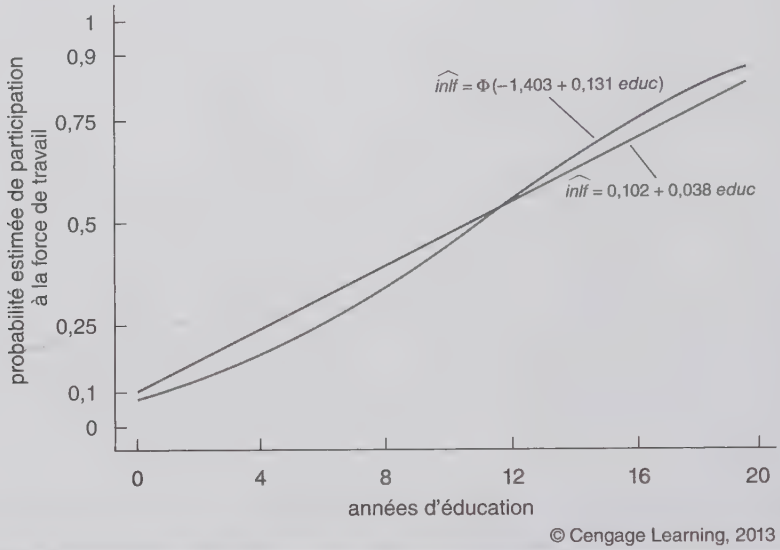


Figure 17.2 *Probabilités de réponse estimée en fonction du niveau d'instruction pour des modèles de probabilités linéaires et probit.*

Pour les modèles logit et probit, des problèmes concernant l'endogénéité des variables explicatives apparaissent également, de manière semblable aux modèles linéaires. Il est possible de tester et de corriger l'endogénéité de variables explicatives en utilisant des méthodes ressemblant aux doubles moindres carrés, même si nous ne les couvrirons pas ici. Evans et Schwab (1995) estiment un modèle probit prédisant si un étudiant va au *college* (NDT : après le lycée), où la variable explicative principale est une variable indicatrice disant si l'étudiant est allé dans un lycée catholique. Evans et Schwab estiment un modèle qui permet de considérer le fait d'être allé dans un lycée catholique comme endogène, par maximum de vraisemblance [Voir Wooldridge (2010, chapitre 15) pour une explication de ces techniques].

Deux autres problèmes liés aux modèles probit ont été étudiés. Le premier est la non normalité de la variable e dans le modèle à variable latente (17.6). Il est évident que si e ne suit pas une loi normale standardisée, la probabilité de réponse n'aura pas une forme probit. Certains auteurs mettent l'accent sur le fait que cela mène à une estimation biaisée de β_j, mais c'est une préoccupation accessoire sauf si on ne s'intéresse qu'au signe des effets. Comme la probabilité de réponse est inconnue, nous ne pourrions pas estimer la valeur des effets marginaux même avec une estimation convergente des β_j.

Le second problème de spécification, également défini en fonction de la variable latente, est l'hétéroscédasticité de e. Si $\text{Var}(e|\mathbf{x})$ dépend de $\mathbf{x}$, la probabilité de réponse n'a plus la forme $G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$; au contraire, elle dépend de la forme de la variance de e et le modèle nécessite une autre forme d'estimation. Ce genre de modèle est rarement utilisé en pratique, car les modèles logit et probit avec des formes fonctionnelles flexibles des variables explicatives marchent généralement bien.

Les modèles à réponse binaire s'appliquent avec peu de modifications à des données en coupe empilées ou à des données où les observations sont indépendantes mais pas nécessairement identiquement distribuées. Des variables indicatrices d'années ou de date sont souvent ajoutées pour prendre en compte les effets temporels agrégés. Exactement comme les modèles linéaires, les modèles logit et probit peuvent être utilisés pour mesurer l'effet de certaines politiques dans le contexte d'expériences naturelles.

Le modèle de probabilités linéaires peut s'appliquer aux données de panel ; il serait généralement estimé avec des effets fixes (voir le chapitre 14). Les modèles logit et probit à effets inobservés sont récemment devenus populaires. Ces modèles sont rendus complexes par la non linéarité des probabilités de réponses, ils sont difficiles à estimer et à interpréter [Voir Wooldridge (2010, chapitre 15)].

17.2 LE MODÈLE TOBIT POUR DES RÉPONSES AVEC SOLUTION EN COIN

Comme nous l'avons mentionné dans le chapitre introductif, les réponses avec des solutions en coin sont une autre forme importante de variable dépendante limitée. Ce type de variable est nul pour une part non négligeable de la population et approximativement continûment distribué sur les valeurs positives. La dépense individuelle en alcool des individus un mois donné pourrait être un exemple. Parmi la population âgée de plus de 21 ans aux États-Unis, cette variable est souvent nulle. Pour une part non négligeable des personnes, la dépense en alcool est nulle. Cette section ignore la vérification de certains détails sur le modèle Tobit [Ils sont donnés dans Wooldridge (2010, chapitre 17)].

Soit y une variable fondamentalement continue sur les valeurs strictement positives mais qui prend la valeur zéro avec une probabilité strictement positive. Rien n'empêche d'utiliser un modèle linéaire pour y. Un modèle linéaire peut en effet être une bonne approximation de $E(y|x_1, x_2, \ldots, x_k)$, surtout si x_j est proche des valeurs moyennes. Mais nous risquerions d'obtenir des valeurs prédites négatives pour y ; ce qui est analogue aux problèmes avec des MPL pour des variables dépendantes binaires. De plus, les variables explicatives apparaissant en niveaux ont un effet partiel sur $E(y|\mathbf{x})$ constant par hypothèse, ce qui peut être trompeur. En pratique, $Var(y|\mathbf{x})$ serait probablement hétéroscédastique, bien que nous puissions facilement traiter l'hétéroscédasticité en calculant des écarts-types estimés et des statistiques de test robustes. Comme la distribution de y a un point de masse en zéro, y ne suit pas une loi normale. Donc l'inférence doit être basée sur une justification asymptotique, comme dans le cas du modèle de probabilités linéaires.

Dans certains cas, il est important d'avoir un modèle qui ne prédit pas de valeurs négatives pour y et dans lequel les effets partiels dépendent des variables explicatives. De plus, on peut vouloir estimer d'autres aspects de la distribution de y conditionnellement à $x_1, \ldots, x_k$ que l'espérance conditionnelle. Le **modèle Tobit** est relativement pratique pour ces objectifs. Le modèle Tobit exprime les réponses observées, y, en fonction d'une variable latente sous-jacente :

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u, \; u|\mathbf{x} \sim \text{Normale}(0, \sigma^2) \qquad \textbf{[17.18]}$$

$$y = \max(0, y^*). \qquad \textbf{[17.19]}$$

La variable latente y^* satisfait l'hypothèse classique de linéarité ; en particulier, elle a une distribution normale et homoscédastique et une espérance conditionnelle linéaire. L'équation (17.19) implique que la variable observée, y, vaut y^* quand $y^* \geq 0$, mais $y = 0$ quand $y^* < 0$. Comme y^* suit une loi normale, y a une distribution continue sur les valeurs strictement positives. En particulier, la densité de y conditionnellement à $\mathbf{x}$ est la même que celle de y^* conditionnellement à $\mathbf{x}$ pour les valeurs positives. De plus,

$$\begin{aligned} P(y = 0|\mathbf{x}) &= P(y^* = 0|\mathbf{x}) = P(u < -\mathbf{x}\boldsymbol{\beta}|\mathbf{x}) \\ &= P(u/\sigma < -\mathbf{x}\boldsymbol{\beta}/\sigma|\mathbf{x}) = \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma) = 1 - \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma), \end{aligned}$$

puisque u/σ suit une distribution normale standardisée et est indépendante de $\mathbf{x}$. Nous avons inclus la constante dans $\mathbf{x}$ pour simplifier la notation. En conséquence, si $(\mathbf{x}_i, y_i)$ est tiré aléatoirement dans la population, la densité de y_i conditionnellement à $\mathbf{x}_i$ s'écrit

$$(2\pi\sigma^2)^{-1/2} \exp[-(y - \mathbf{x}_i\boldsymbol{\beta})^2/(2\sigma^2)] = (1/\sigma)\phi[(y - \mathbf{x}_i\boldsymbol{\beta})/\sigma], \; y > 0 \qquad \textbf{[17.20]}$$

$$P(y_i = 0|\mathbf{x}_i) = 1 - \Phi(\mathbf{x}_i\beta/\sigma), \quad [17.21]$$

où $\phi(.)$ est la fonction de densité de la loi normale standardisée.

Nous pouvons obtenir la fonction de log-vraisemblance de chaque observation i à partir de (17.20) et (17.21) :

$$\ell_i(\beta,\sigma) = 1(y_i = 0)\log[1 - \Phi(\mathbf{x}_i\beta/\sigma)] + 1(y_i > 0)\log\{(1/\sigma)\phi[(y_i - \mathbf{x}_i\beta/\sigma)]\}; \quad [17.22]$$

remarquez qu'elle dépend de σ, l'écart-type de u, ainsi que de β_j. La log-vraisemblance d'un échantillon aléatoire de taille n est obtenue en calculant la somme de (17.22) sur tous les i. Les estimateurs du maximum de vraisemblance de β et σ sont obtenus en maximisant la log-vraisemblance ; cela nécessite des méthodes numériques, bien que ce soit généralement facile à faire avec des programmes pré-codés.

Comme dans le cas des logit et probit, chaque estimation de Tobit a un écart-type d'échantillonnage, et ces écarts-types peuvent être estimés et utilisés pour construire des statistiques t pour chaque $\hat{\beta}_j$; la formule matricielle utilisée pour trouver les écarts-types estimés est compliquée et n'est pas présentée ici [Voir, par exemple, Wooldridge (2010, chapitre 17)].

Il est facile de faire des tests de restrictions d'exclusion multiples avec un test de Wald ou du rapport de vraisemblance.

Le test de Wald a une forme similaire à celle des modèles logit ou probit ; le test *RV* est toujours donné par (17.12), en remplaçant bien entendu les fonctions de log-vraisemblance par celle du Tobit des modèles restreint et non restreint.

Pour aller plus loin 17.3

Soit y le nombre de liaisons extraconjugales d'une femme mariée dans la population américaine ; nous voudrions expliquer cette variable avec les autres caractéristiques de la femme – par exemple, si elle travaille hors du foyer, son mari et sa famille. Est-ce un bon exemple pour un modèle Tobit ?

Interpréter les estimations du modèle Tobit

Avec les ordinateurs modernes, il n'est généralement guère plus difficile d'obtenir une estimation d'un modèle Tobit par maximum de vraisemblance que d'obtenir l'estimateur des MCO d'un modèle linéaire. De plus, les résultats d'un modèle Tobit et des MCO sont souvent similaires. Cela pourrait donner envie d'interpréter les coefficients d'un modèle Tobit comme ceux d'une régression linéaire. Les choses ne sont malheureusement pas si simples.

L'équation (17.18) nous permet de voir que les β_j mesurent les effets partiels des x_j sur $E(y^*|\mathbf{x})$, y^* étant la variable latente. Parfois, y^* a un sens économique, mais plus souvent, elle n'en a pas. La variable que nous cherchons à expliquer est y, la variable dépendante observée (comme le temps de travail, ou le montant des dons aux œuvres de charité). Par exemple, nous pourrions être intéressés par la sensibilité du temps de travail aux taux d'imposition marginaux dans une perspective de politique publique.

Nous pouvons estimer $P(y = 0|\mathbf{x})$ à partir de (17.21) ; cela permet évidemment d'estimer $P(y > 0|\mathbf{x})$. Que se passe-t-il si nous cherchons à estimer l'espérance de y en fonction de $\mathbf{x}$? Dans les modèles Tobit, il est utile de séparer deux espérances : $E(y|y > 0, \mathbf{x})$, qui est parfois appelée « espérance conditionnelle » car elle est conditionnelle à $y > 0$, et $E(y|\mathbf{x})$, qui est malheureusement parfois appelée « espérance non conditionnelle » (Elle est conditionnelle aux variables explicatives). L'espérance nous donne, pour des valeurs données

de $\mathbf{x}$, l'espérance de y dans la sous-population où y est positive. Nous pouvons facilement calculer $E(y|\mathbf{x})$ en fonction de $E(y|y > 0,\mathbf{x})$:

$$E(y|\mathbf{x}) = P(y > 0|\mathbf{x}) \cdot E(y|y > 0,\mathbf{x}) = \Phi(\mathbf{x}\beta/\sigma) \cdot E(y|y > 0,\mathbf{x}). \qquad [17.23]$$

Nous utilisons un résultat propre aux variables à distribution normale pour trouver $E(y|y > 0,\mathbf{x})$: si $z \sim \text{Normale}(0,1)$, alors $E(z|z > c) = \phi(c)/[1 - \Phi(c)]$ pour toute constante c. Mais $E(y|y > 0,\mathbf{x}) = \mathbf{x}\beta + E(u|u > -\mathbf{x}\beta) = \mathbf{x}\beta + \sigma E[(u/\sigma)|(u/\sigma) > -\mathbf{x}\beta] = \mathbf{x}\beta + \sigma\phi(\mathbf{x}\beta/\sigma)/\Phi(\mathbf{x}\beta/\sigma)$, car $\phi(-c) = \phi(c)$, $1 - \Phi(-c) = \Phi(c)$, et u/σ suit une loi normale standardisée indépendamment de $\mathbf{x}$.

Nous pouvons résumer ceci :

$$E(y|y > 0,\mathbf{x}) = \mathbf{x}\beta + \sigma\lambda(\mathbf{x}\beta/\sigma), \qquad [17.24]$$

où $\lambda(c) = \phi(c)/\Phi(c)$ est appelé **ratio de Mills inversé** ; c'est le rapport entre la densité de la loi normale standardisée et sa fonction de répartition, toutes deux évaluées en c.

L'équation (17.24) est importante : elle montre que l'espérance de y conditionnelle à $y > 0$ est égale à $\mathbf{x}\beta$ plus un terme strictement positif, qui est σ fois l'inverse du ratio de Mills évalué en $\mathbf{x}\beta/\sigma$. Cette équation montre aussi qu'utiliser les MCO uniquement sur les observations pour lesquelles $y_i > 0$ ne mesurera pas toujours β de manière convergente ; l'inverse du ratio de Mills est fondamentalement une variable omise, et est corrélée aux éléments de $\mathbf{x}$ dans le cas général.

En combinant (17.23) et (17.24), on obtient

$$E(y|\mathbf{x}) = \Phi(\mathbf{x}\beta/\sigma)[\mathbf{x}\beta + \sigma\lambda(\mathbf{x}\beta/\sigma)] = \Phi(\mathbf{x}\beta/\sigma)\mathbf{x}\beta + \sigma\phi(\mathbf{x}\beta/\sigma), \qquad [17.25]$$

où la seconde égalité vient du fait que $\Phi(\mathbf{x}\beta/\sigma)\lambda(\mathbf{x}\beta/\sigma) = \phi(\mathbf{x}\beta/\sigma)$. Cette équation montre que si y suit un modèle Tobit, $E(y|\mathbf{x})$ est une fonction non linéaire de $\mathbf{x}$ et β. Bien que ce ne soit pas évident, on peut montrer que le terme de droite de l'équation (17.25) est positif pour toute valeur de $\mathbf{x}$ et β. En conséquence, une fois que nous avons une estimation de β, nous pouvons être sûrs que les valeurs prédites pour y – c'est-à-dire l'estimation de $E(y|\mathbf{x})$ – sont positives. L'inconvénient d'éviter des prédictions d'y négatives est que l'équation (17.25) est plus compliquée qu'un modèle linéaire prédisant $E(y|\mathbf{x})$. Surtout, les effets partiels de (17.25) sont plus compliqués que dans un modèle linéaire. Comme nous le verrons, l'effet partiel de x_j sur $E(y|y > 0, \mathbf{x})$ et $E(y|\mathbf{x})$ ont le même signe que le coefficient β_j, mais le signe des effets dépend de la valeur de *toutes* les variables explicatives. Comme σ apparaît dans (17.25), il n'est pas étonnant que les effets partiels dépendent aussi de σ.

Si x_j est une variable continue, les effets partiels peuvent être calculés. Tout d'abord,

$$\partial E(y|y > 0,\mathbf{x})/\partial x_j = \beta_j + \beta_j \cdot \frac{d\lambda}{dc}(\mathbf{x}\beta/\sigma),$$

en supposant que x_j ne soit pas une fonction des autres régresseurs. En différenciant $\lambda(c) = \phi(c)/\Phi(c)$ et en utilisant $d\Phi/dc = \phi(c)$ et $d\phi/dc = -c\phi(c)$, on peut montrer que $d\lambda/dc = -\lambda(c)[c + \lambda(c)]$. En conséquence,

$$\partial E(y|y > 0,\mathbf{x})/\partial x_j = \beta_j\{1 - \lambda(\mathbf{x}\beta/\sigma)[\mathbf{x}\beta/\sigma + \lambda(\mathbf{x}\beta/\sigma)]\}. \qquad [17.26]$$

Cela montre que l'effet partiel de x_j sur $E(y|y > 0, \mathbf{x})$ n'est pas seulement une fonction de β_j. Le facteur d'ajustement est le terme entre accolades, $\{\cdot\}$, et dépend d'une fonction linéaire de $\mathbf{x}$, $\mathbf{x}\beta/\sigma = (\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)/\sigma$. On peut démontrer que ce facteur d'ajustement est toujours strictement compris entre 0 et 1, En pratique, (17.26) peut s'estimer en utilisant les estimations de β_j et σ par EMV. Comme dans les modèles logit et probit, il faut aussi avoir des valeurs des x_j, par exemple les valeurs moyennes ou d'autres valeurs d'intérêt. L'équation (17.26) montre un détail subtil qui est parfois oublié quand on utilise les modèles Tobit pour des réponses avec solution en coin : le paramètre σ apparaît directement dans les effets

partiels, avoir une estimation de σ est donc nécessaire pour l'estimation des effets partiels. L'estimation de σ pourrait sembler accessoire. Bien que la valeur de σ n'affecte pas les signes des effets partiels, elle affecte la valeur des effets, et l'importance économique de l'effet des variables explicatives n'est pas accessoire. En conséquence, le paramètre σ n'est pas accessoire ; si certains considèrent ce paramètre comme accessoire, cela vient d'une confusion entre le modèle Tobit avec des solutions en coin et les applications à une censure des données (voir la section 17.4 pour ce dernier cas).

Toutes les quantités économiques habituelles, comme les élasticités, peuvent être calculées. Par exemple, l'élasticité de y par rapport à x_1, conditionnellement à $y > 0$, est

$$\frac{\partial \mathrm{E}(y|y>0,\mathbf{x})}{\partial x_1} \cdot \frac{x_1}{\mathrm{E}(y|y>0,\mathbf{x})}. \qquad [17.27]$$

Cela peut être calculé quelle que soit la forme fonctionnelle de x_1 : linéaire, logarithmique ou quadratique, par exemple.

Si x_1 est une variable binaire, l'effet d'intérêt est la différence entre $\mathrm{E}(y|y > 0, \mathbf{x})$ avec $x_1 = 1$ et avec $x_1 = 0$. Les effets partiels d'autres variables discrètes (comme le nombre d'enfants) peuvent être calculés de manière similaire.

(17.25) permet de trouver la dérivée partielle de $\mathrm{E}(y|\mathbf{x})$ par rapport à un x_j continu. Cette dérivée prend en compte le fait que des gens commençant avec $y = 0$ peuvent choisir $y > 0$ quand x_j change :

$$\frac{\partial \mathrm{E}(y|\mathbf{x})}{\partial x_j} = \frac{\partial \mathrm{P}(y>0|\mathbf{x})}{\partial x_j} \cdot \mathrm{E}(y|y>0,\mathbf{x}) + \mathrm{P}(y>0|\mathbf{x}) \cdot \frac{\partial \mathrm{E}(y|y>0,\mathbf{x})}{\partial x_j} \qquad [17.28]$$

Comme $\mathrm{P}(y > 0|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$,

$$\frac{\partial \mathrm{P}(y>0|\mathbf{x})}{\partial x_j} = (\beta_j/\sigma)\phi(\mathbf{x}\boldsymbol{\beta}/\sigma), \qquad [17.29]$$

et nous pouvons donc estimer chaque terme de (17.28), une fois que nous avons les EMV de β_j ainsi que des valeurs particulières de x_j.

De manière remarquable, si nous remplaçons (17.26) et (17.29) dans (17.28), et en utilisant le fait que $\Phi(c)\lambda(c) = \phi(c)$ pour tout c, nous trouvons

$$\frac{\partial \mathrm{E}(y|\mathbf{x})}{\partial x_j} = \beta_j \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma). \qquad [17.30]$$

L'équation (17.30) nous permet de comparer grossièrement des estimations par MCO et venant d'un Tobit [L'équation (17.30) peut également être tirée directement de (17.25) en utilisant l'égalité $d\phi(z)/dz = -z\phi(z)$.]. Appelons $\hat{\gamma}_j$, les coefficients de la régression de y_i sur x_{i1}, x_{i2}, ..., x_{ik} par les MCO pour $i = 1, \ldots, n$, c'est-à-dire sur tout l'échantillon. Ces $\hat{\gamma}_j$, sont une estimation directe de $\partial \mathrm{E}(y|\mathbf{x})/\partial x_j$. Pour les rendre comparables au coefficient Tobit $\hat{\beta}_j$, il faut multiplier $\hat{\beta}_j$ par un facteur d'ajustement.

Comme dans les cas logit et probit, il y a deux approches habituelles pour trouver le facteur d'ajustement donnant des effets partiels – du moins pour les variables explicatives. Elles viennent toutes deux de l'équation (17.30). Premièrement, l'effet marginal au point moyen, EMPM, est obtenu en calculant $\Phi(\mathbf{x}\hat{\boldsymbol{\beta}}/\hat{\sigma})$, que nous noterons $\Phi(\bar{\mathbf{x}}\hat{\boldsymbol{\beta}}/\hat{\sigma})$. Nous pouvons alors utiliser ce facteur pour multiplier les coefficients des variables explicatives continues. L'EMPM a les mêmes inconvénients que pour les modèles logit et probit : nous ne sommes pas forcément intéressés par un effet pour une personne « moyenne », cette moyenne n'est pas toujours intéressante et peut même ne pas avoir de sens. Par ailleurs, il faut choisir entre faire la moyenne des fonctions non linéaires ou prendre les fonctions non linéaires au point moyen.

L'effet marginal moyen, EMM, est le plus souvent préférable. Dans ce cas, le facteur d'échelle est $n^{-1}\sum_{i=1}^{n}\Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$. Contrairement à l'EMPM, l'EMM ne nécessite pas de prendre une moyenne qui peut ne pas exister, et il n'y a pas de décision à prendre sur le calcul des moyennes de fonctions non linéaires. Comme celui de l'EMPM, le facteur d'échelle de l'EMM est toujours entre zéro et un, puisque $0 < \Phi(\mathbf{x}\hat{\boldsymbol{\beta}}/\hat{\sigma}) < 1$ pour toute valeur des variables explicatives. En effet, $\hat{\mathrm{P}}(y_i > 0|\mathbf{x}_i) = \Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$, donc les facteurs d'échelle de l'EMM et de l'EMPM tendent à être plus proches de un quand il y a peu d'observations avec $y_i = 0$. Dans le cas où $y_i > 0$ pour tout i, les paramètres du Tobit et des MCO sont identiques [Bien entendu, si $y_i > 0$ pour tout i, il n'est pas justifié d'utiliser un Tobit. Prédire $\log(y_i)$ avec un modèle de régression linéaire semble bien plus adapté].

Malheureusement, avec des variables explicatives discrètes, comparer les estimations des MCO et d'un Tobit n'est pas si facile (même si l'utilisation du facteur d'échelle pour les variables explicatives continues est souvent une bonne approximation). Pour un Tobit, l'effet partiel d'une variable explicative discrète – par exemple binaire – doit être calculé en estimant $\mathrm{E}(y|\mathbf{x})$ dans l'équation (17.25). Par exemple, si x_1 est binaire, nous devons faire le calcul pour $x_1 = 1$ et ensuite pour $x_1 = 0$. Si nous mettons les autres valeurs explicatives à leur moyenne de l'échantillon, nous obtenons une valeur analogue à (17.16) dans les cas logit et probit. Si nous calculons la différence des valeurs prédites pour chaque individu avant de calculer des moyennes, nous obtenons un EMM analogue à (17.17). Heureusement, beaucoup de logiciels statistiques récents calculent les EMM de modèles relativement sophistiqués, dont le modèle Tobit, et ce aussi bien pour des variables explicatives continues que discrètes.

EXEMPLE 17.2
Offre de travail annuelle des femmes mariées

Le fichier MROZ inclut des données sur le temps de travail de 753 femmes mariées (en heures), dont 428 ont reçu un salaire hors du foyer durant l'année ; 325 de ces femmes n'ont pas travaillé. Pour les femmes ayant travaillé, la durée du travail est très variable, de 12 heures à 4 950 heures. En conséquence, la durée annuelle du travail semble adaptée à un modèle Tobit. Nous pouvons aussi estimer un modèle linéaire par les MCO (sur toutes les 753 observations), et calculer les écarts-types robustes à l'hétéroscédasticité. Les résultats sont donnés dans le tableau 17.3.

Plusieurs aspects des résultats de ce tableau sont importants. Premièrement, le signe des coefficients du modèle Tobit sont les mêmes que ceux des MCO, et la significativité statistique des coefficients est similaire (Les coefficients de *nwifeinc* et *kidsge6* pourraient faire exception, mais les statistiques t sont similaires). Par ailleurs, les valeurs des coefficients des MCO et du modèle Tobit ne sont pas comparables, même s'il serait tentant de le faire. Nous devons être attentifs et ne pas penser que parce que le coefficient de *kidsge6* dans le modèle Tobit est environ le double de celui des MCO, le modèle Tobit implique un effet des jeunes enfants sur le temps de travail plus important.

Nous pouvons multiplier les coefficients estimés du modèle Tobit par les facteurs d'ajustement correspondants pour les rendre approximativement comparables aux MCO. Le facteur EMM $n^{-1}\sum_{i=1}^{n}\Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$ est dans ce cas environ 0,589, nous pouvons l'utiliser pour trouver l'effet partiel correspondant à l'estimation Tobit. Si, par exemple, le coefficient d'*educ* est multiplié par 0,589, nous trouvons $0{,}589(80{,}65) \approx 47{,}50$ (soit 47,5 heures de plus), ce qui est nettement plus que l'effet partiel des MCO – environ 28,8 heures. Le tableau 17.4 contient les EMM de toutes les variables, les EMM du modèle linéaire sont simplement les coefficients des MCO à part pour la variable *exper* qui est prise en compte de manière quadratique. Les EMM et leurs écarts-types, obtenus avec Stata®, sont arrondis à deux chiffres après la virgule. À cause des arrondis, les résultats peuvent être légèrement différents d'une multiplication par 0,589 des coefficients du modèle Tobit. Les EMM Tobit de *nwinfeinc*, *educ* et *kidslt6* sont nettement plus grands que ceux des coefficients des MCO en valeur absolue. Les EMM d'*exper* et *age* sont similaires, et pour *kidsge6*, qui est très loin d'être statistiquement significatif, l'EMM Tobit est plus petit en valeur absolue.

Si, en revanche, nous cherchons à calculer l'effet d'une année supplémentaire d'instruction en partant de la valeur moyenne de toutes les variables explicatives, nous obtenons le facteur d'échelle d'un EMPM $\Phi(\bar{\mathbf{x}}\hat{\boldsymbol{\beta}}/\hat{\sigma})$. Il est ici d'environ 0,645 [si on utilise le carré de l'expérience moyenne $(\overline{exper})^2$, au lieu de la moyenne d'$exper^2$]. Cet effet partiel, d'environ 52 heures, est presque le double de l'estimateur des MCO. À l'exception de *kidsge6*, les coefficients du modèle Tobit impliquent tous des effets partiels plus grands que les coefficients des MCO correspondants.

Nous avons rapporté un *R*-carré à la fois pour la régression linéaire et le modèle Tobit. Le *R*-carré des MCO est le *R*-carré habituel. Pour le Tobit, le *R*-carré est le carré de la corrélation entre y_i et $\hat{y}_i$, où $\hat{y}_i = \Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})\mathbf{x}_i\hat{\boldsymbol{\beta}} + \hat{\sigma}\phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$ est l'estimateur de $E(y|\mathbf{x} = \mathbf{x}_i)$. En effet, le *R*-carré des MCO est égal au carré de la corrélation entre y_i et les valeurs prédites [voir l'équation (3.29)]. Dans des modèles non linéaires comme le modèle Tobit, le carré de la corrélation n'est pas égal à un *R*-carré basé sur la somme des carrés des résidus comme dans (3.28). Les valeurs prédites définies plus haut et les résidus $y_i - \hat{y}_i$, sont corrélés dans l'échantillon. Un *R*-carré défini comme le carré de la corrélation entre y_i et $\hat{y}_i$ a l'avantage d'être toujours entre zéro et un, ce qui ne serait pas forcément le cas d'un *R*-carré basé sur la somme des carrés des résidus.

Tableau 17.3 *Estimation du temps de travail annuel par MCO et Tobit*

Variable dépendante : hours (*temps de travail en heures*)		
Variables explicatives	**Modèle linéaire (MCO)**	**Tobit (EMV)**
nwifeinc	−3,45 (2,54)	−8,81 (4,46)
educ	28,76 (12,95)	80,65 (21,58)
exper	65,67 (9,96)	131,56 (17,28)
$exper^2$	−0,700 (0,325)	−1,86 (0,54)
age	−30,51 (4,36)	−54,41 (7,42)
kidslt6	−442,09 (58,85)	−894,02 (111,88)
kidsge6	−32,78 (23,18)	−16,22 (38,64)
constant	1 330,48 (270,78)	965,31 (446,44)
Valeur de la log-vraisemblance *R*-carré $\hat{\sigma}$	– 0,266 750,18	−3 819,09 0,274 1 122,02

Tableau 17.4 *Effets marginaux moyens de modèles prédisant le nombre d'heures travaillées*

Variables explicatives	Modèle linéaire	Tobit
nwifeinc	−3,45 (2,24)	−5,19 (2,62)
educ	28,76 (13,04)	47,47 (12,62)
exper	50,78 (4,45)	48,79 (3,59)
age	−30,51 (4,24)	−32,03 (4,29)
kidslt6	−442,09 (57,46)	−526,28 (64,71)
kidsge6	−32,78 (22,80)	−9,55 (22,75)

Nous pouvons voir que, si on se base sur les R-carrés, l'espérance conditionnelle du Tobit prédit le temps de travail à peine mieux que les MCO. Cependant, rappelons-nous que les estimations du modèle Tobit ne sont pas choisies pour maximiser un R-carré – elles maximisent une log-vraisemblance – alors que l'estimateur des MCO est celui qui produit le plus grand R-carré parmi les estimateurs linéaires.

Par construction, toutes les valeurs prédites du modèle Tobit pour *hours* sont positives. En revanche, 39 valeurs prédites par les MCO sont négatives. Bien que ces prédictions négatives puissent être un problème, elles concernent 39 observations sur 753, soit environ 5 %. Il n'est pas évident de comprendre comment ces valeurs prédites pour les MCO affectent les différences d'effets partiels. La figure (17.3) représente E(*hours*|**x**) en fonction du niveau d'instruction ; pour le modèle Tobit, les autres variables explicatives sont mises à leur moyenne. Pour le modèle linéaire, l'équation est $\widehat{hours} = 387{,}19 + 28{,}76\,educ$. Pour le modèle Tobit, l'équation est $\widehat{hours} = \Phi[(-694{,}12 + 80{,}65\,educ)/1\,122{,}02] \cdot (-694{,}12 + 80{,}65\,educ) + 1\,122{,}02 \cdot \Phi[(-694{,}12 + 80{,}65\,educ)/1\,122{,}02]$. Comme on peut le voir sur la figure, le modèle linéaire donne des estimations du temps de travail plus élevées, y compris à des niveaux d'instruction relativement élevés. Par exemple, à huit années d'instruction, les MCO prédisent 617,5 heures de travail, le modèle Tobit environ 423,9. À 12 années d'instruction, les prédictions sont respectivement 732,7 et 598,3. Les deux lignes de prédictions se croisent après 17 années d'instruction, mais aucune femme de l'échantillon n'a plus de 17 années d'instruction. La pente croissante de la courbe du modèle Tobit indique clairement que celui-ci prédit que l'instruction a un effet marginal croissant sur le temps de travail espéré.

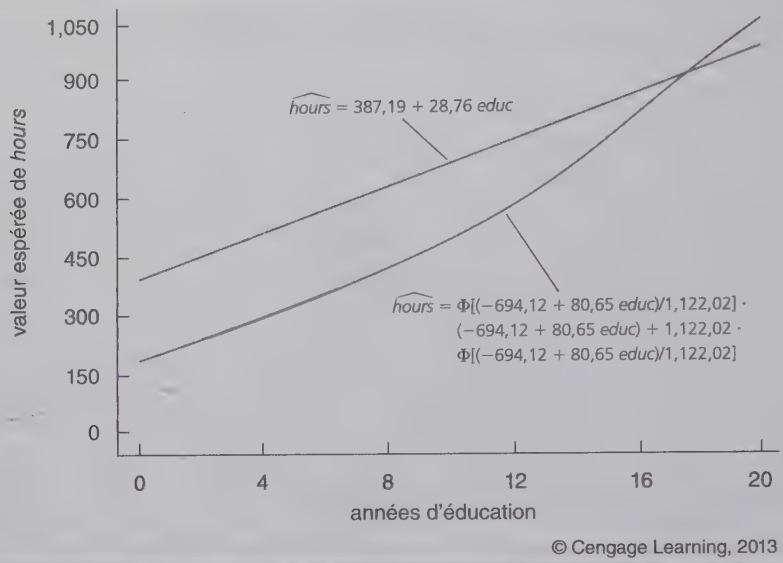


Figure 17.3 *Valeurs espérées de hours en fonction du niveau d'instruction pour des modèles linéaires et Tobit.*

Problèmes de spécification dans les modèles Tobit

Le modèle Tobit dépend de la normalité et de l'hétéroscédasticité des résidus du modèle à variable latente sous-jacent, surtout les formules pour les espérances de (17.24) et (17.25). Si $E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$, nous savons depuis le chapitre 5 que la normalité de y ne joue pas de rôle sur l'absence de biais, la convergence ou l'inférence asymptotique. L'hétéroscédasticité n'affecte pas l'absence de biais ou la convergence des MCO, même si elle demande de calculer des écarts-types estimés et des statistiques de test robustes pour faire de l'inférence. Dans un modèle Tobit, si une des hypothèses de (17.18) est fausse, il est difficile de savoir ce que l'EMV du modèle Tobit estime. Néanmoins, pour des violations modérées de ces hypothèses, le modèle Tobit donne probablement de bonnes estimations des effets partiels sur les moyennes conditionnelles. Il est possible d'avoir des hypothèses un peu moins restrictives dans (17.18), mais ces modèles sont bien plus difficiles à estimer et à interpréter.

Une limite du modèle Tobit pouvant être importante dans certaines applications est que l'espérance conditionnelle à $y > 0$ est très liée à la probabilité que $y > 0$. C'est évident dans les équations (17.26) et (17.29). En particulier, l'effet de x_j sur $P(y > 0|\mathbf{x})$ est proportionnel à β_j, comme l'effet sur $E(y|y > 0, \mathbf{x})$, et les deux fonctions multipliant β_j sont positives et dépendent de $\mathbf{x}$ seulement via $\mathbf{x}\boldsymbol{\beta}/\sigma$. Cela limite les possibilités du modèle Tobit. Par exemple, considérons la relation entre la valeur de l'assurance-vie et l'âge d'une personne. Les personnes jeunes ont probablement moins de chances d'avoir une assurance-vie, la probabilité que $y > 0$ augmenterait donc avec l'âge (au moins jusqu'à un certain point). Conditionnellement à avoir une assurance-vie, la valeur pourrait décroître avec l'âge, puisque l'assurance-vie devient moins importante quand les gens s'approchent de la fin de leur vie. Le modèle Tobit n'offre pas cette possibilité.

Un moyen de tester grossièrement si un modèle Tobit est adapté est d'estimer un probit où la variable binaire prédite, appelons-la w, vaut 1 si $y > 0$ et $w = 0$ si $y = 0$. Dans ce cas, d'après (17.21), w suit un modèle probit, où le coefficient de x_j est $\gamma_j = \beta_j/\sigma$. Cela veut dire qu'on peut estimer le ratio entre β_j et σ par probit pour tout j. Si le modèle Tobit est valable, l'estimateur probit, $\hat{\gamma}_j$, devrait être « proche » de $\hat{\beta}_j/\hat{\sigma}$, où $\hat{\beta}_j$ et σ sont les estimations du modèle Tobit. Elles ne seront jamais identiques à cause de la marge d'erreur des

estimateurs. Mais il est possible de regarder s'il y a des signes qui posent problèmes. Par exemple, si $\hat{\gamma}_j$, est négatif et significatif alors que $\hat{\beta}_j$ est positif, un modèle Tobit n'est peut-être pas adapté. Ou alors, si $\hat{\gamma}_j$ et $\hat{\beta}_j$ ont le même signe, mais $\hat{\beta}_j/\hat{\sigma}$ est beaucoup plus grand ou beaucoup plus petit que $\hat{\gamma}_j$, cela peut aussi indiquer un problème. Il ne faut pas trop s'inquiéter de changements de signe ou de valeurs des coefficients qui ne sont significatifs dans aucun des deux modèles.

Dans l'exemple du temps de travail annuel, $\hat{\sigma} = 1\,122{,}02$. Quand nous divisons les coefficients Tobit de *nwifeinc* par σ, nous trouvons $-8{,}81/1\,122{,}02 \approx -0{,}0079$; le coefficient probit de *nwifeinc* est environ $-0{,}012$, ce qui est différent mais comparable. Pour *kidslt6*, le coefficient estimé de $\hat{\beta}_j/\hat{\sigma}$ est environ $-0{,}797$, à comparer avec un estimateur probit de $-0{,}868$. Encore une fois, ce n'est pas une grande différence, mais cela indique qu'avoir peu d'enfants a plus d'effet sur le fait de travailler que sur le temps de travail d'une femme une fois qu'elle travaille (En pratique, le Tobit calcule une moyenne de ces deux effets mélangés). Nous ne pouvons pas savoir si ces effets sont statistiquement différents, mais ils ont des ordres de grandeur similaires.

Que se passe-t-il si nous concluons que le modèle Tobit n'est pas adapté ? Il y a des modèles qui peuvent être utilisés quand le modèle Tobit ne marche pas. Ils ont tous la propriété que $P(y > 0|\mathbf{x})$ et $E(y|y > 0, \mathbf{x})$ dépendent de paramètres différents, donc x_j peut avoir des effets différents sur ces deux fonctions [voir *hurdle models* et *two-part models* dans Wooldridge (2010, chapitre 17)].

17.3 LE MODÈLE DE RÉGRESSION DE POISSON

Les **variables de comptage** sont un autre type de variables dépendantes non négatives qui peuvent prendre des valeurs entières non négatives : {0, 1, 2, ...}. Nous sommes particulièrement intéressés aux cas où y prend assez peu de valeurs différentes, dont zéro. On peut donner comme exemple le nombre d'enfants d'une femme, le nombre de fois qu'une personne est arrêtée par la police dans l'année, le nombre de brevets déposés par une entreprise sur l'année. Pour les mêmes raisons que les modèles à réponse binaire et les Tobit, un modèle linéaire pour $E(y|x_1, \ldots, x_k)$ ne donne pas forcément les meilleures prédictions pour toutes les valeurs des variables explicatives (Il peut néanmoins être intéressant de commencer par un modèle linéaire, comme nous l'avons fait dans l'exemple 3.5).

Comme dans le modèle Tobit, nous ne pouvons pas prendre le logarithme d'une variable de comptage, car elle prend souvent la valeur zéro. Une approche pertinente est de modéliser son espérance comme une exponentielle :

$$E(y|x_1, x_2, ..., x_k) = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k). \qquad [17.31]$$

Comme $\exp(\cdot)$ est toujours positive, (17.31) s'assure que les valeurs prédites pour y seront aussi positives. La fonction exponentielle est représentée en figure A.5 de l'Annexe A.

Bien que (17.31) soit plus compliqué qu'un modèle linéaire, nous savons déjà comment interpréter les coefficients. En prenant le log de (17.31), on trouve :

$$\log[E(y|x_1, x_2, \ldots, x_k)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \qquad [17.32]$$

donc le log de la valeur espérée est linéaire. En utilisant les propriétés des approximations de la fonction log utilisées dans les chapitres précédentes, on trouve donc

$$\%\,\Delta E(y|\mathbf{x}) \approx (100\beta_j)\Delta x_j.$$

En d'autres termes, $100\beta_j$ est approximativement le changement de $E(y|\mathbf{x})$ en pourcentage pour une augmentation de x_j d'une unité. Parfois, une estimation plus précise est nécessaire, et il est facile de regarder

l'effet d'un changement discret des valeurs espérées. En gardant toutes les variables explicatives à part x_k constantes, le rapport des espérances pour une augmentation de x_k^0 à x_k^1 s'écrit

$$[\exp(\beta_0 + \mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} + \beta_k x_k^1)/\exp(\beta_0 + \mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} + \beta_k x_k^0)] - 1 = \exp(\beta_k \Delta x_k) - 1,$$

où $\mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1}$ est une abréviation de $\beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1}$, et $\Delta x_k = x_k^1 - x_k^0$. Quand $\Delta x_k = 1$ (par exemple si x_k est une variable binaire que nous changeons de zéro à un) le changement est $\exp(\beta_k) - 1$. Étant donné $\hat{\beta}_k$, nous obtenons $\exp(\hat{\beta}_k) - 1$ et multiplions ceci par 100 pour transformer le changement en proportions en changement en pourcentages.

Si $x_j = \log(z_j)$ pour une variable $z_j > 0$, alors son coefficient β_j s'interprète comme l'élasticité par rapport à z_j. Techniquement, il s'agit de l'élasticité de *l'espérance* de y par rapport à z_j car nous ne pouvons calculer de changement en pourcentages quand $y = 0$. Dans notre cas, cette distinction n'a pas d'importance. En pratique, l'essentiel ici est que nous pouvons interpréter les coefficients de l'équation (17.31) comme si nous avions un modèle linéaire, avec $\log(y)$ comme variable dépendante. Il y a quelques petites différences que nous n'étudierons pas ici.

Comme (17.31) n'est pas linéaire en fonction des paramètres – souvenez-vous que $\exp(\cdot)$ n'est pas une fonction linéaire –, nous ne pouvons pas utiliser de technique de régression linéaire. Nous pourrions utiliser les *moindres carrés non linéaires* qui, comme les MCO, minimisent la somme des carrés des résidus. Cependant, la plupart des données de comptage sont hétéroscédastiques, et les moindres carrés non linéaires n'exploitent pas cette hétéroscédasticité [voir Wooldridge (2010, chapitre 12)]. Nous nous servirons donc plutôt du maximum de vraisemblance et de la technique d'*estimation par quasi-maximum de vraisemblance*.

Dans le chapitre 4, nous avions dit que la loi normale était la loi de distribution standard pour les régressions linéaires. Cette hypothèse de normalité est raisonnable pour des variables dépendantes (presque) continues qui peuvent prendre de nombreuses valeurs. Une variable de comptage ne peut suivre une loi normale (puisque la loi normale décrit des variables continues qui peuvent prendre toutes les valeurs décimales), et si elle prend peu de valeurs, la distribution peut être très loin d'être normale. La loi de distribution la plus utilisée pour les variables de comptage est la **distribution de Poisson**.

Comme nous cherchons l'effet des variables explicatives sur y, nous devons avoir une distribution de Poisson conditionnellement à $\mathbf{x}$. La distribution de Poisson est entièrement déterminée par sa moyenne, donc il suffit de spécifier $\mathrm{E}(y|\mathbf{x})$. Supposons qu'elle a la forme de (17.31), que nous écrirons de manière abrégée $\exp(\mathbf{x}\boldsymbol{\beta})$. La probabilité que y prenne la valeur h conditionnellement à x est alors

$$\mathrm{P}(y = h|\mathbf{x}) = \exp[-\exp(\mathbf{x}\boldsymbol{\beta})][\exp(\mathbf{x}\boldsymbol{\beta})]^h/h!, \; h = 0,1, \ldots,$$

où $h!$ est la factorielle (voir l'annexe B). Cette distribution, qui est la base du **modèle de régression de Poisson**, nous permet de calculer la probabilité conditionnelle quelle que soit la valeur des variables explicatives. Par exemple, $\mathrm{P}(y = 0|\mathbf{x}) = \exp[-\exp(\mathbf{x}\boldsymbol{\beta})]$. Une fois que nous avons estimé les β_j, nous pouvons calculer les probabilités pour toutes les valeurs de $\mathbf{x}$.

Pour un échantillon $\{(\mathbf{x}_i, y_i) : i = 1, 2, \ldots, n\}$ donné, nous pouvons maintenant calculer la log-vraisemblance :

$$\mathscr{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \mathbf{x}_i \boldsymbol{\beta} - \exp(\mathbf{x}_i \boldsymbol{\beta})\}, \qquad \textbf{[17.33]}$$

où le terme $-\log(y_i!)$ a été supprimé car il ne dépend pas de $\boldsymbol{\beta}$. Cette fonction de log-vraisemblance est facile à maximiser, même si l'EMV de Poisson n'a pas de solution en forme fermée (pas de solution analytique).

Les écarts-types estimés des estimateurs de Poisson $\hat{\beta}_j$ sont faciles à calculer après que la log-vraisemblance ait été maximisée. La formule est dans l'annexe 17B. Ils sont donnés en même temps que les $\hat{\beta}_j$ par tout logiciel de statistiques.

Comme pour les modèles logit, probit et Tobit, nous ne pouvons pas comparer directement les coefficients des estimations d'un modèle de Poisson avec les estimations par MCO d'une fonction linéaire. Cependant, une comparaison approximative est possible au moins pour les variables explicatives continues. Sous l'hypothèse (17.31), l'effet partiel de x_j sur $E(y|x_1, x_2, ..., x_k)$ est $\partial E(y|x_1, x_2, \ldots, x_k)/x_j = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \cdot \beta_j$. Cette expression vient de la dérivation des fonctions composées et du fait que la dérivée de la fonction exponentielle est la fonction exponentielle. Si nous appelons $\hat{\gamma}_j$ le coefficient de la régression de y sur $x_1, x_2, ..., x_k$ par les MCO, nous pouvons comparer grossièrement la valeur de $\hat{\gamma}_j$ avec l'effet partiel moyen d'une régression avec une fonction exponentielle. Le facteur d'échelle de l'EMM s'écrit en effet dans ce cas $n^{-1}\sum_{i=1}^{n}\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}) = n^{-1}\sum_{i=1}^{n}\hat{y}_i$, ce qui est simplement la moyenne des y_i, en définissant les valeurs prédites comme $\hat{y}_i = \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$. En d'autres termes, pour une régression de Poisson avec une espérance exponentielle, la moyenne des valeurs prédites est la même que la moyenne des variables dépendantes y_i – comme dans une régression linéaire. Cela simplifie le calcul du facteur d'échelle des estimations de Poisson $\hat{\beta}_j$ pour les rendre comparables aux estimations par MCO $\hat{\gamma}_j$: pour une variable explicative continue, on peut comparer $\hat{\gamma}_j$ à $\bar{y} \cdot \hat{\beta}_j$.

Bien que l'EMV de Poisson soit une première étape naturelle pour les données de comptage, elle est souvent bien trop restrictive. Toutes les probabilités et les moments d'une distribution de Poisson sont totalement déterminés par la moyenne. En particulier, la variance est égale à la moyenne :

$$\text{Var}(y|\mathbf{x}) = E(y|\mathbf{x}). \qquad \textbf{[17.34]}$$

Ceci est restrictif, et on a pu montrer que c'était faux dans de nombreuses applications. Heureusement, la loi de Poisson a des propriétés de robustesse très utiles : que la distribution suive une loi de Poisson ou non, les estimateurs de β_j sont toujours convergents et asymptotiquement normaux [Voir Wooldridge (2010, chapitre 18) pour plus de détails]. C'est similaire à l'estimateur des MCO, qui est convergent et asymptotiquement normal que les résidus soient normaux ou non ; mais l'estimateur des MCO est l'EMV seulement si les résidus sont normaux.

En utilisant l'EMV de Poisson alors que nous ne supposons pas nécessairement que la distribution de Poisson est totalement vraie, nous appelons cette analyse **estimation par quasi-maximum de vraisemblance** (**EQMV**). L'EQMV de Poisson est très pratique car il est programmé dans beaucoup de logiciels d'économétrie. Cependant, les écarts-types estimés doivent être ajustés sauf si l'hypothèse de variance des distributions de Poisson (17.34) est vraie.

Un ajustement simple des écarts-types estimés est possible si on suppose que la variance est proportionnelle à la moyenne :

$$\text{Var}(y|\mathbf{x}) = \sigma^2 E(y|\mathbf{x}), \qquad \textbf{[17.35]}$$

où $\sigma^2 > 0$ est un paramètre inconnu. Quand $\sigma^2 = 1$, nous obtenons la variance de Poisson (17.34). Quand $\sigma^2 > 1$, la variance est plus grande que la moyenne pour tout $\mathbf{x}$; cela s'appelle de la **sur-dispersion** car la variance est plus grande que dans le cas de Poisson, et c'est le cas de beaucoup de modèles de comptage. Le cas $\sigma^2 < 1$, appelé *sous-dispersion*, est moins commun mais possible dans (17.35).

Avec (17.35), il est facile d'ajuster les écarts-types estimés de l'EMV de Poisson habituel. Appelons $\hat{\beta}_j$ l'EQMV de Poisson et définissons ses résidus $\hat{u}_i = y_i - \hat{y}_i$, où $\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})$ est la valeur prédite. Comme d'habitude, le résidu entre l'observation i est la différence entre y_i et sa valeur prédite. $(n - k - 1)^{-1}\sum_{i=1}^{n}\hat{u}_i^2/\hat{y}_i$, est un estimateur convergent de σ^2, la division par $\hat{y}_i$ est l'ajustement

approprié pour l'hétéroscédasticité, et $n - k - 1$ est le nombre de degrés de libertés avec n observations et $k + 1$ coefficients estimés $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$. Si $\hat{\sigma}$ est la racine carrée de $\hat{\sigma}^2$, nous multiplions les écarts-types estimés habituels de Poisson par $\hat{\sigma}$. Si $\hat{\sigma}$ est nettement plus grand que un, les écarts-types estimés corrigés peuvent être nettement plus grands que les écarts-types estimés de Poisson, généralement incorrects.

Même (17.35) n'est pas totalement général. Comme dans le cas du modèle linéaire, nous ne pouvons pas obtenir d'EQMV de Poisson si nous ne restreignons pas du tout la variance [Voir Wooldridge (2010, chapitre 18) pour plus de détails].

Sous l'hypothèse de distribution de Poisson, la statistique du rapport de vraisemblance peut être utilisée pour tester des restrictions d'exclusion, qui ont, comme toujours, la forme de (17.12). Si nous avons q restrictions d'exclusion, la statistique suit approximativement un χ^2_q sous l'hypothèse nulle. Sous l'hypothèse moins restrictive (17.35), un ajustement simple est disponible (et la statistique est alors appelée **statistique du rapport des quasi-vraisemblances**) : nous divisons (17.12) par $\hat{\sigma}^2$, $\hat{\sigma}^2$ étant obtenu avec le modèle non restreint.

EXEMPLE 17.3
Régression de poisson prédisant le nombre d'arrestations

Appliquons maintenant le modèle de régression de Poisson aux données de criminalité contenues dans CRIME1, utilisées entre autres dans l'exemple 9.1. La variable dépendante, *narr86*, est le nombre d'arrestations par personne en 1986. Cette variable vaut zéro pour 1 970 des 2 725 hommes de l'échantillon, et *narr86* est plus grand que cinq dans huit cas seulement. Dans ce cas, une régression de Poisson est donc plus adaptée qu'une régression linéaire. Le tableau 17.5 présente également les résultats d'un modèle de régression linéaire.
Les écarts-types estimés reportés pour les MCO sont classiques ; donner des écarts-types estimés robustes à l'hétéroscédasticité aurait aussi été possible. Les écarts-types estimés de la régression de Poisson sont les écarts-types estimés habituels du maximum de vraisemblance. Comme $\hat{\sigma} = 1{,}232$, les écarts-types estimés de la régression de Poisson devraient être augmentés de ce facteur (les écarts-types estimés corrigés seraient augmentés de 23 %. Par exemple, un écart-type estimé plus fiable de *tottime* est $1{,}23(0{,}015) \approx 0{,}0185$, qui donne une statistique t d'environ 1,3. L'ajustement des écarts-types estimés diminue la significativité de toutes les variables, mais certaines d'entre elles sont très significatives statistiquement.

Tableau 17.5 *Déterminants du nombre d'arrestations chez les jeunes hommes*

Variable dépendante : *narr86*		
Variables explicatives	**Modèle linéaire (MCO)**	**Modèle exponentiel (EQMV de Poisson)**
pcnv	−0,132 (0,040)	−0,402 (0,085)
avgsen	−0,011 (0,012)	−0,024 (0,020)
tottime	0,012 (0,009)	0,024 (0,015)
ptime86	−0,041 (0,009)	−0,099 (0,021)
qemp86	−0,051 (0,014)	−0,038 (0,029)

Variable dépendante : *narr86*		
Variables explicatives	**Modèle linéaire (MCO)**	**Modèle exponentiel (EQMV de Poisson)**
inc86	−0,0015 (0,0003)	−0,0081 (0,0010)
black	0,327 (0,045)	0,661 (0,074)
hispan	0,194 (0,040)	0,500 (0,074)
born60	−0,022 (0,033)	−0,051 (0,064)
constant	0,577 (0,038)	−0,600 (0,067)
Log-vraisemblance *R*-carré $\hat{\sigma}$	– 0,073 0,829	−2 248,76 0,077 1,232

Les coefficients MCO et de Poisson ne sont pas directement comparables, et ils ont des significations très différentes. Par exemple, le coefficient de *pcnv* dans le modèle linéaire implique que si $\Delta pcnv = 0,10$, le nombre d'arrestations diminue de 0,013 (*pcnv* est la proportion des arrestations précédentes ayant débouché sur une condamnation). Le coefficient de Poisson implique que si $\Delta pcnv = 0,10$, le nombre d'arrestations se réduit d'environ 4 % [0,402(0,10) = 0,0402, que nous multiplions par 100 pour avoir un pourcentage]. Cela suggère donc une diminution du nombre d'arrestations de 4 % environ si les chances de condamnation augmentent de 10 %.

Le coefficient de Poisson de *black* implique que, toutes choses égales par ailleurs, le nombre d'arrestations espérées d'un homme noir serait environ $100 \cdot [\exp(0,661) - 1] \approx 93.7$ % plus élevé que pour un homme blanc avec les mêmes caractéristiques.

Comme pour le modèle Tobit dans le tableau 17.3, nous rapportons un *R*-carré dans la régression de Poisson : le carré de la corrélation entre y_i et $\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})$. La justification de cette mesure de qualité de la prédiction est la même que pour le modèle Tobit. Nous pouvons voir que le modèle exponentiel estimé par l'EQMV de Poisson colle un peu mieux aux données. Rappelez-vous que les estimations par les MCO maximisent le *R*-carré, au contraire des estimations de Poisson (qui maximisent une log-vraisemblance).

D'autres modèles de régressions sur données de comptage ont été proposés et utilisés dans diverses applications, ils généralisent de plusieurs manières les régressions de Poisson. Si vous vous intéressez aux effets de x_j sur la réponse moyenne, il y a peu de raisons de chercher plus loin que la régression de Poisson : ce modèle est simple, donne généralement de bons résultats, et nous avons discuté plus haut de ses propriétés de robustesse. Nous pourrions même appliquer une régression de Poisson à un y qui correspond au modèle Tobit, du moment que (17.31) tient toujours. Cela pourrait donner une bonne estimation des effets moyens. Les extensions des régressions de Poisson sont plus utiles quand on cherche à estimer des probabilités, comme $P(y > 1|\mathbf{x})$ [voir par exemple Cameron et Trivedi (1998)].

17.4 LES MODÈLES DE RÉGRESSIONS TRONQUÉES OU CENSURÉES

Les modèles des sections 17.1, 17.2 et 17.3 s'appliquent à différentes sortes de variables dépendantes limitées souvent utilisées en économétrie. Pour choisir un modèle, il faut se souvenir qu'on utilise un modèle probit ou logit pour les réponses binaires, un modèle Tobit pour les réponses avec solutions en coin, et une régression de Poisson pour les données de comptage : nous devons utiliser un modèle qui correspond à la distribution de y. Il n'y a pas de problème d'observation des données. Par exemple, dans l'application de l'exemple 17.2 avec un modèle Tobit prédisant l'offre de travail des femmes, le problème n'est pas d'observer le nombre d'heures travaillées ; simplement, un nombre important des femmes mariées n'a pas de travail salarié. Dans l'application de la régression de Poisson pour prédire le nombre annuel d'arrestations, nous observons la variable dépendante pour beaucoup de jeunes hommes, mais cette variable peut être nulle ou prendre de petites valeurs entières.

Malheureusement, la distinction entre les caractéristiques de la variable dépendante (comme valoir zéro pour une part importante de la population) et les problèmes de censure des données peut parfois induire en erreur. C'est souvent le cas dans les applications du modèle Tobit. Dans ce livre, le modèle Tobit standard décrit dans la section 17.2 ne traite que les solutions en coin. Mais la littérature utilise souvent les modèles Tobit pour d'autres types de situation : quand la variable dépendante a été censurée au-dessus ou en-dessous d'un seuil. Généralement, cette censure est due au questionnaire de l'enquête produisant les données, et parfois à des contraintes institutionnelles. Au lieu de traiter les problèmes de censure des données comme des solutions en coin, nous les résolvons en utilisant un **modèle de régression censurée**. Fondamentalement, le problème résolu par un modèle de régression censurée est qu'il y a des valeurs manquantes de la variable dépendante y. Bien que nous puissions tirer au sort des unités dans la population et observer les variables explicatives pour toutes ces unités, la variable dépendante y_i est manquante pour certains i. Bien entendu, nous savons si les valeurs manquantes sont au-dessus ou en-dessous d'un certain seuil, et cette information est utile pour estimer les paramètres du modèle.

Un **modèle de régression tronquée** traite les situations où nous excluons, sur la base de y, une partie de la population de l'échantillon. En d'autres termes, nous n'avons pas d'échantillon aléatoire de la population sous-jacente, mais nous savons quelle règle a été utilisée pour choisir les unités d'observation. Cette règle dépend y, selon qu'il soit en-dessous ou au-dessus d'un certain seuil. Nous expliquerons plus en détail la différence entre modèles de régression tronquée et censurée plus loin.

Modèles de régression censurée

Les modèles de régressions censurées peuvent être définis sans hypothèse distributionnelle, mais nous étudierons dans cette sous-section le **modèle de régression censurée normal**. La variable que nous voudrions expliquer, y, vient d'un modèle linéaire classique. L'indice i représente une observation de la population :

$$y_i = \beta_0 + \mathbf{x}_i\boldsymbol{\beta} + u_i,\ u_i | \mathbf{x}_i, c_i \sim \text{Normale}\ (0,\sigma^2) \qquad \textbf{[17.36]}$$

$$w_i = \min(y_i, c_i). \qquad \textbf{[17.37]}$$

Au lieu d'observer y_i, nous l'observons si et seulement si il est plus petit qu'une valeur de censure c_i. Notez que (17.36) inclut l'hypothèse que u_i est indépendant de c_i (Pour rendre l'explication concrète, nous considérons uniquement une censure du maximum ici, ou *censure à droite* ; le problème de la censure du minimum ou *censure à gauche* se traite de manière similaire).

Un exemple de censure à droite des données est la **coupure à droite**. Quand une variable est coupée à droite, sa valeur n'est connue que jusqu'à un seuil donné. Quand les valeurs sont plus grandes que ce seuil, on sait seulement que la variable est égale ou supérieure au seuil. Par exemple, la richesse du ménage est coupée à droite dans certaines enquêtes. Supposons que les répondants donnent leur richesse, mais qu'on autorise les répondants à répondre « plus de 500 000 € ». Dans ce cas, nous observons la richesse des répondants qui possèdent moins de 500 000 €, mais pas la richesse de ceux qui possèdent plus de 500 000 €. Dans ce cas, le seuil de censure, c_i, est le même pour tous les i. Dans beaucoup de situations, le seuil de censure peut changer en fonction des individus ou des caractéristiques familiales.

Pour aller plus loin 17.5

Soit mvp_i la valeur de la productivité marginale du travailleur i, c'est-à dire la productivité marginale du travailleur multipliée par le prix du bien produit. Supposons que mvp_i soit une fonction linéaire d'une erreur inobservée et de variables explicatives, comme, entre autres, le niveau d'instruction et l'expérience. S'il y a concurrence parfaite et sans contrainte institutionnelle, chaque travailleur est payé à la valeur de sa production marginale. Soit $minwage_i$ le salaire minimum de l'employé i, qui dépend de l'État américain. Nous observons $wage_i$, qui est le plus grand terme entre mvp_i et $minwage_i$. Écrivez le modèle approprié pour les salaires observés.

Si nous observions un échantillon aléatoire de $(\mathbf{x}, y)$, nous pourrions simplement estimer $\boldsymbol{\beta}$ par les MCO, et l'inférence statistique serait standard (Nous incluons encore la constante dans x pour simplifier). La censure pose un problème : en utilisant un argument similaire au modèle Tobit, une régression des MCO sur les observations non censurées – celles pour lesquelles $y_i < c_i$ – produit des estimateurs non convergents des β_j. Une régression de w_i sur x_i par les MCO en utilisant toutes les observations n'estime pas β_j de manière convergente, sauf s'il n'y a pas de censure. C'est similaire au modèle Tobit, mais le problème de départ est très différent. Dans le modèle Tobit, on modélisait un comportement économique qui donne souvent une variable dépendante nulle. Dans une régression censurée, il y a un problème de collection des données qui sont censurées.

Sous les hypothèses (17.36) et (17.37), nous pouvons estimer $\boldsymbol{\beta}$ (et σ^2) par maximum de vraisemblance sur un échantillon aléatoire de (x_i, w_i). Pour cela, nous avons besoin de la densité de w_i conditionnellement à (x_i, c_i). Pour les observations non censurées, $w_i = y_i$, et la densité de w_i est la même que celle d'y_i : Normale$(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$. Pour les observations censurées, nous avons besoin de la probabilité que w_i soit égale à la valeur de censure, c_i, étant donné x_i :

$$P(w_i = c_i|\mathbf{x}_i) = P(y_i \geq c_i|\mathbf{x}_i) = P(u_i \geq c_i - \mathbf{x}_i\boldsymbol{\beta}) = 1 - \Phi[(c_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma].$$

Nous pouvons combiner ces deux parties pour obtenir la densité de w_i étant donnés $\mathbf{x}_i$ et c_i :

$$f(w|\mathbf{x}_i c_i) = 1 - \Phi[(c_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma], \; w = c_i, \qquad \textbf{[17.38]}$$

$$= (1/\sigma)\phi\,[(w - \mathbf{x}_i\boldsymbol{\beta}/\sigma], \; w < c_i. \qquad \textbf{[17.39]}$$

La log-vraisemblance de l'observation i est obtenue en prenant le logarithme népérien de la densité pour chaque i. Nous pouvons maximiser la somme de ces termes sur tous les i par rapport à β_j et σ, pour obtenir les EMV.

Il est important de comprendre que nous pouvons interpréter les β_j comme dans un modèle de régression linéaire avec échantillonnage aléatoire. C'est très différent des modèles Tobit appliqués aux solutions en coin, où l'espérance du terme d'intérêt est une fonction non linéaire des β_j.

Une application importante des modèles de régression censurée est l'**analyse des durées**. Une variable de durée est une variable qui mesure le temps avant qu'un événement donné n'arrive. Par exemple, nous pourrions vouloir prédire le nombre de jours avant qu'un criminel relâché de prison soit arrêté. Pour certains

criminels, cela peut ne pas arriver, ou arriver après tellement longtemps que nous devions censurer la durée pour analyser les données.

Dans les applications à des durées de la régression censurée normale, la variable dépendante est souvent prise en logarithme népérien, ce qui veut dire qu'il faut également prendre le log du seuil de censure dans (17.37). Comme nous l'avons vu dans tout ce livre, l'utilisation d'une transformation en log pour la variable dépendante mène à une interprétation des paramètres en termes de pourcentages. De plus, comme dans le cas de beaucoup de variables positives, le log d'une durée a généralement une distribution plus proche de la loi normale (conditionnelle) que la durée elle-même.

EXEMPLE 17.4
Durée avant la récidive

Le fichier RECID contient des données sur la durée en mois avant qu'un ex-prisonnier des prisons de Caroline du Nord ne soit arrêté après avoir été relâché ; appelons cette variable *durat*. Certains prisonniers ont participé à un programme de travail en prison. Nous prenons également en compte plusieurs variables démographiques, ainsi que des mesures d'emprisonnement et de l'histoire criminelle.

Sur 1 445 prisonniers, 893 n'avaient pas été arrêtés tant qu'ils étaient suivis ; ces informations sont donc censurées. Les durées de censures diffèrent entre prisonniers, entre 70 et 81 mois.

Le tableau 17.6 donne les résultats d'une régression normale censurée de log(*durat*). Tous les coefficients, en les multipliant par 100, donnent des estimations de l'effet *ceteris paribus* d'une unité supplémentaire des variables explicatives en pourcentage de la durée espérée.

Plusieurs coefficients du tableau 17.6 sont intéressants. Les variables *priors* (le nombre de condamnations précédentes) et *tserved* (le nombre total de mois passés en prison) ont un effet négatif sur la durée avant l'arrestation suivante. Cela suggère que ces variables mesurent la tendance criminelle des individus plus que l'effet dissuasif de la prison. Par exemple, un prisonnier avec une condamnation en plus a une durée avant l'arrestation suivante plus basse de 14 %. Une année supplémentaire d'emprisonnement réduit cette durée d'environ $100 \cdot 12(0{,}019) = 22{,}8$ %. Un résultat surprenant est qu'une personne emprisonnée pour crime a une durée espérée d'environ 56 % [$\exp(0{,}444) - 1 \approx 0{,}56$] *plus grande* qu'un homme emprisonné pour un délit.

Les personnes avec une histoire de consommation de drogues ou d'alcool ont une durée nettement plus courte avant l'arrestation suivante (Les variables *alcohol* et *drugs* sont binaires). Les hommes vieux, et ceux qui étaient mariés au moment de l'emprisonnement, ont des durées plus longues avant d'être ré-arrêtés. Les hommes noirs ont une durée nettement plus courte, de l'ordre de 42 % [$\exp(-0{,}543) - 1 \approx 0{,}42$].

La variable d'intérêt politique clé, *workprg*, n'a pas l'effet désiré. L'estimation est que, toutes choses égales par ailleurs, les hommes qui participent au programme ont des durées avant la récidive plus courtes d'environ 6,3 % que ceux qui n'ont pas participé. La statistique t est petite, il faudrait probablement plutôt conclure que le programme n'a pas d'effet. Cela pourrait être dû à un problème d'auto-sélection, ou à une conséquence de la manière dont les hommes sont inclus dans le programme. Bien entendu, il se peut aussi que le programme soit inefficace.

Tableau 17.6 *Estimation d'une régression censurée prédisant la récidive*

Variable dépendante : log(*durat*)	
Variables explicatives	**Coefficient (Écart-type)**
workprg	−0,063 (0,120)
priors	−0,137 (0,021)

Variable dépendante : log(*durat*)	
Variables explicatives	**Coefficient (Écart-type)**
tserved	−0,019 (0,003)
felon	0,444 (0,145)
alcohol	−0,635 (0,144)
drugs	−0,298 (0,133)
black	−0,543 (0,117)
married	0,341 (0,140)
educ	0,023 (0,025)
age	0,0039 (0,0006)
constant	4,099 (0,348)
Log-vraisemblance	−1 597,06
$\hat{\sigma}$	1,810

Dans cet exemple, il est indispensable de prendre en compte la censure, puisque presque 62 % des observations sont censurées. Si nous appliquons directement les MCO à tout l'échantillon et que nous traitons les durées censurées comme si elles ne l'étaient pas, les coefficients estimés seraient très différents. De fait, ils sont tous plus proches de zéro. Par exemple, le coefficient de *priors* devient −0,059 ($\hat{\sigma}$ = 0,009), et celui d'*alcohol* devient −0,262 ($\hat{\sigma}$ = 0,060). Bien que les signes des effets soient les mêmes, l'importance de ces variables est largement diminuée. Les coefficients de la régression censurée sont bien plus fiables.

Il y a d'autres manières de mesurer les effets des variables explicatives du tableau 17.6 sur la durée que de se focaliser uniquement sur la durée espérée. Un traitement de l'analyse moderne des durées est au-delà de ce qui est couvert ici [Pour une introduction, voir Wooldridge (2010, chapitre 22)].

Si une des hypothèses de la régression normale censurée est violée – en particulier, s'il y a hétéroscédasticité ou non normalité des résidus u_i – l'EMV est non convergent en général. Cela montre que la censure peut être très coûteuse, puisque les MCO sur un échantillon non censuré ne requièrent ni la normalité ni l'homoscédasticité pour être convergents. Il y a des méthodes plus avancées qui ne requièrent pas de supposer de distribution pour les résidus [Voir Wooldridge (2010, chapitre 19)].

Modèle de régression tronquée

Le modèle de régression tronquée diffère du modèle de régression censurée par un aspect important. Dans le cas des données censurées, les observations *sont* tirées aléatoirement dans la population. Le problème de censure est que, alors que nous observons les variables explicatives pour toutes les observations tirées au sort, nous observons

la variable dépendante y seulement quand elle n'est pas censurée au-dessus ou en-dessous d'un seuil connu. Sur des données tronquées, on restreint son attention à une partie de la population avant l'échantillonnage ; il y a donc une partie de la population que nous n'observons pas. Pour celle-ci, nous n'observons pas les variables explicatives (ni aucune variable). Un échantillonnage tronqué vient généralement du fait qu'une enquête vise en particulier une partie de la population et ignore entièrement d'autres parties de celle-ci, généralement pour des raisons de coût. Des chercheurs peuvent ensuite vouloir utiliser l'échantillon tronqué pour répondre à des questions valables pour toute la population, mais il faut reconnaître que l'échantillon n'a pas été tiré aléatoirement dans la population entière.

Par exemple, Hausman et Wise (1977) ont utilisé des données venant d'une expérience de taxation négative sur le revenu pour étudier différents déterminants de celui-ci. Les familles inclues dans l'enquête gagnaient nécessairement moins de 1,5 fois le seuil de pauvreté américain de 1967, ce seuil dépendant de la composition du ménage. Hausman et Wise voulaient utiliser ces données pour estimer une équation de revenus valable pour toute la population.

Le modèle de régression tronquée normale commence avec un modèle sous-jacent sur la population entière basé sur les hypothèses classiques du modèle linéaire :

$$y = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u,\ u|\mathbf{x} \sim \text{Normale}(0,\sigma^2). \qquad \textbf{[17.40]}$$

Souvenons-nous que ce sont des hypothèses fortes : u ne doit pas seulement être indépendant de $\mathbf{x}$, mais aussi suivre une loi normale. Nous nous focalisons sur ce modèle car il est difficile d'en relâcher les hypothèses.

Quand (17.40) est vrai, nous savons que sur un échantillon aléatoire de la population, les MCO sont la procédure d'estimation la plus efficace. Le problème vient du fait que nous n'observons pas d'échantillon aléatoire : l'hypothèse RLM.2 est violée. En particulier, un échantillon aléatoire $(\mathbf{x}_i, y_i)$ est observé seulement si $y_i \leq c_i$, où c_i est le seuil de troncature qui dépend des variables explicatives – notamment de $\mathbf{x}_i$ (Dans l'exemple de Hausman et Wise, c_i dépend de la taille du ménage). Cela veut dire que si $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ est notre échantillon *observé*, y_i est nécessairement inférieur ou égal à c_i. C'est la différence avec le modèle de régression censurée, où nous observons $\mathbf{x}_i$ pour un échantillon aléatoirement tiré dans la population ; dans le modèle tronqué, nous observons $\mathbf{x}_i$ seulement si $y_i \leq c_i$.

Pour estimer les β_j (ainsi que s), nous avons besoin de la distribution de y_i, conditionnelle à $y_i \leq c_i$ et à $\mathbf{x}_i$. Elle s'écrit

$$g(y|\mathbf{x}_i, c_i) = \frac{f(y|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)}{F(c_i|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)},\ y \leq c_i, \qquad \textbf{[17.41]}$$

où $f(y|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ est la densité de la loi normale de moyenne $\beta_0 + \mathbf{x}_i\boldsymbol{\beta}$ et de variance σ^2, et $F(c_i|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ est la fr (fonction de répartition) de la loi normale avec les mêmes moyenne et variance, évaluée en c_i. Cette expression pour la densité conditionnelle à $y_i \leq c_i$ semble intuitive : c'est la densité de y dans la population conditionnellement à $\mathbf{x}$, divisée par la probabilité F que y_i soit plus petite ou égale à c_i (étant donné $\mathbf{x}_i$), $\text{P}(y_i \leq c_i|\mathbf{x}_i)$. En effet, nous « re-normalisons » la densité en divisant par la partie de la surface sous $f(\cdot|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ située à gauche de c_i.

Si nous prenons le log de (17.41), que nous en calculons la somme sur tous les i, et maximisons le résultat par rapport à β_j et σ^2, nous obtenons un estimateur du maximum de vraisemblance. Il mène à un estimateur convergent et asymptotiquement normal. L'inférence, dont les écarts-types estimés et la log-vraisemblance, est standard et traitée par Wooldridge (2010, chapitre 19).

Nous pourrions analyser les données de l'exemple 17.4 comme un échantillon tronqué en supprimant toutes les données sur les observations censurées. Cela nous donneraient 552 observations d'une distribution normale tronquée, où le point de troncature dépendrait de i. Cependant, ce serait une mauvaise idée pour analyser des données de durées (ou coupées à droite), car cela élimine des informations utiles. Le fait que nous

connaissions une borne inférieure pour 893 observations, ainsi que leurs variables explicatives est une information utile ; les régressions censurées utilisent cette information, contrairement aux régressions tronquées.

Hausman et Wise (1977) donnent un meilleur exemple de régression tronquée, où ils mettent en avant le fait que les MCO appliqués à un échantillon tronqué par le haut produisent généralement des estimateurs biaisés vers zéro. Intuitivement, cela semble logique. Supposons que la relation d'intérêt soit celle entre revenu et niveau d'instruction. Si nous observons uniquement les personnes dont le revenu est en-dessous d'un seuil, nous perdons les plus hauts revenus. Cela tend à aplatir la relation estimée par rapport à la relation dans la population entière. La figure 17.4 illustre ce problème quand le revenu est tronqué au-dessus de 50 000 $. En effet, toutes les observations représentées par les cercles en noir seraient ramenées sur la ligne horizontale à *income* = 50.

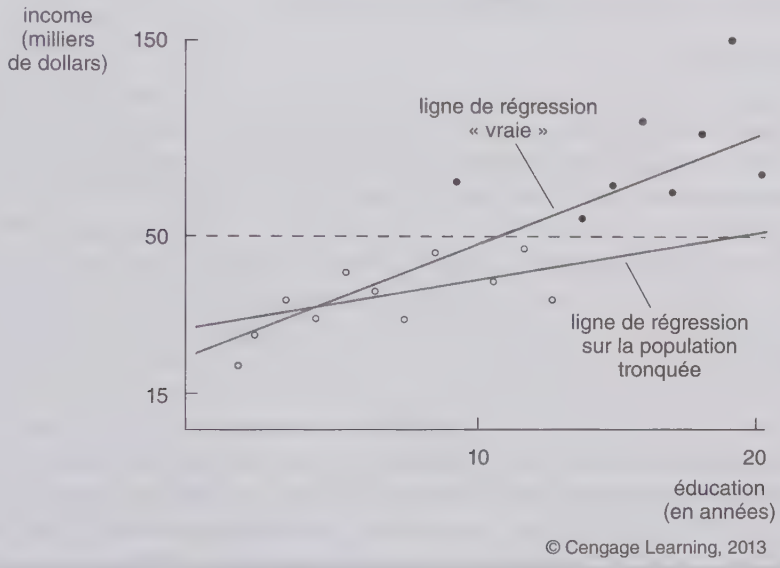


Figure 17.4 *Ligne de régression « vraie » sur toute la population et ligne de régression incorrecte pour une régression tronquée où les revenus sont observés en-dessous de 50 000 $.*

Comme pour la régression censurée, si l'hypothèse de normalité et d'hétéroscédasticité de (17.40) est violée, l'EMV normal tronqué est biaisé et non convergent. Des méthodes qui ne nécessitent pas ces hypothèses sont disponibles ; voir Wooldridge (2010, chapitre 19) pour une discussion et des références.

17.5 CORRECTION POUR LA SÉLECTION DE L'ÉCHANTILLON

Les régressions tronquées sont un cas particulier du cas plus général de **sélection non aléatoire de l'échantillon**. La conception des enquêtes est loin d'être la seule cause de sélection non aléatoire. Des répondants peuvent ne pas répondre à toutes les questions, ce qui rend des données manquantes pour certaines variables explicatives ou dépendantes. Comme nous ne pouvons pas utiliser ces observations, nous devons nous demander si les supprimer biaise nos estimateurs.

Un autre exemple fréquent est généralement appelé **troncature auxiliaire**. Dans ce cas, nous ne pouvons pas observer y à cause d'une autre variable. L'exemple générique est l'estimation de la *fonction d'offre de salaire potentiel* en économie du travail. L'objectif est d'observer dans quelle mesure différents facteurs, comme le niveau d'instruction, affectent le salaire qu'une personne peut obtenir. Nous observons le salaire

potentiel de tous ceux qui appartiennent à la population active, qui est leur salaire actuel. Mais pour ceux qui ne font pas partie de la population active, nous n'observons pas de salaire potentiel. Le fait de travailler peut être corrélé avec des caractéristiques inobservables qui affectent le salaire potentiel, donc utiliser seulement la population active peut mener à des estimateurs de l'équation de salaire potentiel biaisés (c'est ce qui a été fait dans tous les exemples jusqu'ici).

Des données de panel peuvent également mener à une sélection non aléatoire de l'échantillon. Dans le cas le plus simple, supposons que nous ayons des données couvrant deux années, mais qu'il y ait de l'attrition, c'est-à-dire que certains individus quittent l'échantillon. C'est particulièrement le cas dans l'analyse des conséquences des politiques publiques, où l'attrition peut être liée à l'efficacité du programme.

Quand les MCO sur l'échantillon sélectionné sont-ils convergents ?

La section 9.4 donnait une brève discussion des types de sélection de l'échantillon que l'on peut ignorer. La distinction clé était entre sélection de l'échantillon *exogène* et *endogène*. Dans le cas du modèle Tobit tronqué, il y a clairement sélection endogène de l'échantillon, et les MCO sont biaisés et non convergents. Cependant, si notre échantillon est uniquement déterminé par une variable exogène, nous avons une sélection exogène de l'échantillon. Les cas entre ces deux extrêmes sont moins intuitifs, et nous apportons donc ici une définition et des hypothèses claires pour les résoudre. Le modèle dans la population s'écrit

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u, \; \mathrm{E}(u|x_1, x_2, ..., x_k) = 0. \qquad \textbf{[17.42]}$$

Il est utile d'écrire ce modèle pour une observation :

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i, \qquad \textbf{[17.43]}$$

où nous utilisons $\mathbf{x}_i\boldsymbol{\beta}$ comme abréviation de $\beta_0 + \beta_1 x_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}$. Supposons maintenant que n soit la taille d'un *échantillon aléatoire* tiré de la population. Si nous pouvions observer y_i et les x_{ij} pour tout i, nous utiliserions les MCO. Supposons que y_i ou certaines variables explicatives ne soit (ou ne soient) pas observée(s) pour certains i pour une raison quelconque. Pour certaines observations, nous observons toutes les variables. Définissons un *indicateur de sélection* s_i pour tout i par $s_i = 1$ si nous observons tous les $(y_i, \mathbf{x}_i)$ et $s_i = 0$ sinon. Donc $s_i = 1$ veut dire que nous pouvons utiliser notre observation ; $s_i = 0$ veut dire que l'observation ne sera pas utilisée. Nous nous intéressons aux propriétés des estimateurs des MCO sur l'**échantillon sélectionné**, c'est-à-dire en utilisant les observations pour lesquelles $s_i = 1$. Donc, nous aurons moins que n observations, disons n_1.

Il se trouve qu'il est assez facile de trouver des conditions pour que les MCO soient convergents (et même non biaisés). En effet, plutôt que d'estimer (17.43), nous pouvons seulement estimer

$$s_i y_i = s_i \mathbf{x}_i\boldsymbol{\beta} + s_i u_i. \qquad \textbf{[17.44]}$$

Quand $s_i = 1$, nous avons (17.43) ; quand $s_i = 0$, nous avons simplement $0 = 0 + 0$, ce qui ne donne clairement aucune information sur $\boldsymbol{\beta}$. Une régression de $s_i y_i$ sur $s_i x_i$ pour $i = 1, 2, ..., n$ est donc la même chose qu'une régression de y_i sur x_i en utilisant les observations pour lesquelles $s_i = 1$. Nous pouvons donc étudier la convergence de $\hat{\beta}_j$ en étudiant (17.44) sur un échantillon aléatoire.

Nous savons depuis l'analyse du chapitre 5 que l'estimateur des MCO de (17.44) est convergent si le terme d'erreur a une moyenne nulle et n'est corrélé avec aucune variable explicative. Dans la population, l'hypothèse de moyenne nulle est $\mathrm{E}(su) = 0$, et l'hypothèse d'absence de corrélation s'écrit

$$\mathrm{E}[(sx_j)(su)] = \mathrm{E}(sx_j u) = 0, \qquad \textbf{[17.45]}$$

où s, x_j et u sont des variables aléatoires représentant la population ; nous avons utilisé l'égalité $s^2 = s$, s étant une variable binaire. La condition (17.45) est différente de celle requise dans un échantillon tiré aléatoirement : $E(x_j u) = 0$. Dans la population, nous avons donc besoin que u ne soit pas corrélé avec sx_j.

La condition clé pour l'absence de biais est $E(su|sx_1, \ldots, sx_k) = 0$. Comme d'habitude, cette condition est plus forte que celle nécessaire à la convergence.

Si s n'est fonction que des variables explicatives, alors sx_j est seulement une fonction de $x_1, x_2, \ldots, x_k$; par l'hypothèse de moyenne conditionnelle formulée en (17.42), sx_j n'est pas non plus corrélée à u. En effet, $E(su|sx_1, \ldots, sx_k) = sE(u|sx_1, \ldots, sx_k) = 0$, puisque $E(u|x_1, \ldots, x_k) = 0$. On peut appeler ceci **sélection exogène de l'échantillon**, puisque $s_i = 1$ est entièrement déterminé par $x_{i1}, \ldots, x_{ik}$. Par exemple, si nous estimons une équation de salaire où les variables explicatives sont le niveau d'instruction, l'expérience totale à ce poste, le sexe, la situation de famille, et ainsi de suite – qui sont supposées exogènes – nous pouvons sélectionner l'échantillon sur la base de toutes les variables explicatives.

Si la sélection de l'échantillon est entièrement aléatoire, c'est-à-dire que s_i est *indépendant* de $(\mathbf{x}_i, u_i)$, alors $E(sx_j u) = E(s)E(x_j u) = 0$, car $E(x_j u) = 0$ sous (17.42). De ce fait, si nous partons d'un échantillon aléatoire et que nous supprimons des observations aléatoirement, les MCO sont toujours convergents. Les MCO ne sont toujours pas biaisés dans ce cas, tant qu'il n'y a pas de multicolinéarité exacte dans l'échantillon sélectionné.

Si s dépend des variables explicatives et de termes d'erreurs supplémentaires indépendants de $\mathbf{x}$ et u, les MCO sont toujours convergents et non biaisés. Par exemple, supposons que le QI soit une variable explicative de l'équation de salaire, mais qu'il manque pour certaines personnes. Supposons que nous pensions que la sélection peut être décrite par $s = 1$ si $QI \geq v$, et $s = 0$ si $QI \leq v$, v étant une variable aléatoire inobservée indépendante de QI, u, et des autres variables explicatives. Nous avons plus de chances d'observer les grands QI, mais il y a toujours une chance d'observer tout QI. Conditionnellement aux variables explicatives, s est indépendant de u, ce qui veut dire que $E(u|x_1, \ldots, x_k, s) = E(u|x_1, \ldots, x_k)$, et la dernière espérance est zéro par hypothèse sur le modèle (17.42). Si en plus nous supposons l'homoscédasticité ($E(u^2|\mathbf{x},s) = E(u^2) = \sigma^2$), alors les écarts-types estimés habituels des MCO et leurs statistiques de test sont valides.

Nous avons montré jusqu'ici plusieurs situations dans lesquelles les estimateurs des MCO sur l'échantillon sélectionné ne sont pas biaisés ou sont au moins convergents. Quand les MCO sur l'échantillon sélectionné ne sont-ils pas convergents ? Nous avons déjà vu un exemple : une régression utilisant un échantillon tronqué. Quand la troncature se fait par le haut, $s_i = 1$ si $y_i \leq c_i$, où c_i est le seuil de troncature. De manière équivalente, $s_i = 1$ si $u_i \leq c_i - \mathbf{x}_i\boldsymbol{\beta}$. Comme si dépend directement de u_i, s_i et u_i seront corrélées, même conditionnellement à x_i. Les MCO sur l'échantillon sélectionné n'estimeront pas β_j de manière convergente. Il y a d'autres raisons moins évidentes pour que s et u soient corrélées ; nous en discuterons dans la sous-section suivante.

Les résultats sur la convergence des MCO se généralisent aux estimations par variables instrumentales. Si les instruments sont notés z_h dans la population, la condition clé pour que les estimateurs des DMC soient convergents est $E(sz_h u) = 0$, ce qui est vrai quand $E(u|\mathbf{z},s) = 0$. En conséquence, si la sélection est entièrement déterminée par les variables exogènes $\mathbf{z}$, ou si s dépend d'autres facteurs indépendants de u et de $\mathbf{z}$, alors les estimateurs des DMC sur l'échantillon sont généralement convergents. Nous avons toujours besoin de supposer que les variables explicatives et les instruments soient suffisamment corrélés dans l'échantillon sélectionné. Wooldridge (2010, chapitre 19) contient une écriture précise de ces hypothèses.

On peut aussi montrer que quand la sélection est entièrement fonction des variables exogènes, l'estimation par maximum de vraisemblance d'un modèle non linéaire – comme le probit ou le logit – produit des estimateurs convergents, asymptotiquement normaux, et que les écarts-types estimés habituels ainsi que les statistiques de test sont valables [Voir encore Wooldridge (2010, chapitre 19)].

Troncature auxiliaire

Comme nous l'avons vu plus haut, la troncature auxiliaire est une forme fréquente de sélection de l'échantillon. Commençons encore avec le modèle (17.42) valable sur toute la population. Nous supposerons cependant que nous observons toujours les variables explicatives x_j. Le problème est que nous n'observons y que pour une partie de l'échantillon. La règle déterminant si nous observons y ne dépend *pas* directement de y. L'exemple le plus fréquent est quand $y = \log(wage^o)$, où $wage^o$ est le *salaire offert*, ou salaire horaire, qu'un individu pourrait avoir s'il travaillait. Si la personne travaille au moment de l'enquête, nous observons le salaire offert puisque nous supposons qu'il s'agit du salaire observé. Mais pour les personnes qui ne travaillent pas, nous ne pouvons pas observer $wage^o$. La troncature du salaire offert est donc *auxiliaire*, puisqu'elle dépend d'une variable auxiliaire, en l'occurrence le fait d'être salarié. Nous observons en général toutes les autres observations sur un individu, comme le niveau d'instruction, l'expérience précédente, la situation familiale, entre autres.

L'approche habituelle pour la troncature auxiliaire est d'ajouter explicitement une équation de sélection au modèle sur la population d'intérêt :

$$y = \mathbf{x}\boldsymbol{\beta} + u, \ \mathrm{E}(u|\mathbf{x}) = 0 \qquad \textbf{[17.46]}$$

$$s = 1[\mathbf{z}\boldsymbol{\gamma} + v \geq 0], \qquad \textbf{[17.47]}$$

où $s = 1$ si nous observons y, et zéro sinon. Nous supposerons que les éléments de $\mathbf{x}$ et $\mathbf{z}$ sont toujours observés et écrivons $\mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ et $\mathbf{z}\boldsymbol{\gamma} = \gamma_0 + \gamma_1 z_1 + ... + \gamma_m z_m$.

L'équation d'intérêt est (17.46), et nous pourrions estimer β par les MCO sur un échantillon aléatoire. L'équation de sélection (17.47) dépend de variables observables z_h et d'une erreur inobservée v. Nous ferons l'hypothèse standard que $\mathbf{z}$ est exogène dans (17.46) :

$$\mathrm{E}(u|\mathbf{x},\mathbf{z}) = 0.$$

En effet, pour que les méthodes proposées ci-dessous marchent bien, nous voudrons que $\mathbf{x}$ soit un sous-ensemble strictement inclus dans $\mathbf{z}$: tout x_j est aussi un élément de $\mathbf{z}$, et il doit y avoir des éléments de $\mathbf{z}$ qui ne sont pas dans $\mathbf{x}$. Nous verrons plus tard pourquoi c'est si important.

Le terme d'erreur v dans l'équation de sélection de l'échantillon est supposé indépendant de $\mathbf{z}$ (et donc de $\mathbf{x}$). Nous supposerons également que v suit une loi normale standardisée. On peut facilement voir qu'une corrélation entre u et v cause un problème de sélection de l'échantillon. Pour le voir, supposons que (u, v) soit indépendant de $\mathbf{z}$. Dans ce cas, en prenant l'espérance de (17.46) conditionnellement à $\mathbf{z}$ et y, et en utilisant le fait que $\mathbf{x}$ est un sous-ensemble de $\mathbf{z}$, on trouve

$$\mathrm{E}(y|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + \mathrm{E}(u|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + \mathrm{E}(u|v),$$

où $\mathrm{E}(u|\mathbf{z},v) = \mathrm{E}(u|v)$ puisque (u, v) est indépendant de $\mathbf{z}$. Si u et v suivent une loi normale jointe (de moyenne nulle), alors il existe ρ tel que $\mathrm{E}(u|v) = \rho v$. Donc

$$\mathrm{E}(y|z,v) = \mathbf{x}\boldsymbol{\beta} + \rho v.$$

Nous n'observons pas v, mais nous pouvons utiliser cette équation pour calculer $\mathrm{E}(y|\mathbf{z},s)$ et ensuite se restreindre à $s = 1$. Nous avons donc

$$\mathrm{E}(y|\mathbf{z},s) = \mathbf{x}\boldsymbol{\beta} + \rho\mathrm{E}(v|\mathbf{z},s).$$

Comme s et v sont liés par (17.47), et comme v suit une loi normale standardisée, nous pouvons montrer que $\mathrm{E}(v|\mathbf{z},s)$ est simplement l'inverse du ratio de Mills, $\lambda(\mathbf{z}\boldsymbol{\gamma})$, quand $s = 1$. Cela donne l'équation importante

$$E(y|\mathbf{z},s = 1) = \mathbf{x}\beta + \rho\lambda(\mathbf{z}\gamma). \quad \textbf{[17.48]}$$

L'équation (17.48) montre que l'espérance de y conditionnellement à $\mathbf{z}$ *et* à l'observation de y est égale à $\mathbf{x}\beta$ plus un terme additif qui dépend de l'inverse du ratio de Mills en $\mathbf{z}\gamma$. Souvenez-vous que nous voulons estimer β. Cette équation montre que nous pouvons le faire en utilisant seulement l'échantillon sélectionné, si nous ajoutons le terme $\lambda(\mathbf{z}\gamma)$ comme régresseur supplémentaire.

Si $\rho = 0$, $\lambda(\mathbf{z}\gamma)$ n'apparaît pas, et les MCO de y sur x sur l'échantillon sélectionné estiment β de manière convergente. Sinon, il y a une variable omise $\lambda(\mathbf{z}\gamma)$, qui est généralement corrélée à $\mathbf{x}$. Quand est-ce que $\rho = 0$? Quand u et v ne sont pas corrélées.

Comme γ est inconnu, nous ne pouvons pas évaluer $\lambda(\mathbf{z}_i\gamma)$ pour chaque i. Cependant, étant donné les hypothèses ci-dessus, s suit un modèle probit en fonction de z :

$$P(s = 1|z) = \Phi(\mathbf{z}\gamma). \quad \textbf{[17.49]}$$

En conséquence, g peut être estimé par un modèle probit de s_i sur z_i, en utilisant tout l'échantillon. Dans une seconde étape, on peut estimer b. Nous résumons cette procédure, récemment surnommée Heckit dans la littérature économétrique suite aux travaux d'Heckman (1976).

Correction pour la sélection de l'échantillon

i. En utilisant toutes les n observations, estimer un modèle probit de s_i sur $\mathbf{z}_i$ et trouver l'estimateur $\hat{\gamma}_h$. Calculer le ratio de Mills inversé $\hat{\lambda}_i = \lambda(\mathbf{z}_i\hat{\gamma})$ pour tous les i (En réalité, nous n'en avons besoin que si $s_i = 1$).

ii. Sur l'échantillon sélectionné, c'est-à-dire quand $s_i = 1$ (pour n_1 observations), estimer la régression

$$y_i \text{ sur } \mathbf{x}_i, \hat{\lambda}_i. \quad \textbf{[17.50]}$$

Les $\hat{\beta}_j$ sont convergents et asymptotiquement normaux.

Un test simple pour l'existence d'un biais de sélection vient de l'équation (17.50). En effet, la statistique t de $\hat{\lambda}_i$ est un test de $H_0 : \rho = 0$. Sous H_0, il n'y a pas de problème de sélection de l'échantillon.

Quand $\rho \neq 0$, les écarts-types estimés habituels des MCO venant de (17.50) ne sont pas corrects. Cela vient du fait qu'ils ne tiennent pas en compte l'estimation de γ, qui utilise les observations de la régression (17.50) parmi d'autres. Certains logiciels d'économétrie calculent des écarts-types estimés corrigés [Malheureusement, ce n'est pas aussi simple que la correction pour l'hétéroscédasticité. Voir Wooldridge (2010, chapitre 6) pour plus de détails]. Dans beaucoup de cas, l'ajustement change peu les résultats, mais il est difficile de le prédire sans vérifier (sauf si $\hat{\rho}$ est petit et n'est pas significatif).

Nous avons dit plus haut que $\mathbf{x}$ devait être strictement inclus dans $\mathbf{z}$. Cela a deux conséquences. Tout d'abord, toute variable explicative de (17.46) doit également être une variable explicative de l'équation de sélection. Bien que dans de rares cas, cela semble logique d'exclure des éléments de l'équation de sélection, inclure tous les éléments de $\mathbf{x}$ dans $\mathbf{z}$ est peu coûteux ; les exclure à tort peut mener à des estimateurs non convergents.

La seconde conséquence est que nous avons au moins un élément de $\mathbf{z}$ qui n'est pas dans $\mathbf{x}$. Cela veut dire que nous avons besoin d'une variable qui affecte la sélection mais n'a *pas* d'effet marginal sur y. Cela n'est pas absolument nécessaire pour appliquer la procédure – nous pouvons faire les deux étapes quand $\mathbf{z} = \mathbf{x}$ – mais les résultats sont généralement peu convaincants s'il n'y a pas de *restriction d'exclusion* dans (17.46). La raison est qu'alors que le ratio de Mills inversé est une fonction non linéaire de $\mathbf{z}$; on peut souvent l'approximer relativement bien par une fonction linéaire. Si $\mathbf{z} = \mathbf{x}$, $\hat{\lambda}_i$ peut être très corrélé avec les $\mathbf{x}_i$. Ce

genre de multicolinéarité peut mener à des écarts-types des $\hat{\beta}_j$ très élevés. Intuitivement, si nous n'avons pas de variable qui affecte la sélection mais pas y, il est très difficile, sinon impossible, de distinguer la sélection de l'échantillon d'une forme fonctionnelle mal spécifiée dans (17.46).

EXEMPLE 17.5
Salaire potentiel des femmes mariées

Nous appliquons la correction pour la sélection de l'échantillon aux données sur les femmes mariées de MROZ. Souvenons-nous que sur les 753 femmes de l'échantillon, seulement 428 ont été salariées durant l'année. L'équation de salaire est standard, log(*wage*) est la variable dépendante et *educ*, *exper* et *exper*^{2} sont les variables explicatives. Pour tester et corriger un éventuel biais de sélection – lié à l'observabilité du salaire potentiel des femmes non-salariées – il nous faut estimer un modèle probit prédisant la participation à la force de travail. En plus de l'éducation et des variables d'expérience, nous incluons dans le tableau 17.1 les autres sources de revenu, l'âge, le nombre de jeunes enfants et d'autres enfants. Ces quatre variables sont exclues de l'équation de salaire par *hypothèse* : nous supposons que, les facteurs de productivité étant donné, *nwifeinc*, *age*, *kidslt6* et *kidsge6* n'affectent pas le salaire potentiel. Au vu des résultats du probit du tableau 17.1, il est clair qu'au moins *age* et *kidslt6* ont un effet sur la participation à la force de travail.

Le tableau 17.7 contient les résultats par les MCO et par la procédure Heckit [Les écarts-types estimés rapportés pour le Heckit sont simplement les écarts-types estimés standard des MCO de la régression (17.50). Il n'y a pas d'indice de problème de sélection de l'échantillon dans l'estimation de l'équation de salaire potentiel. Le coefficient de $\hat{\lambda}$ a une statistique t très faible (0,239), nous ne pouvons donc pas rejeter $H_0 : \rho = 0$. Tout aussi important, il n'y a en pratique pas de grandes différences entre les coefficients des autres variables dans le tableau 17.7. Les rendements de l'instruction estimés ne diffèrent que d'un dixième de point de pourcentage.

Tableau 17.7 *Salaire potentiel des femmes mariées*

Variable dépendante : log(*wage*)		
Variables explicatives	**MCO**	**Heckit**
Educ	0,108 (0,014)	0,109 (0,016)
exper	0,042 (0,012)	0,044 (0,016)
exper^2	−0,00081 (0,00039)	−0,00086 (0,00044)
constant	−0,522 (0,199)	−0,578 (0,307)
$\hat{\lambda}$	–	0,032 (0,134)
Nombre d'observations *R*-carré	428 0,157	428 0,157

Une alternative à la procédure en deux étapes ci-dessus est l'estimation complète par maximum de vraisemblance. Cette approche est plus compliquée, car elle demande de trouver la distribution jointe de y et s. Il est parfois intéressant de tester s'il y a un problème de sélection avec la procédure ci-dessus ; s'il n'y a pas de signe de problème de sélection, il n'y a pas de raison d'aller plus loin. Si nous détectons un problème de sélection de l'échantillon, nous pouvons soit utiliser les estimateurs en deux étapes, soit estimer la régression et l'équation de sélection de manière jointe par EMV [Voir Wooldridge (2010, chapitre 19)].

Dans l'exemple 17.5, nous ne savons pas seulement si une femme a travaillé durant l'année : nous savons aussi combien d'heures a travaillé chaque femme. Il est également possible d'utiliser cette information dans une autre procédure d'estimation. Au lieu d'utiliser le ratio de Mills inversé $\hat{\lambda}_i$, nous pouvons utiliser les résidus du modèle Tobit $\hat{v}_i$, donnés par $\hat{v}_i = y_i - x_i\hat{\beta}$ quand $y_i > 0$. On pourrait montrer que la régression (17.50) en remplaçant $\hat{\lambda}_i$ par $\hat{v}_i$ donne également des estimateurs convergents de β_j, et que la statistique t de $\hat{v}_i$ est un test valide pour l'existence d'un biais de sélection. Cette approche a l'avantage d'utiliser plus d'information, elle n'est cependant pas toujours applicable [Voir Wooldridge (2010, chapitre 19)].

Il y a beaucoup d'autres modèles concernant les biais de sélection. Il est utile de mentionner également les modèles avec des variables explicatives endogènes *en plus* d'un biais de sélection possible. On peut écrire un modèle avec une seule variable explicative endogène :

$$y_1 = \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\beta}_1 + u_1, \qquad \textbf{[17.51]}$$

où y_1 est observé seulement quand s = 1, et y_2 peut éventuellement n'être observé que quand y_1 est observé. Par exemple, y_1 peut être le pourcentage de votes reçus par le candidat sortant, y_2 le pourcentage des dépenses de campagne qu'il a causées. Pour les sortants qui ne se représentent pas, on ne peut observer ni y_1 ni y_2. Si des facteurs exogènes affectent la décision de se représenter et sont corrélés avec les dépenses de campagne, il est possible d'estimer α_1 et les éléments de $\boldsymbol{\beta}_1$ de manière convergente par les variables instrumentales. Pour être convaincant, la méthode requiert *deux* variables exogènes n'apparaissant pas dans (17.51). En effet, une doit affecter la décision de participation, et l'autre doit être corrélée à y_2 [la condition nécessaire pour estimer (17.51) par les DMC]. Brièvement, la méthode est d'estimer l'équation de sélection par probit, où *toutes* les variables exogènes apparaissent dans l'équation du probit. Ensuite, on ajoute l'inverse du ratio de Mills à (17.51), et on estime cette équation par les DMC. L'inverse du ratio de Mills est son propre instrument, comme il dépend seulement de variables exogènes. Les variables exogènes sont les autres instruments. Comme précédemment, la statistique t de $\hat{\lambda}_i$ peut constituer un test d'existence d'un biais de sélection [Voir Wooldridge (2010, chapitre 19) pour plus d'informations].

RÉSUMÉ

Dans ce chapitre, nous avons couvert plusieurs méthodes avancées souvent utilisées dans les applications, surtout en microéconomie. Les modèles logit et probit sont utilisés pour les variables à réponse binaire. Ces modèles ont plusieurs avantages par rapport aux modèles de probabilités linéaires : les probabilités prédites sont toujours entre zéro et un, et les effets marginaux ne sont pas constants. L'inconvénient principal des modèles logit et probit est qu'ils sont plus difficiles à interpréter.

Le modèle Tobit s'applique aux variables dépendantes positives avec un point de masse en zéro, et qui prennent aussi un certain nombre de valeurs positives. Beaucoup de variables individuelles, comme l'offre de travail, la valeur des assurances-vie, le montant de l'épargne de retraite, ont ces caractéristiques. Comme pour les logit et probit, les valeurs espérées de y conditionnellement à $\mathbf{x}$ (que ce soit conditionnellement à $y > 0$ ou non) dépendent à la fois de $\mathbf{x}$ et de $\boldsymbol{\beta}$ de manière non linéaire. Nous avons donné les expressions de ces espérances ainsi que des formules pour mesurer les effets partiels des x_j sur ces espérances. Ces grandeurs peuvent être estimées après qu'un modèle Tobit ait été estimé par maximum de vraisemblance.

Quand la variable dépendante est une variable de comptage – une variable qui prend des valeurs entières positives – une régression de Poisson est plus adaptée. La valeur espérée de y en fonction de x_j est exponentielle. Les paramètres s'interprètent donc comme des semi-élasticités ou comme des élasticités, selon que x_j soit ou non sous forme logarithmique. Les paramètres s'interprètent *comme s'*ils étaient dans un modèle

linéaire avec une variable dépendante log(y). Les paramètres peuvent être estimés par EMV. Cependant, comme la distribution de Poisson impose l'égalité de la variance et de la moyenne, il est souvent nécessaire de calculer des écarts-types estimés qui permettent de la sur-dispersion ou de la sous-dispersion. Ce sont des ajustements simples des écarts-types estimés et des statistiques correspondant aux EMV.

Les modèles de régression censurée et tronquée traitent des types spécifiques de problèmes de données manquantes. Dans les régressions censurées, la variable dépendante est censurée au-dessus ou en-dessous d'un seuil donné. Nous pouvons utiliser de l'information sur les données censurées car nous observons toujours les variables explicatives, comme dans les applications à des durées ou avec des observations censurées à droite. On peut utiliser une régression tronquée quand une partie de l'échantillon est totalement exclue : nous n'observons rien au sujet des individus qui ne sont pas dans le plan d'échantillonnage. C'est un cas particulier des problèmes de sélection.

La section 17.5 donne un traitement systématique des problèmes de sélection non aléatoire de l'échantillon. Nous avons montré qu'une sélection exogène de l'échantillon n'affecte pas la convergence des estimateurs des MCO appliqués sur l'échantillon disponible, mais que la sélection endogène l'affecte. Nous avons montré comment on peut corriger le biais de sélection dans le cas de la troncature auxiliaire, où des observations de y ne sont pas disponibles à cause d'une autre variable (comme la participation à la force de travail). La méthode d'Heckman est relativement facile à mettre en œuvre dans ces cas.

MOTS-CLÉS

Analyse des durées p. 707
Censure à droite p. 706
Distribution de Poisson p. 702
Échantillon sélectionné p. 712
Effet Marginal Moyen (EMM) p. 688
Effet Marginal au Point Moyen (EMPM) p. 687
Effet Partiel Moyen (EPM) p. 688
Estimateur du Maximum de Vraisemblance (EMV) p. 684
Estimateur du Quasi-Maximum de Vraisemblance (EQMV) p. 703
Fonction de log-vraisemblance p. 684
Inverse du ratio de Mills p. 695
Méthode Heckit p. 715
Modèles à réponse binaire p. 681
Modèle à Variable Latente (MVL) p. 681
Modèle de régression censurée p. 706
Modèle de régression censurée normale p. 706
Modèle de régression de Poisson p. 702
Modèle de régression tronquée p. 706
Modèle de régression tronquée normale p. 710
Modèle logit p. 681
Modèle probit p. 681
Modèle Tobit p. 693
Pourcentage de prédictions correctes p. 686
Probabilité de réponse p. 681
Pseudo R-Carré p. 687
Réponse avec solution en coin p. 680
Sélection de l'échantillon exogène p. 713
Sélection non aléatoire de l'échantillon p. 711

PROBLÈMES

1. i. Pour une réponse binaire y, soit $\bar{y}$ la proportion de un dans l'échantillon (qui est la moyenne de y_i). Soit $\hat{q}_0$ le pourcentage de prédictions correctes quand $y = 0$ et $\hat{q}_1$ le pourcentage de prédictions correctes quand $y = 1$. Si $\hat{p}$ est le pourcentage total de prédictions correctes, montrez que $\hat{p}$ est une moyenne pondérée de $\hat{q}_0$ et $\hat{q}_1$:

$$\hat{p} = (1 - \bar{y})\hat{q}_0 + \bar{y}\hat{q}_1.$$

ii. Dans un échantillon de 300 observations, supposons que $\bar{y} = 0{,}070$, il y a donc 210 observations avec $y_i = 1$ et 90 avec $y_i = 0$. Supposons que le pourcentage de prédictions correctes quand $y = 0$ est 80 %, et qu'il est de 40 % quand $y = 1$. Trouver le pourcentage total de prédictions correctes.

2. Soit *grad* une variable indicatrice prédisant si un étudiant athlète d'une université américaine obtient son diplôme en cinq ans. Soient *hsGPA* et *SAT* respectivement sa moyenne de lycée et son score aux examens d'entrée. Soit *study* le nombre d'heures passées par semaine dans la salle d'études. Supposons que l'on obtienne le modèle logit suivant en utilisant les données de 420 étudiants athlètes :

$$\hat{P}(grad = 1|hsGPA, SAT, study) = \Lambda(-1{,}17 + 0{,}24\ hsGPA + 0{,}00058\ SAT + 0{,}073\ study),$$

où $\Lambda(z) = \exp(z)/[1 + \exp(z)]$ est la fonction logistique. Si *hsGPA* est fixé à 3,0 et *SAT* est fixé à 1 200, calculez la différence de probabilité d'obtenir son diplôme entre une personne passant 10 heures par semaine en salle d'étude et une personne y passant 5 heures par semaine.

3. (Nécessite des Calculs)

i. Supposons dans un modèle Tobit que $x_1 = \log(z_1)$ et que ce soit la seule fois que z_1 apparaît dans $\mathbf{x}$. Montrez que

$$\frac{\partial E(y|y > 0, \mathbf{x})}{\partial z_1} = (\beta_1/z_1)\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\}, \qquad \textbf{[17.52]}$$

où β_1 est le coefficient de $\log(z_1)$.

ii. Si $x_1 = z_1$ et $x_2 = z_1^2$, montrez que

$$\frac{\partial E(y|y > 0, \mathbf{x})}{\partial z_1} = (\beta_1 + 2\beta_2 z_1)\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\},$$

où β_1 est le coefficient de z_1 et β_2 est le coefficient de z_1^2.

4. Soit mvp_i la valeur marginale du travailleur i, c'est-à-dire le produit de sa productivité marginale par le prix unitaire du bien fabriqué. Supposons que

$$\log(mvp_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$
$$wage_i = \max(mvp_i, minwage_i),$$

où les variables explicatives incluent entre autres l'éducation et l'expérience, et $minwage_i$ est le salaire minimum versé par l'État américain à la personne i. Écrivez $\log(wage_i)$ en fonction de $\log(mvp_i)$ et de $\log(minwage_i)$.

5. (Nécessite des Calculs) Soit *patents* le nombre de brevets déposés par une entreprise sur une année. Supposons que l'espérance de *patents* étant donné *sales* et *RD* soit

$$\mathrm{E}(patents|sales, RD) = \exp[\beta_0 + \beta_1 \log(sales) + \beta_2 RD + \beta_3 RD^2],$$

où *sales* est le montant annuel des ventes de l'entreprise, et *RD* la dépense totale en recherche et développement sur les 10 dernières années.

i. Comment estimeriez-vous β_j ? Justifiez votre réponse en discutant de la forme de la variable *patents*.

ii. Comment interpréteriez-vous β_1 ?

iii. Trouvez l'effet partiel de *RD* sur $\mathrm{E}(patents|sales, RD)$.

6. Considérons une fonction d'épargne familiale valable pour les familles des États-Unis :

$$sav = \beta_0 + \beta_1 inc + \beta_2 hhsize + \beta_3 educ + \beta_4 age + u,$$

où *hhsize* est la taille du ménage, *educ* est le nombre d'années d'éducation du chef de ménage, et *age* est l'âge du chef de ménage. Supposons que $\mathrm{E}(u|inc,hhsize,educ,age) = 0$.

i. Supposons que l'échantillon n'inclue que les familles dont le chef a plus de 25 ans. Si nous utilisons les MCO sur cet échantillon, obtiendra-t-on des estimateurs biaisés de β_j ? Expliquez.

ii. Maintenant, supposons que l'échantillon n'inclue que les couples mariés sans enfants. Peut-on estimer tous les paramètres de l'équation d'épargne ? Lesquels peut-on estimer ?

iii. Supposons que les familles qui épargnent plus de 25 000 $ par an soient exclues de l'échantillon. Les MCO produisent-ils des estimateurs convergents de β_j ?

7. Supposons que vous soyez embauché par une université pour étudier les facteurs qui déterminent si un étudiant admis dans l'université assiste aux cours. Vous obtenez un grand échantillon aléatoire d'étudiants admis l'année précédente. Vous savez si chacun de ces étudiants s'est rendu en cours, les notes qu'il a obtenues au lycée, le revenu familial, l'obtention de bourses, l'origine ethnique, et des variables géographiques. Quelqu'un vous explique : « Toute analyse menée sur ces données sera biaisée parce que ce n'est pas un échantillon aléatoire des candidats à l'université, mais seulement de ceux qui candidatent à cette université. » Que pensez-vous de cette critique ?

EXERCICES SUR ORDINATEUR

C1. Utilisez les données de PNTSPRD pour cet exercice.

i. La variable *favwin* est une variable binaire prenant la valeur un si le match est gagné par l'équipe favorite – selon les cotes de paris à Las Vegas. Un modèle de probabilité linéaire estimant les chances que l'équipe favorite gagne en fonction de l'écart des cotes *spread* est

$$\mathrm{P}(favwin = 1|spread) = \beta_0 + \beta_1 spread.$$

Expliquez pourquoi, si l'écart des cotes inclut toutes les informations pertinentes, on s'attend à ce que $\beta_0 = 0{,}5$.

ii. Estimez le modèle de la question (i) par les MCO. Testez $\mathrm{H}_0 : \beta_0 = 0{,}5$ contre une hypothèse alternative bilatérale. Utilisez à la fois les écarts-types estimés par défaut et ceux robustes à l'hétéroscédasticité.

iii. Est-ce que *spread* est statistiquement significatif ? Quelle est la probabilité estimée que l'équipe favorite gagne quand *spread* = 10 ?

iv. Maintenant, estimons un modèle probit prédisant P(*favwin* = 1|*spread*). Interprétez et testez l'hypothèse nulle que la constante est nulle [Indice : souvenez-vous que $\Phi(0) = 0{,}5$].

v. Utilisez le modèle probit pour estimer la probabilité que l'équipe favorite gagne quand *spread* = 10. Comparez ceci avec le MPL de la question (iii).

vi. Ajoutez les variables *favhome*, *fav25* et *und25* au modèle probit et testez la significativité jointe de ces variables en utilisant un test du rapport des vraisemblances (Combien de *dl* y a-t-il dans la distribution du chi-deux ?). Interprétez ce résultat en discutant si l'écart des cotes incorpore toute l'information disponible avant un match.

C2. Utilisez les données de LOANAPP dans cet exercice ; voyez aussi l'Exercice sur Ordinateur C8 du chapitre 7.

i. Estimez un modèle probit prédisant *approve* à partir de *white*. Trouvez la probabilité qu'un prêt soit approuvé pour les blancs et les non-blancs. Comment ces estimations se comparent-elles à celles d'un modèle de probabilités linéaires ?

ii. Ajoutez maintenant les variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2*, et *vr* au modèle probit. Y a-t-il des indices statistiquement significatifs de discrimination contre les non-blancs ?

iii. Estimez le modèle de la question (ii) avec un logit. Comparez le coefficient de *white* avec celui estimé par probit.

iv. Utilisez l'équation (17.17) pour mesurer l'ampleur des discriminations par logit et probit.

C3. Utilisez les données de FRINGE pour cet exercice.

i. Pour quel pourcentage des travailleurs de l'échantillon *pension* est-il nul ? Quelle est la plage des valeurs prises par *pension* pour les travailleurs avec des retraites strictement positives ? Pourquoi un modèle Tobit est-il adapté pour modéliser *pension* ?

ii. Estimez un modèle Tobit pour prédire *pension* à partir de *exper*, *age*, *tenure*, *educ*, *depends*, *married*, *white* et *male*. Les blancs et les hommes ont-ils des retraites espérées significativement plus grandes ?

iii. Utilisez les résultats de la question (ii) pour estimer la différence de retraite espérée entre un homme blanc et une femme non-blanche ayant tous deux 35 ans, célibataires et sans personne à charge, ayant 16 années d'éducation et 10 années d'expérience.

iv. Ajoutez *union* au modèle Tobit et commentez sa significativité.

v. Appliquez le modèle Tobit de la partie (iv) avec *peratio*, le ratio retraite sur revenu d'activités, comme variable dépendante (Notez que cette fraction est entre zéro et un, mais bien qu'elle prenne souvent la valeur zéro, elle ne s'approche jamais de un. Le modèle Tobit est donc une bonne approximation ici). Le sexe ou l'origine ethnique affectent-ils le ratio *peratio* ?

C4. Dans l'Exemple 9.1, nous avions ajouté les termes quadratiques $pcnv^2$, $ptime86^2$, et $inc86^2$ à un modèle linéaire prédisant *narr86*.

i. Utilisez les données de CRIME1 pour ajouter ces termes à la régression de Poisson de l'Exemple 17.3.

ii. Calculez l'estimateur de σ^2 donné par $\hat{\sigma}^2 = (n - k - 1)^{-1} \sum_{i=1}^{n} \hat{u}_i^2 / \hat{y}_i$. Y a-t-il des preuves de sur-dispersion ? Comment les écarts-types estimés de l'EMV de Poisson devraient-ils être ajustés ?

iii. Utilisez les résultats des parties (i) et (ii) et du tableau 17.5 pour calculer la statistique du rapport des quasi-vraisemblances pour la significativité jointe des trois termes quadratiques. Qu'en concluez-vous ?

C5. Référez-vous au tableau 13.1 du chapitre 13. Pour le construire, nous avons utilisé les données de FERTIL1 pour estimer un modèle linéaire prédisant *kids*, le nombre d'enfants d'une femme.

i. Estimez une régression de Poisson prédisant *kids*, en utilisant les mêmes variables que dans le tableau 13.1. Interprétez le coefficient de *y82*.

ii. En gardant les autres variables constantes, quelle est la différence de fertilité en pourcentages entre une femme noire et une femme non-noire ?

iii. Trouvez $\hat{\sigma}$. Y a-t-il des preuves de sous-dispersion ou sur-dispersion ?

iv. Calculez les valeurs prédites de la régression de Poisson et trouvez le *R*-carré qui est le carré de la corrélation entre $kids_i$ et $\widehat{kids}_i$. Comparez ceci avec le *R*-carré du modèle de régression linéaire.

C6. Utilisez les données de RECID pour estimer le modèle de l'Exemple 17.4 par les MCO, en utilisant *seulement* les 552 durées non-censurées. Comparez ces estimations avec celles du tableau 17.6 et commentez ces comparaisons.

C7. Utilisez MROZ pour cet exercice.

i. En utilisant les 428 femmes qui étaient dans la population active, estimez les rendements de l'éducation par les MCO en incluant comme variables explicatives *exper*, $exper^2$, *nwifeinc*, *age*, *kidslt6*, et *kidsge6*. Donnez l'estimation de l'effet d'*educ* et son écart-type estimé.

ii. Estimez maintenant les rendements de l'éducation par Heckit, avec toutes les variables exogènes dans la régression de seconde étape. En d'autres termes, la régression d'intérêt prédit log(*wage*) à partir de *educ*, *exper*, $exper^2$, *nwifeinc*, *age*, *kidslt6*, *kidsge6*, et $\hat{\lambda}$. Comparez les rendements estimés de l'éducation *et* leurs écarts-types estimés avec ceux de la question (i).

iii. En utilisant les 428 femmes actives, régressez $\hat{\lambda}$ sur *educ*, *exper*, $exper^2$, *nwifeinc*, *age*, *kidslt6*, et *kidsge6*. Le *R*-carré est-il grand ? Cela aide-t-il à expliquer le résultat de la question (ii) ? (*Indice* : pensez à la multicolinéarité)

C8. Le fichier JTRAIN2 contient des données sur une expérience de formation à la recherche d'emploi pour un groupe d'hommes. Les hommes pouvaient intégrer le programme de janvier 1976 à mi-1977 environ. Le programme s'est arrêté en décembre 1977. Le but est de tester si la participation au programme de formation a eu un effet sur les chances de chômage et les revenus en 1978.

i. La variable *train* indique la participation au programme. Combien d'hommes de l'échantillon ont-ils participé au programme ? Quelle est la plus longue durée de participation d'un individu au programme ?

ii. Régressez linéairement *train* sur les variables démographiques et préprogramme *unem74*, *unem75*, *age*, *educ*, *black*, *hisp*, et *married*. Ces variables sont-elles significatives de manière jointe au seuil de 5 % ?

iii. Estimez une version probit du modèle de la question (ii). Calculez le test de rapport de vraisemblance pour la significativité jointe de toutes les variables. Que concluez-vous ?

iv. À partir des réponses aux questions (ii) et (iii), la participation au programme de formation peut-elle être considérée comme exogène pour expliquer le chômage en 1978 ? Expliquez.

v. Faites une régression simple de *unem78* sur *train* et rapportez le résultat sous forme d'équation. Quel est l'effet estimé de la participation au programme sur la probabilité d'être au chômage en 1978 ? Est-ce statistiquement significatif ?

vi. Faites un probit de *unem78* sur *train* et reportez les résultats sous forme d'équation. Y a-t-il un sens à comparer le coefficient de *train* avec celui du modèle linéaire de la question (v) ?

vii. Trouvez les probabilités prédites des questions (v) et (vi). Expliquez pourquoi elles sont identiques. Quelle approche utiliseriez-vous pour mesurer l'effet et la significativité statistique du programme de formation ?

viii. Ajoutez toutes les variables de la question (ii) comme variables explicatives aux modèles des questions (v) et (vi). Les probabilités prédites sont-elles identiques maintenant ? Quelle est la corrélation entre les deux ?

ix. En utilisant le modèle de la question (viii), estimez l'effet partiel moyen de *train* sur la probabilité de chômage en 1978. Utilisez (17.17) avec $c_k = 0$. Comparez les estimations avec les coefficients de la question (viii).

C9. Utilisez les données d'APPLE pour cet exercice. Ce sont des données d'enquête téléphonique essayant de mesurer la demande pour (hypothétique) une pomme « écologique ». On présentait à chaque famille un jeu de prix (tiré au sort) pour des pommes normales et des pommes labellisées écologiques. On leur demandait combien ils achèteraient de chaque type de pommes.

i. Sur les 660 familles de l'échantillon, combien disent-elles ne pas vouloir de pomme labellisée écologique au prix proposé ?

ii. La variable *ecolbs* semble-t-elle avoir une distribution continue sur les valeurs strictement positives ? Quelle est l'implication de votre réponse pour la validité d'un modèle Tobit pour prédire *ecolbs* ?

iii. Estimez un modèle Tobit prédisant *ecolbs* à partir des variables explicatives *ecoprc*, *regprc*, *faminc*, et *hhsize*. Quelles variables sont-elles significatives au seuil de 1 % ?

iv. *faminc* et *hhsize* sont-elles significatives de manière jointe ?

v. Les signes des coefficients des variables de prix sont-ils ceux que vous attendiez dans la question (iii) ? Expliquez.

vi. Soit β_1 le coefficient d'*ecoprc* et β_2 le coefficient de *regprc*. Testez l'hypothèse $H_0: -\beta_1 = \beta_2$ contre une alternative bilatérale. Donnez la *p*-valeur du test (Vous pouvez vous rapporter à la section 4.4 si votre logiciel d'économétrie ne calcule pas ce test directement).

vii. Trouvez les estimations d'E(*ecolbs*|**x**) pour toutes les observations de l'échantillon [Voir l'équation (17.25)]. Appelez-les $\widehat{ecolbs}_i$. Quelles sont les plus petites et les plus grandes valeurs prédites ?

viii. Calculez le carré de la corrélation entre $ecolbs_i$ et $\widehat{ecolbs}_i$.

ix. Estimez maintenant un modèle linéaire prédisant *ecolbs* à partir des variables explicatives de la question (iii). Pourquoi les estimations des MCO sont-elles si proches des estimations du Tobit ? En terme de qualité de la prédiction, le modèle Tobit est-il meilleur que le modèle linéaire ?

x. Que pensez-vous de la phrase suivante : « Comme le *R*-carré du modèle Tobit est tout petit, les effets-prix estimés ne sont probablement pas convergents. »

C10. Utilisez les données de SMOKE pour cet exercice.

i. La variable *cigs* est le nombre de cigarettes fumées par jour. Combien de personnes de l'échantillon ne fument pas du tout ? Quelle proportion des gens dit fumer 20 cigarettes par jour ? Pourquoi pensez-vous que beaucoup de personnes disent fumer exactement 20 cigarettes ?

ii. Au vu de vos réponses à la question (i), *cigs* est-il un bon candidat pour suivre une distribution de Poisson conditionnelle ?

iii. Estimez un modèle de régression de Poisson de *cigs*, en incluant les variables explicatives log(*cigpric*), log(*income*), *white*, *educ*, *age*, et age^2. Quelles sont les élasticités-prix et revenu estimées ?

iv. En utilisant les écarts-types estimés du maximum de vraisemblance, les variables de prix et de revenu sont-elles significatives au seuil de 5 % ?

v. Obtenez l'estimateur de σ^2 décrit après l'équation (17.35). Qu'est $\hat{\sigma}$? Comment ajusteriez-vous les écarts-types estimés de la question (iv) ?

vi. En utilisant les écarts-types estimés ajustés de la question (v), les élasticités-prix et revenu sont-elles maintenant significativement différentes de zéro ? Expliquez.

vii. Les variables d'âge et d'éducations sont-elles significatives en utilisant les écarts-types estimés robustes ? Comment interprétez-vous le coefficient d'*educ* ?

viii. Trouvez les valeurs prédites de la régression de Poisson, $\hat{y}_i$. Trouvez les valeurs maximales et minimales, et discutez si le modèle prédit bien les grandes consommations de cigarettes.

ix. En utilisant les valeurs prédites de la question (viii), trouvez le carré de la corrélation entre $\hat{y}_i$ et y_i.

x. Estimez un modèle linéaire pour prédire *cigs* par les MCO, en utilisant les variables explicatives (et les formes fonctionnelles) de la question (iii). Le modèle linéaire donne-t-il de meilleures prédictions que le modèle exponentiel ? L'un des deux *R*-carré est-il particulièrement grand ?

C11. Utilisez les données de CPS91 pour cet exercice. Ces données concernent les femmes mariées, et on trouve aussi des informations sur le revenu et les caractéristiques de leur mari.

i. Quelle proportion des femmes disent-elles travailler ?

ii. En utilisant seulement les données sur les femmes qui travaillent – vous n'avez pas le choix – estimez l'équation de salaire

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 black + \beta_5 hispanic + u$$

par les moindres carrés ordinaires. Reportez les résultats sous leur forme habituelle. Semble-t-il y avoir des différences de salaire significatives par origine ethnique ?

i. Estimez un modèle probit prédisant *inlf* qui inclut les variables explicatives de l'équation de salaire de la question (ii) ainsi que *nwifeinc* et *kidlt6*. Les coefficients de ces deux dernières variables ont-ils le signe attendu ? Sont-ils statistiquement significatifs ?

ii. Expliquez pourquoi, pour tester le biais de sélection et éventuellement corriger l'équation de salaire pour la participation à la force de travail, il est important que *nwifeinc* et *kidlt6* prédisent *inlf*. Que doit-on supposer sur le rôle de *nwifeinc* et *kidlt6* dans l'équation de salaire ?

iii. Calculez l'inverse du ratio de Mills (de chaque observation) et ajoutez-le en tant que régresseur supplémentaire dans l'équation de salaire de la question (ii). Quelle est la *p*-valeur non-directionnelle ? Est-ce si petit avec 3 286 observations ?

iv. L'ajout de l'inverse du ratio de Mills change-t-il beaucoup les coefficients de l'équation de salaire ? Expliquez.

C12. Utilisez les données de CHARITY pour répondre à ces questions.

i. La variable *respond* est une variable binaire qui vaut un si une personne a donné à la dernière sollicitation. La base de données ne contient que des personnes qui ont donné au moins une fois dans le passé. Quelle proportion de personnes ont-elles donné la dernière fois ?

ii. Estimez un modèle probit prédisant *respond* à partir des variables explicatives *resplast*, *weekslast*, *propresp*, *mailsyear*, et *avggift*. Quelles sont les variables explicatives statistiquement significatives ?

iii. Trouvez l'effet partiel moyen de *mailsyear* et comparez ce dernier avec le coefficient d'un modèle de probabilités linéaires.

iv. Utilisez les mêmes variables explicatives, estimez un modèle Tobit prédisant *gift*, le montant du don le plus récent (en florins néerlandais). Quelles variables explicatives sont-elles maintenant statistiquement significatives ?

v. Comparez l'EPM Tobit de *mailsyear* avec celui d'une régression linéaire. Sont-ils similaires ?

vi. Les estimations des questions (ii) et (iv) sont-elles vraiment compatibles avec un modèle Tobit ? Expliquez.

C13. Utilisez les données de HTV pour faire cet exercice.

i. En utilisant les MCO sur tout l'échantillon, estimez un modèle prédisant log(*wage*) à partir des variables explicatives *educ*, *abil*, *exper*, *nc*, *west*, *south*, et *urban*. Reportez le rendement de l'éducation estimé et son écart-type estimé.

ii. Estimez maintenant l'équation de la question (i) en utilisant seulement les personnes avec *educ* < 16. Quel pourcentage de l'échantillon perdez-vous ? Quel est maintenant le rendement estimé d'une année d'éducation ? Comment cela se compare-t-il à la question (i) ?

iii. Supprimez maintenant toutes les informations avec *wage* ≥ 20, de sorte à ce que tous ceux qui restent dans l'échantillon gagnent moins de 20 $ par heure. Faites la régression de la partie (i) et commentez le coefficient d'*educ* (Comme le modèle de régression normale tronquée suppose que y est continue, en théorie, peu importe que nous supprimions les observations avec *wage* ≥ 20 ou *wage* > 20. En pratique, y compris dans cette application, cela peut changer un peu car quelques personnes gagnent exactement 20 $ par heure).

iv. En utilisant l'échantillon de la question (iii), utilisez une régression tronquée [la troncature par le haut étant en log(20)]. La régression tronquée donne-t-elle les rendements de l'éducation sur la population entière, en supposant que les estimations de la question (i) soient convergentes ? Expliquez.

C14 Utilisez les données de HAPPINESS pour cet exercice. Voyez également l'Exercice sur ordinateur C15 du chapitre 13.

i. Estimez un modèle probit prédisant *vhappy* à partir de *ocattend* et *regattend*, et incluez toutes les variables indicatrices d'années. Trouvez l'effet partiel moyen de *occattend* et *regattend*. Comparez ces effets avec ceux tirés de l'estimation d'un modèle de probabilités linéaires.

ii. Définissez une variable, *highinc*, qui prend la valeur un si le revenu est au-dessus de 25 000 $. Incluez *highinc*, *unem*10, *educ*, et *teens* au modèle probit de la question (ii). L'effet partiel de *regattend* change-t-il beaucoup ? Et la significativité statistique ?

iii. Discutez l'EPM et la significativité statistique des quatre nouvelles variables de la question (ii). La taille des coefficients est-elle crédible ?

iv. En prenant en compte les facteurs de la question (ii), y a-t-il des différences de bonheur déclaré par sexe ou par origine ethnique ? Justifiez votre réponse.

C15. Utilisez les données de ALCOHOL, tirées de Terza (2002), pour cet exercice. Les données, sur 9 822 hommes, donnent les caractéristiques professionnelles, informent sur la consommation d'alcool, et comprennent également des variables démographiques et de contexte. Dans cette question, nous étudierons les effets de l'abus d'alcool sur *employ*, qui est une variable binaire valant un si la personne a un emploi. Si *employ* = 0, l'homme est soit inactif soit au chômage.

i. Quelle proportion de l'échantillon est-elle employée au moment de l'enquête ? Quelle fraction a abusé de l'alcool ?

ii. Faites une régression simple de *employ* sur *abuse* et rapportez les résultats sous la forme habituelle, en obtenant les écarts-types estimés robustes à l'hétéroscédasticité. Interprétez l'équation estimée. Vous attendiez-vous à ce résultat ? Est-il statistiquement significatif ?

iii. Faites un probit d'*employ* sur *abuse*. Le signe et la significativité statistique ont-ils changé depuis la question (ii) ? Comparez l'effet partiel moyen du probit avec celui d'un modèle de probabilités linéaires.

iv. Calculez les valeurs prédites du MPL estimé à la question (ii) et donnez les pour *abuse* = 0 et pour *abuse* = 1. Cela ressemble-t-il aux valeurs prédites du modèle probit ? Pourquoi ?

v. Ajoutez les variables *age*, *agesq*, *educ*, *educsq*, *married*, *famsize*, *white*, *northeast*, *midwest*, *south*, *centcity*, *outercity*, *qrt*1, *qrt*2, et *qrt*3 au MPL de la question (ii). Qu'advient-il du coefficient d'*abuse* et de sa significativité statistique ?

vi. Estimez un modèle probit en utilisant les variables de la question (v). Trouvez l'EPM d'*abuse* et sa statistique *t*. L'effet estimé est-il maintenant identique à celui du modèle linéaire ? Est-il « proche » ?

vii. Des variables décrivant l'état de santé général des hommes sont aussi inclues dans les données. Est-il évident qu'il faudrait ajouter ces variables comme variables explicatives ? Pourquoi ?

viii. Pourquoi *abuse* pourrait-il être considéré comme endogène dans l'équation prédisant *employ* ? Pensez-vous que les variables *mothalc* et *fathalc*, qui indiquent si la mère et le père de l'individu étaient alcooliques, sont des variables instrumentales possibles pour *abuse* ?

ix. Estimez un MPL correspondant à la question (v) par les DMC, où *mothalc* et *fathalc* sont les instruments d'*abuse*. La différence entre les coefficients des MCO et des DMC est-elle grande ?

x. Utilisez le test décrit en section 15.6 pour tester si *abuse* est endogène dans le MPL.

C16. Utilisez les données de CRIME1 pour répondre à cette question.

(i) Pour les estimations des MCO données dans le tableau 17.7, trouvez les écarts-types robustes à l'hétéroscédasticité. Cela crée-t-il des différences notables sur la significativité des coefficients ?

(ii) Retrouvez les écarts-types vraiment robustes – c'est-à-dire ceux qui ne reposent pas sur l'équation (17.35) – pour les estimations de régression de Poisson de la seconde colonne. (Cela nécessite que votre logiciel de statistiques calcule ces écarts-types vraiment robustes) Comparez les intervalles de confiance vraiment robustes au seuil de 95 % pour b_{pouv} à ceux obtenus de manière habituelle dans le tableau 17.7.

(iii) Calculez l'effet marginal moyen de chaque variable du modèle de régression de Poisson. Utilisez la formule pour les variables binaires pour *black*, *hispan*, et *born60*. Comparez les EMM de *gemp86* et *inc86* avec les coefficients des MCO correspondants.

(iv) Si votre logiciel de statistiques donne les écarts-types robustes des EMM de la question (iii), comparez la statistique *t* de l'estimation par les MCO de b_{pcnv} avec la statistique *t* robuste de l'EMM de *pcnv*.

ANNEXE 17A

17A.1 Estimation par maximum de vraisemblance avec des variables explicatives

L'Annexe C passe en revue l'estimation par maximum de vraisemblance dans le cas le plus simple estimant les paramètres d'une distribution non conditionnelle. Or la plupart des modèles d'économétrie ont des variables explicatives, que nous estimions ces modèles par MCO ou EMV. Cette dernière technique est indispensable pour les modèles non linéaires, et nous apportons ici une discussion très brève de cette approche dans le cas général.

Tous les modèles couverts dans ce chapitre peuvent être mis sous la forme suivante. Soit $f(y|\mathbf{x},\beta)$ la fonction de densité du tirage aléatoire dans la population y_i, conditionnellement à $\mathbf{x}_i = \mathbf{x}$. L'estimateur du maximum de vraisemblance (EMV) de β maximise la fonction de log-vraisemblance

$$\max_{\mathbf{b}} \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i, \mathbf{b}), \qquad [17.53]$$

où le vecteur **b** est l'argument du problème de maximisation. Dans la plupart des cas, l'EMV, que nous noterons $\hat{\beta}$, est convergent et a approximativement une distribution normale sur les grands échantillons. C'est vrai même quand nous ne pouvons pas donner de formule pour $\hat{\beta}$, sauf dans des cas très particuliers.

Pour les modèles à réponse binaire (logit et probit), la densité conditionnelle est déterminée par deux valeurs, $f(1|\mathbf{x},\beta) = \mathrm{P}(y_i = 1|\mathbf{x}_i) = G(\mathbf{x}_i\beta)$ and $f(0|\mathbf{x},\beta) = \mathrm{P}(y_i = 0|\mathbf{x}_i) = 1 - G(\mathbf{x}_i\beta)$. Une manière abrégée d'écrire cette densité est $f(y|\mathbf{x},\beta) = [1 - G(\mathbf{x}\beta)]^{(1-y)}[G(\mathbf{x}\beta)]^{y}$ pour $y = 0, 1$. Nous pouvons donc réécrire (17.53) :

$$\max_{b} \sum_{i=1}^{n} \{(1 - y_i)\log[1 - G(\mathbf{x}_i\mathbf{b})] + y_i \log[G(\mathbf{x}_i\mathbf{b})]\}. \qquad [17.54]$$

D'une manière générale, les solutions de (17.54) sont trouvées rapidement par des ordinateurs modernes en utilisant des méthodes itératives pour maximiser une fonction. Le temps de calcul est généralement assez faible même pour de gros échantillons.

Les fonctions de log-vraisemblance du modèle Tobit ainsi que des régressions censurées et tronquées sont seulement un peu plus compliquées, elles dépendent d'un paramètre de variance en plus de β. Elles se tirent facilement des densités trouvées dans le texte. Voir Wooldridge (2010) pour plus de détails.

ANNEXE 17B

17B.1 Écarts-types asymptotiques des modèles à variables dépendantes limitées

Le calcul des écarts-types estimés asymptotiques des modèles et des méthodes introduites dans ce chapitre sont bien au-delà de ce qui est couvert ici. Elles nécessitent de l'algèbre matriciel, mais aussi une théorie asymptotique détaillée pour les estimations non linéaires. Les bases nécessaires à une analyse systématique de ces méthodes et de plusieurs de leurs variantes sont données dans Wooldridge (2010).

Il est cependant instructif de voir les formules donnant les écarts-types estimés asymptotiques au moins pour certaines des méthodes. Pour le modèle de réponse binaire $\mathrm{P}(y = 1|\mathrm{x}) = G(\mathbf{x}\beta)$, où $G(\cdot)$ est la fonction logit ou probit, et β est le vecteur de paramètres $k \times 1$, la matrice de variance asymptotique de $\hat{\beta}$ est donnée par

$$\widehat{\mathrm{Avar}}(\hat{\beta}) \equiv \left(\sum_{i=1}^{n} \frac{[g(\mathbf{x}_i\hat{\beta})]^2 \mathbf{x}_i'\mathbf{x}_i}{G(\mathbf{x}_i\hat{\beta})[1 - G(\mathbf{x}_i\hat{\beta})]} \right)^{-1}, \qquad [17.55]$$

qui est une matrice $k \times k$ (Voir l'annexe D pour quelques notions d'algèbre matriciel). A part les termes incluant $g(\cdot)$ et $G(\cdot)$, cette formule ressemble beaucoup à la matrice de variance-covariance de l'estimateur des MCO, sans le terme $\hat{\sigma}^2$. L'expression (17.55) tient compte de la nature non linéaire de la probabilité – c'est-à-dire la nature non linéaire de $G(\cdot)$ – ainsi que de la forme particulière d'hétéroscédasticité des modèles à réponse binaire : $Var(y|\mathrm{x}) = G(\mathrm{x}\beta)[1 - G(\mathrm{x}\beta)]$.

La racine carrée des éléments diagonaux de (17.55) sont les écarts-types estimés asymptotiques des $\hat{\beta}_j$ et ils sont rapportés par défaut par les logiciels d'économétrie qui analysent les modèles logit et probit. Une fois que nous les avons, les statistiques t (asymptotiques) et les intervalles de confiance sont trouvés de manière habituelle.

La matrice (17.55) est également la base du test de Wald pour les restrictions multiples sur $\boldsymbol{\beta}$ [Voir Wooldridge (2010, chapitre 15)].

La matrice de variance asymptotique du modèle Tobit est plus compliquée mais elle a une structure similaire. Remarquez que nous pouvons également obtenir un écart-type estimé de $\hat{\sigma}$. La variance asymptotique de la régression de Poisson, en autorisant $\sigma^2 \neq 1$ comme dans (17.35), ressemble beaucoup plus à (17.55) :

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \left(\sum_{i=1}^{n} \exp(\mathbf{x}_i \hat{\boldsymbol{\beta}}) \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \qquad \textbf{[17.56]}$$

Les racines carrées des éléments diagonaux de cette matrice sont les écarts-types estimés asymptotiques. Si l'hypothèse de Poisson tient, on peut supprimer $\hat{\sigma}^2$ de la formule (puisque $\sigma^2 = 1$).

La formule de la matrice de variance-covariance totalement robuste est donnée dans Wooldridge (2010, chapitre 18) :

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{n} \exp(\mathbf{x}_i \hat{\boldsymbol{\beta}}) \mathbf{x}_i' \mathbf{x}_i \right]^{-1} \left(\sum_{i=1}^{n} \hat{u}_i^2 \mathbf{x}_i' \mathbf{x}_i \right) \left[\sum_{i=1}^{n} \exp(\mathbf{x}_i \hat{\boldsymbol{\beta}}) \mathbf{x}_i' \mathbf{x}_i \right]^{-1},$$

où $\hat{u}_i = y_i - \exp(\mathbf{x}_i \hat{\boldsymbol{\beta}})$ sont les résidus de la régression de Poisson. Cette expression a une structure similaire à la matrice habituelle de variance-covariance robuste à l'hétéroscédasticité pour les MCO, et elle est calculée automatiquement par de nombreux logiciels statistiques pour trouver les écarts-type vraiment robustes.

Les écarts-types estimés asymptotiques des régressions censurées, tronquées, et de la correction pour la sélection de l'échantillon Heckit sont plus compliqués, bien qu'il y ait des points communs avec les formules ci-dessus [Voir Wooldridge (2010) pour plus de détails].

CHAPITRE 18

MATIÈRES AVANCÉES DANS L'ANALYSE DES SÉRIES TEMPORELLES

Traduction de Michel Beine

Dans ce chapitre nous couvrons quelques matières avancées supplémentaires en économétrie des séries temporelles. Dans les chapitres 10, 11 et 12 nous avons insisté à plusieurs occasions sur le fait qu'utiliser des données de séries temporelles dans une analyse de régression requiert beaucoup de soin dû à la présence d'une tendance et à la nature persistante de beaucoup de séries temporelles économiques. En plus d'étudier des matières telles que les modèles à retards distribués infinis et la prévision, nous discutons des avancées récentes dans l'analyse des processus de séries temporelles avec racines unitaires.

Dans la section 18. 1, nous décrivons les modèles à retards distribués infinis qui permettent qu'une variation dans la variable explicative affecte toutes les valeurs futures de la variable dépendante. Conceptuellement, ces modèles sont des extensions directes des modèles à retards distribués finis vus dans le chapitre 10, mais l'estimation de ces modèles pose certains défis intéressants.

Dans la section 18. 2, nous montrons comment tester formellement la présence d'une racine unitaire dans les séries temporelles. Rappelons-nous que dans le chapitre 11, nous avons exclu les processus à racine unitaire de manière à appliquer la théorie asymptotique habituelle. Dans la mesure où la présence d'une racine unitaire implique qu'un choc a un impact de long terme, savoir si un processus présente une racine unitaire est intéressant en soi-même.

Nous couvrons la notion de régression fallacieuse entre deux processus de séries temporelles, l'un et l'autre présentant une racine unitaire. Le résultat principal est que même si deux séries à racine unitaire sont indépendantes, il est très probable que la régression de l'une sur l'autre générera une statistique t statistiquement significative. Cela illustre les conséquences potentiellement sérieuses d'utiliser l'inférence standard lorsque les variables dépendantes et indépendantes sont des processus intégrés.

La notion de cointégration s'applique lorsque deux séries sont I(1) mais qu'une combinaison linéaire des deux est I(0). Dans ce cas la régression de l'une sur l'autre n'est pas fallacieuse mais nous indique par ailleurs quelque chose concernant la relation de long terme entre elles. La cointégration entre deux séries implique aussi l'existence d'un modèle particulier appelé modèle à correction d'erreur utilisé pour étudier la dynamique de court terme. Nous couvrirons ces modèles dans la section 18.4.

Dans la section 18.5 nous apporterons un aperçu de la prévision et nous couvrirons tous les outils vus dans ce chapitre et les chapitres précédents pour montrer comment les méthodes de régression peuvent être utilisées pour prédire les résultats futurs des séries temporelles. La littérature sur la prévision est vaste, si bien que nous insisterons seulement sur les méthodes les plus classiques en matière de régression. Nous évoquerons également un domaine conjoint, à savoir la causalité à la Granger.

18.1 MODÈLES À RETARDS DISTRIBUÉS INFINIS

Soit $\{(y_t, z_t) : t = \ldots, -2, -1, 0, 1, 2, \ldots\}$ un processus bivarié de séries temporelles (qui est seulement observé partiellement). Un **modèle à retards distribués infinis (MRDI)** reliant y_t à la valeur contemporaine et à toutes les valeurs passées de z est donné par :

$$y_t = \alpha + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \ldots + u_t, \qquad \textbf{[18.1]}$$

dans lequel la somme des z retardés s'étend de manière infinie dans le passé. Ce modèle est seulement une approximation de la réalité puisqu'aucun processus économique n'est initié de manière infinie dans le passé. En comparaison avec un modèle à retard distribué fini, un MRDI ne requiert pas que nous tronquions le retard à une valeur particulière.

Pour que le modèle (18.1) ait un sens, les coefficients de retard δ_j doivent tendre vers zéro lorsque j tend vers l'infini. Cela ne veut pas dire que δ_2 est plus petit en valeur que δ_1. Cela veut dire seulement que l'effet de z_{t-j} sur y_t doit éventuellement devenir petit lorsque j augmente. Dans la plupart des applications,

cela a également un sens économique : le passé distant de z devrait être moins important pour expliquer y que le passé récent de z.

Même si nous décidons que (18.1) est un modèle utile, nous ne pouvons clairement pas l'estimer sans restriction. Tout d'abord nous observons seulement une histoire finie des données. L'équation (18.1) implique un nombre infini de paramètres δ_0, δ_1, δ_2, qui ne peuvent pas être estimés sans restriction. Plus tard, nous imposerons des restrictions sur les δ_j qui nous permettront d'estimer (18.1).

Comme pour les modèles à retards distribués finis (RDF), l'impact de court terme dans l'équation (18.1) est donné par δ_0 (voir chapitre 10). Généralement, les δ ont la même interprétation que dans un modèle RDF. Supposons que $z_s = 0$ pour tout $s < 0$ et que $z_0 = 1$ et $z_s = 0$ pour tout $s > 1$; en d'autres termes, au temps $t = 0$, z augmente temporairement d'une unité et puis revient vers sa valeur initiale de zéro. Pour tout $h \geq 0$, nous avons $y_h = \alpha + \delta_h + u_h$ pour $h \leq 0$ et donc

$$E(y_h) = \alpha + \delta_h, \qquad \textbf{[18.2]}$$

dans lequel nous utilisons l'hypothèse habituelle que u_h a une moyenne égale à zéro. Il s'ensuit que δ_h est la variation de $E(y_h)$ étant donné un changement temporaire d'une unité de z au temps zéro. Pour que le modèle RDI fasse sens, δ_h doit tendre vers zéro lorsque h augmente. Cela signifie qu'un changement temporaire de z n'a aucun effet de long terme sur le y attendu : $E(y_h) = \alpha + \delta_h \rightarrow \alpha$ *quand* $h \rightarrow \infty$.

Nous avons supposé que le processus z débute à $z_s = 0$ et que l'augmentation d'une unité intervient en $t = 0$. Ces hypothèses sont là uniquement dans un but d'illustration. Plus généralement, si z augmente temporairement d'une unité (à partir de n'importe quel niveau initial) au temps t, alors δ_h mesure la variation de la valeur attendue de y après h périodes. La distribution des retards, qui est la projection de δ_h en fonction de h, montre la dynamique attendue que les y futurs suivent étant donné une augmentation temporaire d'une unité de z.

L'impact de long terme (ILT) dans le modèle 18.1 est la somme de tous les coefficients de retard :

$$ILT = \delta_0 + \delta_1 + \delta_2 + \delta_3 + \ldots \qquad \textbf{[18.3]}$$

Dans lequel nous supposons que la somme infinie est bien définie. Puisque les δ_j doivent converger vers zéro, l'ILT peut souvent être bien approximé par une somme finie de la forme $\delta_0 + \delta_1 + \ldots + \delta_p$ pour p suffisamment élevé. Pour interpréter l'ILT, supposons que le processus z est stationnaire en $z_s = 0$ pour $s = 0$. En $t = 0$, le processus augmente de manière permanente d'une unité. Par exemple, si z_t est le changement en pourcentage de l'offre de monnaie et si y_t est le taux d'inflation, alors nous sommes intéressés par les effets d'une augmentation permanente d'un point de pourcentage dans la croissance de l'offre de monnaie. Ensuite en substituant $z_s = 0$ pour $s < 0$ et $z_t = 1$ pour $t \geq 0$, nous avons

$$y_h = \alpha + \delta_0 + \delta_1 + \ldots + \delta_h + u_h,$$

où $h \geq 0$ est n'importe quel horizon. Parce que u_t a une moyenne égale à zéro pour tout t, nous avons :

$$E(y_h) = \alpha + \delta_0 + \delta_1 + \delta_2 + \ldots + \delta_h. \qquad \textbf{[18.4]}$$

[il est utile de comparer (18.4) et (18.2)]. Lorsque l'horizon augmente, c'est-à-dire lorsque $h \rightarrow \infty$, le terme de droite de (18.4) est par définition l'effet de long terme plus α. Donc l'ILT mesure la variation de long terme dans la valeur attendue de y étant donné une augmentation *permanente* d'une unité de z.

Pour aller plus loin 18.1

Supposons que $z_s = 0$ pour $s < 0$ et que $z_0 = 1$, $z_1 = 1$, et $z_s = 0$ pour $s > 1$. Trouvez $E(y_{t-1})$, $E(y_0)$, et $E(y_h)$ pour $h \geq 1$. Qu'arrive-t-il si $h \rightarrow \infty$?

La dérivation précédente de l'IPT et l'interprétation de δ_j ont utilisé le fait que les erreurs ont une moyenne égale à zéro ; comme d'habitude, ce n'est pas vraiment une hypothèse, moyennant le fait qu'une constante est incluse dans le modèle. Un examen plus approfondi de notre raisonnement montre que nous avons supposé que la variation de z à n'importe quelle période de temps n'a pas d'effet sur la valeur attendue de u_t.

C'est la version du retard distribué infini de l'hypothèse d'exogénéité stricte que nous avons introduite dans le chapitre 10 (en particulier l'hypothèse TS3).

Formellement,

$$E(u_t | \ldots, z_{t-2}, z_{t-1}, z_t, z_{t+1}, \ldots) = 0, \qquad \textbf{[18.5]}$$

si bien que la valeur attendue de u_t ne dépend pas de z à n'importe quelle période de temps. Bien que (18.5) soit naturelle dans certaines applications, cela exclut d'autres possibilités importantes. En effet (18.5) ne permet pas un effet en retour de y_t vers les z futurs parce que z_{t+h} doit être non corrélé avec u_t pour $h > 0$. Dans l'exemple inflation-croissance de l'offre de monnaie, où y_t est l'inflation et z_t la croissance de l'offre de monnaie, (18.5) exclut que les variations futures de la croissance de l'offre de monnaie soient liées aux variations du taux d'inflation aujourd'hui. Étant donné que la politique monétaire s'évertue à garder le taux d'inflation et les taux d'intérêt à certains niveaux, cela peut être irréaliste.

Une approche pour estimer les δ_j que nous couvrirons dans la sous-section suivante impose l'hypothèse d'exogénéité stricte de manière à produire des estimateurs consistants de δ_j. Une hypothèse plus faible est :

$$E(u_t | z_t, z_{t-1}, \ldots) = 0. \qquad \textbf{[18.6]}$$

Sous (18.6), l'erreur est non corrélée avec les z contemporains et passés mais elle peut être corrélée avec les z futurs ; ceci permet à z_t d'être une variable qui suit des règles de politique qui dépendent des y passés. Parfois (18.6) est suffisant pour estimer les δ_j ; nous expliquerons cela dans la sous-section suivante.

Une chose qu'il faut garder à l'esprit est que ni (18.5), ni (18.6) ne disent rien à propos de la corrélation sérielle de $\{u_t\}$ (tout comme dans les modèles à retards distribués finis). Au minimum, nous pouvons nous attendre à ce que $\{u_t\}$ soit corrélé sériellement parce que (18.1) n'est généralement pas complet dynamiquement dans le sens exposé dans la section 11.4. Nous étudierons le problème de corrélation sérielle plus tard.

Comment interpréter les coefficients de retard de l'ILT si (18.6) est valide mais que (18.5) ne l'est pas ? La réponse est la suivante : de la même manière que précédemment. Nous pouvons toujours procéder au raisonnement fictif précédent (ou contrefactuel) même si les données que nous observons sont générées par un certain feedback entre y_t et les z futurs. Par exemple nous pouvons nous demander ce qu'est l'effet de long terme d'une augmentation permanente de la croissance de l'offre de monnaie sur l'inflation même si les données relatives à la croissance de l'offre de monnaie ne peuvent pas être caractérisées comme strictement exogènes.

Les retards distribués géométriquement (ou à la Koyck)

Parce qu'il y a généralement un nombre infini de δ_j, nous ne pouvons les estimer de manière consistante sans restriction. La version la plus simple de (18.1) qui fait que le modèle dépend toujours d'un nombre infini de retards, est le modèle à retards distribués de manière géométrique (RDG) (à la Koyck). Dans ce modèle les δ_j dépendent seulement de deux paramètres :

$$\delta_j = \gamma\rho^j, \ |\rho| < 1, \quad j = 0, 1, 2, \ldots. \qquad \textbf{[18.7]}$$

Les paramètres γ et ρ peuvent être positifs ou négatifs mais ρ doit être plus petit que un en valeur absolue. Cela garantit que $\delta_j \to 0$ lorsque $j \to \infty$. En fait, cette convergence intervient à un taux très rapide (par exemple avec $\rho = 0{,}5$ et $j = 10$, $\rho^j = 1/1024 < 0{,}001$).

L'impact de court terme (ICT) dans le RDG est simplement $\delta_0 = \gamma$ si bien que le signe de l'ICT est déterminé par le signe de γ. Si $\gamma > 0$ et si $\rho > 0$, alors tous les coefficients de retard sont positifs. Si $\rho < 0$, les coefficients de retard alternent en signe (ρ^j est négatif pour j impair). L'impact de long terme est plus difficile à obtenir mais nous pouvons utiliser un résultat standard concernant la somme d'une série géométrique : pour $|\rho| < 1$, $1 + \rho + \rho^2 + \ldots + \rho^j + \ldots = 1/(1 - \rho)$, si bien que

$$ILT = \gamma/(1 - \rho).$$

L'ILT a le même signe que γ.

Si nous insérons (18.7) dans (18.1), nous avons toujours un modèle qui dépend des z s'échelonnant dans un passé indéfini. Cependant une simple soustraction nous donne un modèle estimable.

Écrivons le RDI au temps t et $t-1$ comme :

$$y_t = \alpha + \gamma z_t + \gamma\rho z_{t-1} + \gamma\rho^2 z_{t-2} + \ldots + u_t \qquad \textbf{[18.8]}$$

et

$$y_{t-1} = \alpha + \gamma z_{t-1} + \gamma\rho z_{t-2} + \gamma\rho^2 z_{t-3} + \ldots + u_{t-1}. \qquad \textbf{[18.9]}$$

Si nous multiplions la seconde équation par ρ et si nous la soustrayons à partir de la première, presque tous les termes disparaissent :

$$y_t - \rho y_{t-1} = (1 - \rho)\alpha + \gamma z_t + u_t - \rho u_{t-1},$$

que nous pouvons écrire comme :

$$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + u_t - \rho u_{t-1}, \qquad \textbf{[18.10]}$$

où $\alpha_0 = (1 - \rho)\alpha$. Cette équation apparaît comme un modèle standard avec une variable dépendante retardée et dans laquelle z_t intervient de manière contemporaine. Parce que γ est le coefficient de z_t et ρ est le coefficient de y_{t-1}, il apparaît que nous pouvons estimer ces paramètres (si pour une raison quelconque nous sommes intéressés par α, nous pouvons toujours obtenir $\hat{\alpha} = \hat{\alpha}_0/(1 - \hat{\rho})$ après avoir estimé ρ et α_0.).

La simplicité de (18.10) est quelque peu trompeuse. Le terme d'erreur dans cette équation, $u_t - \rho u_{t-1}$ est généralement corrélé avec y_{t-1}. À partir de (18.9), il est très clair que u_{t-1} et y_{t-1} sont corrélés. Par conséquent si nous écrivons (18.10) comme

$$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + v_t, \qquad \textbf{[18.11]}$$

où $v_t = u_t - \rho u_{t-1}$, alors nous avons généralement une corrélation entre v_t et y_{t-1}. Sans hypothèse supplémentaire, l'estimation MCO de (18.11) produits des estimateurs inconsistants de γ et ρ.

Un cas où v_t *doit* être corrélé avec y_{t-1} intervient lorsque u_t est indépendant de z_t et de *toutes* les valeurs passées de z et y. Alors, (18.8) est dynamiquement complet si bien que u_t est non corrélée avec y_{t-1}. À partir de (18.9), la covariance entre v_t et y_{t-1} est égale à $-\rho\text{Var}(u_{t-1}) = -\rho\sigma_u^2$, qui est égale à zéro seulement si $\rho = 0$. Nous pouvons voir facilement que v_t est corrélée sériellement : parce que $\{u_t\}$ est non corrélée sériellement, $\text{E}(v_t v_{t-1}) = \text{E}(u_t u_{t-1}) - \rho\text{E}(u_{t-1}^2) - \rho\text{E}(u_t u_{t-2}) + \rho^2\text{E}(u_{t-1}u_{t-2}) = -\rho\sigma_u^2$. Pour $j > 1$, $\text{E}(v_t v_{t-j}) = 0$. Donc $\{v_t\}$ est un processus moyenne mobile d'ordre un (voir section 11.1). Ceci et l'équation (18.11) donnent l'exemple d'un modèle, dérivé du modèle original, avec une variable dépendante retardée et une forme particulière de corrélation sérielle.

Si nous posons l'hypothèse stricte d'exogénéité (18.5), alors z_t est non corrélé avec u_t, et par conséquent avec v_t. Donc si nous trouvons une variable instrumentale adéquate pour y_{t-1}, alors nous pouvons estimer (18.11) par VI. Quel est un bon instrument potentiel pour y_{t-1} ? Par hypothèse u_t et u_{t-1}, sont tous les deux non corrélés avec z_{t-1} si bien que v_t est non corrélé avec z_{t-1}. Si $\gamma \neq 0$, z_{t-1} et y_{t-1} sont corrélés, même

après élimination de z_t. Par conséquent, nous pouvons utiliser les instruments (z_t, z_{t-1}) pour estimer (18.11). Généralement les écarts-types doivent être ajustés pour la corrélation sérielle dans le processus $\{v_t\}$, comme nous en avons discuté dans la section 15.7.

Une alternative à l'estimation par VI utilise le fait que $\{u_t\}$ peut contenir une forme particulière de corrélation sérielle. En particulier, en plus de (18.6), supposons que $\{u_t\}$ suive le processus AR(1) :

$$u_t = \rho u_{t-1} + e_t \qquad \textbf{[18.12]}$$

$$\mathrm{E}(e_t|z_t, y_{t-1}, z_{t-1}, \ldots) = 0. \qquad \textbf{[18.13]}$$

Il est important de remarquer que le ρ qui apparaît dans (18.12) est le même paramètre qui multiplie y_{t-1} dans (18.11). Si (18.12) et (18.13) sont valables nous pouvons écrire l'équation (18.10) comme

$$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + e_t, \qquad \textbf{[18.14]}$$

qui est un modèle complet dynamiquement sous (18.13). Depuis le chapitre 11, nous pouvons obtenir des estimateurs des paramètres par MCO qui sont convergents et asymptotiquement normaux. C'est très utile car il n'y a pas besoin de s'occuper de la corrélation sérielle dans les erreurs. Si e_t satisfait l'hypothèse d'homoscédasticité, $\mathrm{Var}(e_t|z_t, y_{t-1}) = \sigma_e^2$, l'inférence habituelle s'applique. Une fois que nous avons estimé γ et ρ, nous pouvons estimer facilement l'ILT : $\widehat{ILT} = \hat{\gamma}/(1-\hat{\rho})$.

La simplicité de cette procédure repose sur l'hypothèse éventuellement forte que $\{u_t\}$ suit un processus AR(1) avec le même ρ apparaissant dans (18.7). Ce n'est pas une procédure plus critiquable que de faire l'hypothèse que les $\{u_t\}$ sont non corrélés sériellement. Néanmoins parce que la convergence des estimateurs repose fortement sur cette hypothèse, c'est une bonne idée de la tester. Un simple test commence par la spécification de $\{u_t\}$ comme un processus AR(1) avec un paramètre *différent*, disons $u_t = \lambda u_{t-1} + e_t$. McClain et Wooldridge (1995) ont proposé un test du multiplicateur de Lagrange de $\mathrm{H}_0 : \lambda = \rho$ qui peut-être calculé après une estimation MCO de (18.14).

Le modèle à retards distribués géométriquement se généralise avec des variables explicatives multiples si bien que nous avons un retard distribué infini pour chaque variable explicative – mais alors nous devons être capables d'écrire le coefficient de $z_{t-j,h}$ comme $\lambda_h \rho^j$. Entre d'autres termes, bien que λ_h soit différent pour chaque variable explicative, ρ est le même. Dès lors nous pouvons écrire

$$y_t = \alpha_0 + \gamma_1 z_{t1} + \ldots + \gamma_k z_{tk} + \rho y_{t-1} + v_t. \qquad \textbf{[18.15]}$$

Les mêmes problèmes qui se sont posés dans le cas d'un z se posent dans le cas de plusieurs z. Comme extension naturelle de (18.12) et (18.13), il faut remplacer z_t juste par $\mathbf{z}_t = (z_{t1}, \ldots, z_{tk})$. L'estimateur des MCO est convergent et asymptotiquement normal. Alternativement, une méthode par VI peut-être utilisée.

Modèles à retards distribués rationnels

Le retard distribué géométriquement implique une distribution des retards relativement restrictive. Quand $\gamma > 0$ et $\rho > 0$, les δ sont positifs et déclinent de manière monotone vers zéro. Il est possible d'avoir des modèles à retards distribués infinis plus généraux. Le RDG est un cas spécial de ce qui est appelé communément un modèle à retards distribués rationnels (RDR). Une couverture générale est au-delà de la portée de notre analyse – Harvey (1990) est une bonne référence – mais nous pouvons proposer une extension simple et utile.

Un tel modèle RDR est décrit de manière aisée en ajoutant un retard de z à l'équation (18.11) :

$$y_t = \alpha_0 + \gamma_0 z_t + \rho y_{t-1} + \gamma_1 z_{t-1} + v_t, \qquad \textbf{[18.16]}$$

où $v_t = u_t - \rho u_{t-1}$ comme précédemment. Par substitution répétée, on peut montrer que (18.16) est équivalent au modèle à retards distribués infinis

$$\begin{aligned} y_t &= \alpha + \gamma_0(z_t + \rho z_{t-1} + \rho^2 z_{t-2} + \ldots) \\ &\quad + \gamma_1(z_{t-1} + \rho z_{t-2} + \rho^2 z_{t-3} + \ldots) + u_t \\ &= \alpha + \gamma_0 z_t + (\rho\gamma_0 + \gamma_1)z_{t-1} + \rho(\rho\gamma_0 + \gamma_1)z_{t-2} \\ &\quad + \rho^2(\rho\gamma_0 + \gamma_1)z_{t-3} + \ldots + u_t, \end{aligned}$$

dans lequel nous avons besoin à nouveau de l'hypothèse $|\rho| < 1$. À partir de cette dernière équation on peut caractériser la distribution des retards. En particulier l'impact de court terme est γ_0, alors que le coefficient de z_{t-h} est $\rho^{h-1}(\rho\gamma_0 + \gamma_1)$ pour $h \geq 1$. Par conséquent ce modèle permet à l'impact de court terme d'avoir un signe différent des autres coefficients de retards même si $\rho > 0$. Néanmoins, si $\rho > 0$, les δ_h ont le même signe que $(\rho\gamma_0 + \gamma_1)$ pour tout $h \geq 1$. La distribution des retards est donnée dans la figure 18.1 pour $\rho = 0{,}5$, $\gamma_0 = -1$, et $\gamma_1 = 1$.

La manière la plus simple de calculer l'impact de long terme est de mettre y et z à leur valeur de long terme pour tout t, disons y^* et z^*, et ensuite de trouver alors la variation de y^* par rapport à z^* (voir aussi le problème 3 dans le chapitre 10). Nous avons $y^* = \alpha_0 + \gamma_0 z^* + \rho y^* + \gamma_1 z^*$ et en résolvant on obtient $y^* = \alpha_0/(1 - \rho) + (\gamma_0 + \gamma_1)/(1 - \rho)z^*$. Maintenant nous utilisons le fait que ILT = $\Delta y^*/\Delta z^*$, ce qui donne

$$ILT = (\gamma_0 + \gamma_1)/(1 - \rho).$$

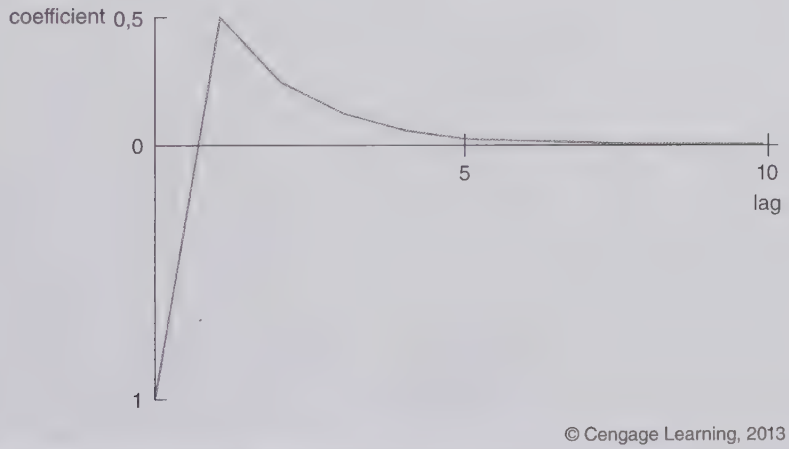


Figure 18.1 *Distribution des retards distribués rationnels (18.16) avec $\rho = 0{,}5$, $\gamma_0 = 21$, et $\gamma_1 = 1$.*

Parce que $|\rho| < 1$, l'ILT a le même signe que $\gamma_0 + \gamma_1$ et l'ILT est égal à zéro si et seulement si $\gamma_0 + \gamma_1 = 0$, comme dans la figure 18.1.

EXEMPLE 18.1
Investissement immobilier et inflation des prix résidentiels

Nous estimons à la fois le modèle à retards distribués géométriques de base et à retards rationnels en appliquant les MCO aux équations (18.14) et (18.16) respectivement. La variable indépendante est *log(invpc)*, après qu'une tendance linéaire ait été enlevée (cela veut dire qu'on a ôté la tendance incluse dans *log(invpc)* de manière linéaire). Pour z_t, nous utilisons la croissance de l'indice des prix. Cela nous permet d'estimer dans quelle mesure l'inflation des prix résidentiels affecte les mouvements dans l'investissement immobilier autour de sa tendance. Les résultats de l'estimation, en utilisant les données dans le fichier HSEINC, sont donnés dans le tableau 18.1.

Le modèle à retards distribués géométriques est clairement rejeté par les données, puisque $gprice_{-1}$ est très significatif. Les R^2 ajustés montrent également que le modèle RDR colle bien mieux aux données. Les deux modèles donnent des valeurs estimées d'ILT très différentes. Si nous utilisons de manière erronée le RDG, l'impact de long terme estimé est proche de cinq : une augmentation d'un point de pourcentage dans l'inflation des prix résidentiels augmente l'investissement immobilier de long terme de 4,7 % (au-delà de sa valeur tendancielle). Economiquement, cela semble improbable. L'ILT estimé à partir du modèle à retards distribués rationnels est en dessous de un. En fait, nous ne pouvons pas rejeter l'hypothèse nulle $\gamma_0 + \gamma_1 = 0$ à n'importe quel degré de significativité raisonnable (*p-valeur* = 0,83), si bien qu'il n'y a aucune preuve que l'ILT soit différent de zéro. Ceci est un bon exemple du fait qu'une mauvaise spécification de la dynamique d'un modèle par omission des retards appropriés peut amener à tirer des conclusions erronées.

Tableau 18.1 *Modèle à retards distribués pour l'investissement immobilier*

Variable dépendante : log(*invpc*), ajustée pour la tendance		
Variables Indépendantes	**RD géométriques**	**RD rationnels**
gprice	3,095 (0,933)	3,256 (0,970)
y_{-1}	0,340 (0,132)	0,547 (0,152)
$gprice_{-1}$	–	−2,936 (0,973)
constante	−0,010 (0,018)	0,006 (0,017)
Impact de long terme	4,689	0,706
Taille de l'échantillon R^2 ajusté	41 0,375	40 0,504

18.2 TESTER LA PRÉSENCE DE RACINES UNITAIRES

Nous abordons maintenant le problème important visant à tester si une série temporelle suit un processus de racine unitaire. Dans le chapitre 11, nous avons donné des directives vagues et nécessairement peu formelles pour décider si une série est I(1) ou non. Dans de nombreux cas, il est utile d'avoir un test formel de la présence d'une racine unitaire. Comme nous le verrons, ces tests doivent être appliqués avec prudence.

L'approche la plus simple pour tester la présence d'une racine unitaire se base sur un modèle AR(1)

$$y_t = \alpha + \rho y_{t-1} + e_t, \; t = 1, 2, \ldots, \qquad \textbf{[18.17]}$$

dans lequel y_0 est la valeur initiale observée. Dans cette section, nous notons $\{e_t\}$ comme un processus qui a une moyenne zéro étant donné les y observés passés :

$$\mathrm{E}(e_t | y_{t-1}, y_{t-2}, \ldots, y_0) = 0. \qquad \textbf{[18.18]}$$

[Sous (18.18), $\{e_t\}$ est appelé une **séquence de différences de martingale** par rapport à $\{y_{t-1}, y_{t-2}, \ldots\}$. Si $\{e_t\}$ est considérée comme i.i.d de moyenne zéro et indépendante de y_0, alors, elle satisfait aussi (18.18).]

Si $\{y_t\}$ suit (18.17), elle possède une racine unitaire, si et seulement si $\rho = 1$. Si $\alpha = 0$ et $\rho = 1$, $\{y_t\}$ suit une marche aléatoire sans dérive [avec des innovations e_t satisfaisant (18.18)]. Si $\alpha \neq 0$ et $\rho = 1$, $\{y_t\}$ suit une marche aléatoire avec dérive, ce qui signifie que $E(y_t)$ est une fonction linéaire de t. Un processus de racine unitaire avec dérive se comporte différemment d'un processus sans dérive. Néanmoins, nous suivons la pratique courante qui est de ne pas spécifier α sous l'hypothèse nulle. Par conséquent, l'hypothèse nulle est que $\{y_t\}$ possède une racine unitaire :

$$H_0 : \rho = 1. \qquad \textbf{[18.19]}$$

Dans la plupart des cas, nous considérons l'hypothèse alternative unilatérale

$$H_1 : \rho < 1. \qquad \textbf{[18.20]}$$

(En pratique, cela implique $0 < \rho < 1$, dans la mesure où $\rho < 0$ serait très rare pour une série suspectée de posséder une racine unitaire.) L'alternative $H_1 : \rho > 1$ n'est habituellement pas considérée car alors y_t est explosif. En fait, si $\alpha > 0$, y_t possède une tendance exponentielle dans sa moyenne si $\rho > 1$.

Quand $|\rho| < 1$, $\{y_t\}$ est un processus AR(1) stable, ce qui signifie qu'il est faiblement dépendant ou asymptotiquement non corrélé. Rappelons-nous à partir du chapitre 11 que $\text{Corr}(y_t, y_{t+h}) = \rho^h \to 0$ lorsque $|\rho| < 1$. Par conséquent, tester (18.19) dans le modèle (18.17), avec l'alternative donnée par (18.20) revient vraiment à tester si $\{y_t\}$ est I(1) contre l'alternative selon laquelle $\{y_t\}$ est I(0). Nous ne considérons pas l'hypothèse nulle comme I(0) dans cette structure parce que $\{y_t\}$ est I(0) pour n'importe quelle valeur de ρ strictement comprise entre -1 et 1, un élément que la procédure classique de test d'hypothèses n'appréhende pas facilement. Il y a des tests pour lesquels l'hypothèse nulle est I(0) contre l'alternative I(1) mais ces tests adoptent une approche différente [voir par exemple Kwiatkowski, Phillips, Schmidt, and Shin (1992)].

Une équation utile pour mener à bien un test de racine unitaire s'obtient en soustrayant y_{t-1} des deux côtés de (18.17) et en redéfinissant $\theta = \rho - 1$:

$$\Delta y_t = \alpha + \theta y_{t-1} + e_t. \qquad \textbf{[18.21]}$$

Sous (18.18), c'est un modèle dynamiquement complet et donc il semble naturel de tester $H_0 : \theta = 0$ contre $H_1 : \theta < 0$.

Le problème est que sous H_0, y_{t-1} est I(1) et donc la distribution asymptotique standard normale de la statistique t qui repose sur le théorème central limite ne s'applique pas : la statistique t n'a pas une distribution approximative standard normale même en grand échantillon. La distribution asymptotique de la statistique t sous H_0 est connue comme la **distribution de Dickey-Fuller**, faisant suite aux travaux de Dickey et Fuller (1979).

Bien que nous ne puissions pas utiliser les valeurs critiques habituelles, nous pouvons utiliser les statistiques t pour $\hat{\theta}$ dans (18.21), du moins une fois que les valeurs critiques appropriées ont été tabulées. Le test qui en résulte est connu sous le nom du **test de Dickey-Fuller (DF)** pour une racine unitaire. La théorie utilisée pour obtenir les valeurs critiques asymptotiques est plutôt complexe et est couverte dans des textes avancés d'économétrie des séries temporelles (voir par exemple Banerjee, Dolado, Galbraith, and Hendry (1993) ou BDGH en bref). Par contre, utiliser ce résultat est plutôt aisé. Les valeurs critiques pour la statistique t ont été tabulées par plusieurs auteurs, en commençant par le travail original de Dickey et Fuller (1979). Le tableau 18.2 contient les valeurs critiques en grand échantillon pour différents niveaux de significativité ; elles sont reprises à partir de BDGH (1993, Tableau 4.2.) (Les valeurs critiques ajustées pour de petits échantillons sont disponibles chez BDGH).

Nous rejetons l'hypothèse nulle $H_0 : \theta = 0$ contre $H_1 : \theta < 0$ si $t_{\hat{\theta}} < c$ où c est une valeur négative dans le tableau 18.2. Par exemple pour mener le test à un niveau de significativité de 5 %, nous rejetons si $t_{\hat{\theta}} < -2{,}86$. Ceci implique une statistique t d'une ampleur beaucoup plus élevée que si nous utilisions

une valeur critique standard normale qui serait égale à – 1,65. Si nous utilisions une valeur critique standard normale pour tester une racine unitaire, nous rejetterions beaucoup plus souvent que 5 % du temps lorsque H_0 est vraie.

Tableau 18.2 *Valeurs critiques asymptotiques de t pour test de racine unitaire sans tendance temporelle.*

Niveau de Significativité	1 %	2,5 %	5 %	10 %
Valeur critique	–3,43	–3,12	–2,86	–2,57

EXEMPLE 18.2
Test de racine unitaire pour les taux des obligations à trois mois.

Nous utilisons les données trimestrielles du fichier INTQRT pour tester la présence d'une racine unitaire dans les taux des obligations à trois mois. Quand nous estimons (18.20), nous obtenons :

$$\widehat{\Delta r3_t} = 0{,}625 - 0{,}091\ r3_{t-1} \quad [18.22]$$
$$(0{,}261)\ (0{,}037)$$
$$n = 123,\ R^2 = 0{,}048,$$

dans laquelle nous gardons la convention de reporter les écarts-types entre parenthèses en dessous des valeurs estimées. Nous devons garder à l'esprit que les écarts-types ne peuvent pas être utilisés pour construire les intervalles de confiance habituels ou pour mener les tests t traditionnels parce qu'ils ne se comportent pas de la manière habituelle lorsqu'il y a une racine unitaire. Le coefficient de $r3_{t-1}$ montre que la valeur estimée de ρ est $\hat{\rho} = 1 + \hat{\theta} = 0{,}909$. Même si celle-ci est plus petite que l'unité, nous ne savons pas si elle est statistiquement plus petite que 1. La statistique t de $r3_{t-1}$ est $-0{,}091/0{,}037 = -2{,}46$. À partir du tableau 18.2, la valeur critique à 10 % est –2,57 et, par conséquent, nous ne rejetons pas $H_0 : \rho = 1$ contre $H_1 : \rho < 1$ contre au niveau de significativité de 10 %.

Pour les autres tests d'hypothèse, lorsque nous ne rejetons pas H_0, nous ne disons pas que nous acceptons H_0. Pourquoi ? Supposons que nous testions $H_0 : \rho = 0{,}9$ dans l'exemple précédent en utilisant un test t standard – ce qui est valable asymptotiquement parce que y_t est I(0) sous H_0 –. Dès lors, nous obtenons $t = 0{,}001/0{,}037$, ce qui est très petit et ne donne aucune preuve à l'encontre de $\rho = 0{,}9$. Pourtant cela n'a pas de sens d'accepter $\rho = 1$ et $\rho = 0{,}9$.

Quand nous ne rejetons pas la racine unitaire, comme dans l'exemple précédent, nous devrions conclure seulement que les données ne fournissent aucune preuve tangible contre H_0. Dans cet exemple, le test donne une certaine évidence contre H_0 parce que la statistique t est proche de la valeur critique à 10 % (Idéalement, nous devrions calculer une telle valeur mais ceci exigerait un logiciel spécial à cause de la distribution non normale). En plus, bien que $\hat{\rho} \approx 0{,}91$ implique un certain degré de persistance pour $\{r3_t\}$, la corrélation entre des observations qui sont espacées de 10 périodes dans un modèle AR(1) est à peu près de 0,35, contre presque 1 si $\rho = 1$.

Quid si maintenant nous utilisons $r3_t$ comme une variable explicative dans une analyse de régression ? Le résultat du test de racine unitaire implique que nous devrions être extrêmement prudents : si $r3_t$ possède une racine unitaire, les approximations asymptotiques habituelles ne sont pas valables (comme nous en avons discuté dans le chapitre 11). Une solution est d'utiliser la première différence de $r3_t$ dans chaque analyse. Mais comme nous le verrons dans la section 18.4, ce n'est pas la seule possibilité.

Nous devons aussi pouvoir mener des tests de racine unitaire dans des modèles avec une dynamique plus compliquée. Si $\{y_t\}$ suit (18.17) avec $\rho = 1$, alors Δy_t est non corrélé sériellement. On peut alors aisément faire en sorte que $\{\Delta y_t\}$ suive un modèle AR en augmentant l'équation (18.21) avec des retards supplémentaires. Par exemple,

$$\Delta y_t = \alpha + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + e_t, \qquad \textbf{[18.23]}$$

Dans lequel $|\gamma| < 1$. Ceci garantit que, sous $H_0 : \theta = 0$, $\{\Delta y_t\}$ suit un modèle stable AR(1). Sous l'alternative $H_1 : \theta < 0$, on peut montrer que $\{y_t\}$ suit un modèle stable AR(2).

Plus généralement, on peut ajouter p retards de Δy_t à l'équation pour tenir compte de la dynamique dans le processus. La manière de tester l'hypothèse nulle de racine unitaire est similaire : on régresse

$$\Delta y_t \text{ sur } y_{t-1}, \Delta y_{t-1}, \ldots, \Delta y_{t-p} \qquad \textbf{[18.24]}$$

et on applique le test t test sur $\hat{\theta}$, le coefficient de y_{t-1}, comme précédemment. Cette version étendue du test de Dickey-Fuller est appelée le test de **Dickey-Fuller augmenté** car la régression a été enrichie des variations retardées, les Δy_{t-h}. Les valeurs critiques et la règle de rejet sont les mêmes qu'auparavant. L'inclusion des variations retardées dans (18.24) vise à ôter toute corrélation sérielle dans Δy_t. Plus nous incluons de retards dans (18.24), plus nous perdons d'observations. Si nous incluons trop de retards, la puissance du test en petit échantillon s'en trouve affectée. Mais si nous décidons d'inclure trop peu de retards, la taille du test sera incorrecte, même asymptotiquement, car la validité des valeurs critiques dans le tableau 18.2. est conditionnelle à une dynamique complètement modélisée. Souvent, la longueur des retards est dictée par la fréquence des données (ainsi que par la taille de l'échantillon). Pour les données annuelles, un ou deux retards sont généralement suffisants. Pour les données mensuelles, on peut inclure 12 retards. Mais il n'y a pas de règle intangible à suivre dans chaque cas.

Il est intéressant de noter que les statistiques t des variations retardées suivent approximativement des distributions t. Les statistiques F testant la significativité conjointe de chaque groupe de termes Δy_{t-h} sont également asymptotiquement valables. (Sous réserve de validité de l'hypothèse d'homoscédasticité discutée dans la section 11.5). Par conséquent, nous pouvons utiliser les tests standards pour déterminer si nous avons assez de variations retardées dans (18.24).

Pour les séries qui ont clairement une tendance temporelle, on doit modifier le test de racine unitaire. Un processus stationnaire autour d'une tendance – qui possède une tendance temporelle dans sa moyenne mais qui est I(0) autour de sa tendance – peut être confondu avec un processus de racine unitaire si on ne contôle pas pour la tendance temporelle dans la régression de Dickey-Fuller. En d'autres termes, si on mène le test habituel DF ou le test DF augmenté sur une série I(0) mais avec tendance, nous aurons sans doute trop peu de puissance pour rejeter la racine unitaire.

EXEMPLE 18.3
Test de racine unitaire pour l'inflation américaine annuelle

Nous utilisons des données sur l'inflation américaine, basée sur l'IPC, afin de tester la présence d'une racine unitaire (voir PHILLIPS). Nous nous focalisons sur les années allant de 1948 à 1996. En incluant un retard de Δinf_t dans la régression de Dickey-Fuller augmentée, nous obtenons :

$$\widehat{\Delta inf}_t = 1{,}36 - 0{,}310\ inf_{t-1} + 0{,}138\ \Delta inf_{t-1}$$
$$(0{,}517)(0{,}103) \qquad (0{,}126)$$
$$n = 47,\ R^2 = 0{,}172.$$

La statistique t pour le test de racine unitaire est $-0{,}310/0{,}103 = -3{,}01$. Comme la valeur critique à 5 % est $-2{,}86$, nous rejetons l'hypothèse nulle d'une racine unitaire au niveau de 5 % de significativité. La valeur estimée de ρ est à peu près de 0,690. Dans l'ensemble, il y a assez de preuves à l'encontre d'une racine unitaire dans l'inflation. Le retard Δinf_{t-1} présente une statistique t d'environ 1,10, si bien que nous n'avons pas à l'inclure, bien qu'on ne pouvait pas le savoir à l'avance. Si nous laissons tomber Δinf_{t-1}, l'évidence à l'encontre de la racine unitaire est un peu renforcée : $\hat{\theta} = -0{,}335$ ($\hat{\rho} = 0{,}665$), et $t_{\hat{\theta}} = -3{,}13$.

Pour prendre en compte les séries avec tendances temporelles, on change l'équation de base en

$$\Delta y_t = \alpha + \delta t + \theta y_{t-1} + e_t, \qquad \textbf{[18.25]}$$

dans laquelle une fois de plus l'hypothèse nulle est $H_0 : \theta = 0$ et l'alternative est $H_1 : \theta < 0$. Sous l'alternative, $\{y_t\}$ est un processus stationnaire autour d'une tendance. Si y_t possède une racine unitaire, alors $\Delta y_t = \alpha + \delta t + e_t$, et donc le *changement* de y_t possède une moyenne linéaire par rapport à t à moins que $\delta = 0$. [On peut montrer que $E(y_t)$ est en fait quadratique par rapport à t.] Il n'est pas inhabituel qu'une série en différence première possède une tendance linéaire, et donc une hypothèse nulle plus appropriée est sans doute $H_0 : \theta = 0$, $\delta = 0$. Bien qu'il soit possible de tester cette hypothèse jointe à l'aide d'un test F – mais avec des valeurs critiques modifiées –, il est habituel de tester $H_0 : \theta = 0$ seulement avec un test t. Nous suivons cette approche ici [Voir BDGH (1993, section 4.4) pour plus de détails sur le test joint].

Lorsque l'on inclut une tendance linéaire dans la régression, les valeurs critiques du test changent. Intuitivement, cela s'explique par le fait qu'enlever la tendance dans un processus racine unitaire le fait ressembler plus à un processus I(0). En conséquence, on exige une valeur plus élevée pour le test t afin de rejeter H_0. Les valeurs critiques du test en t de Dickey-Fuller qui incluent une tendance linéaire sont reproduites dans le tableau 18.3 et sont reprises de BDGH (1993, Table 4.2).

Tableau 18.3 *Valeurs critiques asymptotiques pour le test en t de racine unitaire avec tendance temporelle.*

Niveau de significativité	1 %	2,5 %	5 %	10 %
Valeur critique	−3,96	−3,66	−3,41	−3,12

Par exemple, pour rejeter une racine unitaire au niveau de 5 %, nous avons besoin d'une statistique t pour $\hat{\theta}$ inférieure à $-3{,}41$, contre $-2{,}86$ sans tendance temporelle.

On peut étendre l'équation (18.25) avec les retards de Δy_t pour tenir compte de la corrélation sérielle, de la même manière que dans le cas sans tendance temporelle.

EXEMPLE 18.4
Racine unitaire dans le log du produit domestique brut réel US.

On peut apliquer le test de racine unitaire avec tendance temporelle aux données du PNB américain dans le fichier INVEN. Les données annuelles couvrent les années allant de 1959 à 1995. On teste si $\log(GDP_t)$ possède une racine unitaire. La série possède une tendance évidente qui semble plus ou moins linéaire. On inclut un seul retard de $\Delta\log(GDP_t)$, qui est tout simplement la croissance du PNB (en format décimal), pour tenir compte de la dynamique :

$$\widehat{gGDP_t} = 1{,}65 + 0{,}0059\, t - 0{,}210 \log(GDP_{t-1}) + 0{,}264\, gGDP_{t-1} \qquad \textbf{[18.26]}$$
$$(0{,}67)\ (0{,}0027)\quad (0{,}087)\qquad\qquad (0{,}165)$$
$$n = 35,\ R^2 = 0{,}268.$$

À partir de cette équation, on obtient $\hat{\rho} = 1 - 0{,}21 = 0{,}79$, qui est clairement inférieur à un. Mais nous ne *pouvons pas* rejeter la racine unitaire pour le log du PNB : la statistique t relative à $\log(GDP_{t-1})$ est $-0{,}210/0{,}087 = -2{,}41$, ce qui est bien au-dessus de la valeur critique à 10 % de $-3{,}12$. La statistique t de $gGDP_{t-1}$ est égale à 1,60, ce qui est presque significatif à 10 % contre une alternative bilatérale.

Que devons nous conclure à propos de la racine unitaire ? Une fois de plus, nous ne pouvons pas rejeter la racine unitaire, mais la valeur estimée de ρ n'est pas spécialement proche de un. Lorsque nous avons un petit échantillon en terme de taille – et $n = 35$ est considéré comme particulièrement petit – il est difficile de rejeter l'hypothèse nulle de la racine unitaire si le processus possède quelque chose proche d'une racine unitaire. En utilisant des données supplémentaires sur des périodes de temps plus longues, beaucoup de chercheurs ont conclu qu'il y a peu de preuves contre l'hypothèse nulle de racine unitaire du $\log(PNB)$. Cela les a amenés à conclure que la croissance du PNB est I(0), ce qui signifie que le $\log(PNB)$ est I(1). Malheureusement, étant donné les tailles actuelles des échantillons, on ne peut pas avoir beaucoup de confiance en cette conclusion.

Si nous omettons la tendance temporelle, on a beaucoup moins de preuves à l'encontre de H_0, puisque $\hat{\theta} = -0{,}023$ et $t_{\hat{\theta}} = -1{,}92$. La valeur estimée de ρ est ici beaucoup plus proche de un, mais c'est trompeur car nous avons omis la tendance linéaire.

Il est tentant de comparer la statistique t relative à la tendance temporelle dans (18.26) à la valeur critique d'une distribution standard normale ou en t, pour voir si la tendance temporelle est significative. Malheureusement, la statistique t relative à la tendance ne possède pas une distribution asymptotique standard normale (sauf si $|\rho| < 1$). La distribution asymptotique de cette statistique t est connue, mais est rarement utilisée. Habituellement, on se repose sur l'intuition (ou sur le grahique des séries temporelles) pour décider si l'on doit inclure une tendance dans le test de DF.

Il existe beaucoup d'autres variantes des tests de racines unitaires. Dans une version qui est seulement applicable si les séries sont sans tendance, l'intercept est omis de la régression ; c'est-à-dire, on impose α égal à zéro dans (18.21). Cette variante du test de Dickey-Fuller est rarement utilisée à cause des biais liés au cas $\alpha \neq 0$. On peut aussi permettre des tendances linéaires plus compliquées, telles que des tendances quadratiques. Une fois de plus, ceci est rarement utilisé.

Une autre classe de tests s'évertue à prendre en compte la correlation sérielle dans Δy_t d'une manière différente qu'en incluant des retards dans (18.21) ou (18.25). Cette approche est liée aux écarts-types robustes à la corrélation sérielle pour les estimateurs MCO que nous avons couverts dans la section 12.5. L'idée est d'être le plus agnostique possible à propos de la corrélation sérielle dans Δy_t. En pratique, le test de Dickey-Fuller (augmenté) se comporte plutôt bien. [Voir BDGH (1993, section 4.3) pour une discussion des autres tests.]

18.3 RÉGRESSION FALLACIEUSE

Dans un contexte en coupes transversales, on utilise le terme « corrélation fallacieuse » pour décrire une situation où deux variables sont liées à travers la corrélation avec une troisième variable. En particulier, si on régresse y sur x, on trouve une relation significative. Mais si on contrôle pour une autre variable, disons z, l'effet partiel de x sur y devient nul. Bien sûr, cela peut arriver aussi dans un contexte de séries temporelles avec des variables I(0).

Comme on en a discuté dans la Section 10.5., il est possible de trouver une relation fallacieuse entre des séries temporelles qui ont des tendances croissantes ou décroissantes. Pour autant que les séries soient faiblement dépendantes au niveau de leurs tendances temporelles, le problème est résolu en incluant une tendance temporelle dans le modèle de régression.

Lorsque nous traitons des processus intégrés d'ordre un, il y a une complication supplémentaire. Même si deux séries ont des moyennes sans tendance, une régression impliquant deux séries I(1) *indépendantes* donnera souvent une statistique en t significative.

Plus précisément, soient $\{x_t\}$ and $\{y_t\}$ des marches aléatoires générées par

$$x_t = x_{t-1} + a_t,\ t = 1, 2, \ldots, \qquad \textbf{[18.27]}$$

Et

$$y_t = y_{t-1} + e_t,\ t = 1, 2, \ldots, \qquad \textbf{[18.28]}$$

où $\{a_t\}$ et $\{e_t\}$ sont des innovations distribuées de manière identique et indépendante, avec une moyenne nulle et des variances respectivement données par σ_a^2 and σ_e^2. Pour faire simple, supposons les valeurs initiales comme $x_0 = y_0 = 0$. Supposons en outre que $\{a_t\}$ et $\{e_t\}$ soient des processus indépendants. Cela implique que $\{x_t\}$ et $\{y_t\}$ sont aussi indépendants. Quid dans ce cas si néanmoins nous effectuons la régression

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t \qquad \textbf{[18.29]}$$

et si nous regardons le résultat du test t habituel pour $\hat{\beta}_1$ et le R-carré classique ? Comme y_t et x_t sont indépendants, nous espérons que plim $\hat{\beta}_1 = 0$. Et même de manière plus importante, si nous testons $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ au seuil de 5 %, nous espérons que la statistique t de $\hat{\beta}_1$ sera non significative dans 95 % des cas. À l'aide d'une simulation, Granger and Newbold (1974) ont montré que ce n'est pas le cas : même si y_t et x_t sont *indépendants*, la régression de y_t sur x_t générera une statistique t significative dans un pourcentage élevé des cas, beaucoup plus élevé que le niveau nominal de significativité.

Granger and Newbold ont appelé cela le **problème de régression fallacieuse** : il n'y a aucune raison que y et x soient liés, mais une régression MCO utilisant une statistique t indiquera souvent une relation.

Des résultats récents de simulation ont été effectués par Davidson and MacKinnon (1993, Table 19.1), où a_t et e_t sont générées comme des variables aléatoires indépendantes et identiquement distribuées selon une loi normale, et dans lesquelles 10 000 échantillons différents sont générés. Pour une taille d'échantillon de $n = 50$ au niveau de significativé de 5 %, la statistique en t standard correspondant à $H_0 : \beta_1 = 0$ contre l'alternative bilatérale rejette H_0 dans 66.2 % des cas, au lieu de 5 % du temps. Lorsque la taille de l'échantillon augmente, les choses empirent : avec $n = 250$, l'hypothèse nulle est rejetée dans 84,7 % des cas !

Exploration supplémentaire 18.2

Dans l'analyse précédente, où $\{x_t\}$ et $\{y_t\}$ sont générés par (18.27) and (18.28) et où $\{e_t\}$ et $\{a_t\}$ sont des séquences i.i.d., quelle est la plim du coefficient de pente, disons $\hat{\gamma}_1$, obtenu dans une régression de Δy_t sur Δx_t ? Décrivez le comportement de la statistique t de $\hat{\gamma}_1$.

Voici la manière de visualiser ce qu'il se passe lorsque l'on régresse le niveau de y sur le niveau de x. Écrivons le modèle sous-jacent (18.29) comme :

$$y_t = \beta_0 + \beta_1 x_t + u_t. \qquad \textbf{[18.30]}$$

Pour que la statistique t de $\hat{\beta}_1$ ait au minimum une distribution approximativement normale standard dans des grands échantillons, $\{u_t\}$ doit être un processus non corrélé sériellement de moyenne zéro. Mais sous $H_0 : \beta_1 = 0$, $y_t = \beta_0 + u_t$, et, parce que $\{y_t\}$ est une marche aléatoire débutant en $y_0 = 0$, l'équation (18.30) tient sous H_0 seulement si $\beta_0 = 0$, et de manière plus importante si $u_t = y_t = \sum_j^t e_j$. En d'autres termes, $\{u_t\}$ est une marche aléatoire sous H_0. Cela viole clairement même la version asymptotique des hypothèses de Gauss-Markov du chapitre 11.

Inclure une tendance temporelle ne change pas vraiment la conclusion. Si y_t ou x_t suivent une marche aléatoire avec dérive et si une tendance temporelle n'est pas incluse, le problème de régression fallacieuse est encore plus grave. Les mêmes conclusions qualitatives s'appliquent si $\{a_t\}$ and $\{e_t\}$ sont des processus I(0) généraux, plutôt ques des séquences i.i.d.

En plus du fait que la statistique t n'ait pas une distribution normale standard – en fait elle augmente vers l'infini lorsque $n \to \infty$ – le comportement du R-carré est non standard. Dans un contexte en coupes transversales ou dans des régressions avec des variables en séries temporelles I(0), le R-carré converge en probabilité vers le R-carré de la population : $1 - \sigma_u^2/\sigma_y^2$. Ce n'est pas le cas dans des régressions fallacieuses avec des processus I(1). Plutôt que d'avoir un R-carré avec une plim bien définie, il converge en fait vers une variable aléatoire. La formalisation de cette notion est au-delà de la portée de ce texte. [Une discussion des propriétés asymptotiques de la statistique t et du R-carré peut être trouvée dans BDGH (section 3.1).] Le résultat est que le R-carré est élevé avec une forte probabilité même lorsque $\{y_t\}$ et $\{x_t\}$ sont des processus de séries temporelles indépendants.

Les mêmes considérations s'appliquent avec des variables indépendantes multiples, chacune d'entre elles pouvant être I(1) ou certaines d'entre elles pouvant être I(0). Si $\{y_t\}$ est I(1) et si au moins certaines des variables explicatives sont I(1), les résultats de la régression peuvent être fallacieux.

La possibilité de régression fallacieuse avec des variables I(1) est très importante et a amené les économistes à revoir beaucoup de régressions de séries temporelles agrégées dont les statistiques t étaient très significatives et dont les R-carré étaient extrêmement élevés. Dans la section suivante, nous montrons que régresser une variable dépendante I(1) sur une variable indépendante I(1) peut-être informatif mais seulement si ces variables sont reliées dans un sens bien précis.

18.4 COINTÉGRATION ET MODÈLES À CORRECTION D'ERREUR

La discussion de la régression fallacieuse dans la section précédente nous rend certainement attentifs concernant l'utilisation des variables I(1) en niveau dans une analyse de régression. Dans les chapitres précédents nous avons suggéré que les variables I(1) doivent d'abord être prises en différences premières avant d'être utilisées dans des modèles de régression linéaire qui peuvent être estimés par MCO ou par variables instrumentales. C'est certainement une bonne procédure à suivre et c'est l'approche utilisée dans beaucoup de régressions en séries temporelles après le papier originel de Granger et Newbold sur le problème de régression fallacieuse. Malheureusement prendre la différence première des variables I(1) de manière récurrente limite la portée des questions que l'on peut aborder.

Cointégration

La notion de **cointégration** qu'Engle et Granger (1987) ont traitée formellement donne du sens aux régressions impliquant des variables I(1). Un traitement complet de la cointégration est exigeant du point de vue mathématique mais nous pouvons décrire les problèmes de base et les méthodes à utiliser dans beaucoup d'applications.

Si $\{y_t : t = 0, 1, \ldots\}$ et $\{x_t : t = 0, 1, \ldots\}$ sont deux processus I(1), alors en général, $y_t - \beta x_t$ est un processus I(1) pour n'importe quelle valeur de β. Néanmoins, il est possible que pour une une certaine valeur de $\beta \neq 0$, $y_t - \beta x_t$ soit un processus I(0), ce qui signifie qu'il possède une moyenne constante, une variance constante, des autocorrélations qui dépendent seulement de la distance temporelle entre deux variables dans la série et qu'il est asymptotiquement non corrélé. Si un tel β existe, on dira que y and x sont *cointégrés*, et nous appellerons β le paramètre de cointégration. [Alternativement, nous pouvons regarder $x_t - \gamma y_t$ pour $\gamma \neq 0$: si $y_t - \beta x_t$ est I(0), alors $x_t - (1/\beta)y_t$ est I(0). Par conséquent la combinaison linéaire de y_t et x_t n'est

pas unique mais si nous fixons le coefficient de y_t à l'unité alors β est unique. Voir problème 3. De manière concrète, nous considérerons des combinaisons linéaires de la forme $y_t - \beta x_t$.]

Exploration supplémentaire 18.3

Soit $\{(y_t, x_t) : t = 1,2,\ldots\}$ deux séries temporelles pour lesquelles chaque série est I(1) sans dérive. Expliquez pouquoi si y_t et x_t sont cointégrés, y_t et x_{t-1} sont aussi cointégrés.

À titre d'exemple, prenons $\beta = 1$, supposons que $y_0 = x_0 = 0$, et écrivons $y_t = y_{t-1} + r_t$, $x_t = x_{t-1} + v_t$, où $\{r_t\}$ et $\{v_t\}$ sont deux processus I(0) avec moyennes nulles. Donc, y_t et x_t ont tendance à s'écarter et à ne pas revenir vers leur valeur initiale de zéro de manière régulière. Par contre, si $y_t - x_t$ est I(0), il a une moyenne égale à zéro et revient bien de manière régulière vers zéro.

En terme d'exemple spécifique, prenons $r6_t$, le taux d'intérêt annuel pour les obligations d'État à six mois (à la fin du trimestre t) et prenons $r3_t$ comme le taux d'intérêt annuel pour les obligations d'État à trois mois (ils sont habituellement appelés en anglais « bond equivalent yields » et sont reportés dans les journaux financiers). Dans l'exemple 18.2, en utilisant les données dans le fichier INTQRT, nous obtenons peu d'évidence contre l'hypothèse selon laquelle $r3_t$ possède une racine unitaire ; la même chose s'applique à $r6_t$. Définissons le spread $spr_t = r6_t - r3_t$ comme la différence de rendements entre les obligations d'État à six mois et à trois mois. Alors, en utilisant l'équation (18.21), la statistique t de Dickey-Fuller pour spr_t est $-7,71$ (avec $\hat{\theta} = -0,67$ ou $\hat{\rho}_1 = 0,33$). Dès lors, nous rejetons fortement la racine unitaire pour spr_t en faveur d'un processus I(0). La leçon à tirer est que bien que $r6_t$ et $r3_t$ apparaissent comme des processus à racine unitaire chacun, la différence entre eux est un processus I(0). En d'autres termes, *r6* et *r3* sont cointégrés.

La cointégration dans cet exemple, comme dans beaucoup d'exemples, possède une interpretation économique. Si *r6* et *r3* n'étaient pas cointegrés, la différence entre les taux d'intérêt pourrait devenir très élevée, avec aucune tendance pour eux de revenir l'un vers l'autre. Sur base d'un argument simple d'arbitrage, ceci semble impossible. Supposons que le spead continue de croître pendant plusieurs périodes de temps, faisant des obligations à six mois un investissement beaucoup plus intéressant. Dans ce cas les investisseurs se détourneraient des obligations à trois mois en faveur des obligations à six mois, poussant vers le haut les prix des obligations à six mois et poussant les rendements des obligations à trois mois vers le bas. Dans la mesure où les taux d'intérêts sont inversement reliés au prix, ceci diminuerait *r6* et augmenterait *r3* jusqu'au moment où le spread est réduit. Par conséquent, on ne peut pas attendre que des déviations importantes entre les taux perdurent : le spread a tendance à revenir vers sa valeur moyenne. (En fait le spread moyen est légèrement positif parce que les investisseurs de long terme sont mieux rémunérés comparativement aux investisseurs de court terme.)

Il y a une autre manière de caractériser le fait que spr_t ne déviera pas durant de longues périodes de temps par rapport à sa valeur moyenne : *r6* et *r3* possèdent une relation de long terme. Pour préciser ce que l'on veut dire par là, supposons que $\mu = \mathrm{E}(spr_t)$ dénote la valeur attendue du spread. Alors nous pouvons écrire :

$$r6_t = r3_t + \mu + e_t,$$

où $\{e_t\}$ a une moyenne égale à zéro et est un processus I(0). L'équilibre ou la relation de long terme intervient quand $e_t = 0$, ou $r6^* = r3^* + \mu$. En tout point du temps, il peut y avoir des déviations par rapport à l'équilibre mais elles seront temporaires : Il y a des forces économiques qui ramènent *r6* et *r3* vers la relation d'équilibre.

Dans l'exemple du taux d'intérêt nous avons utilisé un raisonnement économique pour assigner une valeur à β si y_t et x_t sont cointégrés. Si nous avons une valeur hypothétique de β, alors tester si deux séries sont cointégrées est simple : on définit simplement une nouvelle variable, $s_t = y_t - \beta x_t$, et on applique le test usuel DF ou le test

DF augmenté à $\{s_t\}$. Si l'on rejette une racine unitaire au niveau de $\{s_t\}$ en faveur d'une alternative I(0), alors, on conclut que y_t et x_t sont cointégrés. En d'autres termes, l'hypothèse nulle est que y_t et x_t sont non cointégrés.

Tester la cointégration est plus compliqué lorsque le paramètre de cointégration (potentiel) β est inconnu. Plutôt que de tester la présence d'une racine unitaire dans $\{s_t\}$, on doit d'abord estimer β. *Si* y_t et x_t sont cointégrés, il s'avère que l'estimateur MCO $\hat{\beta}$ de la régression

$$y_t = \hat{\alpha} + \hat{\beta} x_t \qquad \textbf{[18.31]}$$

est convergent pour β. Sous l'hypothèse nulle, les deux séries ne sont pas cointégrées, ce qui signifie que, sous H_0, on mène une regression fallacieuse. Heureusement, il est possible de tabuler les valeurs critiques même lorsque β est estimé, pour lesquelles on applique le test Dickey-Fuller ou Dickey-Fuller augmenté aux résidus à savoir, $\hat{u}_t = y_t - \hat{\alpha} - \hat{\beta} x_t$, à partir de (18.31). La seule différence est que ces valeurs critiques tiennent compte que β est estimé. Le test qui en résulte est appelé le test d'**Engle-Granger**, et les valeurs critiques asymptotiques sont données dans le tableau 18.4. Elles sont tirées de Davidson and MacKinnon (1993, Tableau 20.2).

Tableau 18.4 *Valeurs critiques asymptotiques du test de cointégration sans tendance temporelle*

Niveau de significativité	1 %	2,5 %	5 %	10 %
Valeur critique	−3,90	−3,59	−3,34	−3,04

Dans le test de base, on régresse $\Delta\hat{u}$ sur $\hat{u}_{t-1}$ et on compare la statistique t relative à $\hat{u}_{t-1}$ à la valeur critique adéquate du Tableau 18.4. Si la statistique t est en-dessous de la valeur critique, nous avons alors la preuve que $y_t - \beta x_t$ est I(0) pour une certaine valeur de β ; c'est-à-dire y_t et x_t sont cointégrés. Nous pouvons ajouter des retards pour tenir compte de la correlation sérielle. Si nous comparons les valeurs critiques du tableau 18.4 avec celles du tableau 18.2, nous avons besoin d'une statistique t beaucoup plus importante en termes de taille afin de conclure en faveur de la cointégration par rapport au cas où nous utiliserions une valeur critique usuelle DF. Ceci intervient parce que les MCO qui minimisent la somme des résidus au carré tendent à produire des résidus qui ressemblent à une séquence I(0) même si y_t et x_t sont non cointégrés.

Comme pour le test usuel Dickey-Fuller, on peut augmenter le test d'Engle-Granger en incluant des retards de $\Delta\hat{u}_t$ comme régresseurs supplémentaires.

Si y_t et x_t sont non cointégrés, une régression de y_t sur x_t est fallacieuse et ne nous dit rien de valable : il n'y a pas de relation de long terme entre y et x. On peut toujours faire tourner une régression impliquant les premières differences, Δy_t et Δx_t, en incluant des retards. Mais nous devons interpréter ces régressions pour ce qu'elles sont : elles expliquent la différence de y en terme de différence de x et n'ont rien à voir nécessairement avec une relation en niveau.

Si y_t et x_t *sont* cointégrés, on peut utiliser cela pour spécifier des modèles dynamiques plus généraux, comme nous le verrons dans la sous-section suivante.

La discussion précédente suppose que ni y_t ni x_t ne possède une tendance. C'est raisonnable pour les taux d'intérêt mais ça ne l'est pas pour d'autres séries temporelles. Si y_t et x_t possèdent une dérive alors $E(y_t)$ et $E(x_t)$ sont des fonctions linéaires habituellement croissantes par rapport au temps. La définition stricte de la cointégration requiert que $y_t - \beta x_t$ soit I(0) sans tendance. Pour voir ce que cela implique, écrivons $y_t = \delta_t + g_t$ et $x_t = \lambda_t + h_t$, où $\{g_t\}$ et $\{h_t\}$ sont des processus I(1), δ est une tendance dans y_t [$\delta = E(\Delta y_t)$], λ est une tendance dans x_t [$\lambda = E(\Delta x_t)$]. Maintenant si y_t et x_t sont cointégrés, il doit exister un β tel que $g_t - \beta h_t$ est I(0). Mais alors,

$$y_t - \beta x_t = (\delta - \beta\lambda)t + (g_t - \beta h_t),$$

qui est généralement un processus *stationnaire autour d'une tendance*. La forme stricte de cointégration exige qu'il n'y ait pas de dérive, ce qui signifie que $\delta = \beta\lambda$. Pour les processus I(1) avec dérive, il est possible que les parties stochastiques – c'est à dire g_t et h_t – soient cointégrées mais que le paramètre β qui fait que $g_t - \beta h_t$ est I(0) n'élimine pas la tendance temporelle linéaire.

On peut tester la cointégration entre g_t and h_t, sans prendre position sur la partie tendance en faisant tourner la régression

$$\hat{y}_t = \hat{\alpha} + \hat{\eta}_t + \hat{\beta}x_t \qquad \textbf{[18.32]}$$

et en appliquant le test DF ou DF augmenté sur les résidus $\hat{u}_t$. Les valeurs critiques asymptotiques sont données dans le tableau 18.5 [à partir de Davidson and MacKinnon (1993, Tableau 20.2)]. Une conclusion en faveur de la cointégration laisse ouverte la possibilité que $y_t - \beta x_t$ ait une tendance linéaire mais permet au moins de rejeter l'hypothèse d'un processus I(1).

Tableau 18.5 *Valeurs critiques asymptotiques du test de cointégration avec tendance temporelle*

Niveau de significativité	1 %	2,5 %	5 %	10 %
Valeur critique	−4,32	−4,03	−3,78	−3,50

EXEMPLE 18.5
Cointégration entre la fertilité et l'exemption personnelle

Dans les chapitres 10 selon et 11, nous avons étudié différents modèles pour estimer la relation entre le taux de fertilité générale (*gfr*) et la valeur réelle de l'exemption d'impôts personnelle (*pe*) aux États-Unis. Les résultats de régression statique en niveau et en différences premières sont sensiblement différents. Les régressions en niveau avec une tendance temporelle inclue donnent un coefficient MCO de *pe* égal à 0,187 (se = 0,035) et $R^2 = 0{,}500$. En différence première (sans tendance), le coefficient de Δpe est −0,043 (se = 0,028), and $R^2 = 0{,}032$. Bien qu'il y ait d'autres raisons expliquant la différence – telles qu'une dynamique des retards distribués mal spécifiée – la différence entre les régressions en niveaux et en variations suggère que nous devrions tester la cointégration. Bien sûr, cela suppose que *gfr* et *pe* sont des processus I(1). Cela semble être le cas : les tests DF augmentés avec une seule variation retardée et une tendance linéaire temporelle donnent chacun des statistiques *t* d'à peu près −1,47 et les coefficients estimés AR(1) sont proches de un.

Lorsque nous obtenons les résidus à partir de la régression de *gfr* sur *t* et *pe* et lorsque nous appliquons les tests DF augmentés avec un retard, nous obtenons une statistique *t* relative à $\hat{u}_{t-1}$ égale à −2,43, ce qui est très loin de la valeur critique à 10 %, −3,50. En conséquence, nous devons conclure qu'il y a peu d'évidence en faveur de la cointégration entre *gfr* et *pe*, même en permettant des tendances séparées. Il est très possible que les résultats de régression précédents que nous avons obtenus en niveaux souffrent du problème de régression fallacieuse. La bonne nouvelle est que lorsque nous utilisons les différences premières et que nous permettons la présence de 2 retards – voir l'équation (11.27) –, nous trouvons un effet de long terme positif et significatif de Δpe sur Δgfr.

Si nous pensons que deux séries sont cointégrées, nous serons souvent désireux de tester des hypothèses à propos du paramètre de cointégration. Par exemple, une théorie peut affirmer que le paramètre de cointégration est égal à un. Idéalement nous pourrions utiliser une statistique *t* pour tester cette hypothèse.

Nous couvrons explicitement le cas sans tendance temporelle bien que l'extension au cas de la tendance linéaire soit immédiate. Lorsque y_t et x_t sont I(1) et cointégrés, on peut écrire

$$y_t = \alpha + \beta x_t + u_t, \qquad \textbf{[18.33]}$$

où u_t est un processus I(0) de moyenne nulle. Généralement, $\{u_t\}$ est sujet à la corrélation sérielle, mais nous savons depuis le chapitre 11 que ceci n'affecte pas la convergence des estimateurs MCO. Comme mentionné précédemment, MCO appliqué à (18.33) estime de manière convergente β (et α). Malheureusement, parce que x_t est I(1), les procédures d'inférence usuelles ne s'appliquent pas nécessairement : MCO n'est pas distribué asymptotiquement de manière normale et la statistique t pour $\hat{\beta}$ n'a pas nécessairement une distribution en t approximative. Nous savons depuis le chapitre 10 que si $\{x_t\}$ est strictement exogène – voir Hypothèse TS.3 – et que les erreurs sont homoscédastiques, non corrélées sériellement, et normalement distribuées, l'estimateur MCO est aussi normalement distribué (conditionnellement aux variables explicatives) et que la stastique t a une distribution en t exacte. Malheureusement ces hypothèses sont trop fortes pour s'appliquer dans la plupart des situations. La notion de cointégration n'implique rien d'autre à propos de la relation entre $\{x_t\}$ et $\{u_t\}$ – en fait elles peuvent être arbitrairement corrélées –. En outre, à part exiger que $\{u_t\}$ soit I(0), la cointégration entre y_t et x_t ne restreint pas la dépendence sérielle de $\{u_t\}$.

Heureusement, la caractéristique de (18.33) qui rend l'inférence plus difficile – le manque d'exogénéité stricte de $\{x_t\}$ – peut être ajustée. Puisque x_t est I(1), la notion propre d'exogénéité stricte est que u_t soit non corrélée avec Δx_s, pour tout t et s. On peut toujours ajuster cela pour un *nouvel* ensemble d'erreurs, au moins approximativement, en écrivant u_t comme une fonction des Δx_s pour tous les s proches de t. Par exemple,

$$u_t = \eta + \phi_0 \Delta x_t + \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + \gamma_1 \Delta x_{t+1} + \gamma_2 \Delta x_{t+2} + e_t, \qquad \textbf{[18.34]}$$

où, par construction, e_t est non corrélé avec chaque Δx_s apparaissant dans l'équation. L'espoir est que e_t soit non corrélé avec des retards et valeurs avancées additionnelles de Δx_s. Nous savons que, quand $|s-t|$ devient grand, la corrélation entre e_t et Δx_s tend vers zéro, parce que ce sont des processus I(0). Dès lors, si nous insérons (18.34) dans (18.33), on obtient

$$y_t = \alpha_0 + \beta x_t + \phi_0 \Delta x_t + \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + \gamma_1 \Delta x_{t+1} + \gamma_2 \Delta x_{t+2} + e_t. \qquad \textbf{[18.35]}$$

Cette équation semble un petit peu étrange parce que les Δx_s futurs apparaissent avec les Δx_t retardés et contemporains. Le point clé est que le coefficient de x_t est toujours β, et par construction, x_t est maintenant strictement exogène dans cette équation. L'hypothèse de stricte exogénéité est la condition importante nécessaire pour obtenir une statistique t approximativement normale pour $\hat{\beta}$. Si u_t est non corrélé avec tous les Δx_s, $s \neq t$, alors nous pouvons laisser tomber les retards et valeurs avancées des variations et simplement inclure le changement contemporain, Δx_t. Alors l'équation que nous estimons ressemble à quelque chose de plus standard et inclut encore la première différence de x_t avec son niveau : $y_t = \alpha_0 + \beta x_t + \phi_0 \Delta x_t + e_t$. En effet, en ajoutant Δx_t, cela résout toute endogénéité contemporaine entre x_t et u_t. (Souvenons-nous que n'importe quelle endogénéité n'entraîne pas la non-convergence. En revanche, on essaye d'obtenir une statistique t asymptotiquement normale.) Savoir si nous devons inclure ou non des retards et des valeurs avancées des variations, et combien, est vraiment un problème empirique. Chaque fois que nous utilisons un retard additionnel ou une valeur avancée additionnelle, nous perdons une observation et ceci peut être coûteux à moins d'avoir une grande base de données.

L'estimateur MCO de β à partir de (18.35) est appelé **l'estimateur à retards et valeurs avancées** de β du fait de la manière dont il emploit Δx. [Voir par exemple Stock and Watson (1993).] Le seul problème dont nous devons nous soucier à propos de (18.35) est de la corrélation sérielle dans $\{e_t\}$. Ceci peut être appréhendé en calculant un écart-type robuste à la corrélation sérielle pour $\hat{\beta}$ (comme décrit dans la section 12.5) ou en utilisant une correction standard AR(1) (comme Cochrane-Orcutt).

EXEMPLE 18.6
Paramètre de cointégration pour les taux d'intérêt.

Auparavant, nous avons testé la cointégration entre *r6* et *r3* – les taux d'intérêts des obligations d'État à six et à trois mois – en supposant que le paramètre de cointégration était égal à un. Ceci nous a amené à conclure en faveur de la cointégration et naturellement à conclure que le paramètre de cointégration était égal à un. Cependant, estimons maintenant le paramètre de cointégration directement et testons $H_0 : \beta = 1$. Nous appliquons l'estimateur à valeurs avancées et retards, avec deux valeurs avancées et deux retards ainsi que le changement contemporain. La valeur estimée de β est $\hat{\beta} = 1{,}038$, et l'écart-type MCO usuel est 0,0081. Par conséquent la statistique t pour $H_0 : \beta = 1$ est $(1{,}038 - 1)/0{,}0081 \approx 4{,}69$, ce qui est un rejet statistique clair de H_0. (Bien sûr, que 1,038 soit économiquement différent de 1 est une considération qui a du sens.) Il y a peu d'évidence de correlation sérielle dans les résidus si bien que nous pouvons utiliser cette statistique t comme ayant une distribution normale approximative. [En comparaison l'estimateur MCO de β sans les valeurs avancées, les retards ou termes contemporains – et en utilisant cinq observations en plus – est égal à 1,026 (se = 0,0077). Mais cette statistique t à partir de (18.33) n'est pas nécessairement valide.]

Il y a beaucoup d'autres estimateurs des paramètres de cointégration et ceci continue à être un domaine très actif de recherche. La notion de cointégration s'applique à plus que deux variables mais l'interprétation, les tests et l'estimation sont beaucoup plus compliqués. Un problème est que même après avoir normalisé un coefficient à l'unité, il peut y avoir plusieurs relations de cointégration. BDGH proposent une discussion et plusieurs réferences.

Modèles à correction d'erreur

En plus de nous apprendre quelque chose sur la relation potentielle de long terme entre deux séries, le concept de cointégration enrichit les types de modèles dynamiques à notre disposition. Si y_t et x_t sont des processus I(1) et ne sont pas cointégrés, nous pouvons alors estimer un modèle dynamique en premières différences. Comme exemple, considérons l'équation

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + u_t, \qquad \textbf{[18.36]}$$

où u_t a une moyenne égale à zéro étant donné Δx_t, Δy_{t-1}, Δx_{t-1}, et des retards supplémentaires. C'est essentiellement l'équation (18.16) mais en premières differences plutôt qu'en niveaux. Si nous considérons cela comme un modèle à retards distribués rationnels, nous pouvons trouver l'impact de court terme, l'impact de long terme et la distribution des retards pour Δy_t comme retard distribué par rapport à Δx.

Si y_t et x_t sont cointégrés avec le paramètre β, alors nous avons des variables supplémentaires I(0) que nous pouvons inclure dans (18.36). Supposons $s_t = y_t - \beta x_t$, si bien que s_t est I(0) et supposons par souci de simplicité que s_t a une moyenne égale à zéro. Maintenant nous pouvons inclure des retards de s_t dans l'équation. Dans le cas le plus simple nous incluons un retard de s_t :

$$\begin{aligned} \Delta y_t &= \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta s_{t-1} + u_t \\ &= \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta(y_{t-1} - \beta x_{t-1}) + u_t, \end{aligned} \qquad \textbf{[18.37]}$$

où $E(u_t | I_{t-1}) = 0$, et I_{t-1} contient l'information sur Δx_t et sur toutes les valeurs passées de x and y. Le terme $\delta(y_{t-1} - \beta x_{t-1})$ est appelé le *terme à correction d'erreur*, et (18.37) est un exemple d'**un modèle à correction d'erreur**. (Dans certains modèles à correction d'erreur le changement contemporain dans x, Δx_t, est omis. Savoir s'il est inclus ou non dépend en partie de l'objectif de l'équation. En prévision, Δx_t est rarement inclus pour les raisons que nous verrons dans la section 18.5)

Un modèle à correction d'erreur nous permet d'étudier la dynamique de court terme dans la relation entre y et x. Par souci de simplification, considérons le modèle sans retard de Δy_t et Δx_t :

$$\Delta y_t = \alpha_0 + \gamma_0 \Delta x_t + \delta (y_{t-1} - \beta x_{t-1}) + u_t, \qquad \textbf{[18.38]}$$

où $\delta < 0$. Si $y_{t-1} > \beta x_{t-1}$, alors y a dépassé l'équilibre durant la période précédente ; parce que $\delta < 0$, le terme à correction d'erreur fait en sorte que y est ramené vers l'équilibre. De manière similaire, si $y_{t-1} > \beta x_{t-1}$, le terme à correction d'erreur induit une variation positive de y vers l'équilibre. Comment estimer les paramètres d'un modèle à correction d'erreur ? Si nous connaissons β, c'est facile. Par exemple dans (18.38), on régresse simplement Δy_t sur Δx_t et s_{t-1}, où $s_{t-1} = (y_{t-1} - \beta x_{t-1})$.

EXEMPLE 18.7
Modèle à correction d'erreur pour rendements

Dans le Problème 6 du chapitre 11, on a régressé $hy6_t$, le rendement à trois mois de détention (en pourcentage) obtenu en achetant une obligation d'État à six mois au temps $t-1$ et en la vendant au temps t comme une obligation d'État à trois mois, sur $hy3_{t-1}$, le rendement à trois mois de détention obtenu en achetant une obligation d'État à trois mois au temps $t-1$. L'hypothèse des anticipations implique que le coefficient de pente ne devrait pas être statistiquement différent de un. Il s'avère qu'il y a une évidence en faveur d'une racine unitaire dans $\{hy3_t\}$, ce qui remet en question l'analyse standard de régression. Nous allons supposer que les deux rendements de détention suivent des processus I(1). L'hypothèse des anticipations implique, au minimum, que $hy6_t$ et $hy3_{t-1}$ sont cointegrés avec β égal à un, ce qui apparaît être le cas (voir l'exercice d'ordinateur C5). Sous cette hypothèse, un modèle à correction d'erreur est

$$\Delta hy6_t = \alpha_0 + \gamma_0 \Delta hy3_{t-1} + \delta(hy6_{t-1} - hy3_{t-1}) + u_t,$$

où u_t a une moyenne zéro étant donné tous les $hy3$ et $hy6$ datés au temps $t-1$ et avant. Les retards sur les variables dans le modèle à correction d'erreur sont dictés par l'hypothèse des anticipations.

Exploration supplémentaire 18.4

Comment testeriez-vous $H_0 : \gamma_0 = 1, \delta = -1$ dans le modèle à correction d'erreurs des rendements de détention ?

En utilisant les données dans INTQRT, on obtient

$$\widehat{\Delta hy6_t} = \underset{(0,043)}{0,090} + \underset{(0,264)}{1,218}\ \Delta hy3_{t-1} - \underset{(0,244)}{0,840}\ (hy6_{t-1} - hy3_{t-1}) \qquad \textbf{[18.39]}$$

$$n = 122,\ R^2 = 0,790.$$

Le coefficient de correction d'erreur est négatif et très significatif. Par exemple, si le rendement de détention de l'obligation d'État à six mois est au-dessus de celui des obligations d'État à trois mois d'un point, alors $hy6$ diminue de 0,84 en moyenne durant le trimestre suivant. De manière intéressante, $\hat{\delta} = -0,84$ n'est pas statistiquement different de -1, comme on peut le voir facilement en calculant l'intervalle de confiance à 95 %.

Dans beaucoup d'autres exemples, le paramètre de cointégration doit être estimé. Dans ce cas, on remplace s_{t-1} par $\hat{s}_{t-1} = y_{t-1} - \hat{\beta} x_{t-1}$, où $\hat{\beta}$ est un estimateur de β. Nous avons couvert l'estimateur standard MCO ainsi que l'estimateur à retards et valeurs avancées. Ceci soulève la question de la manière dont la variation d'échantillonnage dans $\hat{\beta}$ influence l'inférence des autres paramètres dans le modèle à correction d'erreurs. Heureusement comme Engle et Granger le montrent, on peut ignorer l'estimation préliminaire de β (asymptotiquement). Cette propriété est très pratique et implique que l'efficacité asymptotique des estimateurs des paramètres dans le modèle à correction d'erreurs n'est pas affectée par le fait qu'on utilise l'estimateur MCO ou l'estimateur à valeurs avancées et retards de $\hat{\beta}$. Bien sûr le choix de $\hat{\beta}$ affectera en général les paramètres estimés du modèle à correction d'erreurs dans un échantillon particulier mais nous n'avons pas de manière systématique de décider quel estimateur préliminaire de β doit être utilisé. La procédure par laquelle on remplace β par est appelée la **procédure en deux étapes d'Engle-Granger**.

18.5 PRÉVISION

Prévoir les séries temporelles économiques est très important dans certains domaines de l'économie et est une discipline qui continue à être activement étudiée. Dans cette section, on se concentre sur les méthodes de prévision basées sur la régression. Diebold (2001) propose une introduction complète à la prévision en incluant des développements récents.

Nous supposons dans cette section que le but principal est de prévoir les valeurs futures des processus de séries temporelles et non pas nécessairement d'estimer des modèles économiques structuraux ou de causalité.

Il est utile de déterminer tout d'abord certains éléments fondamentaux de prévision qui ne dépendent pas d'un modèle spécifique. Supposons qu'au temps t, nous voulions prévoir le résultat de y au temps $t + 1$, ou y_{t+1}. La période de temps peut correspondre à une année, un trimestre, un mois, une semaine ou même un jour. Soit I_t, qui dénote l'information que nous pouvons observer au temps t. Cet **ensemble d'informations** inclut y_t, les valeurs précédentes de y, et souvent d'autres variables qui sont datées au temps t ou avant. On peut combiner cette information de manière infinie pour prévoir y_{t+1}. Y a-t-il une procédure optimale ?

La réponse est oui, moyennant le fait que nous spécifions la perte associée à l'erreur de prévision. Soit f_t, qui dénote la prévision faite au temps t. On appelle f_t la **prévision une étape plus loin**. L'erreur de prévision est $e_{t+1} = y_{t+1} - f_t$, que nous observons une fois le résultat sur y_{t+1} observé. La mesure la plus courante de la perte est la même que celle relative à l'estimation par moindres carrés ordinaires d'un modèle de régression linéaire multiple : l'erreur au carré, e_{t+1}^2. L'erreur de prévision au carré considère les erreurs de prédictions positives et négatives de manière symétrique et les erreurs de prédiction plus importantes reçoivent un poids relatif plus important. Par exemple les erreurs de $+2$ et -2 donnent la même perte et génèrent une perte quatre fois plus grande que les erreurs de prédictions $+1$ et -1. L'erreur de prévision au carré est un exemple de **fonction de perte**. Une autre fonction de perte populaire est la valeur absolue de l'erreur de prediction $|e_{t+1}|$. Pour des raisons que nous verrons bientôt, nous nous nous concentrons maintenant sur la perte de l'erreur au carré.

Étant donnée la fonction de perte de l'erreur au carré, on peut déterminer la meilleure manière d'utiliser l'information au temps t pour prévoir y_{t+1}. Mais nous devons admettre qu'au temps t, nous ne connaissons pas e_{t+1} : c'est une variable aléatoire, parce que y_{t+1} est une variable aléatoire. Dès lors, tout critère pertinent pour choisir f_t doit être basé sur ce que nous savons au temps t. Il est normal de choisir la prévision qui minimise l'erreur de prévision au carré attendue, étant donné I_t :

$$\mathrm{E}(e_{t+1}^2 | I_t) = \mathrm{E}[(y_{t+1} - f_t)^2 | I_t]. \qquad \textbf{[18.40]}$$

Un fait de base en probabilité (voir Propriété CE.6 dans l'annexe B) est que l'espérance conditionnelle, $\mathrm{E}(y_{t+1}|I_t)$, minimise (18.40). En d'autres termes, si on veut minimiser l'erreur de prévision au carré attendue étant donné l'information au temps t, notre prévision devrait être la valeur attendue de y_{t+1} étant donné les variables que nous connaissons au temps t.

Pour beaucoup de processus de séries temporelles populaires, l'espérance conditionnelle est facile à obtenir. Supposons que $\{y_t : t = 0, 1, \ldots\}$ est une séquence de différences de martingale (SDM) et soit $I_t\{y_t, y_{t-1}, \ldots, y_0\}$, le passé observé de y. Par définition, $\mathrm{E}(y_{t+1}|I_t) = 0$ pour tout t ; la meilleure prévision de y_{t+1} au temps t est toujours zéro. Rappelons-nous depuis la section 18.2 qu'une séquence i.i.d. avec moyenne nulle est une séquence de différences de martingale.

Une séquence de différences de martingale est une séquence pour laquelle le passé n'est pas utile pour prédire le futur. Les rendements d'actions sont généralement perçus comme bien approximés par une SDM, ou en tout cas, avec une moyenne positive. Le point-clé est que $\mathrm{E}(y_{t+1}|y_t, y_{t-1}, \ldots) = \mathrm{E}(y_{t+1})$: la moyenne conditionnelle est égale à la moyenne non conditionnelle pour laquelle le passé de y n'aide pas à prédire le y futur.

Un processus $\{y_t\}$ est une **martingale** si $E(y_{t+1}|y_t, y_{t-1}, \ldots, y_0) = y_t$ pour tout $t \geq 0$. [Si $\{y_t\}$ est une martingale, alors $\{\Delta y_t\}$ est une séquence de différences de martingale, ce qui explique le terme.] La valeur prédite de y pour la période suivante est toujours la valeur de y pour cette période.

Un exemple plus compliqué est

$$E(y_{t+1}|I_t) = \alpha y_t + \alpha(1 - \alpha)y_{t-1} + \ldots + \alpha(1 - \alpha)^t y_0, \qquad \textbf{[18.41]}$$

où $0 < \alpha < 1$ est un paramètre que nous devons choisir. Cette méthode de prévision est appellée **lissage exponentiel** parce que les poids relatifs aux y retardés décroissent de manière exponentielle vers zéro.

La raison pour laquelle on écrit l'espérance comme dans (18.41) est que cela mène à une relation récurrente simple. Posons $f_0 = y_0$. Alors, pour $t \geq 1$, les prévisions peuvent être obtenues comme

$$f_t = \alpha y_t + (1 - \alpha)f_{t-1}.$$

En d'autres termes, la prévision de y_{t+1} est une moyenne pondérée de y_t et de la prévision de y_t faite au temps $t - 1$. Le lissage exponentiel convient seulement à certaines séries temporelles très spécifiques et exige de choisir α. Les méthodes de régression vers lesquelles nous nous tournons maintenant sont plus flexibles.

La discussion précédente s'est focalisée sur la prévision de y seulement une période à l'avance. Les problèmes généraux qui s'appliquent dans la prévision de y_{t+h} au temps t, où h est n'importe quel entier positif sont similaires. En particulier, si nous utilisons l'erreur de prévision au carré attendue comme notre mesure de pertes, le meilleur prédicteur est $E(y_{t+h}|I_t)$. Lorsque l'on traite une **prévision à plusieurs étapes à l'avance**, on utilise la notation $f_{t,h}$ pour indiquer la prévision de y_{t+h} faite au temps t.

Types de modèles de régression utilisés pour la prévision

Il y a beaucoup de modèles de régression différents que l'on peut utiliser pour prédire les valeurs futures des séries temporelles. Le premier modèle de régression pour données en séries temporelles du chapitre 10 était un modèle statique. Pour voir comment nous pouvons prévoir à partir de ce modèle, supposons que nous ayions une seule variable explicative :

$$y_t = \beta_0 + \beta_1 z_t + u_t. \qquad \textbf{[18.42]}$$

Supposons pour le moment que les paramètres β_0 et β_1 soient connus. Écrivons cette équation au temps $t + 1$ comme $y_{t+1} = \beta_0 + \beta_1 z_{t+1} + u_{t+1}$. Maintenant, si z_{t+1} est connu au temps t, si bien qu'il est un élément de I_t et que $E(u_{t+1}|I_t) = 0$, alors

$$E(y_{t+1}|I_t) = \beta_0 + \beta_1 z_{t+1},$$

où I_t inclut $z_{t+1}, y_t, z_t, \ldots, y_1, z_1$. Le terme de droite de cette équation est la prévision de y_{t+1} au temps t. Cette sorte de prévision est appelée habituellement une **prévision conditionnelle** parce qu'elle est conditionnelle au fait que l'on connaisse la valeur de z au temps $t + 1$. Malheureusement, en tout point du temps, on connaît rarement la valeur des variables explicatives pour les périodes de temps futures. Les exceptions incluent les tendances temporelles et les variables indicatrices saisonnières que nous couvrirons explicitement plus tard mais, à part cela, la connaissance de z_{t+1} au temps t est rare. De ce fait, nous voudrions générer des prévisions conditionnelles pour certaines valeurs de z_{t+1}.

Un autre problème avec (18.42) comme modèle de prévision est que $E(u_{t+1}|I_t) = 0$ implique que $\{u_t\}$ ne peut pas contenir de la corrélation, et c'est quelque chose que nous avons vu comme souvent violé dans la plupart des modèles de régression statiques. [Le problème 8 vous demande de dériver la prévision dans un modèle simple à retards distribués avec des erreurs AR(1).]

Si z_{t+1} n'est pas connue au temps t, nous ne pouvons pas l'inclure dans I_t. Alors, nous avons

$$E(y_{t+1}|I_t) = \beta_0 + \beta_1 E(z_{t+1}|I_t).$$

Cela signifie que pour prévoir y_{t+1}, nous devons d'abord prévoir z_{t+1}, sur base du même ensemble d'informations. Ceci est habituellement appelé **prévision non conditionnelle** parce qu'on ne suppose pas qu'on connaît z_{t+1} au temps t. Malheureusement, c'est un terme quelque peu galvaudé parce que notre prévision est toujours conditionnelle à l'information de I_t. Mais le nom est ancré dans la littérature sur la prévision.

Pour la prévision, à moins que nous soyons liés à un modèle statique comme dans (18.42) pour d'autres raisons, il est préférable de spécifier un modèle qui dépend seulement des valeurs retardées de y et z. Cela nous épargne une étape supplémentaire de prévision de la variable du terme de droite avant de prévoir y. Le type de modèle que nous avons en tête est

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + u_t$$
$$\mathrm{E}(u_t | I_{t-1}) = 0, \qquad \textbf{[18.43]}$$

où I_{t-1} contient y et z est daté au temps $t-1$ et avant. Maintenant, la prévision de y_{t+1} au temps t est $\delta_0 + \alpha_1 y_t + \gamma_1 z_t$; si nous connaissons les paramètres, nous pouvons juste insérer les valeurs de y_t et de z_t.

Si nous voulons seulement utiliser le passé de y pour prédire le futur de y, alors nous pouvons laisser tomber z_{t-1} à partir de (18.43). Évidemment nous pouvons ajouter plus de retards de y ou de z et des retards des autres variables. En particulier pour la prévision une étape à l'avance, ces modèles peuvent être très utiles.

Prévision une étape à l'avance

Obtenir une prévision pour une période suivant la fin de l'échantillon est relativement direct en utilisant des modèles tels que (18.43). Comme d'habitude, soit n la taille de l'échantillon. La prévision de y_{n+1} est

$$\hat{f}_n = \hat{\delta}_0 + \hat{\alpha}_1 y_n + \hat{\gamma}_1 z_n, \qquad \textbf{[18.44]}$$

où on suppose que les paramètres ont été estimés par MCO. Nous utilisons un accent circonflexe sur f_n pour insister sur le fait que nous avons utilisé les paramètres du modèle de régression. (Si nous connaissions les paramètres, il n'y aurait pas d'erreur d'estimation dans la prévision.) L'erreur de prévision que nous ne connaitrons pas avant $n + 1$ – est

$$\hat{e}_{n+1} = y_{n+1} - \hat{f}_n. \qquad \textbf{[18.45]}$$

Si nous ajoutons plus de retards de y ou de z à l'équation de prévision, nous perdons tout simplement plus d'observations au début de l'échantillon.

La prévision $\hat{f}_n$ de y_{n+1} est habituellement appelée une **prévision ponctuelle**. Nous pouvons aussi obtenir un intervalle de prévision. Un intervalle de prévision est essentiellement la même chose qu'un intervalle de fluctuation que nous avons étudié dans la section 6.4. À ce moment-là, nous avons montré comment, sous les hypothèses du modèle classique linéaire, obtenir un intervalle de prédiction exact à 95 %. Un intervalle de prevision est obtenu exactement de la même manière. Si le modèle ne satisfait pas les hypothèses du modèle classique linéaire – par exemple s'il contient des variables dépendantes retardées comme dans (18.44), l'intervalle de prévision est toujours valable approximativement moyennant le fait que u_t, étant donné I_{t-1}, est approximativement normalement distribué avec moyenne nulle et variance constante. (Cela garantit que les estimateurs MCO sont distribués approximativement normalement avec les variances habituelles MCO et que u_{n+1} est indépendent des estimateurs MCO avec moyenne nulle et variance σ^2.)

Soit $\mathrm{se}(\hat{f}_n)$ l'erreur standard de prévision et soit $\hat{\sigma}$ l'erreur standard de régression. [Depuis la section 6.4, on peut obtenir $\hat{f}_n$ et $\mathrm{se}(\hat{f}_n)$ comme l'intercept et son erreur standard à partir de la régression de y_t sur $(y_{t-1} - y_n)$ et $(z_{t-1} - z_n)$, $t = 1, 2, \ldots, n$; en clair, on retranche la valeur de y indicée par n de chaque y retardé, et de même pour z, avant de mener la régression.] Alors,

$$\mathrm{se}(\hat{e}_{n+1}) = \{[\mathrm{se}(\hat{f}_n)]^2 + \hat{\sigma}^2\}^{1/2}, \qquad \textbf{[18.46]}$$

et l'intervalle de prévision (approximatif) à 95 % est

$$\hat{f}_n \pm 1{,}96 \cdot \text{se}(\hat{e}_{n+1}). \qquad \textbf{[18.47]}$$

Parce que $\text{se}(\hat{f}_n)$ est à peu près proportionnel à $1/\sqrt{n}$, $\text{se}(\hat{f}_n)$ est habituellement petit par rapport à l'incertude relative à l'erreur u_{n+1}, comme mesuré par $\hat{\sigma}$. [Certains logiciels économétriques calculent les intervalles de prévision de manière habituelle, mais d'autres exigent des manipulations simples pour obtenir (18.47).]

EXEMPLE 18.8
Prévision du taux de chômage américain

Nous utilisons les données dans PHILLIPS, mais seulement pour les années allant de 1948 à 1996, pour prévoir le taux de chômage civil aux États-Unis pour 1997. On utilise deux modèles. Le premier est un modèle simple AR(1) pour *unem* :

$$\widehat{unem}_t = 1{,}572 + 0{,}732\ unem_{t-1}$$
$$(0{,}577)\ \ (0{,}097) \qquad \textbf{[18.48]}$$
$$n = 48,\ \bar{R}^2 = 0{,}544,\ \hat{\sigma} = 1{,}049.$$

Dans le second modèle, on ajoute l'inflation avec un retard d'une année :

$$\widehat{unem}_t = 1{,}304 + 0{,}647\ unem_{t-1} + 0{,}184\ inf_{t-1}$$
$$(0{,}490)\ \ (0{,}084) \qquad\quad (0{,}041) \qquad \textbf{[18.49]}$$
$$n = 48,\ \bar{R}^2 = 0{,}677,\ \hat{\sigma} = 0{,}883.$$

Le taux d'inflation retardé est très significatif dans (18,49) ($t \approx 4{,}5$), et le R-carré ajusté à partir de la seconde équation est beaucoup plus élevé que dans la première. Par contre, cela ne signifie pas nécessairement que la seconde équation produira une meilleure prévision pour 1997. Tout ce qu'on peut dire jusqu'à présent est que, en utilisant les données jusqu'en 1996, un retard de l'inflation aide à expliquer la variation du taux de chômage.

Pour obtenir des prévisions pour 1997, nous devons connaître *unem* et *inf* en 1996. Ils s'élèvent respectivement à 5,4 et 3,0. Dès lors, la prévision de $unem_{1997}$ à partir de l'équation (18.48) est 1,572 + 0,732(5,4), ou environ 5,52. La prévision à partir de l'équation (18.49) est 1,304 + 0,647(5,4) + 0,184(3,0), ou environ 5,35. Le taux de chômage civil effectif pour 1997 a été de 4,9, si bien que les deux équations surprédisent le taux effectif. La seconde équation donne une prévision légèrement meilleure.

On peut obtenir facilement un intervalle de prévision à 95 %. Quand on régresse $unem_t$ sur $(unem_{t-1} - 5{,}4)$ et $(inf_{t-1} - 3{,}0)$, on obtient 5,35 comme intercept – qu'on a déjà calculé comme prévision – et se $(\hat{f}_n) = 0{,}137$. Dès lors, comme $\hat{\sigma} = 0{,}883$, on a $\text{se}(\hat{e}_{n+1}) = [(0{,}137)^2 + (0{,}883)^2]^{1/2} \approx 0{,}894$. L'intervalle de prévision à 95 % à partir de (18.47) est 5,35 ± 1,96 (0,894), ou à peu près [3,6, 7,1]. C'est un intervalle large, et la valeur réalisée pour 1997 (4,9) est bien comprise dans cet intervalle. Comme attendu, l'écart-type de u_{n+1}, qui égale 0,883, est une fraction élevée de $\text{se}(\hat{e}_{n+1})$.

Un prévisionniste professionnel doit habituellement produire une prévision pour chaque période de temps. Par exemple, au temps n, il ou elle produit une prévision de y_{n+1}. Alors quand y_{n+1} et z_{n+1} deviennent disponibles, il ou elle doit prédire y_{n+2}. Même si le prévisionniste s'appuie sur son modèle (18.43), il a deux choix pour prédire y_{n+2}. Le premier est d'utiliser $\hat{\delta}_0 + y_{n+1} + \hat{\gamma}_1 z_{n+1}$, où les paramètres ont été estimés en utilisant les n premières observations. La seconde possibilité est de réestimer les paramètres en utilisant toutes les $n + 1$ observations et d'utiliser alors la même formule pour prédire y_{n+2}. Pour prédire pour les périodes de temps suivantes, on peut généralement utiliser les valeurs estimées

des paramètres obtenues à partir des n observations initiales, ou nous pouvons mettre à jour les paramètres de régression chaque fois que nous obtenons une nouvelle observation. Bien que cette deuxième approche exige plus de calcul, la charge supplémentaire est relativement modeste et cela peut (ou ne peut pas) mieux fonctionner parce que les coefficients de régression s'ajustent quelque peu, au moins aux nouvelles données.

Comme exemple spécifique, supposons que nous voulions prédire le taux de chômage pour 1998 en utilisant le modèle avec un seul retard de l'année de *unem* and *inf*. La première possibilité est d'insérer juste les valeurs du chômage et l'inflation de 1997 dans le membre de droite de (18.49). Avec *unem* = 4,9 et *inf* = 2,3 en 1997, nous avons une prévision pour $unem_{1998}$ d'à peu près 4,9 (le fait d'obtenir la même valeur que le taux de chômage pour 1997 est juste une coïncidence). La seconde possibilité est de réestimer l'équation en ajoutant l'observation relative à 1997 et ensuite d'utiliser cette nouvelle équation (voir exercice d'ordinateur C6).

Le modèle dans l'équation (18.43) est une équation de ce qu'on appelle un **modèle autorégressif vectoriel** (**VAR**). On sait ce qu'est un modèle autoregressif vectoriel depuis le chapitre 11 : on modélise des séries uniques, $\{y_t\}$, en fonction de leur propre passé. Dans des modèles autorégressifs vectoriels, on modélise plusieurs séries – qui, si vous êtes familier avec l'algèbre linéaire, sont à l'origine du mot vecteur – en termes de leur propre passé. Si nous avons deux séries, y_t and z_t, une autorégression vectorielle consiste en des équations qui ont la forme

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + \alpha_2 y_{t-2} + \gamma_2 z_{t-2} + \ldots \qquad \textbf{[18.50]}$$

et

$$z_t = \eta_0 + \beta_1 y_{t-1} + \rho_1 z_{t-1} + \beta_2 y_{t-2} + \rho_2 z_{t-2} + \ldots,$$

Chaque équation contient une erreur qui a une valeur attendue égale à zéro étant donné son information passée concernant y et z.

Dans l'équation (18.43) – et dans l'exemple estimé dans (18.49) – on a supposé qu'un retard de chaque variable capturait toute la dynamique. (Un test F joint de significativité de $unem_{t-2}$ and inf_{t-2} confirme que seulement un retard de chacune des variables est nécessaire).

Comme l'exemple 18.8 l'illustre, les modèles VAR peuvent être utiles pour la prévision. Dans beaucoup de cas, nous sommes intéressés par la prévision d'une seule variable, y, auquel cas nous devons estimer et analyser seulement l'équation relative à y. Rien ne nous empêche d'ajouter d'autres variables retardées, disons, w_{t-1}, w_{t-2}, ..., à l'équation (18.50). De telles équations sont estimées de manière efficiente par MCO, à condition que nous incluions assez de retards de toutes les variables et que l'équation satisfasse l'hypothèse d'homoscédasticité pour les régressions de séries temporelles.

Des équations telles que (18.50) nous permettent de tester si, après avoir contrôlé pour l'effet des y passés, les z passés aident à prévoir y_t. Généralement, on dit que *z cause y au sens de Granger* si

$$\mathrm{E}(y_t|I_{t-1}) \neq \mathrm{E}(y_t|J_{t-1}), \qquad \textbf{[18.51]}$$

où I_{t-1} contient l'information passée sur y *et* z, et J_{t-1} contient seulement l'information sur les y passés. Quand (18.51) est valide, les z passés sont utiles, en plus des y passés, pour prédire y_t. Le terme « cause » dans « cause au sens de Granger » doit être interprété avec prudence. Le seul sens dans lequel z « cause » y est donné par (18.51). En particulier, cela n'a rien à voir avec la causalité contemporaine entre y and z, si bien que cela ne nous permet pas de déterminer si z_t est une variable exogène ou endogène dans une équation liant y_t à z_t. (C'est aussi pourquoi la notion de **causalité à la Granger** ne s'applique pas dans des contextes purs de coupes transversales.)

Une fois qu'on suppose un modèle linéaire et qu'on décide combien de retards de y on devrait inclure dans $E(y_t|y_{t-1}, y_{t-2}, \ldots)$, on peut tester aisément l'hypothèse nulle que z ne cause pas y au sens de Granger. Pour être plus spécifique, supposons que $E(y_t|y_{t-1}, y_{t-2}, \ldots)$ dépende seulement de 3 retards :

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + u_t$$
$$E(u_t|y_{t-1}, y_{t-2}, \ldots) = 0.$$

Maintenant, sous l'hypothèse nulle que z ne cause pas y au sens de Granger, tout retard qui peut être ajouté à l'équation doit avoir un coefficient dans la population égal à zéro. Si l'on ajoute z_{t-1}, alors nous pouvons simplement faire un test t sur z_{t-1}. Si nous ajoutons deux retards de z, alors nous pouvons faire un test F de significativité jointe de z_{t-1} et z_{t-2} dans l'équation

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + \gamma_1 z_{t-1} + \gamma_2 z_{t-2} + u_t$$

(S'il y a hétéroscédasticité, nous pouvons utiliser une forme robuste du test. Il ne peut y avoir de corrélation sérielle sous H_0 puisque le modèle est complet dynamiquement.)

Concrètement, comment décider combien de retards de y et z inclure ? Tout d'abord nous commençons par estimer un modèle autorégressif pour y et nous menons des tests en t et en F pour déterminer combien de retards de y doivent apparaître. Avec des données annuelles, le nombre de retards est typiquement modeste, disons un ou deux. Avec des données trimestrielles ou mensuelles, il y a habituellement beaucoup plus de retards. Une fois que le modèle autoregressif pour y a été choisi, on peut tester les retards de z. Le choix des retards de z est moins important parce que lorsque z ne cause pas au sens de Granger y, aucun ensemble de retards de z ne doit être significatif. Avec des données annuelles, un ou deux retards sont habituellement utilisés ; avec des données trimestrielles, habituellement 4 ou 8 ; et avec des données mensuelles 6, 12 ou peut-être même 24 s'il y a assez de données.

Nous avons déjà couvert un exemple de test de causalité à la Granger dans l'équation (18.49). Le modèle autorégressif qui s'ajuste le mieux au chômage est AR(1). Dans l'équation (18.49), nous avons ajouté un seul retard de l'inflation et cela s'est avéré très significatif. Dès lors, l'inflation cause, au sens de Granger, le chômage.

Il y a une définition étendue de la causalité à la Granger qui est souvent utile. Soit $\{w_t\}$ une troisième série (ou qui représente plusieurs séries supplémentaires). Alors, *z cause y conditionnellement à w au sens de Granger* si (18.51) est valide, mais maintenant I_{t-1} contient l'information passée sur y, z, et w, tandis que J_{t-1} contient l'information passée sur y et w. Il est certainement possible que z cause au sens de Granger y, mais z ne cause pas y au sens de Granger conditionnellement à w. Un test de l'hypothèse nulle selon laquelle z ne cause pas y conditionnellement à w au sens de Granger s'obtient en testant la significativité de z dans un modèle de y qui dépend aussi des y retardés et des w retardés. Par exemple pour tester si la croissance dans l'offre de monnaie cause au sens de Granger la croissance du PIB réel conditionnellement à la variation des taux d'intérêt, nous régresserions $gGDP_t$ sur les retards de $gGDP$, Δ_{int}, and gM *et nous* mènerions des tests de significativité sur les retards de gM. [Voir, par exemple, Stock and Watson (1989).]

Comparaison des prévisions une étape à l'avance

Dans la plupart des problèmes de prévisions, il y a plusieurs méthodes en concurrence pour la prévision. Même lorsque l'on restreint son attention au modèle de régression, on a beaucoup de possibilités. Quelles variables inclure et avec combien de retards ? Devons-nous utiliser des logs, les niveaux des variables ou les premières différences ?

Afin de décider d'une méthode de prévision, nous avons besoin d'une manière de choisir la méthode la plus appropriée. En résumé, on peut distinguer entre des critères intra-échantillon et des critères hors

échantillon. Dans un contexte de régression, les critères intra-échantillon incluent le R-carré et en particulier le R-carré ajusté. Il y a beaucoup d'autres statistiques de sélection de modèles mais nous ne couvrirons pas celles-ci ici [voir par exemple Ramanathan (1995, Chapter 4)].

Pour la prévision, il est préférable d'utiliser des critères hors échantillon parce que la prévision est essentiellement un problème hors échantillon. Un modèle peut donner un bon ajustement par rapport à y dans l'échantillon utilisé pour estimer les paramètres. Mais cela peut ne pas se concrétiser en une bonne performance de prévision. Une comparaison hors échantillon consiste à utiliser la première partie de l'échantillon pour estimer les paramètres du modèle et de garder la dernière partie de l'échantillon pour évaluer ses capacités de prévision. C'est ce que nous aurions dû faire si nous n'avions pas su déjà les valeurs futures des variables.

Supposons que nous ayons $n + m$ observations, où nous utilisons les n premières observations pour estimer les paramètres de notre modèle et que nous gardions les dernières m observations pour la prévision. Soit $\hat{f}_{n+h}$ la prévision une étape à l'avance de y_{n+h+1} pour $h = 0, 1, \ldots, m-1$. Les m erreurs de prévision sont $\hat{e}_{n+h+1} = y_{n+h+1} - \hat{f}_{n+h}$. Comment devrions-nous mesurer dans quelle mesure notre modèle prédit bien y en dehors de l'échantillon ? Deux mesures sont en usage. La première est **la racine de l'erreur moyenne au carré (RMSE)** :

$$RMSE = \left(m^{-1} \sum_{h=0}^{m-1} \widehat{e^2_{n+h+1}} \right)^{1/2} \qquad \textbf{[18.52]}$$

C'est essentiellement l'écart-type d'échantillon des erreurs de prévision (sans aucun ajustement pour les degrés de liberté). Si nous calculons la RMSE pour deux ou plusieurs méthodes en concurrence, alors nous privilégierons la méthode avec la plus petite RMSE hors échantillon.

Une seconde mesure populaire est l'**erreur absolue moyenne (MAE)** qui est la moyenne des erreurs absolues de prévision :

$$MAE = m^{-1} \sum_{h=0}^{m-1} |\widehat{e_{n+h+1}}| \qquad \textbf{[18.53]}$$

Une fois de plus nous préférons une MAE plus petite. D'autres critères possibles incluent la minimisation des valeurs absolues des erreurs de prévision.

EXEMPLE 18.9
Comparaison hors échantillon des prévisions de chômage

Dans l'Exemple 18.8, nous avons trouvé que l'équation (18.49) s'ajuste bien mieux sur les années allant de 1948 à 1996 que l'équation (18.48), et au moins en ce qui concerne la prévision du chômage en 1997, le modèle qui inclut l'inflation retardée fonctionne mieux. Maintenant, nous utilisons les deux modèles, toujours estimés sur les données allant jusqu'en 1996, pour comparer les prévisions une étape à l'avance entre 1997 et 2003. Cela laisse sept observations hors échantillon ($n = 48$ et $m = 7$) à utiliser dans les équations (18.52) et (18.53). Pour le modèle AR(1), RMSE = 0,962 et MAE = 0,778. Pour le modèle qui ajoute l'inflation retardée (un modèle VAR d'ordre un), RMSE = 0,673 et MAE = 0,628. Donc, en vertu des deux mesures, le modèle qui inclut inf_{t-1} produit les meilleures prévisions hors échantillon de 1997 à 2003. Dans ce cas, les critères intra-échantillon et hors échantillon choisissent le même modèle.

Plutôt que d'utiliser seulement les n premières observations pour estimer les paramètres du modèle, on peut réestimer les modèles chaque fois que nous ajoutons une nouvelle observation et utilisons le nouveau modèle pour prévoir la période suivante.

Prévisions plusieurs étapes en avant

Prévoir plusieurs périodes à l'avance est généralement plus difficile que prévoir une période à l'avance. Nous pouvons formaliser ceci de la manière suivante. Supposons que nous considérions la prévision de y_{t+1} au temps t et à une période en avant (si bien que $s < t$). Alors $\text{Var}[y_{t+1} - \text{E}(y_{t+1}|I_t)] \leq \text{Var}[y_{t+1} - \text{E}(y_{t+1}|I_s)]$, où l'inégalité est habituellement stricte. Nous ne prouverons pas ce résultat de manière générale mais intuitivement, il tient la route : la variance de l'erreur de prévision en prédisant y_{t+1} est plus grande lorsque nous basons cette prévision sur moins d'information.

Si $\{y_t\}$ suit un modèle AR(1) (qui inclut une marche aléatoire, potentiellement avec dérive), on peut montrer aisément que la variance de l'erreur augmente avec l'horizon de prévision. Le modèle est

$$y_t = \alpha + \rho y_{t-1} + u_t$$

$$\text{E}(u_t|I_{t-1}) = 0, I_{t-1} + \{y_{t-1}, y_{t-2}, \ldots\},$$

et $\{u_t\}$ a une variance contante σ^2 conditionelle à I_{t-1}. Au temps $t + h - 1$, notre prévision de y_{t+h} est $\alpha + \rho y_{t+h-1}$ et l'erreur de prévision est simplement u_{t+h}. Dès lors, la variance de la prévision une étape à l'avance est simplement σ^2. Pour trouver les prévisions plusieurs étapes à l'avance, on obtient par substitutions répétées,

$$y_{t+h} = (1 + \rho + \ldots + \rho^{h-1})\alpha + \rho^h y_t + \rho^{h-1} u_{t+1} + \rho^{h-2} u_{t+2} + \ldots + u_{t+h}.$$

Au temps t, la valeur attendue de u_{t+j}, pour tout $j \geq 1$, est zéro. Donc

$$\text{E}(y_{t+h}|I_t) = (1 + \rho + \ldots + \rho^{h-1})\ \alpha + \rho^h y_t, \qquad \textbf{[18.54]}$$

et l'erreur de prévision est $e_{t,h} = \rho^{h-1} u_{t+1} + \rho^{h-2} u_{t+2} + \ldots + u_{t+h}$. C'est une somme de variables aléatoires non corrélées, et donc la variance de la somme est la somme des variances :

$\text{Var}(e_{t,h}) = \sigma^2\ [\rho^{2(h-1)} + \rho^{2(h-2)} + \ldots + \rho^2 + 1]$. Dans la mesure où $\rho^2 > 0$, chaque terme multipliant σ^2 est positif, si bien que la variance de l'erreur de prévision est h. Quand $\rho^2 < 1$ lorsque h devient grand, la variance de l'erreur de prévision converge vers $\sigma^2/(1 - \rho^2)$ qui est juste la variance non conditionnelle de y_t. Dans le cas d'une marche aléatoire ($\rho = 1$), $f_{t,h} = \alpha h + y_t$ et $\text{Var}(e_{t,h}) = \sigma^2 h$: la variance de prévision augmente sans borne avec l'augmentation de l'horizon h. Ceci montre qu'il est très difficile de prévoir une marche aléatoire, avec ou sans dérive à un horizon futur élevé. Par exemple, les prévisions de taux d'intérêt lointaines deviennent drastiquement moins précises.

L'équation (18.54) montre qu'utiliser le modèle AR(1) pour de la prévision multi-étapes est aisé, une fois que ρ a été estimé par OLS. La prévision de y_{n+h} au temps n est

$$\hat{f}_{n,h} = (1 + \hat{\rho} + \ldots + \hat{\rho}^{h-1})\ \hat{\alpha} + \hat{\rho}^h y_n \qquad \textbf{[18.55]}$$

Obtenir des intervalles de prévision est plus dur, à moins que $h = 1$, parce qu'obtenir l'écart-type de $\hat{f}_{n,h}$ est difficile. Néanmoins, l'écart-type de $\hat{f}_{n,h}$ est habituellement petit par rapport à l'écart-type du terme d'errreur, et ce dernier peut être estimé comme $\hat{\sigma}\,[\hat{\rho}^{2(h-1)} + \hat{\rho}^{2(h-1)} + \hat{\rho}^{2(h-2)} + \ldots + \hat{\rho}^2 + 1]^{1/2}$, où $\hat{\sigma}$ est l'erreur standard de régression à partir de la régression AR(1). Nous pouvons utiliser cela pour obtenir un intervalle de confiance approximatif. Par exemple, quand $h = 2$, un intervalle de confiance à 95 % (pour n grand) est

$$\hat{f}_{n,2} \pm 1.96\ \hat{\sigma}(1 + \hat{\rho}^2)^{1/2}. \qquad \textbf{[18.56]}$$

Parce que nous sous-estimons l'écart-type de y_{n+h}, cet intervalle est trop étroit, mais pas tant que cela dans le cas où n est grand.

Une approche moins traditionnelle mais utile est d'estimer un modèle différent pour chaque horizon de prévision. Par exemple, supposons que l'on veuille prévoir y 2 périodes à l'avance. Si I_t dépend seulement de y à travers le temps on peut supposer que $\text{E}(y_{t+2}|I_t) = \alpha_0 + \gamma_1 y_t$ [ce qui comme nous l'avons vu avant est

valide si $\{y_t\}$ suit un modèle AR(1)]. Nous pouvons estimer α_0 et γ_1 en régressant y_t sur un intercept et sur y_{t-2}. Même si les erreurs de cette équation contiennent de la correlation sérielle – les erreurs dans des périodes contiguës sont corrélées – on peut obtenir des estimateurs convergents et approximativement normaux de α_0 et de y_1. La prévision de y_{n+2} au temps n est simplement $\hat{f}_{n,2} = \hat{\alpha}_0 + \hat{\gamma}_1 y_n$. En outre et de manière relativement importante, l'erreur standard de régression est justement ce dont nous avons besoin pour calculer l'intervalle confiance de la prévision. Malheureusement pour obtenir l'écart-type de $\hat{f}_{n,2}$, en utilisant la procédure de la prévision une étape à l'avance exige que nous obtenions l'écart-type robuste à la corrélation sérielle comme décrit dans la section 12.5. Cet écart-type tend vers zéro lorsque n devient grand tandis que la variance de l'erreur est constante. Dès lors, on peut obtenir un intervalle approximatif en utilisant (18.56) et en insérant l'ESR à partir de la régression y_t sur y_{t-2} en lieu et place de $\hat{\sigma}(1 + \hat{\rho})^{1/2}$. Mais nous devons nous souvenir que ceci ignore l'erreur d'estimation de $\hat{\alpha}_0$ et $\hat{\gamma}_1$.

Nous pouvons aussi calculer des prévisions plusieurs étapes à l'avance avec des modèles autorégressifs plus compliqués. Par exemple, supposons que $\{y_t\}$ suive un modèle AR(2) et qu'au temps n, on veuille prévoir y_{n+2}. On a $y_{n+2} = \alpha + \rho_1 y_{n+1} + \rho_2 y_n + u_{n+2}$, et donc

$$\mathrm{E}(y_{n+2}|I_n) = \alpha + \rho_1 E(y_{n+1}|I_n) + \rho_2 y_n$$

On peut écrire cela comme

$$f_{n,2} = \alpha + \rho_1 f_{n,1} + \rho_2 y_n,$$

si bien que la prévision deux périodes à l'avance au temps n peut être effectuée une fois que celle une période à l'avance a été obtenue. Si les paramètres du modèle AR(2) ont été estimés par MCO, alors on met en pratique cela de la manière suivante :

$$\hat{f}_{n,2} = \hat{\alpha} + \hat{\rho}_1 \hat{f}_{n,1} + \hat{\rho}_2 y_n. \qquad \textbf{[18.57]}$$

Ainsi, $\hat{f}_{n,1} = \hat{\alpha} + y_n + \hat{\rho}_2 y_{n-1}$, que l'on calcule au temps n. Ensuite, on insère cela dans (18.57), avec y_n, pour obtenir $\hat{f}_{n,2}$. Pour tout $h > 2$, obtenir n'importe quelle prévision h étapes à l'avance à partir d'un modèle AR(2) est aisé et peut se calculer de manière récursive : $\hat{f}_{n,h} = \hat{\alpha} + \hat{\rho}_1 \hat{f}_{n,h-1} + \hat{\rho}_2 \hat{f}_{n,h-2}$

Un raisonnement similaire peut s'appliquer pour obtenir des prévisions plusieurs étapes à l'avance à partir de modèles VAR. Pour illustrer, supposons que nous ayons

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + u_t \qquad \textbf{[18.58]}$$

et

$$z_t = \eta_0 + \beta_1 y_{t-1} + \rho_1 z_{t-1} + v_t.$$

Si nous voulons prévoir y_{n+1} au temps n, on utilise simplement $\hat{f}_{n,1} = \hat{\delta}_0 + \hat{\alpha}_1 y_n + \hat{\gamma}_1 z_n$. De manière similaire, la prévision de z_{n+1} au temps n est $\hat{g}_{n,1} = \hat{\eta}_0 + \hat{\beta}_1 y_n + \hat{\rho}_1 z_n$. Supposons maintenant que nous voulions obtenir une prévision de y deux étapes à l'avance au temps n. À partir de (18.58), on a

$$\mathrm{E}(y_{n+2}|I_n) = \delta_0 + \alpha_1 \mathrm{E}(y_{n+1}|I_n) + \gamma_1 \mathrm{E}(z_{n+1}|I_n)$$

[vu que $\mathrm{E}(u_{n+2}|I_n) = 0$], et donc on peut écrire la prévision comme

$$\hat{f}_{n,2} = \hat{\delta}_0 + \hat{\alpha}_1 \hat{f}_{n,1} + \hat{\gamma}_1 \hat{g}_{n,1}. \qquad \textbf{[18.59]}$$

Cette équation montre que la prévision deux étapes à l'avance de y dépend de la prévision une étape à l'avance de y et z. Généralement on peut construire des prévisions plusieurs étapes à l'avance de y en utilisant la formule récursive

$$\hat{f}_{n,h} = \hat{\delta}_0 + \hat{\alpha}_1 \hat{f}_{n,h-1} + \hat{\gamma}_1 \hat{g}_{n,h-1}, \; h \geq 2.$$

EXEMPLE 18.10
Prévisions deux ans à l'avance du taux de chômage

Afin d'utiliser l'équation (18.49) pour prévoir le chômage deux ans à l'avance – disons en utilisant les données jusqu'en 1996 pour prévoir le taux en 1998 – nous avons besoin d'un modèle expliquant l'inflation. Le meilleur modèle pour *inf* en termes de *inf* et *unem* retardés semble être un modèle simple AR(1) (le chômage n'est pas significatif lorsque il est ajouté à la régression) :

$$\widehat{inf}_t = (1,277) + (0,665)inf_{t-1}$$
$$(0,558) \quad (0,107)$$
$$n = 48, R^2 = 0,457, \bar{R}^2 = 0,445.$$

Si nous insérons la valeur de 1996 de *inf* dans cette équation, on obtient la prévision de *inf* pour 1997 : $\widehat{inf}_{1997} = 3,27$. Ensuite on peut insérer cela, avec $\widehat{unem}_{1997} = 5,35$ (que nous avons obtenu précédemment), dans (18.59) pour prévoir $unem_{1998}$:

$$\widehat{unem}_{1998} = 1,304 + 0,647(5,35) + 0,184(3,27) \approx 5,37.$$

Souvenons-nous, cette prévision n'utilise l'information que jusqu'en 1996. La prévision une période à l'avance de $unem_{1998}$, obtenue en insérant dans (18.48) les valeurs de 1997 pour *unem* et *inf* était environ de 4,90. Le taux de chômage effectif en 1998 était 4,5 %, ce qui signifie que dans ce cas, la prévision une période à l'avance fait un peu mieux que la prévision deux périodes à l'avance.

Comme pour la prévision une période à l'avance, une racine de l'erreur moyenne au carré hors échantillon ou une erreur moyenne absolue hors échantillon peuvent être utilisées pour choisir entre les méthodes de prévisions plusieurs périodes à l'avance.

Prévoir les Processus avec tendance, saisonnalité et processus intégrés

Nous pouvons maintenant traiter les prévisions des séries qui ont des tendances, qui ont de la saisonnalité ou qui possèdent une racine unitaire. Rappelons que dans les chapitres 10 et 11 on a vu que l'approche permettant de traiter les variables dépendantes ou indépendantes avec tendance dans les modèles de régression est d'inclure des tendances temporelles, la plus populaire d'entre elles étant la tendance linéaire. Les tendances peuvent aussi être inclues dans les équations de prévision bien qu'elles doivent être traitées avec prudence.

Dans le cas le plus simple, supposons que $\{y_t\}$ possède une tendance linéaire mais est imprévisible autour de cette tendance. On peut alors écrire

$$y_t = \alpha + \beta t + u_t, \ \mathrm{E}(u_t|I_{t-1}) = 0, \ t = 1, 2, \ldots, \qquad \textbf{[18.60]}$$

où comme d'habitude, I_{t-1} contient l'information observée jusqu'en $t - 1$ (qui inclut au moins le passé de y). Comment prévoir y_{n+h} au temps n pour tout $h \geq 1$? C'est simple parce que $\mathrm{E}(y_{n+h}|I_n) = \alpha + \beta(n + h)$. La variance de l'erreur de prévision est simplement $\sigma^2 = \mathrm{Var}(u_t)$ (en supposant une variance constante dans le temps). Si nous estimons α et β par MCO en utilisant les n premières observations, alors notre prévision de y_{n+h} au temps n est $\hat{f}_{n,h} = \hat{\alpha}_0 + \hat{\beta}(n + h)$. En d'autres termes, on insère simplement la période de temps correspondant à y dans la fonction estimée de la tendance. Par exemple, si nous utilisons les $n = 131$ observations dans BARIUM pour prédire les importations mensuelles depuis la Chine vers les États-Unis de chlorure de barium chinois, nous obtenons $\hat{\alpha} = 249,56$ et $\hat{\beta} = 5,15$. La période d'échantillon finit en décembre 1988 si bien que la prévision des importations de chlorure de barium chinois six mois plus tard est

249,56 + 5,15(137) = 955,11, mesurées en tonnes. En comparaison la valeur de décembre 1988 est 1 087,81, si bien qu'elle est plus grande que la valeur prédite six mois plus tard. La série et sa droite de tendance estimée sont reproduites dans la figure 18.2.

Comme discuté dans le chapitre 10, la plupart des séries économiques sont caractérisées au moins approximativement par un taux de croissance constant, ce qui suggère que $\log(y_t)$ suit une tendance linéaire temporelle. Supposons que nous utilisions les n observations pour obtenir l'équation

$$\log(y_t) = \hat{\alpha} + \hat{\beta}t, \; t = 1, 2, \ldots, n. \qquad \textbf{[18.61]}$$

Alors, pour prédire $\log(y)$ pour toute période future $n + h$, nous insérons simplement $n + h$ dans l'équation de la tendance, comme précédemment. Mais cela ne nous permet pas de prédire y, qui est ce que nous cherchons la plupart du temps. Il est tentant de prendre l'exposant de $\hat{\alpha} + \hat{\beta}(n + h)$ pour obtenir une prévision de y_{n+h}, mais c'est faux, pour les mêmes raisons exposées dans la Section 6.4. On doit tenir compte explicitement de l'erreur implicite dans (18.61). La manière la plus simple est d'utiliser les n observations pour régresser y_t sur $\exp(\widehat{\log(y_t)})$ *sans* intercept. Soit $\hat{\gamma}$ le coefficient de pente de $\exp(\widehat{\log(y_t)})$. Alors la prévision de y en période $n + h$ est simplement

$$\hat{f}_{n,h} = \hat{\gamma} \exp[\hat{\alpha} + \hat{\beta}(n + h)] \qquad \textbf{[18.62]}$$

Investigation supplémentaire 18.5

Supposons qu'on modélise $\{y_t : t = 1, 2, \ldots, 46\}$ comme une tendance linéaire temporelle, avec des données débutant en 1950 et finissant en 1995. Définissons la variable $year_t$ qui s'observe de 1950 pour $t = 1$ à 1995 pour $t = 46$. Si on estime l'équation $\hat{y}_t = \hat{\gamma} + \hat{\delta} year_t$, comment est-ce que $\hat{\gamma}$ et $\hat{\delta}$ se comparent avec $\hat{\alpha}$ et $\hat{\beta}$ dans $\hat{y}_t = \hat{\alpha} + \hat{\beta}t$? Comment est-ce que les prévisions de ces deux équations se comparent entre elles ?

Comme exemple, si nous utilisons les 687 premières semaines des données relatives à l'indice du New York Stock Exchange dans NYSE, nous obtenons $\hat{\alpha} = 3{,}782$ et $\hat{\beta} = 0{,}0019$ [en régressant $\log(price_t)$ sur une tendance linéaire temporelle] ; cela suggère que l'indice croît d'à peu près 0,2 % par semaine en moyenne. Quand on régresse *price* sur les valeurs ajustées et prises en exposant, nous obtenons $Y = 1{,}018$. Maintenant, nous prédisons quatre semaines à l'avance, qui est la dernière semaine dans l'échantillon, en utilisant (18.62) : 1,018 exp [3,782 + 0,0019 (691)] ≈166,12. La valeur effective s'avère être 164,25, si bien que nous avons quelque peu surprédit. Mais ce résultat est bien meilleur que si l'on estime une tendance linéaire temporelle pour les 687 premières semaines : la valeur prédite pour la semaine 691 est 152,23, ce qui est une sous-prédiction considérable.

Bien que les modèles de tendance puissent être utiles pour la prévision, ils doivent être utilisés avec prudence, en particulier pour prédire loin dans le futur les séries intégrées qui possèdent une dérive. Ce problème éventuel peut être appréhendé en considérant une marche aléatoire avec dérive. Au temps $t + h$, on peut écrire y_{t+h} comme

$$y_{t+h} = \beta h + y_t + u_{t+1} + \ldots + u_{t+h},$$

où β est la dérive (habituellement $\beta > 0$), et pour lequel chaque u_{t+j} a une moyenne zéro étant donné I_t et une variance contsante σ^2. Comme nous l'avons vu auparavant, la prévision de y_{t+h} au temps t est $E(y_{t+h}|I_t) = \beta h + y_t$ et la variance de l'erreur de prévision est $\sigma^2 h$.

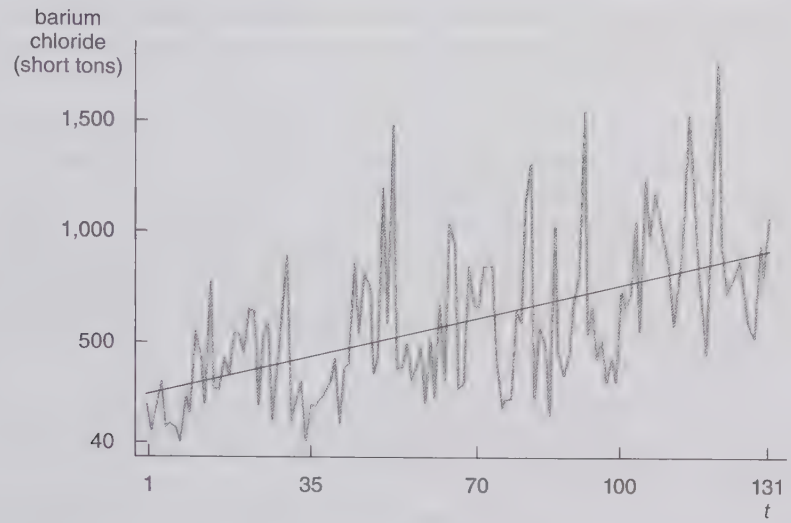


Figure 18.2 *Importations des États-Unis de chlorure de barium chinois (en tonnes) et sa droite de tendance estimée, 249,56 + 5,15t.*

Qu'arrive-t-il si on utilise un modèle avec tendance linéaire ? Soit y_0 la valeur initiale du processus au temps zéro, que nous considérons comme non aléatoire. Alors on peut écrire

$$\begin{aligned} y_{t+h} &= y_0 + \beta\,(t+h) + u_1 + u_2 + \ldots + u_{t+h} \\ &= y_0 + \beta\,(t+h) + v_{t+h}. \end{aligned}$$

Cela ressemble à une tendance linéaire mais avec un intercept $\alpha = y_0$. Néanmoins, l'erreur a une variance $\sigma^2(t+h)$ avec une moyenne égale à zéro. Dès lors, si on utilise une tendance linéaire $y_0 + \beta(t+h)$ pour prédire y_{t+h} en t, la variance de l'erreur de prévision est $\sigma^2(t+h)$, comparée à $\sigma^2 h$ quand on utilise $\beta h + y_t$. Le ratio des variances de prévision est $(t+h)/h$, qui peut être important lorsque t est élevé. Le point important est que nous ne devons pas utiliser une tendance linéaire pour prédire une marche aléatoire avec dérive. (l'Exercice d'ordinateur C8 vous demande de comparer les prévisions à partir d'une tendance cubique et celles d'une simple marche aléatoire pour le taux de fécondité des États-Unis).

Les tendances déterministes peuvent aussi produire des prévisions problématiques si les paramètres de tendance sont estimés en utilisant de vieilles données et que le processus est affecté d'un changement important dans la droite de tendance. Parfois, des chocs exogènes – tels que les crises pétrolières des années 70 – peuvent changer la trajectoire des variables sujettes à la tendance. Si une droite tendancielle ancienne est utilisée pour prédire le futur, les prévisions peuvent diverger substantiellement. Le problème peut être atténué en utilisant les données les plus récentes disponibles pour obtenir les paramètres de la droite tendancielle.

Rien ne nous empêche de combiner les tendances avec d'autres modèles pour la prévision. Par exemple nous pouvons ajouter une tendance linéaire à un modèle AR(1), ce qui peut fonctionner correctement pour prédire des séries avec tendances mais qui sont aussi des processus AR stables autour de la tendance.

Il est assez immédiat de prédire des processus avec une saisonnalité déterministe (séries mensuelles ou trimestrielles). Par exemple le fichier BARIUM contient la production mensuelle d'essence aux États-Unis entre 1978 et 1988. La série n'a pas de tendance marquée mais est fortement sujette à un schéma de saisonnalité. (La production d'essence est plus forte les mois d'été et en décembre.) Dans le modèle le plus simple, nous régresserions *gas* (mesuré en gallons) sur 11 variables muettes mensuelles, disons de février à décembre. Dans ce cas, la prévision pour n'importe quel mois futur est simplement la constante plus le

coefficient de la variable muette appropriée. (Pour janvier, la prévision est juste l'intercept de la régression.) On peut aussi ajouter des retards des variables et des tendances temporelles pour traiter des séries générales avec saisonnalité.

Prédire des processus avec racine unitaire exige également un traitement spécial. Auparavant, nous avons obtenu la valeur attendue d'une marche aléatoire conditionnelle à l'information jusqu'au temps n. Pour prédire une marche aléatoire avec dérive possible α, h périodes de temps dans le futur au temps n, on utilise $\hat{f}_{n,h} = \hat{\alpha}h + y_n$, où $\hat{\alpha}$ est la moyenne d'échantillon de Δy_t jusque $t = n$. (s'il n'y a pas de dérive, on impose = 0). Une alternative serait d'utiliser le modèle AR(1) pour $\{y_t\}$ et d'utiliser la formule de prévision (18.55). Cette approche n'impose pas la racine unitaire mais si une racine est présente, $\hat{\rho}$ converge en probabilité vers un lorsque n devient large. Néanmoins, $\hat{\rho}$ peut être substantiellement différent de un, notamment si la taille de l'échantillon n'est pas élevée.

Le fait que cette approche fournit de meilleures prévisions hors échantillon est une question empirique. Si dans le modèle AR(1), ρ est inférieur à un, même de peu, alors le modèle AR(1) tendra à produire des prévisions meilleures à long terme.

Généralement il existe deux approches pour produire des prévisions pour des processus I(1). La première est d'imposer une racine unitaire. Pour une prévision une période à l'avance, on obtient un modèle pour prédire la variation de y, Δy_{t+1} étant donné l'information jusqu'au temps t. Alors, puisque $y_{t+1} = \Delta y_{t+1} + y_t$, $E(y_{t+1} + I_t) + y_t$. Dès lors, notre prévision de y_{n+1} au temps n est juste

$$\hat{f}_n = \hat{g}_n + y_n,$$

où $\hat{g}_n$ est la prévision de Δy_{n+1} au temps n. Typiquement, un modèle AR(1) qui est nécessairement stable est utilisé pour Δy_t, ou une autorégression vectorielle est utilisée.

Ceci peut être étendu à des prévisions plusieurs étapes à l'avance en écrivant y_{n+h} comme

$$y_{n+h} = (y_{n+h} - y_{n+h-1)} + (y_{n+h-1} - y_{n+h-2}) + \ldots + (y_{n+1} - y_n) + y_n,$$

ou

$$y_{n+h} = \Delta y_{n+h} + \Delta y_{n+h-1} + \ldots + \Delta y_{n+1} + y_n$$

Dès lors, la prévision de y_{n+h} au temps n est

$$\hat{f}_{n,h} = \hat{g}_{n,h} + \hat{g}_{n,h-1} + \ldots + \hat{g}_{n,1} + y_n \qquad \textbf{[18.63]}$$

où $\hat{g}_{n,j}$ est la prévision de Δy_{n+j} au temps n. Par exemple, on peut modéliser Δy_t comme un AR(1) stable, obtenir des prévisions plusieurs étapes à l'avance à partir de (18.55) (mais avec $\hat{\alpha}$ et $\hat{\rho}$ obtenus à partir de Δy_t sur Δy_{t-1}, et y_n remplacé par Δy_n), et ensuite introduire cela dans (18.63).

La seconde approche pour prédire des variables I(1) est d'utiliser un processus AR général ou un modèle VAR pour $\{y_t\}$. Ceci n'impose pas une racine unitaire. Par exemple si on utilise un modèle AR(2),

$$y_t = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + u_t \qquad \textbf{[18.64]}$$

alors $\rho_1 + \rho_2 = 1$. Si on insère dans $\rho_1 = 1 - \rho_2$ et si on réarrange, on obtient $\Delta y_t = \alpha - \rho_2 \Delta y_{t-1} + u_t$, ce qui est un modèle AR(1) stable, à la difference que cela nous ramène à la première approche décrite plus haut. Rien ne nous empêche d'estimer (18.64) directement par MCO. Une chose intéressante à propos de cette régression est qu'on peut l'utiliser la statistique t relative à $\hat{\rho}_2$ pour déterminer si y_{t-2} est significatif. (Ceci suppose que l'hypothèse d'homoscédasticité est valide ; sinon, on peut utiliser la forme robuste à l'hétéroscédasticité.) Nous ne montrerons pas cela formellement, mais intuitivement il découle en réecrivant l'équation comme $y_t = \alpha + \gamma y_{t-1} - \rho_2 \Delta y_{t-1} + u_t$ où $\gamma = \rho_1 + \rho_2$. Même si $\gamma = 1$, ρ_2 est l'opposé du coefficient d'un processus $\{\Delta y_{t-1}\}$ stationnaire, faiblement dépendant. Comme les résultats de régression seront identiques à ceux de (18.64), on peut utiliser (18.64) directement.

Comme exemple, estimons un modèle AR(2) pour le taux de fécondité général dans FERTIL3, en utilisant les données jusqu'1979. (Dans l'Exercice d'ordinateur C8, on vous demande d'utiliser ce modèle pour la prévision, ce qui explique pourquoi on a laissé certaines observations à la fin de l'échantillon.)

$$\widehat{gfr}_t = 3{,}22 + 1{,}272\ gfr_{t-1} + 0{,}311\ gfr_{t-2} \quad \textbf{[18.65]}$$
$$(2{,}92)\ (0{,}120)\qquad (0{,}121)$$
$$n = 65,\ R^2 = 0{,}949,\ \bar{R}^2 = 0{,}947.$$

La statistique t du second retard est à peu près $-2{,}57$, ce qui est statistiquement différent de zéro à un niveau d'à peu près de 1 % (le premier retard a une statistique en t très significative, qui a une distribution approximative en t suivant le même raisonnement que pour $\hat{\rho}_2$). Le R-carré, ajusté ou non, n'est pas spécialement informatif en tant que mesure d'ajustement parce qu'apparemment, *gfr* contient une racine unitaire, et il n'y a donc pas de raison de se demander quelle part de la variance on explique.

Les coefficients des 2 retards dans (18.65) somment à 0,961, qui est proche et non statistiquement différent de un (comme on peut le vérifier en appliquant le test de Dickey-Fuller à l'équation $\Delta gfr_t = \alpha + \theta fr_{t-1} + \delta_1 \Delta gfr_{t-1} + u_t$). Même si on n'a pas imposé la restriction inhérente à la racine unitaire, on peut toujours utiliser (18.65) pour la prévision, comme discuté préalablement.

Avant de finir cette section, nous mettons en évidence l'amélioration potentielle dans la prévision des modèles autorégressifs avec variables I(1). Soient $\{y_t\}$ et $\{z_t\}$ deux processus I(1). Une approche pour obtenir des prévisions de y est d'estimer une autorégression bivariée portant sur les variables Δy_t et Δz_t et d'utiliser ensuite (18.63) pour générer des prévisions une ou plusieurs étapes à l'avance ; c'est essentiellement la première approche que nous avons décrite précédemment. Cependant si y_t et z_t sont cointégrés, nous avons plus de variables stationnaires, stables dans l'ensemble d'informations qui peuvent être utilisées pour prévoir Δy : en particulier, les retards de $y_t - \beta z_t$ où β est le paramètre de cointégration. Un simple modèle à correction d'erreur est

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_1 \Delta z_{t-1} + \delta_1 (y_{t-1} + \beta z_{t-1}) + e_t,$$
$$\mathrm{E}(e_t | I_{t-1}) = 0. \quad \textbf{[18.66]}$$

Pour prévoir y_{n+1}, on utilise les observations jusque n pour estimer le paramètre de cointégration β et ensuite pour estimer les paramètres du modèle à correction d'erreur par MCO, comme décrit dans la section 18.4. Prévoir Δy_{n+1} est simple : on insère juste Δy_n, Δz_n, et $y_n - \beta z_n$ dans l'équation estimée. Ayant obtenu la prévision de Δy_{n+1}, on l'ajoute à y_n.

En réarrangeant le modèle à correction d'erreur, on écrit

$$y_t = \alpha_0 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t, \quad \textbf{[18.67]}$$

où $\rho = 1 + \alpha_1 + \delta$, $\rho_2 = -\alpha_1$ et ainsi de suite, qui est la première équation du modèle-VAR pour y_t et z_t. Notez que ceci dépend de cinq paramètres, autant que dans le modèle à correction d'erreur. Le point est que pour la tâche de prévision, le modèle VAR en niveaux et le modèle à correction d'erreur sont essentiellement les mêmes. Ce n'est pas le cas dans des modèles à correction d'erreur plus généraux. Par exemple, supposons que $\alpha_1 = \gamma_1 = 0$ dans (18.66), mais que nous disposions d'un second terme à correction d'erreur, $\delta_2(y_{t-2} - \beta z_{t-2})$. Alors, le terme à correction d'erreur inclut seulement quatre paramètres alors que (18.67) – qui a le même ordre de retards pour y et z – contient cinq paramètres. Donc les modèles à correction d'erreur peuvent économiser des paramètres, ils sont plus parcimonieux que les VAR en niveaux.

Si y_t et z_t sont I(1) mais non cointégrés, le modèle approprié est (18.66) sans terme à correction d'erreurs. Ceci peut être utilisé pour prévoir Δy_{n+1} et nous pouvons ajouter cela à y_n pour prévoir y_{n+1}.

RÉSUMÉ

Les matières des séries temporelles couvertes dans ce chapitre sont habituellement utilisées en macroéconomie empirique, en finance empirique et dans d'autres domaines appliqués. Nous commencons par montrer comment les modèles à retards distribués infinis peuvent être interprétés et estimés. Ils peuvent fournir des distributions de retards flexibles avec moins de paramètres qu'un modèle à retards distribués finis similaire. Le retard distribué géométrique, et, plus généralement, les modèles à retards distribués rationnels sont les plus populaires. Ils peuvent être estimés avec des procédures économétriques standard ou avec des équations dynamiques simples.

Tester la présence d'une racine unitaire est devenu très courant en économétrie des séries temporelles. Si une série possède une racine unitaire, alors, dans de nombreux cas, les approximations normales usuelles en grand échantillon ne sont plus valables. En outre, un processus avec une racine unitaire possède la propriété qu'une innovation a un effet qui dure dans le temps, ce qui est intéressant en soi. Même s'il y a beaucoup de tests de racine unitaire, le test en t de Dickey-Fuller est probablement le plus populaire et le plus simple à mettre en oeuvre. Il peut accommoder une tendance linéaire lors du test de la racine unitaire, en ajoutant une tendance à la régression de Dickey-Fuller.

Quand une série I(1), y_t, est régressée sur une autre série, x_t, il y a un risque important de régression fallacieuse, même si les séries ne contiennent pas de tendances marquées. Cela a été étudié en profondeur dans le cas de la marche aléatoire : même si deux marches aléatoires sont indépendantes, le test en t classique de significativité du coefficient de pente, basé sur les valeurs critiques habituelles, rejettera beaucoup plus souvent que la valeur nominale du test ne l'indique. En outre, le R^2 tend vers une variable aléatoire plutôt que vers zéro (comme cela serait le cas si on régressait la différence de y_t sur la différence de x_t).

Quand les séries sont cointégrées, une régression impliquant des variables I(1) n'est pas fallacieuse. Cela signifie qu'une fonction linéaire de deux variables I(1) est I(0). Si y_t et x_t sont I(1) mais que $y_t - x_t$ est I(0), y_t et x_t ne peuvent s'écarter arbitrairement l'un de l'autre. Il existe des tests simples de l'hypothèse nulle de non cointégration contre l'alternative de cointégration, un de ceux-ci étant basé sur un test de racine unitaire à la Dickey-Fuller appliqué sur les résidus d'une régression statique. Il existe aussi des estimateurs simples du paramètre de cointégration qui donnent des statistiques en t avec des distributions approximativement standard normales (et des intervalles de confiance asymptotiquement valables). Nous avons couvert l'estimateur avec retards et valeurs avancées dans la section 18.4.

La cointégration entre y_t et x_t implique que les termes à correction d'erreurs peuvent intervenir dans un modèle reliant Δy_t à Δx_t ; les termes à correction d'erreurs sont des retards de $y_t - \beta x_t$, où β est le paramètre de cointégration. Une procédure d'estimation en deux étapes est valable pour estimer les modèles à correction d'erreur. D'abord, β est estimé en utilisant une régression statique (ou la régression avec retards et valeurs avancées). Ensuite les MCO sont utilisés pour estimer un modèle dynamique simple sur les premières différences qui inclut les termes à correction d'erreur.

La section 18.5 inclut une introduction à la prévision en insistant sur les méthodes de prévision basées sur la régression. Les modèles statiques, ou plus généralement, les modèles qui contiennent des variables explicatives contemporaines de la variable dépendante, présentent des limites car les variables explicatives doivent être prédites. Si l'on insère des valeurs hypothétiques dans les variables futures inconnues des variables explicatives, on obtient des prévisions conditionnelles. Les prévisions non conditionnelles sont identiques à une modélisation simple de y_t en fonction de l'information *passée* que l'on a observée au moment où la prévision est faite. Les modèles dynamiques de régression, y compris les autorégressions et les autorégressions vectorielles, sont couramment utilisés. En plus d'obtenir des prévisions ponctuelles une étape à l'avance, nous avons aussi discuté de la construction des intervalles de confiance des prévisions, qui ressemblent fort à des intervalles de prédiction.

Différents critères peuvent être utilisés pour choisir parmi les méthodes de prévision. Les mesures de performances les plus courantes sont la racine de l'erreur moyenne au carré et l'erreur absolue moyenne.

Les deux critères estiment la taille de l'erreur moyenne de prévision. C'est la plus informative pour calculer ces mesures en utilisant des prévisions hors échantillon

Les prévisions à plusieurs étapes à l'avance impliquent de nouveaux défis et sont sujettes à des variances d'erreurs de prévision plus grandes. Néanmoins, pour des modèles tels que les autorégressions ou autorégressions vectorielles, des prévisions plusieurs étapes à l'avance peuvent être calculées, et des intervalles de confiance approchés peuvent être obtenus.

Prévoir des séries I(1) avec tendance exige une attention spéciale. Les processus avec tendances deterministes peuvent être prédits en incluant des tendances déterministes dans les modèles de régression. Un désavantage potentiel est que les tendances déterministes peuvent générer des mauvaises prévisions à des horizons de prédiction lointains : une fois estimée, une tendance linéaire continue à croître ou à décroître. L'approche typique pour prédire un processus I(1) est de prédire sa différence et d'ajouter le niveau de la variable à sa différence prédite. Alternativement, les modèles autorégressifs vectoriels peuvent être utilisés sur les séries en niveau. Si les séries sont cointégrées, des modèles à correction d'erreur peuvent être plutôt utilisés.

MOTS-CLÉS

Causalité à la Granger p. 754
Cointégration p. 743
Critères hors Echantillon p. 755
Critères intra-Echantillon p. 755
Distribution de Dickey-Fuller p. 737
Ensemble d'information p. 750
Erreur Absolue Moyenne (MAE) p. 756
Erreur de prévision p. 750
Estimateur à retards et valeurs avancées p. 747
Fonction de perte p. 750
Intervalle de prévision p. 752
Lissage Exponentiel p. 751
Martingale p. 751
Modèle Autorégressif Vectoriel (VAR) p. 754
Modèle à retards distribués infinis p. 730
Modèle à retards distribués rationnels p. 734
Modèles à correction d'erreur p. 748
Prévision Conditionnelle p. 751
Prévision non conditionnelle p. 752
Prévision plusieurs étapes à l'avance p. 751
Prévision ponctuelle p. 752
Prévision une étape à l'avance p. 750
Problème de Régression Fallacieuse p. 742
Procédure en deux étapes d'Engle-Granger p. 749
Racine de l'erreur moyenne au carré (RMSE) p. 756
Racines unitaires p. 736
Retards Distribués Géométriques (ou de Koyck) p. 732
Séquence de Différences de Martingale p. 736
Test de Dickey-Fuller Augmenté p. 739
Test de Dickey-Fuller (DF) p. 737
Test d'Engle-Granger p. 745

PROBLÈMES

1. Considérez l'équation (18.15) avec $k = 2$. En utilisant une approche par VI pour estimer γ_h et ρ, qu'utiliseriez-vous comme instruments de y_{t-1} ?

2. Un modèle économique intéressant qui mène à un modèle économétrique avec une variable dépendante retardée relie y_t à la *valeur attendue* de x_t, disons x^*_t, où l'anticipation est basée sur toute l'information observée jusqu'au temps $t-1$:

$$y_t = \alpha_0 + \alpha_1 x^*_t + u_t. \qquad \textbf{[18.68]}$$

Une hypothèse naturelle concernant $\{u_t\}$ est que $E(u_t|I_{t-1}) = 0$, où I_{t-1} dénote toute l'information sur y et x observée jusqu'au temps $t-1$; cela veut dire que $E(y_t|I_{t-1}) = \beta_0 + \beta_1 x^*_t$. Pour compléter ce modèle, nous avons besoin d'une hypothèse sur la manière dont l'anticipation x^*_t est formée. Nous avons vu un exemple simple d'anticipations adaptatives dans la section 11.2 où $x^*_t - x^*_{t-1} = 1$. Un schéma d'anticipations adaptatives plus compliqué est

$$x^*_t - x^*_{t-1} = \lambda (x_{t-1} - x^*_{t-1}), \qquad \textbf{[18.69]}$$

où $0 < \lambda < 1$. Cette équation implique que le changement dans les anticipations s'ajuste par rapport au fait que la valeur réalisée de la période précédente était au-dessus ou en-dessous de son anticipation. L'hypothèse $0 < \lambda < 1$ signifie que le changement dans les anticipations est une fraction de l'erreur de dernière période.

i. Montrez que les deux équations impliquent que $y_t = \lambda \alpha_0 + (1 - \lambda)y_{t-1} + \lambda\alpha_1 x_{t-1} + u_t + (1 - \lambda)u_{t-1}$. [*Conseil* : Retardez l'équation (18.68) d'une période, multipliez la par $(1 - \lambda)$, retirez cela de (18.68). Après utilisez (18.69).]

ii. Sous $E(u_t|I_{t-1}) = 0$, $\{u_t\}$ est non corrélé sériellement. Qu'est ce que cela implique à propos des nouvelles erreurs ?

iii. Si on écrit l'équation à partir de (i) comme $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_{t-1} + v_t$, comment estimeriez-vous de manière consistante β_1 ?

iv. Étant donné des estimateurs convergents de β_j, comment estimeriez-vous de manière consistante α_1 ?

3. Supposons que $\{y_t\}$ et $\{z_t\}$ soient des séries I(1), mais que $y_t - \beta z_t$ soit I(0) pour une certaine valeur $\beta \neq 0$. Montrez que pour tout $\delta \neq \beta$, $y_t - \delta z_t$ doit être I(1).

4. Considérez le modèle à correction d'erreur de l'équation (18.37). Montrez que si vous ajoutez un autre retard au terme à correction d'erreur, $y_{t-2} - \beta x_{t-2}$, l'équation souffre de collinéarité parfaite. (*Conseil* : montrez que $y_{t-2} + \beta x_{t-2}$ est une fonction linéaire parfaite de $y_{t-1} - \beta x_{t-1}$, Δy_{t-1} et Δx_{t-1})

5. Supposons le processus $\{(x_t, y_t) : t = 0, 1, 2, \ldots\}$ qui suit les équations

$$y_t = \beta x_t + u_t$$

et

$$\Delta x_t = \gamma \Delta x_{t-1} + v_t,$$

où $E(u_t|I_{t-1}) = E(v_t|I_{t-1}) = 0$, et où I_{t-1} contient l'information sur x et y datés au temps $t - 1$ et avant, $\beta \neq 0$ et $|\gamma| < 1$ [si bien que x_t, et donc y_t, sont I(1)]. Montrer que ces deux équations impliquent un modèle à correction d'erreur du type

$$\Delta y_t = \gamma_1 \Delta x_{t-1} + \delta(y_{t-1} - \beta x_{t-1}) + e_t,$$

où $\gamma_1 = \beta\gamma$, $\delta = -1$ et $e_t = u_t + \beta v_t$ (*conseil* : retirez d'abord y_{t-1} des deux termes de l'équation. Ensuite, ajoutez et retirez βx_{t-1} du terme de droite et réarrangez. Enfin, utilisez la seconde équation pour obtenir le modèle à correction d'erreur qui contient Δx_{t-1}.)

6. En utilisant les données mensuelles de VOLAT, le modèle suivant a été estimé :

$$\widehat{pcip} = \underset{(0,56)}{(1,54)} + \underset{(0,042)}{(0,344)}pcip_{-1} + \underset{(0,045)}{(0,074)}pcip_{-2} + \underset{(0,042)}{(0,073)}pcip_{-3} + \underset{(0,013)}{(0,031)}pcsp_{-1}$$

$$n = 554,\ R^2 = 0,174,\ \bar{R}^2 = 0,168,$$

où *pcip* est la variation en pourcentage dans la production industrielle, à un taux annualisé, et *pcsp* la variation en pourcentage de l'indice Standard & Poor 500, aussi à un taux annualisé.

i. Si les valeurs de *pcip* des 3 derniers mois sont zéro et si $pcsp_{-1} = 0$, quelle est la croissance prédite de la production industrielle de ce mois ? Est-ce statistiquement différent de zéro ?

ii. Si les valeurs de *pcip* des 3 derniers mois sont égales à zéro et mais que $pcsp_{-1} = 0$, quelle est la croissance prédite de la production industrielle de ce mois ?

7. Soit gM_t le taux de croissance annuel de l'offre de monnaie et soit $unem_t$ le taux de chômage. En supposant que $unem_t$ suit un processus AR(1) stable, expliquez en details comment vous testeriez si *gM* cause *unem* au sens de Granger.

8. Supposons que y_t suive le modèle

$$y_t = \alpha + \delta_1 z_{t-1} + u_t$$
$$u_t = \rho u_{t-1} + e_t$$
$$\mathrm{E}(e_t | I_{t-1}) = 0,$$

où I_{t-1} contient y et z datés en $t-1$ et avant.

i. Montrez que $\mathrm{E}(y_{t+1}|I_t) = (1-\rho)\alpha + \rho y_t - \rho\delta_1 z_{t-1}$ (*Conseil* : écrivez et insérez ceci dans la seconde équation ; ensuite insérez le résultat dans la première équation et prenez l'espérance conditionnelle.)

ii. Supposons que vous utilisez n observations pour estimer α, δ_1 et ρ. Écrivez l'équation qui prédit y_{n+1}.

iii. Expliquez pourquoi le modèle avec un retard de z et une corrélation sérielle AR(1) est un cas spécial du modèle

$$y_t = \alpha_0 + \rho y_{t-1} + \gamma_1 z_{t-1} + \gamma_2 z_{t-2} + e_t.$$

iv. Qu'est ce que le point (iii) suggère à propos de l'utilisation de modèles avec corrélation sérielle AR(1) à des fins de prévision ?

9. Soit $\{y_t\}$ une séquence I(1). Supposons que $\hat{f}_n = \hat{g}_n + y_n$ est la prévision à une étape de Δy_{n+1} et soit $\hat{f}_n = \hat{g}_n + y_n$ la prévision à une étape de y_{n+1}. Expliquez pourquoi les erreurs de prévision liées à la prévision de Δy_{n+1} et de y_{n+1} sont identiques.

EXERCICES SUR ORDINATEUR

C1. Utilisez les données de WAGEPRC pour cet exercice. Le Problème 5 du chapitre 11 a fourni des valeurs estimées d'un modèle à retards distribués finis de *gprice* sur *gwage*, où 12 retards de *gwage* ont été utilisés.

i. Estimez un modèle simple à RD géométriques de *gprice* sur *gwage*. En particulier, estimez l'équation (18.11) par MCO. Quel est l'effet de court terme et l'effet de long terme ? Décrivez la distribution estimée des retards.

ii. Comparez l'effet de court terme et l'effet de long terme à ceux obtenus dans le problème 5 du chapitre 11. Comment est-ce que les distributions de retards estimées se comparent entre elles ?

iii. Maintenant, estimez le modèle à retards distribués rationnels à partir de (18.16). Décrivez la distribution estimée des retards et comparez l'effet de court terme estimé et l'effet de long terme à ceux obtenus dans la partie (ii).

C2. Utilisez les données de HSEINV pour cet exercice.

i. Testez la racine unitaire de log(*invpc*), en incluant une tendance linéaire temporelle et deux retards de $\Delta\log(invpc_t)$. Utilisez un niveau de significativité de 5 %.

ii. Utilisez l'approche de (i) pour tester une racine unitaire de log(*price*).

iii. Étant donné les résultats de (i) et (ii), est-ce qu'il convient de tester la cointégration entre log(*invpc*) et log(*price*) ?

C3. Utilisez les données de VOLAT pour cet exercice.

i. Estimez un modèle AR(3) pour *pcip*. Ensuite, ajoutez un quatrième retard et vérifiez qu'il est significatif.

ii. Ajoutez trois retards de *pcsp* au modèle AR(3) pour tester si *pcsp* cause au sens de Granger *pcip*. Énoncez avec prudence votre conclusion.

iii. Ajoutez trois retards de la variation de *i3* (le taux des obligations à 3 mois) au modèle de la partie (ii). Est-ce que *pcsp* cause au sens de Granger *pcip* conditionnellement au passé de Δi3 ?

C4. En testant la cointégration entre *gfr* et *pe* dans l'Exemple 18.5, ajoutez t^2 à l'équation (18.32) pour obtenir les résidus MCO. Incluez un retard dans le test DF augmenté. La valeur critique à 5 % de ce test est −4,15.

C5. Utilisez les données de INTQRT pour cet exercice.

i. Dans l'Exemple 18.7, on a estimé un modèle à correction d'erreur pour les rendements à détention des obligations à 6 mois, dans lequel un retard de ces rendements était une variable explicative. On a supposé que le paramètre de cointégration de l'équation $hy6_t = \alpha + \beta hy3_{t-1} + u_t$ était égal à un. Maintenant, ajoutez une variation avancée $\Delta hy3_t$, la variation contemporaine, $\Delta hy3_{t-1}$ et la variation retardée, $\Delta hy3_{t-2}$ de $\Delta hy3_{t-1}$. Bref, estimez l'équation

$$hy6_t = \alpha + \beta hy3_{t-1} + \phi_0 \Delta hy3_t + \phi 1 \Delta hy3_{t-1} + \rho 1 \Delta hy3_{t-2} + e_t$$

et reportez les résultats de l'équation. Testez $H_0 : \beta = 1$ contre une alternative bilatérale. Supposez que le retard et la variation avancée sont suffisants de sorte que $\{hy3_{t-1}\}$ est strictement exogène dans cette équation et ne vous préoccupez pas de la corrélation sérielle.

ii. Ajoutez $\Delta hy3_{t-2}$ et $(hy6_{t-2} - hy3_{t-3})$ au modèle à correction d'erreur de (18.39). Est-ce que ces termes sont conjointement significatifs ? Que concluez-vous à propos du modèle à correction d'erreur ?

C6. Utilisez les données de PHILLIPS pour cet exercice.

i. Estimez les modèles de (18.48) et (18.49) en utlisant les données jusqu'en 1997. Est-ce que les valeurs estimées des paramètres changent fortement par rapport à (18.48) et (18.49) ?

ii. Utilisez les nouvelles équations pour prédire $unem_{1998}$; arrondissez à deux décimales. Quelle équation produit-elle la meilleure prévision ?

iii. Comme vu dans le texte, la prévision de $unem_{1998}$ en utilisant (18.49) est 4,90. Comparez ceci à la prévision obtenue avec les données jusqu'en 1997. Est-ce que l'utilisation de cette année supplémentaire pour l'estimation produit une meilleure prévision ?

iv. Utilisez le modèle estimé de (18.48) pour obtenir une prévision de *unem* deux étapes à l'avance. En clair, prévoyez $unem_{1998}$ en utilisant l'équation (18.55) avec $\hat{\alpha} = 1{,}572$, $\hat{\rho} = 0{,}732$ et $h = 2$. Est-ce mieux ou moins bien qu'une prévision une étape à l'avance obtenue en insérant $unem_{1997} = 4{,}9$ dans (18.48) ?

C7. Utilisez les données de BARIUM pour cet exercice.

i. Estimez le modèle à tendance linéaire $chnimp_t = \alpha + \beta t + u_t$, en utilisant les 119 premières observations (ceci exclut les 12 derniers mois d'observations pour 1988). Quelle est l'erreur standard de régression ?

ii. Maintenant estimez un modèle AR(1) pour *chnimp*, une fois de plus en utilisant toutes les données sauf les 12 derniers mois. Comparez l'erreur standard de régression avec celle de (i). Quel modèle donne le meilleur ajustement ?

iii. Utilisez les modèles de (i) et (ii) pour calculer les erreurs de prévision une étape à l'avance pour les 12 mois de 1988. (Vous devriez obtenir 12 erreurs de prévision pour chaque méthode.) Calculez et comparez les RMSEs et les MAE pour les deux méthodes. Quelle méthode de prévision fonctionne le mieux hors échantillon pour ces prévisions une étape à l'avance ?

iv. Ajoutez des variables muettes mensuelles à la régression de (i). Est-ce qu'elles sont conjointement significatives ? (Ne vous préoccupez pas de la corrélation sérielle modeste dans les erreurs quand vous faites le test joint.)

C8. Utilisez les données de FERTIL3 pour cet exercice.

i. Tracez le graphe de *gfr* en fonction du temps. Est-ce qu'il contient une tendance ascendante ou descendante marquée sur la période complète d'échantillon ?

ii. En utilisant les données jusqu'en 1979, estimez un modèle à tendance temporelle cubique pour *gfr* (en clair, régressez *gfr* sur t, t^2, et t^3, avec un intercept). Commentez le R-carré de la régression.

iii. En utilisant le modèle de (ii), calculez l'erreur absolue moyenne des erreurs de prévision une étape à l'avance pour les années allant de 1980 à 1984.

iv. En utilisant les données jusqu'en 1979, régressez Δgfr_t sur une constante. Est-elle statistiquement différente de zéro ? Convient-il de supposer que tout terme de dérive est zéro, si on suppose que gfr_t suit une marche aléatoire ?

v. Prévoyez maintenant *gfr* de 1980 à 1984, en utilisant le modèle de marche aléatoire : la prévision de gfr_{n+1} est simplement gfr_n. Trouvez la MAE. Comment cela se compare-t-il à la MAE de (iii) ? Quelle méthode de prévision préférez-vous ?

vi. Estimez maintenant un modèle AR(2) pour *gfr*, une fois de plus en utilisant les données jusqu'en 1979. Est-ce que le second retard est significatif ?

vii. Obtenez les MAE de 1980 à 1984, en utilisant le modèle AR(2). Est-ce que ce modèle général marche mieux hors échantillon que la marche aléatoire ?

C9. Utilisez CONSUMP pour cet exercice.

i. Soit y_t le revenu disponible réel par tête. Utilisez les données jusqu'en 1989 pour estimer le modèle $y_t = \alpha + \beta t + \rho y_{t-1} + u_t$ et reportez les résultats de la manière habituelle.

ii. Utilisez l'équation estimée de (i) pour prévoir y en 1990. Quelle est l'erreur de prévision ?

iii. Calculez l'erreur absolue moyenne des prévisions une étape à l'avance pour les années 90, en utilisant les paramètres estimés de (i).

iv. Maintenant, calculez MAE sur la même période, mais retirez y_{t-1} de l'équation. Est-ce mieux d'inclure ou non y_{t-1} dans le modèle ?

C10. Utilisez les données de INTQRT pour cet exercice.

i. En utilisant toutes les données sauf des 4 dernières années (16 trimestres), estimez un modèle AR(1) pour $\Delta r6_t$ (On utilise la différence car $r6_t$ semble posséder une racine unitaire.) Utilisez la RMSE des prévisions une étape à l'avance de $\Delta r6_t$ en utlisant les 16 derniers trimestres.

ii. Ajoutez maintenant le terme à correction d'erreurs $spr_{t-1} = r6_{t-1} - r3_{t-1}$ à l'équation du point (i). (On suppose que le paramètre de cointégration est égal à un.) Calculez la RMSE pour les 16 trimestres. Est-ce que le terme à correction d'erreur améliore la prévision hors échantillon dans ce cas ?

iii. Estimez maintenant le paramètre de cointégration, plutôt que de le supposer égal à un. Utilisez à nouveau les 16 derniers trimestres pour générer une RMSE hors échantillon. Comment cela se compare-t-il par rapport aux points (i) et (ii) ?

iv. Est-ce que vos conclusions changeraient si vous prédisiez *r6* plutôt que $\Delta r6$. Expliquez.

C11. Utilisez les données de VOLAT pour cet exercice.

i. Confirmez que $lsp500 = \log(sp500)$ et $lip = \log(ip)$ semblent avoir une racine unitaire. Utilisez les tests de Dickey-Fuller avec 4 variations retardées et faites les tests avec et sans tendance linéaire.

ii. Faites une régression de *lsp*500 sur *lip*. Discutez de la taille de la statistique en *t* et du *R*-carré.

iii. Utilisez les résidus du point (ii) pour tester si *lsp*500 et *lip* sont cointégrés. Utilisez le test standard de Dickey-Fuller et le test ADF avec deux retards. Que concluez-vous ?

iv. Ajoutez une tendance linéaire à la régression du point (ii) et testez la cointégration en utilisant les mêmes tests que dans le point (iii).

v. Est-ce que le prix des actions et l'activité économique réelle semblent avoir une relation de long terme ?

C12. Utilisez aussi les données de VOLAT pour cet exercice. L'Exercice d'ordinateur C11 analyse la relation de long terme entre les prix des actions et l'activité économique réelle. Ici, vous analyserez la question de la causalité à la Granger en utilisant des variations en pourcentage.

i. Estimez un modèle AR(3) pour $pcip_t$, la variation en pourcentage de la production industrielle (donnée en taux annualisé). Montrez que le second et le troisième retards sont conjointement significatifs au niveau de 2,5 %.

ii. Ajoutez un retard de $pcsp_t$ à l'équation estimée au point (i). Le retard est-il statistiquement significatif ? Qu'est-ce que cela vous apprend à propos de la causalité à la Granger entre le prix des actions et l'activité économique réelle ?

iii. Refaites le point (ii) mais obtenez une statistique en *t* robuste à l'hétéroscédasticité. Est-ce que ce test robuste modifie les conclusions du point (ii) ?

C13. Utilisez les données de TRAFFIC2 pour cet exercice. Ces données mensuelles, sur les accidents de la route en Californie entre 1981 et 1989, ont été utilisées dans l'Exercice d'ordinateur C11.

i. En utilisant une régression standard de Dickey-Fuller, testez si $ltotacc_t$ possède une racine unitaire. Pouvez-vous rejeter une racine unitaire au niveau de 2,5 % ?

ii. Ajoutez maintenant deux variations retardées au test du point (i) et calculez le test Dickey-Fuller augmenté. Que concluez-vous ?

iii. Ajoutez une tendance linéaire temporelle à la régression ADF du point (ii). Qu'arrive-t-il ?

iv. Étant donné les résultats des points (i) à (iii), quelle serait pour vous la meilleure caractérisation pour $ltotacc_t$: un processus I(1) ou un processus I(0) autour d'une tendance linéaire temporelle ?

v. Testez la présence d'une racine unitaire pour le pourcentage de victimes $prcfat_t$, en utilisant deux retards dans la régression ADF. Dans ce cas, est-ce que cela importe qu'on introduise une tendance linéaire temporelle ou non ?

C14. Utilisez les données MINWAGE.DTA du secteur 232 pour répondre aux questions suivantes.

i. Confirmez que $lwage232_t$ et $lemp232_t$ sont bien caractérisés par des processus I(1). Utilisez le test DF augmenté avec un retard de $gwage232$ et de $gemp232$, et une tendance linéaire temporelle. N'y a-t-il pas de doute sur le fait que ces séries sont supposées avoir des racines unitaires ?

ii. Régressez $lemp232_t$ sur $lwage232_t$ et testez la cointégration, avec et sans tendance temporelle, avec deux retards dans le test d'Engle-Granger augmenté. Que concluez-vous ?

iii. Régressez $lemp232_t$ sur le log du taux de salaire réel, $lrwage232_t = lwage232_t - lcpi_t$, et une tendance temporelle. Trouvez-vous de la cointégration ? Sont-ils plus proches d'être cointégrés lorsque vous utilisez les salaires réels plutôt que les salaires nominaux ?

iv. Quels sont les facteurs qui pourraient manquer dans la relation de cointégration du point (iii) ?

C15. Cette question vous demande d'étudier ce qu'on appelle la courbe de Beveridge dans une perspective de cointégration. Les données américaines mensuelles entre décembre 2000 et février 2012 sont dans BEVERIDGE.

i. Testez la racine unitaire dans *urate* en utilisant le test de Dickey-Fuller (avec une constante) et le DF augmenté avec deux retards de *curate*. Que concluez-vous ? Est-ce que les retards de *curate* dans le test DF augmenté sont significatifs ? Est-ce que cela importe pour le résultat concernant la racine unitaire ?

ii. Répétez le point (i) mais avec le taux de vacance, *vrate*.

iii. En supposant que *urate* et *vrate* sont tous les deux I(1), la courbe de Beveridge

$$urate_t = \alpha + \beta\, vrate_t + u_t,$$

n'a du sens que si *urate* et *vrate* sont cointégrés (avec un paramètre de cointégration $\beta < 0$). Testez la cointégration en utilisant le test d'Engle-Granger sans retard. Est-ce que *urate* et *vrate* sont cointégrés au seuil de significativité de 10 % ? Qu'en est-il pour le seuil à 5 % ?

iv. Calculez l'estimateur à retards et valeurs avancées avec $cvrate_t$, $cvrate_{t-1}$, et $cvrate_{t+1}$ comme variables explicatives I(0) ajoutées à l'équation du point (ii). Calculez l'écart-type de Newey-West de $\hat{\beta}$ en utilisant 4 retards (donc $g = 4$ en termes de notation de la Section 12.5). Quel est l'intervalle de confiance à 95 % pour β ? Quelle comparaison avec un intervalle de confiance qui n'est pas robuste à la corrélation sérielle (ou à l'hétéroscédasticité) ?

v. Refaites le test d'Engle-Granger mais avec deux retards dans la régression DF augmentée. Que concluez-vous concernant le caractère robuste de l'affirmation selon laquelle *urate* et *vrate* sont cointégrés ?

CHAPITRE 19

MENER À BIEN UN PROJET EMPIRIQUE

Traduction de Michel Beine

Dans ce chapitre, nous expliquons quels éléments sont nécessaires pour réussir une analyse empirique, en insistant sur la capacité à terminer un travail à rendre. En plus de vous rappeler les problèmes importants qui ont été soulevés dans cet ouvrage, nous insisterons sur des thèmes récurrents qui sont importants pour la recherche appliquée. Nous ferons également quelques suggestions de sujets de recherche de manière à stimuler votre imagination. Nous intégrons en référence des listes de journaux et de données économiques.

19.1 POSER UNE QUESTION

Il ne faut pas sous-estimer l'importance de poser une question très précise à laquelle, en principe, on peut apporter une réponse quand on dispose des données adéquates. On ne peut pas savoir par où commencer si l'objectif de l'analyse n'a pas été rendu explicite. La disponibilité importante de données renforce la tentation de se lancer dans une collecte de données basée sur des idées à moitié construites mais c'est souvent contre-productif. Si les hypothèses et le modèle à estimer ne sont pas formulés de manière soignée, il est probable que vous oublierez de réunir des informations sur les variables importantes, que vous obtiendrez un échantillon de la mauvaise population, ou que vous collecterez des données relatives à la mauvaise période de temps.

Cela ne signifie pas que votre question soit sans fondement. En particulier, si le projet a une échéance, vous ne devez pas être trop ambitieux. De ce fait, lorsque vous choisissez un sujet, vous devez être relativement certain que les données existent de manière à vous permettre de répondre à la question en temps utile.

Au moment de choisir un sujet, vous devez choisir un domaine de l'économie ou des sciences sociales qui vous intéresse. Par exemple, si vous avez suivi un cours en économie du travail, vous avez probablement vu des théories qui peuvent être testées empiriquement ou des relations qui peuvent avoir une portée de politique économique. Les économistes du travail proposent constamment de nouvelles variables qui peuvent expliquer les différentiels de salaire. On peut citer par exemple : la qualité de l'école secondaire [Card et Krueger (1992) ou Betts (1995)], la quantité de maths et de sciences suivies à l'école secondaire [Levine et Zimmerman (1995)], l'apparence physique [Hamermesh et Biddle (1994), Averett et Korenman (1996), Biddle et Hamermesh (1998), ou encore Hamermesh et Parker (2005)]. Les chercheurs en finances publiques fédérales ou locales étudient comment l'activité économique locale dépend de variables de politique économique telles que les taxes sur l'habitat, les taxes sur la vente, le niveau et la qualité des services (tels que les écoles, la police, les pompiers), et ainsi de suite. [Voir par exemple, White (1986), Papke (1987), Bartik (1991), Netzer (1992), ou Mark, McGuire, et Papke (2000).]

Les économistes qui étudient les problèmes d'éducation s'intéressent à la manière dont les dépenses affectent la performance scolaire [Hanushek (1986)], dans quelle mesure fréquenter certaines catégories d'écoles améliore la réussite scolaire [par exemple, Evans et Schwab (1995)] et quels facteurs affectent le lieu où les écoles privées choisissent de se localiser [Downes et Greenstein (1996)].

Les macroéconomistes sont intéressés par les relations entre différentes séries temporelles agrégées, telles que le lien entre la croissance du produit domestique brut et la croissance de l'investissement fixe ou de l'investissement en équipements [voir De Long et Summers (1991)] ou encore l'effet des taxes sur les taux d'intérêt [par exemple, Peek (1982)].

Bien entendu, il est aussi utile d'estimer des modèles qui sont essentiellement descriptifs. Par exemple, quand on évalue les taxes sur l'habitat, on utilise des modèles (appelés modèles de prix hédoniques) pour estimer la valeur immobilière des maisons qui n'ont pas été vendues récemment. Ce type de modèle associe le prix d'une maison à ses caractéristiques (taille, nombre de chambres à coucher, nombre de salles de bain, et ainsi de suite). Pour autant, ce type de question ne constitue pas vraiment un sujet intéressant pour un projet avec échéance : il est peu probable que l'on apprenne des choses surprenantes et une telle analyse n'a pas d'applications précises en termes de politique économique. Par contre, ajouter le taux de criminalité dans

le voisinage comme variable explicative permettrait de déterminer dans quelle mesure le facteur crime a une influence sur le prix des maisons, ce qui peut s'avérer utile pour estimer le coût de la criminalité.

Certaines relations essentiellement descriptives ont été estimées en utilisant des données macro-économiques. Par exemple, une fonction d'épargne agrégée peut être utilisée pour estimer la propension marginale agrégée à épargner, ainsi que la réponse en termes d'épargne au rendement des actifs (tel que les taux d'intérêt). On peut rendre ce type d'analyse encore plus intéressante en utilisant des données de séries temporelles d'un pays qui a traversé une période de troubles politiques afin de déterminer dans quelle mesure les taux d'épargne diminuent au cours des périodes d'incertitude politique.

Une fois que vous avez choisi un domaine de recherche, il y a différentes manières d'identifier les articles spécifiques du domaine. Le *Journal of Economic Literature* (JEL) dispose d'un système de classification détaillée grâce auquel chaque article possède une sorte de code d'identification qui le situe dans certains sous-domaines de l'économie. Le JEL contient également une liste d'articles publiés dans une grande variété de journaux, organisée par thèmes et il propose même des résumés courts de certains articles.

Très utiles pour trouver les articles publiés dans différents domaines, il existe des services sur Internet, tel qu'*EconLit*, auxquels beaucoup d'universités souscrivent. *EconLit* permet aux utilisateurs de faire une recherche complète à partir d'à peu près tous les journaux en économie, par auteur, par sujet, par mot dans le titre, et ainsi de suite. Le *Social Sciences Citation Index* est utile pour trouver les articles publiés dans un grand nombre de domaines des sciences sociales, ou encore les articles populaires qui ont souvent été cités par d'autres travaux publiés.

Google Scholar est quant à lui un moteur de recherche sur Internet qui peut s'avérer très utile pour explorer la recherche dans différents domaines ou la recherche d'un auteur en particulier, notamment pour prendre connaissance des travaux qui n'ont pas été publiés dans un journal académique ou qui n'ont pas encore été publiés.

Au moment de choisir un sujet, vous devez garder à l'esprit un certain nombre de choses. Tout d'abord, pour qu'une question soit intéressante, elle ne doit pas avoir nécessairement des applications de politique économique générales. Par exemple vous pourriez être intéressé par le fait de savoir dans quelle mesure vivre dans une communauté au sein de votre université amène les étudiants à avoir des notes moyennes plus élevées ou plus basses. Cette question ne suscitera peut-être pas l'intérêt des gens qui n'appartiennent pas à votre université, mais elle intéressera probablement au moins un certain nombre de personnes au sein de l'université. D'un autre côté, vous pouvez étudier un problème qui, au départ, est d'un intérêt local mais qui s'avère avoir un intérêt général : par exemple, la détermination des facteurs qui affectent l'abus d'alcool sur les campus des universités et quelles politiques l'université peut mener.

Par ailleurs, il peut être très difficile, en particulier pour un projet sur un trimestre ou un semestre, de mener une recherche vraiment originale en utilisant les agrégats macro-économiques classiques de l'économie américaine. Par exemple, la question de savoir si la croissance de la monnaie ou la croissance de la dépense publique affecte la croissance économique a été, et continue à être, étudiée par des macroéconomistes professionnels. La question de savoir si les rendements d'actions ou d'autres actifs peuvent être systématiquement prédits en utilisant l'information connue a été, pour des raisons évidentes, étudiée très attentivement. Cela ne signifie pas que vous deviez éviter d'estimer des modèles macro-économiques ou des modèles de finance empirique car l'utilisation de données récentes peut amener quelque chose de constructif au débat. En outre, vous pouvez parfois trouver une nouvelle variable qui a un effet important sur les agrégats économiques ou sur les rendements financiers ; une telle découverte peut être grisante.

Il faut comprendre que des exercices tels que l'intégration de quelques années supplémentaires pour estimer une courbe de Philips standard ou une fonction de consommation agrégée, que ce soit pour l'économie américaine ou pour tout autre économie importante, ne vont probablement pas apporter de nouveaux

résultats, bien qu'ils puissent être instructifs pour un étudiant. Il est peut être plus intéressant d'utiliser les données d'une économie plus petite pour estimer une courbe de Philips statique ou dynamique ou encore une courbe de Beveridge (en permettant par exemple que les pentes des courbes dépendent de l'information qui est connue avant la période de temps contemporaine), ou tester l'hypothèse des marchés efficients, etc…

Au niveau macro-économique, il y a également beaucoup de questions qui peuvent être étudiées en détail. Par exemple, les économistes du travail ont publié beaucoup d'articles dont le but est d'estimer le rendement de l'éducation. Cette question est toujours étudiée parce qu'elle est très importante. De nouvelles bases de données, ainsi que de nouvelles approches économétriques, continuent à être développées. Par exemple, comme nous l'avons vu dans le chapitre 9, certaines données permettent mieux que d'autres d'approximer certaines variables comme la capacité non observée. (Comparez WAGE1 et WAGE2.) Dans d'autres cas, on peut obtenir des données de panel ou des données tirées d'une expérience naturelle – voir chapitre 13 – qui permettent d'appréhender une vieille question avec une nouvelle perspective.

On peut aussi citer l'exemple des criminologues qui étudient les effets de différentes lois sur la criminalité. La question de savoir si la peine capitale a un effet de découragement a été longuement débattue. De même, les économistes ont été intéressés par la question de l'effet potentiel des taxes sur le tabac et l'alcool sur leur consommation (comme toujours dans un sens « toutes autres choses égales par ailleurs »). Puisque des données sur des années supplémentaires deviennent disponibles au niveau de l'État, une base de données plus riche, en panel, peut être créée et cela peut nous aider à appréhender d'une meilleure manière des questions majeures de politique économique. En outre, l'efficacité de certaines innovations relativement récentes de lutte contre la criminalité – telle que la politique au niveau de la communauté – peut être évaluée empiriquement.

Lorsque vous formulez votre question, il est utile de discuter vos idées avec vos condisciples, enseignants et amis. Vous devriez être capable de convaincre les gens que la réponse à la question est d'un certain intérêt. (Bien sûr, savoir si vous pouvez proposer une réponse de manière persuasive est un autre problème, mais vous devez commencer par avoir une question intéressante.) Si quelqu'un s'intéresse au sujet de l'article sur lequel vous travaillez et que la réponse apportée est « je fais un article sur le crime » ou « je fais un article sur les taux d'intérêt », il est fort probable que vous avez choisi un problème général sans formuler une vraie question. Vous devriez plutôt être capable de dire quelque chose comme « j'étudie les effets de la politique de communauté sur les taux de criminalité urbains aux États-Unis » ou « j'étudie comment la volatilité de l'inflation influence les taux d'intérêt à court terme au Brésil. »

19.2 REVUE DE LA LITTÉRATURE

Tous les articles, même s'ils sont relativement courts, doivent contenir une revue de la littérature appropriée. Il est rare que l'on essaye d'entamer un projet empirique sur lequel il n'y a pas eu de publication auparavant. Si vous cherchez dans les journaux ou si vous utilisez un **service de recherche en ligne** tel que *EconLit* pour aborder un sujet, alors vous êtes déjà bien avancé dans votre revue de la littérature. Si vous choisissez un sujet – tel que l'étude des effets de l'utilisation des drogues sur la performance des étudiants de licence dans votre université – alors vous devrez probablement travailler un peu plus dur. Mais les services de recherche en ligne rendent ce travail bien plus facile, car vous pouvez chercher par mots-clés, par mot dans le titre, par auteur, etc. Vous pouvez alors lire les résumés d'articles pour voir dans quelle mesure ils sont reliés à votre propre travail.

Lorsque vous faites une recherche sur la littérature, vous devez réfléchir aux sujets reliés qui peuvent ne pas apparaître dans une recherche qui utilise une série de mots-clés. Par exemple, si vous étudiez les effets de l'utilisation des drogues sur les salaires ou sur la moyenne des notes, vous devriez aussi regarder la littérature qui étudie l'effet de la consommation d'alcool sur ces facteurs. Savoir comment mener une

recherche approfondie de la littérature est une compétence qui s'acquiert, mais vous pouvez déjà aller très loin en réfléchissant avant d'entamer la recherche.

Il n'existe pas de consensus parmi les chercheurs sur la manière dont une revue de la littérature doit être incorporée dans un article. Certains aiment avoir une section appelée « revue de la littérature », alors que d'autres préfèrent inclure la revue de la littérature dans l'introduction. Ceci est essentiellement une question de goût, bien qu'une revue de la littérature approfondie mérite probablement sa propre section. Si le projet à remettre tient une place centrale dans le cours – par exemple dans un séminaire senior ou dans un cours d'économétrie avancée – alors votre revue de la littérature doit être probablement conséquente. Les travaux à rendre à la fin d'un premier cours sont habituellement plus courts et les revues de la littérature sont plus brèves.

19.3 COLLECTE DES DONNÉES

La décision concernant la base de données appropriée

Collecter des données pour un travail à échéance peut être instructif, excitant et parfois même frustrant. Vous devez tout d'abord décider quel type de données est nécessaire pour répondre à la question posée. Comme nous en avons discuté dans l'introduction et comme nous l'avons couvert à travers cet ouvrage, les bases de données apparaissent sous différentes formes. Les types les plus habituels sont les coupes transversales, les séries temporelles, les coupes transversales empilées et les données de panel.

Beaucoup de questions peuvent être abordés en utilisant chacune des structures de données que nous avons décrites. Par exemple, pour étudier si l'application des lois diminue la criminalité, nous pouvons utiliser une coupe transversale de villes, une série temporelle pour une ville donnée, ou des données de panel de villes – qui consistent en des données de mêmes villes observées sur deux ou plusieurs années.

Décider quel type de données collecter dépend souvent de la nature de l'analyse. Pour répondre à des questions au niveau individuel ou familial, nous avons souvent accès seulement à une seule coupe transversale ; typiquement, celles-ci sont obtenues via des enquêtes. Dans ce cas, nous devons nous demander si nous pouvons obtenir des données suffisamment riches pour mener une analyse *ceteris paribus* convaincante. Par exemple, supposons que nous voulions savoir si les familles qui épargnent via des comptes individuels de retraite (*individual retirement accounts* – IRA)[1] – qui possèdent certains avantages fiscaux – font moins d'épargne non-IRA. En d'autres termes, est-ce que l'épargne via les IRA se substitue simplement à d'autres formes d'épargne ? Il existe des bases de données, telles que l'enquête sur les finances des consommateurs, (*Survey of Consumer Finances*) qui contiennent des informations sur différentes formes d'épargne pour un échantillon de familles, différent chaque année. Certains problèmes apparaissent dans l'utilisation de telles données. Sans doute, le problème le plus important est de savoir s'il y a assez de variables de contrôle – notamment le revenu, des caractéristiques démographiques ainsi que des variables capturant la préférence pour l'épargne – pour mener une analyse *ceteris paribus* raisonnable. S'il s'agit là des seules données disponibles, alors nous devons faire avec ce que nous avons.

Les mêmes problèmes se posent avec des données en coupe transversale sur les entreprises, les villes, les États, et ainsi de suite. Dans la plupart des cas, ce n'est pas clair que nous puissions être capables de mener une analyse *ceteris paribus* avec une seule coupe transversale. Par exemple, toute étude sur les effets de l'application de la loi sur la criminalité doit tenir compte de l'endogénéité des dépenses pour l'application de la loi. Lorsqu'on utilise des méthodes de régression standard, il peut être très difficile de proposer une analyse convaincante *ceteris paribus*, peu importe le nombre de contrôles que nous avons. (Voir la section 19.4 pour une discussion supplémentaire.)

1 Note de la traduction : plan d'épargne retraite individuel par capitalisation aux États-Unis.

Si vous avez lu les chapitres avancés sur les méthodes de données de panel, vous savez que l'observation des mêmes unités pour deux ou plusieurs points dans le temps peut permettre de tenir compte des effets non observés, constants dans le temps, qui normalement biaiseraient les résultats dans une régression sur une coupe transversale unique. Les bases de données en panel sont relativement difficiles à obtenir pour des individus ou familles – bien que certaines bases importantes existent, telles que le *Panel Study of Income Dynamics*[2] – mais elles peuvent être utilisées de manière très convaincante. Des bases de données de panel sur des entreprises existent aussi. Par exemple, *Compustat* et le *Center for Research in Security Prices* (CRSP) gèrent de très grandes bases de données en panel qui contiennent de l'information financière sur les entreprises. Il est plus facile d'obtenir des données de panel sur des unités plus larges, telles que des écoles, des villes, des comtés, et des États, parce que ceux-ci disparaissent rarement et parce que les agences gouvernementales sont chargées de la collecte de l'information sur les mêmes variables chaque année. Par exemple, le Bureau Fédéral d'Investigation (Federal Board of Investigation – FBI) collecte et fournit de l'information détaillée sur les taux de criminalité au niveau des villes. Les sources de données sont indiquées à la fin de ce chapitre.

Les données se présentent sous diverses formes. Certaines bases de données, particulièrement les bases historiques, sont disponibles seulement sous format papier. Pour de petites bases de données, saisir les données vous-même à partir de la forme papier est possible et pratique. Parfois, les articles sont publiés avec les bases de données modestes – particulièrement dans les applications en séries temporelles. Celles-ci peuvent être utilisées dans une étude empirique, peut-être en ajoutant des données supplémentaires pour les années plus récentes.

Beaucoup de bases de données sont disponibles en format électronique. Différentes agences gouvernementales fournissent des données sur leur site Internet. Parfois, des sociétés privées compilent des bases de données pour les rendre plus aisées pour l'utilisateur et les proposent alors moyennant une cotisation. Les auteurs des articles sont souvent enclins à fournir leurs bases de données en format électronique. De plus en plus de bases de données sont disponibles sur Internet. Le Web est une ressource importante de **bases de données en ligne**. De nombreux sites Internet contenant des données économiques et d'autres types de données ont été créés. D'autres sites contiennent des liens vers des bases de données qui sont intéressantes pour les économistes ; certains de ces sites sont indiqués à la fin de ce chapitre. Généralement, rechercher des sources de données sur Internet est facile et deviendra de plus en plus aisé dans le futur.

Saisir et conserver des données

Une fois que vous avez décidé du type de données et que vous avez déterminé une source de données, vous devez arranger les données dans un format utilisable. Si les données sont fournies dans un format électronique, elles sont déjà dans un certain format, idéalement pour une utilisation générale. La manière la plus flexible d'obtenir des données dans un format électronique est le fichier en **texte standard (ASCII)**. Tous les logiciels statistiques et économétriques permettent de conserver les données brutes de cette manière. Typiquement, on peut directement lire un fichier texte dans un logiciel économétrique, moyennant le fait que le fichier est structuré de manière appropriée. Les fichiers de données que nous avons utilisés dans ce texte fournissent quelques exemples de la manière dont des données transversales, temporelles, empilées, et en panel peuvent être conservées. Une règle à suivre est que les données doivent avoir une forme tabulée : chaque observation est présentée par une ligne différente, les différentes variables sont données par les colonnes. De temps en temps, vous pourrez rencontrer une base de données dans laquelle chaque colonne représente une observation et chaque ligne une variable différente. Ceci n'est pas idéal, mais la plupart des logiciels permettent aux données d'être lues de cette manière et ensuite d'être restructurées. Naturellement, il est crucial de savoir comment les données sont organisées avant de les lire dans votre logiciel économétrique.

2 Note de la traduction : base de données de panel commencé en 1968, sur un échantillon de ménages américains.

Pour les données temporelles, il y a seulement une manière raisonnable d'entrer et de conserver les données, à savoir de manière chronologique, avec la période de temps la plus ancienne donnée par la première observation et la période de temps la plus récente par la dernière observation. Il est souvent utile d'inclure des variables indiquant l'année et, si cela a du sens, le trimestre ou le mois. Cela facilite l'estimation d'un ensemble de modèles, en incluant la possibilité de tenir compte de la saisonnalité et de changements structurels pour différentes périodes de temps. Pour les données empilées, il est habituellement optimal d'avoir la coupe transversale de l'année la plus ancienne correspondant au premier bloc d'observations, suivie par la coupe transversale de la seconde année, et ainsi de suite. (Voir FERTIL1 comme exemple.) Cet arrangement n'est pas crucial, mais il est très important d'avoir une variable indiquant l'année correspondant à chaque observation.

Pour les données de panel, comme nous en avons discuté dans la section 13.5, il est optimal que toutes les années pour chaque unité de coupe transversale soient adjacentes et soient dans l'ordre chronologique. Cette façon d'ordonner les données permet d'utiliser toutes les méthodes de données de panel vues dans les chapitres 13 et 14. Avec les données de panel, il est important d'inclure un identifiant unique pour chaque unité en coupe transversale ainsi qu'une variable indiquant l'année.

Si vous obtenez vos données en format papier, vous avez différentes options pour les saisir informatiquement. Tout d'abord, vous pouvez créer un fichier texte en utilisant un éditeur de texte standard. (C'est de cette manière que certaines bases de données brutes inclues dans cet ouvrage ont été initialement créées.) Typiquement, il est nécessaire que chaque ligne commence par une nouvelle observation, que chaque ligne contienne le même ordre des variables – en particulier, chaque ligne devrait avoir le même nombre d'entrées – et que les valeurs soient séparées par un espace. Parfois, des séparateurs différents, tels qu'une virgule, sont meilleurs, mais ceci dépend du logiciel que vous utilisez. Si vous avez des observations manquantes sur certaines variables, vous devez décider comment les renseigner, laisser juste un blanc n'est pas approprié. Beaucoup de logiciels de régression acceptent un point-virgule comme symbole d'une valeur manquante. Certaines personnes préfèrent utiliser un nombre – probablement une valeur impossible pour une variable d'intérêt – pour renseigner les valeurs manquantes. Si vous n'êtes pas attentifs, ceci peut être dangereux, comme nous en discuterons plus loin.

Si les données que vous avez ne sont pas numériques – vous voulez par exemple inclure les noms dans un échantillon de collèges ou de villes – alors, il vous faudra vérifier, pour le logiciel économétrique que vous utilisez, quelle est la meilleure manière de saisir ces variables (souvent appelées *strings*, ou chaînes de caractères). Souvent, les chaînes de caractères sont indiquées entre des guillemets simples ou doubles. Parfois, le fichier de texte peut suivre un formatage rigide des chaînes de caractères, ce qui requiert habituellement un petit programme que l'on doit lire dans le fichier texte. Mais vous devez vérifier votre logiciel économétrique pour les détails.

Une autre option, souvent disponible, est d'utiliser un tableur pour saisir vos données, tels qu'Excel. Cette solution présente certains avantages par rapport à un fichier texte. Tout d'abord, puisque chaque observation de chaque variable est une cellule, il est moins probable que des nombres vont être accolés (comme cela arriverait si vous oubliez d'introduire un espace dans un fichier texte). Deuxièmement, les tableurs offrent la possibilité de manipuler des données, notamment de trier ou de calculer des moyennes. Cet avantage est moins important si vous utilisez un logiciel qui permet une gestion sophistiquée des données. Beaucoup de logiciels, y compris Stata ou Eviews, le permettent. Si vous utilisez un tableur pour la saisie initiale des données, alors vous devrez souvent exporter les données dans un format qui peut être lu par votre logiciel économique. Généralement, cela se fait sans problème, puisque les tableurs exportent vers des fichiers texte en proposant différents formats.

Une troisième alternative est d'entrer les données directement dans votre logiciel économétrique. Bien que cela évite le recours à un éditeur de texte ou à un tableur, cela peut être plus périlleux si vous ne pouvez pas vous déplacer librement entre les différentes observations pour faire des corrections ou des additions.

Les données téléchargées sur Internet peuvent apparaître dans un format différent. Souvent, les données sont sous la forme de fichiers textes, mais différentes conventions sont utilisées pour séparer les variables. Pour les bases de données en panel, les conventions sur la manière d'ordonner les données peuvent être différentes. Certaines bases de données venant d'Internet sont présentées sous la forme de fichiers tableurs, auquel cas vous devrez utiliser un tableur approprié pour les lire.

Examiner, nettoyer et décrire vos données

Il est extrêmement important de devenir familier avec toute base de données que vous utilisez dans une analyse empirique. Si vous saisissez les données vous-même, vous serez forcé de tout savoir à propos des données. Mais si vous obtenez les données d'une source extérieure, vous devrez passer du temps pour comprendre sa structure et ses conventions. Même les bases de données les plus souvent utilisées et soigneusement documentées peuvent contenir des problèmes. Si vous utilisez une base de données que l'auteur d'un article vous a fournie, vous devez être conscient que des règles utilisées pour la construction de la base de données peuvent avoir été oubliées.

Nous avons précédemment vu les manières habituelles de stocker les différentes bases de données. Vous devez aussi savoir comment les valeurs manquantes ont été codées. Il est préférable d'indiquer les valeurs manquantes par un caractère non numérique, tel qu'un point-virgule. Si un nombre est utilisé comme code de valeurs manquantes, tel que « 999 » or « –1 », vous devez alors être très prudent quand vous utilisez ces observations pour calculer toute statistique. Votre logiciel économétrique ne saura probablement pas qu'un nombre particulier représente une valeur manquante : il est probable que ces observations seront utilisées dans le calcul, puisqu'elles sont valides, et ceci peut produire des résultats particulièrement trompeurs. La meilleure approche est de transformer tous les codes numériques pour les valeurs manquantes en d'autres caractères (tel qu'un point-virgule) qui ne peuvent pas être confondus avec des données réelles.

Vous devez aussi connaître la nature des variables de la base de données. Quelles sont les variables binaires ? Quelles sont les variables ordinales (telles qu'une notation de crédit) ? Par exemple, est-ce que les valeurs monétaires sont exprimées en dollars, milliers de dollars, millions de dollars, ou encore en une autre unité ? Les variables qui représentent un taux – tels que les taux d'éviction de l'école, les taux d'inflation, les taux de syndicalisation, ou les taux d'intérêt – sont-elles mesurées en pourcentage ou en proportion ?

En particulier, pour les données temporelles, il est crucial de savoir si les valeurs monétaires sont exprimées en termes nominaux (en dollars courants) ou en termes réels (en dollar constants). Si les valeurs sont en termes réels, quelle est l'année de base ou la période de base ?

Si vous recevez une base de données d'un auteur, certaines variables peuvent avoir déjà été transformées d'une certaine manière. Par exemple, parfois seulement le log de la variable (tel que le salaire ou la rémunération) est reporté dans la base de données.

Il est nécessaire de détecter les erreurs dans les données pour préserver l'intégrité de toute analyse de données. Il est souvent utile de trouver le minimum et maximum, les moyennes et les écarts-types de toutes les variables ou au moins les plus importantes parmi celles qui seront analysées. Par exemple, si vous trouvez que la valeur minimum du niveau d'instruction dans votre échantillon est –99, vous saurez alors qu'au moins une valeur dans votre échantillon doit être transformée en une valeur manquante. Si, après analyse supplémentaire, vous trouvez qu'une série d'observations possèdent la valeur –99 comme niveau d'éducation, vous pouvez alors être sûr que vous êtes tombés sur le code de la valeur manquante pour l'éducation. Considérons un autre exemple : si vous trouvez que le taux moyen de condamnation pour meurtre dans un échantillon de villes est égal à 0,632, vous pouvez en déduire que le taux de condamnation est exprimé en proportion et non en pourcentage. Par contre, si la valeur maximum est supérieure à un, il est probable qu'il s'agisse là d'une erreur typographique. (Il n'est pas rare de trouver des bases de données dans lesquelles la

plupart des valeurs d'une variable en taux ont été introduites comme un pourcentage, mais où certaines ont été introduites comme une proportion, et vice-versa. Des erreurs d'encodage de données peuvent être difficiles à détecter mais il est important d'essayer de les repérer.)

On doit aussi être prudent dans l'utilisation des données temporelles. Si nous utilisons des données mensuelles ou trimestrielles, nous devons savoir quelles variables, s'il y en a, ont été ajustées pour la saisonnalité. Transformer des données requiert aussi beaucoup d'attention. Supposons que nous ayons une base de données mensuelles et que nous voulions créer la variation d'une variable d'un mois à l'autre. Afin de mener à bien cette opération, nous devons nous assurer que les données ont été rangées par ordre chronologique, de la période la plus ancienne à la plus récente. Si pour quelque raison ce n'est pas le cas, le calcul de la différence résultera en une catastrophe. Afin de s'assurer que les données ont été ordonnées de manière appropriée, il est utile d'avoir un indicateur de la période de temps. Avec des données annuelles, il suffit certes de connaître l'année, mais nous devrions savoir si l'année a été codée en quatre chiffres ou en deux chiffres (par exemple, 1998 contre 98). Avec des données mensuelles ou trimestrielles, il est aussi utile d'avoir une variable ou des variables indiquant le mois ou le trimestre. Avec des données mensuelles, nous pouvons avoir un ensemble de variables indicatrices (11 ou 12 variables) ou une variable indiquant le mois (de 1 à 12 ou une variable *string*, telle que *jan*, *fév*, et ainsi de suite).

Avec ou sans un indicateur annuel, mensuel, ou trimestriel, nous pouvons facilement construire des tendances temporelles dans tous les logiciels économétriques. Il est facile de construire des variables muettes saisonnières quand le mois ou le trimestre est indiqué. Nous devons savoir, au minimum, le mois ou le trimestre de la première observation.

Manipuler des données de panel peut être encore plus difficile. Dans le chapitre 13, nous avons discuté de l'approche générale qui consiste en des MCO empilés sur des données différenciées pour prendre en compte les effets non observés. En construisant les données différenciées, nous devons être attentifs à ne pas créer des observations fantômes. Supposons que nous ayons un panel équilibré de villes de 1992 à 1997. Même si les données ont été rangées par ordre chronologique pour chaque unité de coupe transversale – ce qui doit être fait avant l'analyse – calculer la différence de manière peu prudente créera une observation pour 1992 pour toutes les villes excepté la première dans l'échantillon. Cette observation sera la valeur en 1992 de la ville i, moins la valeur en 1997 de la ville $i - 1$, ce qui n'a clairement pas de sens. Par conséquent, nous devons nous assurer que l'année 1992 est manquante pour toutes les variables différenciées.

19.4 ANALYSE ÉCONOMÉTRIQUE

Cet ouvrage s'est concentré sur l'analyse économétrique et nous n'allons pas fournir une revue des méthodes économétriques dans cette section. Cependant, nous pouvons donner un certain nombre de renseignements généraux sur le type de problèmes que l'on doit considérer dans une analyse empirique.

Comme nous en avons discuté précédemment, après avoir décidé d'un sujet, nous devons collecter une base de données appropriée. En supposant que ceci ait été fait également, nous devons ensuite décider des méthodes économétriques appropriées.

Si votre cours s'est concentré sur l'estimation par moindres carrés ordinaires d'un modèle de régression linéaire multiple, soit sur des données transversales ou temporelles, l'approche économétrique a été essentiellement décidée pour vous. Ce n'est pas nécessairement une faiblesse, car les MCO constituent la méthode économétrique la plus utilisée. Bien sûr, vous devez encore décider s'il est nécessaire d'utiliser une variante des MCO – telle que les moindres carrés pondérés ou la correction pour la corrélation sérielle dans une régression sur données temporelles.

De manière à justifier les MCO, vous devez montrer que les hypothèses-clés des MCO sont satisfaites dans votre modèle grâce à des arguments convaincants. Comme nous en avons discuté largement, le premier problème est de savoir si le terme d'erreur n'est pas corrélé avec les variables explicatives. Idéalement, vous avez été capables de tenir compte de suffisamment de facteurs supplémentaires pour supposer que ce qui subsiste dans le terme d'erreur n'a pas de lien avec les régresseurs. En particulier, lorsque vous étudiez des données transversales individuelles, familiales, ou des données d'entreprise, vous serez souvent confrontés au problème d'auto-sélection – que nous avons discuté dans les chapitres 7 et 15. Par exemple, dans l'exemple sur les IRA de la section 19.3, il se peut que les familles pour lesquelles on n'observe pas de préférence pour l'épargne soient aussi celles qui ont ouvert des comptes IRA. Vous devez aussi être en mesure d'argumenter que les autres sources potentielles d'endogénéité – en particulier, les erreurs de mesures et la simultanéité – ne constituent pas un problème majeur.

Lorsque l'on spécifie un modèle, il faut aussi prendre une décision quant à la forme fonctionnelle. Est-ce que certaines variables doivent apparaître sous une forme logarithmique ? (Dans les applications économétriques, la réponse est souvent oui.) Est-ce qu'il faut inclure le niveau mais aussi le carré de certaines variables, afin de pouvoir capter un effet décroissant ? Comment est-ce que les facteurs qualitatifs doivent apparaître ? Est-ce qu'il est suffisant d'inclure seulement des variables binaires pour les différents attributs et les différents groupes ? Est-ce qu'ils doivent être combinés avec des variables quantitatives ? (Voir le chapitre 7 pour les détails.)

Une erreur habituelle, en particulier chez les débutants, est d'inclure de manière incorrecte des variables explicatives dans un modèle de régression, celles-ci étant reportées avec une valeur numérique qui n'a pas de signification quantitative. Par exemple, dans une base de données d'individus qui contient de l'information sur le salaire, l'éducation, l'expérience professionnelle, et d'autres variables, une variable « profession » peut être inclue. Typiquement, ces valeurs sont juste des codes arbitraires qui ont été attribués à différents métiers. Le fait qu'un instituteur reçoive, par exemple, la valeur 453 alors qu'un technicien en informatique reçoive la valeur 751, n'a de sens qu'en ce que cela permet de distinguer les deux métiers. Cela n'a pas de sens d'inclure la variable « profession » telle quelle dans un modèle de régression. (Quel sens y a-t-il de mesurer l'effet d'une augmentation d'occupation d'une unité lorsqu'une augmentation d'une unité n'a pas d'interprétation quantitative ?) En lieu et place, différentes variables muettes devraient être définies pour différentes professions (ou groupe de profession s'il y en a beaucoup). Dans ce cas, les variables muettes peuvent être inclues dans le modèle de régression. Un problème moins évident apparaît lorsqu'une variable qualitative ordonnée est inclue comme une variable explicative. Supposons que dans une base de données sur les salaires, on inclut une variable qui mesure « la satisfaction au travail », définie sur une échelle allant de 1 à 7, avec 7 correspondant à la satisfaction la plus importante. À condition d'avoir assez de données, nous voudrions définir un ensemble de six variables muettes, par exemple pour les niveaux de satisfaction au travail allant de deux à sept, en laissant le niveau 1 de satisfaction comme groupe de base. En incluant les six variables dans la régression, nous permettons d'avoir une relation complètement flexible entre la variable de réponse et la satisfaction au travail. En incluant la variable « satisfaction travail » telle quelle, on suppose implicitement qu'une augmentation d'une unité dans la variable ordinale a une signification quantitative. Alors que la direction de l'effet est souvent estimée de manière appropriée, l'interprétation du coefficient sur une variable ordinale est difficile. Quand une variable ordinale prend beaucoup de valeurs, alors nous pouvons définir un ensemble de variables muettes pour un intervalle de valeurs. Voir la section 7.3 comme exemple.

Parfois, nous voulons expliquer une variable qui correspond à une réponse ordinale. Par exemple, nous pouvons envisager d'utiliser la variable « satisfaction au travail », du type de celle décrite plus haut, comme variable dépendante dans un modèle de régression, en faisant apparaître les caractéristiques à la fois du travailleur et de l'employeur parmi les variables indépendantes. Malheureusement, avec la variable « satisfaction au travail » dans sa forme originale, les coefficients du modèle sont difficiles à interpréter : chaque coefficient mesure la variation de la satisfaction au travail étant donnée une augmentation d'une unité de la variable indépendante. Certains modèles – parmi lesquels le *Logit ordonné* et le *Probit ordonné* sont les plus courants – conviennent parfaitement à l'étude des réponses ordonnées. Ces modèles étendent essentiellement

les modèles *Probit* et *Logit* binaires que nous avons discutés dans le chapitre 17. [Voir Wooldridge (2010, Chapitre 15) pour un traitement des modèles à réponse ordonnée.] Une solution simple est de transformer toute réponse ordonnée en une réponse binaire. Par exemple nous pouvons définir une variable égale à un si la satisfaction au travail est au moins de 4, sinon 0. Malheureusement, la création d'une variable binaire fait perdre de l'information et exige d'utiliser un seuil quelque peu arbitraire.

Pour l'analyse de données transversales, un problème secondaire, mais néanmoins important, est de savoir si on est en présence d'hétéroscédasticité. Dans le chapitre 8, nous avons expliqué comment on pouvait traiter cela. La manière la plus simple est de calculer des statistiques robustes à l'hétéroscédasticité.

Comme nous avons insisté dans les chapitres 10, 11 et 12, l'analyse de séries temporelles requiert des précautions supplémentaires. L'équation doit-elle être estimée en niveau ? Si les niveaux sont utilisés, est-il nécessaire d'inclure des tendances temporelles ? Est-il plus approprié de différencier les données ? Si les données sont mensuelles ou trimestrielles, faut-il tenir compte de la saisonnalité ? Si vous introduisez de la dynamique – par exemple la dynamique de retards distribués – combien de retards doivent être inclus ? Vous devez débuter avec un certain nombre de retards basés sur l'intuition ou sur le sens commun, mais à la fin, ceci demeure un problème empirique.

S'il est possible que votre modèle soit mal spécifié, en raison par exemple de variables omises, et si vous utilisez les MCO, vous devriez essayer de mener une sorte d'analyse de mauvaise spécification, du type évoqué dans les chapitres 3 et 5. Pouvez-vous déterminer à partir d'un certain nombre d'hypothèses raisonnables, le signe du biais éventuel des estimateurs ?

Si vous avez étudié la méthode des variables instrumentales, vous savez qu'elle peut être utilisée pour résoudre différentes formes d'endogénéité, notamment les variables omises (Chapitre 15), les erreurs de mesure sur les régresseurs (Chapitre 15) et la simultanéité (Chapitre 16). Bien sûr, vous devez réfléchir de manière approfondie à la validité des variables instrumentales que vous considérez.

Les bons articles en sciences sociales empiriques proposent une **analyse de sensibilité**. En bref, cela signifie que vous devez estimer votre modèle original et ensuite le modifier dans des directions qui semblent raisonnables. Idéalement, les conclusions importantes ne changent pas. Par exemple, si vous utilisez comme variable explicative une mesure de la consommation d'alcool (par exemple, dans une équation expliquant les notes moyennes), obtenez-vous des résultats qualitativement similaires si vous remplacez la mesure quantitative par une variable muette indiquant la consommation d'alcool ? Si la variable de consommation binaire est significative mais la variable indiquant la quantité d'alcool consommée ne l'est pas, il se peut que la consommation d'alcool reflète certains attributs non observés qui affectent le score GPA et qui sont aussi corrélés avec la consommation d'alcool. Néanmoins, cela doit être examiné au cas par cas.

Si quelques observations sont très différentes de la majorité de l'échantillon – vous avez par exemple quelques entreprises dans l'échantillon qui sont beaucoup plus grandes que les autres entreprises – vos résultats changent-ils significativement lorsque ces observations sont exclues de l'estimation ? Si c'est le cas, vous devez sans doute changer les formes fonctionnelles pour prendre en compte ces observations de manière adéquate ou apporter des arguments pour dire qu'elles suivent un modèle complètement différent. Le problème des valeurs aberrantes a été discuté dans le chapitre 9.

L'utilisation des données de panel soulève un certain nombre de problèmes économétriques supplémentaires. Supposons que vous ayez deux périodes. Il y a au moins quatre manières d'utiliser deux périodes de données panel, sans recourir aux variables instrumentales. Vous pouvez empiler les deux années et mener une analyse standard par MCO, comme discuté dans le chapitre 13. Bien que cela augmente la taille de l'échantillon comparativement à une coupe transversale unique, cela ne permet pas de neutraliser l'effet des variables non observées qui ne varient pas dans le temps. En outre, les termes d'erreur, dans une telle équation, sont presque toujours corrélés de manière sérielle à cause d'un effet non observé. L'estimation des effets

aléatoires corrige le problème de corrélation sérielle et produit des estimateurs asymptotiquement efficaces, à condition que l'effet non observé soit une moyenne nulle étant donné la valeur des variables explicatives pour toutes les périodes de temps.

Une autre possibilité est d'inclure la variable dépendante retardée dans l'équation pour la seconde année. Dans le chapitre 9, nous avons présenté ceci comme une manière d'atténuer quelque peu le problème des variables omises, puisque dans tous les cas, nous considérons que la valeur de la variable dépendante en première période est fixe. Ceci mène souvent à des résultats similaires que l'étude des différences premières, comme nous l'avons vu dans le chapitre 13.

Avec des données de panel présentant des années supplémentaires, les mêmes alternatives s'offrent à nous, plus une alternative supplémentaire. Nous pouvons en effet appliquer la transformation par effets fixes pour éliminer les effets non observés. (Avec deux années de données, ceci revient au même que de calculer les différences premières.) Dans le chapitre 15, nous avons montré comment les méthodes d'estimation par variables instrumentales peuvent être combinées aux transformations de données de panel pour relâcher un peu plus les hypothèses d'exogénéité. D'une manière générale, c'est une bonne idée d'appliquer un certain nombre de méthodes économétriques raisonnables et de comparer les résultats obtenus. Ceci nous permet souvent de déterminer quelles hypothèses sont susceptibles d'être fausses.

Même si vous êtes très attentifs au moment du choix de votre sujet, de l'écriture de votre modèle, de la collecte de vos données, et de l'analyse économétrique, il est tout à fait possible que vous obteniez des résultats déconcertants – au moins au début. On a souvent tendance à essayer différents modèles, différentes techniques d'estimation ou peut-être différents sous-ensembles de données jusqu'au moment où les résultats correspondent plus ou moins à ce à quoi vous vous attendiez. Tous les chercheurs qui font de l'analyse empirique explorent virtuellement différents modèles avant de trouver le meilleur modèle. Malheureusement, cette pratique du **data mining** viole les hypothèses que nous avons faites dans notre analyse économétrique. Les résultats d'absence de biais des MCO et d'autres estimateurs, de même que les distributions de Student et de Fisher que nous avons dérivées pour les tests d'hypothèses, s'appuient sur l'hypothèse que nous observons un échantillon qui découle du modèle de la population et que nous estimions ce modèle une seule fois. En estimant des modèles qui sont des variantes de votre modèle original, vous violez cette hypothèse parce que vous utilisez le même jeu de données dans une recherche de spécification. En effet, on utilise alors le résultat des tests en utilisant les données pour spécifier le modèle. Les valeurs estimées et les tests qui sont issus de différents spécifications de modèle ne sont pas indépendants les uns par rapport aux autres.

Certaines recherches de spécification ont été programmées dans des logiciels classiques. Une méthode populaire de recherche est connue sous le nom de *régression en plusieurs étapes* (ou régression pas à pas, *stepwise regression*). Elle consiste à utiliser différentes combinaisons de variables explicatives dans une analyse en régression multiple de manière à arriver au meilleur modèle. Une régression en plusieurs étapes peut être menée de plusieurs manières et nous n'avons pas l'intention de couvrir ces techniques ici. L'idée générale est soit de démarrer avec un modèle contenant de nombreuses variables et de garder celles dont les p-valeurs sont en-dessous d'un certain niveau de significativité ou, alternativement, de partir d'un modèle contenant peu de variables et d'ajouter les variables qui ont des p-valeurs significatives. Parfois, des groupes de variables sont testées avec un test de Fisher. Malheureusement, le modèle final dépend souvent de l'ordre dans lequel les variables ont été éliminées ou ajoutées. [Pour en savoir plus sur les régressions en plusieurs étapes, voir Draper et Smith (1981).] En outre, ceci est une forme grave de *data mining*, et il est difficile d'interpréter les statistiques de Student et de Fisher dans le modèle final. On pourrait argumenter que la régression en plusieurs étapes automatise simplement ce que les chercheurs font de toute manière dans la recherche de différents modèles. Cependant, dans la plupart des études empiriques, une ou deux variables explicatives sont d'un intérêt prioritaire, et donc le but est de voir dans quelle mesure les coefficients de ces variables sont robustes à l'ajout ou au retrait d'autres variables, ou à la modification de la forme fonctionnelle.

En principe, il est possible d'intégrer les effets du data mining dans notre inférence statistique. En pratique, ceci est très difficile et rarement effectué, en particulier dans le travail empirique élaboré. [Voir Leamer (1983) pour une présentation intéressante de ce problème.] En revanche, nous pouvons essayer de recourir le moins possible au *data mining*, il faut pour cela ne pas explorer un grand nombre de modèles ou de méthodes d'estimation jusqu'à ce que un résultat significatif soit obtenu pour ensuite ne reporter que ce résultat. Si une variable est statistiquement significative dans seulement une petite partie des modèles estimés, il est fort probable que la variable n'a pas d'effet au niveau de la population.

19.5 RÉDIGER UN ARTICLE EMPIRIQUE

Rédiger un article qui présente une analyse économétrique est un grand défi, mais qui peut être aussi gratifiant. Un article réussi associe une analyse de données approfondie et convaincante à un exposé et des explications de qualité. Par conséquent, vous devez avoir une bonne connaissance de votre problème, une bonne compréhension des méthodes économétriques et de solides qualités de rédaction. Ne soyez pas découragés si vous trouvez qu'écrire un article empirique est difficile, la plupart des chercheurs professionnels ont passé de nombreuses années à apprendre à mener une analyse empirique et à présenter les résultats d'une manière convaincante.

Si le style d'écriture varie, la plupart des papiers suivent le même plan général. Dans les paragraphes suivants, nous proposons des pistes pour intituler les sections et nous expliquons ce que chaque section doit contenir. Ce sont seulement des suggestions qui ne doivent pas forcément être suivies de manière stricte. Dans l'article final, chaque section est numérotée, on peut éventuellement commencer par un pour l'introduction.

Introduction

L'introduction établit les objectifs de base de l'étude et explique pourquoi cette analyse est importante. Elle contient généralement une revue de la littérature qui présente ce qui a été fait et dans quelle direction les travaux précédents peuvent être améliorés. (Comme discuté dans la section 19.2, une revue de la littérature fouillée peut être proposée dans une section séparée.) Présenter des statistiques simples ou des graphes qui révèlent une relation apparemment paradoxale est une bonne façon d'introduire le sujet de l'étude. Par exemple, supposons que vous écriviez un travail sur les facteurs affectant la fécondité dans un pays en voie de développement, avec une attention particulière sur les niveaux d'éducation des femmes. Une manière intéressante d'introduire le sujet serait de présenter un tableau et des graphiques montrant que la fécondité a baissé dans le temps et une explication succincte de la manière dont vous espérez analyser les facteurs qui ont contribué à ce déclin. À ce stade, vous pouvez déjà savoir que, *ceteris paribus*, les femmes plus éduquées ont moins d'enfants et que les niveaux d'éducation moyens ont augmenté à travers le temps.

La plupart des chercheurs aiment bien résumer les résultats de leur article dans l'introduction. Cela peut être une bonne façon de capturer l'attention du lecteur. Par exemple, vous pouvez affirmer que votre meilleure valeur estimée de l'effet du fait de manquer 10 sur 30 heures de cours durant un semestre est à peu près d'un demi-point. Néanmoins, le résumé ne doit pas être trop précis parce que ni les méthodes, ni les données utilisées pour obtenir la valeur estimée n'ont encore été introduites.

Structure conceptuelle (ou théorique)

Dans cette section, vous décrivez l'approche générale pour répondre à la question que vous avez posée. Il peut s'agir de théorie économique formelle, mais il s'agit souvent d'une explication intuitive des problèmes conceptuels qui se posent lors de cette analyse.

À titre d'exemple, supposons que vous étudiez les effets des débouchés économiques et de la sévérité des peines sur le comportement criminel. Afin d'expliquer la participation à une activité criminelle, on peut spécifier un problème de maximisation d'utilité dans lequel l'individu choisit combien de temps il consacre à des activités légales et illégales, étant donné les taux de salaire dans chacune des deux activités, ainsi que les variables mesurant la probabilité d'être condamné et la sévérité des peines pour les activités criminelles. Un tel exercice sert à suggérer quelles variables doivent être inclues dans l'analyse empirique et à donner une idée de comment les variables doivent apparaître dans le modèle économétrique (sans l'indiquer de façon précise).

Dans la plupart des cas, il n'est pas nécessaire d'écrire explicitement une théorie économique. Pour l'analyse économétrique de politique économique, le bon sens est généralement suffisant pour spécifier un modèle. Supposons par exemple que vous vous intéressiez à l'estimation des effets de la participation au programme *Aid to Families with Dependent Children* (AFDC)[3] sur les effets de la performance à l'école de l'enfant. L'AFDC apporte un revenu complémentaire, mais facilite également l'accès à *Medicaid*[4] et à d'autres transferts. La partie difficile d'une telle analyse est de décider l'ensemble des variables dont il faut neutraliser l'effet. Dans cet exemple, nous pourrions tenir compte du revenu familial (incluant l'AFDC et tout autre aide financière), l'éducation de la mère, si la famille vit dans une zone urbaine, et d'autres variables. Dans ce cas, l'inclusion d'un indicateur de participation à l'AFDC mesure (idéalement) les bénéfices non monétaires de l'AFDC. Expliquer quels facteurs il faut prendre en compte et par quels mécanismes l'AFDC pourrait améliorer la performance scolaire se substitue à une théorie économique formelle.

Modèles économétriques et méthodes d'estimation

Il est très utile d'avoir une section qui contient une série d'équations du type de celles que vous estimez et de les présenter dans la section « résultats » du papier. Ceci vous permet d'être clair sur la variable qui est la variable-clé et les facteurs dont vous allez neutraliser l'effet. L'écriture d'équations contenant des termes d'erreur vous permet de discuter la question de la fiabilité des MCO comme méthode d'estimation.

La distinction entre le modèle et une méthode d'estimation doit être effectuée dans cette section. Un modèle représente une relation de la population (définie grossièrement pour introduire des équations en séries temporelles). Par exemple vous devriez écrire

$$colGPA = \beta_0 + \beta_1 alcohol + \beta_2 hsGPA + \beta_3 SAT + \beta_4 female + u \qquad \textbf{[19.1]}$$

pour décrire la relation entre les notes à l'université (*colGPA*) et la consommation d'alcool (*alcohol*), avec un certain nombre de variables dont on veut tenir compte dans l'équation. En toute logique, cette équation représente une population, telle que celle de tous les étudiants en Licence dans une université particulière. Il n'y a pas de chapeau sur les β_j ou sur *colGPA* parce que nous ne connaissons pas (et nous ne connaîtrons jamais) ces nombres. Plus tard, nous les estimerons. Dans cette section, n'anticipez pas la présentation de vos résultats empiriques. En d'autres termes, ne commencez pas avec un modèle général pour ensuite écrire que vous avez omis certaines variables parce qu'elles s'avéraient être non significatives. De telles discussions devraient être maintenues dans la section « résultats ».

Un modèle de séries temporelles qui relie, au niveau de la ville, les vols de voitures au taux de chômage et au taux de condamnation pourrait ressembler à

$$\begin{aligned} thefts_t = {} & \beta_0 + \beta_1 unem_t + \beta_2 unem_{t-1} + \beta_3 cars_t + \beta_4 convrate_t \\ & + \beta_5 convrate_{t-1} + u_t, \end{aligned} \qquad \textbf{[19.2]}$$

3 Note de la traduction : aide pour les familles avec enfants dépendants.

4 Note de la traduction : Medicaid est un programme de politique publique dont le but est de fournir une assurance maladie aux ménages à bas revenus.

où l'indice t est utile pour insister sur la dynamique dans cette équation (dans ce cas, on permet au chômage et au taux de condamnation pour vol d'automobile d'avoir des effets retardés).

Après avoir spécifié un modèle ou des modèles, il est approprié de discuter des méthodes d'estimation. Dans la plupart des cas, ce sera MCO, mais, par exemple, dans une équation en séries temporelles, vous pourriez utiliser MCQG pour procéder à une correction pour la corrélation sérielle (comme dans le chapitre 12). Cependant, la méthode pour estimer un modèle est bien distincte du modèle lui-même. Il n'est pas approprié, par exemple, de parler d'un « modèle MCO ». Les moindres carrés ordinaires sont une méthode d'estimation, de même que les moindres carrés pondérés, Cochrane-Orcutt, et ainsi de suite. Il y a habituellement différentes manières d'estimer n'importe quel modèle. Vous devriez expliquer pourquoi la méthode que vous avez choisie est appropriée.

Toutes les hypothèses qui sont utilisées pour obtenir un modèle économétrique estimable à partir d'un modèle économique sous-jacent doivent être discutées clairement. Par exemple, dans l'exemple sur la qualité de l'école secondaire mentionné dans la section 19.1, le problème central dans l'analyse est de savoir comment mesurer la qualité de l'école. Doit-elle être basée sur les scores moyens SAT, le pourcentage d'élèves accédant à l'université, le ratio étudiants – professeurs, le niveau d'éducation moyen des enseignants, une certaine combinaison de ces variables, ou éventuellement d'autres mesures ?

Nous devons toujours faire des hypothèses à propos de la forme fonctionnelle, qu'il y ait un modèle théorique qui ait été présenté ou non. Comme vous le savez, les modèles à élasticité constante ou à semi-élasticité constante sont intéressants parce que les coefficients sont faciles à interpréter (comme effets en pourcentage). Il n'y a pas de règles strictes sur la manière de choisir la forme fonctionnelle, mais les directives discutées dans la section 6.2 semblent fonctionner correctement en pratique. Vous n'avez pas besoin d'une discussion complète sur la forme fonctionnelle, mais il est utile de mentionner si vous estimez des élasticités ou une semi-élasticité. Par exemple, si vous estimez l'effet d'une certaine variable sur le salaire ou la rémunération, la variable dépendante sera presque toujours en forme logarithmique, et vous pouvez inclure celle-ci aussi dans n'importe quelle équation depuis le début. Vous ne devez pas présenter chacune, et même la plupart, des variations de forme fonctionnelle que vous reporterez plus tard dans la section « résultats ».

Souvent, les données utilisées en économie empirique sont au niveau de la ville ou du comté. Par exemple, supposons que pour la population des villes de petite à moyenne taille, vous désiriez tester l'hypothèse selon laquelle posséder une équipe de base-ball de ligue mineure entraîne un taux de divorce plus faible pour la ville. Dans ce cas, vous devez tenir compte du fait que les plus grandes villes auront plus de divorces. Une manière de tenir compte de la taille de la ville est de remettre les divorces à l'échelle de la population adulte ou de la population de la ville. Par conséquent, un modèle raisonnable est :

$$\log(div/pop) = \beta_0 + \beta_1 mlb + \beta_2 perCath + \beta_3 \log(inc/pop) + other\ factors, \qquad [19.3]$$

où *mlb* est une variable muette égale à un si la ville a une équipe de base-ball en ligue mineure et *perCath* est le pourcentage de la population qui est catholique (donc un nombre tel que 34,6 signifie 34,6 %). Notez que *div/pop* est un taux de divorce, qui est généralement plus facile à interpréter que le nombre absolu de divorces.

Une autre manière de tenir compte de la population est d'estimer le modèle

$$\log(div) = \gamma_0 + \gamma_1 mlb + \gamma_2 perCath + \gamma_3 \log(inc) + \gamma_4 \log(pop) + other\ factors. \qquad [19.4]$$

Le paramètre d'intérêt, γ_1, lorsqu'il est multiplié par 100, donne la différence en pourcentage entre les taux de divorce en gardant la population, le pourcentage de catholiques, le revenu, et n'importe quel autre facteur, constants. Dans l'équation (19.3), β_1 mesure l'effet en pourcentage du base-ball en ligue mineure, qui

peut varier soit parce que le nombre de divorces change, soit parce que la population change. En utilisant le fait que $\log(div/pop) = \log(div) - \log(pop)$ et $\log(inc/pop) = \log(inc) - \log(pop)$, on peut réécrire (19.3) comme

$$\log(div) = \beta_0 + \beta_1 mlb + \beta_2 perCath + \beta_3 \log(inc) + (1 - \beta_3) \log(pop) + other\ factors.$$

Ce qui montre que (19.3) est un cas spécial de (19.4) avec $\gamma_4 = (1 - \beta_3)$ et $\gamma_j = \beta_j$, $j = 0,1,2,3$. Alternativement, (19.4) est équivalent au fait d'ajouter la population comme variable explicative supplémentaire à (19.3). Cela facilite le fait de tester pour en effet séparé de la population sur le taux de divorce.

Si vous utilisez une méthode d'estimation plus élaborée, tel que les doubles moindres carrés, vous devez fournir une série de raisons de l'utiliser. Si vous utilisez les DMC, vous devez proposer une discussion approfondie des raisons pour lesquelles vos choix en termes d'instruments pour la variable (ou les variables) explicative(s) endogène(s) sont valables. Comme on l'a mentionné dans le chapitre 15, il y a deux conditions pour qu'une variable soit considérée comme une bonne VI. Tout d'abord, elle doit être omise et exogène par rapport à l'équation d'intérêt (équation structurelle). Ceci est une hypothèse. Deuxièmement, elle doit avoir une corrélation partielle avec la variable explicative endogène. Ceci peut être testé. Par exemple, dans l'équation, vous pourriez utiliser une variable binaire qui indique qu'un étudiant vit dans un dortoir comme une VI pour la consommation d'alcool. Cela exige que la situation de vie n'a pas d'impact direct sur *colGPA* – si bien qu'elle est omise de (19.1) – et qu'elle est non corrélée avec les facteurs inobservés dans *u* qui ont un effet sur *colGPA*. Nous devrions également vérifier que *dorm* est partiellement corrélé avec *alcohol* en régressant *alcohol* sur *dorm*, *hsGPA*, *SAT*, et *female*. (Voir le chapitre 15 pour plus de détails.)

Vous pouvez tenir compte du problème des variables omises (ou de l'hétérogénéité omise) en utilisant des données de panel. Une fois de plus, c'est facilement décrit en écrivant une équation ou deux. En fait, il est utile de montrer comment différencier les équations dans le temps pour ôter les effets non observables constants dans le temps ; ceci mène à une équation qui peut être estimée par MCO ; alternativement, si vous utilisez une estimation avec effets fixes, il suffit simplement de l'expliquer.

Comme exemple simple, supposons que vous testiez si des taux de taxation au niveau du comté plus élevés réduisent l'activité économique, mesurée par la production manufacturière par tête. Supposons que pour les années 1982, 1987, et 1992, le modèle est

$$\log(manuf_{it}) = \beta_0 + \delta_1 d87_t + \delta_2 d92_t + \beta_1 tax_{it} + \ldots + a_i + u_{it},$$

où $d87_t$ et $d92_t$ sont des variables temporelles et tax_{it} est le taux de taxation pour le comté i au temps t (en pourcentage). Nous aurions d'autres variables qui varient à travers le temps dans l'équation, y compris des mesures du coût des affaires (telles que les salaires moyens), des mesures de la productivité du travailleur (mesurée par l'éducation moyenne), et ainsi de suite. Le terme a_i est l'effet fixe, contenant tous les facteurs qui ne varient pas dans le temps, et u_{it} est le terme d'erreur idiosyncratique. Pour ôter a_i, nous pouvons soit prendre la différence entre les années ou utiliser la différence de moyenne dans le temps (la transformation par effets fixes).

Les données

Vous devriez toujours avoir une section qui décrit de manière approfondie les données utilisées dans l'analyse empirique. C'est particulièrement important si vos données sont non standards et n'ont pas été beaucoup utilisées par d'autres chercheurs. Une information suffisante devrait être présentée de manière à ce qu'un chercheur puisse, en principe, obtenir les données et refaire votre analyse. En particulier, toutes les sources de données disponibles publiquement devraient être inclues dans les références, et les bases de données succinctes doivent être reportées dans une annexe. Si vous utilisez votre propre enquête pour collecter des données, une copie du questionnaire doit être présentée dans une annexe.

Parallèlement à une discussion des sources de données, assurez-vous de commenter les unités de chacune des variables (par exemple, est-ce que le revenu est mesuré en centaines ou en milliers de dollars ?). Inclure un tableau des définitions de variables est aussi très utile pour le lecteur. Les noms dans le tableau devraient correspondre aux noms utilisés pour décrire les résultats économétriques dans la section suivante.

Il est aussi très informatif de présenter un tableau des statistiques descriptives, tel que les valeurs maximales et minimales, les moyennes, les écarts-types pour chaque variable. Un tel tableau facilite l'interprétation des valeurs estimées des coefficients dans la section suivante, et cela permet d'insister sur les unités de mesure des variables. Pour les variables binaires, la seule statistique de base nécessaire est la proportion de « uns » dans l'échantillon (ce qui est le même que la moyenne d'échantillon). Pour les variables présentant une tendance, des éléments comme les moyennes sont moins intéressants. Il est souvent utile de calculer le taux de croissance moyen d'une variable à travers les années dans votre échantillon.

Vous devriez toujours expliquer clairement combien d'observations vous avez. Pour les bases de données en séries temporelles, identifiez les années que vous utilisez dans votre analyse, incluez une description de toutes les périodes particulières de l'histoire (telles que la seconde Guerre Mondiale). Si vous utilisez des données de panel ou des coupes transversales empilées, assurez-vous de reporter combien d'unités en coupe transversale (personnes, villes, et ainsi de suite) vous possédez pour chaque année.

Résultats

La section « résultats » devrait inclure vos valeurs estimées de tous les modèles présentés dans la section « modèles ». Vous pouvez débuter avec une analyse très simple. Par exemple, supposons que le pourcentage d'étudiants d'une classe terminale (*percoll*) fréquentant l'université soit utilisé comme une mesure de la qualité de l'école secondaire que cette personne a fréquentée. Dans ce cas, une équation à estimer est

$$\log(wage) = \beta_0 + \beta_1 percoll + u.$$

Bien entendu, ceci ne tient pas compte de l'effet d'autres facteurs qui peuvent déterminer les salaires et qui peuvent être corrélés avec *percoll*. Mais une analyse simple peut attirer l'attention du lecteur sur une analyse plus élaborée et révéler l'importance de neutraliser l'effet des autres facteurs.

Si seulement quelques équations sont estimées, vous pouvez présenter les résultats sous la forme d'une équation avec les écarts-types entre parenthèses en-dessous des coefficients estimés. Si votre modèle inclut plusieurs variables explicatives et si vous présentez quelques variations du modèle général, il est préférable de donner les résultats sous une forme de tableau plutôt que sous la forme d'une équation. La plupart des papiers devraient avoir au moins un tableau, qui devrait toujours inclure au moins le R-carré et le nombre d'observations pour chaque équation. D'autres statistiques, telles que le R-carré ajusté, peuvent aussi être présentées.

La chose la plus importante est de discuter l'interprétation et la force de vos résultats empiriques. Est-ce que les coefficients ont les signes attendus ? Est-ce qu'ils sont statistiquement significatifs ? Si un coefficient est statistiquement significatif mais présente un signe contre-intuitif, est-ce que cela peut être vrai ? Cela peut révéler un problème de données ou de méthode économétrique (par exemple, les MCO peuvent être inappropriés à cause de problèmes de variables omises).

Assurez-vous de décrire l'ampleur des coefficients des variables explicatives importantes. Souvent, une ou deux variables de politique sont centrales dans l'étude. Leurs signes, tailles, et significativités statistiques doivent être traités en détails. Veillez à faire une distinction entre la significativité statistique et économique. Si la statistique t est petite, est-ce parce que le coefficient est petit ou parce que son écart-type est élevé ?

En plus de discuter les valeurs estimées du modèle le plus général, vous pouvez fournir les cas spéciaux intéressants, en particulier ceux nécessaires pour tester certaines hypothèses multiples. Par exemple,

dans une étude visant à déterminer le différentiel de salaire entre industries, vous pouvez présenter l'équation sans les variables muettes industrielles ; cela permet aux lecteurs de tester aisément si les différentiels entre industries sont significatifs statistiquement (en utilisant la forme du R-carré du test en F). N'hésitez pas à éliminer différentes variables pour trouver la « meilleure » combinaison des variables explicatives. Comme nous l'avons mentionné précédemment, c'est une tâche difficile et à vrai dire pas très bien définie. Cela est important seulement si l'élimination d'une série de variables modifie de manière substantielle les tailles et la significativité des coefficients d'intérêt. L'élimination d'un groupe de variables – telles que des interactions ou des termes quadratiques – pour simplifier le modèle peut être justifiée via un test en F.

Si vous avez utilisé au moins deux méthodes différentes – telles que les MCO ou les DMC, ou des séries temporelles en niveaux ou en différences, la différenciation avec des données panel ou des MCO empilés –, vous devez alors commenter toutes les différences importantes. Si les MCO donnent des résultats contre-intuitifs, est-ce que les DMC ou les méthodes en données de panel améliorent les valeurs estimées ? Ou, est-ce l'inverse qui se passe ?

Conclusions

Ceci peut être une section courte qui résume ce que vous avez appris. Par exemple, vous pouvez présenter la taille d'un coefficient qui était d'un intérêt particulier. La conclusion doit aussi discuter les précautions par rapport aux conclusions tirées, et peut même proposer des directions pour la recherche ultérieure. Il est utile d'imaginer que les lecteurs vont d'abord se tourner vers la conclusion avant de décider s'ils vont lire le reste du papier.

Conseils de styles

Vous devez donner un titre à votre papier qui reflète son sujet, mais assurez-vous que le titre n'est pas trop long au point d'être difficile à lire. Le titre devrait être sur une page de titre séparée qui inclut aussi votre nom, l'affiliation, et – si c'est approprié – le numéro du cours. La page de titre doit inclure aussi un court résumé, et le résumé peut être aussi proposé sur une page séparée.

Les articles doivent être tapés en double espace. Toutes les équations doivent commencer sur une nouvelle ligne, et doivent être centrées et numérotées de manière consécutive, à savoir, (1), (2), (3), et ainsi de suite. Les figures et les tableaux importants peuvent être inclus après le corps du texte. Dans le texte, citez les articles par auteur et par date, par exemple, White (1980). La section bibliographique à la fin du papier doit être écrite en format standard. Un certain nombre d'exemples sont donnés en référence à la fin du texte.

Quand vous introduisez une équation dans la section consacrée aux modèles économétriques, vous devez décrire les variables importantes : la variable dépendante et la ou les variable(s) indépendante(s) importante(s). Pour se focaliser sur une seule variable indépendante, vous pouvez écrire une équation, telle que

$$GPA = \beta_0 + \beta_1 alcohol + \boldsymbol{x\delta} + u$$

Ou

$$\log(wage) = \beta_0 + \beta_1 educ + \boldsymbol{x\delta} + u,$$

où la notation $\boldsymbol{x\delta}$ est un résumé pour différentes variables explicatives supplémentaires. À ce stade, vous devez seulement décrire ces variables de manière générale ; elles peuvent être décrites de manière spécifique dans la section « données » dans un tableau. Par exemple, dans une étude des facteurs influençant les salaires des cadres, vous pourriez inclure un tableau comme le Tableau 19.1.

Un tableau de statistiques synthétiques, obtenu à partir du Tableau I dans Papke et Wooldridge (1996) et similaire aux données de la base de données 401K, peut être établi comme illustré dans le Tableau 19.2.

Dans la section résultats, vous pouvez écrire les coefficients estimés, soit sous une forme équation, comme nous l'avons fait souvent, soit dans un tableau. En particulier, quand plusieurs modèles ont été estimés avec différents ensembles de variables explicatives, les tableaux sont très utiles. Si vous écrivez la valeur estimée dans une équation, par exemple,

$$\widehat{\log(salary)} = 2{,}45 + 0{,}236\log(sales) + 0{,}008\,roe + 0{,}061\,ceoten$$
$$(0{,}93)\ (0{,}115)\qquad (0{,}003)\qquad (0{,}028)$$
$$n = 204,\ R^2 = 0{,}351,$$

assurez-vous de mentionner aux alentours de la première équation que les écarts-types sont entre parenthèses. Il est acceptable de reporter les statistiques t pour tester $H_0 : \beta_j = 0$, ou leur valeur absolue, mais il est très important de mentionner ce que vous faites.

Tableau 19.1 *Description des variables*

salary	Salaire annuel (incluant les bonus) en 1990 (en milliers)
sales	Ventes de la firme en 1990 (en millions)
roe	Rendement moyen sur action, 1988-1990 (en pour cent)
pcsal	Changement en pourcentage du salaire, 1988-1990
pcroe	Changement en pourcentage du roe, 1988-1990
indust	= 1 si une société industrielle, 0 sinon
finance	= 1 si une société financière, 0 sinon
consprod	= 1 si une société de produits de consommation, 0 sinon
util	= 1 si une société d'équipements, 0 sinon
ceoten	Nombre d'années comme PDG de la société

Tableau 19.2 *Statistiques de synthèse*

Variable	Moyenne	Écart-type	Minimum	Maximum
prate	0,869	0,167	0,023	1
mrate	0,746	0,844	0,011	5
employ	4 621,01	16 299,64	53	443 040
age	13,14	9,63	4	76
sole	0,415	0,493	0	1
Nombre d'observations = 3 784				

Si vous reportez vos résultats sous la forme d'un tableau, assurez-vous que les variables dépendantes et indépendantes sont clairement indiquées. Une fois de plus, précisez si les écarts-types ou les statistiques *t* sont en-dessous des coefficients (la dernière option est préférée). Certains auteurs préfèrent utiliser des astérisques pour indiquer la significativité statistique à différents niveaux de significativité (par exemple, une étoile signifie significatif à 5 %, deux étoiles signifient significatif à 10 % mais pas à 5 %, et ainsi de suite). Ce n'est pas nécessaire si vous discutez précisément la significativité des variables explicatives dans le texte.

Un exemple de tableau de résultats, tiré de Papke et Wooldridge (1996), est reporté dans le tableau 19.3.

Tableau 19.3 *Résultats MCO. Variable Dépendante : Taux de participation*

Variables Indépendantes	(1)	(2)	(3)
mrate	0,156 (0,012)	0,239 (0,042)	0,218 (0,342)
$mrate^2$	–	– 0,087 (0,043)	– 0,096 (0,073)
log(*emp*)	– 0,112 (0,014)	– 0,112 (0,014)	– 0,098 (0,111)
$\log(emp)^2$	0,0057 (0,0009)	0,0057 (0,0009)	0,0052 (0,0007)
age	0,0060 (0,0010)	0,0059 (0,0010)	0,0050 (0,0021)
age^2	– 0,00007 (0,00002)	– 0,00007 (0,00002)	– 0,00006 (0,00002)
sole	– 0,0001 (0,0058)	0,0008 (0,0058)	0,0006 (0,0061)
constant	1,213 (0,051)	0,198 (0,052)	0,085 (0,041)
Muettes industries ?	non	non	oui
Observations *R*-carrés	3 784 0,143	3 784 0,152	3 784 0,162

Note : les valeurs entre parenthèses en-dessous des coefficients estimés sont les écarts-types.

Vos résultats devront être faciles à lire et à interpréter si vous choisissez les unités de vos variables indépendantes et de la variable dépendante de telle manière à ce que les coefficients ne soient pas trop élevés ou trop petits. Vous ne devez jamais reporter des nombres tels que 1,051e – 007 ou 3,524e + 006 pour vos coefficients ou vos écarts-types, et vous ne devez pas utiliser la notation scientifique. Si les coefficients sont soit extrêmement petits, soit extrêmement larges, remettez à l'échelle les variables indépendantes et dépendante, comme nous en avons discuté dans le chapitre 6. Vous devez limiter le nombre de chiffres reportés après la virgule décimale de manière à ne pas donner une impression erronée de précision. Par exemple, si votre logiciel de régression estime un coefficient à 0,54821059, vous devez reporter cela comme 0,548 ou même comme 0,55 dans l'article.

En principe, les commandes que votre logiciel économétrique particulier utilise pour produire les résultats ne doivent pas apparaître dans le papier ; seuls les résultats sont importants. Si une commande spéciale a été utilisée pour mener à bien une certaine méthode d'estimation, ceci peut être reporté dans une annexe. Une annexe est aussi un bon endroit pour inclure des résultats supplémentaires qui vont dans le sens de votre analyse mais qui ne sont pas centraux par rapport à elle.

RÉSUMÉ

Dans ce chapitre, nous avons discuté des ingrédients d'une étude empirique réussie et nous avons fourni des conseils qui peuvent améliorer la qualité de l'analyse. Finalement, le succès de toute étude dépend spécialement du soin et des efforts consentis.

MOTS-CLÉS

Analyse de mauvaise spécification p. 783
Analyse de sensibilité p. 783
Bases de données en ligne p. 778
Data Mining p. 784
Éditeur de texte p. 779
Fichier Texte (ASCII) p. 778
Internet p. 775
Services de recherche en ligne p. 776
Tableur p. 779

ÉCHANTILLON DE PROJETS EMPIRIQUES

Dans le texte, nous avons vu des exemples d'analyse économétrique qui, soit proviennent de, soit sont motivés par des travaux publiés. Nous espérons que ceux-ci vous ont donné une bonne idée sur la portée d'une analyse empirique. Nous incluons dans la liste suivante des exemples supplémentaires de questions que d'autres ont trouvées ou sont susceptibles de trouver intéressantes. Ils ont pour but de stimuler votre imagination ; il n'y a pas d'ambition ici à fournir tous les détails de modèles spécifiques, d'exigences de données, ou de méthodes d'estimation alternatives. Il doit être possible de terminer ce projet en un semestre.

1. Menez à bien votre propre enquête de campus pour répondre à une question qui intéresse votre université. Par exemple : quel est l'effet du travail sur les notes moyennes en Bachelier ? Vous pouvez interroger les étudiants à propos de leur résultats moyens à l'école secondaire, le résultat moyen au bac, les scores ACT ou SAT, les heures de travail par semaine, la participation aux activités de sports, la branche principale, le genre, l'origine ethnique, et ainsi de suite.[5] Ensuite, utilisez ces variables pour créer un modèle qui explique la note moyenne. Quel effet, s'il y en a, exerce une heure supplémentaire travaillée par semaine sur le score moyen ? Un problème ici est que les heures travaillées peuvent être endogènes : cela peut être corrélé avec des facteurs inobservés qui influencent le score moyen au BAC, ou un score moyen plus faible peut amener les étudiants à travailler plus.

Une meilleure approche serait de collecter le score moyen cumulé avant le semestre et ensuite d'obtenir le score moyen du semestre le plus récent, en même temps que la quantité travaillée durant ce semestre,

5 Note de l'éditeur : les scores ACT (American College Testing) et SAT sont des résultats de tests standardisés menés à la fin de l'école secondaire aux États-Unis et utilisés dans les critères de sélection d'admission dans les universités américaines.

et les autres variables. Dans ce cas, le score moyen cumulé peut être utilisé comme une variable explicative de contrôle dans l'équation.

2. Il y a beaucoup de variantes sur le sujet précédent. Vous pouvez étudier les effets de la consommation de drogue ou d'alcool, ou le fait de vivre dans une communauté, sur la note moyenne scolaire. Vous pourrez alors tenir compte de l'effet de nombreuses variables liées au contexte familial, ainsi que des variables sur la performance passée.

3. Est-ce que les lois sur le contrôle des armes appliquées au niveau de la ville réduisent le nombre de crimes violents ? De telles questions peuvent être difficiles à traiter avec une coupe transversale unique parce que les lois en application au niveau de la ville et de l'État sont souvent endogènes. [Voir Kleck et Patterson (1993) pour un exemple. Ils utilisent des données en coupe transversale et des méthodes de variables instrumentales, mais leurs VIs sont discutables.] Les données de panel peuvent être très utiles pour inférer la causalité dans ce contexte. Au minimum, vous pourriez contrôler pour l'effet du taux de criminalité violente d'une année précédente.

4. Low et McPheters (1983) utilisent des données transversales sur les taux de salaire et estiment le risque de décès des officiers de police, en même temps que l'effet d'autres variables dont ils veulent neutraliser l'effet. L'idée est de déterminer si les officiers de police ont des compensations salariales pour le fait de travailler dans des villes avec un risque plus élevé de blessures au travail ou de décès.

5. Est-ce que les lois sur le consentement parental augmentent le taux de naissance chez les adolescents ? Vous pouvez utiliser des données au niveau de l'État pour mener cette analyse : soit une série temporelle pour un État donné, ou, même mieux, des données de panel par État. Est-ce que les mêmes lois réduisent les taux d'avortement parmi les adolescents ? Le *Statistical Abstract of the United States* contient toutes sortes de données au niveau de l'État. Levine, Trainor, et Zimmerman (1996) ont étudié les effets des restrictions sur le financement de l'avortement sur des résultats similaires. D'autres facteurs, tels que l'accès aux avortements, peuvent influencer le taux de naissance chez les adolescentes et les taux d'avortement.

6. Est-ce que les modifications du code de la route affectent le nombre de victimes de la circulation routière ? McCarthy (1994) inclut une analyse de données en séries temporelles mensuelles pour l'État de Californie. Un ensemble de variables muettes peut être utilisé pour indiquer les mois durant lesquels certaines lois ont été en place. Le fichier TRAFFIC2 contient les données utilisées par McCarthy (1994). Une alternative est d'obtenir un ensemble de données de panel sur les États aux États-Unis, dans laquelle vous pouvez exploiter la variation au niveau des règlements entre les États, ainsi que dans le temps. Freedman (2007) est un bon exemple d'une analyse au niveau de l'État ; celle-ci utilise 25 années de données qui retracent les modifications des lois sur la conduite en état d'ivresse, le port de la ceinture de sécurité, et les limitations de vitesse dans différents États. Les données peuvent être trouvées dans le fichier DRIVING.

Mullahy et Sindelar (1994) utilisent des données au niveau individuel qui correspondent aux lois de l'État et aux taxes sur l'alcool pour estimer les effets des lois et des taxes sur la probabilité de conduite en état d'ivresse.

7. Est-ce que les gens d'origine ethnique sont discriminés négativement sur le marché du crédit ?

Hunter et Walker (1996) ont abordé cette question ; en fait, nous avons utilisé leurs données dans les Exercices d'ordinateur C.8 du Chapitre 7 et C.2 du Chapitre 17.

8. Y a-t-il une prime liée au mariage pour les athlètes professionnels ? Korenman et Neumark (1991) ont trouvé une prime de salaire significative pour les hommes mariés après avoir utilisé une série de méthodes économétriques, mais leur analyse est limitée, parce qu'ils ne peuvent pas observer directement la productivité. (De plus, Korenman et Neumark ont restreint leur analyse à des hommes ayant certains métiers.) Les athlètes professionnels fournissent un groupe intéressant pour lequel on peut étudier la prime de mariage parce que

l'on peut facilement collecter des données sur différentes mesures de productivité, en plus du salaire. La base de données concernant les joueurs de la National Basketball Association (NBA) est un exemple. Pour chaque joueur, nous avons des informations sur les points marqués, les rebonds, les passes décisives, le temps de jeu, et la démographie. Comme dans l'exercice d'ordinateur C.9 dans le chapitre 6, on peut utiliser une analyse de régression multiple pour tester si les mesures de productivité diffèrent par statut marital. On peut aussi utiliser ce type de données pour tester si les hommes mariés sont mieux payés après avoir tenu compte des différences de productivité. (Par exemple, les propriétaires de la NBA peuvent penser que les hommes mariés apportent de la stabilité à l'équipe, ou sont plus favorables pour l'image de l'équipe.) Pour les sports individuels – tels que le golf et le tennis – les rémunérations annuelles reflètent directement la productivité. De telles données, de même que l'âge et l'expérience, sont relativement faciles à collecter.

9. Répondez à cette question : est-ce que les fumeurs de cigarettes sont moins productifs ? Une variante de cette question est : est-ce que les travailleurs qui fument ont plus de jours de maladies (toutes choses égales par ailleurs) ? Mullahy et Portney (1990) utilisent des données au niveau individuel pour évaluer cette question. On pourrait utiliser, disons, des données au niveau de la métropole. Quelque chose comme la productivité moyenne dans le secteur manufacturier peut être relié au pourcentage de fumeurs du secteur manufacturier. On peut tenir compte de l'effet d'autres variables, telles que l'éducation moyenne du travailleur, le capital par travailleur, et la taille de la ville (vous pouvez imaginer encore plus de choses).

10. Est-ce que le salaire minimum diminue le taux de pauvreté ? Vous pouvez utiliser des données au niveau de l'État ou du comté pour répondre à cette question. L'idée est que le salaire minimum varie entre les États parce que certains États ont des minimums plus élevés que le minimum fédéral. En outre, il y a des changements dans le temps dans le minimum nominal à l'intérieur d'un État, certains dus au changement au niveau fédéral et d'autres à cause des changements au niveau de l'État. Neumark et Wascher (1995) ont utilisé des données de panel sur les États pour estimer les effets du salaire minimum sur les taux d'emploi des jeunes travailleurs, ainsi que sur les taux de participation scolaire.

11. Quels facteurs influencent la performance de l'étudiant dans les écoles publiques ? Il est relativement aisé d'obtenir des données au niveau de l'école ou au moins au niveau du district dans la plupart des États. Est-ce que la dépense par étudiant a une influence ? Est-ce que les rapports étudiants – enseignants ont un effet quelconque ? Il est difficile d'estimer les effets *ceteris paribus* parce que la dépense est liée à d'autres facteurs, tels que les revenus familiaux ou les taux de pauvreté. La base de données, MEAP93, concernant les écoles secondaires du Michigan, contient une mesure des taux de pauvreté. Une autre possibilité est d'utiliser des données de panel, ou au moins de contrôler pour l'effet de la performance de l'année passée (tel qu'un résultat de tests moyens ou le pourcentage des étudiants ayant réussi un examen).

Vous pouvez examiner des facteurs moins évidents qui influencent la performance scolaire. Par exemple, après avoir tenu compte de l'effet du revenu, est-ce que la structure familiale a une influence ? Il est possible que les familles avec deux parents, avec un seul d'entre eux travaillant de manière rémunérée, aient un effet positif sur la performance. (Il pourrait y avoir au moins deux canaux : les parents passent plus de temps avec les enfants, et ils peuvent aussi faire du travail volontaire à l'école.) Qu'en est-il de l'effet des ménages avec un seul parent, en contrôlant pour le revenu et les autres facteurs ? Vous pouvez fusionner des données du recensement pour une ou deux années avec les données de district scolaire.

Est-ce que les écoles publiques pour lesquelles il y a des écoles privées dans le voisinage éduquent mieux leurs étudiants à cause de la concurrence ? C'est un problème de simultanéité difficile parce que les écoles privées sont probablement localisées dans des zones où les écoles publiques sont déjà pauvres. Hoxby (1994) a utilisé une approche par variables instrumentales, dans laquelle les proportions de la population de différentes religions sont des VIs pour le nombre d'écoles privées.

Rouse (1998) a étudié une question différente : est-ce que les étudiants qui ont été capables de fréquenter une école privée grâce au programme de repas du Milwaukee s'en sortent mieux que ceux qui ne

l'ont pas fait ? Elle a utilisé des données de panel et a été capable de tenir compte d'un effet étudiant non observé. Un sous-ensemble des données de Rouse est inclus dans le fichier VOUCHER.

12. Est-ce que les rendements excédentaires d'une action, ou d'un indice sur actions, peuvent être prédits par le rapport prix/dividende retardé ? Ou par les taux d'intérêt retardés ou la politique monétaire hebdomadaire ? Il serait intéressant de sélectionner un indice étranger sur actions, ou un des indices américains moins connu. Cochrane (1997) propose une revue intéressante des théories récentes et des résultats empiriques expliquant les rendements excédentaires sur action.

13. Y a-t-il une discrimination raciale dans le marché des cartes de base-ball ? Ceci implique de relier les prix des cartes de base-ball aux facteurs qui peuvent influencer leur prix, tels que les statistiques de la carrière, si le joueur a été dans le Hall of Fame, et ainsi de suite. En gardant les autres facteurs fixes, est-ce que les cartes des joueurs ayant des origines ethniques ou hispaniques se vendent avec une réduction ?

14. Vous pouvez tester si le marché des paris sur les sports est efficace. Par exemple, est-ce que la différence sur les parties de football ou de basket-ball contient toute l'information utilisable pour parier sur l'écart de score ? La base de données PNTSPRD contient de l'information sur les matchs masculins de basket-ball universitaire. La variable de résultat est binaire. Est-ce que la différence est couverte ou pas ? Ensuite, vous pouvez essayer de trouver l'information qui était connue avant chaque match à jouer de manière à prédire si l'écart est couvert. (Bonne chance !) Un site internet utile qui contient des écarts historiques et les résultats pour les parties de basketball et de football universitaires masculins est www.goldsheet.com.

15. Quel est l'effet, s'il y en a un, du sport au collège sur les autres aspects de l'université (les candidatures, la qualité des étudiants, la qualité des départements non sportifs) ? McCormick et Tinsley (1987) ont regardé les effets du succès en sport dans les collèges importants sur les variations des résultats SAT des étudiants entrant. La séquence chronologique est ici importante : en toute logique, c'est le succès passé récent qui affecte les candidatures contemporaines et la qualité de l'étudiant. On doit neutraliser l'effet de beaucoup d'autres facteurs – tels que les mesures de la qualité de l'école et du suivi des étudiants – pour rendre l'analyse convaincante parce que, sans tenir compte des autres facteurs, il y a une corrélation négative entre la performance académique et sportive. Une analyse plus récente du lien entre performance académique et sportive est proposée par Tucker (2004), qui regarde également dans quelle mesure les contributions des anciens élèves sont influencées par les résultats sportifs.

Une variante est de mettre ensemble des rivaux naturels en football ou en basket-ball masculin et de regarder les différences entre écoles comme une fonction de l'école qui a gagné la rencontre de football ou un ou plusieurs matchs de basket-ball. ATHLET1 et ATHLET2 sont des petites bases de données qui peuvent être étendues et mises à jour.

16. Collectez des taux de meurtre pour un échantillon de comtés (disons, à partir des *Uniform Crime Reports* du FBI) pour deux années. Faites en sorte que la dernière année soit telle que des variables économiques et démographiques sont faciles à obtenir à partir du *County and City Data Book*. Vous pouvez obtenir le nombre total des personnes sujettes à exécution pour les années en question au niveau du comté. Si les années sont 1990 et 1985, vous pouvez estimer

$$mrdrte_{90} = \beta_0 + \beta_1 mrdrte_{85} + \beta_2 executions + autres\ facteurs,$$

où l'intérêt réside dans le coefficient relatif à *executions*. Le taux de meurtre retardé et d'autres facteurs permettent de tenir compte de leur effet. Si plus de deux années de données sont obtenues, alors les méthodes de données de panel du chapitre 13 et 14 peuvent être utilisées.

D'autres facteurs peuvent aussi agir comme des repoussoirs du crime. Par exemple, Cloninger (1991) a présenté une analyse en coupe transversale des effets des réponses mortelles de la police sur les taux de criminalité.

Sur un plan différent, quels sont les facteurs qui affectent les taux de criminalité sur les campus universitaires ? Est-ce que la proportion d'étudiants qui vivent en communauté a un effet ? Est-ce que la taille des forces de police a une influence, ainsi que le type de police utilisée ? (Soyez prudents sur l'inférence de la causalité ici.) Est-ce que la mise en œuvre d'un programme d'escorte permet de réduire le crime ? Qu'en est-il des taux de criminalité dans les communautés voisines ? Récemment les collèges et les universités ont dû reporter les statistiques de criminalité ; durant les années précédentes, les transmissions de données étant volontaires.

17. Quels sont les facteurs qui influencent la productivité manufacturière au niveau de l'État ? En plus des niveaux de capital et d'éducation des travailleurs, vous pouvez regarder les taux de syndicalisation. Une analyse en données de panel serait des plus convaincantes ici, en utilisant des années multiples sur les données du recensement, disons en 1980, 1990, 2000, et 2010. Clark (1984) propose une analyse de la manière dont la syndicalisation influence la performance au niveau de l'entreprise et la productivité. Quelles autres variables peuvent expliquer la productivité ?

Les données au niveau de la firme peuvent être obtenues à partir de *Compustat*. Par exemple, en fixant d'autres facteurs, est-ce que les variations au niveau de la syndicalisation influencent le prix de l'action d'une firme ?

18. Utilisez des données au niveau de l'État – ou du comté –, ou si possible des données au niveau du district scolaire pour regarder les facteurs qui influencent la dépense d'éducation par étudiant. Une question intéressante est la suivante : toute autre chose égale par ailleurs (telle que le revenu ou les niveaux d'éducation des résidents), est-ce que les districts avec un plus grand pourcentage de personnes handicapées dépensent moins dans les écoles ? Les données de recensement peuvent être appariées avec les données de dépenses par district scolaire pour obtenir une coupe transversale de grande taille. Le ministère de l'éducation américain compile de telles données.

19. Quels sont les effets des règlements par État, tels que les lois sur le port du casque à moto, sur le nombre de victimes à moto ? Ou est-ce que les différences dans les règlements sur la navigation – tel que l'âge minimum requis pour la conduite – permettent d'expliquer les taux d'accidents en bateau ? Le ministère des Transports américain compile de telles informations. Celles-ci peuvent être fusionnées avec des données obtenues à partir du *Statistical Abstract of the United States*. Une analyse en données de panel semble possible ici.

20. Quels sont les facteurs qui affectent le taux de croissance de la production ? Deux facteurs d'intérêt sont l'inflation et l'investissement [voir par exemple, Blomström, Lipsey, et Zejan (1996)]. Vous pouvez utiliser des données en séries temporelles d'un pays que vous trouvez intéressant. Alternativement, vous pouvez utiliser une coupe transversale de pays, comme dans De Long et Summers (1991). Friedman et Kuttner (1992) trouvent que, au moins dans les années 80, l'écart entre le taux sur le papier commercial et le taux sur les obligations d'État influence la production réelle.

21. Quel est le comportement des fusions dans l'économie américaine (ou toute autre économie) ? Shughart et Tollison (1984) caractérisent (le log) des fusions annuelles dans l'économie américaine comme une marche aléatoire en montrant que la différence dans les logs – approximativement, le taux de croissance – est impossible à prédire étant donné les taux de croissance passés. Est-ce toujours valable ? Est-ce que c'est valable dans différentes industries ? Quelle mesure passée de l'activité économique peut être utilisée pour prédire les fusions ?

22. Quels sont les facteurs qui peuvent expliquer les différences ethniques et de genre dans les salaires et l'emploi ? Par exemple, Holzer (1991) a couvert les résultats sur « l'hypothèse de non-appariement spatial » pour expliquer les différences dans les taux d'emploi entre les noirs et les blancs. Korenman et Neumark (1992) ont examiné les effets de l'éducation des enfants sur les salaires des femmes, alors que Hersch et Stratton (1997) ont regardé les effets des responsabilités dans le ménage sur les salaires des hommes et des femmes.

23. Collectez des données mensuelles ou trimestrielles sur les taux d'emploi des adolescents, le salaire minimum, et les facteurs qui affectent l'emploi des adolescents pour estimer les effets du salaire minimum sur l'emploi des adolescents. Solon (1985) a utilisé des données américaines trimestrielles tandis que Castillo-Freeman et Freeman (1992) ont utilisé des données annuelles sur Porto Rico. Il peut être intéressant d'analyser des données en séries temporelles dans un État à bas salaires des États-Unis – les variations du salaire minimum sont susceptibles d'avoir l'effet le plus important.

24. Au niveau de la ville, estimez un modèle en séries temporelles pour la criminalité. Un exemple est Cloninger et Sartorius (1979). Comme alternative, vous pouvez estimer les effets de la politique de communauté ou des programmes de basket-ball de minuit – des innovations relativement récentes – dans la lutte contre la criminalité. L'identification de la causalité est difficile. Inclure une variable dépendante retardée peut être utile. Puisque vous utilisez des données en séries temporelles, vous devez être conscient du problème de régression fallacieuse.

25. Grogger (1990) a utilisé des données sur le nombre d'homicides journaliers pour estimer les effets de découragement exercés par la peine capitale. Peut-il y avoir d'autres facteurs – tels que l'information concernant la riposte mortelle de la part de la police – qui peuvent avoir un effet sur le nombre de crimes journaliers ?

26. Y a-t-il des effets de l'utilisation de l'ordinateur en termes de productivité moyenne ? Vous avez besoin de collecter des données en séries temporelles, sans doute au niveau national, sur la productivité, le pourcentage d'employés utilisant des ordinateurs, et d'autres facteurs. Qu'en est-il des dépenses en recherche et développement (probablement en proportion des ventes totales) ? Quels sont les facteurs sociologiques qui peuvent affecter la productivité, tels que par exemple la consommation d'alcool ou les taux de divorce ?

27. Quels sont les facteurs qui influencent les salaires des cadres exécutifs supérieurs ? Les fichiers CEOSAL1 et CEOSAL2 sont des bases de données qui incluent différentes mesures de performance de la firme ainsi que des informations telles que la permanence et le niveau d'éducation. Vous pouvez certainement mettre à jour ces fichiers de données et examiner d'autres facteurs intéressants. Rose et Shepard (1997) ont considéré la diversification de la firme comme un déterminant important de la rémunération du PDG.

28. Est-ce que les différences dans les codes fiscaux entre les États influencent le montant de l'investissement direct étranger ? Hines (1996) a étudié les effets des taxes sur l'entreprise au niveau de l'État, ainsi que la possibilité de souscrire à des crédits d'impôt étranger sur l'investissement fait à partir de l'extérieur des États-Unis.

29. Quels sont les facteurs qui influencent les résultats des élections ? Est-ce que les dépenses ont de l'importance ? Est-ce que les votes sur certaines matières spécifiques ont de l'importance ? Est-ce que l'état de l'économie locale a de l'importance ? Voir, par exemple Levitt (1994), et les bases de données VOTE1 et VOTE2. Fair (1996) a mené une analyse en séries temporelles sur les élections présidentielles américaines.

30. Tester ici si les magasins ou les restaurants pratiquent la discrimination par les prix basée sur la race ou l'origine ethnique. Graddy (1997) a utilisé des données sur les restaurants de restauration rapide dans le New Jersey et en Pennsylvanie, en même temps que les caractéristiques par zone définie à l'échelle du code postal, pour voir si les prix varient en fonction des caractéristiques de la population locale. Elle a trouvé que les prix des biens standard, tel que les sodas, augmentent lorsque la proportion de résidents d'origine ethnique augmente. (Ses données sont contenues dans le fichier DISCRIM.) Vous pouvez collecter des données similaires dans votre région en enquêtant dans les restaurants et dans les magasins au niveau des prix des biens standards et en les croisant avec les données récentes du recensement. Voir le papier de Graddy pour les détails de son analyse.

31. Menez votre propre étude « d'audit » pour tester la discrimination en termes d'origines ethniques ou de genre dans le recrutement. (Une telle étude est décrite dans l'exemple C.3. de l'annexe C.) Trouver des paires d'amis de même niveau de qualification, disons un homme et une femme, qui candidatent à des ouvertures

de postes dans des bars ou des restaurants locaux. Vous pouvez leur donner des curriculums vitae par téléphone qui reflètent chaque fois le même niveau d'expérience professionnelle, avec la seule différence étant le genre (ou l'origine ethnique). Ensuite, vous pouvez noter qui obtient les interviews et les offres de travail. Neumark (1996) a décrit une telle étude menée à Philadelphie. Une variante serait de tester si l'attractivité physique générale ou une caractéristique spécifique, tel que le fait d'être obèse, avoir des tatouages visibles ou des piercings sur le corps, a une influence dans les décisions de recrutement. Vous pouvez utiliser le même genre dans les paires appariées, et il peut ne pas être facile d'obtenir des volontaires pour une telle étude.

32. Suivant Hamermesh et Parker (2005), essayez d'établir un lien entre l'apparence physique des enseignants au collège et les évaluations des étudiants. Ceci peut être fait sur le campus via une enquête. Des données quelque peu brutes peuvent être obtenues à partir des sites internet qui permettent aux étudiants de classer leurs professeurs et qui fournissent des informations sur l'apparence. Idéalement par contre, toute évaluation de la beauté ne doit pas être faite par les étudiants contemporains ou anciens, car ces évaluations peuvent être influencées par les notes obtenues.

33. Utilisez des données de panel pour étudier les effets de différentes politiques économiques sur la croissance économique régionale. Étudier l'effet des taxes ou des dépenses est naturel, mais d'autres politiques peuvent être intéressantes. Par exemple, Craig, Jackson, et Thomson (2007) étudient les effets des programmes de Garantie du Crédit de l'Association des Petites Entreprises sur la croissance du revenu par tête.

34. Blinder et Watson (2014) ont récemment étudié les différences systématiques que l'on peut observer aux États-Unis, en particulier au niveau du PIB réel, en fonction du parti politique duquel provient le président en fonction. Il est possible de mettre à jour les données en ajoutant les trimestres plus récents en en étudiant d'autres variables économiques, comme le taux de chômage.

LISTE DES JOURNAUX

La liste suivante est une liste non exhaustive des journaux populaires contenant de la recherche empirique en économie, management, et d'autres sciences sociales. Une liste complète des journaux peut être trouvée sur Internet à http://www.econlit.org

American Economic Review
American Journal of Agricultural Economics
American Political Science Review
Applied Economics
Brookings Papers on Economic Activity
Canadian Journal of Economics
Demography
Economic Development and Cultural Change
Economic Inquiry
Economica
Economics Letters
Empirical Economics
Federal Reserve Bulletin
International Economic Review
International Tax and Public Finance
Journal of Applied Econometrics
Journal of Business and Economic Statistics
Journal of Development Economics
Journal of Economic Education

Journal of Empirical Finance
Journal of Environmental Economics and Management
Journal of Finance
Journal of Health Economics
Journal of Human Resources
Journal of Industrial Economics
Journal of International Economics
Journal of Labor Economics
Journal of Monetary Economics
Journal of Money, Credit and Banking
Journal of Political Economy
Journal of Public Economics
Journal of Quantitative Criminology
Journal of Urban Economics
National Bureau of Economic Research Working Papers Series
National Tax Journal
Public Finance Quarterly
Quarterly Journal of Economics
Regional Science & Urban Economics
Review of Economic Studies
Review of Economics and Statistics

SOURCES DE DONNÉES

De nombreuses sources de données sont valables à travers le monde. Les gouvernements de la plupart des pays compilent de nombreuses données ; certaines sources de données générales et facilement accessibles pour les États-Unis, telles que l'*Economic Report of the President*, the *Statistical Abstract of the United States*, et le *County and City Data Book*, on déjà été mentionnées. Les données financières internationales de beaucoup de pays sont publiées de manière annuelle dans *International Financial Statistics*. Différents magazines, comme *BusinessWeek* and *U.S. News and World Report*, publient souvent des statistiques – telles que les salaires des PDG et la performance de la firme, le classement des programmes académiques – qui sont nouveaux et qui peuvent être utilisées dans l'analyse économique.

Plutôt que d'essayer de fournir une liste ici, nous fournissons plutôt certaines adresses Internet qui constituent des sources complètes pour les économistes. Un site très utile pour les économistes est à SUNY, Oswego. L'adresse est

http://www.rfe.org.

Ce site propose des liens vers les journaux, les sources de données, et des listes d'économistes professionnels et académiques. Il est vraiment simple à utiliser. Un autre site très utile est

http://econometriclinks.com,

qui contient des liens vers de nombreuses sources de données ainsi que d'autres sites d'intérêt pour les économistes empiriques.

En outre, le *Journal of Applied Econometrics* et le *Journal of Business and Economic Statistics* possèdent des archives de données qui contiennent des bases de données utilisées dans la plupart des articles publiés dans les journaux durant les dernières années. Si vous trouvez une base de données qui vous intéresse, c'est une bonne manière de débuter, car le nettoyage et le formatage de données ont déjà été accomplis. Le

hic est que certaines de ces bases de données sont utilisées pour des analyses économétriques qui sont plus avancées que ce que nous avons couvert dans cet ouvrage. D'un autre côté, il est souvent utile d'estimer des modèles plus simples en utilisant des méthodes économétriques standard à des fins de comparaison.

Beaucoup d'universités, telles que l'université de California – Berkeley, l'université de Michigan, et l'université du Maryland, entretiennent des bases de données très importantes ainsi que des liens vers une série de base de données. Votre propre bibliothèque contient peut-être une série importante de liens vers des bases de données en management, économie, et d'autres sciences sociales. Les banques centrales régionales, telles que celles à Saint-Louis, gèrent une série de données. Le National Bureau of Economic Research publie une série de données utilisées par certains de ses chercheurs. Les gouvernements régionaux et fédéraux publient aujourd'hui beaucoup de données qui peuvent être disponibles via Internet. Les données de recensement sont disponibles publiquement à partir du U.S. Census Bureau. (Deux publications utiles sont *Economic Census*, publié les années finissant par deux et sept, et le *Census of Population and Housing*, publié au début de chaque décennie.) D'autres agences, telles que le département de la justice américaine, rendent aussi des données disponibles auprès du public.

ANNEXE

A

OUTILS MATHÉMATIQUES DE BASE

Traduction de Alain Durré

Cette annexe couvre les outils mathématiques de base utilisés dans l'analyse économétrique. Nous résumerons ici diverses propriétés des opérateurs de sommation, nous étudierons les propriétés de certaines équations linéaires et non-linéaires, et nous reverrons des notions de proportions et de pourcentages. Nous présenterons également certaines fonctions spéciales qui apparaissent souvent dans les applications économétriques, comme les fonctions quadratiques et le logarithme naturel. Les quatre premières sections de cette annexe nécessitent simplement des connaissances de base en algèbre. La section A.5 contient un bref aperçu du calcul différentiel ; bien qu'une connaissance de ce type de calcul ne soit pas nécessaire afin de comprendre l'essentiel du texte, nous utiliserons ces calculs dans certaines annexes en fin de chapitre et dans plusieurs des chapitres plus avancés de la troisième partie de cet ouvrage.

A.1 OPÉRATEUR DE SOMMATION ET STATISTIQUES DESCRIPTIVES

L'**opérateur de sommation** est une notation utile pour manipuler des expressions impliquant des sommes de nombres, et joue un rôle clé dans les statistiques et l'analyse économétrique. Si l'expression $\{x_i : i = 1, \ldots, n\}$ désigne une suite de n nombres, alors nous écrivons la somme de ces nombres de la manière suivante :

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \ldots + x_n. \qquad \textbf{[A.1]}$$

Avec cette définition, il est simple de démontrer que l'opérateur de sommation a les propriétés suivantes :

Propriété 1 : Pour toute constante c,

$$\sum_{i=1}^{n} c = nc. \qquad \textbf{[A.2]}$$

Propriété 2 : Pour toute constante c,

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i. \qquad \textbf{[A.3]}$$

Propriété 3 : Si $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$ est un ensemble de n paires de nombres, et pour a et b deux constantes, alors

$$\sum_{i=1}^{n} (ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i. \qquad \textbf{[A.4]}$$

Il est aussi important d'être conscient des opérations qui **ne peuvent pas** être réalisées à partir de l'opérateur de sommation. Si $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$ est un ensemble de n paires de nombre, avec $y_i \neq 0$ pour tout i. Alors,

$$\sum_{i=1}^{n} (x_i/y_i) \neq \left(\sum_{i=1}^{n} x_i\right) \Big/ \left(\sum_{i=1}^{n} y_i\right).$$

En d'autres termes, la somme des ratios n'est pas égale au ratio des sommes. En prenant par exemple $n = 2$, une application algébrique simple nous montre bien que cette égalité n'est pas valable ; en effet $x_1/y_1 + x_2/y_2 \neq (x_1 + x_2)/(y_1 + y_2)$. De la même façon, la somme des carrés n'est pas égale au carré

de la somme : $\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i \right)^2$, sauf dans certains cas spécifiques. Le fait que le carré de la somme ne soit pas égal à la somme des carrés peut être montré simplement en considérant $n = 2$: $x_1^2 + x_2^2 \neq (x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2$.

Pour n nombres $\{x_i : i = 1, \ldots, n\}$, la **moyenne** se calcule en additionnant l'ensemble des données, puis en divisant par n :

$$\bar{x} = (1/n)\sum_{i=1}^{n} x_i. \qquad \textbf{[A.5]}$$

Lorsque x_i est un échantillon d'une variable donnée (par exemple du nombre d'années d'études), la moyenne est souvent appelée *moyenne de l'échantillon*, pour souligner le fait que ce chiffre est calculé à partir d'un ensemble particulier de données. La moyenne de l'échantillon est un exemple de **statistique descriptive** ; dans ce cas, la statistique décrit la tendance centrale de l'ensemble des points x_i.

Il existe des propriétés basiques à propos des moyennes qu'il est important de connaître. Tout d'abord, supposons que nous calculions une variable d pour chaque observation, égale à la valeur de l'observation moins la moyenne de l'échantillon : $d_i \equiv x_i - \bar{x}$ (où « d » représente *l'écart* par rapport à la moyenne). Dans ce cas, la somme des écarts est toujours égale à 0 :

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

Nous résumons cela en écrivant :

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0. \qquad \textbf{[A.6]}$$

Une application numérique simple permet de vérifier cela. Supposons $n = 5$ et $x_1 = 6$, $x_2 = 1$, $x_3 = -2$, $x_4 = 0$, et $x_5 = 5$. Alors, $\bar{x} = 2$, et l'écart par rapport à la moyenne pour chaque observation est donc $\{4, -1, -4, -2, 3\}$. La somme des ces valeurs est égale à zéro, comme vu dans l'équation (A.6).

Dans le cadre du chapitre 2 à propos de l'analyse de régression, il est essentiel de connaître certaines propriétés algébriques supplémentaires impliquant des écarts par rapport aux moyennes de l'échantillon. Une propriété importante est que la somme du carré des écarts est égale à la somme des carrés des x_i moins n fois le carré de $\bar{x}$:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2. \qquad \textbf{[A.7]}$$

Cela peut être démontré en utilisant les propriétés de base de l'opérateur de sommation :

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x})^2 &= \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n(\bar{x})^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2. \end{aligned}$$

Pour un ensemble de données et deux variables, $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$, nous pouvons aussi montrer que :

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i(y_i - \bar{y}) \qquad [A.8]$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i y_i - n(\bar{x} \cdot \bar{y}) ;$$

qui est une généralisation de (A.7). (où, $y_i = x_i$ pour tout i.)

La moyenne est une mesure de la tendance centrale d'une série que nous allons utiliser dans la majeure partie de cet ouvrage. Cependant, il est parfois utile d'utiliser la médiane (ou médiane de l'échantillon) afin de décrire la tendance centrale d'une série. Pour obtenir la médiane de n nombres $\{x_1, \ldots, x_n\}$, nous devons tout d'abord classer les valeurs x_i de la plus petite à la plus grande. Ensuite, si n est impair, alors la médiane de l'échantillon est le nombre se trouvant au milieu de la série triée par ordre croissant. Par exemple, en considérant la série $\{-4, 8, 2, 0, 21, -10, 18\}$, la médiane est égale à 2 (car la série triée par ordre croissant donne $\{-10, -4, 0, 2, 8, 18, 21\}$). Si nous changeons la dernière valeur de cette liste, par exemple si 21 devient 42, alors la médiane reste la même, tandis que la moyenne dans cette situation aurait augmenté de 5 à 8. Généralement, la médiane est moins sensible que la moyenne aux variations des valeurs extrêmes (positives ou négatives) dans une liste de nombres triés. C'est pourquoi le « salaire médian » ou la « valeur médiane des biens immobiliers » sont souvent utilisés, plutôt que la moyenne, afin de synthétiser les salaires ou la valeur de l'immobilier dans une ville ou dans un pays donné.

Si n est pair, il n'existe pas une méthode unique afin de déterminer la médiane, étant donné qu'il existe deux nombres « au milieu ». Souvent, la médiane est définie comme la moyenne des deux nombres du milieu (une fois la liste triée du plus petit au plus grand). En utilisant cette règle, la médiane pour la suite de nombre $\{4, 12, 2, 6\}$ serait égale à $(4 + 6)/2 = 5$.

A.2 PROPRIÉTÉ DES FONCTIONS LINÉAIRES

Les fonctions linéaires jouent un rôle important dans l'analyse économétrique car elles sont simples à manipuler et à interpréter. Si x et y sont deux variables reliées par la relation suivante :

$$y = \beta_0 + \beta_1 x, \qquad [A.9]$$

alors on dit que y est une **fonction linéaire de** x, et que β_0 et β_1 sont les deux paramètres (nombres) décrivant cette relation. β_0 est l'**ordonnée à l'origine** et β_1 la pente.

La caractéristique principale d'une fonction linéaire est que le changement de y est toujours égal β_1 multiplié par le changement de x :

$$\Delta y = \beta_1 \Delta x, \qquad [A.10]$$

où Δ signifie la « variation ». En d'autres termes, l'**effet marginal** de x sur y est constant, et est égal à β_1.

EXEMPLE A.1
Fonction linéaire des dépenses de logement

Supposons que la relation entre les dépenses mensuelles de logement et le salaire mensuel s'écrive sous la forme :

$$logement = 164 + 0{,}27\,revenu \qquad \textbf{[A.11]}$$

Dans ce cas, pour chaque hausse du salaire de 1 dollar, 27 centimes sont dépensés pour le logement. Si le salaire de la famille augmente de 200 $, alors les dépenses de logement augmenteront de (0,27)200 = 54 $. Cette fonction est tracée dans le graphique A.1.

Selon l'équation (A.11), une famille n'ayant aucun revenu dépense 164 $ pour son logement, ce qui ne peut pas être littéralement vrai. Pour les familles avec de faibles revenus, cette fonction linéaire ne permettra pas de décrire de manière adéquate la relation entre le *salaire* et les *dépenses de logement*, et c'est pour cela que nous devrons éventuellement utiliser d'autres types de fonctions pour décrire ce type de relation.

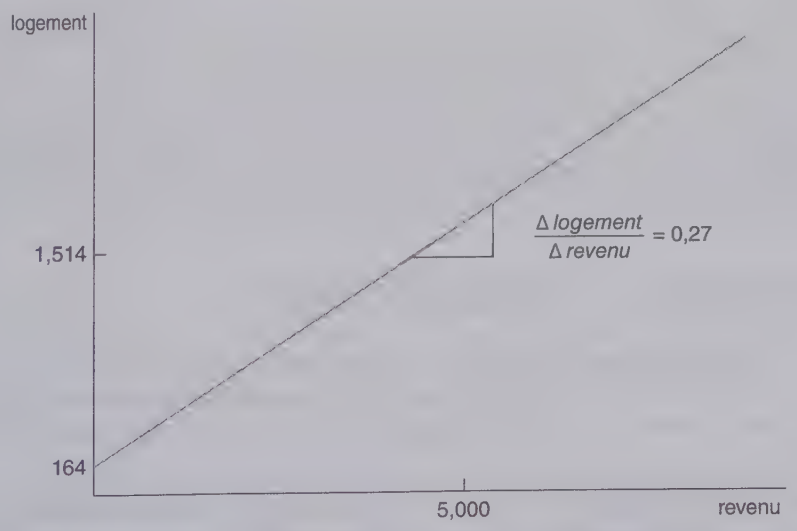
Dans (A.11), la *propension marginale à consommer* en dépenses de logement en fonction du salaire est égale à 0,27. Ce chiffre est très différent de la *propension moyenne à consommer*, qui est égale à

$$\frac{logement}{revenu} = 164/revenu + 0{,}27$$

La propension moyenne à consommer n'est pas constante et est toujours supérieure à la propension marginale à consommer. La propension moyenne à consommer tend vers la propension marginale à consommer lorsque le revenu augmente.

Les fonctions linéaires peuvent être composées de plusieurs variables. Supposons que y est fonction de x_1 et de x_2, tel que

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \qquad \textbf{[A.12]}$$



Graphique A.1 *Graphique de l'équation* ***logement = 164 + 0,27 revenu.***

Il est difficile d'imaginer la forme de cette fonction, car il faudrait la représenter en trois dimensions. Cependant, β_0 est toujours la valeur de l'ordonnée à l'origine (la valeur de y lorsque $x_1 = 0$ et $x_2 = 0$), et β_1 et β_2 mesurent la pente de chaque variable. À partir de (A.12), le changement de y, pour tout changement de x_1 et x_2, peut s'écrire

$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_2. \qquad \textbf{[A.13]}$$

Si x_2 ne change pas, c'est-à-dire si $\Delta x_2 = 0$, alors nous avons

$$\Delta y = \beta_1 \Delta x_1 \text{ si } \Delta x_2 = 0$$

β_1 est donc bien la pente de x_1 :

$$\beta_1 = \frac{\Delta y}{\Delta x_1} \text{ si } \Delta x_2 = 0$$

Étant donné que β_1 mesure la variation de y suite à un changement de x_1 lorsque x_2 est fixe, β_1 est souvent appelé **effet partiel** de x_1 sur y. La notion d'effet marginal impliquant que les autres facteurs soient fixes, cela se rapproche fortement de la notion de **ceteris paribus** (toutes choses égales par ailleurs). Nous pouvons faire la même interprétation pour le paramètre β_2 : $\beta_2 = \frac{\Delta y}{\Delta x_2}$ si $\Delta x_1 = 0$, de telle sorte que β_2 représente l'effet marginal de x_2 sur y.

EXEMPLE A.1
La demande de compact-disc

Supposons que la demande mensuelle en « compact-disc » (CD) soit liée au prix des CD et au revenu disponible mensuel :

$$\textit{Demande de CD} = 120 - 9{,}8\ \textit{prix} + 0{,}03\ \textit{revenu},$$

où *prix* est le prix en dollar d'un CD et *revenu* est mesuré en dollar. La courbe de demande montre la relation entre la *quantité* et le *prix*, avec la variable *revenu* et les autres facteurs fixes. Cette relation est représentée en deux dimensions dans le graphique A.2, pour un niveau de salaire mensuel fixé à 900 $. La pente de la courbe de demande est égale à −9,8, ce qui représente *l'effet marginal* du prix sur les quantités (en fixant le salaire). Si le prix d'un CD augmente de 1 dollar, alors la demande diminuera de 9,8 (en omettant le fait qu'il ne soit pas possible d'acheter un nombre non-entier de CD). Une augmentation du salaire ne fait que déplacer la courbe de demande (changement de l'ordonnée à l'origine) mais ne modifie en rien la pente de cette courbe.

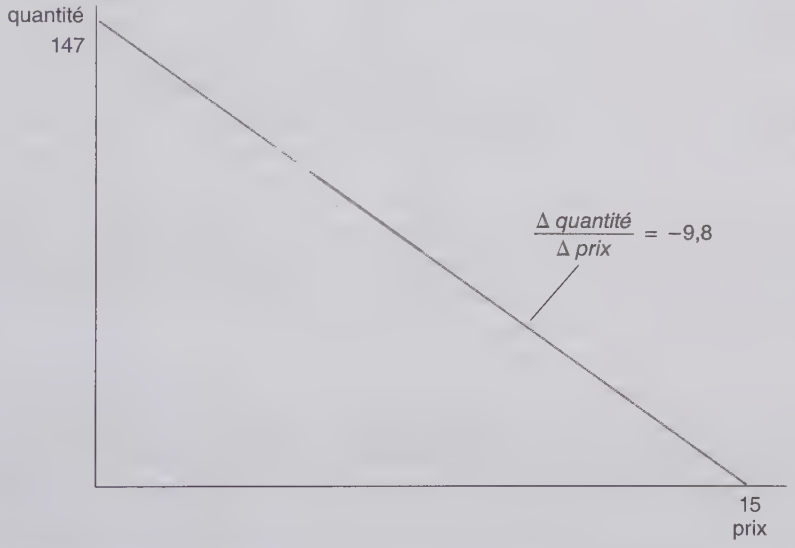
A.3 PROPORTIONS ET POURCENTAGES

Les proportions et pourcentages jouent un rôle tellement important en économie appliquée qu'il est essentiel d'être parfaitement à l'aise avec ces notions. De nombreuses quantités publiées dans la presse le sont sous la forme de pourcentage, comme par exemple le taux d'intérêt, le taux de chômage ou bien encore le taux de réussite au lycée.

Il est important d'être capable de transformer des proportions en pourcentages, et vice versa. Pour obtenir un pourcentage, il suffit de multiplier une proportion par 100. Par exemple, si la proportion d'adultes ayant obtenu le baccalauréat est de 0,82, on dit alors que 82 % (82 pourcents) des adultes ont le baccalauréat. Une autre manière de voir les pourcentages et les proportions est de voir une proportion comme la forme décimale d'un pourcentage. Par exemple, si le taux marginal de taxation d'une famille ayant un revenu de 30 000 $ est de 28 %, alors la proportion du prochain dollar de salaire qui devra être payée dans le cadre de l'impôt sur le revenu est de 0,28 (ou 28 centimes).

Lorsque nous utilisons des pourcentages, nous avons souvent besoin de les convertir sous forme décimale. Par exemple, si la taxe de vente dans un pays est de 6 % et que 200 $ sont dépensés en biens taxables, alors la taxe de vente payée est égale à 200*(0,06) = 12 $. Si le rendement annuel d'un certificat de dépôt (CD) est de 7.6 % et que vous placez 3 000 $ sur ce compte au début de l'année, alors à la fin de l'année, vous recevrez des intérêts pour un montant 3 000 × (0,076) = 228 $. Malheureusement pour vous, les revenus des intérêts ne s'obtiennent pas en multipliant 3 000 par 7,6.

Nous devons faire attention au fait que des proportions sont parfois incorrectement appelées pourcentage dans la presse. Si vous lisez par exemple « Le pourcentage d'étudiant du secondaire qui boivent de l'alcool est de 0,57 », nous savons qu'il s'agit en réalité de 57 % (et non pas 0,57 %, comme une interprétation littéraire pourrait le faire penser). Les fans de volley-ball ont sûrement déjà dû lire des articles de presse contenant des déclarations telles que « son pourcentage de frappe est de 0,372 ». En réalité, son pourcentage de frappe est de 37,2 % et non de 0,372.



Graphique A.2 *Graphique de l'équation* ***quantité*** *= 120 – 9,8 prix + 0,03 revenu, avec un salaire constant de $ 900.*

En économétrie, nous mesurons souvent des variations exprimées dans des unités différentes. Définissons x comme étant une variable quelconque, telle que le salaire d'un individu, le nombre de crimes commis par une communauté ou bien le profit d'une entreprise. Soit x_0 et x_1 les deux premières valeurs de x : x_0 est le valeur initiale, et x_1 la valeur suivante. Par exemple, x_0 pourrait être le salaire d'un individu en 1994, et x_1 le salaire de ce même individu en 1995. La **variation proportionnelle** de x lorsque x passe de x_0 à x_1, parfois appelée la **variation relative**, est simplement égale à

$$(x_1 - x_0)/x_0 = \Delta x/x_0, \qquad \textbf{[A.14]}$$

en supposant bien évidemment que $x_0 \neq 0$. En d'autres termes, pour obtenir la variation proportionnelle, il suffit de diviser la variation de x par sa valeur initiale. C'est un moyen de standardiser la variation, afin que celle-ci soit sans unité. Par exemple, si le salaire d'un individu passe de 30 000 $ par an à 36 000 $ par an, alors la variation proportionnelle est égale à 6 000/30 000 = 0,20.

Il est plus fréquent d'exprimer les variations en pourcentage. Le **pourcentage de variation** de x lorsque x passe de x_0 à x_1 est simplement égal à 100 fois la variation proportionnelle :

$$\%\Delta x = 100(\Delta x/x_0)\ ; \qquad \textbf{[A.15]}$$

la notation « $\%\Delta x$ » se lit « pourcentage de variation de x. » Par exemple, lorsque le salaire d'un individu passe de 30 000 $ à 33 750 $, alors son salaire augmente de 12,5 %. On obtient ce résultat en multipliant la variation proportionnelle, 0,125, par 100.

De nouveau, nous devons faire attention au fait que les termes « variation proportionnelle » et « pourcentage de variation » sont parfois utilisés de manière incorrecte par les médias. Dans l'exemple précédent par exemple, déclarer que le pourcentage de variation de salaire est de 0,125 est faux et peut prêter à confusion.

Lorsque nous nous intéressons par exemple aux variations de la population ou à des variations exprimées en dollars, il n'existe pas d'ambiguïté sur le sens d'un pourcentage de variation. Mais lorsque la variable pour laquelle nous souhaitons calculer un pourcentage de variation est elle-même un pourcentage, comme c'est d'ailleurs souvent le cas en économie et en sciences sociales, cela peut poser quelques problèmes. Pour illustrer cela, notons x le pourcentage d'adulte ayant un diplôme d'études secondaires dans une ville donnée. Supposons que $x_0 = 24$ (24 % ont un diplôme d'études secondaires), et que ce pourcentage passe à $x_1 = 30$. Il est possible de calculer deux valeurs pour décrire la variation du pourcentage de personnes avec un diplôme d'études secondaires. La première consiste à calculer la variation x, Δx, soit dans notre exemple, $\Delta x = x_1 - x_0 = 6$. On dit alors que le pourcentage d'adultes ayant un diplôme d'études secondaires a augmenté de six points de pourcentage. Il est aussi possible de calculer le pourcentage de variation de x, en utilisant la formule de l'équation (A.15) : $\%\Delta x = 100[(30 - 24)/24] = 25$.

Dans cet exemple, le pourcentage de variation et la variation en point de pourcentage donnent deux résultats très différents. La **variation en point de pourcentage** représente la différence entre le pourcentage final et le pourcentage initial. Le pourcentage de variation correspond à un changement relatif par rapport à une valeur initiale. Il est important de bien distinguer ces deux valeurs, ce que fait le chercheur rigoureux, mais ce qui n'est malheureusement pas toujours le cas dans la presse populaire.

EXEMPLE A.2
Augmentation de la taxe de vente dans le Michigan

En mars 1994, le Michigan a voté une loi pour augmenter la taxe de vente de 4 % à 6 %. Lors des débats politiques, les partisans de cette mesure présentaient cela comme une augmentation de deux points de pourcentage, c'est-à-dire une augmentation de deux cents par dollar. À l'inverse, les opposants dénonçaient la mesure en parlant d'une augmentation de 50 % de cette taxe. Les deux camps avaient raison ; ils utilisaient simplement deux méthodes différentes pour mesurer la hausse de la taxe de vente. Naturellement, chaque groupe utilisait le chiffre qui renforçait le plus sa position.

Pour une variable telle que le salaire, cela ne fait aucun sens de parler de « point de pourcentage de salaire », étant donné que le salaire n'est pas mesuré en pourcentage. Une variation de salaire peut être décrite soit en dollar, soit en pourcentage de variation.

A.4 PRÉSENTATION DE QUELQUES FONCTIONS SPÉCIALES ET DE LEURS PROPRIÉTÉS

Dans la section A.2, nous avons revu les propriétés de base des fonctions linéaires. Nous avons présenté la caractéristique des fonctions du type $y = \beta_0 + \beta_1 x$: une variation d'une unité de x entraîne toujours la *même* variation de y, peu importe le niveau initial de x. Comme nous l'avons noté précédemment, cela revient à dire que l'effet marginal de x sur y est constant, ce qui n'est pas réaliste dans le cas de nombreuses relations économiques. Par exemple, la notion économique de *rendement marginal décroissant* n'est pas compatible avec une relation linéaire.

Afin de modéliser une grande variété de phénomènes économiques, nous avons besoin d'étudier différentes fonctions non-linéaires. Une **fonction non-linéaire** est caractérisée par le fait que la variation de y pour un changement donné de x dépend de la valeur initiale de x. Certaines fonctions non-linéaires apparaissent fréquemment en économie, et il est donc important de savoir interpréter ce type de fonction. Une compréhension complète des fonctions non-linéaires demande des connaissances approfondies en calcul différentiel. Ici, nous verrons simplement les aspects les plus importants de ces fonctions, en laissant de côté les détails de certaines dérivées pour la section A.5.

Fonction quadratique

Une façon simple de capturer le phénomène de rendement marginal décroissant consiste à ajouter un terme quadratique à la relation linéaire. Considérons par exemple l'équation

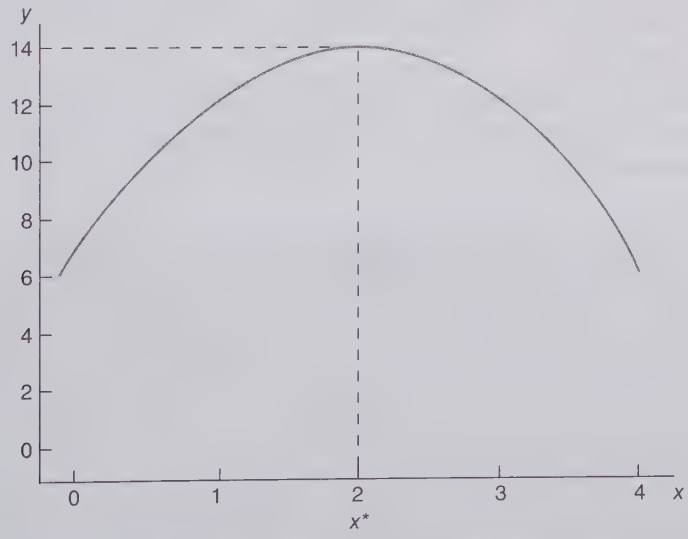
$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \qquad \textbf{[A.16]}$$

où β_0, β_1, et β_2 sont des paramètres. Lorsque $\beta_1 > 0$ et $\beta_2 < 0$, alors la relation entre y et x est de forme parabolique. Le graphique A.3 illustre cela avec $\beta_0 = 6$, $\beta_1 = 8$, et $\beta_2 = -2$.

Lorsque $\beta_1 > 0$ et $\beta_2 < 0$, il peut être montré (en utilisant les calculs présentés dans la prochaine section) que le maximum de la fonction se trouve au point

$$x^* = \beta_1/(-2\beta_2). \qquad \textbf{[A.17]}$$

Par exemple, si $y = 6 + 8x - 2x^2$ (donc avec $\beta_1 = 8$ et $\beta_2 = -2$), alors la plus grande valeur de y se trouve au point $x^* = 8/4 = 2$, et cette valeur est égale à $6 + 8(2) - 2(2)^2 = 14$ (voir graphique A.3).



Graphique A.3 *Graphique de l'équation* $y = 6 + 8x - 2x^2$.

L'équation (A.16) implique un effet de **rendement marginal décroissant** de x sur y, ce qui peut facilement être vu graphiquement. Supposons que nous commencions avec une valeur faible de x, puis que nous augmentions x par une constante appelée c. Cela aura un effet plus grand sur y que si nous avions commencé avec une valeur de x plus élevée, puis si nous avions augmenté x par la même constante c. À partir d'un certain niveau $x > x^*$, une hausse de x entraîne même une diminution de y.

L'assertion indiquant que x a un effet marginal décroissant sur y consiste à dire que la pente de la fonction du graphique A.3 est décroissante lorsque x augmente. Bien que ce phénomène soit facilement visible graphiquement, nous voulons être capable de quantifier rapidement la variation de la pente d'une fonction. Mathématiquement, une approximation de la pente d'une fonction quadratique s'écrit :

$$pente = \frac{\Delta y}{\Delta x} \approx \beta_1 + 2\beta_2 x \qquad \textbf{[A.18]}$$

pour de « petites » variations de x. [La partie droite de l'équation (A.18) est la **dérivée** en fonction de x de la fonction de l'équation (A.16).] Il est aussi possible d'écrire cette relation sous la forme

$$\Delta y \approx (\beta_1 + 2\beta_2 x)\Delta x \qquad \textbf{[A.19]}$$

pour des valeurs petites de Δx.

Pour mesurer la précision de cette approximation, considérons de nouveau la fonction $y = 6 + 8x - 2x^2$. Selon l'équation (A.19), $\Delta y \approx (8 - 4x)\Delta x$. Maintenant, supposons que nous commencions à $x = 1$ puis que nous augmentions x de 0,1 unité $\Delta x = 0{,}1$. En utilisant (A.19), $\Delta y \approx (8 - 4) \times 0{,}1 = 0{,}4$. Bien évidemment, nous aurions pu aussi calculer la variation de y en calculant y pour $x = 1$, puis pour $x = 1{,}1$, soit $y_0 = 6 + 8(1) - 2(1)^2 = 12$ et $y_1 = 6 + 8(1{,}1) - 2(1{,}1)^2 = 12{,}38$, avec donc une variation exacte de y de 0,38. On voit bien que l'approximation en utilisant l'équation (A.19) est proche de la variation exacte de y.

Maintenant, supposons que nous commencions toujours à $x = 1$ mais que nous augmentions x par une valeur plus grande, égale par exemple à 0,5 : $\Delta x = 0{,}5$. Dans cette situation, une approximation de la variation est $\Delta y \approx 4 \times (0{,}5) = 2$. Il est possible de calculer la variation exacte de y en appliquant le même raisonnement que précédemment, en calculant donc y pour $x = 1$ puis pour $x = 1{,}5$. La valeur initiale de y est alors de 12, tandis que la valeur finale est égale à $6 + 8(1{,}5) - 2(1.5)^2 = 13{,}5$: la variation réelle est donc égale à 1,5 (et non à 2 comme calculé avec l'approximation). Dans cet exemple, on voit bien que l'approximation est plus éloignée car la variation de x est plus grande.

Dans de nombreuses applications, l'équation (A.19) est utilisée pour approximer l'effet marginal de x sur y pour toute valeur initiale de x et pour une faible variation de x. Il est aussi possible de calculer la variation exacte si nécessaire.

EXEMPLE A.3
Une fonction quadratique du salaire

Supposons que la relation entre le salaire horaire et le nombre d'années de travail (*exper*) soit de la forme

$$salaire = 5{,}25 + 0{,}48\,exper - 0{,}008\,exper^2 \qquad \textbf{[A.20]}$$

Cette fonction a une forme similaire à celle du graphique A.3. En utilisant (A.17), *exper* a donc un effet positif sur le salaire jusqu'au point de retournement $exper^* = 0{,}48/[2(0{,}008)] = 30$. La première année d'expérience rapporte environ 0,48, ou 48 centimes [voir (A.19) avec $x = 0$ et $\Delta x = 1$]. Chaque année d'expérience supplémentaire entraîne une augmentation de salaire, mais une augmentation inférieure à celle de l'année précédente, ce qui reflète donc le rendement marginal décroissant de l'expérience. À partir de la 30ème année, une année supplémentaire d'expérience diminue même le salaire. Cette situation n'est pas très réaliste, mais est la conséquence de l'utilisation d'une fonction quadratique afin de capturer l'effet de rendement marginal décroissant : ce type de fonction atteint un maximum pour un point donné, puis est décroissante après cette valeur (ici n = 30). Pour des raisons pratiques, le point de retournement peut être très élevé, et il est possible de regarder uniquement la partie de la courbe avant le point de retournement (mais ce n'est pas toujours le cas).

Le graphique de la fonction quadratique de (A.16) exhibe une forme en U si $\beta_1 < 0$ et $\beta_2 > 0$, et dans ce cas nous avons un rendement marginal croissant. Le minimum de cette fonction correspond alors au point $-\beta_1/(2\beta_2)$.

Logarithme Naturel

La fonction non-linéaire qui joue le rôle le plus important dans l'analyse économétrique est le **logarithme naturel**. Dans ce texte, nous utilisons la notation « log(x) » lorsque nous utilisons le logarithme naturel, aussi appelé simplement « **fonction log** »

$$y = \log(x). \qquad \textbf{[A.21]}$$

Vous vous rappelez peut-être avoir appris différentes notations concernant le logarithme naturel : $\ln(x)$ et $\log_e(x)$ étant les plus communes. Ces différentes notations sont utiles lorsque des logarithmes avec des bases différentes sont utilisés. Pour nos besoins, nous utiliserons simplement le logarithme naturel, et donc nous utiliserons la notation $\log(x)$ dans l'ensemble de cet ouvrage. Cela correspond à la notation utilisée dans de nombreux logiciels statistiques, bien que certains utilisent $\ln(x)$ [tout comme la plupart des calculatrices]. Les économistes utilisent quant à eux les deux notations, $\log(x)$ et $\ln(x)$ (ce qui peut vous être utile lorsque vous lirez des papiers académiques en économie).

La fonction $y = \log(x)$ est définie pout tout $x > 0$, et est représentée dans le graphique A.4. Ce n'est pas très important de savoir comment les valeurs de $\log(x)$ sont obtenues. Pour nos besoins, nous pouvons voir la fonction comme une « boîte noire », qui renvoie une valeur $\log(x)$ pour tout $x > 0$.

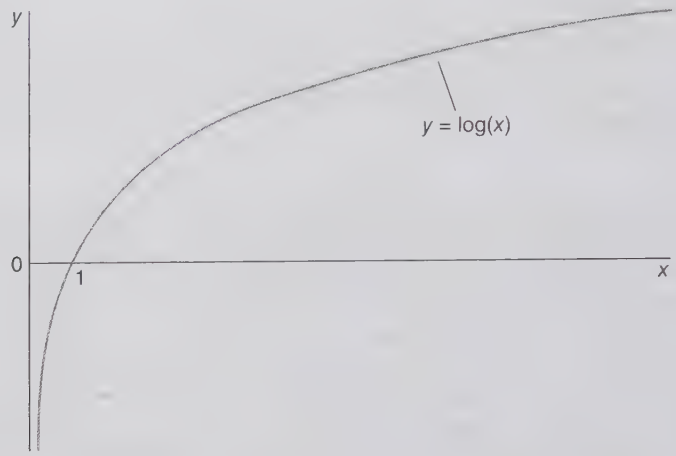
Plusieurs choses sont apparentes à partir du graphique A.4. Premièrement, lorsque $y = \log(x)$, la relation entre y et x montre une situation de rendement marginal décroissant. Une différence importante entre la fonction quadratique du graphique A.3 et le graphique A.4 est que pour $y = \log(x)$, l'effet de x sur y ne devient jamais négatif : la pente de la fonction tend vers 0 lorsque x devient très grand, mais la pente n'est jamais égale à 0 et n'est jamais négative.

À partir du graphique A.4, nous pouvons de plus noter que :

$$\log(x) < 0 \text{ pour } 0 < x < 1$$
$$\log(1) = 0$$
$$\log(x) > 0 \text{ pour } x > 1$$

En particulier, $\log(x)$ peut être positif ou négatif. Certaines propriétés de la fonction log nous seront utiles par la suite, à savoir :

$$\log(x_1 \cdot x_2) = \log(x_1) + \log(x_2),\ x_1, x_2 > 0$$
$$\log(x_1/x_2) = \log(x_1) - \log(x_2),\ x_1, x_2 > 0$$
$$\log(x^c) = c\log(x),\ x > 0,\ c \text{ un nombre quelconque.}$$



Graphique A.4 *Graphique de l'équation* **y** = *log*(**x**).

Le logarithme peut être utilisé pour approximer diverses relations apparaissant dans les applications économétriques. Tout d'abord, $\log(1 + x) \approx x$ pour $x \approx 0$. Vous pouvez essayer pour $x = 0{,}02$, $x = 0{,}1$, et $x = 0{,}5$ afin de voir que la précision de cette estimation se détériore au fur et à mesure que x augmente. Plus utile encore, il est possible d'utiliser la différence entre deux logarithmes pour approximer une variation relative. Soit x_0 et x_1 deux valeurs positives. Dans ce cas, nous pouvons démontrer en utilisant quelques calculs que :

$$\log(x_1) - \log(x_0) \approx (x_1 - x_0)/x_0 = \Delta x/x_0 \qquad \textbf{[A.22]}$$

pour des petites variations de x. Si nous multiplions (A.22) par 100 et définissons $\Delta\log(x) = \log(x_1) - \log(x_0)$, alors

$$100 \cdot \Delta\log(x) \approx \%\Delta x \qquad \textbf{[A.23]}$$

pour des petites variations de x. Le sens de « petites variations » dépend du contexte, et nous verrons différents exemples dans la suite de cet ouvrage.

Pourquoi estimer le pourcentage de variation en utilisant (A.23) alors que la formule exacte pour calculer le pourcentage de variation est très simple ? Avant d'expliquer pourquoi la relation (A.23) est très utile en économétrie, regardons tout d'abord la précision de cette approximation avec deux exemples.

Premièrement, supposons que $x_0 = 40$ et $x_1 = 41$. Dans cette situation, le pourcentage de variation de x, lorsque x passe de x_0 à x_1, est de $100(x_1 - x_0)/x_0 = 2{,}5$ %. Avec le logarithme, nous obtenons alors, $\log(41) - \log(40) = 0{,}0247$, et si l'on multiplie ce nombre par 100, nous obtenons un nombre très proche de 2,5. L'approximation marche donc très bien dans cet exemple. Maintenant, considérons un changement beaucoup plus important, en définissant : $x_0 = 40$ et $x_1 = 60$. Dans cette situation, la variation en pourcentage exacte est égale à 50 %. Cependant, avec la formule du log, nous obtenons $\log(60) - \log(40) \approx 0{,}4055$, ce qui nous donne donc une approximation de 40,55 %, qui est très éloignée de la valeur exacte.

Mais pourquoi la relation (A.23) est utile simplement pour les petites variations ? Pour répondre à cette question, nous devons tout d'abord définir l'**élasticité** de y par rapport à x, avec la formule :

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \frac{\%\Delta y}{\%\Delta x}. \qquad \textbf{[A.24]}$$

L'élasticité de y par rapport à x est donc le pourcentage de variation de y lorsque x augmente de 1 %. Cette notion est une notion de base de l'économie.

Si y est une fonction linéaire de x, $y = \beta_0 + \beta_1 x$, alors l'élasticité est :

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{\beta_0 + \beta_1 x}, \qquad \textbf{[A.25]}$$

qui dépend donc clairement de la valeur de x. (ceci est une généralisation bien connue résultant de la théorie de la demande : l'élasticité n'est pas constante sur une courbe de demande linéaire).

Les élasticités sont d'une importance toute particulière en économie appliquée, et non pas seulement pour la théorie de la demande. Dans de nombreuses situations, il est utile d'avoir des modèles à élasticité constante, et la fonction log permet justement cela. Si nous utilisons l'approximation (A.23) pour x et y, alors l'élasticité est donc approximativement égale à $\Delta\log(y)/\Delta\log(x)$. Un modèle à élasticité constante est donc approximé en utilisant l'équation

$$\log(y) = \beta_0 + \beta_1 \log(x), \qquad \textbf{[A.26]}$$

Avec β_1 l'élasticité de y par rapport à x (pour tout x, $y > 0$).

EXEMPLE A.4
Fonction de demande à élasticité constante

Si q est la quantité demandée et p le prix, et que ces variables sont reliées par la relation

$$\log(q) = 4{,}7 - 1{,}25\log(p)$$

alors l'élasticité-prix est environ égale à $-1{,}25$. Une augmentation de 1 % du prix entraîne donc une baisse de la quantité demandée d'environ 1,25 %.

Pour nos besoins, le fait que β_1 dans (A.26) soit une approximation de l'élasticité n'est pas important. En réalité, lorsque l'élasticité est calculée à partir d'une dérivée, comme dans la section A.5, alors sa définition est exacte. Dans le cadre de l'analyse économétrique, l'équation (A.26) définit donc un modèle à élasticité constante. Ces modèles sont d'ailleurs largement utilisés en économie appliquée.

Les travaux empiriques peuvent utiliser d'autres types de fonction log. Supposons que $y > 0$ et définissons

$$\log(y) = \beta_0 + \beta_1 x. \qquad \textbf{[A.27]}$$

Dans ce cas, $\Delta\log(y) = \beta_1 \Delta x$, donc $100 \cdot \Delta\log(y) = (100 \cdot \beta_1)\Delta x$. Donc, lorsque y et x sont reliés selon une équation du type (A.27), nous avons

$$\%\Delta y \approx (100 \cdot \beta_1)\Delta x. \qquad \textbf{[A.28]}$$

EXEMPLE A.5
Équation logarithmique du salaire

Supposons que le salaire horaire et le nombre d'années d'éducation soient reliés par la relation :

$$\log(salaire) = 2{,}78 + 0{,}094\, educ$$

En utilisant (A.28), nous obtenons donc,

$$\%\Delta salaire \approx 100(0{,}094)\Delta educ = 9{,}4\Delta educ$$

Dans cette situation, une année supplémentaire d'expérience entraîne une hausse du salaire horaire d'environ 9,4 %.

En général, la quantité $\%\Delta y/\Delta x$ est appelé **semi-élasticité** de y par rapport à x. La semi-élasticité représente la variation en pourcentage de y lorsque x augmente d'une unité. Dans le modèle (A.27), nous avons montré que la semi-élasticité est constante et égale à $100 \times \beta_1$. Dans l'exemple A.6, nous pouvons résumer la relation entre le salaire et le nombre d'années d'études en montrant que – peu importe le nombre d'années d'études de départ – une année supplémentaire d'étude entraîne une hausse de salaire d'environ 9,4 %.

Une autre formule utile en économie appliquée est

$$y = \beta_0 + \beta_1 \log(x), \qquad \textbf{[A.29]}$$

où $x > 0$. Comment peut-on interpréter cette relation ? Si l'on s'intéresse à la variation de y, nous obtenons $\Delta y = \beta_1 \Delta \log(x)$, qui peut être réécrit sous la forme $\Delta y = (\beta_1/100)[100 \cdot \Delta \log(x)]$. Dans cette situation, en utilisant l'équation (A.23), nous avons donc :

$$\Delta y \approx (\beta_1/100)(\%\Delta x). \qquad \textbf{[A.30]}$$

En d'autres termes, y augmente de $\beta_1/100$ lorsque x augmente de 1 %.

EXEMPLE A.6
Fonction d'offre de travail

Supposons que l'offre de travail d'un employé s'écrive sous la forme

$$heures = 33 + 45{,}1 \log(salaire),$$

où *salaire* est le salaire horaire et *heures* est le nombre d'heures travaillées par jour. Alors, en utilisant (A.30),

$$\Delta heures \approx (45{,}1/100)(\%\Delta salaire) = 0{,}451\%\Delta salaire$$

En d'autres termes, une augmentation de 1 % de *salaire* augmente le nombre d'heures travaillées par semaine d'environ 0,45, soit une augmentation d'un petit peu moins d'une demi-heure. Si le salaire augmente de 10 %, alors $\Delta heures = 0{,}451 \times (10) = 4{,}51$, soit une augmentation d'environ 4 h 30.

La fonction exponentielle

La dernière fonction de cette partie est une fonction spéciale liée au log. Considérons l'équation (A.27). Dans cette équation, $\log(y)$ est une fonction linéaire de x. Mais comment pouvons-nous exprimer y en fonction de x ? Pour cela, nous devons alors utiliser ce que l'on appelle la **fonction exponentielle**.

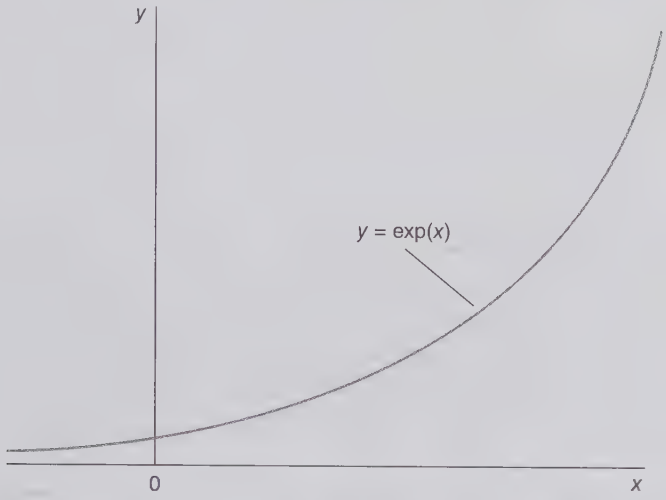
Une fonction exponentielle s'écrit sous la forme $y = \exp(x)$, et une représentation de cette fonction est tracée dans le graphique A.5. À partir de ce graphique, nous pouvons voir que $\exp(x)$ est défini pour toutes les valeurs de x, et est toujours supérieur à 0. La fonction exponentielle peut aussi s'écrire sous la forme $y = e^x$, mais nous n'utiliserons pas cette notation ici. Il existe deux valeurs importantes de la fonction exponentielle : $\exp(0) = 1$ et $\exp(1) = 2{,}7183$ (arrondi à la 4ème décimale).

La fonction exponentielle est l'inverse dans la fonction log, dans le sens où $\log[\exp(x)] = x$ pour tout x, et $\exp[\log(x)] = x$ pour tout $x > 0$. En d'autres termes, le log supprime l'exponentielle, et inversement (c'est pourquoi la fonction exponentielle est parfois appelée « fonction anti-log »). En particulier, notons que $\log(y) = \beta_0 + \beta_1 x$ est équivalent à

$$y = \exp(\beta_0 + \beta_1 x).$$

Si $\beta_1 > 0$, la relation entre x et y est de la même forme que celle du graphique A.5. Donc, si $\log(y) = \beta_0 + \beta_1 x$ pour $\beta_1 > 0$, alors x a un effet marginal croissant sur y. En prenant l'exemple A.6, cela signifie qu'une année supplémentaire d'étude à partir d'un niveau donné entraîne une hausse de salaire supérieure à la hausse de salaire de l'année précédente.

Deux relations sont particulièrement utiles à propos de la manipulation des fonctions exponentielles, à savoir : $\exp(x_1 + x_2) = \exp(x_1)\exp(x_2)$ et $\exp[c.\log(x)] = x^c$.



Graphique A.5 *Graphique de l'équation* y = exp(x).

A.5 LE CALCUL DIFFÉRENTIEL

Dans la section précédente, nous avons utilisé certaines approximations ayant comme fondement le calcul différentiel. Soit $y = f(x)$ une fonction f. Pour une faible variation de x, nous avons,

$$\Delta y \approx \frac{df}{dx} \cdot \Delta x, \qquad \textbf{[A.31]}$$

où df/dx est la dérivée de la fonction f, évaluée au point initial x_0. Nous pouvons aussi écrire la dérivée sous la forme dy/dx.

Par exemple, si $y = \log(x)$, alors $dy/dx = 1/x$. En utilisant (A.31), avec dy/dx évalué en x_0, cela nous donne $\Delta y \approx (1/x_0)\Delta x$, ou $\Delta\log(x) \approx \Delta x/x_0$, ce qui correspond à l'approximation donnée par l'équation (A.22).

En économétrie appliquée, il est important de connaître les dérivées des fonctions usuelles, car celles-ci sont utilisées pour calculer la pente d'une fonction en un point donné. Nous pouvons alors utiliser (A.31) pour obtenir une approximation de la variation de y suite à un petit changement de x. Pour une fonction linéaire, la dérivée est simplement la pente de la droite, comme nous pouvons le voir en calculant la dérivée de cette fonction ; si $y = \beta_0 + \beta_1 x$, alors $dy/dx = \beta_1$.

Si $y = x^c$, alors $dy/dx = cx^{c-1}$. La dérivée d'une somme de deux fonctions est égale à la somme de chaque dérivée, soit : $d[f(x) + g(x)]/dx = df(x)/dx + dg(x)/dx$. La dérivée d'une fonction multipliée par une constante est égale à cette constante multipliée par la dérivée de la fonction, soit : $d[cf(x)]/dx = c[df(x)/dx]$. Ces règles simples permettent de calculer les dérivées de fonctions plus compliquées. D'autres règles, concernant le produit, le quotient ou bien encore les fonctions composées sont sûrement familières à nombreux d'entre vous, mais nous ne reverrons pas cela ici.

Voici une liste non-exhaustive de fonctions souvent utilisées en économie, accompagnées de leurs dérivées :

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \; ; \; dy/dx = \beta_1 + 2\beta_2 x$$
$$y = \beta_0 + \beta_1 / x \; ; \; dy/dx = -\beta_1 / (x^2)$$
$$y = \beta_0 + \beta_1 \sqrt{x} \; ; \; dy/dx = (\beta_1 / 2) x^{1/2}$$
$$y = \beta_0 + \beta_1 \log(x) \; ; \; dy/dx = \beta_1 / x$$
$$y = \exp(\beta_0 + \beta_1 x) \; ; \; dy/dx = \beta_1 \exp(\beta_0 + \beta_1 x).$$

Si $\beta_0 = 0$ et $\beta_1 = 1$ dans la dernière expression, nous obtenons $dy/dx = \exp(x)$, lorsque $y = \exp(x)$.

Dans la section A.4, nous avions souligné que l'équation (A.26) permettait de définir un modèle à élasticité constante lorsque le calcul différentiel était utilisé. D'un point de vue calculatoire, l'élasticité se calcule avec la formule $(dy/dx)\cdot(x/y)$. En utilisant les propriétés du logarithme et de l'exponentiel, il est possible de montrer que pour (A.26), nous obtenons bien une élasticité constante $(dy/dx)\cdot(x/y) = \beta_1$.

Lorsque y est fonction de plusieurs variables, la notion de **dérivée partielle** devient importante. Supposons que

$$y = f(x_1, x_2). \qquad \textbf{[A.32]}$$

Il existe alors deux dérivées partielles, une par rapport à x_1 et l'autre par rapport à x_2. La dérivée partielle de y en fonction de x_1, que nous notons $\partial y / \partial x_1$, est simplement égale à la dérivée usuelle de (A.32) par rapport à x_1, lorsque x_2 est considérée comme une constante. De la même façon, $\partial y / \partial x_2$ est simplement la dérivée (A.32) par rapport à x_2, avec x_1 constant.

Les dérivées partielles sont utiles pour les mêmes raisons que les dérivées classiques. Il est possible d'approximer la variation de y

$$\Delta y \approx \frac{\partial y}{\partial x_1} \cdot \Delta x_1, \text{ avec } x_2 \text{ constant.} \qquad \textbf{[A.33]}$$

Cette relation permet d'étudier les effets partiels dans un modèle non-linéaire de la même façon que ce que nous avons vu pour le modèle linéaire. Si par exemple

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

alors

$$\frac{\partial y}{\partial x_1} = \beta_1, \frac{\partial y}{\partial x_2} = \beta_2.$$

Cela correspond donc aux effets partiels vu dans la section A.2.

Considérons maintenant un exemple un peu plus compliqué

$$y = 5 + 4x_1 + x_1^2 - 3x_2 + 7x_1 \cdot x_2. \qquad \textbf{[A.34]}$$

À présent, la dérivée de (A.34), par rapport à x_1 (avec x_2 constant), est simplement égale à :

$$\frac{\partial y}{\partial x_1} = 4 + 2x_1 + 7x_2 \; ;$$

et dépend donc à la fois de x_1 et x_2. La dérivée de (A.34), par rapport à x_2, est égale à $\partial y / \partial x_2 = -3 + 7x_1$, et donc ne dépend que de x_1.

EXEMPLE A.7
Fonction du salaire avec terme d'interraction

Considérons l'équation d'une fonction reliant le salaire au nombre d'années d'études

$$salaire = 3{,}10 + 0{,}41\,educ + 0{,}19\,exper - 0{,}004\,exper^2 + 0{,}007\,educ \cdot exper \qquad \textbf{[A.35]}$$

L'effet partiel de la variable *exper* sur la variable *salaire* est la dérivée partielle de (A.35) :

$$\frac{\partial salaire}{\partial exper} = 0{,}19 - 0{,}008\,exper + 0{,}007\,educ$$

Cette équation permet d'approximer la variation de salaire lorsque l'expérience augmente d'une année. L'effet partiel dépend donc ici du niveau initial de *exper*, mais aussi de celui de *educ*. Par exemple, pour un travailleur ayant les caractéristiques *educ* = 12 et *exper* = 5, une année d'expérience supplémentaire augmente son salaire d'environ 0,19 – 0,008(5) + 0,007(12) = 0,234, ou 23,4 centimes par heure. La variation exacte peut être définie en calculant le salaire dans (A.35) pour *exper* = 5, *educ* = 12, puis pour *exper* = 6, *educ* = 12, et en calculant ensuite la différence. Avec cette méthode, nous obtenons 0,23, ce qui est très proche de notre estimation en utilisant les dérivées partielles.

Le calcul différentiel joue un rôle important dans la minimisation et la maximisation de fonctions à une ou plusieurs variables. Si $f(x_1, x_2, \ldots, x_k)$ est une fonction différentiable de k-variables, alors une condition nécessaire pour que $x_1^*, x_2^*, \ldots, x_k^*$ soit un minimum ou un maximum de f pour tous les x_j est que

$$\frac{\partial f}{\partial x_j}(x_1^*, x_2^*, \ldots, x_k^*) = 0,\ j = 1, 2, \ldots, k. \qquad \textbf{[A.36]}$$

En d'autres termes, toutes les dérivées partielles doivent être égales à 0 en x_h^*. C'est ce que l'on appelle les conditions du premier ordre, qui permettent donc de minimiser ou de maximiser une fonction. D'un point de vue pratique, nous souhaitons donc résoudre l'équation (A.36) pour tous les x_h^*. Il est ensuite possible d'utiliser d'autres critères pour déterminer s'il s'agit alors d'un minimum ou d'un maximum, mais nous n'aurons pas besoin de cela ici. [Voir Sydsaeter et Hammond (1995) pour une discussion sur le calcul différentiel à plusieurs variables et l'optimisation de ces fonctions.]

RÉSUMÉ

Les outils mathématiques présentés ici sont très importants pour être capable d'analyser le résultat des régressions, ainsi que pour comprendre les probabilités et statistiques qui seront couvertes dans les annexes B et C. La compréhension des fonctions non-linéaires – spécialement les fonctions quadratiques, logarithmiques et exponentielles – est cruciale pour comprendre la recherche en économie appliquée actuelle. Le niveau de compréhension requis de ces fonctions ne demande pas de connaissances approfondies en calcul différentiel, bien que cela soit utile pour certaines dérivées.

MOTS-CLÉS

Dérivée p. 812
Dérivée partielle p. 818
Effet marginal p. 806, 807
Effet marginal décroissant p. 812
Élasticité p. 814
Fonction exponentielle p. 816

Fonction linéaire p. 806
Fonction logarithmique p. 813
Fonction non-linéaire p. 811
Logarithme naturel p. 813
Médiane p. 806
Modèle à élasticité constante p. 815
Moyenne p. 805
Opérateur de sommation p. 804
Ordonnée à l'origine p. 806
Pente p. 806
Pourcentage de variation p. 809
Semi-élasticité p. 816
Statistique descriptive p. 805
Toutes choses égales par ailleurs p. 808
Variation en point de pourcentage p. 810
Variation proportionnelle p. 809
Variation relative p. 809

PROBLÈMES

1. Le tableau suivant contient la dépense mensuelle de logement pour 10 familles

Famille	Dépense Mensuelle de Logement (Dollars)
1	300
2	440
3	350
4	1 100
5	640
6	480
7	450
8	700
9	670
10	530

i. À combien est égale la moyenne de la dépense de logement ?

ii. À combien est égale la médiane de la dépense de logement ?

iii. Si les dépenses de logement étaient exprimées en centaines de dollars, plutôt qu'en dollars, quelles seraient les valeurs de la dépense moyenne et de la dépense médiane ?

iv. Supposons que la famille 9 voit sa dépense de logement augmenter à 900 $, et que les dépenses de toutes les autres familles restent les mêmes. Calculer alors la moyenne et la médiane pour la variable « dépense de logement ».

2. Supposez que l'équation suivante décrive la relation entre le nombre moyen de cours manqués durant un semestre (*missed*) et la distance entre le logement principal et l'école (*distance*, mesurée en miles) :

$$missed = 3 + 0{,}2\ distance.$$

i. Tracez cette droite, en nommant correctement les axes. Comment interprétez-vous l'ordonnée à l'origine dans cette équation ?

ii. Quel est le nombre moyen de cours manqués pour un étudiant habitant à 5 miles de l'école ?

iii. Combien de cours de plus un élève habitant à 20 miles de l'école aura-t-il manqué par rapport à un élève habitant à 10 miles ?

3. Dans l'exemple A.2, la quantité de « compact-disc » acheté était une fonction du prix des CD et du salaire de l'individu, en suivant la relation $quantité = 120 - 9{,}8\ prix + 0{,}03\ revenu$. Quelle est la demande de CD si $prix = 15$ et $revenu = 200$? Qu'est-ce que cela suggère sur l'utilisation d'une fonction linéaire pour décrire une courbe de demande ?

4. Supposons que le taux de chômage aux États-Unis diminue de 6,4 % à 5,6 % en une année.

i. Exprimez la baisse du taux de chômage en nombre de point de pourcentage

ii. Exprimez la baisse du taux de chômage en pourcentage de variation

5. Supposons que le rendement associé à la détention d'une action d'une entreprise donnée soit de 15 % une année, et de 18 % l'année suivant. La majorité des actionnaires prétendent que le rendement de l'action a augmenté de 3 %, alors que le chef de la direction annonce que le rendement a augmenté de 20 %. Réconciliez ces deux visions.

6. Supposons que la personne A gagne 35 000 $ par an et que la personne B gagne 42 000 $.

i. Trouvez le pourcentage exact permettant de montrer la différence entre le salaire de B et celui de A.

ii. Maintenant, utilisez la différence du logarithme naturel afin d'approximer cette différence (en pourcentage).

7. Supposons que le modèle suivant décrive la relation entre le salaire annuel (*salaire*) et le nombre d'années d'expérience sur le marché du travail (*exper*) : $\log(salaire) = 10{,}6 + 0{,}027\ exper$.

i. À combien s'élève la variable *salaire* lorsque $exper = 0$? Et lorsque $exper = 5$? (*Astuce* : Vous devez utiliser la fonction exponentielle.)

ii. En utilisant l'équation (A.28), approximez la variation en pourcentage de salaire lorsque *exper* augmente de 5 (années).

iii. Utilisez le résultat de (i) pour calculer le pourcentage de différence exact de salaire lorsque $exper = 0$ et $exper = 5$. Commentez ce résultat par rapport au résultat de l'approximation de (ii).

8. Notons *grthemp* la croissance proportionnelle de l'emploi, au niveau d'un pays, entre 1990 et 1995. Notons *salestax* le taux de taxe de vente, en proportion. Interprétez l'ordonnée à l'origine et la pente de l'équation

$$grthemp = 0{,}043 - 0{,}78\ salestax.$$

9. Supposons que le rendement de certaines récoltes (en boisseau par acre) soit relié au volume d'engrais utilisés (en livres par acre)

$$yield = 120 + 0{,}19\sqrt{fertilizer}.$$

i. Tracez cette fonction dans un graphique, en utilisant quelques valeurs de la variable *fertilizer*.

ii. Décrivez la forme de cette fonction, et comparez-la avec ce qu'aurait été une fonction linéaire décrivant la relation entre *yield* et *fertilizer*.

10. Supposons que dans un État, un test standardisé soit effectué par tous les étudiants diplômés. Notons *score* le score de l'étudiant à ce test. Le résultat à ce test dépend de la taille de la classe de l'étudiant, et suit une relation quadratique telle que :

$$score = 45{,}6 + 0{,}082\,class - 0{,}000147\,class^2,$$

où *class* correspond au nombre d'étudiants par classe.

i. Comment interprétez-vous littéralement la valeur 45,6 dans l'équation ci-dessus ? Cette variable a-t-elle un intérêt particulier ? Justifiez.

ii. À partir de l'équation, définissez la taille optimale d'une classe (c'est-à-dire la taille qui maximise le résultat du test), en arrondissant à l'unité la plus proche ? Quel est le meilleur résultat possible à ce test ?

iii. Tracez un graphique pour illustrer vos réponses à la question (ii).

iv. Pensez-vous que la relation entre les variables *score* et *class* soit déterministique ? Est-il réaliste de penser qu'à partir du moment où vous connaissez la taille de la classe pour un étudiant donné, vous pouvez déterminer avec certitude son score au test ? Justifiez.

11. Considérez la droite

$$y = \beta_0 + \beta_1 x.$$

(i) Soit (x_1, y_1) et (x_2, y_2) deux points sur la droite. Montrez que si $\bar{x} = (x_1 + x_2)/2$ et $\bar{y} = (y_1 + y_2)/2$, alors $(\bar{x}, \bar{y})$ est aussi sur la droite.

(ii) Étendez les résultats de la question (i) à n points sur la droite, $\{(x_i, y_i) : i = 1, \ldots, n\}$.

ANNEXE

B

ÉLÉMENTS DE PROBABILITÉS

Traduction de Michel Beine

Cette annexe couvre les concepts clés de base en probabilités. Les annexes B et C constituent prioritairement des revues de la matière ; le but n'est pas qu'elles remplacent un cours en probabilités et statistiques. Cependant, tous les concepts de probabilités et de statistiques que nous utilisons dans le texte sont couverts dans ces annexes.

La théorie des probabilités est une discipline d'intérêt pour les étudiants en management, en économie, ainsi que dans les autres sciences sociales. Par exemple, considérons le problème d'une compagnie aérienne essayant de fixer le nombre de réservations qu'elle doit accepter pour un vol possédant cent sièges disponibles. Si moins de cent personnes veulent procéder à une réservation, alors ces réservations doivent toutes être acceptées. Mais quid si plus de 100 personnes demandent une réservation ? Une solution sans aucun risque est d'accepter au plus cent réservations. Cependant, parce que certaines personnes font des réservations et ensuite ne se présentent pas pour le vol, il est possible que l'avion ne soit pas plein même si cent réservations ont été enregistrées. Cela résulte alors dans des pertes de revenus pour la compagnie aérienne. Une stratégie différente est de réserver plus de cent sièges et d'espérer que certaines personnes ne se présenteront pas, de sorte que le nombre final des passagers soit aussi proche que possible de cent. Cette politique court le risque que la compagnie aérienne ait à dédommager les personnes qui seront inévitablement exclues du vol du fait d'un nombre de réservations excédant la capacité (soit la pratique de« surréservation » ou « surbooking »).

Une question naturelle dans ce contexte est la suivante : peut-on déterminer le nombre optimal de réservations que la compagnie aérienne devrait accepter ? C'est un problème non trivial. Cependant, étant donné une certaine information (sur les coûts des compagnies aériennes et la fréquence à laquelle les gens se présentent après réservation), nous pouvons utiliser les bases de la théorie des probabilités pour arriver à une solution.

B.1 VARIABLES ALÉATOIRES ET LEURS DISTRIBUTIONS DE PROBABILITÉ

Supposons que nous lancions une pièce 10 fois et que nous comptions le nombre de fois où la pièce tombe du côté face. Ceci est un exemple d'une **expérience [aléatoire]**. Généralement une expérience [aléatoire] consiste en toute procédure pouvant, au moins en théorie, être répétée à l'infini, et possédant un ensemble clairement établi de résultats [possibles]. Nous pourrions, en principe, continuer la procédure du lancer de pièces un grand nombre de fois. Avant de lancer la pièce, nous savons que le nombre de faces qui apparaissent est un nombre entier compris entre 0 et 10, si bien que les résultats de l'expérience [aléatoire] sont bien définis.

Une **variable aléatoire** est une variable qui prend des valeurs numériques et possède un résultat qui est déterminé par une expérience [aléatoire]. Dans l'exemple du lancer de la pièce, le nombre de faces qui apparaissent dans 10 lancers de la pièce est un exemple d'une variable aléatoire. Avant de lancer la pièce 10 fois, nous ne savons pas combien de fois la pièce tombera du côté face. Une fois que nous avons lancé la pièce 10 fois et que nous comptons le nombre de faces, nous obtenons le résultat de la variable aléatoire pour ce tirage particulier de l'expérience [aléatoire]. Un autre tirage peut produire un résultat différent.

Dans l'exemple des réservations de la compagnie aérienne que nous avons introduit précédemment, le nombre de personnes se présentant pour le vol est une variable aléatoire : avant chaque vol particulier, nous ne savons pas combien de personnes se présenteront.

Pour analyser les données collectées dans les sciences du management et les sciences sociales, il est important d'avoir une connaissance de base des variables aléatoires et de leurs propriétés. En suivant les conventions habituelles en probabilités et statistiques dans les annexes B et C, nous notons les variables aléatoires par des lettres en majuscule, habituellement W, X, Y, et Z ; les réalisations spécifiques de ces variables étant notées par les lettres correspondantes en minuscule, w, x, y, et z. Par exemple, dans l'expérience du lancer de la pièce, on va noter X le nombre de faces apparaissant dans un lancer de 10 pièces. Dans ce cas,

X n'est pas associé à une valeur quelconque particulière mais nous savons que X prendra une valeur dans l'ensemble {0, 1, 2, ..., 10}. Un résultat particulier est, disons, $x = 6$.

Nous indiquons des ensembles larges de variables aléatoires en utilisant des suffixes. Par exemple, si nous enregistrons le revenu de l'année dernière pour 20 ménages choisis aléatoirement aux États-Unis, nous pouvons noter ces variables aléatoires par X_1, X_2, ..., X_{20} ; les réalisations particulières seraient alors notées x_1, x_2, ..., x_{20}.

Comme indiqué dans la définition, les variables aléatoires sont toujours définies de manière à prendre des valeurs numériques, même lorsqu'elles décrivent des événements qualitatifs. Par exemple, si l'on considère le lancer d'une pièce unique, pour lequel les deux résultats possibles sont pile et face. On définit alors une variable aléatoire de la manière suivante : $X = 1$ si la pièce tombe sur face, et $X = 0$ si la pièce tombe sur pile.

Une variable aléatoire qui peut prendre les valeurs zéro et un est appelée **une variable aléatoire de Bernoulli (binaire)**. Dans la théorie des probabilités, il est habituel d'appeler l'événement $X = 1$ un « succès » et l'événement $X = 0$ un « échec ». Pour une application particulière, la nomenclature succès – échec peut ne pas correspondre à notre notion personnelle d'un succès ou d'un échec, mais c'est une terminologie utile que nous adopterons.

Variables aléatoires discrètes

Une **variable aléatoire discrète** peut prendre seulement un nombre fini ou un nombre infini dénombrable de valeurs possibles. La notion d'« infini dénombrable » signifie que même si la variable aléatoire peut prendre un nombre infini de valeurs, ces valeurs peuvent être mises dans une correspondance directe avec des entiers positifs. Parce que la distinction entre « infini dénombrable » et « infini non dénombrable » est quelque peu subtile, nous nous concentrerons seulement sur des variables aléatoires discrètes qui peuvent prendre un nombre fini de valeurs. Larsen et Marx (1986, Chapitre 3) fournissent un traitement détaillé de cette question.

Une variable aléatoire de Bernoulli est l'exemple le plus simple d'une variable aléatoire discrète. La seule chose dont nous avons besoin pour décrire complètement le comportement d'une variable aléatoire de Bernoulli est la probabilité de prendre la valeur un. Dans l'exemple du lancer de pièce, si la pièce est « non biaisée », alors $P(X = 1) = 1/2$ (qu'il convient de lire comme suit : « la probabilité que X égale un est un demi »). Parce que les probabilités doivent sommer à un, nous obtenons en outre que $P(X = 0) = 1/2$.

Les chercheurs en sciences sociales sont intéressés par d'autres choses que des lancers de pièce, si bien que nous devons tenir compte de situations plus générales. Une fois de plus, considérez l'exemple dans lequel la compagnie aérienne doit décider combien de personnes doivent réserver un vol qui possède 100 sièges disponibles. Ce problème peut être analysé dans le contexte de plusieurs variables aléatoires de Bernoulli comme suit : pour un client sélectionné aléatoirement, définissez une variable aléatoire de Bernoulli comme $X = 1$ si la personne se présente pour la réservation, et $X = 0$ sinon.

Il n'y a pas de raison de penser que la probabilité qu'un client particulier se présente est 1/2 ; en principe, la probabilité peut être n'importe quel nombre entre zéro et un. Appelons ce nombre θ si bien que

$$P(X = 1) = \theta \qquad \textbf{[B.1]}$$

$$P(X = 0) = 1 - \theta. \qquad \textbf{[B.2]}$$

Par exemple, si, $\theta = 0{,}75$, alors il y a 75 % de chances qu'un consommateur se présente après avoir fait une réservation et 25 % de chances que le client ne se présente pas. Intuitivement, la valeur de θ est cruciale pour déterminer la stratégie de la compagnie aérienne dans la prise des réservations. Les méthodes pour estimer θ, étant donné les données historiques relatives aux réservations des compagnies aériennes, constituent une question d'intérêt en statistiques, une discipline que nous couvrirons dans l'annexe C.

Plus généralement, toute variable aléatoire discrète est entièrement décrite en établissant ses valeurs possibles et les probabilités associées que la variable prenne chaque valeur. Si X prend les k valeurs possibles $\{x_1, \ldots, x_k\}$, alors les probabilités sont définies par

$$p_j = P(X = x_j), j = 1, 2, \ldots, k, \qquad \textbf{[B.3]}$$

où chaque p_j est entre 0 et 1 et

$$p_1 + p_2 + \ldots + p_k = 1. \qquad \textbf{[B.4]}$$

L'équation (B.3) se lit comme : « la probabilité que X prenne la valeur x_j est égale à p_j. »

Les équations (B.1) et (B.2) montrent que les probabilités de succès et d'échec pour une variable aléatoire de Bernoulli sont déterminées totalement par la valeur de θ. Dans la mesure où les variables aléatoires de Bernoulli sont si courantes, nous réservons une notation spéciale pour elles : $X \sim \text{Bernoulli}(\theta)$ se lit comme « X suit une distribution de Bernoulli avec probabilité de succès égale à θ. »

La **fonction de densité de probabilité** (**FDP**) de X résume l'information concernant les réalisations possibles de X et les probabilités correspondantes :

$$f(x_j) = p_j, j = 1, 2, \ldots, k, \qquad \textbf{[B.5]}$$

avec $f(x) = 0$ pour tout x différent de x_j pour un j quelconque. En d'autres termes, pour tout nombre réel x, $f(x)$ est la probabilité que la variable aléatoire X prenne la valeur particulière x. Lorsque l'on considère plus d'une variable aléatoire, il est parfois utile d'indexer la FDP en question : f_X est la FDP de X, f_Y est la FDP de Y et ainsi de suite.

Étant donné la FDP de toute variable aléatoire discrète, il est simple de calculer la probabilité de n'importe quel événement impliquant cette variable aléatoire. Par exemple, supposons que X soit le nombre de lancers-francs réussis par un joueur de basket-ball bénéficiant de deux essais, si bien que X prend trois valeurs $\{0,1,2\}$. Supposons que la FDP de X est donnée par

$$f(0) = 0{,}20, f(1) = 0{,}44, \text{ et } f(2) = 0{,}36$$

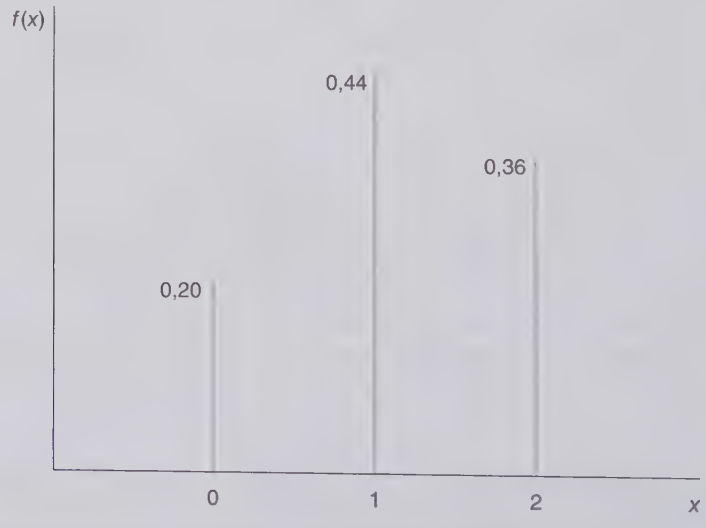


Figure B.1 *La FDP du nombre de lancers francs réussis après deux essais.*

Les trois probabilités somment à un, et c'est obligatoire. En utilisant cette FDP, on peut calculer la probabilité que le joueur réussira au moins un lancer franc : $P(X \geq 1) = P(X = 1) + P(X = 2) = 0{,}44 + 0{,}36 = 0{,}80$. La FDP de X est reproduite à la figure B.1.

Variables aléatoires continues

Une variable X est une variable aléatoire continue si elle prend n'importe quelle valeur réelle avec une probabilité zéro. Cette définition est quelque peu contre-intuitive parce que dans n'importe quelle application, nous observons un résultat donné pour une variable aléatoire. L'idée est qu'une variable aléatoire continue X prend tellement de valeurs possibles que nous ne pouvons pas les dénombrer ou les faire correspondre avec des entiers positifs, si bien que la cohérence logique nous amène à ce que X peut prendre n'importe quelle valeur avec une probabilité zéro. Si les mesures sont toujours discrètes en pratique, les variables aléatoires qui prennent des valeurs numériques se traitent avantageusement comme continues. Par exemple, la mesure la plus fine du prix d'un bien est exprimée en termes de centimes. On peut imaginer établir toutes les valeurs possibles du prix (même si la liste peut continuer de manière indéfinie), ce qui techniquement considère le prix comme une variable aléatoire discrète. Cependant, il y a tant de valeurs possibles du prix qu'utiliser la mécanique des variables aléatoires discrètes n'est pas possible.

Nous pouvons définir une fonction de densité de probabilité pour les variables aléatoires continues, et, comme pour les variables aléatoires discrètes, la FDP fournit l'information sur les résultats possibles de la variable aléatoire. Cependant, comme cela n'a pas de sens de discuter la probabilité qu'une variable aléatoire continue prenne une valeur particulière, nous utilisons la FDP d'une variable aléatoire continue seulement pour calculer des événements impliquant un intervalle de valeurs. Par exemple, si a et b sont des constantes, pour lesquelles $a < b$, la probabilité que X s'établisse entre les nombres a et b, $P(a \leq X \leq b)$, est l'aire située sous la FDP entre les points a et b comme indiqué à la figure B.2. Si vous êtes familiers avec l'algèbre, vous reconnaîtrez ceci comme l'intégrale de la fonction f entre les points a et b. L'aire totale en-dessous de la FDP doit toujours être égale à un.

Lorsque l'on calcule les probabilités pour des variables aléatoires continues, le plus facile est de travailler avec la **fonction de distribution cumulée (FDC)**. Si X est n'importe quelle variable aléatoire, alors sa FDC est définie pour tout nombre réel x par

$$F(x) = P(X \leq x). \qquad \textbf{[B.6]}$$

Pour les variables aléatoires discrètes, (B.6) est obtenue en établissant la somme des FDP pour toutes les valeurs x_j telles que $x_j \leq x$. Pour une variable aléatoire continue, $F(x)$ est l'aire sous la FDP, f, à gauche du point x. Parce que $F(x)$ est simplement une probabilité, sa valeur est toujours comprise entre 0 et 1. En outre, si $x_1 < x_2$, alors $P(X \leq x_1) \leq P(X \leq x_2)$, à savoir, $F(x_1) \leq F(x_2)$. Cela signifie qu'une FDC est une fonction croissante (ou du moins non décroissante) de x.

Deux propriétés importantes des FDC utiles pour calculer les probabilités sont les suivantes :

$$\textit{Pour tout nombre } c,\ P(X > c) = 1 - F(c). \qquad \textbf{[B.7]}$$

$$\textit{Pour tous les nombres } a < b,\ P(a < X \leq b) = F(b) - F(a). \qquad \textbf{[B.8]}$$

Dans notre analyse de l'économétrie, nous utiliserons les FDC pour calculer les probabilités seulement pour les variables aléatoires continues, ce qui implique que cela n'a pas d'importance si les inégalités dans les probabilités sont strictes ou non. Dès lors, pour une variable aléatoire continue X,

$$P(X \geq c) = P(X > c), \qquad \textbf{[B.9]}$$

et

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b). \qquad \textbf{[B.10]}$$

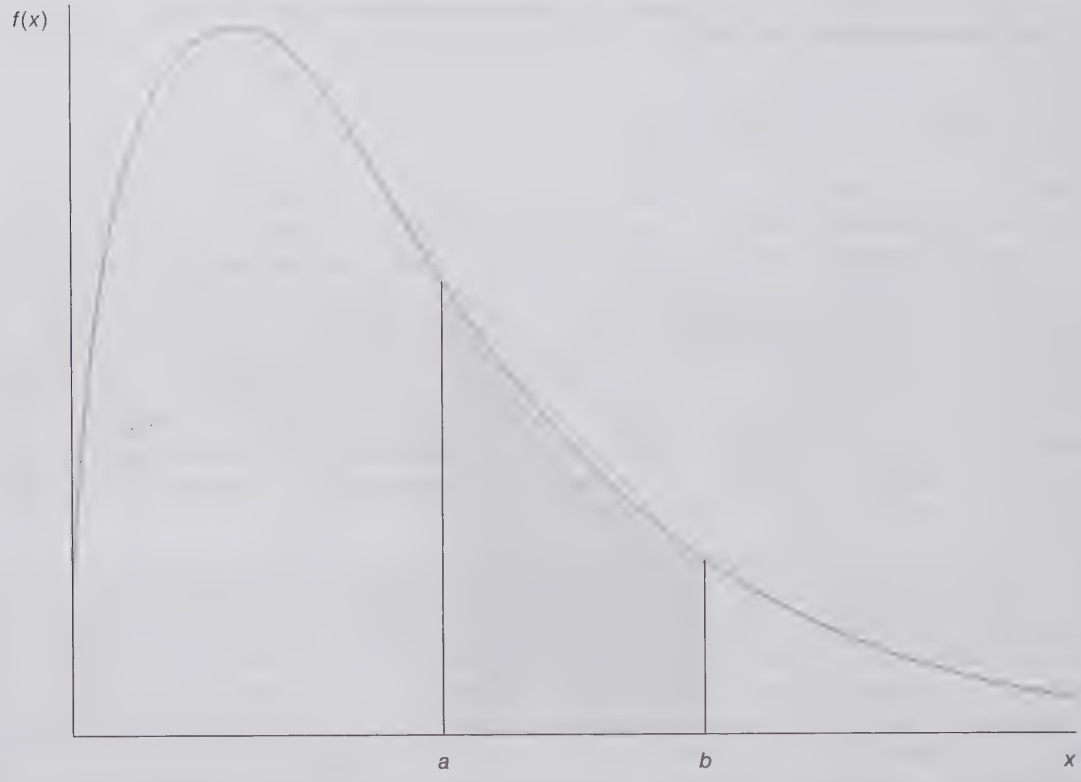


Figure B.2 *La probabilité que X se situe entre les points a et b.*

Combinées avec (B.7) et (B.8), les équations (B.9) et (B.10) étendent significativement les calculs de probabilité qui peuvent être faits en utilisant les FDC continues.

Les fonctions de distribution cumulées ont été tabulées pour toutes les distributions continues importantes en probabilité et statistique. La plus connue d'entre elles est la distribution normale, que nous couvrirons ainsi que des distributions associées, dans la section B.5.

B.2 DISTRIBUTIONS JOINTES, DISTRIBUTIONS CONDITIONNELLES, ET INDÉPENDANCE

En économie, nous nous intéressons souvent à l'occurrence d'événements impliquant plus d'une variable aléatoire. Par exemple, dans l'exemple de réservation de la compagnie aérienne mentionné auparavant, la compagnie aérienne peut s'intéresser à la probabilité qu'une personne qui fait une réservation se présente et est un voyageur d'affaires ; ceci est un exemple d'une *probabilité jointe*. Alternativement, la compagnie aérienne peut s'intéresser à la *probabilité conditionnelle* suivante : conditionnellement au fait que la personne est un voyageur d'affaires, quelle est la probabilité qu'il ou elle se présente ? Dans les deux sous-sections suivantes, nous formalisons les notions de distribution jointe et conditionnelle et la notion importante d'*indépendance* des variables aléatoires.

Distributions jointes et indépendance

Soit X et Y des variables aléatoires discrètes. Alors, (X,Y) possède une **distribution jointe**, qui est totalement décrite par la *fonction de densité de probabilité jointe* de (X,Y) :

$$f_{XY}(x,y) = P(X = x, Y = y), \qquad \textbf{[B.11]}$$

où le terme de droite est la probabilité que $X = x$ et $Y = y$. Lorsque X et Y sont continues, une FDP jointe peut aussi être définie, mais nous ne couvrirons pas de tels détails parce que les FDP jointes pour les variables aléatoires continues ne sont pas utilisées explicitement dans ce texte.

Dans un cas, il est facile d'obtenir la FDP jointe si on nous donne les FDP de X et de Y. En particulier, les variables aléatoires X et Y sont dites indépendantes, si et seulement si,

$$f_{XY}(x,y) = f_X(x)f_Y(y) \qquad \textbf{[B.12]}$$

pour tout x et y, où f_X est la FDP de X et f_Y est la FDP de Y. Dans le cas de plus d'une variable aléatoire, les FDP sont souvent appelées *fonctions de densité de probabilité marginales* pour les distinguer de la FDP jointe f_{XY}. Cette définition de la dépendance est valable pour les variables aléatoires discrètes et continues.

Pour comprendre le sens de (B.12), il est plus facile de traiter le cas discret. Si X et Y sont des variables aléatoires discrètes, alors (B.12) est le même que

$$P(X = x, Y = y) = P(X = x)P(Y = y)\ ; \qquad \textbf{[B.13]}$$

En d'autres termes, la probabilité que $X = x$ et $Y = y$ est le produit des deux probabilités $P(X = x)$ et $P(Y = y)$. Une implication de (B.13) est que les probabilités jointes sont relativement faciles à calculer, puisqu'elles ne demandent que la connaissance de $P(X = x)$ et $P(Y = y)$.

Si des variables aléatoires ne sont pas indépendantes, alors on dit qu'elles sont dépendantes.

EXEMPLE B.1
Tirs de lancers francs

Considérons un joueur de basket-ball lançant deux lancers francs. Soit X une variable aléatoire de Bernoulli égale à un s'il réussit le premier lancer franc, 0 sinon. Soit Y une variable aléatoire de Bernoulli égale à un s'il réussit le second lancer franc, 0 sinon. Supposons qu'il est un lanceur de lancer franc avec 80 % de réussite, si bien que $P(X = 1) = P(Y = 1) = 0{,}8$. Quelle est la probabilité que le joueur réussisse ses deux lancers-francs ?

Si X et Y sont indépendants, nous pouvons répondre facilement à cette question : $P(X = 1,Y = 1) = P(X = 1)P(Y = 1) = (0{,}8)(0{,}8) = 0{,}64$. Dès lors, il y a une 64 % de chances de réussir les deux lancers francs. Si la probabilité de réussir le second lancer franc dépend de ce qui s'est passé pour le premier – à savoir, X et Y ne sont pas indépendants – alors ce calcul simple n'est pas valable.

L'indépendance des variables aléatoires est un concept très important. Dans la sous-section suivante, nous montrerons que si X et Y sont indépendants, alors connaître le résultat de X ne change pas les probabilités des résultats possibles de Y, et vice-versa. Un fait utile à propos de l'indépendance est que si X et Y sont indépendants et si nous définissons de nouvelles variables aléatoires pour n'importe quelles fonctions g et h, alors ces nouvelles variables aléatoires $g(X)$ et $h(Y)$ sont aussi indépendantes.

Il n'y a pas lieu de s'arrêter à deux variables aléatoires. Si $X_1, X_2, \ldots, X_n$ sont des variables aléatoires discrètes, alors leur FDP jointe est $f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. Les variables aléatoires $X_1, X_2, \ldots, X_n$ sont des **variables aléatoires indépendantes** si, et seulement si, leur FDP jointe est le produit des FDP individuelles pour tout $(x_1, x_2, \ldots, x_n)$. Cette définition de l'indépendance est aussi valable pour les variables aléatoires continues.

La notion d'indépendance joue un rôle important pour obtenir certaines des distributions classiques en probabilités et statistiques. Auparavant, nous avons défini une variable aléatoire de Bernoulli comme une variable aléatoire 0-1 indiquant si un certain événement était survenu. Nous nous intéressons souvent au nombre de succès dans une séquence de tirages de Bernoulli indépendants. Un exemple standard de tirages indépendants est le lancer d'une pièce un grand nombre de fois. Dans la mesure où le résultat de n'importe quel lancer particulier n'a rien à voir avec les résultats des autres lancers, l'indépendance est une hypothèse correcte.

L'indépendance est souvent une approximation raisonnable dans des situations plus compliquées. Dans l'exemple de réservation de la compagnie aérienne, supposons que la compagnie aérienne accepte n réservations pour un vol particulier. Pour chaque $i = 1, 2, \ldots, n$, dénotons par Y_i la variable aléatoire de Bernoulli indiquant si un consommateur i se présente : $Y_i = 1$ si le consommateur se présente, et $Y_i = 0$ sinon. Si θ dénote une fois de plus la probabilité de succès (utiliser la réservation), chaque Y_i suit une distribution de Bernoulli(θ). Comme approximation, nous pouvons supposer que les Y_i sont indépendants les uns des autres, bien que ce ne soit pas exactement vrai en réalité : certaines personnes voyagent en groupe, ce qui signifie que le fait qu'une personne se présente ou non n'est pas vraiment indépendant du fait que les autres personnes se présentent. Modéliser ce type de dépendance est néanmoins complexe, si bien que nous pouvons être enclins à utiliser l'indépendance comme une approximation.

La variable d'intérêt primaire est le nombre total de consommateurs se présentant à partir des n réservations ; appelons cette variable X. Puisque chaque Y_i est un quand une personne se présente, nous pouvons écrire $X = Y_1 + Y_2 + \ldots + Y_n$. Maintenant, en supposant que chaque Y_i a une probabilité de succès θ et que les Y_i sont indépendants, on peut montrer que X suit une **distribution binomiale**. À savoir, la fonction de densité de probabilité de X est

$$f(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \; x = 0, 1, 2, \ldots, n, \qquad \textbf{[B.14]}$$

où $\binom{n}{x} = \dfrac{n!}{x!\,(n-x)!}$ et pour tout entier n, $n!$ (lisez « factorielle n ») est défini comme $n! = n \cdot (n-1) \cdot (n-2) \ldots 1$. Par convention, $0! = 1$. Quand une variable aléatoire X possède la FDP donnée par (B.14), on écrit $X \sim \text{Binomiale}(n, \theta)$. L'équation (B.14) peut être utilisée pour calculer $P(X = x)$ pour toute valeur de x de 0 à n.

Si le vol possède 100 sièges disponibles, la compagnie aérienne s'intéresse à $P(X > 100)$. Supposons qu'initialement, n est égal à 120, si bien que la compagnie aérienne accepte 120 réservations, et supposons que la probabilité que chaque personne se présente est $\theta = 0{,}85$. Alors, $P(X > 100) = P(X = 101) + P(X = 102) + \ldots + P(X = 120)$ et chacune des probabilités dans la somme peut être trouvée à partir de l'équation (B.14) avec $n = 120$, $\theta = 0{,}85$, et la valeur appropriée de x (101 à 120). C'est un calcul difficile à la main, mais beaucoup de logiciels statistiques ont des commandes pour calculer ce type de probabilité. Dans ce cas, la probabilité que plus de 100 personnes se présentent et à peu près de 0,659, ce qui est probablement un risque de surréservation trop risqué pour être toléré par une compagnie aérienne. Si en lieu et place, le nombre de réservations est de 110, la probabilité que plus de 100 passagers se présentent est seulement d'à peu près 0,024.

Distributions conditionnelles

En économétrie, nous nous intéressons habituellement à la manière dont une variable aléatoire, appelons-la Y, est reliée à une ou plusieurs autres variables. Pour l'instant, supposons qu'il n'y ait qu'une variable, appelons-la X et que nous nous intéressions à ses effets. Le maximum que nous puissions savoir à propos de la manière dont X affecte Y est contenu dans la **distribution conditionnelle** de Y étant donné X. Cette information est résumée par la *fonction de densité de probabilité conditionnelle*, définie par

$$f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x) \qquad \textbf{[B.15]}$$

pour toutes les valeurs de x telles que $f_X(x) > 0$. L'interprétation est aisée lorsque X et Y sont discrètes. Dans ce cas,

$$f_{Y|X}(y|x) = P(Y = y|X = x), \qquad \textbf{[B.16]}$$

où le terme de droite se lit comme « la probabilité que $Y = y$ étant donné que $X = x$. » Quand Y est continue, $f_{Y|X}(y|x)$ n'est pas directement interprétable comme une probabilité, pour les raisons exposées ci-dessus, mais les probabilités conditionnelles s'obtiennent en calculant les aires en-dessous des FDP conditionnelles.

Une caractéristique importante des distributions conditionnelles est que, si X et Y sont des variables aléatoires indépendantes, la connaissance de la valeur prise par X ne nous apprend rien sur la probabilité que Y prenne différentes valeurs (et vice-versa). En clair, $f_{Y|X}(y|x) = f_Y(y)$, et $f_{X|Y}(x|y) = f_X(x)$.

EXEMPLE B.2
Tir de lancers-francs

On considère à nouveau l'exemple du lancer au basket-ball, dans lequel deux lancers francs sont tentés. Supposons que la densité conditionnelle est

$$f_{Y|X}(1|1) = 0{,}85, f_{Y|X}(0|1) = 0{,}15$$
$$f_{Y|X}(1|0) = 0{,}70, f_{Y|X}(0|0) = 0{,}30.$$

Cela signifie que la probabilité qu'un joueur réussisse le second lancer franc dépend du fait que le premier lancer franc a été réussi : si le premier lancer franc est réussi, la chance de réussir le second est 0,85 ; si le premier lancer franc est raté, la chance de réussir le second est 0,70. Ceci implique que X et Y ne sont *pas* indépendants ; ils sont dépendants.

Nous pouvons toujours calculer $P(X = 1, Y = 1)$ à condition que l'on connaisse $P(X = 1)$. Supposons que la probabilité de réussir le premier lancer franc est 0,8, c'est-à-dire, $P(X = 1) = 0{,}8$. Alors, à partir de (B.15), nous avons

$$P(X = 1, Y = 1) = P(Y = 1|X = 1) \cdot P(X = 1) = (0{,}85)(0{,}8) = 0{,}68.$$

B.3 CARACTÉRISTIQUES DES DISTRIBUTIONS DE PROBABILITÉ

Donc beaucoup de cas, nous nous intéressons seulement à quelques aspects des distributions des variables aléatoires. Les caractéristiques qui sont intéressantes peuvent être classées dans trois catégories : des mesures de tendance centrale, des mesures de variabilité et d'écart, et des mesures d'association entre deux variables aléatoires. Nous couvrons la dernière de ces mesures dans la section B.4.

Une mesure de tendance centrale : la valeur espérée

La valeur espérée est un des concepts probabilistes les plus importants que nous rencontrerons dans notre étude de l'économétrie. Si X est une variable aléatoire, la **valeur espérée** (espérance) de X, dénotée $E(X)$ et souvent μ_X ou simplement μ, est la moyenne pondérée de toutes les valeurs possibles prises par X. Les poids sont déterminés par la fonction densité de probabilité. Parfois, la valeur espérée est appelée la *moyenne de la population*, spécialement lorsque nous voulons insister sur le fait que X représente une certaine variable dans la population.

La définition précise de la valeur espérée est la plus simple dans le cas où X est une variable aléatoire discrète prenant un nombre fini de valeurs, disons, $\{x_1, \ldots, x_k\}$. Soit $f(x)$ la fonction de densité de probabilité de X. La valeur espérée de X est la moyenne pondérée

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_k f(x_k) \equiv \sum_{j=1}^{k} x_j f(x_j). \qquad \textbf{[B.17]}$$

Celle-ci se calcule aisément étant donné les valeurs de la FDP pour chaque résultat possible de X.

EXEMPLE B.3
Calcul d'une valeur espérée

Supposons que X prenne les valeurs –1, 0, et 2 avec probabilités 1/8, 1/2, et 3/8, respectivement. Dans ce cas,

$$E(X) = (-1)\cdot(1/8) + 0\cdot(1/2) + 2\cdot(3/8) = 5/8.$$

Cet exemple illustre quelque chose de curieux à propos des valeurs espérées : la valeur espérée de X peut être un nombre qui n'est pas un résultat possible de X. Nous savons que X prend les valeurs –1, 0, ou 2, et pourtant sa valeur espérée est 5/8. Ceci rend la valeur espérée impropre à résumer la tendance centrale de certaines variables aléatoires discrètes, mais les calculs du type de ceux que nous avons étudiés peuvent être utiles, comme nous le verrons plus tard.

Si X est une variable aléatoire continue, alors $E(X)$ est définie comme une intégrale :

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx, \qquad \textbf{[B.18]}$$

que nous supposons bien définie. Ceci peut tout de même être interprété comme une moyenne pondérée. Pour les distributions continues les plus courantes, $E(X)$ est un nombre qui est un résultat possible de X. Dans ce texte, nous n'avons pas besoin de calculer les valeurs espérées en utilisant les intégrales ; nous nous appuierons néanmoins sur des résultats bien connus de probabilité pour les valeurs espérées des variables aléatoires spéciales.

Étant donné une variable aléatoire X et une fonction $g(\bullet)$, nous pouvons créer une nouvelle variable aléatoire $g(X)$. Par exemple, si X est une variable aléatoire, alors X^2 et $\log(X)$ (si $X > 0$) le sont aussi. La valeur espérée de $g(X)$ est, une fois de plus, simplement une moyenne pondérée :

$$E[g(X)] = \sum_{j=1}^{k} g(x_j) f_x(x_j) \qquad \textbf{[B.19]}$$

ou, pour une variable aléatoire continue,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_x(x) dx. \qquad \textbf{[B.20]}$$

EXEMPLE B.4
Valeur espérée de X^2

Pour une variable aléatoire de l'exemple B.3, soit $g(X) = X^2$. Dans ce cas,

$$E(X^2) = (-1)^2(1/8) + (0)^2(1/2) + (2)^2(3/8) = 13/8.$$

Dans l'exemple B.3, nous avons calculé $E(X) = 5/8$, si bien que $[E(X)]^2 = 25/64$. Ceci montre que $E(X^2)$ n'est *pas* le même que $[E(X)]^2$. En fait, pour une fonction non linéaire $g(X)$, $E[g(X)] \neq g[E(X)]$ (sauf dans des cas très spécifiques).

Si X et Y sont des variables aléatoires, alors $g(X,Y)$ est une variable aléatoire de n'importe quelle fonction g, et de ce fait nous pouvons définir son espérance. Quand X et Y sont toutes deux discrètes, et

qu'elles prennent respectivement les valeurs $\{x_1, x_2, \ldots, x_k\}$ et $\{y_1, y_2, \ldots, y_m\}$, la valeur espérée est donnée par :

$$E[g(X,Y)] = \sum_{h=1}^{k}\sum_{j=1}^{m} g(x_h, y_j) f_{X,Y}(x_h, y_j)$$

où $f_{X,Y}$ est la FDP jointe de (X,Y). La définition est plus compliquée pour des variables aléatoires continues dans la mesure où cela implique l'intégration ; nous n'en avons pas besoin ici. L'extension à plus de deux variables aléatoires est directe.

Propriété des valeurs espérées

En économétrie, nous ne sommes pas tellement intéressés par le calcul de valeurs espérées à partir de différentes distributions ; les calculs les plus importants ont été faits très souvent et nous les considérerons comme acquis. Nous devrons manipuler certaines valeurs espérées en utilisant quelques règles simples. Elles sont tellement importantes que nous leur donnons ici un intitulé :

Propriété E.1 : Pour toute constante c, $E(c) = c$.

Propriété E.2 : Pour toutes constantes a et b, $E(aX + b) = aE(X) + b$.

Une implication utile de E.2 est que, si $\mu = E(X)$, et si nous définissons une nouvelle variable aléatoire comme $Y = X - \mu$, alors $E(Y) = 0$; dans E.2, prenons $a = 1$ et $b = -\mu$.

Comme exemple de la propriété E.2, soit X la température mesurée en degrés Celsius à midi un jour particulier en un endroit donné ; supposons que la valeur attendue est $E(X) = 25$. Si Y est la température mesurée en degrés Fahrenheit, alors $Y = 32 + (9/5)X$. À partir de la propriété E.2, la température attendue en degrés Fahrenheit est donnée par : $E(Y) = 32 + (9/5)$ soit $E(X) = 32 + (9/5)\cdot 25 = 77$ degrés Farenheit.

En règle générale, il est aisé de calculer la valeur attendue d'une fonction linéaire de plusieurs variables aléatoires.

Propriété E.3 : Si $\{a_1, a_2, \ldots, a_n\}$ sont des constantes, $\{X_1, X_2, \ldots, X_n\}$ des variables aléatoires, alors :

$$E(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \ldots + a_nE(X_n).$$

Ou, en utilisant la notation de sommation,

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i). \qquad \textbf{[B.21]}$$

Comme cas particulier, nous avons (avec chaque $a_i = 1$)

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i). \qquad \textbf{[B.22]}$$

de telle sorte que la valeur espérée de la somme est la somme des valeurs espérées. Cette propriété est souvent utilisée pour les dérivations en statistiques.

EXEMPLE B.5
Trouver le revenu espéré

Soit X_1, X_2, et X_3 le nombre de pizzas de petite, moyenne, et grande taille respectivement, vendues durant une journée dans une échoppe de pizzas. Ce sont des variables aléatoires avec les valeurs espérées $E(X_1) = 25$, $E(X_2) = 57$, et $E(X_3) = 40$. Les prix des pizzas de petite, moyenne et grande taille sont respectivement 5,50, 7,60, et 9,15 USD. Dès lors, le revenu espéré des ventes de pizzas pour une journée donnée est

$$\begin{aligned} E(5{,}50\,X_1 + 7{,}60\,X_2 + 9{,}15\,X_3) &= 5{,}50\,E(X_1) + 7{,}60\,E(X_2) + 9{,}15\,E(X_3) \\ &= 5{,}50(25) + 7{,}60(57) + 9{,}15(40) = 936{,}70, \end{aligned}$$

à savoir, 936,70 USD. Le revenu réel de n'importe quel jour différera généralement de cette valeur, mais ceci est le revenu *espéré*.

Nous pouvons aussi utiliser la propriété E.3 pour montrer que si $X \sim \text{Binomiale}(n,\theta)$, alors $E(X) = n\theta$. À savoir, le nombre attendu de succès dans un tirage de Bernoulli à n tirages est simplement le nombre de tirages multiplié par la probabilité de succès de tout tirage particulier. Ceci se voit aisément en écrivant X comme $X = Y_1 + Y_2 + \ldots + Y_n$, où chaque $Y_i \sim \text{Binomiale}(\theta)$. Dans ce cas,

$$E(X) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} \theta = n\theta.$$

Nous pouvons appliquer ceci à l'exemple des réservations de la compagnie aérienne, où la compagnie aérienne enregistre $n = 120$ réservations, et où la probabilité de se présenter est $\theta = 0{,}85$. Le nombre attendu de personnes se présentant est $120(0{,}85) = 102$. Dès lors, s'il y a 100 sièges disponibles, le nombre attendu de personnes se présentant est trop élevé ; ceci permet à la compagnie de savoir si c'est une bonne idée d'enregistrer 120 réservations.

En fait, ce que la compagnie devrait faire c'est définir une fonction de profit tenant compte à la fois du revenu net généré par chaque siège vendu et par le coût par passager de ne pas être retenu pour le vol. La fonction de profit est aléatoire du fait que le nombre réel de personnes se présentant est aléatoire. Soit r le revenu net de chaque passager. (Vous pouvez imaginer ceci comme le prix du ticket pour faire simple.) Soit c la compensation donnée à chaque voyageur non retenu pour le vol. Ni r ni c ne sont aléatoires ; ils sont supposés connus de la compagnie aérienne. Soit Y les profits du vol. Dès lors, avec 100 sièges disponibles,

$$\begin{aligned} Y &= rX \text{ si } X \leq 100 \\ &= 100r - c(X - 100) \text{ si } X > 100. \end{aligned}$$

La première équation donne le profit si au plus 100 personnes se présentent pour le vol ; la seconde équation est le profit si plus de 100 personnes se présentent. (Dans le dernier cas, le revenu net de vente des tickets est $100r$, puisque tous les 100 sièges sont vendus, et ensuite $c(X - 100)$ et le coût d'admettre plus de 100 réservations.) En utilisant le fait que X suit une distribution $Binomiale(n,0{,}85)$, où n et le nombre de réservations enregistrées, les profits attendus, $E(Y)$, peuvent être trouvés comme une fonction de n (et de r et de c). Calculer directement $E(Y)$ serait relativement difficile, mais cela peut se faire rapidement en utilisant l'ordinateur. Une fois pour les valeurs de r et c données, la valeur de n qui maximise les profits attendus peut être trouvée en cherchant le résultat pour les différentes valeurs de n.

Une autre mesure de tendance centrale : la médiane

La valeur attendue constitue seulement une possibilité pour définir la tendance centrale d'une variable aléatoire. Une autre mesure de tendance centrale est la médiane. Une définition générale de la médiane n'a pas d'intérêt réel pour ce que nous voulons traiter dans le cadre de cet ouvrage. Si X est continue, alors la médiane

de X, disons m, est la valeur telle que la moitié de l'aire sous la FDP est à droite de m et l'autre moitié de l'aire est à gauche de m.

Quand X est discrète et prend un nombre fini impair de valeurs, la médiane est obtenue en ordonnant les valeurs possibles de X et en sélectionnant la valeur du milieu. Par exemple, si X prend les valeurs $\{-4,0,2,8,10,13,17\}$, alors la valeur médiane de X est 8. Si X prend un nombre pair de valeurs, il y a alors deux valeurs médianes. Parfois, on en prend la moyenne pour obtenir une valeur médiane unique. Dès lors, si X prend les valeurs $\{-5,3,9,17\}$, alors les valeurs médianes sont 3 et 9 ; si nous prenons la moyenne de ces valeurs, nous obtenons une médiane égale à 6.

En général, la médiane, parfois notée $Med(X)$, et la valeur espérée, $E(X)$, sont différentes. Aucune n'est « meilleure » que l'autre comme mesure de tendance centrale ; elles sont toutes les deux valables pour mesurer le centre de la distribution de X. Dans un cas particulier, la médiane est la valeur attendue (ou la moyenne) sont les mêmes. Si X a une **distribution symétrique** autour de la valeur μ, alors μ est à la fois la valeur espérée et la médiane. Mathématiquement, la condition s'écrit $f(\mu + x) = f(\mu - x)$ *pour* tout x. Ce cas est illustré à la figure B.3.

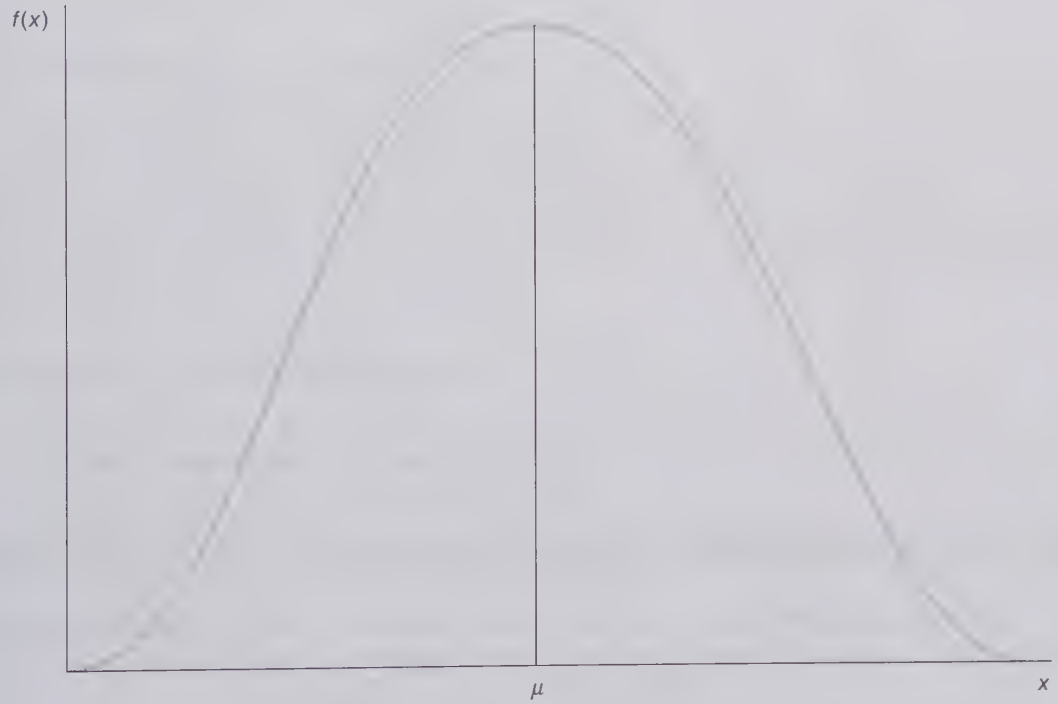


Figure B.3 *Une distribution de probabilité symétrique.*

Mesures de variabilité : variance et écart-type

Bien que la tendance centrale d'une variable aléatoire constitue une mesure utile, elle ne nous renseigne pas intégralement sur les caractéristiques de la distribution d'une variable aléatoire. La figure B.4 montre les FDP de deux variables aléatoires avec la même moyenne. Clairement, la distribution de X est centrée plus étroitement autour de sa moyenne que la distribution de Y. Nous voudrions avoir une manière simple de résumer les différences dans les dispersions des distributions.

Variance

Soit $\mu = E(X)$ pour une variable aléatoire X. Il y a différentes manières de mesurer dans quelle mesure X est éloignée de sa valeur espérée, mais la manière la plus simple est de travailler algébriquement avec la différence au carré, $(X - \mu)^2$. (Le carré élimine le signe de la mesure de distance ; la valeur positive qui en résulte correspond à notre notion intuitive de la distance, et traite les valeurs au-dessus et en-dessous de μ symétriquement.) La distance est elle-même une variable aléatoire puisqu'elle change avec les résultats de X. Comme nous avons besoin d'un nombre pour mesurer la valeur centrale de X, nous avons également besoin d'un nombre qui nous dit dans quelle mesure X est éloigné en moyenne de μ. Un tel nombre est la *variance*, qui nous dit la distance attendue de X par rapport à sa moyenne :

$$Var(X) \equiv E[(X - \mu)^2] \qquad \textbf{[B.23]}$$

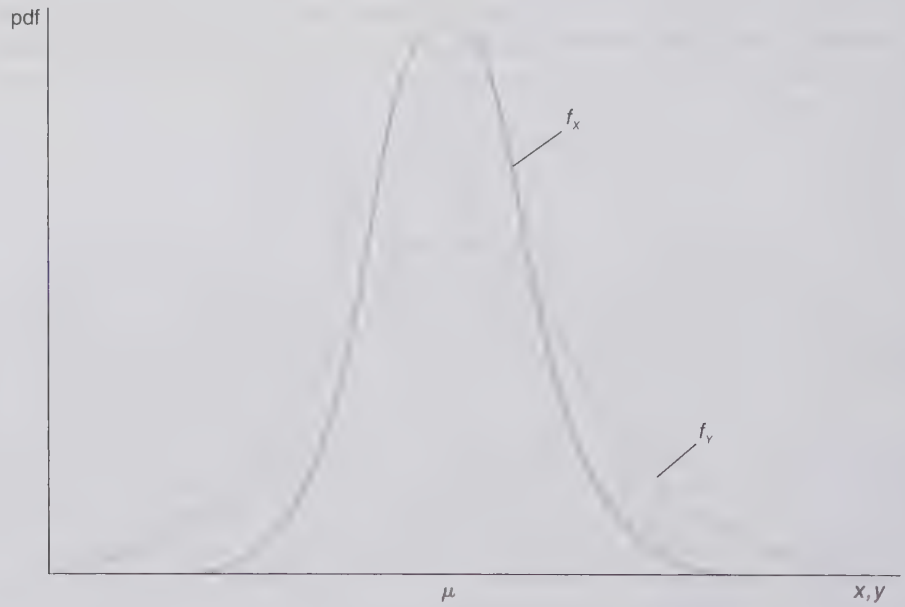


Figure B.4 *Variables aléatoires dont la moyenne est identique mais dont la distribution diffère.*

La variance est parfois dénotée σ_X^2 ou simplement σ^2, lorsque le contexte est clair. À partir de (B.23), il s'ensuit que la variance est toujours non négative. Comme instrument de calcul, il est utile d'observer que :

$$\sigma^2 = E(X^2 - 2X\mu + \mu^2) = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2. \qquad \textbf{[B.24]}$$

En utilisant soit (B.23) ou (B.24), nous ne devons pas distinguer entre des variables aléatoires discrètes et continues : la définition de la variance est la même dans les deux cas. La plupart du temps, nous calculons tout d'abord $E(X)$, et ensuite $E(X^2)$, et puis enfin la formule (B.24). Par exemple si $X \sim Bernoulli(\theta)$, alors $E(X) = \theta$, et, puisque $X^2 = X$, $E(X^2) = \theta$. À partir de l'équation (B.24), il suit que $Var(X) = E(X^2) - \theta^2 = \theta - \theta^2 = \theta(1 - \theta)$.

Deux propriétés importantes de la variance suivent.

Propriété VAR.1 : $Var(X) = 0$ si et seulement si, il y a une constante c telle que $P(X = c) = 1$, auquel cas $E(X) = c$.

Cette première propriété dit que la variance de toute constante est zéro et si une variable aléatoire a une variance égale à zéro, alors elle est constante.

Propriété VAR.2 : Pour toutes constantes a et b, $Var(aX + b) = a^2Var(X)$.

Cela signifie qu'ajouter une constante à une variable aléatoire ne change pas la variance, mais multiplier une variable aléatoire par une constante accroît la variance par un facteur égal au *carré* de cette constante. Par exemple, si X dénote la température en degrés Celsius et $Y = 32 + (9/5)X$ est la température en degrés Fahrenheit, alors $Var(Y) = (9/5)^2Var(X) = (81/25)Var(X)$.

Écart-type

L'**écart-type** d'une variable aléatoire, dénoté $\sigma(X)$, est simplement la racine carrée positive de la variance : $\sigma(X) \equiv +\sqrt{Var(X)}$. L'écart-type est parfois dénoté σ_X ou simplement σ, lorsque la variable aléatoire est bien comprise. Deux propriétés relatives à l'écart-type suivent immédiatement à partir des propriétés VAR.1 et VAR.2.

Propriété SD.1 : Pour toute constante c, $\sigma(c) = 0$.

Propriété SD.2 : Pour toutes constantes a et b,

$$\sigma(aX + b) = |a|\sigma(X).$$

En particulier, si $a > 0$, alors $\sigma(aX) = a \cdot \sigma(X)$.

Cette dernière propriété rend l'écart-type plus naturel à manipuler que la variance. Par exemple, supposons que X soit une variable aléatoire mesurée en milliers de dollars, disons, le revenu. Si nous définissons $Y = 1\,000X$, alors Y est le revenu mesuré en dollars. Supposons que $E(X) = 20$, et $\sigma(X) = 6$. Alors, $E(Y) = 1\,000E(X) = 20\,000$, et $\sigma(Y) = 1\,000$, $\sigma(X) = 6\,000$, si bien que la valeur espérée et l'écart-type augmentent tous les deux du même facteur, 1000. Si nous avions travaillé avec la variance, nous aurions obtenu, $Var(Y) = (1{,}000)^2Var(X)$, si bien que la variance de Y est 1 million de fois plus élevée que la variance de X.

Standardiser une variable aléatoire

Comme application des propriétés de la variance et de l'écart-type – et c'est une matière d'intérêt pratique en soi –, supposons qu'étant donné une variable aléatoire X, nous définissions une nouvelle variable aléatoire en soustrayant sa moyenne μ et en divisant par son écart-type σ :

$$Z \equiv \frac{X - \mu}{\sigma}, \qquad \textbf{[B.25]}$$

que nous pouvons écrire comme $Z = aX + b$, où $a \equiv (1/\sigma)$ et $b \equiv -(\mu/\sigma)$. Dans ce cas, à partir de la propriété E.2,

$$E(Z) = aE(X) + b = (\mu/\sigma) - (\mu/\sigma) = 0.$$

À partir de la propriété VAR.2,

$$Var(Z) = a^2Var(X) = (\sigma^2/\sigma^2) = 1.$$

Dès lors, la variable aléatoire Z a une moyenne de zéro et une variance (et dès lors un écart-type) égaux à un. Cette procédure est parfois appelée *standardisation* de la variable aléatoire X, et Z est appelée une variable aléatoire standardisée. (Dans les cours d'introduction aux statistiques, cette méthode est parfois appelée *transformation-z* de X.) Il est important de se rappeler que l'écart-type, et non pas la variance, apparaît au dénominateur de (B.25). Comme nous le verrons cette transformation est fréquemment utilisée en inférence statistique.

Comme cas particulier, supposons que $E(X) = 2$, et que $Var(X) = 9$. Alors, $Z = (X - 2)/3$ a une valeur attendue de zéro et une variance de un.

Coefficients d'asymétrie et d'aplatissement

Nous pouvons utiliser la version standardisée d'une variable aléatoire pour définir d'autres caractéristiques de la distribution d'une variable aléatoire. Ces caractéristiques sont décrites en utilisant ce qu'on appelle les *moments d'ordre supérieur*. Par exemple, le troisième moment de la variable aléatoire Z de (B.25) est utilisé pour déterminer si une distribution est symétrique autour de sa moyenne. On peut écrire

$$E(Z^3) = E[(X - \mu)^3]/\sigma^3$$

Si X a une distribution symétrique autour de μ, alors Z a une distribution symétrique autour de zéro. (La division par σ^3 ne change pas, que la distribution soit symétrique ou non.) Cela signifie que la densité de Z en deux points quelconques z et $-z$ est la même, ce qui signifie que, en calculant $E(Z^3)$, les valeurs positives z^3 quand $z > 0$ sont exactement compensées par les valeurs négatives $(-z)^3 = -z^3$. Il s'ensuit que, si X est symétrique autour de zéro, alors $E(Z) = 0$. Généralement, $E[(X - \mu)^3]/\sigma^3$ est vu comme une mesure de symétrie dans la distribution de X. Dans un problème statistique, nous pouvons utiliser des données pour estimer $E(Z^3)$ pour déterminer si la distribution de population sous-jacente apparaît comme symétrique. (L'exercice sur ordinateur C5.4 dans le chapitre 5 en fournit une illustration.)

Il peut être informatif de calculer le quatrième moment de Z,

$$E(Z^4) = E[(X - \mu)^4]/\sigma^4.$$

Parce que $Z^4 \geq 0$, $E(Z^4) \geq 0$ (et, dans tout cas intéressant, strictement supérieur à zéro). Sans avoir une valeur de référence, il est difficile d'interpréter les valeurs de $E(Z^4)$, mais des valeurs plus élevées signifient que les queues de la distribution de X sont plus épaisses. Le quatrième moment $E(Z^4)$ est appelé le **kurtosis** (ou coefficient d'aplatissement) de la distribution de X. Dans la section B.5., nous obtiendrons $E(Z^4)$ pour la distribution normale.

B.4 CARACTÉRISTIQUES DES DISTRIBUTIONS JOINTES ET CONDITIONNELLES

Mesures d'association : covariance et corrélation

Alors que la FDP jointe de deux variables aléatoires décrit complètement la relation entre elles, il est utile d'avoir des mesures de synthèse de la manière dont, en moyenne, deux variables aléatoires varient l'une par rapport à l'autre. Comme pour la valeur attendue et la variance, ce nombre unique permet de résumer un aspect d'une distribution complète, ce qui est dans ce cas une distribution jointe de deux variables aléatoires.

Covariance

Soit $\mu_X = E(X)$ et $\mu_Y = E(Y)$; considérons la variable aléatoire $(X - \mu_X)(Y - \mu_Y)$. Maintenant, si X est au-dessus de sa moyenne et Y est au-dessus de sa moyenne, alors $(X - \mu_X)(Y - \mu_Y) > 0$. Ceci est vrai si $X < \mu_X$ et $Y < \mu_Y$. D'un autre côté, si $X > \mu_X$ et $Y < \mu_Y$, et vice-versa, alors $(X - \mu_X)(Y - \mu_Y) < 0$. Comment, dans ce cas, ce produit peut-il nous apprendre quelque chose à propos de la relation entre X et Y ?

La covariance entre deux variables aléatoires X et Y, parfois appelée la *covariance de la population* pour insister sur le fait que cela concerne la relation entre deux variables décrivant une population, est définie comme la valeur attendue du produit $(X - \mu_X)(Y - \mu_Y)$:

$$Cov(X,Y) \equiv E[(X - \mu_X)(Y - \mu_Y)], \qquad \textbf{[B.26]}$$

qui est parfois dénoté σ_{XY}. Si $\sigma_{XY} > 0$, alors, en moyenne, quand X est au-dessus de sa moyenne, Y est aussi au-dessus de sa moyenne. Si $\sigma_{XY} < 0$, alors, en moyenne, quand X est au-dessus de sa moyenne, Y est quant à lui en-dessous de la sienne.

Plusieurs expressions utiles pour calculer $Cov(X,Y)$ sont les suivantes :

$$\begin{aligned} Cov(X,Y) &= E[(X - \mu_X)(Y - \mu_Y)] = E[(X - \mu_X)Y] \\ &= E[X(Y - \mu_Y)] = E(XY) - \mu_X\mu_Y. \end{aligned} \quad \textbf{[B.27]}$$

Il s'ensuit à partir de (B.27), que si $E(X) = 0$ ou $E(Y) = 0$, alors $Cov(X,Y) = E(XY)$.

La covariance mesure le degré de dépendance linéaire entre deux variables aléatoires. Une covariance positive indique des variables aléatoires évoluant dans la même direction, tandis qu'une covariance négative indique qu'elles évoluent dans des directions opposées. Interpréter l'ampleur de la covariance peut être quelque peu délicat, comme nous le verrons bientôt.

Parce que la covariance est une mesure de la manière dont deux variables aléatoires sont reliées, il est naturel de se demander comment la covariance est reliée à la notion d'indépendance. Ceci est donné par la propriété suivante.

Propriété COV.1 : Si X et Y sont indépendants, alors $Cov(X,Y) = 0$.

Cette propriété découle de l'équation (B.27) et le fait que $E(XY) = E(X)E(Y)$ quand X et Y sont indépendants. Il est important de se rappeler que l'inverse de COV.1 n'est *pas* vrai : une covariance égale à zéro entre X et Y n'implique pas que X et Y sont indépendants. En fait il y a des variables aléatoires X telles que, si $Y = X^2$, $Cov(X,Y) = 0$. [N'importe quelle variable aléatoire avec $E(X) = 0$ et $E(X^3) = 0$ a cette propriété.] Si $Y = X^2$, alors X et Y sont clairement non indépendantes : une fois que nous connaissons X, nous connaissons Y. Il semble plutôt étrange que X et X^2 puissent avoir une covariance égale à zéro, et cela révèle une faiblesse de la covariance comme mesure générale de l'association entre des variables aléatoires. La covariance est utile dans des contextes où les relations sont au minimum approximativement linéaires.

La seconde propriété majeure de la covariance concerne les covariances entre les fonctions linéaires.

Propriété COV.2 : Pour toutes les constantes a_1, b_1, a_2, et b_2,

$$Cov(a_1X + b_1, a_2Y + b_2) = a_1a_2Cov(X,Y). \quad \textbf{[B.28]}$$

Une application importante de COV.2 est que la covariance entre deux variables aléatoires peut être simplement altérée en multipliant une ou les deux variables aléatoires par une constante. Ceci est important en économie car les variables monétaires comme les taux d'inflation peuvent être définies dans différentes unités de mesure sans que l'on change leur interprétation.

Finalement, il est utile de savoir que la valeur absolue de la covariance entre deux variables aléatoires quelconques est bornée par le produit de leur écart-type ; ceci est connu comme l'*inégalité de Cauchy-Schwartz*.

Propriété COV.3 : $|Cov(X,Y)| \leq \sigma(X)\sigma(Y)$.

Coefficient de corrélation

Supposons que nous voulions connaître la relation entre le degré d'éducation et les revenus annuels dans la population active. Nous notons par X l'éducation et par Y le revenu et nous calculons leur covariance. La réponse que nous obtenons dépendra de la manière dont on mesure l'éducation et les revenus. La propriété COV.2 implique que la covariance entre l'éducation et le revenu dépend du fait que les revenus sont mesurés en dollars ou en milliers de dollars, ou du fait que l'éducation est mesurée en mois ou en années. Il est très clair que la manière dont nous mesurons ces variables n'a pas d'implication sur la force avec laquelle les deux variables sont reliées. Mais la covariance entre elles dépend des unités de mesure.

Le fait que la covariance dépend des unités de mesure est une faiblesse qui est surmontée par le **coefficient de corrélation** entre X et Y :

$$Corr(X,Y) \equiv \frac{Cov(X,Y)}{\sigma(X)\cdot\sigma(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}; \qquad \textbf{[B.29]}$$

le coefficient de corrélation entre X et Y est parfois dénoté ρ_{xy} (et est parfois appelé la *corrélation de la population*).

Parce que σ_X et σ_Y sont positifs, $Cov(X,Y)$ et $Corr(X,Y)$ ont toujours le même signe, et $Corr(X,Y) = 0$ si et seulement si, $Cov(X,Y) = 0$. Certaines des propriétés de la covariance s'appliquent à la corrélation. Si X et Y sont indépendants, alors $Corr(X,Y) = 0$, mais une corrélation de zéro n'implique pas l'indépendance des variables. (Comme la covariance, le coefficient de corrélation est aussi une mesure d'indépendance linéaire.) Cependant, la taille du coefficient de corrélation est plus facile à interpréter que la taille de la covariance grâce à la propriété suivante.

Propriété CORR.1 : $-1 \leq Corr(X,Y) \leq 1$.

Si $Corr(X,Y) = 0$, ou de manière équivalente $Cov(X,Y) = 0$, alors il n'y a pas de relation linéaire entre X et Y, et X et Y sont dits des **variables aléatoires non corrélées** ; sinon X et Y sont *corrélés*. $Corr(X,Y) = 1$ implique une relation linéaire positive parfaite, ce qui signifie que pouvons écrire $Y = a + bX$ pour une certaine constante a et une certaine constante $b > 0$. $Corr(X,Y) = -1$ implique une relation linéaire négative parfaite, si bien que $Y = a + bX$ pour un certain $b < 0$. Les cas extrêmes de 1 positifs ou négatifs arrivent rarement. Les valeurs de ρ_{xy} plus proches de 1 ou de -1 indiquent des relations linéaires plus fortes.

Comme mentionné précédemment, la corrélation entre X et Y est invariante aux unités de mesure, soit de X ou de Y. Ceci est établi de manière générale comme suit.

Propriété CORR.2 : Pour des constantes a_1, b_1, a_2, et b_2, avec $a_1 a_2 > 0$,

$$Corr(a_1X + b_1, a_2Y + b_2) = Corr(X,Y).$$

Si $a_1 a_2 < 0$, alors

$$Corr(a_1X + b_1, a_2Y + b_2) = -Corr(X,Y).$$

Comme exemple, supposons que la corrélation entre les revenus et l'éducation dans la population active est 0,15. Cette mesure ne dépend pas du fait que les revenus sont mesurés en dollars, milliers de dollars, ou dans toute autre unité ; cela ne dépend pas non plus du fait que l'éducation est mesurée en années, trimestres, mois, et ainsi de suite.

Variance d'une somme de variables aléatoires

Maintenant que nous avons défini la covariance et la corrélation, nous pouvons terminer notre liste des propriétés majeures de la variance.

Propriété VAR.3 : Pour les constantes a et b,

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y).$$

Il s'en suit immédiatement que, si X et Y sont décorrélés – si bien que $Cov(X,Y) = 0$ – alors

$$Var(X + Y) = Var(X) + Var(Y) \qquad \textbf{[B.30]}$$

et

$$Var(X - Y) = Var(X) + Var(Y). \qquad \textbf{[B.31]}$$

Dans le dernier cas, notons dans quelle mesure la variance de la différence est la *somme des variances*, et non pas la différence des variances.

Comme illustration de (B.30), soit X les profits gagnés par un restaurant durant un vendredi soir et soit Y les profits gagnés le samedi soir suivant. Alors, $Z = X + Y$ représente les profits des deux soirs. Supposons que X et Y ont tous les deux une valeur espérée de 300 USD et un écart-type de 15 USD (si bien que la variance est 225). Les profits espérés pour les deux soirs sont $E(Z) = E(X) + E(Y) = 2 \cdot (300) = 600$ USD. Si X et Y sont indépendants, et de ce fait non corrélés, alors la variance des profits totaux est la somme des variances : $Var(Z) = Var(X) + Var(Y) = 2 \cdot (225) = 450$. Il s'ensuit que l'écart-type des profits totaux est $\sqrt{450}$ ou à peu près 21,21 USD.

Les expressions (B.30) et (B.31) se généralisent à plus de deux variables aléatoires. Pour établir cette extension, nous avons besoin d'une définition. Les variables aléatoires $\{X_1, \ldots, X_n\}$ sont des **variables aléatoires non corrélées deux à deux** si chaque variable dans l'ensemble est non corrélée avec chaque autre variable dans l'ensemble.

À savoir, $Cov(X_i, X_j) = 0$, pour tout $i \neq j$.

Propriété VAR.4 : Si $\{X_1, \ldots, X_n\}$ sont des variables aléatoires non corrélées deux à deux et si $\{a_i : i = 1, \ldots, n\}$ sont des constantes, alors

$$Var(a_1X_1 + \ldots + a_nX_n) = a_1^2\, Var(X_1) + \ldots + a_n^2\, Var(X_n).$$

En notation avec l'opérateur de somme, nous pouvons écrire

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2\, Var(X_i). \qquad \textbf{[B.32]}$$

Un cas spécial de la propriété VAR.4 intervient lorsque nous prenons $a_i = 1$ pour tous les i. Alors pour les variables aléatoires non corrélées deux à deux, la variance de la somme est la somme des variances :

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i). \qquad \textbf{[B.33]}$$

Parce que des variables aléatoires indépendantes sont non corrélées (voir propriété COV.1), la variance d'une somme de variables aléatoires indépendantes est la somme des variances.

Si les X_i ne sont pas non corrélés deux à deux, alors l'expression de $Var\left(\sum_{i=1}^{n} a_i X_i\right)$ est beaucoup plus compliquée ; nous devons ajouter au terme de droite de (B.32) les termes $2a_i a_j Cov(x_i, x_j)$ pour tous les $i > j$.

Nous pouvons utiliser (B.33) pour dériver la variance d'une variable aléatoire binomiale. Soit $X \sim$ Binomiale (n, θ) et écrivons $X = Y_1 + \ldots + Y_n$, où les Y_i sont des variables aléatoires de Bernoulli(θ) indépendantes. Dans ce cas, via (B.33), $Var(X) = Var(Y_1) + \ldots + Var(Y_n) = n\theta(1 - \theta)$. Dans l'exemple de réservation de la compagnie aérienne avec $n = 120$ et $\theta = 0{,}85$, la variance du nombre de passagers arrivant pour la réservation est $n = 120$ et $\theta = 0{,}85$, si bien que $120(0{,}85)(0{,}15) = 15{,}3$ et que l'écart-type est d'à peu près 3,9.

Espérance conditionnelle

La covariance et la corrélation mesurent la relation linéaire entre deux variables aléatoires et les traitent de manière symétrique. Souvent en sciences sociales, nous voudrions expliquer une variable, appelée Y, en terme d'une autre variable disons, X. Ensuite, si Y est relié à X d'une manière non linéaire, nous voudrions le savoir.

Appelons Y la variable expliquée et X la variable explicative. Par exemple, Y pourrait être le salaire horaire, et X pourrait être le nombre d'années d'éducation formelle.

Nous avons déjà introduit la notion de la fonction de densité de probabilité conditionnelle de Y étant donné X. Dès lors, nous voudrions voir comment la distribution du salaire varie avec le niveau d'éducation. Cependant, nous voulons comme d'habitude avoir une manière simple de résumer cette distribution. Un simple nombre ne suffira plus, car la distribution de Y étant donné $X = x$ dépend généralement de la valeur de x. Cependant nous pouvons résumer la relation entre Y et X en regardant l'**espérance conditionnelle** de Y étant donné X, appelée parfois la *moyenne conditionnelle*. L'idée est la suivante. Supposons que nous sachions que X a pris une valeur particulière, disons x. Alors, nous pouvons calculer la valeur attendue de Y, étant donné ce que nous savons à propos du résultat de X. Nous notons cette valeur espérée par $E(Y|X = x)$, ou parfois $E(Y|X)$ par raccourci. Généralement, lorsque x change, $E(Y|x)$ change aussi.

Lorsque le Y est une variable aléatoire discrète prenant les valeurs $\{y_1, \ldots, y_m\}$, alors

$$E(Y|x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x).$$

Quand Y est continue, $E(Y|x)$ est défini en intégrant $yf_{Y|X}(y|x)$ sur toutes les valeurs possibles de Y. Comme dans le cas des espérances non conditionnelles, l'espérance conditionnelle est une moyenne pondérée de toutes les valeurs possibles de Y, mais désormais, les poids reflètent le fait que X a pris une certaine valeur. Dès lors, $E(Y|x)$ est simplement une certaine fonction de x, qui nous dit comment la valeur espérée de Y varie avec x.

Comme exemple, soit (X,Y) représentant la population de tous les individus qui travaillent, où X est le nombre d'années d'éducation et Y le salaire horaire. Dans ce cas, $E(Y|X=12)$ est le salaire horaire moyen pour toutes les personnes dans la population avec 12 années d'éducation (soit à peu près l'éducation à l'issue du secondaire). $E(Y|X=16)$ est le salaire horaire moyen pour toutes les personnes avec 16 années d'éducation. Tracer la valeur espérée pour les différents niveaux d'éducation fournit une information importante sur la manière dont les salaires et l'éducation sont liés. Voir la figure B.5 pour une illustration.

En principe, la valeur espérée du salaire horaire peut être trouvée pour chaque niveau d'éducation, et ces espérances peuvent être résumées dans un tableau. Parce que l'éducation peut varier fortement – elle peut même être mesurée en fraction d'années – ceci est une manière compliquée de montrer la relation entre le salaire moyen et la quantité d'éducation. En économétrie, nous spécifions habituellement des fonctions simples qui capturent cette relation. Comme exemple, supposons que la valeur espérée de *WAGE* étant donné *EDUC* est la fonction linéaire

$$E(WAGE|EDUC) = 1{,}05 + 0{,}45\ EDUC.$$

Si cette relation est valable dans la population des personnes qui travaillent, le salaire moyen pour les personnes avec 8 années d'éducation est $1{,}05 + 0{,}45(8) = 4{,}65$, et 4,65 USD. Le salaire moyen pour les personnes avec 16 années d'éducation est de 8,25, et 8,25 USD. Le coefficient de *EDUC* implique que chaque année d'éducation augmente le salaire horaire espéré de 0,45, ou 45 cents (de dollar).

Les espérances conditionnelles peuvent aussi être des fonctions non linéaires. Par exemple, supposons que X est une variable aléatoire qui est toujours plus grande que zéro. Cette fonction est représentée à la figure B.6. Ceci pourrait représenter une fonction de demande, où Y est la quantité demandée et X est le prix. Si Y et X sont reliés de cette manière, une analyse d'association linéaire, telle qu'une analyse de corrélation, serait incomplète.

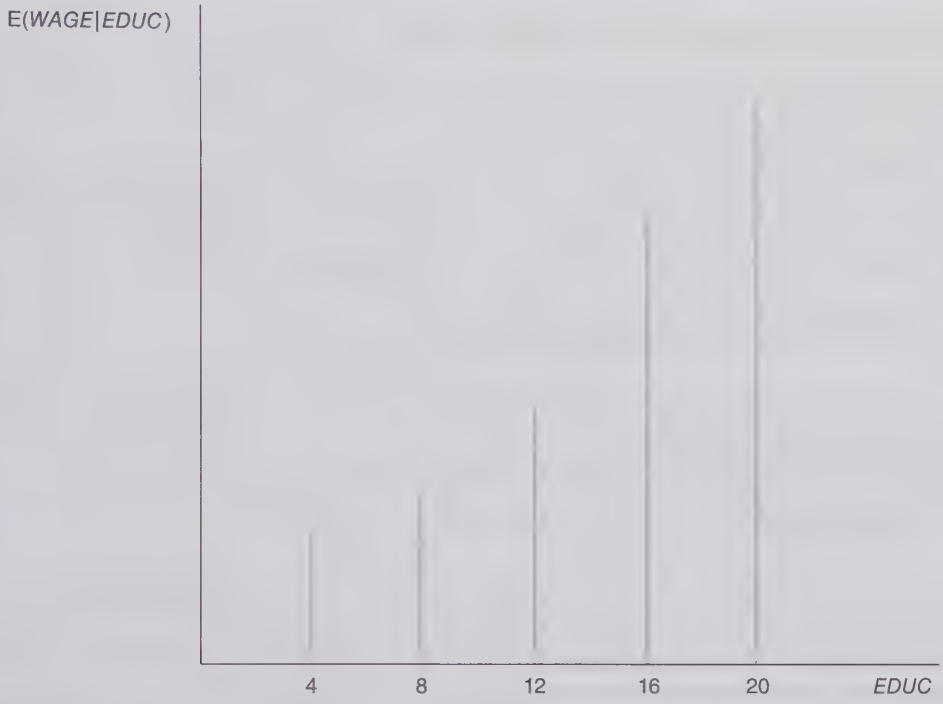


Figure B.5 *L'espérance du salaire horaire étant donné différents niveaux d'instruction.*

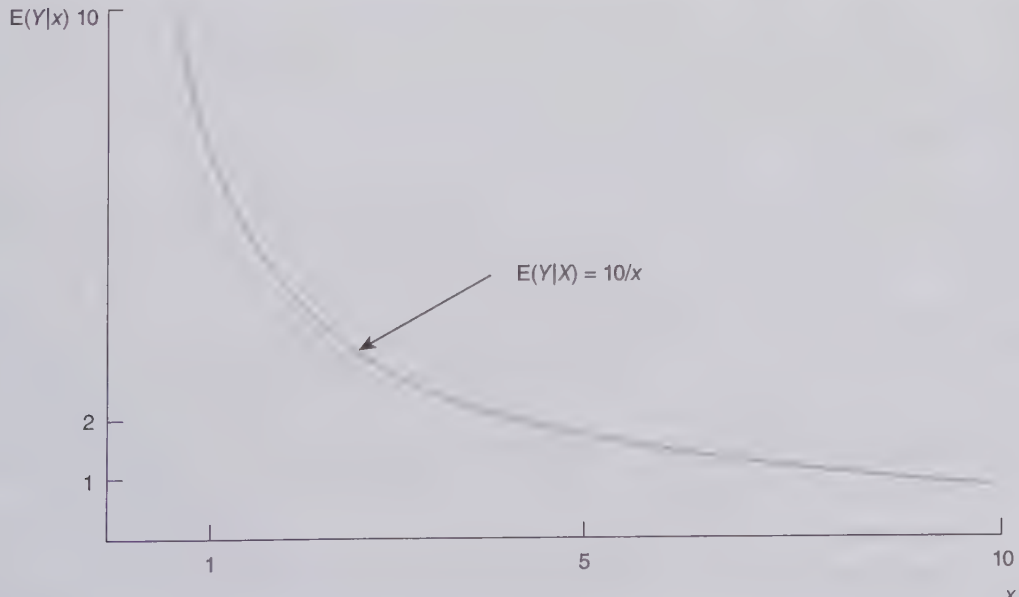


Figure B.6 *Graphique de E(Y|X) = 10/x*

Propriétés de l'espérance conditionnelle

Une série de propriétés de base des espérances conditionnelles sont utiles pour certaines dérivations utiles en économétrie.

Propriété CE.1 : $E[c(X)|X] = c(X)$ pour toute fonction $c(X)$.

Cette première propriété signifie que les fonctions de X se comportent comme des constantes quand nous calculons les espérances conditionnelles par rapport à X. Par exemple, $E(X^2|X) = X^2$. Intuitivement, cela signifie simplement que si nous connaissons X, alors nous connaissons aussi X^2.

Propriété CE.2 : Pour les fonctions $a(X)$ et $b(X)$,

$$E[a(X)Y + b(X)|X] = a(X)E(Y|X) + b(X).$$

Par exemple, nous pouvons calculer facilement l'espérance conditionnelle d'une fonction telle que $XY + 2X^2$: $E(XY + 2X^2|X) = XE(Y|X) + 2X^2$.

La propriété suivante lie ensemble les notions d'indépendance et d'espérances conditionnelles.

Propriété CE.3 : si X et Y sont indépendants, alors, $E(Y|X) = E(Y)$.

Cette propriété signifie que, si X et Y sont indépendants, alors la valeur espérée de Y étant donné X ne dépend pas de X, auquel cas, $E(Y|X)$ est toujours égal à la valeur espérée (non conditionnelle) de Y. Dans l'exemple des salaires et de l'éducation, si les salaires étaient indépendants de l'éducation, alors les salaires moyens de l'école secondaire et des diplômés de l'université seraient le même. Comme ceci est presque certainement faux, nous ne pouvons pas supposer que le salaire et l'éducation sont indépendants.

Un cas particulier de la propriété est le suivant : si U et X sont indépendants et $E(U) = 0$, alors, $E(U|X) = 0$.

Il existe aussi des propriétés de l'espérance conditionnelle qui ont trait au fait que $E(Y|X)$ est une fonction de X, disons, $E(Y|X) = \mu(X)$. Parce que X est une variable aléatoire, $\mu(X)$ est aussi une variable aléatoire. En outre, $\mu(X)$ a une distribution de probabilité et dès lors une valeur espérée. Généralement, la valeur espérée de $\mu(X)$ peut être très difficile à calculer directement. La **loi des espérances itérées** établit que la valeur espérée de $\mu(X)$ est simplement égale à la valeur espérée de Y. Nous écrivons ceci ainsi.

Propriété CE.4 : $E[E(Y|X)] = E(Y)$.

Cette propriété est quelque peu difficile à appréhender au début. Elle signifie que, si nous obtenons tout d'abord $E(Y|X)$ comme une fonction de X et que nous en prenons la valeur espérée (par rapport à la distribution de X, bien entendu), alors nous obtenons $E(Y)$. Ceci n'est pas évident au premier abord, mais cela peut être dérivé en utilisant la définition des valeurs espérées.

Comme exemple de la manière d'utiliser la propriété, CE.4, supposons $Y = WAGE$ et $X = EDUC$, où $WAGE$ est mesurée en heures et $EDUC$ mesurée en années. Supposons que la valeur espérée de $WAGE$ étant donné $EDUC$ est $E(WAGE|EDUC) = 4 + 0{,}60\ EDUC$. En outre, $E(EDUC) = 11{,}5$. Dans ce cas, la loi des espérances itérées implique que $E(WAGE) = E(4 + 0{,}60\ EDUC) = 4 + 0{,}60\ E(EDUC) = 4 + 0{,}60(11{,}5) = 10{,}90$, ou 10,90 USD par heure.

La propriété suivante établit une version plus générale de la loi des espérances itérées.

Propriété CE.4' : $E(Y|X) = E[E(Y|X,Z)|X]$.

En d'autres termes, on peut trouver $E(Y|X)$ en deux étapes. Premièrement, trouver pour toute autre variable aléatoire Z. Ensuite, trouver la valeur de $E(Y|X,Z)$, conditionnellement à X.

Propriété CE.5 : Si $E(Y|X) = E(Y)$, alors $Cov(X,Y) = 0$ [et ainsi $Corr(X,Y) = 0$].

En fait, *chaque* fonction de X est non corrélée avec Y.

Cette propriété implique que, si la connaissance de X ne change pas la valeur espérée de Y, alors X et Y doivent être non corrélés, ce qui implique que si X et Y sont corrélées, alors $E(Y|X)$ doit dépendre de X. L'inverse de la propriété n'est pas vrai : si X et Y ne sont pas corrélées, $E(Y|X)$ pourrait toujours dépendre de X. Par exemple, supposons $Y = X^2$. Alors, $E(Y|X) = X^2$, qui est clairement une fonction de X. Cependant, comme nous l'avons mentionné dans notre discussion sur la covariance la corrélation, il est possible que X et X^2 soient non corrélées. L'espérance conditionnelle capture la relation non linéaire entre X et Y qu'une analyse de corrélation raterait complètement.

Les propriétés CE.4 et CE.5 ont deux implications importantes : si U et X sont des variables aléatoires telles que $E(U|X) = 0$, alors $E(U) = 0$, et U et X sont non corrélés.

Propriété CE.6 : Si $E(Y^2) < \infty$ et $E[g(X)^2] < \infty$ pour une certaine fonction g, alors $E\{[Y - \mu(X)]^2|X\} \leq E\{[Y - g(X)]^2|X\}$ et $E\{[Y - \mu(X)]^2\} \leq E\{[Y \mid g(X)]^2\}$.

La propriété CE.6 est très utile dans des contextes de prédiction ou de prévision. La première inégalité dit que, si nous mesurons la prédiction de manière imprécise comme l'erreur de prédiction au carré attendue, conditionnellement à X, alors la moyenne conditionnelle est meilleure que n'importe quelle autre fonction de X pour prédire Y. La moyenne conditionnelle minimise aussi l'erreur de prédiction au carré attendue non conditionnelle.

Variance conditionnelle

Soit deux variables aléatoires X et Y. La variance de Y, conditionnelle à $X = x$, est égale à la variance associée à la distribution conditionnelle de Y, étant donné que $X = x$. Sous forme mathématique, $Var(Y|X = x) = E\{[Y - E(Y|x)]^2|x\}$. Néanmoins, la formule

$$Var(Y|X = x) = E(Y^2|x) - [E(Y|x)]^2$$

est plus fréquemment utilisée lorsqu'il s'agit de calculer la variance conditionnelle. Même si nous n'aurons pas à la calculer très souvent, il nous faudra poser des hypothèses à son sujet. Nous serons également amenés à adapter cette formule pour traiter de certaines problématiques liées à l'analyse de régression.

Par exemple, supposons que $Y = EPARGNE$ et $X = REVENU$, ces deux variables étant mesurées sur base annuelle pour la population de toutes les familles. Soit $Var(EPARGNE|REVENU) = 400 + 0{,}25\ REVENU$. Cette formule implique que la variance de l'épargne augmente lorsque le revenu s'élève. Notez bien que cette relation entre la *variance* de l'épargne par rapport au revenu n'a rien à voir avec celle qui caractérise la *valeur attendue* de l'épargne étant donné le revenu.

Nous présentons une propriété intéressante de la variance conditionnelle (VC).

Propriété VC.1 : Si X et Y sont indépendants, alors $Var(Y|X) = Var(Y)$.

Cette propriété est assez évidente puisque, dans ce cas, la distribution de Y étant donné X ne dépend pas de X et $Var(Y|X)$ n'est qu'une caractéristique de cette distribution.

B.5 LES DISTRIBUTIONS STATISTIQUES INCONTOURNABLES

La distribution normale

La distribution normale et celles qui en découlent sont les distributions les plus couramment utilisées en économétrie. L'hypothèse selon laquelle les variables aléatoires de la population sont distribuées selon une loi normale simplifie le calcul des probabilités. Nous allons également recourir très fréquemment à ces distributions pour mener des tests d'hypothèses ou construire des intervalles de confiance, même lorsque la population sous-jacente ne suit pas une loi normale. À ce stade de l'analyse, il est impossible d'approfondir la

discussion mais vous pourrez constater que ces distributions interviennent à de multiples reprises dans cet ouvrage.

Une variable aléatoire, dite « normale », est une variable aléatoire continue, qui peut prendre n'importe quelle valeur. Sa FDP est représentée par la célèbre « courbe en cloche », représentée à la figure B.7.

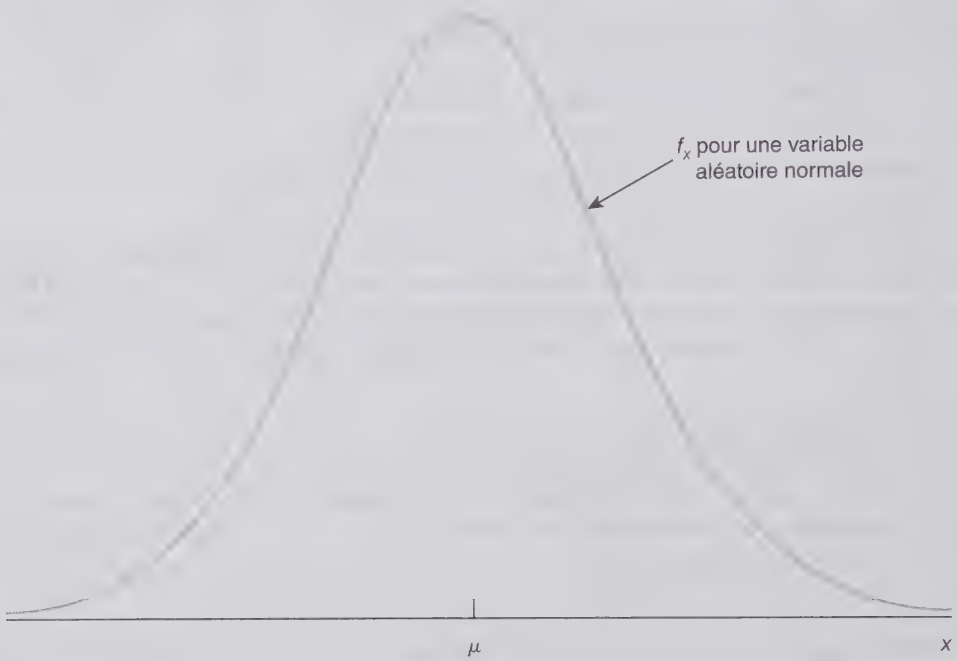


Figure B.7 *La forme générale de la fonction de densité de probabilité.*

Sur le plan mathématique, la FDP de X peut s'écrire sous la forme

$$f(x) = 1/(\sigma\sqrt{2\pi})\exp[-(x-\mu)^2/2\sigma^2], -\infty < x < \infty \quad \textbf{[B.34]}$$

où $\mu = E(X)$ et $\sigma^2 = Var(X)$. Dans ce cas de figure, nous disons que X suit une **distribution normale** dont l'espérance est μ et la variance σ^2, soit $X \sim N(\mu, \sigma^2)$. Vu que la distribution normale est symétrique autour de μ, μ correspond également à la médiane X. La distribution normale est parfois appelée *distribution Gaussienne*, en l'honneur des travaux réalisés par le célèbre mathématicien C. F. Gauss.

Certaines variables aléatoires suivent une distribution plus ou moins normale. Par exemple, le poids et la taille des êtres humains, les résultats aux examens, ou le taux de chômage régional ont des FDP dont la forme est proche de celle décrite à la figure B.7. D'autres distributions, comme la distribution des revenus, ne suivent pas la FDP de la loi normale, en règle générale. Dans la plupart des pays, le revenu n'est pas symétriquement distribué ; la distribution est asymétrique, affichant une extrémité droite plus longue que la gauche. [Le plus bas des revenus est nul, alors qu'il n'y a pas de plafond au revenu le plus élevé]. Dans certains cas, il est possible de transformer une variable aléatoire pour la rendre « normale ». Une transformation populaire est le log naturel (ou logarithme népérien), qui convient aux variables aléatoires positives. Si X est une variable aléatoire positive, comme le revenu, et que $Y = log(X)$ suit une distribution normale, alors X suit une *distribution log-normale*. Il s'avère que la distribution log-normale caractérise assez bien la distribution des revenus que l'on observe dans de nombreux pays. La distribution d'autres variables, comme les prix de biens et services, est également bien décrite par la distribution log-normale.

La distribution normale standard

Un cas particulier de la distribution normale est représenté par la distribution dont l'espérance est nulle et la variance (et l'écart-type) est l'unité. Lorsqu'une variable aléatoire Z suit une distribution $N(0,1)$, nous disons qu'elle suit une **distribution normale standard**. Z est une variable normale standard [ou encore une variable normale centrée et réduite]. La FDP d'une variable normale standard est notée $\phi(z)$. Si nous utilisons $\mu = 0$ et $\sigma^2 = 1$ dans (B.34), nous obtenons

$$\phi(z) = (1/\sqrt{2\pi})\exp(-z^2/2), -\infty < z < \infty. \quad \textbf{[B.35]}$$

La fonction de distribution cumulée (FDC) standard est notée $\Phi(z)$ et correspond à l'aire sous ϕ, à gauche de z (voir la figure B.8). Rappelez-vous que $\Phi(z) = P(Z \leq z)$; comme Z est continue, nous pouvons également écrire $\Phi(z) = P(Z < z)$.

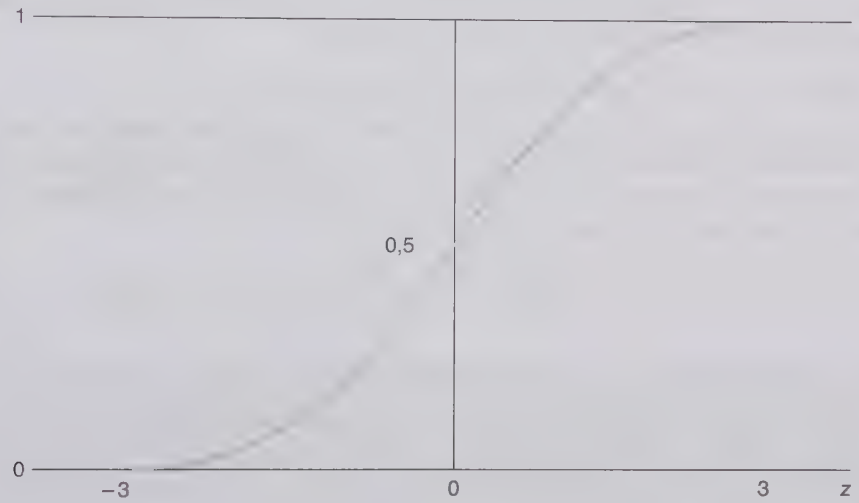


Figure B.8 *La fonction de distribution cumulée de la loi normale standard.*

Aucune formule simple ne peut être utilisée pour obtenir les valeurs de $\Phi(z)$. [En effet, $\Phi(z)$ correspond à une intégrale, sans solution analytique, de la fonction (B.35)]. Les valeurs de $\Phi(z)$. peuvent néanmoins être aisément tabulées, comme dans le tableau G.1 de l'annexe G, pour des valeurs de z comprises entre $-3,1$ et 3,1. Pour $z \leq -3,1$, $\Phi(z)$ est inférieure à 0,001 ; pour $z \geq -3,1$, $\Phi(z)$ est supérieure à 0,999. De nos jours, tous les logiciels disposent de lignes de commande qui permettent de calculer les valeurs d'une FDC standard ; il est donc souvent inutile d'imprimer une table complète pour obtenir les probabilités de z.

En recourant à quelques propriétés élémentaires de probabilités [en particulier, les probabilités (B.7) et (B.8)], nous sommes capables d'utiliser la FDC standard pour calculer la probabilité d'occurrence de n'importe quel événement relatif à une variable aléatoire normale standard. Les formules les plus importantes sont

$$P(Z > z) = 1 - \Phi(z), \quad \textbf{[B.36]}$$

$$P(Z > -z) = P(Z > z), \quad \textbf{[B.37]}$$

et

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a). \quad \textbf{[B.38]}$$

Étant donné que Z est une variable aléatoire continue, ces trois formules sont vérifiées, que les inégalités soient ou non strictes. Par exemple, $P(Z > 0{,}44) = 1 - 0{,}67 = 0{,}33$; $P(Z < -0{,}92) = P(Z > 0{,}92) = 1 - 0{,}821 = 0{,}179$; et $P(-1 < Z \leq 0{,}5) = 0{,}692 - 0{,}159 = 0{,}533$.

Pour tout $c > 0$, une autre expression utile est

$$\begin{aligned} P(|Z| > c) &= P(Z > c) + P(Z < -c) \\ &= 2 \cdot P(Z > c) = 2[1 - \Phi(c)]. \end{aligned} \qquad \textbf{[B.39]}$$

Par conséquent, la probabilité que la valeur absolue de Z soit plus élevée qu'une constante positive c, est égale au double de la probabilité $P(Z > c)$; ce résultat est une suite logique de la propriété de symétrie de la distribution normale standard.

La plupart du temps, nous commençons à travailler avec une variable aléatoire distribuée selon une loi normale, soit $X \sim N(\mu,\sigma^2)$, où $\mu \neq 0$ et $\sigma^2 \neq 1$. Toute variable X peut ensuite être transformée en une variable normale centrée réduite grâce à la propriété suivante.

Propriété 1 de *N*. Si $X \sim N(\mu,\sigma^2)$, alors $(X - \mu)/\sigma \sim N(0,1)$.

Cette propriété montre qu'il est facile de transformer n'importe quelle variable aléatoire normale en une variable normale centrée réduite [soit une variable centrée (par rapport à la moyenne) et réduite (par rapport à l'écart-type)]. Par exemple, si nous cherchons à calculer $P(X \leq 1)$ alors que $X \sim N(3,4)$, nous devons transformer X en une variable normale centrée réduite :

$$\begin{aligned} P(X \leq 1) &= P(X - 3 \leq 1 - 3) = P((X - 3)/2 \leq -1) \\ &= P(Z \leq -1) = \Phi(-1) = 0{,}159. \end{aligned}$$

EXEMPLE B.6
Calcul de probabilités pour une variable aléatoire normale

Calculons d'abord $P(2 < X \leq 6)$ lorsque $X \sim N(4,9)$. Pour rappel, le choix entre les signes $<$ ou $\leq$ n'a pas d'importance puisque X est une variable aléatoire continue. Donc,

$$\begin{aligned} P(2 < X \leq 6) &= P\left(\frac{2-4}{3} < \frac{X-4}{3} \leq \frac{6-4}{3}\right) = P(-2/3 < Z \leq 2/3) \\ &= \Phi(0{,}67) - \Phi(-0{,}67) = 0{,}749 - 0{,}251 = 0{,}498. \end{aligned}$$

Calculons maintenant $P(|X| > 2)$.

$$\begin{aligned} P(|X| > 2) &= P(X > 2) + P(X < -2) \\ &= P[(X - 4)/3 > (2 - 4)/3] + P[(X - 4)/3 < (-2 - 4)/3] \\ &= 1 - \Phi(-2/3) + \Phi(-2) \\ &= 1 - 0{,}251 + 0{,}023 = 0{,}772. \end{aligned}$$

Les autres propriétés de la distribution normale

Nous concluons cette section en identifiant les autres propriétés désirables de la distribution normale que nous exploiterons dans le reste de l'ouvrage.

Propriété 2 de *N*. Si $X \sim N(\mu, \sigma^2)$, alors $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

Par conséquent, si $X \sim N(1,9)$, alors $Y = 2X + 3$ est distribuée selon une loi normale de moyenne égale à $2E(X) + 3 = 5$ et de variance égale à $2^2 \cdot 9 = 36$. Enfin, $\sigma(Y) = 2\sigma(X) = 2 \cdot 3 = 6$.

Nous avons précédemment montré qu'absence de corrélation et indépendance n'étaient pas équivalents. Dans le cas particulier des variables aléatoires distribuées selon une loi normale, il s'avère qu'une corrélation nulle suffit à garantir l'indépendance.

Propriété 3 de *N*. Si X et Y sont toutes deux distribuées selon une loi normale, elles sont indépendantes si, et seulement si, $Cov(X,Y) = 0$.

Propriété 4 de *N*. Toute combinaison de variables aléatoires indépendantes et identiquement distribuées selon une loi normale, suit elle-même une distribution normale.

Prenons un exemple. Soit une variable aléatoire indépendante, X_i pour $i = 1, 2$, et 3, distribuée selon $N(\mu,\sigma^2)$. Si $W = X_1 + 2X_2 - 3X_3$, alors W est distribuée selon une loi normale ; il ne nous reste qu'à calculer sa moyenne et sa variance.

$$E(W) = E(X_1) + 2E(X_2) - 3E(X_3) = \mu + 2\mu - 3\mu = 0$$

et

$$Var(W) = Var(X_1) + 4Var(X_2) + 9Var(X_3) = 14\sigma^2.$$

La propriété 4 implique que la moyenne de variables aléatoires, indépendantes et normalement distribuées, suit elle-même une distribution normale. Si $Y_1, Y_2, \ldots, Y_n$ sont des variables aléatoires indépendantes et que chacune est distribuée selon une loi $N(\mu, \sigma^2)$, alors

$$\bar{Y} \sim N(\mu,\sigma^2/n). \qquad \textbf{[B.40]}$$

Ce résultat est essentiel lorsqu'il s'agit d'inférer les propriétés statistiques de la moyenne d'une population normalement distribuée.

D'autres propriétés de la distribution normale sont également intéressantes, même si elles ne jouent pas un rôle central dans cet ouvrage. Étant donné qu'une variable aléatoire normale est symétrique par rapport à la moyenne, son degré d'asymétrie est nul, soit $E[(X - \mu)^3] = 0$. Il est également possible de démontrer que le coefficient d'aplatissement (ou kurtosis) de la distribution normale standard est égale à 3, soit

$$E[(X - \mu)^4]/\sigma^4 = 3,$$

De manière équivalente, sachant que Z suit une distribution normale standard, $E(Z^4) = 3$. [Une distribution dont le coefficient d'aplatissement vaut 3 est dite mésokurtique]. Comme cette distribution est très fréquemment utilisée en probabilités et statistiques, le coefficient d'aplatissement de *n'importe quelle* variable aléatoire (dont le quatrième moment de la distribution existe, naturellement) est mesurée relativement à celui de la distribution normale standard. Si $E[(X - \mu)^4]/\sigma^4 > 3$, la distribution de X possède des extrémités plus épaisses que la distribution normale. [On parle de distribution leptokurtique.] C'est souvent le cas de la distribution de Student que nous allons introduire prochainement. Si $E[(X - \mu)^4]/\sigma^4 < 3$, la distribution de X dispose d'extrémités plus minces (ou moins épaisses) que la distribution normale. [On parle alors de distribution platikurtique].

La distribution du chi-deux

La distribution du chi-deux (ou chi deux, le tiret étant facultatif) est obtenue directement à partir de variables aléatoires normales centrées et réduites. Définissons Z_i, $i = 1, 2, \ldots, n$, comme des variables aléatoires indépendantes, chacune suivant une distribution normale standard. Soit une nouvelle variable aléatoire X correspondant à la somme des carrés des Z_i :

$$X = \sum_{i=1}^{n} Z_i^2. \qquad \textbf{[B.41]}$$

Dans ce cas, X suit une **distribution du chi deux** avec n **degrés de liberté** (ou *ddl*, pour faire bref). Nous écrivons $X \sim \chi^2_n$. Les *ddl* de la distribution du chi deux correspond au nombre de termes inclus dans la somme (B.41). Le concept de degrés de liberté joue un rôle important dans les analyses statistiques et économétriques.

La FDP de la distribution du chi deux, pour différents degrés de liberté, est représentée par la figure B.9. Comme la formule de cette FDP ne nous sera pas utile, nous ne la reproduisons pas dans le texte. L'équation (B.41) montre clairement qu'une variable aléatoire distribuée selon une chi deux sera toujours nulle ou positive ; contrairement à la distribution normale, la distribution du chi deux n'est pas symétrique, quel que soit le point de référence. Lorsque $X \sim \chi^2_n$, nous pouvons montrer que l'espérance de X est égale à n, soit le nombre de termes compris dans (B.41), et que sa variance est égale à $2n$.

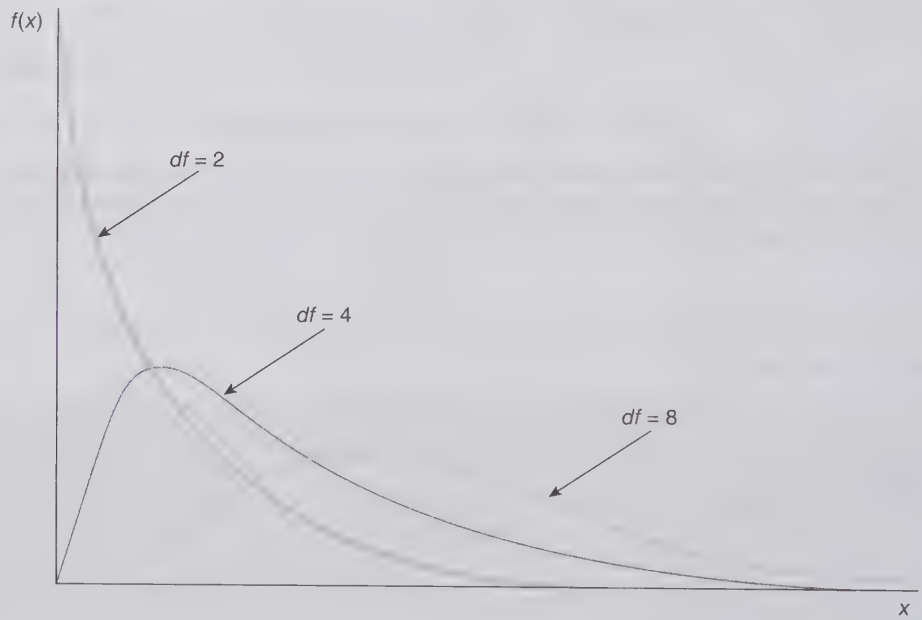


Figure B.9 *La distribution du chi deux pour différents degrés de liberté.*

La distribution t de Student

La distribution t de Student est l'outil de prédilection de l'analyse statistique et de la régression linéaire. Nous obtenons une distribution t à l'aide de deux variables. La première, Z, suit une distribution normale standard et la seconde, X, suit une distribution du chi deux à n degrés de liberté. Par ailleurs, Z et X sont des variables indépendantes. Dans ce cas, la variable aléatoire

$$T = Z/(\sqrt{X/n}) \qquad \textbf{[B.42]}$$

suit une **distribution** t à n degrés de liberté, soit $T \sim t_n$. Les degrés de liberté de la distribution t sont déterminés en fonction de ceux de la variable du dénominateur, X, distribuée selon la loi du chi deux.

La FDP de la distribution t ressemble à celle de la normale, mais la première est plus étirée et l'aire totale située dans ses extrémités est plus grande. [Les extrémités de la distribution sont souvent identifiées pour des valeurs absolues de t supérieures à 2 environ]. L'espérance d'une variable aléatoire distribuée selon une loi de Student est égale à zéro, à condition que $n > 1$, et sa variance est égale à $n/(n-2)$, pour tout $n > 2$.

(La variance tend vers l'infini lorsque $n \leq 2$, la distribution devenant trop étirée.) La FDP de la distribution t est représentée sur la figure B.10 pour différents degrés de liberté. Au fur et à mesure que les degrés de liberté augmentent, la distribution t se rapproche de la distribution normale standard.

La distribution F de Fisher-Snedecor

Une autre distribution très importante en statistiques et en économétrie est la distribution F. Elle est particulièrement utile pour réaliser des tests d'hypothèses dans le contexte des régressions linéaires multiples.

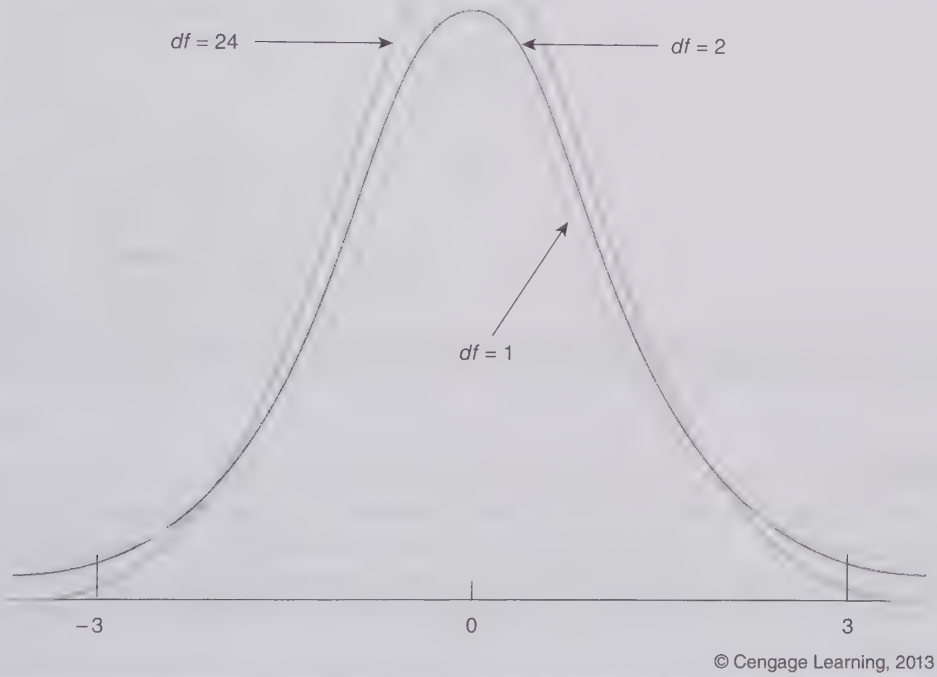


Figure B.10 *La distribution t de Student pour différents degrés de liberté.*

Pour définir une variable distribuée selon une loi de Fisher-Snedecor, nous avons besoin de deux variables indépendantes distribuées selon une loi du chi deux, soit $X_1 \sim \chi^2_{k_1}$ et $X_2 \sim \chi^2_{k_2}$. La variable aléatoire

$$F = (X_1/k_1)/(X_2/k_2) \qquad \textbf{[B.43]}$$

suit une **distribution *F*** à (k_1,k_2) degrés de liberté, soit $F \sim F_{k_1,k_2}$. La FDP d'une distribution F est reprise à la figure B.11, pour différents degrés de liberté.

L'ordre des degrés de liberté dans F_{k_1,k_2} revêt toute son importance. Le nombre entier k_1 représente les *degrés de liberté du numérateur*, car ce nombre est associé à la variable du chi deux qui se trouve au numérateur de (B.43). De manière équivalente, le nombre entier k_2 mesure les *degrés de liberté du dénominateur*, car ce nombre est associé à la variable du chi deux qui se trouve au dénominateur de (B.43). Cela mérite toute notre attention car (B.43) peut également s'écrire sous la forme $(X_1k_2)/(X_2k_1)$ et, dans ce cas, k_1 apparaît au dénominateur. Retenez simplement que les *ddl* du numérateur et du dénominateur sont respectivement donnés par les nombres entiers associés aux variables du numérateur et du dénominateur de (B.43).

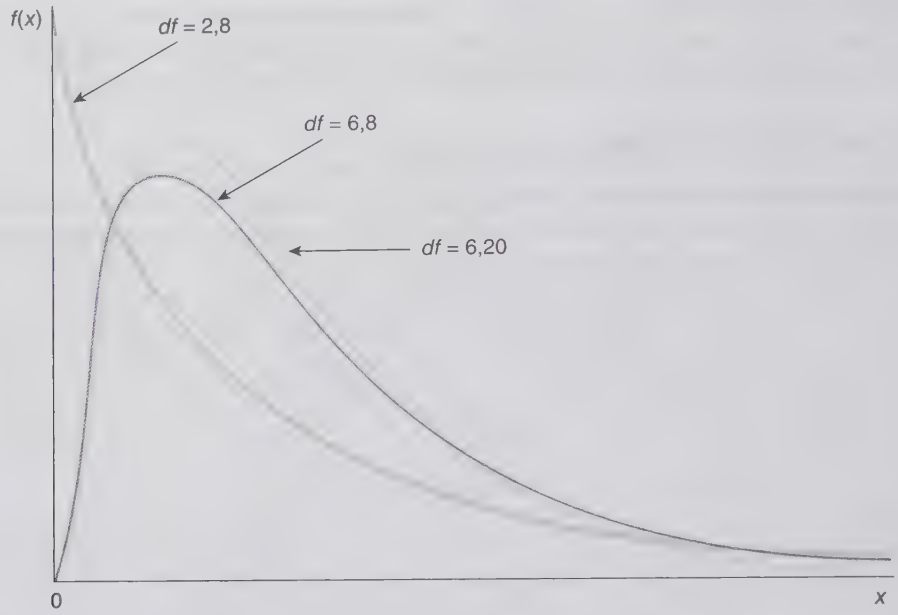


Figure B.11 *La distribution* F_{k_1,k_2} *pour différents degrés de liberté,* k_1 *et* k_2.

RÉSUMÉ

Dans cette annexe, nous avons passé en revue tous les éléments de probabilités dont nous avons besoin pour nous lancer dans l'étude de l'économétrie. La plupart de ces concepts sont d'ailleurs couverts dans les cours d'introduction aux probabilités et statistiques. Les thèmes plus avancés, tels que les propriétés des espérances conditionnelles, seront étudiés en temps voulu, notamment lorsque nous aborderons l'analyse de régression dans la partie 1.

Dans un cours d'introduction à la statistique, l'accent est mis sur le calcul des moyennes, variances, covariances, etc., propres à différentes distributions. Dans la partie 1 de cet ouvrage, nous n'aurons pas besoin de recourir à ces calculs ; par contre, nous utiliserons très souvent les propriétés que nous avons identifiées dans cette annexe et dont bénéficient les espérances et les variances.

MOTS-CLÉS

Asymétrie p. 838
Coefficient de corrélation p. 840
Coefficient d'aplatissement (ou kurtosis) p. 838
Covariance p. 838
Degrés de liberté p. 850
Distribution binomiale p. 830
Distribution du chi deux p. 850
Distribution conditionnelle p. 830
Distribution *F* p. 851

Distribution t p. 850
Distribution jointe p. 829
Distribution normale p. 846
Distribution normale standard p. 847
Distribution symétrique p. 835
Écart-type p. 837
Espérance conditionnelle p. 842
Expérience p. 824
Fonction de densité de probabilité (FDP) p. 826
Fonction de distribution cumulée (FDC) p. 827
Kurtosis (ou coefficient d'aplatissement) p. 838
Loi des espérances itérées p. 844
Valeur espérée p. 831
Variable aléatoire p. 824
Variable aléatoire continue p. 827
Variable aléatoire de Bernoulli p. 825
Variable aléatoire discrète p. 825
Variables aléatoires indépendantes p. 829
Variables aléatoires non corrélées p. 840
Variables aléatoires non corrélées deux à deux p. 841
Variables aléatoires standardisées p. 837
Variance p. 836

EXERCICES

1. Supposons qu'une étudiante américaine désire passer le « SAT », qui est un examen standardisé au niveau national. Expliquez la raison pour laquelle son résultat peut être considéré comme une variable aléatoire.

2. Soit une variable aléatoire X distribuée selon $N(5,4)$. Calculez les probabilités des événements suivants :

i. $P(X \leq 6)$.

ii. $P(X > 4)$.

iii. $P(|X - 5)| > 1)$.

3. La presse financière fait souvent grand cas de la capacité de certains gestionnaires de fonds communs de placement à battre le marché année après année. (Cela signifie que le rendement obtenu en investissant dans ces fonds gérés activement est plus élevé que celui obtenu en détenant un portefeuille investi dans un indice boursier de référence, comme le S&P 500). Pour être plus concret, considérez une période de 10 ans et tenez compte de la population des 4 170 fonds communs de placement que le *Wall Street Journal* identifiait au 1er janvier 1995. Lorsque nous disons que la performance d'un gestionnaire de fonds est aléatoire par rapport au marché, nous voulons signifier que chaque gestionnaire a 1 chance sur 2 de battre le marché chaque année, et que cette capacité à battre le marché est indépendante de la performance passée.

i. Si la performance du gestionnaire relativement à celle du marché est véritablement aléatoire, quelle est la probabilité qu'un fonds batte le marché chaque année pendant 10 ans ?

ii. Calculez la probabilité qu'*au moins* un de ces 4 170 fonds batte le marché chaque année pendant 10 ans. Quel enseignement en tirez-vous ?

iii. À l'aide d'un tableur ou d'un logiciel de statistique, calculez la probabilité binomiale qu'au moins 5 fonds battent le marché chaque année pendant 10 ans.

4. Soit X le taux d'emploi des adultes âgés de plus de 65 ans dans un pays sélectionné au hasard. Si le taux d'emploi est donné en décimales, la variable X est comprise entre 0 et 1. Supposons que la FDC de X soit $F(x) = 3x^2 - 2x^2$, pour $0 \leq x \leq 1$. Calculez la probabilité que le taux d'emploi des seniors soit au moins 0,6 (60 %).

5. Juste avant le début du procès pour meurtre du célèbre joueur de base-ball afro-américain, O. J. Simpson, en 1995, un sondage avait trouvé qu'environ 20 % de la population des adultes pensait qu'il était innocent. (L'existence de preuves ADN avait filtré dans la presse.) En guise d'illustration, ignorez le fait que ces 20 % ne représentent qu'une estimation obtenue à partir d'un sous-échantillon de la population. Supposons plutôt qu'il s'agisse de la vraie valeur du pourcentage d'adultes qui auraient innocenté Simpson juste avant le début du procès et la sélection des jurés. Supposons également que les 12 membres du jury soient sélectionnés de manière aléatoire et indépendante au sein de la population. (En réalité, ce ne fut pas le cas ; les avocats des deux parties sélectionnèrent 10 femmes et 2 hommes, dont 9 personnes afro-américaines, 2 de type caucasien, et 1 hispanique.)

i. Calculez la probabilité que le jury contienne un membre qui, avant l'ouverture du procès, défende l'innocence de Simpson. [*Astuce* : Définissez une variable aléatoire X distribuée selon une loi binomiale de paramètres $n = 12$ et $p = 20\%$, pour représenter le nombre de jurés qui, avant le début du procès, sont en faveur de O.J. Simpson.]

ii. Calculez la probabilité que le juré soit composé d'*au moins* deux membres qui croient en l'innocence de Simpson. [*Astuce* : $P(X \geq 2) = 1 - P(X \leq 1)$, et $P(X \leq 1) = P(X = 0) + P(X = 1)$.

6. (Cet exercice requiert quelques éléments de calcul) Soit X la peine de prison, en années, à laquelle est condamnée une personne coupable de vol de voiture dans un pays donné. Supposons que la FDP de X soit représentée par

$$f(x) = (1/9)x^2,\ 0 < x < 3.$$

Calculez l'intégrale de cette fonction pour trouver la peine de prison attendue.

7. Si un joueur de basket-ball réussit 74 % de ses lancers francs, combien de lancers francs est-il susceptible de réussir en moyenne dans un match durant lequel huit lancers francs ont lieu ?

8. Une étudiante à l'université suit trois cours : un cours à deux crédits, un cours à trois crédits et un cours à quatre crédits. Le résultat attendu pour un cours à deux crédits est égal à 3,5 [sur 4] alors qu'il est égal à 3 pour les cours à trois et quatre crédits. Quelle est la moyenne globale des résultats à laquelle elle peut s'attendre ? (*Astuce* : Chaque résultat est pondéré par la part des crédits que le cours représente dans le total.)

9. Soit X le salaire annuel des professeurs d'université aux États-Unis, en milliers de dollars. Si le salaire moyen est de 52,3 et que l'écart-type est de 14,6, que deviennent ces estimations lorsque le salaire est mesuré en dollars ?

10. Dans une grande université, la relation entre la moyenne générale obtenue à l'université (*GPA*) et le résultat obtenu au test d'admission à l'université (*SAT*) peut être synthétisée par l'espérance conditionnelle suivante : $E(GPA|SAT) = 0{,}70 + 0{,}002\ SAT$.

i. Calculez la valeur attendue de *GPA* lorsque $SAT = 800$. Calculez $E(GPA|SAT = 1\,400)$. Quelle différence observez-vous entre les deux ? Commentez.

ii. Si le résultat au test d'admission (*SAT*) est 1 100 *en moyenne* à l'université, quelle est la moyenne attendue de *GPA* au niveau de l'université ? (*Astuce* : Utilisez la propriété CE.4.)

iii. Si le résultat d'un étudiant au test d'admission (*SAT*) est 1 100, peut-on dire que sa moyenne générale à l'université (*GPA*) sera celle obtenue au point (ii) ? Expliquez.

ANNEXE

C

ÉLÉMENTS DE STATISTIQUE MATHÉMATIQUE

Traduction de Cédric Heuchenne et Marion Leturcq

C.1 POPULATIONS, PARAMÈTRES ET ÉCHANTILLONNAGE ALÉATOIRE

L'inférence statistique consiste à découvrir les caractéristiques d'une population à partir d'un échantillon issu de cette population. Par **population**, nous entendons un ensemble défini de sujets, tels que des personnes, des entreprises, des villes ; etc. Les caractéristiques qui peuvent être découvertes grâce à l'inférence statistique proviennent des *estimations* et des *tests d'hypothèses*.

Prenons deux exemples en guise d'illustration. Beaucoup d'économistes du travail s'intéressent au rendement du niveau d'instruction, mesuré par l'augmentation moyenne (en pourcentage) des rémunérations, due à une année supplémentaire d'études. Comme il serait impossible et trop coûteux de collecter des informations concernant la rémunération et le niveau d'instruction de chaque membre de la population active des États-Unis, les économistes du travail collectent des données relatives à un sous-ensemble de la population, sur base duquel ils peuvent obtenir leur meilleure *estimation* du rendement du niveau d'instruction. Cette estimation peut être *ponctuelle* (par exemple, 7.5 %) ou *par intervalle* (entre 5.6 % et 9.4 %, par exemple).

Les économistes s'intéressent aussi à la criminalité urbaine et cherchent à déterminer l'effet des programmes de surveillance dans les quartiers urbanisés sur le taux de criminalité. Il leur faut tirer un échantillon de la population et comparer les taux de criminalité dans des quartiers avec et sans programmes de surveillance. Ces économistes aboutissent ensuite à une des deux conclusions suivantes : soit les programmes de surveillance n'ont pas d'effet sur les taux de criminalité, soit ils en ont. Ce sont les *tests d'hypothèses* qui permettent de trancher.

La première étape d'une procédure d'inférence statistique est l'identification de la population d'intérêt. Cela peut sembler trivial mais il est important d'être très précis. Lorsque cette population est bien définie, il est possible de spécifier un modèle pour caractériser les relations qui nous intéressent dans cette population. Ces modèles reposent sur des distributions de probabilité (ou sur leurs caractéristiques, à tout le moins) qui dépendent de certains paramètres inconnus. Les paramètres représentent des constantes qui déterminent la direction et la force des relations entre les variables. Dans le premier exemple concernant les économistes du travail, le paramètre d'intérêt était le rendement du niveau d'instruction dans la population.

Échantillonnage

Pour réviser les bases de l'inférence statistique, nous adoptons le cadre de travail le plus simple possible. Soit la variable aléatoire Y qui représente une population avec une fonction de densité de probabilité $f(y ; \theta)$. [En anglais, cette fonction est appelée « *pdf* » pour « *probability density function* ».] Elle ne dépend ici que d'un seul paramètre θ. Nous supposons que cette fonction de densité est connue, à l'exception de la valeur de θ. Différentes valeurs de θ impliquent différentes distributions de la population ; c'est précisément la raison pour laquelle nous sommes intéressés par cette valeur de θ. Si nous pouvons obtenir des échantillons particuliers de la population, nous pouvons déduire de l'information sur la valeur de θ. Le système d'échantillonnage le plus simple est l'échantillonnage aléatoire.

Échantillonnage aléatoire. Si $Y_1, Y_2, \ldots, Y_n$ sont des variables aléatoires indépendantes dont la fonction de densité est $f(y ; \theta)$, alors $\{Y_1, Y_2, \ldots, Y_n\}$ est un échantillon aléatoire de $f(y ; \theta)$ ou de la population représentée par $f(y ; \theta)$.

Lorsque $\{Y_1, Y_2, \ldots, Y_n\}$ est un échantillon aléatoire de densité $f(y ; \theta)$, on dit aussi que les Y_i sont des variables aléatoires *indépendantes* et *identiquement distribuées* (ou *i.i.d.*) de $f(y ; \theta)$. Dans certains cas, il n'est d'ailleurs pas nécessaire de spécifier entièrement la distribution.

Dans la définition d'échantillonnage aléatoire, le caractère aléatoire de $Y_1, Y_2, \ldots, Y_n$ vient du fait que nous ne pouvons pas anticiper les résultats obtenus avant de procéder à l'échantillonnage. Par exemple, si nous obtenons les revenus familiaux d'un échantillonnage de $n = 100$ familles aux États-Unis, ces revenus ne seront pas les mêmes d'un échantillon à l'autre, même si la taille reste la même. Lorsque l'échantillon est disponible, nous disposons d'un ensemble de nombres, disons $\{y_1, y_2, \ldots, y_n\}$, qui constituent les données sur lesquelles nous pouvons travailler. Pour déterminer si un échantillon est issu d'un processus d'échantillonnage aléatoire, il est nécessaire de comprendre ce que représente un processus d'échantillonnage.

Les échantillons aléatoires qui suivent une distribution de Bernoulli sont souvent utilisés pour illustrer des concepts statistiques et traiter certaines applications concrètes. Si $Y_1, Y_2, \ldots, Y_n$ sont des variables aléatoires indépendantes dont la distribution *Bernoulli*(θ) est $P(Y_i = 1) = \theta$ et $P(Y_i = 0) = 1 - \theta$, alors $\{Y_1, Y_2, \ldots, Y_n\}$ est un échantillon aléatoire qui suit une distribution de *Bernoulli*(θ). Considérons l'exemple des réservations d'avion de l'annexe B. Chaque Y_i indique si le client i s'est présenté à l'aéroport ; $Y_i = 1$ si le client s'est présenté et $Y_i = 0$ sinon. Dans ce contexte, θ est la probabilité qu'un client, tiré au hasard dans la population de tous les clients, se présente à l'aéroport.

Dans beaucoup d'autres cas, nous pouvons aussi supposer que l'échantillon aléatoire est tiré au hasard à partir d'une distribution normale. Soit $\{Y_1, Y_2, \ldots, Y_n\}$ un échantillon aléatoire provenant d'une population *Normale*(μ, σ^2). Dans ce cas, la population est caractérisée par deux paramètres : la moyenne μ et la variance σ^2. La moyenne μ a un intérêt prépondérant mais σ^2 est également très important : sans information sur σ^2, il est impossible d'inférer quoi que ce soit sur μ.

C.2 ESTIMATEURS – PROPRIÉTÉS EN ÉCHANTILLONS FINIS

Dans cette section, nous étudions les propriétés en échantillons finis des estimateurs. Le terme « échantillon fini » vient du fait que ces propriétés sont valables pour tous les échantillons, quelle que soit leur taille. On parle parfois de « propriétés en petits échantillons ». La section C.3 couvre les « propriétés asymptotiques », qui décrivent le comportement des estimateurs lorsque la taille de l'échantillon augmente jusqu'à l'infini.

Estimateurs et Estimations

Avant d'étudier les propriétés des estimateurs, nous devons définir ce que représente un estimateur. Étant donné un échantillon aléatoire $\{Y_1, Y_2, \ldots, Y_n\}$ tiré au hasard à partir d'une population dont la distribution dépend d'un paramètre θ inconnu, un **estimateur** de θ est une règle de calcul qui permet d'attribuer une valeur de θ à chaque échantillon possible. Cette règle est spécifiée avant d'effectuer l'échantillonnage ; elle reste la même, quelles que soient les données effectivement obtenues.

Par exemple, si $\{Y_1, Y_2, \ldots, Y_n\}$ est un échantillon aléatoire qui est issu d'une population avec une moyenne μ, un estimateur naturel de μ est la moyenne empirique de l'échantillon aléatoire :

$$\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i. \qquad \textbf{[C.1]}$$

Soit $\bar{Y}$ la **moyenne de l'échantillon** ; contrairement à la moyenne définie dans l'annexe A (où elle correspondait à une statistique descriptive), $\bar{Y}$ est ici considéré comme un *estimateur*. Pour n'importe quelle réalisation des variables aléatoires $Y_1, \ldots, Y_n$, on utilise la même règle pour estimer μ. Pour les données réelles $y_1, y_2, \ldots, y_n$, l'**estimation** de μ est égal au nombre correspondant à la moyenne des données récoltées dans cet échantillon en particulier, soit $\bar{y} = (y_1 + y_2 + \ldots + y_n) / n$.

EXEMPLE C.1
Taux de chômage

Considérons ci-dessous un échantillon composé de différents taux de chômage dans 10 villes aux États-Unis.

Ville	Taux de Chômage
1	5,1
2	6,4
3	9,2
4	4,1
5	7,5
6	8,3
7	2,6
8	3,5
9	5,8
10	7,5

Notre estimation du taux de chômage moyen aux États-Unis est $\bar{y} = 6{,}0$. En règle générale, chaque échantillon conduit à une estimation différente mais la règle qui permet de calculer l'estimation est toujours la même, indépendamment des villes qui se trouvent dans l'échantillon ou de leur nombre.

Plus généralement, un estimateur W d'un paramètre θ peut être représenté par une formule mathématique abstraite :

$$W = h(Y_1, Y_2, \ldots, Y_n) \qquad \textbf{[C.2]}$$

pour des fonctions h des variables aléatoires $Y_1, Y_2, \ldots, Y_n$. Comme dans le cas spécial de la moyenne de l'échantillon, W est une variable aléatoire parce qu'elle dépend de l'échantillon aléatoire : si nous obtenons différents échantillons aléatoires de la population, la valeur de W peut changer. Lorsqu'un ensemble particulier de données $\{y_1, y_2, \ldots, y_n\}$ est inséré dans la fonction h, on obtient une *estimation* de θ, notée $w = h(y_1, y_2, \ldots, y_n)$. W est parfois appelée *estimateur ponctuel* et w *estimation ponctuelle* pour les distinguer des estimateurs et estimations *par intervalles*, présentés dans la section C.5.

Pour évaluer la qualité de la procédure d'estimation, nous présentons différentes propriétés de la distribution de probabilité de la variable aléatoire W. La distribution d'un estimateur est souvent appelée **distribution d'échantillonnage**, parce que cette distribution décrit la vraisemblance des différents résultats de W provenant des différents échantillons aléatoires. Puisqu'il y a un nombre illimité de règles pour combiner les données et estimer les paramètres, nous avons besoin de critères raisonnables pour opérer une sélection parmi les différents estimateurs disponibles ou, à tout le moins, écarter certains d'entre eux. Nous devons donc quitter le domaine des statistiques descriptives dans lequel la moyenne de l'échantillon, pour prendre un exemple, ne servait qu'à synthétiser un ensemble de données. L'étude de la distribution d'échantillonnage des estimateurs s'effectue à l'aide d'outils de statistique mathématique.

Biais

Étant donné la distribution de probabilité de Y_i et la fonction h, nous pouvons en principe obtenir toute la distribution d'échantillonnage de W. Il est habituellement plus facile de se concentrer sur certaines caractéristiques de la distribution de W lorsque W est évalué comme estimateur de θ. La première propriété importante d'un estimateur concerne son espérance.

Estimateur sans biais. Un estimateur W de θ est un **estimateur sans biais** si

$$E(W) = \theta \qquad \textbf{[C.3]}$$

pour toutes les valeurs possibles de θ.

Si un estimateur est sans biais, alors l'espérance de sa distribution de probabilité est égale au paramètre qu'il est supposé estimer. Disposer d'un estimateur sans biais *ne* veut *pas* dire que l'estimation obtenue à partir d'un échantillon quelconque sera égale à θ ou proche de θ. Nous ne pouvons pas le savoir. Cela veut plutôt dire que nous obtenons θ en effectuant la moyenne des estimations que nous pourrions obtenir en tirant un nombre illimité d'échantillons aléatoires dans la population. Comme, dans la plupart des cas, un seul échantillon aléatoire est disponible, cette réflexion reste abstraite mais la distinction entre estimation et estimateur reste tout aussi importante.

Si un estimateur est biaisé, le **biais** est défini comme suit.

Biais d'un estimateur. Si W est un **estimateur biaisé** de θ, le biais est

$$Biais(W) \equiv E(W) - \theta. \qquad \textbf{[C.4]}$$

La figure C.1 montre deux estimateurs ; le premier est sans biais et le second est biaisé vers le haut (ou positivement).

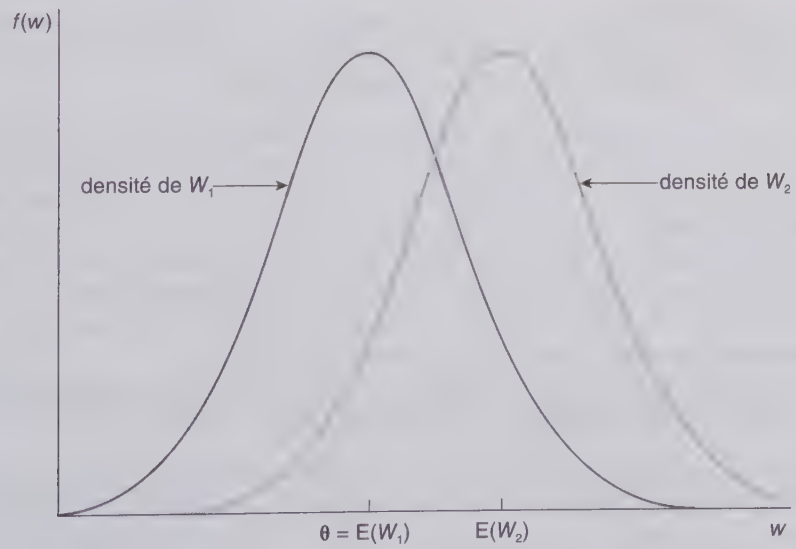


Figure C.1 *Un estimateur sans biais, W_1, et un estimateur biaisé vers le haut, W_2*

La présence et la taille d'un biais dépendent de la distribution de Y et de la fonction h. Nous n'avons généralement aucun contrôle sur la distribution de Y, qui peut être déterminée par la nature ou par une

dynamique sociale. Nous avons néanmoins le choix du *modèle* pour cette distribution, en particulier de la règle h que nous devons choisir en vue d'obtenir un estimateur sans biais.

Nous pouvons démontrer, de manière très générale, que certains estimateurs sont sans biais. Démontrons, par exemple, que la moyenne d'un échantillon $\bar{y}$ est un estimateur sans biais de la moyenne μ de la population, indépendamment de la distribution de la population sous-jacente. On utilise les propriétés de l'espérance (E.1 et E.2) couvertes dans la section B.3 :

$$E(Y) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \left(\frac{1}{n}\right)E\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu.$$

Pour effectuer des tests d'hypothèses, nous devons également estimer la variance σ^2 d'une population dont la moyenne est μ. Si $\{Y_1, \ldots, Y_n\}$ est l'échantillon aléatoire de la population dont $E(Y) = \mu$ et $Var(Y) = \sigma^2$, l'estimateur de la variance est défini par

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad \textbf{[C.5]}$$

qui est habituellement appelé **variance de l'échantillon**. On peut démontrer que S^2 n'est pas biaisé pour σ^2 : $E(S^2) = \sigma^2$. La division par $n - 1$, et non par n, tient compte du fait que la moyenne μ inconnue est estimée. Si μ était connue, un estimateur non biaisé de σ^2 serait $n^{-1}\sum_{i=1}^{n}(Y_i - \mu)^2$ mais μ est rarement connue dans la pratique.

Bien que l'absence de biais soit une propriété désirable pour un estimateur (un estimateur biaisé est clairement connoté négativement), elle ne résout pas tous les problèmes. En effet, certains estimateurs, même biaisés, peuvent bien se comporter. Nous en verrons un bref exemple.

Il existe également des estimateurs sans biais qui ne donnent pas de bons résultats. Considérons, par exemple, la moyenne μ d'une population. Au lieu d'utiliser la moyenne de l'échantillon $\bar{Y}$ pour estimer μ, supposons que nous écartions toutes les observations, sauf la première, après avoir collecté un échantillon de taille n. Autrement dit, l'estimateur de μ est simplement $W = Y_1$. Cet estimateur n'est pas biaisé parce que $E(Y_1) = \mu$. On comprend néanmoins que le fait de négliger toutes les observations, sauf une, n'est pas une bonne méthode d'estimation puisque la plus grande partie de l'information contenue dans l'échantillon est perdue. Par exemple, cela revient à n'utiliser qu'une seule observation pour estimer $E(Y)$ alors que 100 différentes observations de la variable aléatoire Y sont disponibles ($n = 100$).

La variance d'échantillonnage de l'estimateur

L'exemple précédent montre que nous avons besoin de critères complémentaires pour bien évaluer un estimateur. L'absence de biais garantit simplement que la valeur moyenne de la distribution d'échantillonnage d'un estimateur sera égale au paramètre qu'il est supposé estimer. Nous devons également nous renseigner sur la dispersion de la distribution de cet estimateur. En moyenne, un estimateur peut être égal à θ, tout en ayant une grande probabilité d'en être très éloigné. Dans la figure C.2, W_1 et W_2 sont tous les deux des estimateurs sans biais de θ, mais la distribution de W_1 est plus concentrée autour de θ : la probabilité que W_1 se situe au-delà de n'importe quelle distance donnée de θ est inférieure à la probabilité que W_2 se situe au-delà de la même distance de θ. En utilisant W_1 comme estimateur, la probabilité d'obtenir une estimation très éloignée de θ à partir d'un échantillon aléatoire est plus faible.

En résumé, la situation présentée à la figure C.2 s'inscrit dans l'étude de la variance (ou de l'écart-type) d'un estimateur. Comme vous le savez, la variance est une mesure de dispersion d'une distribution. La variance d'un estimateur est souvent appelée **variance d'échantillonnage** parce qu'elle est la variance associée à une distribution d'échantillonnage. Pour rappel, la variance d'échantillonnage n'est pas une variable aléatoire : il s'agit d'une constante, qui est parfois inconnue.

Nous obtenons la variance de la moyenne de l'échantillon (qui sert d'estimateur à la moyenne μ de la population) de la manière suivante :

$$Var(\bar{Y}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i)$$
$$= (1/n^2)\sum_{i=1}^{n} \sigma^2 = (1/n^2)(n\sigma^2) = \sigma^2/n. \qquad \textbf{[C.6]}$$

Notez que nous avons utilisé les propriétés de la variance (VAR.2 et VAR.4) des sections B.3 et B.4, ainsi que celle d'indépendance des Y_i. En résumé, si $\{Y_i : i = 1, 2, \ldots, n\}$ est un échantillon aléatoire tiré d'une population dont la moyenne est μ et la variance σ^2, alors $\bar{Y}$ a la même moyenne μ que celle de la population mais sa variance d'échantillonnage est égale à la variance de la population, σ^2, divisée par la taille de l'échantillon.

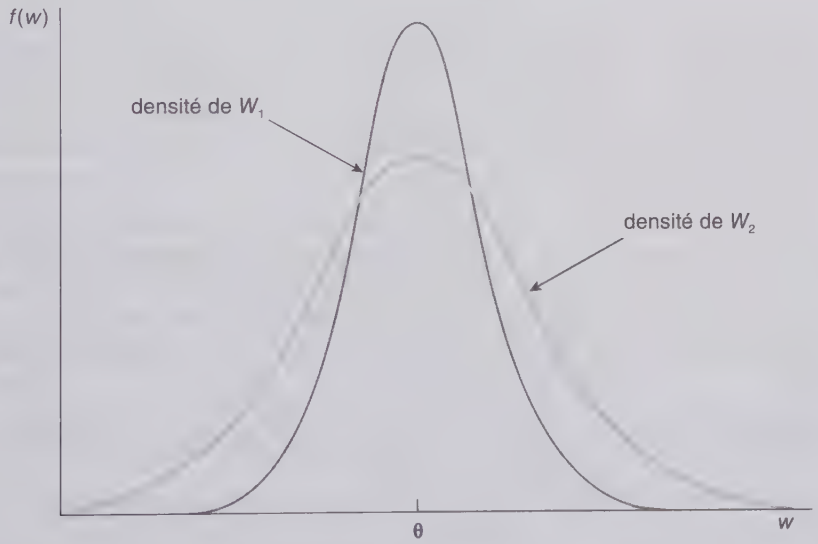


Figure C.2 *Les distributions d'échantillonnage de deux estimateurs sans biais de θ.*

Il est important de noter que la variance d'échantillonnage $Var(\bar{Y}) = \sigma^2/n$ peut s'approcher de zéro quand la taille d'échantillon augmente. C'est une caractéristique cruciale pour un estimateur « raisonnable » (ou désirable) et nous y reviendrons dans la section C.3.

Comme le suggère la figure C.2, parmi tous les estimateurs sans biais, nous préférons celui dont la variance est la plus faible. Cela conduit à l'élimination de plusieurs types d'estimateurs. Pour un échantillon aléatoire tiré d'une population dont la moyenne est μ et la variance σ^2, nous savons que $\bar{Y}$ n'est pas biaisé et que $Var(\bar{Y}) = \sigma^2/n$. Qu'en est-il de l'estimateur Y_1, celui qui correspond simplement à la première observation tirée ? Puisque la valeur de Y_1 est issue d'un tirage aléatoire dans la population, $Var(Y_1) = \sigma^2$. Par conséquent, il peut exister une différence considérable entre $Var(Y_1)$ et $Var(\bar{Y})$, même pour des échantillons

de petite taille. Si $n = 10$, alors $Var(Y_1)$ est 10 fois plus grande que $Var(\bar{Y}) = \sigma^2/10$. Nous avons donc une raison formelle d'écarter Y_1 comme estimateur désirable de μ.

Tableau C.1 *Simulations des estimateurs pour une distribution Normale(μ, 1) avec $\mu = 2$*

Réalisation	y_1	$\bar{y}$
1	−0,64	1,98
2	1,06	1,43
3	4,27	1,65
4	1,03	1,88
5	3,16	2,34
6	2,77	2,58
7	1,68	1,58
8	2,98	2,23
9	2,25	1,96
10	2,04	2,11
11	0,95	2,15
12	1,36	1,93
13	2,62	2,02
14	2,97	2,10
15	1,93	2,18
16	1,14	2,10
17	2,08	1,94
18	1,52	2,21
19	1,33	1,16
20	1,21	1,75

Pour insister sur ce point, le tableau C.1 présente les résultats d'un petit exercice de simulations. Au moyen du logiciel statistique Stata®, 20 échantillons aléatoires de taille 10 ont été générés à partir d'une distribution normale de paramètres $\mu = 2$ et $\sigma^2 = 1$. Nous nous intéressons à l'estimation de μ. Pour chacun des 20 échantillons aléatoires, deux estimations, y_1 et $\bar{y}$, sont calculées : leurs valeurs sont listées dans le tableau C.1. Nous pouvons constater que les valeurs de y_1 sont plus dispersées que celles de $\bar{y}$. En fait, y_1 varie entre −0,64 et 4,27 et $\bar{y}$ varie seulement entre 1,16 et 2,58. En outre, dans 16 cas sur 20, $\bar{y}$ est plus proche de $\mu = 2$ que ne l'est y_1. En considérant toutes les simulations, la moyenne de y_1 vaut environ 1,89, tandis que celle de $\bar{y}$ vaut 1,96. Ces valeurs, qui sont proches de 2, montrent que les deux estimateurs ne sont pas biaisés (on pourrait d'ailleurs obtenir des valeurs encore plus proches de 2 si le nombre de simulations passait à 20). Par contre, si nous ne considérons que les moyennes des différentes estimations, nous ignorons le fait que la moyenne de l'échantillon $\bar{Y}$, en tant qu'estimateur de μ, est nettement supérieure à Y_1.

Efficacité

La comparaison des variances de $\bar{Y}$ et de Y_1, que nous venons de faire, constitue un exemple d'approche générale que nous pouvons adopter pour comparer différents estimateurs sans biais.

Efficacité relative. Si W_1 et W_2 sont deux estimateurs sans biais de θ, W_1 est plus efficace que W_2 lorsque $Var(W_1) \leq Var(W_2)$ pour tout θ, avec une inégalité stricte pour au moins une valeur de θ.

Lors de l'estimation de la moyenne d'une population, nous avons précédemment montré que $Var(\bar{Y}) < Var(Y_1)$ pour n'importe quelle valeur de σ^2 lorsque $n > 1$. Par conséquent, $\bar{Y}$ est plus efficace que Y_1 pour estimer μ. Cependant, même à ce stade, nous ne pouvons toujours pas choisir le meilleur estimateur sans biais sur base du critère de la plus petite variance. Étant donnés deux estimateurs sans biais de θ, un estimateur peut en effet avoir une variance plus petite pour *certaines* valeurs de θ et une variance plus grande pour *d'autres* valeurs de θ.

Si nous considérons la catégorie particulière des estimateurs sans biais, nous pouvons montrer que la moyenne de l'échantillon aura la plus petite variance. Le problème C.2 consiste à vérifier que $\bar{Y}$ affiche effectivement la variance la plus petite parmi tous les estimateurs sans biais, qui sont fonctions linéaires de $Y_1, Y_2, \ldots, Y_n$. Dans ce contexte, tous les Y_i ont la même moyenne, la même variance, et ne sont pas corrélés deux à deux.

Si nous ne limitons pas notre attention aux estimateurs sans biais, la comparaison des variances n'a pas vraiment de sens. Par exemple, pour estimer la moyenne μ de la population, nous pourrions utiliser un estimateur égal à zéro, indépendamment de l'échantillon sélectionné. Naturellement, la variance de cet estimateur est nulle (puisque l'estimateur est le même pour tous les échantillons aléatoires). Néanmoins, le biais de cet estimateur est égal à $-\mu$. Cet estimateur, égal à zéro, ne peut donc pas être un bon estimateur pour de grandes valeurs de $|\mu|$.

Une manière de comparer des estimateurs qui ne sont pas nécessairement sans biais est de calculer leur **erreur quadratique moyenne**, que nous noterons EQM. Si W est un estimateur de θ, alors l'EQM de W est définie par $\text{EQM}(W) = E[(W - \theta)^2]$. L'EQM mesure la distance qui sépare en moyenne l'estimateur W du paramètre θ. On peut démontrer que $\text{EQM}(W) = Var(W) + [Biais(W)]^2$. L'EQM($W$) dépend donc de la variance et du biais. Le calcul de l'EQM nous permet de comparer deux estimateurs qui peuvent être biaisés.

C.3 PROPRIÉTÉS ASYMPTOTIQUES DES ESTIMATEURS

Dans la section précédente, nous avons montré que l'estimateur Y_1 de la moyenne μ de la population n'était pas un bon estimateur. Même s'il n'est pas biaisé, cet estimateur Y_1 peut afficher une variance bien plus grande que la variance calculée à partir de la moyenne de l'échantillon. Notez aussi que la variance de l'estimateur Y_1 sera toujours la même, quelle que soit la taille de l'échantillon [puisque $n = 1$]. Par contre, la précision de l'estimateur $\bar{Y}$ s'améliore au fur et à mesure que n augmente, puisque sa variance diminue.

Ce raisonnement démontre que nous pouvons écarter certains estimateurs simples en étudiant leurs propriétés asymptotiques (ou en grands échantillons). De plus, en procédant de la sorte, nous pouvons également identifier des avantages liés à l'utilisation d'estimateurs potentiellement biaisés, dont les variances sont difficiles à calculer.

L'analyse asymptotique consiste à approximer les caractéristiques de la distribution d'échantillonnage d'un estimateur. Ces approximations dépendent de la taille de l'échantillon. Malheureusement, il est difficile d'identifier la taille de l'échantillon qui rend l'analyse asymptotique appropriée ; en réalité, cela dépend de la distribution sous-jacente de la population. Il est néanmoins couramment admis que les approximations en grands échantillons fonctionnent bien pour des échantillons dont la taille est au minimum $n = 20$.

Convergence

La première propriété asymptotique d'un estimateur concerne la distance entre les valeurs possibles de l'estimateur et le paramètre qu'il est supposé estimer, lorsque la taille de l'échantillon augmente indéfiniment.

Convergence. Si W_n est un estimateur de θ basé sur l'échantillon $Y_1, Y_2, \ldots, Y_n$ de taille n, alors W_n est un **estimateur convergent** si pour tout $\varepsilon > 0$,

$$P(|W_n - \theta| > \varepsilon) \rightarrow 0 \text{ si } n \rightarrow \infty. \qquad \textbf{[C.7]}$$

Quand W_n est convergent, θ représente la limite en probabilité de W_n, écrite $\text{plim}(W_n) = \theta$.

Contrairement au biais (qui est une caractéristique d'un estimateur pour une taille d'échantillon donnée), la convergence concerne le comportement de la distribution d'échantillonnage de l'estimateur quand la taille d'échantillon devient grande. Pour souligner cet aspect de l'analyse, nous avons ajouté à l'estimateur un indice qui représente la taille de l'échantillon ; cette convention est adoptée de manière systématique dans cette section.

L'équation (C.7) peut sembler assez technique et difficile à démontrer sur la base de principes fondamentaux de probabilité. Par contre, son interprétation est simple. Elle signifie que la distribution de W_n devient de plus en plus concentrée autour de θ lorsque $n \rightarrow \infty$. [Autrement dit, la dispersion de cette distribution autour du paramètre de la population diminue au fur et à mesure que la taille de l'échantillon grandit.] Cela signifie que, pour des échantillons de taille croissante, il est de moins en moins probable que W_n soit très éloigné de θ. Cette tendance est illustrée dans la figure C.3.

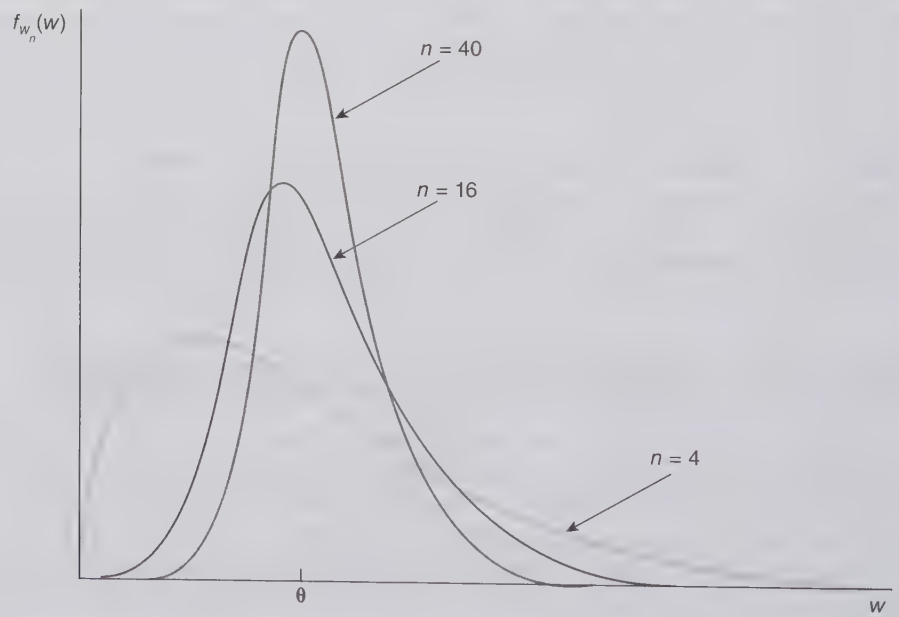


Figure C.3 *Les distributions d'échantillonnage d'un estimateur convergent pour des trois échantillons de tailles différentes.*

Lorsqu'un estimateur n'est pas convergent, il ne concourt pas à la découverte de θ, même dans l'hypothèse où une quantité illimitée de données serait disponible. C'est la raison pour laquelle la convergence est considérée comme une propriété indispensable dont doivent, à tout le moins, disposer les estimateurs utilisés en statistique ou en économétrie. Cette propriété de convergence dépend souvent de certaines hypothèses qui peuvent ne pas être vérifiées. Quand des estimateurs ne parviennent pas à converger, on peut habituellement

trouver leurs limites en probabilité ; il est alors important de connaître les distances qui séparent ces limites en probabilité du paramètre θ.

Comme nous l'avons indiqué précédemment, les estimateurs sans biais ne sont pas nécessairement convergents ; ils le sont néanmoins lorsque leur variance s'approche de zéro lorsque la taille de l'échantillon augmente. De manière plus formelle, si W_n est un estimateur sans biais de θ et $Var(W_n) \to 0$ quand $n \to \infty$, alors $\text{plim}(W_n) = \theta$. Les estimateurs sans biais, qui utilisent toute l'information présente dans l'échantillon de données, seront habituellement convergents car leur variance se rapproche généralement de zéro lorsque la taille de l'échantillon augmente.

Un bon exemple d'estimateur convergent est la moyenne d'un échantillon aléatoire tiré d'une population de moyenne μ et de variance σ^2. Nous avons déjà montré que la moyenne de l'échantillon n'est pas biaisée pour μ. Dans l'équation (C.6), nous avons montré que $Var(\bar{Y}_n) = \sigma^2 / n$ pour n'importe quelle taille n. Par conséquent, en plus d'être sans biais, $\bar{Y}_n$ est un estimateur convergent de μ puisque $Var(\bar{Y}_n) \to 0$ lorsque $n \to \infty$,

En conclusion, $\bar{Y}_n$ est un estimateur convergent de μ, même lorsque « $Var(\bar{Y}_n)$ disparaît ». Ce résultat classique est appelé **loi des grands nombres** (**LGN**).

Loi des grands nombres. Si $Y_1, Y_2, \ldots, Y_n$ sont des variables aléatoires indépendantes et identiquement distribuées de moyenne μ, alors

$$\text{plim}(\bar{Y}_n) = \mu \qquad \textbf{[C.8]}$$

La loi des grands nombres montre que, si nous nous intéressons à l'estimation de la moyenne μ d'une population, on peut obtenir des valeurs arbitrairement proches de μ en choisissant un échantillon suffisamment grand. Il est possible de combiner ce résultat fondamental avec les propriétés élémentaires des limites en probabilité pour démontrer que des estimateurs relativement complexes sont convergents.

Propriété PLIM.1 : Soit $\gamma = g(\theta)$ un paramètre défini par une fonction g continue du paramètre θ. Si $\text{plim}(W_n) = \theta$ et que $G_n = g(W_n)$ est un estimateur de γ, alors

$$\text{plim}(G_n) = \gamma. \qquad \textbf{[C.9]}$$

Souvent, on écrit

$$\text{plim}\ g(W_n) = g(\text{plim}(W_n)) \qquad \textbf{[C.10]}$$

pour une fonction continue $g(\theta)$.

L'hypothèse de continuité de la fonction $g(\theta)$ est une exigence technique qui a souvent été vulgarisée de la manière suivante : $g(\theta)$ est « une fonction qui peut être entièrement dessinée sans jamais soulever le stylo ». Dans cet ouvrage, toutes les fonctions que nous utilisons sont des fonctions continues ; il est donc inutile de donner une définition formelle de la continuité d'une fonction. Parmi les nombreux exemples de fonction continue, nous avons : $g(\theta) = a + b\theta$ (a et b étant des constantes), $g(\theta) = \theta^2$, $g(\theta) = \sqrt{\theta}$, $g(\theta) = \exp(\theta)$, y compris leurs variantes.

Comme exemple important d'estimateur convergent mais biaisé, considérons l'estimation de l'écart-type, σ, d'une population de moyenne μ et de variance σ^2. Nous avons déjà expliqué que la variance de l'échantillon $S_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$ n'était pas un estimateur biaisé de σ^2. En utilisant la loi des grands nombres et quelques manipulations algébriques, nous pouvons également démontrer que S_n^2 est convergent. L'estimateur naturel de $\sigma = \sqrt{\sigma^2}$ est $S_n = \sqrt{S_n^2}$ (où la racine carrée est toujours une racine carrée positive). S_n représente l'**écart-type de l'échantillon**, mais il *ne* s'agit *pas* d'un estimateur sans biais de σ, puisque l'espérance de la racine carrée n'est pas la racine carrée de l'espérance (voir section B.3).

Néanmoins, en utilisant PLIM.1, $\text{plim}(S_n) = \sqrt{\text{plim}(S_n^2)} = \sqrt{\sigma^2} = \sigma$, l'estimateur S_n est donc un estimateur convergent de σ.

Citons des autres propriétés utiles des limites en probabilité :

Propriété PLIM.2 : Si $\text{plim}(T_n) = \alpha$ et $\text{plim}(U_n) = \beta$, alors

(i) $\text{plim}(T_n + U_n) = \alpha + \beta$;

(ii) $\text{plim}(T_n U_n) = \alpha\beta$;

(iii) $\text{plim}(T_n / U_n) = \alpha/\beta$, si $\beta \neq 0$.

Ces trois propriétés des limites en probabilité permettent de combiner des estimateurs convergents pour obtenir d'autres estimateurs convergents. Par exemple, considérons $\{Y_1, Y_2, \ldots, Y_n\}$, un échantillon aléatoire de n revenus annuels tirés au hasard dans la population des travailleurs dont le diplôme le plus élevé correspond à celui des études secondaires ; notons la moyenne de cette population μ_Y. Considérons également $\{Z_1, Z_2, \ldots, Z_n\}$, un échantillon aléatoire de revenus annuels de travailleurs diplômés de l'université ; notons la moyenne de cette population μ_Z. Supposons que nous voulions estimer la différence en pourcentage entre les revenus annuels moyens de ces deux catégories de travailleurs, soit $\gamma = 100\,(\mu_Z - \mu_Y) / \mu_Y$. (Il s'agit donc de l'écart entre les revenus moyens de ces deux populations, exprimé en pourcentage du revenu moyen des travailleurs diplômés du secondaire). Comme $\overline{Y}_n$ et $\overline{Z}_n$ sont respectivement des estimateurs convergents de μ_Y et de μ_Z, nous pouvons déduire de PLIM.1 et de la partie (iii) de PLIM.2 que

$$G_n \equiv 100(\overline{Z}_n - \overline{Y}_n) / \overline{Y}_n$$

est un estimateur convergent de γ. Comme G_n correspond simplement à la différence en pourcentage entre $\overline{Z}_n$ et $\overline{Y}_n$ dans l'échantillon, G_n est un estimateur naturel de γ. Même si G_n est un estimateur biaisé de γ, il s'agit d'un estimateur utile dans de nombreux cas, sauf éventuellement quand n est petit.

Normalité asymptotique

La convergence est une propriété des estimateurs ponctuels[1]. Elle nous indique que la distribution de l'estimateur se resserre autour du paramètre quand la taille de l'échantillon s'accroît. Elle *ne* nous renseigne *pas* sur la *forme* que prend cette distribution pour une taille d'échantillon donnée. Pour construire des intervalles d'estimation et tester des hypothèses, nous devons pouvoir approximer les distributions des estimateurs. Quand l'échantillon est grand, la plupart des estimateurs en économétrie ont des distributions qui sont bien approximées par une distribution normale, ce qui nous conduit à énoncer la définition suivante.

Normalité asymptotique. Soit $\{Z_n : n = 1, 2, \ldots\}$ une suite de variables aléatoires, telle que pour tout nombre z,

$$P(Z_n \leq z) \to \Phi(z), \text{ quand } n \to \infty \qquad \textbf{[C.11]}$$

où $\Phi(z)$ est la fonction de répartition d'une loi normale centrée réduite. On dit alors que Z_n a une *distribution asymptotique normale centrée réduite*. Dans ce cas, on écrira souvent $Z_n \overset{a}{\sim} \text{Normal}(0,1)$. (Le « a » au-dessus du tilde signifie « asymptotiquement » ou « approximativement ».) La propriété (C.11) signifie que la fonction de répartition de Z_n est de plus en plus proche de la fonction de répartition d'une loi normale centrée réduite quand la taille de l'échantillon n s'accroît. Lorsque n est grand et que la **normalité asymptotique** est vérifiée, nous obtenons l'approximation suivante : $P(Z_n \leq z) \approx \Phi(z)$. Par conséquent,

1 Un estimateur ponctuel signifie qu'étant donné l'échantillon, l'estimateur ne fournira qu'une seule valeur (un point) du paramètre de la population envisagée.

les probabilités associées à Z_n peuvent être approximées par des probabilités d'une loi normale centrée réduite.

Le **théorème central limite (TCL)** est un des résultats les plus puissants en probabilité et en statistiques. Ce théorème affirme que la moyenne centrée réduite d'un échantillon aléatoire suit asymptotiquement une loi normale centrée réduite, *quelle que soit la population* dont cet échantillon est issu (à condition que la variance soit finie naturellement).

Théorème central limite. Soit $\{Y_1, Y_2, \ldots, Y_n\}$ un échantillon aléatoire de moyenne μ et de variance σ^2. Alors,

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma / \sqrt{n}} \qquad \textbf{[C.12]}$$

suit asymptotiquement une loi normale centrée réduite.

La variable Z_n dans [C.12] est la version centrée et réduite de $\bar{Y}_n$. On lui a soustrait $E(\bar{Y}_n) = \mu$ et on a divisé par $\sigma(\bar{Y}_n) = \sigma / \sqrt{n}$, où $\sigma(\bar{Y}_n)$ représente l'écart-type de $\bar{Y}_n$. Par conséquent, quelle que soit la distribution de la population Y, Z_n est de moyenne nulle et de variance unitaire, ce qui coïncide avec la moyenne et la variance d'une loi normale centrée réduite. Il faut remarquer ici que toute la distribution de Z_n devient arbitrairement proche de la distribution d'une loi normale centrée réduite quand n s'accroît.

On peut écrire la variable centrée réduite de l'équation [C.12] sous la forme $\sqrt{n}(\bar{Y}_n - \mu) / \sigma$, ce qui nous indique qu'il est nécessaire de multiplier la différence entre la moyenne de l'échantillon et la moyenne de la population par la racine carrée de la taille de la population afin d'obtenir une distribution asymptotique qui peut nous être utile. Sans la multiplication par $\sqrt{n}$, nous aurions seulement $(\bar{Y}_n - \mu) / \sigma$, ce qui converge en probabilité vers zéro. En d'autres termes, quand $n \to \infty$, la distribution de $(\bar{Y}_n - \mu) / \sigma$ se rétrécie autour d'un point qui ne représente pas une bonne approximation de la distribution de $(\bar{Y}_n - \mu) / \sigma$ pour des échantillons de taille raisonnable. Multiplier par $\sqrt{n}$ nous permet de garder la variance de Z_n constante. En pratique, on considère souvent que $\bar{Y}_n$ est approximativement distribuée selon une loi normale de moyenne μ et de variance σ^2 / n, ce qui nous permet de mettre en place des procédures statistiques correctes puisque qu'on aboutit à la variable centrée réduite de l'équation [C.12].

La plupart des estimateurs rencontrés en statistiques et en économétrie peuvent s'écrire en fonction des moyennes de l'échantillon, auquel cas on peut appliquer la loi des grands nombres et le théorème central limite. Quand deux estimateurs convergents suivent asymptotiquement une distribution normale, on choisit l'estimateur dont la variance est la plus faible.

En plus de la moyenne centrée réduite de l'échantillon présentée en [C.12], de nombreuses autres statistiques qui dépendent des moyennes de l'échantillon se trouvent être asymptotiquement normales. Un cas particulier important est la statistique obtenue en remplaçant σ par son estimateur convergent S_n dans l'équation [C.12] :

$$\frac{\bar{Y}_n - \mu}{S_n / \sqrt{n}} \qquad \textbf{[C.13]}$$

qui a également une distribution asymptotique normale centrée réduite quand n est grand. Pour un échantillon de taille finie, les distributions exactes des variables décrites en [C.12] et en [C.13] ne sont assurément pas les mêmes, mais la différence est souvent suffisamment petite pour qu'on puisse l'ignorer quand n est grand.

Tout au long de cette section, nous avons écrit chaque estimateur en ajoutant la lettre souscrite n, afin d'insister sur la nature de l'analyse asymptotique et de l'analyse de grands échantillons. Cette convention alourdit l'écriture sans rien apporter de plus, une fois que nous avons compris les fondements de l'analyse

asymptotique. Nous enlevons par la suite la lettre souscrite n et nous faisons confiance au lecteur pour se rappeler que les estimateurs dépendent de la taille de l'échantillon et que les propriétés de convergence et de normalité asymptotique font référence à l'accroissement sans limite de la taille de l'échantillon.

C.4 APPROCHES GÉNÉRALES DE L'ESTIMATION DE PARAMÈTRES

Nous avons, jusqu'à présent, utilisé la moyenne de l'échantillon pour illustrer les propriétés des estimateurs sur des échantillons de taille finie ou de grande taille. Il est naturel de se poser la question : existe-t-il des approches plus générales de l'estimation, qui nous donnent des estimateurs ayant les bonnes propriétés, c'est-à-dire qui ne soient pas biaisés, qui soient convergents et efficaces ?

La réponse est oui. Une explication détaillée des différentes approches de l'estimation irait au-delà des objectifs de cet ouvrage. Nous présentons ici une discussion informelle. Pour une discussion rigoureuse, voir Larsen et Marx (1986, Chapitre 5).

Méthode des moments

Considérons un paramètre θ d'une distribution définie sur une population. Il existe en général plusieurs façons d'obtenir des estimateurs sans biais et convergents de θ. Il est en pratique impossible de recourir à toutes ses méthodes pour comparer ensuite leur estimateur sur base des critères présentés dans les sections C.2 et C.3. Heureusement, il a été montré que certaines méthodes ont de bonnes propriétés générales et pour la plupart, la logique sur laquelle ils sont construits est séduisante car intuitive.

Dans la section précédente, nous avons vu que la moyenne empirique est un estimateur non biaisé de l'espérance et que la variance empirique est un estimateur non biaisé de la variance de la population. Ces estimateurs sont des exemples d'estimateurs construits sur la méthode des moments. En général, une estimation basée sur la méthode des moments procède de la façon suivante : on montre que le paramètre θ est relié à quelques valeurs espérées de la distribution de Y, généralement $E(Y)$ ou $E(Y^2)$ (même si on fait parfois des choix plus exotiques). Imaginons par exemple que le paramètre d'intérêt soit relié à la moyenne de la population $\theta = g(\mu)$ où g est une fonction connue. Puisque la moyenne empirique $\bar{Y}$ est un estimateur non biaisé et convergent de μ, il est naturel de remplacer μ par $\bar{Y}$, ce qui nous donne l'estimateur $g(\bar{Y})$ pour θ. L'estimateur $g(\bar{Y})$ est convergent vers θ et si, de plus, $g(\mu)$ est une fonction linéaire en μ, alors $g(\bar{Y})$ est également un estimateur non biaisé. Nous avons donc remplacé le moment de la population μ par sa contrepartie empirique $\bar{Y}$, d'où l'appellation « méthode des moments ».

Nous allons maintenant présenter deux estimateurs supplémentaires obtenus grâce à la méthode des moments qui seront utiles au moment de discuter l'analyse par régression. La covariance entre deux variables aléatoires X et Y est définie par $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$. La méthode des moments nous suggère d'estimer σ_{XY} par $n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$. C'est en effet un estimateur convergent de σ_{XY}, mais il se trouve être biaisé, pour les mêmes raisons que la variance empirique est biaisée si on utilise comme dénominateur n, plutôt que $n - 1$. La covariance empirique est définie par

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}). \qquad \textbf{[C.14]}$$

On peut montrer qu'il s'agit là d'un estimateur sans biais de σ_{XY}. (Remplacer n par $n - 1$ ne fait aucune différence quand la taille de l'échantillon s'accroît indéfiniment, l'estimateur est donc encore convergent.)

Comme nous l'avons présenté dans la section B.4 de l'annexe B, la covariance entre deux variables est souvent difficile à interpréter. En général, on préfère s'intéresser aux corrélations. Du fait que la corrélation de la population s'écrit $\rho_{XY} = \sigma_{XY} / (\sigma_X \sigma_Y)$, la méthode des moments nous suggère d'estimer ρ_{XY} par :

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{1/2}\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)^{1/2}}, \qquad \textbf{[C.15]}$$

que nous pouvons appeler **coefficient de corrélation empirique** (ou *corrélation empirique*, ou coefficient de corrélation de l'échantillon). Notez que nous avons simplifié le quotient par $n - 1$ dans l'expression de la covariance empirique et des écarts-types empiriques. En fait, nous aurions pu diviser chacun de ces termes par n, et nous serions arrivés à la même formule.

On peut montrer que le coefficient de corrélation empirique appartient toujours à l'intervalle $[-1,1]$, comme il se doit. Puisque S_{XY}, S_X et S_Y sont convergents pour les paramètres de population auxquels ils correspondent, R_{XY} est un estimateur convergent de la corrélation de la population, ρ_{XY}. Cependant, R_{XY} est un estimateur biaisé pour deux raisons. D'abord, S_X et S_Y sont des estimateurs biaisés de σ_X et σ_Y respectivement. Ensuite, R_{XY} est un quotient d'estimateurs ; il serait donc biaisé, même si S_X et S_Y ne l'étaient pas. Étant donné nos objectifs, cela n'a pas d'importance pour nous, même si le fait qu'il n'existe pas d'estimateur non biaisé de ρ_{XY} est un résultat classique de statistique mathématique.

Maximum de vraisemblance

Une autre approche générale de l'estimation est la méthode du maximum de vraisemblance, un thème couvert dans la plupart des cours d'introduction à la statistique. Nous nous contenterons ici de présenter un bref résumé pour le cas le plus simple. Soit $\{Y_1, Y_2, \ldots, Y_n\}$ un échantillon aléatoire tiré dans une population de distribution $f(y ; \theta)$. Puisque nous faisons l'hypothèse que l'échantillon est tiré aléatoirement, la distribution jointe de $\{Y_1, Y_2, \ldots, Y_n\}$ est simplement le produit des densités : $f(y_1 ; \theta) f(y_2 ; \theta) \ldots f(y_n ; \theta)$. Dans le cas discret, elle s'écrit $P(Y_1 = y_1, Y_1 = y_1, \ldots, Y_n = y_n)$. On définit maintenant la *fonction de vraisemblance* de la façon suivante :

$$L(\theta ; Y_1, \ldots, Y_n) = f(Y_1 ; \theta) f(Y_2 ; \theta) \ldots f(Y_n ; \theta)$$

Il s'agit là d'une variable aléatoire puisqu'elle dépend du résultat d'un échantillon aléatoire $\{Y_1, Y_2, \ldots, Y_n\}$. L'**estimateur par maximum de vraisemblance** de θ, que l'on appelle ici W, est la valeur de θ qui maximise la vraisemblance. (C'est pourquoi L est une fonction de θ, suivi d'un échantillon aléatoire.) Cette valeur dépend clairement de l'échantillon aléatoire. Le principe du maximum de vraisemblance est qu'il faut choisir pour θ la valeur qui, parmi toutes les valeurs possibles de θ, rend la vraisemblance des données observées la plus grande possible. C'est intuitivement une approche valable pour estimer θ.

En général, il est plus pratique de travailler avec la fonction de log-vraisemblance, qu'on obtient en prenant le logarithme naturel de la fonction de vraisemblance :

$$\log(L(\theta ; Y_1, \ldots, Y_n)) = \sum_{i=1}^{n} \log(f(Y_i ; \theta)) \qquad \textbf{[C.16]}$$

en utilisant le fait que le log d'un produit est la somme des logs. Puisque [C.16] est la somme de variables aléatoires indépendantes et identiquement distribuées, l'analyse d'estimateurs issus de [C.16] est relativement aisée.

L'estimateur par maximum de vraisemblance (EMV) est généralement convergent et parfois sans biais, mais c'est aussi le cas de nombreux autres estimateurs. Quand le modèle $f(y ; \theta)$ est correctement spécifié, l'EMV est généralement l'estimateur le plus efficace asymptotiquement, ce qui le rend souvent attrayant. De plus, l'EMV est parfois l'estimateur sans biais dont la variance est la plus petite, c'est-à-dire qu'il est l'estimateur dont la variance est la plus petite parmi tous les estimateurs non biaisés de θ. [Voir Larsen et Marx (1986, Chapitre 5) pour vérifier ces assertions.]

Dans le chapitre 17, nous aurons besoin du maximum de vraisemblance pour estimer les paramètres de modèles économétriques poussés. En économétrie, nous sommes presque toujours intéressés par la distribution de Y conditionnelle à un ensemble de variables explicatives, qu'on appelle ici $X_1, X_2, \ldots, X_n$. On remplace donc la densité en [C.16] par $f(Y_i \mid X_{i1}, \ldots, X_{ik} ; \theta_1, \ldots, \theta_p)$, où la densité peut dépendre de p paramètres, $\theta_1, \ldots, \theta_p$. Heureusement, pour mettre en œuvre correctement les méthodes de maximum de vraisemblance, nous n'avons pas besoin de plonger trop profondément dans des problèmes de calculs ou dans la théorie statistique des grands échantillons. Voir Wooldridge (2010, Chapitre 13) pour une présentation de l'estimation par maximum de vraisemblance.

Moindres Carrés

Le troisième type d'estimateur, qui joue un rôle central dans cet ouvrage, est **l'estimateur par moindres carrés**. Un exemple de moindres carrés a déjà été présenté : la moyenne empirique, $\bar{Y}$, est l'estimateur par moindres carrés de l'espérance, μ. On sait déjà que $\bar{Y}$ est un estimateur par la méthode des moments. Comment sait-on qu'il s'agit aussi d'un estimateur par moindres carrés ? On peut montrer que la plus petite somme possible des carrés des différences

$$\sum_{i=1}^{n} (Y_i - m)^2$$

est obtenue pour la valeur de m telle que $m = \bar{Y}$. La démonstration n'est pas difficile, mais nous ne présentons pas ici le calcul.

Pour certaines distributions importantes, en particulier les distributions normales et de Bernoulli, la moyenne empirique $\bar{Y}$ est aussi l'estimateur du maximum de vraisemblance de l'espérance μ. Ainsi, les principes des moindres carrés, de la méthode des moments et du maximum de vraisemblance aboutissent fréquemment au *même* estimateur. Avec d'autres distributions, les estimateurs sont similaires mais pas identiques.

C.5 ESTIMATION D'INTERVALLE ET INTERVALLES DE CONFIANCE

Une estimation ponctuelle obtenue sur la base d'un échantillon particulier n'apporte pas, en elle-même, suffisamment d'information pour tester des théories économiques ou pour éclairer les discussions sur la mise en place d'une politique. Une estimation ponctuelle peut représenter la meilleure idée que se fait le chercheur de la valeur d'un paramètre dans la population, mais par nature, elle ne nous dit en rien si l'estimation est *potentiellement* proche du paramètre de la population. Par exemple, prenons le cas d'un chercheur qui annoncerait, en se basant sur un échantillon aléatoire de salariés, que des subventions à la formation professionnelle augmentent le salaire horaire de 6,4 %. Dans quelle mesure peut-on savoir si cela est proche ou pas de l'augmentation qu'obtiendrait la population des salariés qui auraient pu recevoir la formation professionnelle ? Puisque nous ne connaissons pas la vraie valeur, nous ne sommes pas en mesure de savoir si l'estimation obtenue pour un échantillon particulier est proche de celle-ci. Nous pouvons cependant dire

quelque chose à ce sujet en nous appuyant sur les probabilités : c'est à ce moment là qu'intervient l'estimation d'intervalles.

Un moyen d'évaluer l'incertitude d'un estimateur est d'estimer son écart-type. Indiquer l'écart-type estimé, en plus de l'estimation ponctuelle, apporte de l'information sur la précision de l'estimateur. Cependant, même si l'on ignore le problème de la dépendance de l'écart-type à l'égard de certains paramètres inconnus de la population, donner l'écart-type estimé en plus de l'estimation ponctuelle ne permet pas de dire s'il est possible que la valeur dans la population soit en relation avec l'estimation. On peut dépasser cette limite en construisant un **intervalle de confiance**.

Nous illustrons le concept d'intervalle de confiance au moyen d'un exemple. Supposons que la population suive une distribution *Normale*(μ, 1) et définissons un échantillon aléatoire issu de cette population, soit $\{Y_1, \ldots, Y_n\}$. (Nous faisons ici l'hypothèse que la variance de la population est connue et égale à l'unité à des fins d'illustration. Nous montrerons ensuite ce qu'il convient de faire dans le cas plus réaliste d'une variance inconnue.) La distribution de la moyenne empirique, $\bar{Y}$, est une loi normale de moyenne μ et de variance $1/n$: $\bar{Y} \sim Normale\left(\mu, \frac{1}{n}\right)$. Sachant cela, nous pouvons centrer et réduire $\bar{y}$. Puisque la distribution de la version centrée et réduite de $\bar{Y}$ est une loi normale standard, nous savons que

$$P\left(-1{,}96 < \frac{\bar{Y} - \mu}{1/\sqrt{n}} < 1{,}96\right) = 0{,}95.$$

L'événement entre parenthèses est équivalent à l'événement $\bar{Y} - 1{,}96 / \sqrt{n} < \mu < \bar{Y} + 1{,}96 / \sqrt{n}$, ce qui nous donne

$$P(\bar{Y} - 1{,}96 / \sqrt{n} < \mu < \bar{Y} + 1{,}96 / \sqrt{n}) = 0{,}95. \qquad \textbf{[C.17]}$$

L'équation [C.17] est intéressante parce qu'elle nous dit que la probabilité que l'intervalle $[\bar{Y} - (1{,}96 / \sqrt{n}), \bar{Y} + (1{,}96 / \sqrt{n})]$ contienne l'espérance μ est de 0,95 ou 95 %. Cette information nous permet d'*estimer un intervalle de confiance* pour μ, qui est obtenu en injectant une observation de la moyenne empirique, c'est-à-dire $\bar{y} = \hat{\bar{Y}}$. Par conséquent, en utilisant cette observation, nous pouvons estimer un intervalle de confiance pour μ :

$$\left[\bar{y} - (1{,}96 / \sqrt{n}) ; \bar{y} + (1{,}96 / \sqrt{n})\right]. \qquad \textbf{[C.18]}$$

Celui-ci est aussi appelé l'intervalle de confiance à 95 %. On peut aussi abréger la notation en l'écrivant : $\bar{y} \pm (1{,}96 / \sqrt{n})$.

L'intervalle de confiance de l'équation [C.18] est facile à calculer lorsque les données de l'échantillon $\{y_1, y_2, \ldots, y_n\}$ sont disponibles. $\bar{y}$ est le seul facteur qui dépende des données. Par exemple, supposez que $n = 16$ et que la moyenne des 16 points est 7,3. Notre estimation de l'intervalle de confiance pour μ est alors $7{,}3 \pm (1{,}96 / \sqrt{16}) = 7{,}3 \pm 0{,}49$, ce qu'on peut écrire sous la forme d'un intervalle [6,81 ; 7,79]. Par construction, $\bar{y} = 7{,}3$ est au centre de cet intervalle.

S'il est facile de comprendre la construction d'un intervalle de confiance, il est plus difficile d'en saisir la signification. Quand nous disons que l'équation [C.18] correspond à un intervalle de confiance à 95 % de μ, nous voulons signifier que l'intervalle *aléatoire*

$$[\bar{Y} - \frac{1{,}96}{\sqrt{n}}, \bar{Y} + \frac{1{,}96}{\sqrt{n}}] \qquad \textbf{[C.19]}$$

contient μ avec une probabilité de 95 %. En d'autres termes, *avant de tirer l'échantillon*, il y a 95 % de chances que [C.19] contient μ. L'équation [C.19] est un exemple d'*estimateur* d'intervalle de confiance. C'est un intervalle aléatoire, puisque les bornes varient avec différents échantillons.

Tableau C.2 *Intervalles de confiance simulés à partir d'une loi normale (μ ; 1). Distribution telle que $\mu = 2$*

Réplication	$\bar{y}$	Intervalle à 95 %	Contient μ ?
1	1,98	(1,36 ; 2,60)	Oui
2	1,43	(0,81 ; 2,05)	Oui
3	1,65	(1,03 ; 2,27)	Oui
4	1,88	(1,26 ; 2,50)	Oui
5	2,34	(1,72 ; 2,96)	Oui
6	2,58	(1,96 ; 3,20)	Oui
7	1,58	(0,96 ; 2,20)	Oui
8	2,23	(1,61 ; 2,85)	Oui
9	1,96	(1,34 ; 2,58)	Oui
10	2,11	(1,49 ; 2,73)	Oui
11	2,15	(1,53 ; 2,77)	Oui
12	1,93	(1,31 ; 2,55)	Oui
13	2,02	(1,40 ; 2,64)	Oui
14	2,10	(1,48 ; 2,72)	Oui
15	2,18	(1,56 ; 2,80)	Oui
16	2,10	(1,48 ; 2,72)	Oui
17	1,94	(1,32 ; 2,56)	Oui
18	2,21	(1,59 ; 2,83)	Oui
19	1,16	(0,54 ; 1,78)	Non
20	1,75	(1,13 ; 2,37)	Oui

Malheureusement, l'intervalle de confiance estimé [C.18] est souvent interprété comme suit : « La probabilité que la vraie valeur de la moyenne, μ, appartienne à [C.18] est égal à 95 % ». Cette interprétation n'est pas correcte. Une fois l'échantillon observé et $\bar{y}$ calculé, les bornes de l'intervalle sont simplement des nombres (6,81 et 7,79 dans l'exemple que nous venons de voir). Le paramètre de la population, μ, même s'il n'est pas connu, est également un nombre. Par conséquent, μ peut très bien ne pas appartenir à l'intervalle [C.18] ; nous ne pouvons d'ailleurs jamais le savoir avec certitude. En réalité, la probabilité n'a pas de sens lorsqu'elle s'applique à un intervalle spécifique, estimé à partir de données. Une bonne interprétation en termes de probabilité est la suivante : 95 % des intervalles de confiance, construits à partir d'échantillons

aléatoires, contiendront μ. [Notez bien que nous ne pouvons pas savoir avec certitude si l'échantillon [C.18] fait partie de ces 95 %.]

Afin de bien comprendre la signification d'un intervalle de confiance, le tableau C.2 présente les calculs menés pour 20 échantillons aléatoires (ou réplications) de taille $n = 10$, tirés dans une distribution suivant une loi normale (2 ; 1). Pour chacun des 20 échantillons, nous calculons $\bar{y}$ puis [C.18] est calculé de la façon suivante : $\bar{y} \pm 1{,}96 / \sqrt{10} = \bar{y} \pm 0{,}62$ (arrondi à deux décimales). Comme on peut le constater, l'intervalle de confiance n'est pas le même pour tous les échantillons aléatoires. 19 des 20 intervalles contiennent la valeur de la moyenne de la population μ. Dans le cas de l'échantillon noté 19, et seulement de celui-là, μ n'appartient pas à l'intervalle de confiance. En d'autres termes, 95 % des échantillons donnent un intervalle qui contient μ. Avec seulement 20 réplications, nous aurions pu ne pas obtenir précisément 95 % des intervalles incluant la valeur μ ; il se trouve que nous avons abouti à 95 % dans cet exemple-ci.

Intervalles de confiance de la moyenne quand la population est distribuée selon une loi normale

L'intervalle de confiance que nous dérivons de l'équation [C.18] nous permet d'illustrer la construction et l'interprétation des intervalles de confiance. En pratique, l'équation [C.18] n'est pas très utile dans le cadre de l'analyse d'une espérance, puisque nous avons supposé que la variance était connue et toujours égale à un. Il est facile d'étendre l'équation [C.18] au cas où l'écart-type σ est connu et peut prendre n'importe quelle valeur. Dans ce cas, l'intervalle de confiance est donné par :

$$\left[\bar{y} - \frac{1{,}96\sigma}{\sqrt{n}} ; \bar{y} + \frac{1{,}96\sigma}{\sqrt{n}}\right] \qquad \textbf{[C.20]}$$

Par conséquent, on construit facilement un intervalle de confiance pour μ quand on connaît la valeur de σ. Pour prendre en compte le fait que σ n'est pas connu, il faut utiliser une estimation s, qui correspond à l'écart-type estimé de σ. Cet écart-type est calculé de la façon suivante :

$$s = \left(\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\right)^{1/2} \qquad \textbf{[C.21]}$$

On obtient ensuite un intervalle de confiance qui dépend entièrement des données observées en remplaçant, dans l'équation [C.20], σ par son estimation s. Malheureusement, on ne conserve pas, en procédant de la sorte, un seuil de confiance à 95 % parce que s dépend d'un échantillon en particulier. En d'autres termes, l'intervalle *aléatoire* $\left[\bar{Y} \pm 1{,}96(S / \sqrt{n})\right]$ ne contient plus μ avec une probabilité de 0,95 parce que la constante σ a été remplacée par une *variable aléatoire S*.

Comment devons-nous procéder ? Plutôt que d'utiliser une distribution normale centrée réduite, il nous faut nous appuyer sur une distribution de Student, que l'on note t_{n-1} quand il s'agit d'une distribution de Student à $n - 1$ degrés de liberté. On s'appuie sur la distribution de Student parce qu'on sait que :

$$\frac{\bar{Y} - \mu}{S / \sqrt{n}} \sim t_{n-1} \qquad \textbf{[C.22]}$$

où $\bar{Y}$ et S représentent respectivement la moyenne et l'écart-type empiriques d'un échantillon aléatoire $\{Y_1, \ldots, Y_n\}$. Nous ne démontrons pas [C.22] ici, mais il est possible de trouver des démonstrations précises dans un grand nombre d'ouvrages (par exemple, Larsen et Marx, 1986, chapitre 7).

Afin de construire l'intervalle de confiance à 95%, on a besoin du 97,5^e centile de la distribution t_{n-1}, que l'on appelle c. En d'autres termes, c est la valeur telle que 95 % de l'aire sous la courbe de t_{n-1} est comprise entre $-c$ et c : $P(-c < t_{n-1} < c) = 0{,}95$. (La valeur de c dépend du nombre de degrés de liberté $n - 1$, même si on ne le rend pas explicite dans les notations.) Le choix de c est illustré par la Figure C.4.

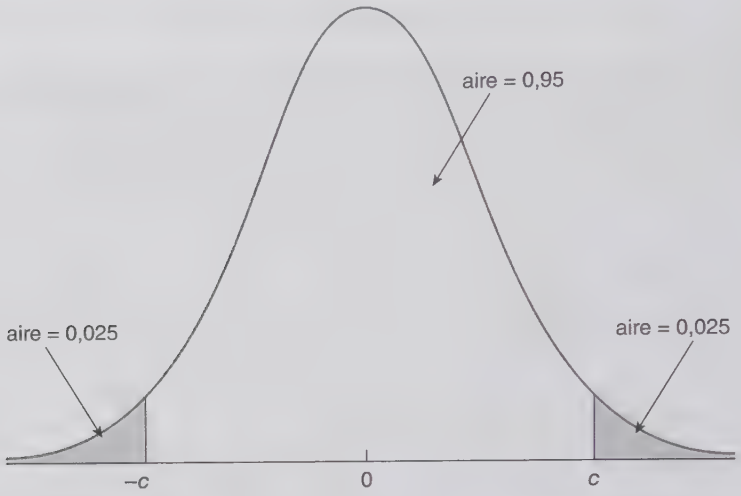


Figure C.4 *Le 97,5e centile, appelé c, d'une distribution t.*

Une fois que c est bien choisi, l'intervalle aléatoire $\left[\bar{Y} - c \cdot s/\sqrt{n}\ ;\ \bar{Y} + c \cdot s/\sqrt{n}\right]$ contient μ avec une probabilité 0,95. Pour un échantillon en particulier, l'intervalle de confiance à 95 % est calculé de la façon suivante :

$$\left[\bar{y} - c \cdot s/\sqrt{n}\ ;\ \bar{y} + c \cdot s/\sqrt{n}\right] \qquad \textbf{[C.23]}$$

Les valeurs de c peuvent être obtenues dans le tableau G.2 de l'annexe G, pour différents degrés de libertés. Par exemple, si $n = 20$, de telle façon que le nombre de degrés de liberté (noté *ddl*) est $ddl = n - 1 = 19$, alors $c = 2{,}093$. Ainsi, l'intervalle de confiance à 95 % est $\left[\bar{y} \pm 2{,}093 \cdot s/\sqrt{20}\right]$, pour lequel $\bar{y}$ et s sont des valeurs obtenues à partir de cet échantillon. Même dans le cas où $s = \sigma$ (ce qui est très peu probable), l'intervalle de confiance calculé avec [C.23] est plus grand que celui calculé avec [C.20] parce que $c > 1{,}96$. Pour un faible nombre de degrés de liberté, [C.23] est beaucoup plus grand que celui calculé avec [C.20].

Dans un cadre plus général, on appelle c_α le $100(1 - \alpha)^e$ centile d'une distribution t_{n-1}. Par conséquent, on construit l'intervalle de confiance à 100 $(1 - \alpha)$ % de la façon suivante :

$$\left[\bar{y} - c_{\alpha/2} \cdot s/\sqrt{n}\ ;\ \bar{y} + c_{\alpha/2} \cdot s/\sqrt{n}\right] \qquad \textbf{[C.24]}$$

Pour obtenir $c_{\alpha/2}$, il faut d'abord choisir α, connaître le nombre de degrés de liberté $n - 1$ et ensuite le tableau G.2 peut être utilisé. Dans la plupart des cas, on se concentrera sur des intervalles de confiance à 95 %.

EXEMPLE C.2
Formation professionnelle subventionnée et productivité des salariés

Holzer, Block, Cheatam et Knott (1993) étudient les effets de subventions à la formation professionnelle sur la productivité du salarié en collectant de l'information sur le taux de rebut pour un échantillon d'entreprises de production du Michigan. Le tableau C.3 établit la liste, pour 20 entreprises, des taux de rebut, qui mesurent le pourcentage d'objets qui ne sont pas utilisables et qui doivent par conséquent être jetés. Chacune de ces entreprises avait reçu en 1988 une subvention à la formation professionnelle et aucune subvention n'avait été accordée en 1987. Nous cherchons à construire un intervalle de confiance pour la *variation* du taux de rebut entre 1987 et 1988 pour la population des entreprises de production qui auraient pu recevoir des subventions.

On fait l'hypothèse que la variation du taux de rebut suit une distribution normale. Puisque $n = 20$, l'intervalle de confiance à 95 % de la moyenne de la variation du taux de rebut μ est donné par : $[\bar{y} \pm 2{,}093 \cdot \hat{\sigma}(\bar{y})]$, où $\hat{\sigma}(\bar{y}) = s / \sqrt{n}$. La valeur 2,093 est le 97,5e centile d'une distribution de Student à 19 degré de libertés, t_{19}. Pour les valeurs particulières de l'échantillon, $\bar{y} = -1{,}15$ et $\hat{\sigma}(\bar{y}) = 0{,}54$ (chacune arrondie à deux décimales), l'intervalle de confiance à 95 % est donc $[-2{,}28 ; -0{,}02]$. La valeur zéro n'appartient pas à cet échantillon, on peut donc en conclure que la variation moyenne du taux de rebut dans la population n'est pas zéro, à un seuil de confiance de 95 %.

En réalité, la construction d'un intervalle de confiance pour la moyenne d'une distribution normale s'effectue en quelques étapes simples. Rappelons-nous que l'écart-type de $\bar{Y}$ est : $\sigma(\bar{Y}) = \sigma / \sqrt{n}$. Par conséquent, $s / \sqrt{n}$ est l'estimation ponctuelle de $\sigma(\bar{Y})$. La variable aléatoire associée, $S / \sqrt{n}$, est aussi appelée **écart-type estimé de** $\bar{Y}$. Puisque les formules contiennent l'estimation ponctuelle $s / \sqrt{n}$ (et non $S / \sqrt{n}$), nous avons besoin de l'**écart-type estimé de** $\bar{y}$, soit $\hat{\sigma}(\bar{y}) = s / \sqrt{n}$. Ensuite, on peut écrire [C.24] de manière abrégée :

$$[\bar{y} \pm c_{\alpha/2} \cdot \hat{\sigma}(\bar{y})] \qquad \textbf{[C.25]}$$

Cette équation montre que la notion d'écart-type estimé joue un rôle important en économétrie.

Tableau C.3 *Taux de rebut pour 20 entreprises de production au Michigan*

Entreprise	1987	1988	Variation
1	10	3	−7
2	1	1	0
3	6	5	−1
4	0,45	0,5	0,05
5	1,25	1,54	0,29
6	1,3	1,5	0,2
7	1,06	0,8	−0,26
8	3	2	−1
9	8,18	0,67	−7,51
10	1,67	1,17	−0,5
11	0,98	0,51	−0,47
12	1	0,5	−0,5
13	0,45	0,61	0,16
14	5,03	6,7	1,67
15	8	4	−4
16	9	7	−2
17	18	19	1
18	0,28	0,2	−0,08
19	7	5	− 2
20	3,97	3,83	−0,14
Moyenne	4,38	3,23	−1,15

À ce niveau-là de l'analyse, l'exemple C.2 est principalement illustratif parce qu'il présente plusieurs lacunes du point de vue économétrique. Il part surtout de l'hypothèse que toute réduction systématique du taux de rebut est due aux subventions pour la formation professionnelle. Cependant, de nombreux événements peuvent avoir lieu au cours de l'année et altérer la productivité du salarié. À partir de cette analyse, nous ne pouvons en aucun cas attribuer la réduction du taux de rebut moyen aux subventions à la formation professionnelle, car d'autres forces extérieures peuvent en être responsables, au moins partiellement.

Une règle générale simple pour construire un intervalle de confiance à 95 %

L'intervalle de confiance de [C.25] peut être calculé pour n'importe quelle taille d'échantillon et pour n'importe quel seuil de confiance. Comme nous l'avons vu dans la section B.5, la distribution t tend vers une distribution normale quand le nombre de degrés de liberté devient grand. En particulier, lorsque $\alpha = 0{,}05$ et $n \to \infty$, nous avons $c_{\alpha/2} \to 1{,}96$, même si $c_{\alpha/2}$ est toujours plus grand que 1,96 pour tout n. En règle générale, on peut toujours approximer un intervalle de confiance à 95 % par

$$[\bar{y} \pm 2 \cdot \hat{\sigma}(\bar{y})] \qquad \textbf{[C.26]}$$

En d'autres termes, on calcule d'abord $\bar{y}$ et son écart-type estimé, puis on calcule plus ou moins deux fois son écart-type estimé pour obtenir l'intervalle de confiance. Cet intervalle de confiance approximatif sera un peu trop grand pour de très grands n et trop petit pour de petits n. Néanmoins, comme on peut le voir avec l'exemple C.2, même avec un n aussi petit que 20, [C.26] est proche de l'intervalle de confiance de la moyenne d'une loi normale. Cela veut dire qu'il est possible de se rapprocher fortement d'un intervalle de confiance à 95 % sans avoir à consulter les tables de Student.

Intervalles de confiance asymptotiques pour des populations non normales

Dans certaines applications, la population n'est clairement pas distribuée selon une loi normale. Un cas classique est celui de la distribution de Bernoulli, où la variable aléatoire prend seulement les valeurs zéro et un. Dans d'autres cas, la population non normale n'a pas de distribution classique. Cela n'a pas d'importance, à condition que la taille de l'échantillon soit suffisamment grande pour que le théorème central limite constitue une bonne approximation de la distribution de la moyenne empirique $\bar{Y}$. Pour n grand, l'intervalle de confiance peut être *approximé* par

$$[\bar{y} \pm 1{,}96 \cdot \hat{\sigma}(\bar{y})] \qquad \textbf{[C.27]}$$

où la valeur 1,96 est le 97,5^{e} centile d'une distribution normale centrée réduite. Mécaniquement, calculer une approximation de l'intervalle de confiance n'est pas différent du cas normal. Une petite différence vient du fait que le nombre utilisé pour multiplier l'écart-type estimé provient de la distribution d'une loi normale et non de la distribution d'une loi de Student, car nous nous situons dans l'analyse asymptotique. Puisque la distribution de Student se rapproche d'une normale centrée réduite quand le nombre de degrés de liberté augmente, l'équation [C.25] peut également être utilisée pour approximer un intervalle à 95 %. Certains préfèrent d'ailleurs utiliser cette dernière à [C.27] car [C.25] donne des résultats exacts pour les populations normales.

EXEMPLE C.3
Discrimination raciale à l'embauche

En 1988 à Washington (District of Columbia), l'*Urban Institute* a réalisé une étude dont l'objectif était d'analyser si la discrimination raciale influençait le processus de recrutement. Dix personnes, séparées en cinq paires, passaient différents entretiens d'embauche. Dans chaque paire, une personne était de type afro-américain et l'autre de type caucasien. On leur inventait chacun un CV équivalent en termes d'expérience, de niveau d'études et d'autres facteurs déterminant leur capacité à obtenir un emploi. L'idée était de rendre les individus les plus semblables possibles, à l'exception de leur couleur de peau. Chaque individu d'une paire passait l'entretien pour le même emploi et les chercheurs notaient lequel des deux se voyait proposer une offre d'emploi. Ceci est un exemple d'analyse de paires appariées, pour laquelle chaque essai est construit à partir de données sur deux individus (ou deux entreprises, deux villes, etc.) que l'on pense être similaires à de nombreux égards sauf pour une caractéristique importante.

Notons θ_B la probabilité que la personne noire se voit offrir un poste et θ_W la probabilité que ce soit la personne blanche. Nous sommes intéressés en premier lieu par la différence $\theta_B - \theta_W$. On appelle B_j une variable de Bernoulli qui prend la valeur 1 si la personne noire se voit proposer une offre d'emploi de la part de l'employeur j, et zéro sinon. De la même façon, pour la personne blanche, on définit $W_i = 1$ si elle se voit proposer une offre d'emploi par l'employeur i, et zéro sinon. En mettant ensemble les cinq paires d'individus, nous obtenons un total de $n = 241$ essais (paires d'entretiens avec les employeurs). Les estimateurs non biaisés de θ_B et de θ_w sont $\overline{B}$ et $\overline{W}$, la part des entretiens pour lesquels les noirs et les blancs, respectivement, se sont vus proposées les offres d'emploi.

Dans le cadre du calcul d'un intervalle de confiance relatif à une moyenne de population, on définit une nouvelle variable $Y_i = B_i - W_i$. Y_i peut prendre trois valeurs : – 1 si la personne blanche a décroché l'emploi (et uniquement elle) ; 0 si les deux personnes ont décroché un emploi ou n'ont rien reçu ; et 1 si la personne noire a obtenu un emploi (et uniquement elle). Alors, $\mu \equiv E(Y_i) = E(B_i) - E(W_i) = \theta_B - \theta_W$.

Bien évidemment, la distribution de Y_i n'est pas normale, la variable est discrète et prend seulement trois valeurs. Néanmoins, une approximation de l'intervalle de confiance pour $\theta_B - \theta_W$ peut être obtenue en utilisant les méthodes pour les grands échantillons.

Les données de l'étude menée par l'*Urban Institute* sont disponibles dans le fichier AUDIT. En utilisant les 241 points observés, $\overline{b} = 0{,}224$ et $\overline{w} = 0{,}357$, donc $\overline{y} = 0{,}224 - 0{,}357 = -0{,}133$. Ainsi, 22,4 % des candidats noirs ont reçu une offre d'emploi, alors que 35,7 % des candidats blancs en ont reçu une. De prime abord, ceci est une preuve de discrimination envers les noirs, mais on peut en apprendre bien davantage en calculant l'intervalle de confiance de μ. Pour calculer une approximation de l'intervalle de confiance à 95 %, nous avons besoin de l'écart-type estimé. Celui-ci est donné par $s = 0{,}482$ (en utilisant l'équation [C.21]). En utilisant l'équation [C.27], on obtient un intervalle de confiance à 95 % pour $\mu = \theta_B - \theta_W$, donné par $-0{,}133 \pm 1{,}96 \cdot \left(\frac{0{,}482}{\sqrt{241}}\right)$ $= -0{,}133 \pm 0{,}031 = [-0{,}164 ; -0{,}102]$. L'approximation de l'intervalle de confiance à 99 % nous donne : $-0{,}133 \pm 2{,}58 \cdot \left(\frac{0{,}482}{\sqrt{241}}\right) = [-0{,}213 ; -0{,}053]$. Bien entendu, ce dernier contient un plus grand ensemble de valeurs que l'intervalle de confiance à 95 %. Dans les deux cas, la valeur zéro n'appartient pas à l'intervalle de confiance. Par conséquent, nous sommes très confiants dans le fait que la différence $\theta_B - \theta_W$, au niveau de la population, n'est pas égale à zéro.

Avant de passer à la question des tests d'hypothèses, il peut être utile de revoir les quantités qui mesurent l'étalement des distributions de la population et des distributions empiriques des estimateurs. Ces quantités apparaissent souvent dans l'analyse statistique et leurs extensions sont importantes pour l'analyse de régression que nous utilisons dans le corps de l'ouvrage. La quantité σ est l'écart-type (inconnu) de la population ; c'est une mesure de l'étalement de la distribution de Y. Quand on divise σ par $\sqrt{n}$, on obtient **l'écart-type empirique de** $\overline{Y}$ ($\overline{Y}$ étant la moyenne de l'échantillon). Alors que σ est fixe dans la population, $\sigma(\overline{Y}) = \sigma/\sqrt{n}$ tend vers zéro lorsque $n \to \infty$: notre estimateur de μ devient de plus en plus précis quand la taille de l'échantillon s'accroît.

L'estimation de σ pour un échantillon particulier, s, est appelée écart-type estimé parce qu'il est obtenu à partir d'un échantillon. (L'écart-type empirique correspond à la variable aléatoire sous-jacente, S, qui varie quand l'échantillon change.) De la même façon que $\overline{y}$ est une estimation de μ, s est considérée comme la meilleure valeur de σ pour un échantillon particulier. La quantité $s/\sqrt{n}$ correspond à l'écart-type estimé de $\overline{y}$ et il s'agit également de notre meilleure estimation pour $\sigma/\sqrt{n}$. Les intervalles de confiance pour le paramètre μ de la population dépendent directement de $\hat{\sigma}(\overline{y}) = s/\sqrt{n}$. Puisque cet écart-type se rapproche de zéro quand la taille de la population croît, un plus grand échantillon va de pair avec un plus petit intervalle de confiance. Par conséquent, on voit clairement que disposer d'un grand échantillon nous permet d'obtenir des intervalles de confiance plus petits. La notion d'écart-type d'une estimation, qui se rapproche de zéro à la vitesse $1/\sqrt{n}$ dans la majorité des cas, joue un rôle fondamental lorsqu'il s'agit de : tester des hypothèses (ce que nous verrons dans la section suivante) ; construire des intervalles de confiance ; et mettre en place des tests dans le contexte des régressions multiples (comme nous l'avons vu dans le chapitre 4).

C.6 TESTS D'HYPOTHÈSES

Jusqu'à présent, nous avons passé en revue les différentes façons de calculer des estimateurs ponctuels ; nous avons également appris à construire et interpréter des intervalles de confiance dans le cas de la moyenne de la population. Nous allons maintenant apprendre à utiliser des tests d'hypothèses pour répondre, de manière affirmative ou négative, à des questions que l'on retrouve fréquemment dans les études empiriques. En voici quelques exemples. (1) Les programmes de formations professionnelles augmentent-ils effectivement la productivité moyenne des salariés (voir exemple C.2) ? (2) Les noirs souffrent-ils de discrimination à l'embauche (voir exemple C.3) ? (3) Des lois plus répressives sur la conduite en état d'ébriété entraînent-elles une diminution du nombre d'arrestations de conducteurs sous l'emprise de l'alcool ?

Les notions de base

Afin d'illustrer les questions associées aux tests d'hypothèses, prenons l'exemple d'une élection. Supposons que deux candidats se présentent à une élection : le candidat A et le candidat B. Les résultats indiquent que 42 % de l'électorat a voté pour le candidat A, et 58 % pour le candidat B. Ceux-ci sont censés représenter les vrais pourcentages de vote dans la population, ce que nous considérons effectivement par la suite.

Le candidat A est convaincu du fait qu'il a reçu plus de votes ; il cherche donc à savoir si l'élection a été truquée. Ayant quelques connaissances de statistiques, le candidat A recrute une agence de conseil pour construire un échantillon de 100 électeurs et demander à chaque personne si elle a ou pas voté pour lui. Imaginons que, pour l'échantillon collecté, 53 personnes ont voté pour le candidat A. Sur cet échantillon, l'estimation du nombre de votes est 53 %, ce qui est clairement plus grand que la valeur enregistrée au niveau de la population, soit 42 %. Le candidat A doit-il pour autant en conclure qu'il y a eu fraude ?

Bien que cette hypothèse soit plausible, suggérant que le décompte des voix s'est effectué en défaveur du candidat A, elle n'en est pas vraie pour autant. Même si 42 % des personnes dans la *population* ont

effectivement voté pour le candidat A, il est possible que, dans un *échantillon* de 100 personnes, 53 aient effectivement voté pour le candidat A. La question peut se formuler de la manière suivante : dans quelle mesure cet échantillon nous apporte la preuve que le résultat officiel de 42 % annoncé est faux ?

Une façon de répondre à cette question est de mettre en place un **test d'hypothèses**. Soit θ la vraie part des votes que le candidat A est parvenu à récolter dans la population. L'hypothèse selon laquelle les résultats annoncés sont justes, peut être énoncée de la façon suivante :

$$H_0 : \theta = 0{,}42 \qquad \textbf{[C.28]}$$

Ceci est un exemple d'**hypothèse nulle**. H_0 correspond toujours à l'hypothèse nulle. Dans les tests d'hypothèses, l'hypothèse nulle joue un rôle similaire à celui de l'accusé lors d'un procès dans de nombreux systèmes judiciaires : de la même façon que l'accusé est présumé innocent jusqu'à ce sa culpabilité soit avérée, l'hypothèse nulle est présumée vraie jusqu'à ce que les données indiquent clairement le contraire. Dans l'exemple en question, le candidat A doit présenter des preuves sérieuses contre [C.28] afin de gagner le droit de recompter les voix.

L'hypothèse alternative, dans l'exemple de l'élection, consiste à supposer que la vraie part des voix pour le candidat A lors de l'élection est plus grande que 0,42 :

$$H_1 : \theta > 0{,}42 \qquad \textbf{[C.29]}$$

Afin de conclure que H_0 est fausse (et que H_1 est vraie), nous devons fournir des preuves contre H_0 « au-delà de tout doute raisonnable ». Combien de votes, sur les 100, sont nécessaires pour conclure que les faits vont clairement à l'encontre de H_0 ? En règle générale, un vote de plus, soit 43 votes sur un échantillon de 100, ne suffit pas pour remettre en cause la fiabilité des résultats de l'élection ; cette différence d'un vote peut très bien résulter de la variation attendue étant donné l'échantillon. Il n'est pas non plus nécessaire d'observer 100 votes en faveur du candidat A pour mettre en doute H_0. Pouvons-nous maintenant considérer que 53 votes sur 100 est un résultat suffisamment probant pour que nous puissions rejeter H_0 ? La réponse dépend de ce qu'on entend par « au-delà de tout doute raisonnable ».

Avant de chercher à quantifier l'incertitude dans le cadre des tests d'hypothèses, il nous faut clarifier un point important. Le lecteur a certainement remarqué que les hypothèses formulées par les équations [C.28] et [C.29] ne reflètent pas l'ensemble des possibles : il se pourrait que θ soit inférieur à 0,42. Pour l'application que nous sommes en train de considérer, cette possibilité ne nous intéresse pas vraiment, puisqu'elle ne remet pas en cause le résultat de l'élection. Par conséquent, nous pouvons simplement déclarer dès le début que nous ne prenons pas en compte les autres θ tels que $\theta < 0{,}42$. Certains auteurs, néanmoins, préfèrent écrire l'hypothèse nulle en incluant ces scénarios, auquel cas notre hypothèse nulle devient $H_0 : \theta \leq 0{,}42$. Écrite de cette façon, l'hypothèse nulle est une hypothèse nulle *composite*, dans le sens où elle permet plus d'une valeur sous H_0. (Par contraste, l'équation [C.28] est un exemple d'hypothèse nulle *simple*.) Pour ce type d'exemple, écrire l'hypothèse nulle comme dans [C.28] ou comme une hypothèse nulle composite n'a pas grande importance : si $\theta \leq 0{,}42$, la valeur la plus difficile à rejeter reste la valeur $\theta = 0{,}42$. (Autrement dit, si on rejette la valeur $\theta = 0{,}42$, contre $\theta > 0{,}42$, alors il nous faut logiquement rejeter toutes les valeurs inférieures à 0,42.) Par conséquent, la procédure de test, basée sur l'hypothèse énoncée en [C.28], aboutit à la même conclusion que lorsque $H_0 : \theta \leq 0{,}42$. Dans cet ouvrage, nous écrivons toujours les hypothèses nulles sous la forme d'une hypothèse simple.

Dans le cadre des tests d'hypothèse, il existe deux types d'erreurs. D'abord, nous pouvons rejeter l'hypothèse nulle alors qu'elle est vraie. Il s'agit d'une **erreur de type I**. Dans l'exemple de l'élection, on commet une erreur de type I si on rejette H_0 alors qu'il y a effectivement 42 % des personnes qui ont voté pour le candidat A. Le deuxième type d'erreur est de ne pas réussir à rejeter H_0 alors qu'elle est fausse. Il s'agit là d'une **erreur de type II**. Dans l'exemple de l'élection, une erreur de type II est commise si nous ne rejetons pas H_0 alors que $\theta > 0{,}42$.

Après avoir pris la décision de rejeter (ou pas) l'hypothèse nulle, il est possible qu'une erreur ait été commise. Certes, il nous est impossible de le savoir avec certitude, mais nous pouvons malgré tout calculer la probabilité de commettre une erreur de type I ou de type II. Les règles des tests d'hypothèses sont établies de telle façon que la probabilité de commettre une erreur de type I est plutôt faible. En général, la probabilité de commettre une erreur de type I, notée α, représente le **seuil de significativité** (ou parfois simplement le *seuil*) d'un test. On peut l'écrire de la façon suivante :

$$\alpha = P(Reject\ H_0 | H_0) \qquad \textbf{[C.30]}$$

Le « seuil α de significativité » (correspondant à la partie gauche de l'égalité) est égal à la « probabilité de rejeter l'hypothèse H_0 alors qu'elle est vraie » (correspondant à la partie droite de l'égalité).

Les tests d'hypothèses classiques nécessitent de spécifier dès le début un seuil de significativité. En spécifiant la valeur de α, nous quantifions notre tolérance à une erreur de type I. On prend souvent pour α les valeurs 0,10, 0,05 et 0,01. En posant $\alpha = 0{,}05$, le chercheur prend le risque de rejeter à tort H_0 dans 5 % des cas pour pouvoir rejeter H_0 à juste titre dans 95 % des cas.

Après avoir choisi le seuil de significativité, le chercheur voudrait naturellement minimiser la probabilité de commettre une erreur de type II. Cela revient à dire qu'il cherche à maximiser la **puissance du test** contre toutes les autres alternatives. La puissance d'un test est égale à 1 moins la probabilité de commettre une erreur de type II. Nous pouvons l'écrire sous la forme mathématique suivante :

$$\pi(\theta) = P(reject\ H_0 | \theta) = 1 - P(type\ \text{II} | \theta)$$

où θ représente la vraie valeur du paramètre. Évidemment, nous aimerions disposer d'une puissance maximale, égale à 1, lorsque l'hypothèse nulle est fausse. C'est malheureusement impossible d'obtenir une puissance aussi élevée tout en conservant un seuil de significativité faible. Pour un seuil de significativité donné, nous préférons choisir des tests qui maximisent la puissance.

Tester des hypothèses sur la moyenne dans une population normale

Afin de tester une hypothèse nulle contre une hypothèse alternative, nous devons choisir une statistique de test (ou, en abrégé, une statistique) ainsi qu'une valeur critique. La statistique et la valeur critique sont choisies pour leur commodité, d'une part, et pour leur capacité à maximiser la puissance du test pour un seuil de significativité donné, d'autre part. Dans cette sous-section, nous étudions les tests d'hypothèses sur la moyenne pour une population tirée d'une loi normale.

Une **statistique de test**, notée T, est une fonction d'un échantillon aléatoire. Quand on calcule la statistique pour un échantillon en particulier, on obtient une valeur particulière de la statistique de test, que l'on note t.

Étant donné la statistique du test, on peut définir une règle qui précise les conditions de rejet de H_0 en faveur de H_1. Dans cet ouvrage, toutes les règles de rejet sont basées sur la comparaison de la statistique de test, t, à une **valeur critique**, c. L'ensemble des valeurs de t qui conduisent à rejeter l'hypothèse nulle forment la **région de rejet**. Pour déterminer la valeur critique, nous devons en premier lieu choisir le seuil de significativité du test. Ensuite, étant donné le seuil α, la valeur critique associée à α est déterminée par la distribution de T, *sous l'hypothèse* que H_0 est vraie. Nous écrivons cette valeur critique c, sans faire apparaître le fait qu'elle dépende de α.

Il est facile de tester des hypothèses sur la moyenne μ à partir d'une population tirée d'une loi normale (μ, σ^2). L'hypothèse nulle s'écrit :

$$H_0 : \mu = \mu_0 \qquad \textbf{[C.31]}$$

où μ_0 est une valeur donnée. Dans la plupart des applications, $\mu_0 = 0$, mais le cas général n'est pas plus compliqué.

La région de rejet que l'on choisit dépend de la nature de l'hypothèse alternative. Les trois alternatives auxquelles on peut être amené à s'intéresser sont :

$$H_1 : \mu > \mu_0, \quad \textbf{[C.32]}$$

$$H_1 : \mu < \mu_0, \quad \textbf{[C.33]}$$

$$H_1 : \mu \neq \mu_0. \quad \textbf{[C.34]}$$

Les équations [C.32] et [C.33] correspondent à des hypothèses unilatérales (aboutissant à un test unilatéral) ; la région de rejet se situe dans une seule extrémité de la distribution. Quand l'hypothèse alternative est [C.32], l'hypothèse nulle composite est $H_0 : \mu \leq \mu_0$, puisque H_0 est rejetée lorsque $\mu > \mu_0$. Cette hypothèse est appropriée quand notre intérêt porte sur une valeur de μ inférieure à μ_0. L'équation [C.34] correspond à une hypothèse bilatérale (aboutissant à un test bilatéral) ; la région de rejet est située aux deux extrémités de la distribution. Cette hypothèse est appropriée quand nous cherchons à identifier tout écart par rapport à l'hypothèse nulle.

Analysons d'abord l'hypothèse alternative [C.32]. Intuitivement, quand l'observation de la moyenne empirique $\bar{y}$ est « suffisamment » grande par rapport à μ_0, nous devrions rejeter H_0 en faveur de H_1. Comment pouvons-nous déterminer si $\bar{y}$ est suffisamment grande pour que nous puissions rejeter H_0 à un seuil de significativité donné ? Pour répondre à cette question, nous avons besoin de calculer la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie. Au lieu de travailler directement avec $\bar{y}$, nous utilisons sa version centrée réduite, dans laquelle on remplace σ par l'écart-type estimé s :

$$t = \sqrt{n}(\bar{y} - \mu_0)/s = (\bar{y} - \mu_0)/\hat{\sigma}(\bar{y}) \quad \textbf{[C.35]}$$

où $\hat{\sigma}(\bar{y}) = s/\sqrt{n}$ représente l'écart-type estimé de $\bar{y}$. Étant donné l'échantillon de données disponible, le calcul de t est aisé. Nous travaillons avec t parce que, sous l'hypothèse nulle, la variable aléatoire suivante suit une distribution de Student à $n - 1$ degrés de liberté, t_{n-1} :

$$T = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{S}$$

Supposons maintenant que le seuil de significativité choisi soit égal à 5 %. La valeur critique c qui résulte de ce choix est déterminée de façon à avoir $P(T > c | H_0) = 0{,}05$. Cela signifie que la probabilité de commettre une erreur de type I est égale à 5 %. Après avoir trouvé c, la règle de rejet est

$$t > c \quad \textbf{[C.36]}$$

dans laquelle c correspond à la valeur du $100(1 - \alpha)$ème centile d'une distribution t_{n-1}. Exprimé en pourcentage, le seuil de significativité est de $(100 \cdot \alpha)$ %. Ceci est un exemple de **test unilatéral**, car sa région de rejet est située sur une extrémité de la distribution t. À un seuil de significativité de 5 %, c est le 95e centile de la distribution t_{n-1}, comme l'illustre la Figure C.5. Un seuil de significativité différent nous donne une valeur critique différente.

La statistique de l'équation [C.35] est souvent appelée **statistique t** pour tester $H_0 : \mu = \mu_0$. La statistique t mesure la distance de $\bar{y}$ à μ_0 relativement à l'écart-type estimé de $\bar{y}$, c'est-à-dire $\hat{\sigma}(\bar{y})$.

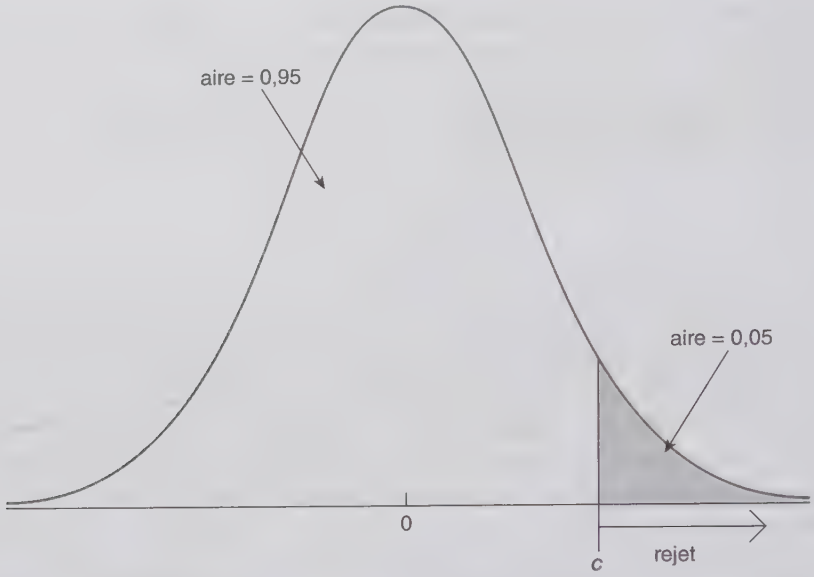


Figure C.5 *Région de rejet d'un test au seuil de significativité de 5 % contre une hypothèse alternative unilatérale* $H_0 : \mu > \mu_0$

EXEMPLE C.4
Zones économiques spéciales et investissements privés

Considérons la population des villes de l'État d'Indiana qui bénéficient de zones économiques spéciales (voir Papke, 1994). L'échantillon comprend 36 villes. La variable Y représente la variation, exprimée en points de pourcentage, des investissements entre l'année qui a précédé l'octroi du statut de zone économique spéciale et celle qui a suivi. On suppose que Y suit une loi normale (μ , σ^2). L'hypothèse nulle selon laquelle les zones économiques spéciales n'ont pas d'effets sur l'investissement privé s'écrit : $H_0 : \mu = 0$. L'hypothèse alternative selon laquelle elles ont un effet positif s'écrit : $H_1 : \mu > 0$. (On suppose ici qu'elles ne peuvent pas avoir d'effet négatif.) Nous voulons tester H_0 au seuil de 5 %. La statistique de test dans ce cas est

$$t = \bar{y} / (s / \sqrt{n}) = \bar{y} / \hat{\sigma}(\bar{y}) \qquad \textbf{[C.37]}$$

La valeur critique c est égale à 1,69 (voir tableau G.2). On rejettera H_0 en faveur de H_1 si $t > 1{,}69$. Vérifions si tel est le cas. L'échantillon nous permet de calculer $\bar{y} = 8{,}2$ et $s = 23{,}9$. Par conséquent, $t \approx 2{,}06$ et H_0 est effectivement rejetée à un seuil de 5 %. En d'autres termes, si nous acceptons de prendre le risque de commettre une erreur de type I dans 5 % des cas, nous pouvons conclure qu'il existe en moyenne une relation positive entre la création de zones économiques spéciales et l'investissement privé. À un seuil de 1 %, la valeur critique devient 2,44 ; H_0 n'est donc plus rejetée. Comme dans l'exemple C.2, nous n'avons pas tenu compte des autres facteurs qui ont pu également influencer les investissements privés entre les deux années de référence. Il serait donc erroné de conclure que le test démontre la présence d'un effet *causal* entre la création de zones économiques spéciales et l'investissement privé.

Lorsque l'hypothèse alternative correspond à [C.33], la règle de rejet est très similaire à celle que nous venons de présenter. Étant donné un seuil de significativité de $100 \cdot \alpha$ %, un test unilatéral dans la partie gauche aboutit au rejet de H_0, en faveur de [C.33], lorsque

$$t < -c \qquad \textbf{[C.38]}$$

En d'autres termes, le rejet de H_0 intervient pour des valeurs négatives de la statistique t suffisamment éloignées de zéro (une valeur négative implique que $\bar{y} < \mu_0$).

Dans le cadre des tests bilatéraux, il convient d'adapter la valeur critique de façon à ce que le seuil de significativité du test reste α. Quand H_1 s'écrit sous la forme $H_1 : \mu \neq \mu_0$, le rejet de H_0 intervient lorsque la *valeur absolue* de $\bar{y}$ est éloignée de μ_0 : un $\bar{y}$ beaucoup plus grand *ou* plus petit que μ_0 augmente la probabilité de rejeter H_0, en faveur de H_1. Dans un tel cas de figure, pour un seuil donné égal à $100 \cdot \alpha$ %, la règle de rejet est

$$|t| > c \qquad \textbf{[C.39]}$$

où $|t|$ est la valeur absolue de la statistique t définie en [C.35]. C'est la raison pour laquelle il s'agit d'un **test bilatéral**. La valeur critique doit être choisie avec prudence : c correspond au $100 \cdot \alpha / 2$ centile d'une distribution t_{n-1}. Par exemple, si $\alpha = 0{,}05$, la valeur critique est celle du 97,5e centile d'une distribution t_{n-1}. Cela nous assure que H_0 est rejetée dans seulement 5 % des cas lorsqu'elle est vraie (voir figure C.6). Par exemple, si $n = 22$, la valeur critique est $c \approx 2{,}08$, correspondant au 97,5e centile d'une distribution t_{21} (voir tableau G.2). La valeur absolue de t doit être plus grande que 2,08 pour pouvoir rejeter H_0 en faveur de H_1 au seuil de 5 %.

Il est important de bien connaître la terminologie des tests d'hypothèse. Lorsque H_0 est valide, il est incorrect de déclarer que : « nous pouvons accepter H_0 au seuil de significativité de 5 % ». Sur le plan épistémologique, il convient de dire que « nous ne parvenons pas à rejeter H_0 en faveur de H_1 au seuil de significativité de 5 % ». En règle générale, lorsqu'un chercheur exploite une base de données, il *ne* parviendra *pas* à rejeter un grand nombre d'hypothèses nulles. Dans l'exemple précédent sur les élections, cela n'aurait eu aucun sens d'« accepter » à la fois $H_0 : \theta = 0{,}42$ et $H_0 : \theta = 0{,}43$, puisqu'il n'existe qu'une seule valeur vraie de θ. Par contre, il est tout à fait envisageable de ne rejeter aucune de ces deux hypothèses, sachant que la valeur vraie de θ n'est pas connue. Pour cette raison, il convient de dire « ne pas (parvenir à) rejeter H_0 » plutôt que « accepter H_0 ».

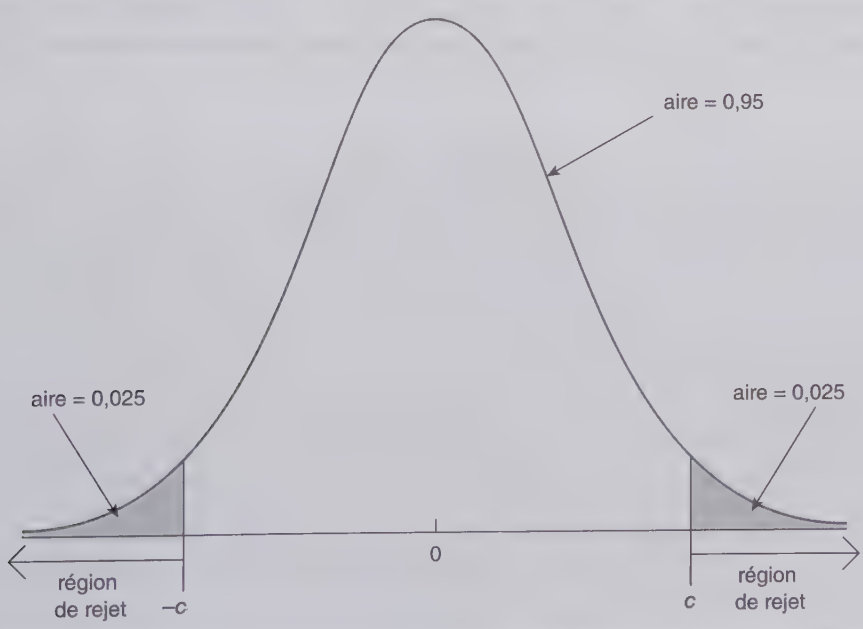


Figure C.6 *Région de rejet d'un test au seuil de significativité de 5 % contre une hypothèse alternative bilatérale* $H_1 : \mu \neq \mu_0$

Tests asymptotiques pour les populations non normales

Lorsque l'échantillon est suffisamment grand, le théorème central limite intervient (voir section C.3). Cela signifie que la procédure de test d'une hypothèse sur la moyenne de la population reste la même, que la population soit ou ne soit pas distribuée selon une loi normale. Sur le plan théorique, cela s'explique par le fait que, sous l'hypothèse nulle,

$$T = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{S} \overset{a}{\sim} Normale(0,1)$$

Par conséquent, lorsque n est grand, nous pouvons comparer les statistiques t de [C.35] aux valeurs critiques d'une distribution normale centrée réduite. Puisque la distribution t_{n-1} converge vers une distribution normale centrée réduite lorsque n s'accroît, les statistiques t sont très proches des valeurs critiques d'une loi normale centrée réduite quand n est extrêmement grand. En se basant sur un n qui s'accroît de manière illimitée, la théorie asymptotique ne peut pas nous dire quand les valeurs critiques d'une distribution normale centrée réduite sont préférables à celles d'une distribution t. Pour des valeurs moyennes de n comprises entre 30 et 60 environ, il est fréquent de recourir à la distribution de Student, puisque nous savons qu'elle convient lorsque la population est distribuée selon une loi normale. Lorsque $n > 120$, le choix entre la distribution normale ou celle de Student est en général sans intérêt, leurs valeurs critiques étant quasiment les mêmes.

Que ce soient les valeurs critiques de la loi normale centrée réduite ou la distribution de Student, ces valeurs ne sont qu'approximativement valides, ce qui implique que les niveaux de significativité que l'on choisit sont également approximatifs. Par exemple, pour des populations qui ne sont pas distribuées selon une loi normale, les niveaux de significativité correspondent à des niveaux de significativité *asymptotiques*. En d'autres termes, si un niveau de significativité de 5 % est choisi alors que la population n'est pas normalement distribuée, le niveau de significativité ne sera, dans les faits, pas *égal à* 5 %, sans que l'on puisse déterminer s'il est effectivement plus grand ou plus petit que 5 %. Quand la taille de l'échantillon est grande, le niveau de significativité sera très *proche de* 5 %. Comme cette distinction n'est pas importante sur le plan pratique, on ignore souvent l'adjectif « asymptotique », ce que nous faisons par la suite.

EXEMPLE C.5
Discrimination raciale à l'embauche

L'étude sur la discrimination à l'embauche de l'*Urban Institute* (voir exemple C.3) vise avant tout à tester H_0 : $\mu = 0$ contre H_1 : $\mu < 0$, sachant que $\mu = \theta_B - \theta_W$ représente la différence entre la probabilité pour une personne de type afro-américain et celle pour une personne de type caucasien de recevoir une offre d'emploi. Pour rappel, μ correspond à l'espérance d'une variable $Y = B - W$, où B et W sont des variables binaires. En utilisant les 241 paires dans la base AUDIT ($n = 241$), on obtient $\bar{y} = -0{,}133$ et $\sigma(\bar{y}) = 0{,}482/\sqrt{241} \approx 0{,}031$. La statistique t du test $H_0 : \mu = 0$ est donc $t = -0{,}133 / 0{,}031 \approx -4{,}29$. Nous avons vu dans l'annexe B qu'il n'est pas possible, pour des raisons pratiques, de distinguer la distribution normale centrée réduite de la distribution de Student à 240 degré de liberté. La valeur $-4{,}29$ se situe tellement loin dans l'extrémité gauche de la distribution que l'on peut rejeter H_0 à n'importe quel seuil de significativité raisonnable. En fait, à un seuil de 0,005 (soit un demi-pourcent), la valeur critique est $-2{,}58$ environ (pour un test unilatéral). En conclusion, une valeur t égale à $-4{,}29$ est une indication très claire à l'encontre de H_0 et en faveur de H_1 ; il y a manifestement discrimination à l'embauche sur le plan statistique.

Calcul et utilisation des p-valeurs

Si les chercheurs choisissent librement leur seuil de significativité (1 %, 5 % ou 10 %), ils peuvent très bien aboutir à des conclusions différentes quant à la significativité de leur test d'hypothèses, même lorsque les données sont identiques. Nous pouvons résoudre en partie ce problème en indiquant clairement le seuil de significativité retenu pour mener le test. Il est néanmoins possible de faire mieux.

Une meilleure information peut être obtenue en répondant à la question suivante : quel est le seuil de significativité le plus petit auquel H_0 est encore rejetée [en partant d'un seuil de significativité égal à 1] ? Réciproquement, quel est le seuil de significativité le plus élevé auquel H_0 n'est toujours pas rejetée [en partant d'un seuil de significativité égal à 0] ? De l'équivalence entre ces questions, nous pouvons déduire que la réponse correspond au seuil de significativité auquel nous sommes indifférents entre le rejet et le non rejet. Ce seuil est appelé la « *p*-valeur » ou la « valeur *p* » d'un test. [Elle mesure la probabilité *exacte* de commettre une erreur de type I]. Calculer la *p*-valeur est plus laborieux que choisir un seuil de significativité à l'avance. Néanmoins, grâce à la puissance de calcul des processeurs actuels, les *p*-valeurs sont souvent calculées de manière automatique.

En guise d'illustration, considérons l'hypothèse $H_0 : \mu = 0$ pour une population distribuée selon une loi normale (μ, σ^2). La statistique du test s'écrit $T = \sqrt{n} \cdot \bar{Y} / S$; si n est suffisamment grand, nous pouvons considérer que T suit une loi normale centrée réduite sous H_0. Imaginons que la valeur de T, observée dans notre échantillon, est $t = 1{,}52$. (Remarquez que nous avons délibérément sauté une étape, celle du choix relatif au seuil de significativité.) Maintenant que nous connaissons la valeur t, nous pouvons calculer la *p*-valeur. Il s'agit en fait du seuil de significativité auquel t est égal à la valeur critique. Puisque la statistique de test T suit une distribution normale centrée réduite sous H_0, nous avons :

$$p - valeur = P(T > 1{,}52 | H_0) = 1 - \Phi(1{,}52) = 0{,}065 \qquad \textbf{[C.40]}$$

$\Phi(\cdot)$ correspond à la fonction de répartition d'une loi normale centrée réduite. [En anglais, il s'agit de la « *cdf* » ou « *cumulative distribution function* »]. En d'autres termes, dans le cadre d'une distribution normale centrée réduite ; la *p*-valeur mesure l'aire qui se situe à droite de la valeur observée de la statistique de test, soit 1,52 dans notre exemple. Nous pouvons le constater sur la figure C.7.

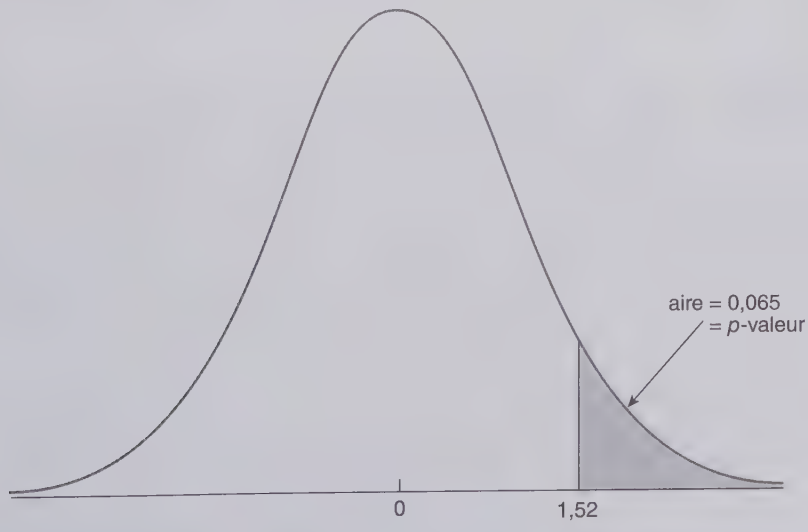


Figure C.7 *Calcul de la p-valeur lorsque t = 1,52 et que l'hypothèse alternative unilatérale est* $H_1 : \mu > \mu_0$

La *p*-valeur est ici égale à 0,065. En d'autres termes, le seuil de significativité le plus élevé auquel l'hypothèse nulle n'est pas rejetée est égal à 6,5 %. Si nous désirons effectuer le test à un niveau de significativité plus faible que 6,5 % (par exemple, 5 %), nous *ne* serons *pas* en mesure de rejeter H_0. Par contre, si nous voulons mener le test à un niveau plus élevé que 6,5 % (par exemple, 10 %), nous serons autorisés à rejeter H_0. Grâce au calcul de la *p*-valeur, nous sommes capables de mener le test à n'importe quel seuil de significativité désiré.

Dans cet exemple, la *p*-valeur offre également une autre interprétation utile : sous l'hypothèse nulle, elle mesure la probabilité d'observer une valeur de T aussi élevée que 1,52. Autrement dit, si l'hypothèse nulle est vraie dans les faits, la probabilité d'observer une valeur de T aussi élevée que 1,52 est égale à 6,5 %. La décision de rejeter H_0 à ce niveau dépend de notre tolérance au risque de commettre une erreur de type I. Comme nous allons le voir, la *p*-valeur s'interprète d'une façon similaire dans tous les autres cas.

En règle générale, une petite *p*-valeur est une preuve en *défaveur* de H_0. Si H_0 est vraie, la *p*-valeur mesure la probabilité d'observer un tel résultat à partir des données disponibles. [Plus la *p*-valeur est proche de 0, moins l'hypothèse nulle est plausible.] Dans l'exemple précédent, si t était plus grand, soit $t = 2{,}85$, alors la *p*-valeur serait de $1 - \Phi(2{,}85) \approx 0{,}002$. Cela signifie que, si l'hypothèse nulle est vraie, la probabilité d'observer une valeur de T aussi élevée que 2,85 est égale à 0,002, soit 0,2 %. Comment peut-on expliquer un tel résultat ? Soit nous sommes plutôt malchanceux car l'échantillon est particulièrement étrange, soit nous avons effectivement repéré une hypothèse nulle qui est fausse. À moins de vouloir faire disparaître toute erreur de type I, nous rejetons l'hypothèse nulle [la *p*-valeur de 0,2 % étant déjà très faible].

Dans le cas contraire, une *p*-valeur élevée constitue une preuve convaincante *en faveur* H_0. [Plus la *p*-valeur est proche de 1, plus l'hypothèse nulle est plausible.] Si nous avions obtenu $t = 0{,}47$ dans l'exemple précédent, alors nous aurions eu : *p*-valeur $= 1 - \Phi(0{,}47) = 0{,}32$. Si H_0 est vraie, la probabilité d'observer une valeur de T aussi élevée que 0,47 est égale à 32 %. Cette valeur est suffisamment grande pour ne pas douter de la plausibilité de H_0, à moins d'accepter de commettre une erreur de type I dans 32 cas sur 100 en la rejetant.

Si nous voulions calculer les *p*-valeurs à l'aide des tables de la distribution de Student, nous aurions besoin de tables particulièrement détaillées. Par exemple, le tableau G.2 ne nous permet que de déterminer des bornes inférieures et supérieures aux *p*-valeurs. Heureusement, le calcul des *p*-valeurs est automatisé dans de nombreux logiciels informatiques ; c'est également le cas pour les fonctions de répartition d'un grand nombre de distributions, dont celle de Student.

EXEMPLE C.6
Formation professionnelle subventionnée et productivité des salariés

Considérons à nouveau les données de Holzer *et al.* (1993) présentées dans l'exemple C.2. Dans une perspective de politique publique, deux questions importantes se posent. D'abord, quelle est notre meilleure estimation de la variation moyenne du taux de rebut, μ ? Nous avons déjà répondu à cette question en utilisant un échantillon de 20 entreprises indiquées au tableau C.3. La moyenne empirique de la variation du taux de rebut est $-1{,}15$. Par rapport au taux de rebut de départ, à savoir celui de 1987, cela représente une baisse du taux de rebut de 26,3 %, soit $-1{,}15/4{,}38 \approx -0{,}263$, ce qui n'est pas négligeable sur le plan économique.

L'autre question est la suivante : faudrait-il offrir un plus grand nombre de formations professionnelles subventionnées à la population des entreprises qui n'en bénéficient toujours pas ? Pour répondre à cette question, nous aimerions savoir si l'échantillon nous apporte une indication solide sur le plan statistique que la productivité des travailleurs (approximée par le taux de rebut) est effectivement liée à l'accès à un programme subventionné de formation professionnelle. L'hypothèse nulle est $H_0 : \mu = 0$ contre l'hypothèse alternative $H_1 : \mu < 0$, où μ est la variation moyenne du taux de rebut. Sous l'hypothèse nulle, il n'y a pas de lien. Dans les faits, l'hypothèse nulle composite est $H_0 : \mu \geq 0$, sachant que le cas $\mu > 0$ ne nous intéresse pas particulièrement.

Puisque $\bar{y} = -1{,}15$ et $\sigma(\bar{y}) = 0{,}54$, $t = -1{,}15/0{,}54 = -2{,}13$. Cette valeur est inférieure à la valeur critique à 5 %, elle-même égale à $-1{,}73$ (d'après la distribution t_{19}). Par contre, la valeur critique à 1 %, égale à $-2{,}54$, est inférieure à t. Calculons maintenant la p-valeur :

$$p - \text{valeur} = P(T_{19} < -2{,}13), \qquad \textbf{[C.41]}$$

où T_{19} représente une variable aléatoire qui suit une loi de Student à 19 degrés de liberté. L'inégalité est renversée par rapport à [C.40] parce que l'hypothèse alternative est de la forme de l'équation [C.33]. La probabilité exprimée en [C.41] représente l'aire située à gauche de $-2{,}13$ pour une distribution t_{19} (voir figure C.8).

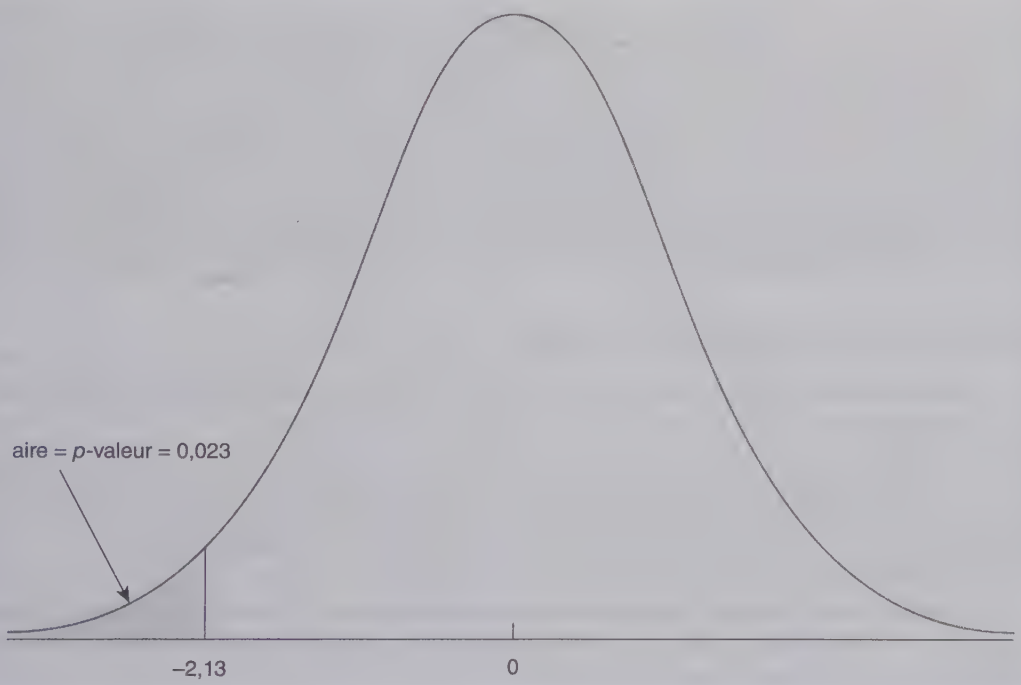


Figure C.8 *Calcul de la p-valeur lorsque $t = -2{,}13$ et que l'hypothèse alternative unilatérale est $H_1 : \mu < 0$*

En consultant le tableau G.2, nous pouvons conclure, tout au plus, que la p-valeur se trouve entre 0,025 et 0,01, tout en étant plus proche de 0,025 (puisque le 97,5e centile est égal à 2,09 environ). En utilisant un logiciel comme Stata, nous pouvons calculer cette p-valeur : elle est égale à 0,023 environ. Cette valeur est suffisamment faible pour contester la plausibilité de H_0. Si nous acceptons de prendre le risque de commettre une erreur de type I dans 2,5 % des cas (et *a fortiori* dans des 5 % des cas), nous pouvons rejeter l'hypothèse nulle selon laquelle la formation professionnelle n'est pas liée au niveau de productivité [puisque la p-valeur est inférieure à ce seuil].

Dans le cadre des tests bilatéraux, le calcul de la p-valeur est similaire. Il faut naturellement tenir compte du fait que la région de rejet se trouve dans les deux extrémités de la distribution. Pour un test de Student portant sur la moyenne d'une population, la p-valeur est calculée de la façon suivante :

$$P(|T_{n-1}| > |t|) = 2P(T_{n-1} > |t|), \qquad \textbf{[C.42]}$$

où t est la valeur de la statistique du test et T_{n-1} est une variable aléatoire qui suit une distribution de Student. (Pour un n grand, il faut remplacer T_{n-1} par une variable aléatoire normale centrée réduite.) La marche à

suivre consiste à calculer la valeur absolue de la statistique t, trouver l'aire à droite de cette valeur pour une distribution t_{n-1}, et multiplier cette aire par 2.

Pour des populations qui ne sont pas distribuées selon une loi normale, la p-valeur exacte peut être difficile à obtenir. Néanmoins, en utilisant les mêmes méthodes de calcul, nous pouvons obtenir les p-valeurs *asymptotiques*. Ces p-valeurs sont valides à condition que l'échantillon soit grand. Pour n plus grand que 120 environ, nous pouvons également utiliser la distribution normale centrée réduite. Le tableau G.1 est suffisamment détaillé pour obtenir une p-valeur précise, mais ce tableau n'est pas très utile puisque le calcul de la p-valeur est automatisé dans les logiciels actuels de statistiques ou d'économétrie.

EXEMPLE C.7
Discrimination raciale à l'embauche

En utilisant les données de l'*Urban Institute* ($n = 241$), pour lesquelles les deux membres de chaque paire sont appariés, nous avons obtenu $t = -4{,}29$. Si Z est une variable aléatoire normale centrée réduite, $P(Z < -4{,}29)$ est pratiquement égale à 0. En d'autres termes, la p-valeur (asymptotique) pour cet exemple est fondamentalement égale à zéro. Cela constitue une indication très solide que H_0 n'est pas plausible et qu'elle doit être rejetée.

L'utilisation de la p-valeur en résumé

i. Choisir une statistique de test T et déterminer le type d'alternative. Cela permet de déterminer si la région de rejet est $t > c$, $t < -c$, ou $|t| > c$.

ii. Utiliser la valeur observée de la statistique t comme valeur critique et calculer le seuil de significativité exact du test. Ce seuil exact correspond précisément à la p-valeur. Si la région de rejet est définie par $t > c$, alors p-valeur = $P(T > t)$. Si la région de rejet prend la forme $t < -c$, alors p-valeur = $P(T < t)$. Enfin, si la région de rejet est définie par $|t| > c$, alors p-valeur = $P(|T| > |t|)$.

iii. Si le seuil de significativité choisi au préalable est égal à α, H_0 est rejetée à un seuil de $100 \cdot \alpha$ % lorsque p-valeur < α. Par contre, si p-valeur ≥ α, H_0 n'est pas rejetée au seuil de $100 \cdot \alpha$ %. Le rejet de l'hypothèse nulle intervient donc pour de petites p-valeurs.

Relation entre un intervalle de confiance et un test d'hypothèses

La construction d'un intervalle de confiance et la conduite d'un test d'hypothèses font appel aux probabilités. Il est donc naturel de penser que ces deux méthodes sont liées, au moins en partie. Il est en effet possible de mener différents tests d'hypothèses après avoir construit un intervalle de confiance.

Les intervalles de confiance que nous avons analysés sont tous bilatéraux par construction. (Nous n'aurons pas besoin de construire des intervalles de confiance unilatéraux dans cet ouvrage.) Les intervalles de confiance peuvent donc être utilisés pour tester une hypothèse nulle contre une hypothèse alternative *bilatérale*. Pour la moyenne de la population, l'hypothèse nulle est donnée par [C.31] et l'alternative est [C.34]. Supposons qu'on ait construit un intervalle de confiance à 95 % pour μ. Si la valeur supposée « vraie » de μ, soit μ_0, n'appartient pas à l'intervalle de confiance, alors $H_0 : \mu = \mu_0$ est rejetée en faveur de $H_1 : \mu \neq \mu_0$ au seuil de 5 %. Si μ_0 appartient à l'intervalle, H_0 ne doit pas être rejetée au seuil de 5 %. Une fois l'intervalle de confiance construit, notez bien que n'importe quelle valeur μ_0 peut être testée. Étant donné qu'un intervalle de confiance contient plus d'une valeur, de nombreuses hypothèses nulles ne seront pas rejetées.

EXEMPLE C.8
Formation professionnelle subventionnée et productivité des salariés

Dans l'exemple emprunté à Holzer *et al.*, nous avons construit un intervalle de confiance à 95 % pour la variation moyenne du taux de rebut μ. Nous avons obtenu : [−2,28 ; −0,02]. Comme la valeur zéro n'appartient pas à l'intervalle, $H_0 : \mu = 0$ est rejetée en faveur de $H_0 : \mu \neq 0$ au seuil de 5 %. Cet intervalle à 95 % implique également que $H_0 : \mu = -2$ ne doit pas être rejetée à un seuil de 5 %. En réalité, pour tout intervalle de confiance, il existe un continuum d'hypothèses nulles qui ne doivent pas être rejetées.

Significativité statistique versus signification pratique

Dans les exemples que nous avons vus jusqu'ici, nous avons utilisé trois moyens pour améliorer notre connaissance, par définition imparfaite, des paramètres de la population : l'estimation ponctuelle, les intervalles de confiance et les tests d'hypothèses. Lorsqu'il s'agit de nous renseigner davantage sur les paramètres de la population, ces trois outils sont tout aussi importants les uns que les autres. Les étudiants ont souvent tendance à attacher plus d'importance aux intervalles de confiance et aux tests d'hypothèses, parce des seuils de confiance ou de significativité y leur sont liés. Néanmoins, il est très important de bien évaluer une estimation ponctuelle et d'en donner une interprétation adéquate sur le plan pratique ou économique.

Le *signe* et *l'ampleur* de l'estimation, soit $\bar{y}$ dans le cas de la moyenne, en déterminent la **signification pratique**. Le signe nous permet de vérifier si une décision prise par un acteur privé ou public, par exemple, s'accompagne d'un effet attendu sur la moyenne ; quant à l'ampleur, elle nous aide à déterminer si cet effet est négligeable ou substantiel. La **significativité statistique** de $\bar{y}$ dépend de la valeur de la statistique t. Pour tester $H_0 : \mu = 0$, la statistique t est simplement $t = \bar{y} / \hat{\sigma}(\bar{y})$. En d'autres termes, la statistique t dépend du ratio de $\bar{y}$ et de son écart-type estimé. Par conséquent, une statistique t peut être grande parce que $\bar{y}$ est élevée ou parce que $\hat{\sigma}(\bar{y})$ est faible. Dans les travaux empiriques, il est tout aussi important de discuter de la significativité statistique que de la signification pratique, en étant conscient qu'une estimation peut être significative sur le plan statistique sans être particulièrement importante d'un point de vue pratique (ou économique). Notez qu'il n'existe pas de règle préétablie qui permette de déterminer la signification d'une estimation sur le plan pratique : elle dépend autant du contexte général que d'une appréciation plus personnelle.

EXEMPLE C.9
Bandes autoroutières et temps de trajet entre domicile et lieu de travail

Soit Y une variable aléatoire qui mesure, en minutes, l'évolution du temps de trajet entre le domicile et le lieu de travail suite à l'élargissement d'une autoroute. Supposons que $Y \sim Normal\ (\mu, \sigma^2)$. Sous l'hypothèse nulle, l'ajout d'une bande autoroutière n'a pas réduit le temps de trajet moyen, soit $H_0 : \mu = 0$. L'hypothèse alternative stipule que le temps de trajet moyen est réduit, soit $H_1 : \mu < 0$. Nous disposons d'un échantillon aléatoire de 900 personnes qui effectuent le trajet chaque jour afin de déterminer si le projet d'élargissement de l'autoroute est efficace. La variation moyenne du temps de trajet est $\bar{y} = -3{,}6$ et l'écart-type estimé est $s = 32{,}7$. Par conséquent, $\hat{\sigma}(\bar{y}) = 32{,}7/\sqrt{900} = 1{,}09$ et $t = -3{,}6/1{,}09 \approx -3{,}30$, ce qui est très significatif d'un point de vue statistique. En effet, la p-valeur est égale à 0,0005 environ. La conclusion est que l'élargissement de l'autoroute s'est accompagné d'une réduction significative, sur le plan statistique, du temps de trajet moyen entre le domicile et le lieu de travail.

Si l'étude devait se contenter d'une analyse de la significativité statistique, nous en tirerions des conclusions très partielles. La significativité statistique ne nous renseigne pas sur le fait que la baisse estimée du temps moyen de trajet est assez faible, soit 3,6 minutes en moyenne. Il est donc indispensable d'indiquer à la fois l'estimation ponctuelle, égale à −3,6, et la p-valeur, égale à 0,05 % ; la première valeur nous renseigne sur la signification pratique alors que la seconde nous informe sur la significativité statistique.

En présence de grands échantillons, il peut arriver que les estimations ponctuelles soient statistiquement significatives mais marginales sur le plan pratique (ou économique). Pour en comprendre la raison, il est utile de définir ce que nous entendons par « test convergent ».

Convergence d'un test. Dans un test convergent, lorsque H_1 est vraie, la probabilité de rejeter H_0 tend vers 1 au fur et à mesure que la taille de l'échantillon croît.

Une autre façon d'exprimer la même idée est de dire qu'un test est convergent si la puissance de ce test s'approche de 1 lorsque H_1 est vraie et que la taille de l'échantillon tend vers l'infini. Tous les tests que nous utilisons dans cet ouvrage bénéficient de cette propriété. Par exemple, un test portant sur la moyenne de la population est convergent car la variance de $\bar{Y}$ converge vers zéro lorsque l'échantillon devient grand. Tester l'hypothèse $H_0 : \mu = 0$ repose sur l'utilisation de $T = \bar{Y} / (S / \sqrt{n})$. Comme $plim(\bar{Y}) = \mu$ et que $plim(S) = \sigma$, il s'en suit que T devient de plus en plus grand (en valeur absolue) lorsque $n \to \infty$. En d'autres termes, *quelle que soit la distance* entre μ et 0, nous pouvons être presque certains de rejeter $H_0 : \mu = 0$ lorsque l'échantillon est suffisamment grand [et que l'hypothèse alternative est correctement spécifiée]. Grâce à cette propriété de convergence, les tests statistiques sont plus précis en grands échantillons et l'hypothèse nulle est plus souvent rejetée. Cependant, cette propriété ne nous aide pas à évaluer l'importance de μ sur le plan pratique (ou économique). [En grands échantillons, il est donc possible d'obtenir des résultats très significatifs sur le plan statistique mais négligeables sur le plan pratique.]

C.7 REMARQUES SUR LA NOTATION

Dans la présentation des probabilités et des statistiques que nous avons faite dans cette annexe et dans la précédente, nous avons veillé à utiliser les conventions classiques de notation pour les variables aléatoires, estimateurs et statistiques de test. Par exemple, nous avons utilisé W quand il s'agissait d'un estimateur (variable aléatoire) et w quand il s'agissait d'une estimation en particulier (ou d'un résultat spécifique) de la variable aléatoire W. La distinction entre un estimateur et une estimation est importante pour bien comprendre les nombreux concepts sur lesquels reposent l'estimation des paramètres et les tests d'hypothèses. Néanmoins, cette distinction devient rapidement lourde à opérer lorsqu'il s'agit d'analyser des modèles économétriques plus sophistiqués, qui impliquent de nombreuses variables aléatoires et paramètres. Dans un tel cas de figure, rester fidèle aux conventions usuelles de probabilité et statistiques requiert l'utilisation de nombreux symboles.

Dans le corps de l'ouvrage, nous avons adopté une convention d'écriture plus simple, qui est largement utilisée en économétrie. Si θ est le paramètre de la population, la notation $\hat{\theta}$ (« thêta chapeau ») sera utilisée pour désigner à la fois l'estimateur et l'estimation de θ. Cette notation est pratique parce qu'elle offre une façon simple de lier l'estimateur au paramètre de population qu'il est sensé estimer. Par exemple, si le paramètre de la population est β, alors $\hat{\beta}$ représente un estimateur ou une estimation de β. Si le paramètre est σ^2, $\hat{\sigma}^2$ est un estimateur ou une estimation de σ^2, et ainsi de suite. Lorsque nous devons comparer deux estimateurs du même paramètre, nous utilisons une nouvelle notation, comme par exemple $\tilde{\theta}$ (on dit « thêta tilde »).

Le fait de ne pas recourir aux notations conventionnelles requiert une plus grande attention de la part du lecteur, mais la distinction entre un estimateur et une estimation n'est pas compliquée à assimiler au bout du compte. Lorsque nous analysons les propriétés *statistiques* de $\tilde{\theta}$ (pour savoir s'il est sans biais ou convergent), notre intérêt porte sur l'estimateur $\tilde{\theta}$. Par contre, si nous obtenons $\tilde{\theta} = 1{,}73$, nous sommes en présence d'une estimation ponctuelle basée sur un échantillon de données. Si le lecteur a bien compris les tenants et aboutissants de notre discussion sur les probabilités et les statistiques, il ne devrait pas y avoir de confusion entre les deux.

RÉSUMÉ

Nous avons présenté les éléments de statistiques mathématiques sur lesquels repose l'analyse économétrique. La notion d'estimateur est fondamentale : il s'agit d'une *règle de calcul* qui permet d'estimer un paramètre de la population à partir de données. Nous avons analysé différentes propriétés des estimateurs. En présence de petits échantillons, les propriétés les plus importantes sont celles d'absence de biais et d'efficacité. Cette dernière est déterminée en comparant la variance des estimateurs sans biais.

L'analyse économétrique s'appuie également sur les propriétés dont bénéficient les estimateurs en présence de grands échantillons. Ces propriétés, dites asymptotiques, sont déterminées en fonction du comportement que les estimateurs adoptent lorsque la taille de l'échantillon croît. Pour être utile, un estimateur doit être convergent. Dans les grands échantillons, le théorème central limite implique que la distribution d'échantillonnage de la plupart des estimateurs est approximativement normale.

La distribution d'échantillonnage d'un estimateur peut être utilisée pour construire des intervalles de confiance. Nous l'avons fait pour la moyenne d'une distribution normale et nous avons également construit des intervalles de confiance asymptotiques lorsque l'hypothèse de normalité n'était pas vérifiée. Nous avons enfin utilisé la méthode classique des tests d'hypothèses, qui requiert la spécification d'une hypothèse nulle, d'une hypothèse alternative et d'un seuil de significativité. Le rejet de l'hypothèse nulle dépend de la comparaison entre la statistique du test et une valeur critique. Quant au calcul de la *p*-valeur, il permet de tester l'hypothèse nulle à n'importe quel seuil de significativité.

MOTS-CLÉS

Biais p. 861
Coefficient de corrélation de l'échantillon (ou coefficient de corrélation empirique) p. 871
Covariance de l'échantillon (ou covariance empirique) p. 871
Distribution d'échantillonnage p. 860
Écart-type de l'échantillon (ou écart-type empirique) p. 867, 880
Échantillon aléatoire p. 858
Erreur quadratique moyenne p. 865
Estimateur p. 859
Estimateur biaisé p. 861
Estimateur convergent p. 866
Estimateur par maximum de vraisemblance p. 871
Estimateur par moindres carrés p. 872
Estimateur sans biais p. 861
Estimateur sans biais à variance minimale p. 871
Estimation p. 859
Estimation d'intervalle p. 872
Erreur de type I p. 881
Erreur de type II p. 881
Hypothèse alternative p. 881
Hypothèse alternative bilatérale p. 883
Hypothèse alternative unilatérale p. 883
Hypothèse nulle p. 881
Intervalle de confiance p. 873
Limite en probabilité p. 866
Loi des grands nombres p. 867

Méthode des moments p. 870
Moyenne de l'échantillon (ou moyenne empirique) p. 859
Non-convergent p. 866
Normalité Asymptotique p. 868
Population p. 858
Puissance d'un test p. 882
p-valeur p. 887
Région de rejet p. 882
Seuil de significativité (ou niveau de significativité) p. 882
Signification pratique p. 891
Statistique de Student (ou statistique t) p. 883
Statistique de test p. 882
Test bilatéral p. 885
Test convergent p. 892
Test d'hypothèse p. 881
Test unilatéral p. 883
Théorème central limite (TCL) p. 869
Valeur critique p. 882
Variance d'échantillonnage p. 863
Variance de l'échantillon (ou variance empirique) p. 862

EXERCICES

1. Y_1, Y_2, Y_3 et Y_4 sont des variables aléatoires indépendantes et identiquement distribuées, issues d'une population de moyenne μ et de variance σ^2. La moyenne de ces quatre variables aléatoires correspond à $\bar{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$.

i. Quelles sont la valeur espérée et la variance de $\bar{Y}$ (en termes de μ et de σ^2) ?

ii. Considérons maintenant un autre estimateur de μ :

$$W = \frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4$$

W est un exemple de moyenne *pondérée* des Y_i. Montrez que W est également un estimateur sans biais de μ. Calculez la variance de W.

iii. En vous appuyant sur vos réponses aux questions (i) et (ii), quel estimateur de μ préférez-vous, $\bar{Y}$ ou W ?

2. Cet exercice présente une version plus générale de l'exercice 1. $Y_1, Y_2, \ldots, Y_n$, sont n variables aléatoires non corrélées (deux à deux), toutes de même moyenne μ et de même variance σ^2. Soit $\bar{Y}$ la moyenne de l'échantillon.

i. On définit la classe des estimateurs linéaires de μ par

$$W_a = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n$$

où les a_i sont des constantes. Quelle restriction doit-on imposer aux a_i pour que W_a soit un estimateur non biaisé de μ ?

ii. Calculez $Var(W_a)$.

iii. Pour tous $a_1, a_2, \ldots, a_n$, l'inégalité suivante est vérifiée : $(a_1 + a_2 + \ldots + a_n)^2 \leq a_1^2 + a_2^2 + \ldots + a_n^2$. En utilisant cette inégalité et en utilisant vos réponses aux questions (i) et (ii), montrez que $Var(W_a) \geq Var(\bar{Y})$ pour tout W_a non biaisé, de sorte que $\bar{Y}$ est le meilleur estimateur linéaire non biaisé. (*Astuce* : Que devient l'inégalité quand a_i satisfait la restriction définie dans la question (i) ?)

3. Soit $\bar{Y}$ la moyenne empirique d'un échantillon aléatoire de moyenne μ et de variance σ^2. Considérez deux estimateurs alternatifs de μ : $W_1 = [(n-1)/n]\bar{Y}$ et $W_2 = \bar{Y}/2$.

i. Montrez que W_1 et W_2 sont deux estimateurs biaisés de μ et calculez le biais pour ces deux estimateurs. Que se passe-t-il lorsque $n \to \infty$? Commentez la différence importante qui existe entre les deux biais lorsque la taille de l'échantillon devient grande.

ii. Calculez les limites en probabilité de W_1 et de W_2. (*Astuce* : utilisez les propriétés PLIM.1 et PLIM.2. Pour W_1, remarquez que $plim[(n-1)/n] = 1$.) Quel estimateur est convergent ?

iii. Calculez $Var(W_1)$ et $Var(W_2)$.

iv. Expliquez la raison pour laquelle W_1 est un meilleur estimateur que $\bar{Y}$ lorsque μ est « proche » de zéro. (Tenez compte du biais et de la variance.)

4. Soient deux variables aléatoires positives, X et Y, pour lesquelles la valeur attendue de Y, étant donné X, est $E(Y|X) = \theta X$. Le paramètre inconnu θ nous montre que la valeur espérée de Y varie avec X.

i. Soit Z une variable aléatoire définie par $Z = Y/X$. Montrez que $E(Z) = \theta$. (*Astuce* : Utilisez les propriétés CE.2 et CE.4 de l'annexe B. Montrez d'abord que $E(Z|X) = \theta$; utiliser CE.4 par la suite.)

ii. Utilisez la question (i) pour montrer que, sur base de l'échantillon aléatoire $\{(X_i, Y_i) : i = 1, 2, \ldots, n\}$, l'estimateur $W_1 = (1/n)\sum_{i=1}^{n} Y_i / X_i$ est un estimateur sans biais de θ.

iii. Expliquez pourquoi l'estimateur $W_2 = \bar{Y} / \bar{X}$ est différent de W_1, sachant que $\bar{Y}$ et $\bar{X}$ représentent les moyennes empiriques. Montrez néanmoins que W_2 est également un estimateur sans biais de θ.

iv. Le tableau suivant contient des données sur les récoltes de maïs pour différents comtés de l'Iowa. Le Ministère de l'Agriculture des États-Unis prédit le nombre d'hectares cultivés de maïs dans chaque comté au moyen de photos satellites. Pour prédire le nombre effectif d'hectares, les chercheurs du Ministère comptent le nombre de « pixels » de maïs sur les photos satellites (par opposition au nombre de pixels de soja ou de terres non cultivées, par exemple). Pour améliorer leur modèle de prédiction, le Ministère interroge ensuite les agriculteurs des comtés sélectionnés afin d'obtenir la production de maïs en hectares. Soit Y_i la production de maïs dans le comté i et X_i le nombre de pixels de maïs dans la photo satellite du comté i. Il y a 17 terrains ($n = 17$) répartis dans 8 comtés. Utilisez cet échantillon pour calculer les estimations de θ évoquées dans les questions (ii) et (iii). Ces estimations sont-elles similaires ?

Terrain	Production de maïs	Pixels de maïs
1	165,76	374
2	96,32	209
3	76,08	253
4	185,35	432
5	116,43	367
6	162,08	361

Terrain	Production de maïs	Pixels de maïs
7	152,04	288
8	161,75	369
9	92,88	206
10	149,94	316
11	64,75	145
12	127,07	355
13	133,55	295
14	77,70	223
15	206,39	459
16	108,33	290
17	118,17	307

5. Soit Y une variable aléatoire qui suit une loi de *Bernoulli*(θ), avec $0 < \theta < 1$. Nous sommes intéressés par l'estimation du « rapport de cotes » (γ), égal à la probabilité de succès (γ) divisée par la probabilité d'échec ($1 - \theta$), soit $\gamma = \theta / (1 - \theta)$. Étant donné un échantillon aléatoire $\{Y_1, \ldots, Y_n\}$, on sait que $\overline{Y}$, la *proportion* de succès sur n essais, est un estimateur sans biais et convergent de θ. Un estimateur intuitif de γ est $G = \overline{Y}/(1 - \overline{Y})$, la proportion de succès divisée par la proportion d'échecs.

i. Expliquez pourquoi G est un estimateur biaisé de γ.

ii. Utilisez PLIM.2 (iii) pour montrer que G est un estimateur convergent de γ.

6. Le gouverneur d'un État vous demande de vérifier si la consommation moyenne de spiritueux dans cet État a diminué suite à l'instauration d'une taxe. Pour un échantillon d'individus sélectionnés aléatoirement, vous obtenez la différence de consommation de spiritueux (en once) entre les années antérieures et postérieures à la mise en place de cette taxe. Pour une personne i tirée aléatoirement dans la population, Y_i représente la variation de consommation de spiritueux. Considérez qu'il s'agit d'un échantillon aléatoire tiré d'une distribution normale (μ, σ^2).

i. L'hypothèse nulle suppose qu'il n'y a eu aucun changement dans la consommation moyenne de spiritueux. Écrivez-la formellement en fonction de μ.

ii. L'hypothèse alternative suppose qu'il y a eu un déclin dans la consommation de spiritueux. Écrivez-la formellement en fonction de μ.

iii. Supposons maintenant que vous disposiez d'un échantillon de taille $n = 900$ et que vous obteniez les estimations suivantes : $\overline{y} = -32{,}8$ et $s = 466{,}4$. Calculez la statistique t pour tester H_0 contre H_1 ; déduisez la p-valeur de ce test. (Puisque l'échantillon est de grande taille, utilisez simplement la distribution d'une loi normale présentée dans le tableau G.1.) Rejetez-vous H_0 au seuil de 5 % ? À celui de 1 % ?

iv. Diriez-vous que l'ampleur de la baisse que vous avez estimée est importante ? Commentez vos résultats sur les plans de la signification pratique (sachant que 1 once ≈ 29,57 ml) et de la significativité statistique.

v. Pour inférer un lien de causalité entre l'instauration de la taxe et la variation de consommation de spiritueux, vous devez faire une hypothèse implicite concernant les autres déterminants de la consommation de spiritueux au cours de cette période. Quelle est-elle ?

7. La nouvelle équipe de direction d'une boulangerie affirme que les salariés sont plus productifs aujourd'hui qu'ils ne l'étaient sous l'équipe dirigeante précédente, ce qui justifierait l'augmentation des salaires. Soit W_i^a le salaire du salarié i sous l'ancienne équipe de direction et W_i^n son salaire sous la nouvelle équipe. La différence est notée $D_i \equiv W_i^n - W_i^a$. Supposez que les D_i constituent un échantillon aléatoire tiré d'une distribution normale (μ, σ^2).

i. En utilisant les données disponibles pour 15 salariés et reprises dans le tableau ci-dessous, construisez un intervalle de confiance exact à 95 % pour μ.

ii. Écrivez l'hypothèse nulle selon laquelle il n'y a pas eu de changement dans le salaire moyen. Plus précisément, que vaut $E(D_i)$ sous H_0 ? Si vous devez évaluer la validité de l'affirmation de la nouvelle équipe de direction, quelle hypothèse alternative allez-vous utiliser, sachant que $\mu = E(D_i)$?

iii. Testez l'hypothèse nulle définie dans la question (ii) contre l'hypothèse alternative proposée, aux seuils de 5 % et de 1 %.

iv. Calculez la p-valeur du test de la question (iii).

Salarié	Ancien salaire	Nouveau salaire
1	8,30	9,25
2	9,40	9,00
3	9,00	9,25
4	10,50	10,00
5	11,40	12,00
6	8,75	9,50
7	10,00	10,25
8	9,50	9,50
9	10,80	11,50
10	12,55	13,10
11	12,00	11,50
12	8,65	9,00
13	7,75	7,75
14	11,25	11,50
15	12,65	13,00

8. Dans son édition du 5 février 1990, le *New York Times* a publié les performances des dix meilleurs joueurs de basket-ball de la NBA aux tirs à trois points. Le tableau suivant résume les données :

Joueur	NTT-NTR
Mark Price	429-188
Trent Tucker	833-345
Dale Ellis	1 149-472
Craig Hodges	1 016-396
Danny Ainge	1 051-406
Byron Scott	676-260
Reggie Miller	416-159
Larry Bird	1 206-455
Jon Sundvold	440-166
Brian Taylor	417-157

Remarque : NTT = nombre de tentatives de tir (FGA = *field goals attempted*) et NTR = nombre de tirs réussis (FGM = *field goal made*)

Pour un joueur i, le résultat d'un tir peut être modélisé par une variable qui suit une loi de Bernoulli : si Y_i est le résultat du tir i, alors $Y_i = 1$ lorsque le tir est réussi et $Y_i = 0$ lorsque le tir est raté. Soit θ la probabilité de réussir une tentative de tir à trois points. Un estimateur naturel pour θ est $\overline{Y} = NTR / NTT$.

i. Estimez θ pour le joueur Mark Price.

ii. Donnez l'écart-type de l'estimateur $\overline{Y}$ de θ et le nombre de tentative de tirs, n.

iii. La variable $(\overline{Y} - \theta)/\sigma(\overline{Y})$ suit asymptotiquement une loi normale centrée réduite, avec $\sigma(\overline{Y}) = \sqrt{\overline{Y}(1-\overline{Y})/n}$. Utilisez cette formule pour tester $H_0 : \theta = 0{,}5$ contre $H_1 : \theta < 0{,}5$ pour Mark Price. Utilisez un seuil de significativité de 1 %.

9. Un dictateur militaire met en place un plébiscite sous la forme d'un vote de confiance qui consiste à approuver ou non la politique qu'il suit. Il affirme être soutenu par 65 % des votants. Un groupe de défense des droits de l'Homme le suspecte d'avoir truqué les élections et vous embauche pour tester la validité des résultats du plébiscite. Le budget dont vous disposez vous permet d'interroger un échantillon aléatoire de 200 votants dans le pays.

i. Soit X le nombre de votes « oui » obtenus au sein de cet échantillon aléatoire tiré de l'ensemble de la population des votants. Quelle serait la valeur espérée de X s'il y avait effectivement 65 % de l'ensemble de la population qui soutient le dictateur ?

ii. Quel serait l'écart-type de X en faisant toujours l'hypothèse que la proportion réelle de votes « oui » est égale à 0,65 ?

iii. Après avoir constitué votre échantillon de 200 votants, vous trouvez qu'il y a effectivement 115 personnes qui ont voté « oui ». En utilisant le théorème central limite, approximez la probabilité de trouver 115 votes, ou moins, en faveur du dictateur, dans un échantillon aléatoire de 200 votants lorsque 65 % de la population a effectivement voté « oui ».

iv. Comment expliqueriez-vous la pertinence du nombre calculé en (iii) à quelqu'un qui n'a aucune connaissance en statistiques ?

10. Avant qu'une grève ne mette prématurément fin à la saison de première ligue de baseball aux États-Unis, Tony Gwynn des San Diego Padres avait réussi à frapper la balle 165 fois sur 419, ce qui représente une moyenne de 0,394 balles frappées. On se demandait cette année-là si Gwynn allait franchir la barre des 0,400 balles frappées. Pour répondre à cette interrogation, vous devez analyser la probabilité pour Gwynn de réussir à frapper une balle en particulier, soit la probabilité θ. Considérez Y_i une variable suivant une loi de *Bernoulli*(θ), qui vaut 1 si Gwynn arrive à frapper une balle et 0 sinon. On obtient alors un échantillon aléatoire tiré d'une distribution de *Bernoulli*(θ), où θ est la probabilité de succès, avec $n = 419$.

Notre meilleure estimation ponctuelle de θ est la moyenne de balles frappées par Gwynn, qui correspond à sa probabilité de succès : $\bar{y} = 0{,}394$. En utilisant le fait que $\sigma(\bar{y}) = \sqrt{\bar{y}(1-\bar{y})/n}$, construisez une approximation de l'intervalle de confiance à 95 % de θ, en utilisant la distribution d'une loi normale centrée réduite. Diriez-vous que les chances pour Gwynn de passer la barre des 0,400 balles frappées étaient fortes ou faibles ? Expliquez.

11. Supposons qu'entre leur première et leur deuxième année d'études universitaires, 400 étudiants soient choisis au hasard et reçoivent une subvention universitaire pour acheter un nouvel ordinateur. y_i mesure la variation dans la note obtenue à l'examen final (*GPA*) entre la première et la deuxième année pour l'étudiant i. Si la variation moyenne est de $\bar{y} = 0{,}132$ avec un écart-type de $s = 1{,}27$, la variation moyenne de la note finale est-elle statistiquement supérieure à zéro ?

ANNEXE

D

NOTIONS DE CALCUL MATRICIEL

Traduction de Alain Durré

Cette annexe résume les principaux concepts de calcul matriciel, et inclut des notions de probabilités. Nous aurons besoin de tout cela pour étudier le modèle de régression multiple avec une notation matricielle dans l'annexe E.

D.1 DÉFINITION DE BASE

Définition D.1 (Matrice). Une **matrice** est un tableau rectangulaire de nombres. Plus précisément, une matrice $m \times n$ comporte m lignes et n colonnes. L'entier positif m est appelé « nombre de lignes »[1], et n est appelé « nombre de colonnes »[2].

Nous utilisons des lettres en caractères gras majuscules pour désigner les matrices. Une matrice $m \times n$ peut s'écrire sous la forme générique

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & & & & \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix},$$

où a_{ij} représente l'élément de la i-ème ligne et de la j-ième colonne. Par exemple, a_{25} correspond au nombre de la deuxième ligne et de la cinquième colonne de **A**. Pour prendre un autre exemple, notons **A** une matrice 2×3 telle que

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \qquad \textbf{[D.1]}$$

où $a_{13} = 7$. L'abbréviation $\mathbf{A} = [a_{ij}]$ est souvent utilisée pour résumer certaines opérations matricielles.

Définition D.2 (Matrice Carrée). Une **matrice carrée** a le même nombre de lignes et de colonnes. La dimension d'une matrice carrée est ce nombre (de lignes et de colonnes).

Définition D.3 (Vecteur)

i. Une matrice $1 \times m$ est appelée un vecteur ligne (de dimension m) et peut s'écrire sous la forme $\mathbf{x} \equiv (x_1, x_2, \ldots, x_m)$.

ii. Une matrice $n \times 1$ est appelée un vecteur colonne, et peut s'écrire sous la forme

$$\mathbf{y} \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Définition D.4 (Matrice Diagonale). Une matrice carrée est appelée matrice diagonale lorsque tous les éléments en dehors de la diagonale principale sont nuls c'est à dire $a_{ij} = 0$ pour tout $i \neq j$. Une matrice diagonale est donc de la forme

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix}.$$

1 Note de traduction : hauteur est également souvent utilisé.
2 Note de traduction : largeur est également souvent utilisé.

Définition D.5 (Matrice Identité et Matrice Nulle)

i. La matrice identité $n \times n$, notée $\mathbf{I}$, ou parfois $\mathbf{I}_n$ pour indiquer sa dimension, est une matrice diagonale dans laquelle les éléments de la diagonale sont tous égaux à 1, et tous les éléments en dehors de la diagonale sont nuls :

$$\mathbf{I} \equiv \mathbf{I}_n \equiv \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

ii. La matrice nulle $m \times n$, notée $\mathbf{0}$, est une matrice $m \times n$ où tous les éléments sont nuls. La matrice nulle n'est pas nécessairement une matrice carrée.

D.2 OPÉRATIONS MATRICIELLES

Addition de Matrices

Deux matrices $\mathbf{A}$ et $\mathbf{B}$ de même dimensions $m \times n$ peuvent être additionnées : $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$. Plus précisément,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & & & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{bmatrix}.$$

Par exemple,

$$\begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -4 \\ 4 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 3 & -1 & 3 \\ 0 & 7 & 3 \end{bmatrix}.$$

Des matrices de dimensions différentes ne peuvent pas être additionnées.

Multiplication Scalaire

Pour tout nombre réel γ (aussi appelé « scalaire »), la multiplication scalaire est définie telle que $\gamma\mathbf{A} \equiv [\gamma a_{ij}]$, où

$$\gamma\mathbf{A} = \begin{bmatrix} \gamma a_{11} & \gamma a_{12} & \cdots & \gamma a_{1n} \\ \gamma a_{21} & \gamma a_{22} & \cdots & \gamma a_{2n} \\ \vdots & & & \\ \gamma a_{m1} & \gamma a_{m2} & \cdots & \gamma a_{mn} \end{bmatrix}.$$

Par exemple, si $\gamma = 2$ et $\mathbf{A}$ est la matrice de l'équation (D.1), alors

$$\gamma\mathbf{A} = \begin{bmatrix} 4 & -2 & 14 \\ -8 & 10 & 0 \end{bmatrix}.$$

Produit Matriciel

Pour multiplier une matrice **A** par une matrice **B** afin de former le produit **AB**, le nombre de colonnes de la matrice **A** doit être égale au nombre de lignes de la matrice **B**. Dans ce cas, si **A** est une matrice de dimension $m \times n$ et **B** une matrice de dimension $n \times p$, alors le produit matriciel **AB** se définit tel que

$$\mathbf{AB} = \left[\sum_{k=1}^{n} a_{ik} b_{kj}\right].$$

En d'autres termes, l'élément (i,j) de la nouvelle matrice **AB** est obtenu en multipliant chaque élément de la $i^{ème}$ ligne de **A** par l'élément correspondant de la $j^{ème}$ colonne de **B**, puis en additionnant ces n produits ensemble. Le schéma ci-dessous résume le fonctionnement du produit matriciel :

$$\begin{array}{ccccc} & \mathbf{A} & & \mathbf{B} & \mathbf{C} \\ i^{ème}\ ligne \longrightarrow & \begin{bmatrix} a_{i1}a_{i2}a_{i3}...a_{in} \end{bmatrix} & \begin{bmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \\ \vdots \\ b_{nj} \end{bmatrix} & = & \begin{bmatrix} \sum_{k=1}^{n} a_{ik}b_{kj} \end{bmatrix}, \\ & & \uparrow & & \uparrow \\ & & j^{ième}\ colonne & & (i,j)^{ième}\ élément \end{array}$$

où, en utilisant l'opérateur de sommation vu dans l'annexe A,

$$\sum_{k=1}^{n} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{in}b_{nj}.$$

Par exemple,

$$\begin{bmatrix} 2 & -1 & 0 \\ -4 & 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 & 6 & 0 \\ -1 & 2 & 0 & 1 \\ 3 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 12 & -1 \\ -1 & -2 & -24 & 1 \end{bmatrix}.$$

Il est également possible de multiplier une matrice par un vecteur. Si **A** est une matrice $n \times m$ et **y** est un vecteur $m \times 1$, alors **Ay** est un vecteur $n \times 1$. Si **x** est un vecteur $1 \times n$, alors **xA** est une vecteur $1 \times m$.

Les additions, les produits scalaires ou les multiplications de matrices peuvent être combinés de multiples façons, et les opérations matricielles suivent un certain nombre de règles semblables aux opérations avec des nombres. Considérons trois matrices **A**, **B**, et **C**, de dimensions appropriées pour réaliser les différentes opérations, et α et β deux nombres réels. La plupart de ces propriétés sont faciles à illustrer à partir des définitions précédentes.

Propriétés du produit matriciel. (1) $(\alpha + \beta)\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$; (2) $\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$; (3) $(\alpha\beta)\mathbf{A} = \alpha(\beta\mathbf{A})$; (4) $\alpha(\mathbf{AB}) = (\alpha\mathbf{A})\mathbf{B}$; (5) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$; (6) $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$; (7) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$; (8) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$; (9) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$; (10) $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$; (11) $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$; (12) $\mathbf{A} - \mathbf{A} = \mathbf{0}$; (13) $\mathbf{A0} = \mathbf{0A} = \mathbf{0}$; et (14) $\mathbf{AB} \neq \mathbf{BA}$.

Cette dernière propriété mérite d'être commentée plus en détail. Si **A** est une matrice $n \times m$ et **B** une matrice $m \times p$, alors **AB** est défini ; mais **BA** n'est défini que si $n = p$ (le nombre de ligne de **A** est

égal au nombre de colonne de **B**). Si **A** est une matrice $m \times n$ et **B** une matrice $n \times m$, alors **AB** et **BA** sont toutes deux définies, mais, **AB** et **BA** sont différentes. En effet, les deux nouvelles matrices n'ont pas la même dimension (sauf si **A** et **B** sont des matrices carrées) : **AB** est une matrice $m \times m$ et **BA** est une matrice $n \times n$. Même lorsque **A** et **B** sont deux matrices carrées, $\mathbf{AB} \neq \mathbf{BA}$ (sauf dans quelques cas particuliers).

Matrice Transposée

Définition D.6 (Transposée). Soit $\mathbf{A} = [a_{ij}]$ une matrice $m \times n$. La **transposée** de **A**, notée $\mathbf{A}'$ (que l'on prononce **A** *prime*), est la matrice $n \times m$ obtenue en interchangeant les lignes et les colonnes de **A**. On peut écrire $\mathbf{A}' \equiv [a_{ji}]$.

Par exemple,

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \quad \mathbf{A}' = \begin{bmatrix} 2 & -4 \\ -1 & 5 \\ 7 & 0 \end{bmatrix}.$$

Propriétés des transposées. (1) $(\mathbf{A}')' = \mathbf{A}$; (2) $(\alpha\mathbf{A})' = \alpha\mathbf{A}'$ pout tout scalaire α ; (3) $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$; (4) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, où **A** est une matrice $m \times n$ et **B** est une matrice $n \times k$; (5) $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2$, où **x** est un vecteur $n \times 1$; (6) Si **A** est une matrice $n \times k$ dont les lignes sont des vecteurs ligne $1 \times k$ $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_n$, on peut donc écrire

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix},$$

alors $\mathbf{A}' = (\mathbf{a}_1' \; \mathbf{a}_2' \; \ldots \; \mathbf{a}_n')$.

Définition D.7 (Matrice symétrique). Une matrice carrée est appelée **matrice symétrique** si et seulement si $\mathbf{A}' = \mathbf{A}$.

Si **X** est une matrice $n \times k$, alors $\mathbf{X}'\mathbf{X}$ est toujours définie et c'est une matrice symétrique, comme on peut le montrer en appliquant la première et la quatrième propriétés des transposées (voir exercice 3).

Multiplication de matrices par blocs

Soit **A** une matrice $n \times k$ dont les lignes sont données par le vecteur $1 \times k$ $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_n$, et **B** une matrice $n \times m$ dont les lignes sont données par le vecteur $1 \times m$ $\mathbf{b}_1$, $\mathbf{b}_2$, ..., $\mathbf{b}_n$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}.$$

Alors,

$$\mathbf{A}'\mathbf{B} = \sum_{i=1}^{n} \mathbf{a}_i' \mathbf{b}_i,$$

où pour tout i, $\mathbf{a}_i'\mathbf{b}_i$ est une matrice $k \times m$. Donc, $\mathbf{A}'\mathbf{B}$ peut être écrit comme la somme de n matrices, étant chacune de dimension $k \times m$. Un cas particulier de cette relation nous donne

$$\mathbf{A}'\mathbf{A} = \sum_{i=1}^{n} \mathbf{a}_i'\mathbf{a}_i,$$

où $\mathbf{a}_i'\mathbf{a}_i$ est une matrice $k \times k$ pour tout i.

Une forme plus générale de la multiplication de matrices par blocs est vérifiée quand les matrices $\mathbf{A}$ $(m \times n)$ et $\mathbf{B}$ $(n \times p)$ peuvent être écrites

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix},$$

où $\mathbf{A}_{11}$ est $m_1 \times n_1$, $\mathbf{A}_{12}$ est $m_1 \times n_2$, $\mathbf{A}_{21}$ est $m_2 \times n_1$, $\mathbf{A}_{22}$ est $m_2 \times n_2$, $\mathbf{B}_{11}$ est $n_1 \times p_1$, $\mathbf{B}_{12}$ est $n_1 \times p_2$, $\mathbf{B}_{21}$ est $n_2 \times p_1$ et $\mathbf{B}_{22}$ est $n_2 \times p_2$. Naturellement, $m_1 + m_2 = m$, $n_1 \times n_2 = n$, et $p_1 + p_2 = p$.

Quand nous écrivons le produit $\mathbf{AB}$, l'expression prend la même forme que si les entrées étaient des scalaires :

$$\mathbf{AB} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{pmatrix}.$$

Notez que chaque multiplication de matrices formant la partition sur la droite est bien définie car les dimensions des lignes et des colonnes sont compatibles pour la multiplication.

Trace

La trace d'une matrice est une opération très simple, qui ne fonctionne qu'avec les matrices carrées.

Définition D.8 (Trace). Pour toute matrice $\mathbf{A}$ $n \times n$ la **trace de la matrice A**, notée tr($\mathbf{A}$), est égale à la somme des éléments de sa diagonale. Mathématiquement,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

Propriété de la Trace. (1) tr($\mathbf{I}_n$) = n ; (2) tr($\mathbf{A}'$) = tr($\mathbf{A}$) ; (3) tr($\mathbf{A} + \mathbf{B}$) = tr($\mathbf{A}$) + tr($\mathbf{B}$) ; (4) tr($\alpha\mathbf{A}$) = αtr($\mathbf{A}$), pour tout scalaire α ; (5) tr($\mathbf{AB}$) = tr($\mathbf{BA}$), où $\mathbf{A}$ est une matrice $m \times n$ et $\mathbf{B}$ une matrice $n \times m$.

Matrice Inverse

La notion de matrice inverse est très importante pour les matrices carrées.

Définition D.9 (Matrice Inverse). Une matrice $n \times n$ $\mathbf{A}$ a une matrice **inverse**, notée $\mathbf{A}^{-1}$ quand $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$ et $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$ Dans ce cas, $\mathbf{A}$ est dite inversible ou non singulière. Dans le cas contraire, $\mathbf{A}$ est dite non-inversible ou singulière.

Propriétés des Matrices Inverses. (1) Si une matrice inverse existe, elle est unique ; (2) $(\alpha\mathbf{A})^{-1} = (1/\alpha)\mathbf{A}^{-1}$, si $\alpha \neq 0$ et $\mathbf{A}$ est inversible ; (3) $(\mathbf{AB})^{-1} = \mathbf{B}^{1}\mathbf{A}^{-1}$ si $\mathbf{A}$ et $\mathbf{B}$ sont toutes deux $n \times n$ et inversibles ; (4) $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

Nous ne verrons pas ici les détails permettant de calculer une matrice inverse. Tout livre d'algèbre matriciel contient des exemples détaillés des différents calculs.

D.3 INDÉPENDANCE LINÉAIRE ET RANG D'UNE MATRICE

Pour un ensemble de vecteurs de même dimension, il est important de savoir si un vecteur peut être exprimé comme une combinaison linéaire des autres vecteurs.

Définition D.10 (Indépendance Linéaire). Soit $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ un ensemble de vecteurs $n \times 1$. Ces vecteurs sont linéairement indépendants si et seulement

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \ldots + \alpha_r\mathbf{x}_r = 0 \qquad \textbf{[D.2]}$$

implique que $\alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$. Si (D.2) est vérifiée pour un ensemble de scalaires qui ne sont pas tous nuls, alors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ est *linéairement dépendant.*

L'affirmation selon laquelle $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ est linéairement dépendant revient à dire qu'au moins un vecteur dans l'ensemble peut être écrit comme une combinaison linéaire d'autres vecteurs de l'ensemble.

Définition D.11 (Rang)

i. Soit **A** une matrice $n \times m$. Le **rang de la matrice A**, que l'on note rang(**A**), est égal au nombre total des colonnes linéairement indépendantes de **A**.

ii. Si **A** est une matrice $n \times m$ et que rang(**A**) = m, alors **A** est une matrice de plein rang.

Si **A** est une matrice $n \times m$, son rang est toujours inférieur ou égal à m. Une matrice est dite de plein rang si ses colonnes forment un ensemble linéairement indépendant. Par exemple, la matrice 3×2 suivante

$$\begin{bmatrix} 1 & 3 \\ 2 & 6 \\ 0 & 0 \end{bmatrix}$$

peut avoir un rang maximum égal à 2. En réalité, le rang de cette matrice est égal à 1, car la seconde colonne égale à 3 fois la première colonne.

Propriété du rang. (1) rang(**A**′) = rang(**A**) ; (2) Si **A** est une matrice $n \times k$, alors rang(**A**) $\leq \min(n,k)$; (3) Si **A** est une matrice $k \times k$ et que rang(**A**) = k, alors **A** est inversible.

D.4 FORME QUADRATIQUE ET MATRICE DÉFINIE POSITIVE

Définition D.12 (Forme Quadratique). Soit **A** une matrice symétrique $n \times n$. La forme quadratique associée à la matrice A est la fonction réelle définie pour tout vecteurs x de dimension $n \times 1$, telle que :

$$f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j>1} a_{ij}x_ix_j.$$

Définition D.13 Matrice Positive Définie et Matrice Positive Semi-Définie

i. Une matrice symétrique **A** est dite **définie positive** si

$$\mathbf{x}'\mathbf{A}\mathbf{x} > 0 \text{ pour tout vecteur } n \times 1 \text{ sauf } \mathbf{x} = \mathbf{0}.$$

ii. Une matrice symétrique **A** est dite **semi-définie positive** si

$$\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0 \text{ pour tout vecteur } n \times 1.$$

Si une matrice est définie positive ou semi-définie positive, alors elle est toujours symétrique.

Propriétés des matrices définies positives et semi-définies positives. (1) Les éléments diagonaux d'une matrice définie positive sont strictement positifs, tandis que les éléments diagonaux d'une matrice semi-définie positive ne sont pas négatifs ; (2) Si **A** est une matrice définie positive, alors $\mathbf{A}^{-1}$ existe et est aussi définie positive ; (3) Si **X** est une matrice $n \times k$, alors $\mathbf{X}'\mathbf{X}$ et $\mathbf{X}\mathbf{X}'$ sont semi-définies positives ; (4) Si **X** est une matrice $n \times k$ de $\text{rang}(\mathbf{X}) = k$, alors $\mathbf{X}'\mathbf{X}$ est définie positive. (et donc non-singulière)

D.5 MATRICES IDEMPOTENTES

Définition D.14 (Matrice Idempotente). Soit **A** une matrice symétrique $n \times n$. **A** est une matrice idempotente si et seulement si, $\mathbf{A}\mathbf{A} = \mathbf{A}$.

Par exemple,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

est une matrice idempotente, comme il est possible de vérifier en multipliant **A** par **A**.

Propriétés des Matrices Idempotentes. Soit **A** une matrice idempotente $n \times n$. (1) $\text{rang}(\mathbf{A}) = \text{tr}(\mathbf{A})$, et (2) **A** une matrice semi-définie positive.

Nous pouvons construire des matrices facilement, en utilisant la formule suivante. Soit **X** une matrice $n \times k$ de rang $k = \text{rang}(\mathbf{X})$. Définissons alors

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I}_n - \mathbf{P}.$$

Alors **P** et **M** sont des matrices symétriques idempotentes, avec $\text{rang}(\mathbf{P}) = k$ et $\text{rang}(\mathbf{M}) = n - k$. Les rangs sont obtenus en utilisant la propriété 1 1 : $\text{tr}(\mathbf{P}) = \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}]$ (à partir de la propriété 5 sur les traces) $= \text{tr}(\mathbf{I}_k) = k$ (propriété 1 sur les traces). Il s'ensuit donc $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{P}) = n - k$.

D.6 DIFFÉRENTIATION DES FORMES LINÉAIRES ET QUADRATIQUES

Pour un vecteur $n \times 1$ donné **a**, considérons la fonction linéaire définie par

$$f(\mathbf{x}) = \mathbf{a}'\mathbf{x},$$

pour tout vecteur **x** $n \times 1$. La dérivée de f par rapport à **x** est un vecteur de dérivées partielles $1 \times n$, *tel que*

$$\partial f(\mathbf{x})/\partial \mathbf{x} = \mathbf{a}'.$$

Pour une matrice symétrique **A** $n \times n$, on peut écrire

$$g(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}.$$

Alors,

$$\partial g(\mathbf{x})/\partial \mathbf{x} = 2\mathbf{x}'\mathbf{A},$$

qui est donc un vecteur $1 \times n$.

D.7 MOMENT ET DISTRIBUTION DE VECTEURS ALÉATOIRES

Afin de dériver l'espérance et la variance des estimateurs des moindres carrés ordinaires (MCO) en utilisant des matrices, nous devons tout d'abord définir l'espérance et la variance d'une vecteur aléatoire. Comme son nom l'indique, un vecteur aléatoire est simplement un vecteur de variables aléatoires. Nous allons aussi définir la loi normale multivariée. Ces concepts sont simplement l'extension des concepts couverts dans l'annexe B.

Espérance

Définition D.15 (Espérance)

i. Si $\mathbf{y}$ est un vecteur aléatoire $n \times 1$, l'espérance **de y**, notée $E(\mathbf{y})$, est un vecteur d'espérance $E(\mathbf{y}) = [E(y_1), E(y_2), \ldots, E(y_n)]'$.

ii. Si $\mathbf{Z}$ est une matrice aléatoire $n \times m$, $E(\mathbf{Z})$ est une matrice $n \times m$ d'espérance : $E(\mathbf{Z}) = [E(z_{ij})]$.

Propriétés de l'Espérance. (1) Si $\mathbf{A}$ est une matrice $m \times n$ et $\mathbf{b}$ un vecteur $n \times 1$, tous deux non-aléatoires, alors $E(\mathbf{Ay} + \mathbf{b}) = \mathbf{A}E(\mathbf{y}) + \mathbf{b}$; (2) Si $\mathbf{A}$ est une matrice $p \times n$ et $\mathbf{B}$ une matrice $m \times k$ tous deux non-aléatoires, alors $E(\mathbf{AZB}) = \mathbf{A}E(\mathbf{Z})\mathbf{B}$.

Variance-Covariance des Matrices

Définition D.16 (Variance-Covariance des Matrices). Si $\mathbf{y}$ est un vecteur aléatoire $n \times 1$ alors sa **matrice de variance-covariance**, notée $\mathrm{Var}(\mathbf{y})$, est égale à ;

$$\mathrm{Var}(\mathbf{y}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & & & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix},$$

où $\sigma_j^2 = \mathrm{Var}(y_j)$ et $\sigma_{ij} = \mathrm{Cov}(y_i, y_j)$. En d'autres termes, les éléments diagonaux d'une matrice de variance-covariance correspondent à la variance de $\mathbf{y}$, et les éléments en dehors de la diagonale correspondent à sa covariance. Comme $\mathrm{Cov}(y_i,y_j) = \mathrm{Cov}(y_j,y_i)$, une matrice de variance-covariance est toujours symétrique.

Propriétés de la matrice de variance-covariance. (1) Si $\mathbf{a}$ est un vecteur non-aléatoire $n \times 1$ alors $\mathrm{Var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'[\mathrm{Var}(\mathbf{y})]\mathbf{a} \geq 0$; (2) Si $\mathrm{Var}(\mathbf{a}'\mathbf{y}) > 0$ pour tout $\mathbf{a} \neq \mathbf{0}$ $\mathrm{Var}(\mathbf{y})$ est définie positive ; (3) $\mathrm{Var}(\mathbf{y}) = E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})']$ où $\boldsymbol{\mu} = E(\mathbf{y})$; (4) Si les éléments de $\mathbf{y}$ ne sont pas corrélés, $\mathrm{Var}(\mathbf{y})$ est une matrice diagonale. Si de plus, $\mathrm{Var}(y_j) = \sigma^2$ pour $j = 1, 2, \ldots, n$, alors $\mathrm{Var}(\mathbf{y}_j) = \sigma^2\mathbf{I}_n$; (5) Si $\mathbf{A}$ est une matrice non-aléatoire $m \times n$ et $\mathbf{b}$ un vecteur non-aléatoire $n \times 1$ alors, $\mathrm{Var}(\mathbf{Ay} + \mathbf{b}) = \mathbf{A}[\mathrm{Var}(\mathbf{y})]\mathbf{A}'$.

Loi Normale Multivariée

La distribution normale d'une variable aléatoire a été largement étudiée dans l'annexe B. Nous devons l'étendre à la distribution des vecteurs aléatoires. Nous ne donnerons pas ici de définition précise concernant cette loi de probabilité, car nous n'en avons pas besoin. Il est important de savoir qu'un vecteur aléatoire normal multivarié est entièrement caractérisé par sa moyenne et sa matrice de variance-covariance.

Par conséquent, si **y** est un vecteur aléatoire normal multivarié $n \times 1$ de moyenne μ et avec une matrice de variance-covariance Σ, alors nous écrivons que y suit une loi normale (μ, Σ) Nous allons maintenant énoncer différentes propriétés utiles de la **distribution normale mutlivariée**.

Propriétés de la Loi Normale Multivariée. (1) Si y~Normale (μ, Σ), alors **y** est normalement distribuée ; (2) Si y~Normale (μ, Σ), y_i et y_j, deux éléments quelconque de **y**, sont indépendants si et seulement si ils ne sont pas corrélés, c'est-à-dire si, $\sigma_{ij} = 0$; (3) Si y~Normale (μ, Σ), alors Ay+b~Normale$(A\mu + b, A\Sigma A')$ où **A** et **b** ne sont pas aléatoires ; (4) Si y~Normale $(0, \Sigma)$ alors, pour des matrices non-aléatoires **A** et **B**, **Ay** et **By** sont indépendants si et seulement si $\mathbf{A}\Sigma\mathbf{B}' = \mathbf{0}$. En particulier, si $\Sigma = \sigma^2\mathbf{I}_n$, alors $\mathbf{AB}' = \mathbf{0}$ est nécessaire et suffisant pour que **Ay** et **By** soient indépendants ; (5) Si y~Normale $(0, \sigma^2 I_n)$ **A** une matrice non-aléatoire $k \times n$, et **B** une matrice symétrique idempotente $n \times n$ alors **Ay** et **y′By** sont indépendants si et seulement si $\mathbf{AB} = \mathbf{0}$; (6) Si y~Normale $(0, \sigma^2 I_n)$ et pour deux matrices symétriques idempotentes **A** et **B** alors **y′Ay** et **y′By** sont indépendants si et seulement si $\mathbf{AB} = \mathbf{0}$.

Loi du Khi-deux

Dans l'annexe B, nous avions défini une variable aléatoire suivant une loi du Khi-deux, comme la somme du carré de variables aléatoires normales indépendantes. En notation vectorielle, si μ~Normale $(0, I_n)$ alors $\mathbf{u}'\mathbf{u} \sim \chi^2_n$.

Propriétés de la Loi du Khi-deux. (1) Si μ~Normale $(0, I_n)$ et **A** est une matrice symétrique idempotente $n \times n$ avec rang(**A**) = q, alors $\mathbf{u}'\mathbf{A}\mathbf{u} \sim \chi^2_q$. (2) Si μ~Normale $(0, I_n)$ et pour **A** et **B** des matrices symétriques idempotentes $n \times n$ définies telle que $\mathbf{AB} = \mathbf{0}$, alors $\mathbf{u}'\mathbf{A}\mathbf{u}$ et $\mathbf{u}'\mathbf{B}\mathbf{u}$ sont des variables indépendantes suivant une loi du Khi-Deux ; et (3) Si z~Normale $(0, C)$ où **C** est une matrice non-singulière $m \times m$ alors $\mathbf{z}'\mathbf{C}^{-1}\mathbf{z} \sim \chi^2_m$.

Loi de Student

Nous avons aussi défini la loi de Student dans l'annexe B. Maintenant, nous allons ajouter une propriété importante.

Propriétés de la Loi de Student. Si μ~Normale $(0, I_n)$, **c** un vecteur non-aléatoire $n \times 1$, **A** une matrice non-aléatoire symétrique idempotente de rang q, et que $\mathbf{Ac} = \mathbf{0}$, alors $\{\mathbf{c}'\mathbf{u}/(\mathbf{c}'\mathbf{c})^{1/2}\}/(\mathbf{u}'\mathbf{A}\mathbf{u}/q)^{1/2} \sim t_q$.

Loi de Fisher

Rappelons-nous qu'une variable aléatoire distribuée selon la loi de Fisher peut être construite comme le quotient de deux variables aléatoires indépendantes distribuées chacune selon une loi du Khi-deux, avec un ajustement dépendant du nombre de degrés de liberté.

Propriétés de la loi de Fisher. Si μ~Normale $(0, I_n)$, avec **A** et **B** des matrices non-aléatoires symétriques idempotentes $n \times n$ avec rang(**A**) = k_1, rang(**B**) = k_2, et que $\mathbf{AB} = \mathbf{0}$, alors $(\mathbf{u}'\mathbf{A}\mathbf{u}/k_1)/(\mathbf{u}'\mathbf{B}\mathbf{u}/k_2) \sim F_{k1,\,k2}$.

RÉSUMÉ

Cette annexe contient un résumé de l'information basique nécessaire pour étudier le modèle linéaire classique en utilisant des matrices. Cette annexe est principalement destinée aux lecteurs familiers avec l'algèbre matriciel et les statistiques multivariées, et qui souhaitent revoir certaines notions qui seront très utilisées dans l'annexe E.

MOTS-CLÉS

Distribution normale multivariée p. 910
Espérance p. 909
Forme quadratique p. 907
Inverse d'une matrice p. 906
Loi de Student p. 910
Matrice p. 902
Matrice carrée p. 902
Matrice définie positive p. 907
Matrice de variance-covariance p. 909
Matrice diagonale p. 902
Matrice idempotente p. 908
Matrice identité p. 903
Matrice nulle p. 903
Matrice semi-définie positive p. 907
Matrice symétrique p. 905
Multiplication scalaire p. 903
Produit matriciel p. 904
Rang d'une matrice p. 907
Trace d'une matrice p. 906
Transposée p. 905
Vecteur aléatoire p. 909
Variable aléatoire suivant une loi de Fisher p. 912
Variable aléatoire suivant une loi du Khi-carré p. 910
Vecteur colonne p. 902
Vecteur ligne p. 902
Vecteur linéairement indépendant p. 907

PROBLÈMES

1. i. Calculez la matrice **AB** en utilisant

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 1 & 6 \\ 1 & 8 & 0 \\ 3 & 0 & 0 \end{bmatrix}.$$

ii. La matrice **BA** existe-t-elle ?

2. Si **A** et **B** sont des matrices diagonales $n \times n$, montrez que $\mathbf{AB} = \mathbf{BA}$.

3. Soit **X** une matrice quelconque $n \times k$. Montrez que $\mathbf{X'X}$ est toujours une matrice symétrique.

4. i. En utilisant les propriétés de la trace d'une matrice, montrez que $\text{tr}(\mathbf{A'A}) = \text{tr}(\mathbf{AA'})$ pour toute matrice **A** $n \times m$.

ii. Pour $\mathbf{A} = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 3 & 0 \end{bmatrix}$, vérifiez que $\text{tr}(\mathbf{A'A}) = \text{tr}(\mathbf{AA'})$.

5. i. En utilisant la définition de l'inverse d'une matrice, montrez que si **A** et **B** sont deux matrices non singulières $n \times n$, alors $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

ii. Si **A**, **B**, et **C** sont toutes des matrices non singulières $n \times n$, exprimez $(\mathbf{ABC}^{-1})$ en fonction de $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$ et $\mathbf{C}^{-1}$.

6. i. Montrez que si **A** est une matrice symétrique définie positive $n \times n$, alors les éléments diagonaux de **A** sont strictement positifs.

ii. Trouvez une matrice symétrique 2×2 dont les éléments diagonaux sont strictement positifs mais qui ne soit pas une matrice définie positive.

7. Soit **A** une matrice symétrique, définie positive $n \times n$. Montrez que pour toute matrice **P** $n \times n$ non singulière, alors **P′AP** est une matrice définie positive.

8. En utilisant la propriété 3, démontrez la propriété 5 de la variance des vecteurs.

9. Soit **a** un vecteur non-aléatoire $n \times 1$ et **u** un vecteur aléatoire $n \times 1$ avec $\mathrm{E}(\mathbf{uu}') = \mathbf{I}_n$. Montrez que $\mathrm{E}[\mathrm{tr}(\mathbf{auu}'\mathbf{a}')] = \sum_{i=1}^{n} a_i^2$.

10. Considérez les propriétés de la loi du Khi-deux décrites dans cette annexe. Combinez-les avec la définition d'une variable aléatoire suivant une loi de Fisher, et retrouvez les propriétés de la loi de Fisher concernant les quotients de formes quadratiques.

11. Soit **X** une $n \times k$ matrice par blocs tel que

$$\mathbf{X} = (\mathbf{X}_1 \ \mathbf{X}_2),$$

où $\mathbf{X}_1$ est $n \times k_1$ et $\mathbf{X}_2$ est $n \times k_2$.

(i) Montrez que

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix},$$

Quelles sont les dimensions des matrices ?

(ii) Soit **b** un $k \times 1$, vecteur par bloc tel que

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix},$$

où $\mathbf{b}_1$ est $k_1 \times 1$ et $\mathbf{b}_2$ est $k_2 \times 1$. Montrez que

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)\mathbf{b}_1 + (\mathbf{X}_1'\mathbf{X}_2)\mathbf{b}_2 \\ (\mathbf{X}_2'\mathbf{X}_1)\mathbf{b}_1 + (\mathbf{X}_2'\mathbf{X}_2)\mathbf{b}_2 \end{pmatrix}.$$

ANNEXE

E

LE MODÈLE DE RÉGRESSION LINÉAIRE SOUS FORME MATRICIELLE

Traduction de Jean-Yves Gnabo

Dans cette annexe, nous retrouvons les résultats liés à l'estimation du modèle de régression linéaire multiple par la méthode des moindres carrés ordinaires en utilisant des notations matricielles et des outils d'algèbre linéaire (voir l'annexe D pour un résumé). Les éléments présentés ci-dessous sont plus avancés que ceux développés dans le corps du texte.

E.1 PRÉSENTATION DU MODÈLE ET DE L'ESTIMATION PAR LES MOINDRES CARRÉS ORDINAIRES

Tout au long de cette annexe, nous désignons par l'indice t une observation issue de l'échantillon de taille n. Commençons par écrire le modèle de régression linéaire multiple composé de k paramètres comme suit :

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t,\ t = 1, 2, \ldots, n, \qquad \textbf{[E.1]}$$

où y_t est la valeur de la variable dépendante pour l'observation t, x_{tj}, $j = 1, 2, \ldots, k$, désignant les variables indépendantes. Comme d'habitude, β_0 désigne la constante et $\beta_1, \ldots, \beta_k$ les paramètres de pente du modèle.

Pour chaque t, on définit un vecteur de dimension $1 \times (k + 1)$, tel que $\mathbf{x}_t = (1, x_{t1}, \ldots, x_{tk})$, avec $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ le vecteur de taille $(k + 1) \times 1$, comprenant l'ensemble des paramètres du modèle. On peut dès lors réécrire le modèle (E.1) comme suit :

$$y_t = \mathbf{x}_t \boldsymbol{\beta} + u_t,\ t = 1, 2, \ldots, n. \qquad \textbf{[E.2]}$$

[À noter que certains auteurs préfèrent définir $\mathbf{x}_t$ comme un vecteur colonne auquel cas $\mathbf{x}_t$ est remplacé par $\mathbf{x}'_t$ dans (E.2). Mathématiquement, il semble cependant plus logique de le définir comme un vecteur ligne.] Nous pouvons maintenant écrire (E.2) entièrement sous forme matricielle en définissant des vecteurs et des matrices permettant de représenter les données de manière adéquate. Soit $\mathbf{y}$ le vecteur de taille $n \times 1$ contenant les observations de y : le $t^{\text{ème}}$ élément de $\mathbf{y}$ correspond à y_t. De même, on note $\mathbf{X}$, la matrice de dimension $n \times (k + 1)$ contenant les observations relatives aux variables explicatives. En d'autres termes, la $t^{\text{ème}}$ ligne de $\mathbf{X}$ correspond au vecteur $\mathbf{x}_t$. De façon détaillée, on a :

$$\underset{n\times(k+1)}{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{X}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{12} & \ldots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

Enfin, on pose $\mathbf{u}$ le vecteur de taille $n \times 1$ des erreurs non observées ou des perturbations. À partir de ces éléments, il est maintenant possible d'écrire (E.2) pour toutes les n observations sous **forme matricielle** comme suit :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \qquad \textbf{[E.3]}$$

Rappelez-vous que comme $\mathbf{X}$ est de dimension $n \times (k + 1)$ et $\boldsymbol{\beta}$ de dimension $(k + 1) \times 1$, il s'en suit que $\mathbf{X}\boldsymbol{\beta}$ est de dimension $n \times 1$.

L'estimation du paramètre inconnu $\boldsymbol{\beta}$ s'obtient en minimisant la somme des carrés des résidus tel que décrit dans la section 3.2. On définit la fonction « somme des carrés des résidus » pour n'importe quelle valeur du vecteur de taille $(k + 1) \times 1$ de paramètres $\mathbf{b}$, comme ceci :

$$\text{SCR}(\mathbf{b}) \equiv \sum_{i=1}^{n} (y_t - \mathbf{x}_t\mathbf{b})^2$$

Le vecteur de taille $(k + 1) \times 1$ des estimateurs des moindres carrés ordinaires, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)'$, est constitué des valeurs de paramètres qui minimisent SCR(**b**) parmi toutes les valeurs possibles de **b**. Nous sommes confrontés à un problème de calcul dans un cadre multivarié. Pour obtenir la valeur du vecteur de paramètres $\hat{\boldsymbol{\beta}}$ qui minimise la somme des carrés des résidus, nous devons résoudre les **conditions du premier ordre** :

$$\partial \text{SCR}(\hat{\boldsymbol{\beta}}) / \partial \mathbf{b} \equiv 0. \qquad \textbf{[E.4]}$$

Étant donné que la dérivée $(y_t - \mathbf{x}_t\mathbf{b})^2$ par rapport à **b** est égale au vecteur de taille $1 \times (k + 1)$ $-2(y_t - \mathbf{x}_t\mathbf{b})\mathbf{x}_t$, l'équation (E.4) peut se réécrire comme suit :

$$\sum_{i=1}^{n} \mathbf{x}_t'(y_t - \mathbf{x}_t\hat{\boldsymbol{\beta}}) \equiv 0. \qquad \textbf{[E.5]}$$

(pour arriver à ce résultat nous avons simplifié l'expression en la divisant par 2 et en prenant la transposée.) On peut écrire les conditions du premier ordre comme suit :

$$\sum_{t=1}^{n} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \ldots - \hat{\beta}_k x_{tk}) = 0$$

$$\sum_{t=1}^{n} x_{t1}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \ldots - \hat{\beta}_k x_{tk}) = 0$$

$$\vdots$$

$$\sum_{t=1}^{n} x_{tk}(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \ldots - \hat{\beta}_k x_{tk}) = 0$$

ce qui revient aux conditions du premier ordre décrites à l'équation (3.13). Nous souhaitons maintenant écrire ces différentes équations sous forme matricielle pour faciliter leur manipulation. En ayant recours à la formule de la multiplication des matrices par blocs présentée en annexe D, on peut voir que (E.5) est équivalent à :

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \qquad \textbf{[E.6]}$$

soit :

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \qquad \textbf{[E.7]}$$

On peut montrer que (E.7) a toujours au moins une solution. Les cas d'éventuelles solutions multiples ne sont pas souhaitables dans la mesure où l'on cherche à identifier un unique ensemble de paramètres estimés par les MCO compte tenu des données à disposition. Sous l'hypothèse que la matrice $\mathbf{X}'\mathbf{X}$, de dimension $(k + 1) \times (k + 1)$, est non singulière, nous pouvons pré-multiplier des deux cotés de l'équation (E.7) par $(\mathbf{X}'\mathbf{X})^{-1}$ afin d'obtenir l'expression de l'estimateur des MCO, $\hat{\boldsymbol{\beta}}$, comme solution du problème :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \qquad \textbf{[E.8]}$$

Cette expression est centrale pour l'analyse matricielle des modèles de régression linéaire multiple. L'hypothèse selon laquelle $\mathbf{X}'\mathbf{X}$ est inversible est équivalente à celle relative au rang de X, rang($\mathbf{X}$) = $(k + 1)$, qui signifie que les colonnes de $\mathbf{X}$ doivent être linéairement indépendantes. Cette hypothèse représente la version matricielle de l'hypothèse RLM.3 du chapitre 3.

Avant d'aller plus loin, l'équation (E.8) mérite que l'on s'y attarde un peu, ce afin de rappeler la nécessité de prendre certaines précautions. En effet, il peut être tentant de simplifier la formule de $\hat{\boldsymbol{\beta}}$ de la manière suivante :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}\mathbf{y}.$$

Le problème dans ce raisonnement est que $\mathbf{X}$ n'est généralement pas une matrice carrée, ce qui fait qu'elle ne peut donc pas être inversée. En d'autres termes, nous ne pouvons pas écrire $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{X}'^{-1}(\mathbf{X})^{-1}$ à moins que $n = (k + 1)$. En pratique, cependant, ce dernier cas de figure ne se rencontre quasiment jamais.

Les vecteurs des valeurs ajustées et des résidus obtenus par application des MCO, de taille $n \times 1$, sont calculés respectivement à partir de :

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}} = \mathbf{y} - \hat{\boldsymbol{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}},$$

En combinant (E.6) avec la définition de û, nous pouvons voir que les conditions du premier ordre pour $\hat{\boldsymbol{\beta}}$ s'écrivent de façon synthétique comme suit :

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}. \qquad \textbf{[E.9]}$$

Puisque la première colonne de $\mathbf{X}$ est composée entièrement de valeurs unitaires, (E.9) implique que la somme des résidus MCO est toujours égale à zéro lorsqu'une constante est introduite dans le modèle. En outre, la covariance entre chaque variable indépendante et les résidus des MCO dans l'échantillon est également égale à zéro. (Rappelons que ces deux propriétés ont déjà fait l'objet d'une discussion dans le chapitre 3.)

La somme des carrés des résidus peut être écrite comme suit :

$$\text{SCR} = \sum_{t=1}^{n}\hat{u}_t^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \qquad \textbf{[E.10]}$$

Toutes les propriétés algébriques du chapitre 3 peuvent être dérivées en utilisant les outils d'algèbre linéaire. Par exemple, nous pouvons montrer que la somme des carrés totaux est égale à la somme des carrés expliqués et des carrés des résidus [voir (3.27)]. Le recours au langage matriciel plutôt qu'aux outils de sommation ne facilite pas la dérivation de ce résultat, nous n'utilisons donc pas cette approche ici.

L'approche matricielle du modèle linéaire multiple peut en revanche s'avérer très utile comme point de départ à l'interprétation géométrique des modèles de régression. Celle-ci implique par ailleurs d'avoir recours à des concepts mathématiques encore plus avancés que ceux abordés dans l'annexe D. [voir Goldberger (1991) ou Greene (1997).]

Théorème de Frisch-Waugh

Dans la section 3-2, nous avions décrit une interprétation alternative des paramètres estimés par les MCO, soit β_1 mesurant la relation entre y et x_1 dans l'échantillon, après avoir purgé $x_1$de l'effet de x_2, dans le cadre d'un exemple à deux variables explicatives. Nous pouvons généraliser ce résultat à l'aide des notations matricielles. Partitionnons la matrice X de format $n \times (k + 1)$ comme suit :

$$X = (X_1 \mid X_2),$$

avec $\mathbf{X}_1$ une matrice de format $n \times (k_1 + 1)$ incluant la constante – bien que cela ne soit pas nécessaire pour aboutir au résultat d'intérêt – et $\mathbf{X}_2$ une matrice de format $n \times k_2$. Nous maintenons l'hypothèse que $\mathbf{X}$ est de plein rang $k + 1$, ce qui implique que $\mathbf{X}_1$ et $\mathbf{X}_2$ soient de plein rangs $k_1 + 1$ et k_2 resp.

Soient les paramètres estimés par la méthode des MCO $\hat{\beta}_1$et $\hat{\beta}_2$ issus de la régression (complète) de $\mathbf{y}$ sur $\mathbf{X}_1$et $\mathbf{X}_2$.

Comme nous le savons, les coefficients de la régression multiple (complète) sur $\mathbf{X}_2$, $\hat{\beta}_2$, diffèrent généralement de $\tilde{\beta}_2$ issus de la régression de $\mathbf{y}$ sur la seule matrice $\mathbf{X}_2$. Pour comprendre ces différences, rappelons que l'on peut obtenir $\hat{\beta}_2$ à partir d'une régression plus « resserrée », en ayant au préalable « purgé » l'influence de $\mathbf{X}_1$ sur $\mathbf{X}_2$. Considérons maintenant cette approche en deux temps :

(i) Régresser (chacune des colonnes) de $\mathbf{X}_2$ sur $\mathbf{X}_1$ et récupérer la matrice des résidus, que l'on note $\ddot{\mathbf{X}}_2$. Il est possible d'écrire $\ddot{\mathbf{X}}_2$ comme suit :

$$\ddot{\mathbf{X}}_2 = \left[\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\right]\mathbf{X}_2 = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2 = \mathbf{M}_1\mathbf{X}_2,$$

avec $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ et $\boldsymbol{M}_1 = \boldsymbol{I}_n - \mathbf{P}_1$ des matrices carrées d'ordre n, symétriques et idempotentes.

(ii) Régresser $\mathbf{y}$ sur $\ddot{\mathbf{X}}_2$ et nommer le vecteur de taille $k_2 \times 1$ des coefficients estimés $\ddot{\beta}_2$.

Le **théorème de Frisch-Waugh (FW)** stipule que :

$$\ddot{\beta}_2 = \hat{\beta}_2$$

Il est important de noter que le théorème de FW ne dit en général rien de l'égalité de l'estimateur issu de la régression multiple complète $\hat{\beta}_2$, et celui issu de la régression « courte », $\tilde{\beta}_2$. En général, $\hat{\beta}_2 \neq \tilde{\beta}_2$. Cependant, dans le cas où $\mathbf{X}_1'\mathbf{X}_2 = 0$ alors $\ddot{\mathbf{X}}_2 = \mathbf{M}_1\mathbf{X}_2 = \mathbf{X}_2$ ce qui conduit à $\ddot{\beta}_2 = \tilde{\beta}_2$ d'où l'on tire $\ddot{\beta}_2 = \hat{\beta}_2$ par application du théorème de FW. Il est en outre possible d'obtenir $\hat{\beta}_2$ en retirant l'influence de $\mathbf{X}_1$ sur $\mathbf{y}$. En d'autres termes, soit $\ddot{\mathbf{y}}$ le résidu issu de la régression de $\mathbf{y}$ sur $\mathbf{X}_1$, de sorte que :

$$\ddot{\mathbf{y}} = \mathbf{M}_1\mathbf{y}$$

Par suite, $\hat{\beta}_2$ s'obtient en régressant $\ddot{\mathbf{y}}$ sur $\ddot{\mathbf{X}}_2$. Il faut bien comprendre que cela n'est pas suffisant pour retirer l'influence de $\mathbf{X}_1$ sur $\mathbf{y}$. L'étape essentielle ici tient à « la purge » de l'effet de $\mathbf{X}_1$ sur $\mathbf{X}_2$. L'exercice 6 en fin de chapitre vous demande de dériver le théorème de FW et d'investiguer certaines questions reliées.

Un autre résultat d'intérêt concerne la régression de $\ddot{\mathbf{y}}$ sur $\ddot{\mathbf{X}}_2$ en prenant soin de conserver le vecteur des résidus, soit $\ddot{\mathbf{u}}$, qui s'avère équivalent aux résidus issus de la régression par les MCO de l'équation originale (complète) :

$$\ddot{\mathbf{y}} = \ddot{\mathbf{X}}_2\hat{\beta}_2 = \ddot{\mathbf{u}} = \hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}_1\hat{\beta}_1 - \mathbf{X}_2\hat{\beta}_2$$

Notons que nous avons eu recours ici au résultat de FW $\ddot{\beta}_2 = \hat{\beta}_2$. Nous ne récupérons pas les résidus originaux issus de la régression par les MCO en régressant $\mathbf{y}$ sur $\ddot{\mathbf{X}}_2$ (mais nous obtenons bien $\hat{\beta}_2$).

Historiquement, et avant le développement des capacités de calcul des ordinateurs que nous connaissons aujourd'hui, le théorème de Frisch-Waugh était utilisé à des fins computationelles. Aujourd'hui, ce résultat est plutôt d'intérêt théorique et s'avère particulièrement intéressant pour comprendre la mécanique des MCO. À titre d'exemple, souvenons-nous, que dans le Chapitre 10, nous avons eu recours à ce théorème pour établir que l'ajout d'une tendance temporelle dans une régression multiple est équivalent mathématiquement à exprimer en première étape toutes les variables en écart à leur tendance avant de procéder à la régression. Le théorème de FW peut également être utilisé dans le Chapitre 14 pour établir l'équivalence de l'estimateur du modèle à effets fixes qu'il soit obtenu à partir de la régression par MCO sur les données exprimées en écart à leur moyenne ou directement sur la régression multiple (complète) intégrant l'ensemble des variables catégorielles.

E.2 PROPRIÉTÉS DES MCO EN ÉCHANTILLON FINI

La dérivation de l'espérance et de la variance de l'estimateur des MCO, $\hat{\beta}$, est facilitée par le recours à l'algèbre linéaire. Cette démarche nécessite cependant de rester prudent dans la manière de poser les hypothèses.

Hypothèses E.1
Linéarité des paramètres

Le modèle peut être écrit comme dans (E.3), où y est un vecteur de taille $n \times 1$ comprenant les valeurs observées [de la variable dépendante], $\mathbf{X}$ une matrice de dimension $n \times (k + 1)$ comprenant les valeurs observées [des variables indépendantes], et $\mathbf{u}$ un vecteur de dimension $n \times 1$ constitué des erreurs ou perturbations non observées.

Hypothèse E.2
Absence de colinéarité parfaite

La matrice $\mathbf{X}$ est de plein rang $(k + 1)$.

Cette formulation permet d'écarter les cas de dépendances linéaires entre les variables explicatives. Sous l'hypothèse E.2, $\mathbf{X'X}$ est inversible, il s'en suit que $\hat{\beta}$ est unique et peut être défini comme dans l'équation (E.8).

Hypothèse E.3
Moyenne conditionnelle nulle

Compte tenu de l'ensemble des éléments de $\mathbf{X}$, chaque erreur u_t possède une moyenne nulle : $E(u_t|\mathbf{X}) = 0$, $t = 1, 2, \ldots, n$.

Sous forme vectorielle, l'hypothèse E.3 s'écrit comme suit :

$$E(\mathbf{u}|\mathbf{X}) = \mathbf{0}. \qquad \textbf{[E.11]}$$

Celle-ci découle directement de RLM.4 sous l'hypothèse d'un échantillonnage aléatoire, RLM.2. Dans les applications en séries temporelles, l'hypothèse E.3 impose l'exogénéité stricte des variables explicatives, une propriété discutée en détail au chapitre 10. Ce cas de figure exclut du modèle les variables explicatives dont les valeurs futures sont corrélées avec le terme d'erreur u_t. En particulier, cette hypothèse impose de ne pas inclure de variables dépendantes retardées. Muni de l'hypothèse E.3, il nous est alors possible de calculer la valeur espérée de $\hat{\beta}$ conditionnellement aux réalisations des x_{tj}.

Théorème E.1
Absence de biais des mco

Sous les hypothèses E.1, E.2 et E.3, l'estimateur des MCO, $\hat{\beta}$, de β est sans biais.

PREUVE : On a recours aux hypothèses E.1 et E.2 et aux règles d'algèbre simples pour pouvoir écrire :

$$\begin{aligned}\hat{\beta} &= (\mathbf{X'X})^{-1}\mathbf{X'y} = (\mathbf{X'X})^{-1}\mathbf{X'}\,(\mathbf{X}\beta + \mathbf{u}) \\ &= (\mathbf{X'X})^{-1}(\mathbf{X'X})\beta + (\mathbf{X'X})^{-1}\mathbf{X'u} = \beta + (\mathbf{X'X})^{-1}\mathbf{X'u}, \qquad \textbf{[E.12]}\end{aligned}$$

où on utilise le fait que $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}_{k+1}$. En prenant l'espérance conditionnelle de $\mathbf{X}$ on obtient :

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0} = \boldsymbol{\beta}, \end{aligned}$$

Comme $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ sous l'hypothèse E.3 et que ce résultat ne dépend pas de la valeur de $\boldsymbol{\beta}$, nous avons démontré l'absence de biais de $\hat{\boldsymbol{\beta}}$.

Afin d'obtenir la forme la plus simple possible de la matrice de variance-covariance de $\hat{\boldsymbol{\beta}}$, nous imposons les hypothèses d'homoscédasticité et d'absence de corrélation sérielle.

Hypothèse E.4
Homoscédasticité et absence de corrélation sérielle

i. $Var(u_t|\mathbf{X}) = \sigma^2$, t = 1, 2, ..., n.
ii. $Cov(u_t,u_s|\mathbf{X}) = 0$, pour tout $t \neq s$. Sous forme matricielle, nous pouvons écrire ces deux hypothèses comme suit :

$$Var(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I}n, \qquad \textbf{[E.13]}$$

où $\mathbf{I}_n$ est la matrice identité de dimension $n \times n$.

La partie (i) de l'hypothèse E.4 correspond à l'hypothèse d'homoscédasticité : la variance de u_t ne dépend d'aucun élément de $\mathbf{X}$, et la variance doit être constante pour toutes les observations t. La partie (ii) est liée à l'hypothèse d'absence de corrélation sérielle : les erreurs ne peuvent être corrélées entre elles. La partie (ii) de l'hypothèse E.4 est vérifiée sous l'hypothèse d'échantillonnage aléatoire, et dans n'importe quel autre contexte où les observations sont indépendantes. Dans le cadre des applications en séries temporelles, la partie (ii) exclut toute corrélation des erreurs au fil du temps (à la fois conditionnellement et inconditionnellement à $\mathbf{X}$).

En raison de (E.13), on dit souvent que $\mathbf{u}$ est caractérisé par une **matrice de variance-covariance sphérique** lorsque l'hypothèse E.4 est vérifiée. Nous pouvons maintenant calculer la **matrice de variance-covariance de l'estimateur des MCO**.

THÉORÈME E.2
Matrice de variance-covariance de l'estimateur des MCO

Muni des hypothèses E.1 à E.4,

$$Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \qquad \textbf{[E.14]}$$

PREUVE : À partir de la dernière formule de l'équation (E.12), nous avons :

$$Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) = Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[Var(\mathbf{u}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Nous utilisons maintenant l'équation E.4 afin d'obtenir :

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\,\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

La formule (E.14) signifie que la variance de $\hat{\beta}_j$ (conditionnellement à $\mathbf{X}$) est obtenue en multipliant σ^2 par le $j^{ème}$ élément de la diagonale de $(\mathbf{X}'\mathbf{X})^{-1}$. Pour les coefficients de pente, nous avons proposé une formule qui peut être facilement interprétée dans l'équation (3.51). L'équation (E.14) nous dit également comment obtenir la covariance entre deux paramètres estimés par les MCO, à savoir multiplier σ^2 par l'élément hors diagonale approprié de $(\mathbf{X}'\mathbf{X})^{-1}$. Dans le chapitre 4, nous avons montré comment éviter de devoir définir les termes de covariances pour obtenir des intervalles de confiance et des statistiques de tests d'hypothèses par le biais d'une réécriture appropriée du modèle.

On peut alors prouver le théorème de Gauss-Markov dans toute sa généralité.

Théorème E.3
Théorème de Gauss-Markov

Sous les hypothèses E.1 à E.4, $\hat{\boldsymbol{\beta}}$ est le meilleur estimateur linéaire sans biais.

PREUVE : Tout autre estimateur linéaire de $\boldsymbol{\beta}$ peut être écrit comme suit :

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{y}, \qquad \textbf{[E.15]}$$

où $\mathbf{A}$ est une matrice de dimension $n \times (k+1)$. Pour que $\tilde{\boldsymbol{\beta}}$ soit sans biais conditionnellement à $\mathbf{X}$, $\mathbf{A}$ peut être composée d'éléments non aléatoires fonctions de $\mathbf{X}$. (Par exemple, $\mathbf{A}$ ne peut pas être une fonction de $\mathbf{y}$.) Afin de mettre en évidence quelles sont les contraintes additionnelles qui doivent être imposées sur $\mathbf{A}$, écrivons :

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = (\mathbf{A}'\mathbf{X})\boldsymbol{\beta} + \mathbf{A}'\mathbf{u}. \qquad \textbf{[E.16]}$$

alors,

$$\begin{aligned} \mathrm{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathrm{E}(\mathbf{A}'\mathbf{u}|\mathbf{X}) \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\mathrm{E}(\mathbf{u}|\mathbf{X}) \text{ car } \mathbf{A} \text{ est une fonction de } \mathbf{X} \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} \text{ car } \mathrm{E}(\mathbf{u}|\mathbf{X}) = 0. \end{aligned}$$

Pour que $\tilde{\boldsymbol{\beta}}$ soit un estimateur sans biais de $\boldsymbol{\beta}$, l'égalité $\mathrm{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$ doit être vérifiée pour tous vecteurs $\boldsymbol{\beta}$ de taille $(k+1) \times 1$, c'est-à-dire :

$$\mathbf{A}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \text{ pour tous vecteurs } \boldsymbol{\beta} \text{ de dimension } (k+1) \times 1. \qquad \textbf{[E.17]}$$

Étant donné que $\mathbf{A}'\mathbf{X}$ est une matrice de dimension $(k+1) \times (k+1)$, (E.17) est vérifiée si et seulement si $\mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1}$. Les équations (E.15) et (E.17) caractérisent la classe des estimateurs linéaires sans biais de $\boldsymbol{\beta}$.

Ensuite, muni de l'hypothèse E.4., on peut écrire la relation suivante à partir de (E.16)

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{A}'[\mathrm{Var}(\mathbf{u}|\mathbf{X})]\mathbf{A} = \sigma^2\mathbf{A}'\mathbf{A},$$

Par conséquent,

$$\begin{aligned} \mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \sigma^2\,[\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2[\mathbf{A}'\mathbf{A} - \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}] \text{ car } \mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1} \\ &= \sigma^2\mathbf{A}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{A} \\ &\equiv \sigma^2\mathbf{A}'\mathbf{M}\mathbf{A}, \end{aligned}$$

où $\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Étant donné que $\mathbf{M}$ est symétrique et idempotente, $\mathbf{A}'\mathbf{M}\mathbf{A}$ est semi-définie positive pour toute matrice $\mathbf{A}$ de taille $n \times (k + 1)$. Cela prouve que l'estimateur des MCO $\hat{\boldsymbol{\beta}}$ est *BLUE*. Pourquoi est-ce important ? Soit $\mathbf{c}$ un vecteur $(k + 1) \times 1$. Considérons par ailleurs, le scalaire donné par la combinaison linéaire $\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + \ldots + c_k\beta_k$. Les estimateurs sans biais de $\mathbf{c}'\boldsymbol{\beta}$ sont $\mathbf{c}'\hat{\boldsymbol{\beta}}$ et $\mathbf{c}'\tilde{\boldsymbol{\beta}}$. Toutefois,

$$\text{Var}(\mathbf{c}'\boldsymbol{\beta}|\mathbf{X}) - \text{Var}(\mathbf{c}'\hat{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{c}'[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})]\mathbf{c} \geq 0,$$

car $[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})]$ est semi-définie positive. Par conséquent, lorsqu'ils sont appliqués pour estimer n'importe quelle combinaison linéaire de b, les MCO fournissent les estimateurs exhibant la plus petite variance. En particulier, $[\text{Var}(\hat{\beta}_j|\mathbf{X}) \leq \text{Var}(\tilde{\beta}|\mathbf{X})]$ pour tout autre estimateur linéaire sans biais de β_j.

L'estimateur sans biais de la variance de l'erreur σ^2 s'écrit :

$$\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} / (n - k - 1),$$

ceci est similaire à l'équation (3.56).

Théorème E.4
Absence de biais de σ^2

Sous les hypothèses E.1 à E.4, $\hat{\sigma}^2$ est sans biais pour σ^2 : $\text{E}(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$ pour tout $\sigma^2 > 0$.

PREUVE : Soit $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{u}$, où $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, la dernière égalité découle de $\mathbf{MX} = \mathbf{0}$. Comme $\mathbf{M}$ est symétrique et idempotente,

$$\hat{\mathbf{u}}\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u}.$$

Puisque $\mathbf{u}'\mathbf{M}\mathbf{u}$ est un scalaire, il est égal à sa trace. Par conséquent,

$$\begin{aligned}\text{E}(\mathbf{u}'\mathbf{M}\mathbf{u}|\mathbf{X}) &= \text{E}[\text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u})|\mathbf{X}] = \text{E}[\text{tr}(\mathbf{M}\mathbf{u}\mathbf{u}')|\mathbf{X}] \\ &= \text{tr}[\text{E}(\mathbf{M}\mathbf{u}\mathbf{u}'|\mathbf{X})] = \text{tr}[\mathbf{M}\text{E}(\mathbf{u}\mathbf{u}'|\mathbf{X})] \\ &= \text{tr}(\mathbf{M}\sigma^2\mathbf{I}_n) = \sigma^2\text{tr}(\mathbf{M}) = \sigma^2(n - k - 1).\end{aligned}$$

La dernière égalité découle de $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - tr[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = n - tr(\mathbf{I}_{k+1}) = n - (k + 1)$ $= n - k - 1$. Il s'en suit que :

$$\text{E}(\hat{\sigma}^2|\mathbf{X}) = \text{E}(\mathbf{u}'\mathbf{M}\mathbf{u}|\mathbf{X})/(\text{n} - \text{k} - 1) = \sigma^2.$$

E.3 INFÉRENCE STATISTIQUE

Muni de la dernière hypothèse relative au modèle linéaire classique, il est possible de montrer que $\hat{\boldsymbol{\beta}}$ suit une distribution normale multivariée. Les statistiques de test standards couvertes dans le chapitre 4 suivent des distributions t et F sous l'hypothèse nulle.

Hypothèse E.5
Normalité des Erreurs

Conditionnellement à $\mathbf{X}$, les erreurs u_t sont indépendantes et identiquement distribuées selon une loi Normale(0, σ^2). De manière équivalente, sachant les valeurs de $\mathbf{X}$, $\mathbf{u}$ est distribué selon une distribution normale multivariée de moyenne nulle et de matrice de variance-covariance $\sigma^2\mathbf{I}_n$: $\mathbf{u} \sim \text{Normale}(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

Sous l'hypothèse E.5, chaque u_t est indépendant des variables explicatives pour tout t. Dans un contexte de séries temporelles, ceci revient essentiellement à faire l'hypothèse d'exogénéité stricte.

Théorème E.5
Normalité de $\hat{\beta}$

Sous les hypothèses du modèle linéaire classique E.1 à E.5, le **théorème E.5 de Normalité de** $\hat{\beta}$ indique que $\hat{\beta}$ suit conditionnellement à $\mathbf{X}$ une loi normale multivariée de moyenne $\boldsymbol{\beta}$ et de matrice de variance-covariance $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$.

Le théorème E.5 est à la base de l'inférence statistique impliquant $\boldsymbol{\beta}$. En effet, muni des propriétés du chi-deux et des distributions t et F que nous avons rappelées à l'annexe D, nous pouvons recourir au théorème E.5 pour établir que les statistiques t et F suivent, sous l'hypothèse nulle, des distributions de Student et Fisher sous les hypothèses E.1 à E.5. Nous illustrons ce point en détaillant la preuve pour les statistiques t.

Théorème E.6
Distribution de la statistique t

Sous les hypothèses E.1 à E.5, $(\hat{\beta}_j - \beta_j)/\hat{\sigma}(\hat{\beta}_j) \sim t_{n-k-1}$, j = 0, 1, ..., k.

PREUVE : La démonstration de ce résultat nécessite de procéder en différentes étapes, l'ensemble des développements tenant conditionnellement à $\mathbf{X}$. Tout d'abord, par application du théorème E.5, $(\hat{\beta}_j - \beta_j)/\sigma(\hat{\beta}_j) \sim$ Normale(0,1), avec $\sigma(\hat{\beta}_j) = \sigma\sqrt{C_{jj}}$, et c_{jj} le $j^{\text{ème}}$ élément diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$. Par suite, sous les hypothèses E.1 à E.5, et conditionnellement à $\mathbf{X}$ on a :

$$(n - k - 1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}. \qquad \textbf{[E.18]}$$

puisque $(n - k - 1)\hat{\sigma}^2/\sigma^2 = (\mathbf{u}/\sigma)'\mathbf{M}(\mathbf{u}/\sigma)$, avec $\mathbf{M}$ la matrice $n \times n$ symétrique et idempotente définie dans le théorème E.4. De plus, comme $\mathbf{u}/\sigma \sim$ Normale(0, $\mathrm{I_n}$) par application de l'hypothèse E.5., il suit que, par application de la propriété 1 d'une distribution du chi-deux tel que stipulé en annexe D, $(\mathbf{u}/\sigma)'\mathbf{M}(\mathbf{u}/\sigma) \sim \mathrm{X}^2_{n-k-1}$ (puisque $\mathbf{M}$ est de rang $n - k - 1$).

Nous devons également montrer que $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}^2$ sont indépendants. Rappelons que $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ et $\sigma^2 = \mathbf{u}'\mathbf{M}\mathbf{u}/(n - k - 1)$. Or $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{M} = 0$ puisque $\mathbf{X}'\mathbf{M} = 0$. D'après la propriété 5 des distributions normales multivariées présentée en annexe D, il suit que, $\hat{\boldsymbol{\beta}}$ et $\mathbf{Mu}$ sont indépendants, et comme σ^2 est une fonction de $\mathbf{Mu}$, $\boldsymbol{\beta}$ et σ^2 le sont également.

$$(\hat{\beta}_j - \beta_j)/\hat{\sigma}(\hat{\beta}_j) = [(\hat{\beta} - \beta_j)/\sigma(\hat{\beta}_j)]/(\hat{\sigma}^2/\sigma^2)^{1/2},$$

soit le ratio d'une normale centrée réduite et de la racine carrée de $\mathrm{x}^2_{n-k-1}/(n - k - 1)$. Nous venons de montrer que ces deux composantes sont indépendantes, dès lors par définition, la statistique t, $(\hat{\beta}_j - \beta_j)/\hat{\sigma}(\hat{\beta}_j)$ suit une distribution t_{n-k-1}. Dans la mesure où cette distribution ne dépend pas de $\mathbf{X}$, il s'agit là également de la distribution non conditionnelle de $(\hat{\beta}_j - \beta_j)/\hat{\sigma}(\hat{\beta}_j)$.

À partir de ce théorème, nous pouvons tester n'importe quelle valeur supposée pour β_j à l'aide de la statistique t.

Sous les hypothèses E.1 à E.5, nous pouvons calculer ce qu'il est commun d'appeler la borne de *Cramer-Rao* pour la matrice de variance-covariance des estimateurs sans biais de $\boldsymbol{\beta}$ (conditionnellement à $\mathbf{X}$) [voir Greene (1997, chapitre 4)]. Il est possible de montrer que cette dernière correspond à $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, soit

exactement l'expression de la matrice de variance-covariance de l'estimateur des MCO. Ce résultat implique que $\hat{\boldsymbol{\beta}}$ est **l'estimateur sans biais de variance minimale de** $\boldsymbol{\beta}$ (conditionnellement à $\mathbf{X}$) : $\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ est une variance semi-définie positive pour tout autre estimateur sans biais de $\boldsymbol{\beta}$; nous n'avons dès lors plus à nous limiter à la classe des estimateurs linéaires en $\mathbf{y}$.

Il est aisé de montrer que l'estimateur des MCO correspond à l'estimateur du maximum de vraisemblance (MV) de $\boldsymbol{\beta}$ sous l'hypothèse E.5. Pour tout t, la distribution de y_t sachant $\mathbf{X}$ est Normale($\mathbf{x}_t\boldsymbol{\beta}$, σ^2). Puisque les y_t sont indépendants conditionnellement à $\mathbf{X}$, la fonction de vraisemblance de l'échantillon est obtenue par le produit des fonctions de densité :

$$\prod_{t=1}^{n}(2\pi\sigma^2)^{-1/2}\exp[-(y_t - \mathbf{x}_t\boldsymbol{\beta})^2 / (2\sigma^2)],$$

où Π désigne l'opérateur « produit ». Maximiser cette expression ou le logarithme népérien de cette expression, en fonction de $\boldsymbol{\beta}$ et σ^2, revient au même :

$$\sum_{t=1}^{n}\left[-\left(\frac{1}{2}\right)\log(2\pi\sigma^2) - (y_t - \mathbf{x}_t\boldsymbol{\beta})^2 / (2\sigma^2)\right].$$

Pour dériver l'expression de $\hat{\boldsymbol{\beta}}$, il suffit de minimiser $\sum_{t=1}^{n}(y_t - \mathbf{x}_t\boldsymbol{\beta})^2$ – la division de cette expression par $2\sigma^2$ n'affectant pas le résultat de l'optimisation – ce qui correspond exactement au problème dont l'estimateur des MCO est solution. L'estimateur de σ^2 que nous avons utilisé, soit $\text{SCR}/(n-k)$, n'est en revanche pas l'estimateur du MV de σ^2 ; celui-ci est en effet donné par SCR/n, et il est biaisé. Dans la mesure où les distributions exactes des statistiques t et F sous l'hypothèse nulle, sont contingentes à l'absence de biais pour l'estimateur de σ^2, l'estimateur des MCO est toujours préféré à celui du MV.

Le fait que l'estimateur des MCO corresponde à l'estimateur du MV sous l'hypothèse E.5 implique une propriété de robustesse intéressante de ce dernier du fait de la normalité. Le raisonnement est simple. Nous savons que l'estimateur des MCO est sans biais sous les hypothèses E.1 à E.3 ; la normalité des erreurs ou l'hypothèse E.A. n'étant à aucun moment requise pour la démonstration de ce résultat. Comme montré dans la section suivante, l'estimateur des MCO est convergent[1] en l'absence de l'hypothèse de normalité, sous condition que la loi des grands nombres puisse s'appliquer (comme c'est en général le cas). Les propriétés statistiques de l'estimateur des MCO impliquent que l'estimateur du MV, qui repose sur la normalité de la fonction de log-vraisemblance, est robuste quelque soit la distribution suivie par les variables aléatoires : la distribution peut être quelconque (à de rares exceptions près), mais nous serons toujours en mesure d'obtenir un estimateur convergent (et, sous les hypothèses E.1 à E.3, sans biais) des paramètres. Comme discuté dans la section 17.3, l'estimateur du maximum de vraisemblance dérivé en l'absence de l'hypothèse de distribution exacte des variables aléatoires est souvent qualifié d'**estimateur du quasi-maximum vraisemblance (EQMV)**.

En règle générale, la convergence de l'estimateur du MV repose sur l'expression de la distribution de probabilité exacte des variables aléatoires dans le but de conclure à la convergence vers la vraie valeur des paramètres. Nous venons de voir que la distribution normale est une exception notable à cette règle. D'autres distributions possèdent également cette propriété, notamment la distribution de Poisson – comme discuté dans la section 17.3. Wooldridge (2010, chapitre 18) présente d'autres exemples utiles.

1 NDT : À l'instar de la terminologie adoptée dans cet ouvrage, la notion de convergence désigne ici la propriété pour l'estimateur de converger en probabilité vers la vraie valeur du paramètre (« *consistancy* » en anglais).

E.4 QUELQUES ÉLÉMENTS D'ANALYSE ASYMPTOTIQUE

L'approche matricielle du modèle de régression linéaire multiple permet également de retrouver les propriétés asymptotiques des MCO de manière plus synthétique. Nous nous proposons maintenant de donner les éléments relatifs à la démonstration des résultats donnés au chapitre 11.

Nous commencerons par démontrer la convergence de l'estimateur des MCO donnée dans le théorème 11.1. Rappelez-vous que les hypothèses nécessaires à la validité de ce théorème contiennent notamment celles relatives à l'analyse en coupes transversales sous échantillonnage aléatoire.

Preuve du théorème 11.1. À l'instar des développements proposés dans le cadre de l'exercice E.1 du présent chapitre et sous les hypothèses SC.1′, nous pouvons écrire l'estimateur des MCO comme suit :

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'y_t\right) = \left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'(\mathbf{x}_t\boldsymbol{\beta}+u_t)\right)\\ &= \boldsymbol{\beta}+\left(\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(\sum_{t=1}^{n}\mathbf{x}_t'u_t\right)\\ &= \boldsymbol{\beta}+\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right). \qquad \textbf{[E.19]}\end{aligned}$$

Par application de la loi des grands nombres, il suit :

$$n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t \xrightarrow{p} \mathbf{A} \text{ et } n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'u_t \xrightarrow{p} \mathbf{0}, \qquad \textbf{[E.20]}$$

$\mathbf{A} = \mathrm{E}(\mathbf{x}_t'\mathbf{x}_t)$ étant une matrice non singulière de dimension $(k+1)\times(k+1)$ sous l'hypothèse SC.2′ et puisque $\mathrm{E}(\mathbf{x}_t'u_t) = 0$ par application de l'hypothèse SC.3′. Nous devons maintenant avoir une version matricielle de la propriété PLIM.1 de l'annexe C. Puisque $\mathbf{A}$ est non singulière, il suit :

$$\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1} \xrightarrow{p} \mathbf{A}^{-1} \qquad \textbf{[E.21]}$$

[Wooldridge (2010, chapitre 3) contient une discussion sur ce type de résultats de convergence.] À partir de (E.19), (E.20) et (E.21) nous pouvons écrire que :

$$\operatorname{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{A}^{-1}.\,\mathbf{0} = \boldsymbol{\beta}.$$

Ce qui complète la démonstration.

Maintenant, esquissons la démonstration du résultat de normalité asymptotique du théorème 11.2.

Preuve du théorème 11.2. À partir de l'équation (E.19), nous pouvons écrire :

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right)$$

$$= \mathbf{A}^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right) + o_p(\mathbf{1}). \qquad \textbf{[E.22]}$$

où le terme « $o_p(1)$ » est un reste qui converge en probabilité vers zéro. Ce terme est égal à $\left[\left(n^{-1}\sum_{t=1}^{n}\mathbf{x}_t'\mathbf{x}_t\right)^{-1} - \mathbf{A}^{-1}\right]\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right)$. Le terme entre parenthèses converge en probabilité vers zéro (pour la même raison que l'argument utilisé lors de la démonstration du théorème 11.1), alors que $\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right)$ est borné en probabilités puisqu'il converge vers une distribution normale multivariée d'après le théorème central limite. Un résultat bien connu en théorie asymptotique est que le produit de tels termes converge en probabilité vers zéro. De plus, $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ hérite de la distribution asymptotique de $\mathbf{A}^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\mathbf{x}_t'u_t\right)$. Voir Wooldridge (2010, chapitre 3) pour plus de détails sur les résultats de convergence utilisés dans cette démonstration.

Par application du théorème central limite, $n^{-1/2}\sum_{t=1}^{n}x_t'u_t$ suit asymptotiquement une distribution normale de moyenne nulle et de matrice de variance-covariance de dimension $(k + 1) \times (k + 1)$ que l'on note ici **B**. $\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ est donc asymptotiquement distribué selon une loi normale de moyenne nulle et de matrice de variance-covariance $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$. Nous montrons maintenant que sous les hypothèses SC.4′ et SC.5′ $\mathbf{B} = \sigma^2\mathbf{A}$. (L'expression générale est utile car elle met en lumière l'expression des écarts-types estimés des estimateurs des MCO robustes à la présence d'hétéroscédasticité et d'autocorrélation, à l'instar des développements discutés dans le chapitre 12.) Dans un premier temps, sous l'hypothèse SC.5′, $\mathbf{x}'_t u_t$ et $\mathbf{x}'_s u_s$ ne sont pas corrélés pour $t \neq s$. Pourquoi est-ce le cas ? Supposons que $s < t$. Alors, par application de la loi des espérances itérées, $\mathrm{E}(\mathbf{x}_t'u_tu_s\mathbf{x}_s) = \mathrm{E}[\mathrm{E}(u_tu_s\mathbf{x}_t'\mathbf{x}_s)|\mathbf{x}_t'\mathbf{x}_s] = \mathrm{E}[\mathrm{E}(u_tu_s|\,\mathbf{x}_t'\mathbf{x}_s)\,\mathbf{x}_t'\mathbf{x}_s] = \mathrm{E}[0\,.\,\mathbf{x}_t'\mathbf{x}_s] = 0$. Les covariances nulles impliquent que la variance de la somme est simplement la somme des variances. On a alors : $\mathrm{Var}(\mathbf{x}_t'u_t) = \mathrm{E}(\mathbf{x}_t'u_tu_t\mathbf{x}_t) = \mathrm{E}(u^2_t\mathbf{x}'_t\mathbf{x}_t)$. D'après la loi des espérances itérées, $\mathrm{E}(u^2_t\mathbf{x}'_t\mathbf{x}_t) = \mathrm{E}[\mathrm{E}(u^2_t\mathbf{x}_t'\mathbf{x}_t|\mathbf{x}_t)] = \mathrm{E}[\mathrm{E}(u^2_t|\mathbf{x}_t)\mathbf{x}'_t\mathbf{x}_t] = \mathrm{E}(\sigma^2\mathbf{x}_t'\mathbf{x}_t) = \sigma^2\mathrm{E}(\mathbf{x}_t'\mathbf{x}_t) = \sigma^2\mathbf{A}$, en utilisant $\mathrm{E}(u^2_t|\mathbf{x}_t) = \sigma^2$ sous les hypothèses SC.3′ et SC.4′. Cela montre que $\mathrm{B} = \sigma^2\mathbf{A}$, et donc sous les hypothèses SC.1′ à SC.5′, nous avons :

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{\mathrm{a}}{\sim} \text{Normale}\,(0,\ \sigma^2\,\mathrm{A}^{-1}). \qquad \textbf{[E.23]}$$

Ce qui complète la démonstration.

À partir de l'équation (E.23), nous considérons $\hat{\boldsymbol{\beta}}$ comme suivant approximativement une distribution normale de moyenne $\boldsymbol{\beta}$ et de matrice de variance-covariance $\sigma^2\mathrm{A}^{-1}/n$. La division par la taille d'échantillon, n, est attendue ici : l'approximation de la matrice de variance-covariance de $\hat{\boldsymbol{\beta}}$ tend vers zéro à la vitesse de $1/n$. Lorsque nous remplaçons σ^2 par un estimateur convergent, soit, $\hat{\sigma}^2 = \mathrm{SCR}/(n - k - 1)$, et faisons de même pour A, soit $(\mathrm{n} - 1)\sum_{t=1}^{n}\mathbf{x}'_t\mathbf{x}_t = \mathbf{X}'\mathbf{X}/n$, nous obtenons alors un estimateur de la variance asymptotique de $\hat{\boldsymbol{\beta}}$:

$$\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}. \qquad \textbf{[E.24]}$$

Notons que les divisions par n se simplifient et que le membre à droite de l'équation (E.24) correspond simplement à l'expression standard de la matrice de variance-variance de l'estimateur des MCO sous les hypothèses de Gauss-Markov. Pour résumer, nous avons démontré que sous les hypothèses SC.1′ à SC.5′ – qui incluent les hypothèses RLM.1 à RLM.5 comme des cas particuliers – les écarts-types estimés et statistiques t sont asymptotiquement valides. Il est donc parfaitement légitime de recourir à la distribution de Student standard pour retrouver des valeurs critiques et p-valeurs dans le cadre des tests d'hypothèses. Il est à noter que dans le cadre général décrit au chapitre 11, l'hypothèse de normalité des erreurs – soit, u_t sachant $\mathbf{x}_t$, u_{t-1}, $\mathbf{x}_{t-1}$, …, u_1, $\mathbf{x}_1$ suit une distribution Normale$(0, \sigma^2)$ – ne nous aide pas plus ici puisque les statistiques t ne suivraient pas de distributions de Student exactes sous cette hypothèse de normalité. En l'absence de l'hypothèse d'exogénéité stricte des variables explicatives, les distributions exactes des paramètres sont si ce n'est impossible, difficiles à obtenir.

En relâchant l'argument précédent, nous pouvons trouver une matrice de variance-covariance robuste à l'hétéroscédasticité. L'élément clé tient à l'estimation séparée de $\mathrm{E}(u^2_t \mathbf{x}'_t \mathbf{x}_t)$ puisque cette matrice n'est plus équivalente à $\sigma^2 \mathrm{E}(\mathbf{x}'_t \mathbf{x}_t)$. En notant $\hat{u}_t$ les résidus des MCO, un estimateur convergent est donné par :

$$(n - k - 1)^{-1} \sum_{t=1}^{n} \hat{\mathbf{u}}^2_t \mathbf{x}'_t \mathbf{x}_t, \qquad \textbf{[E.25]}$$

en divisant par $n - k - 1$ plutôt que n de façon à obtenir des meilleures propriétés en échantillon finis pour l'estimateur. Lorsque nous introduisons cette expression dans l'équation (E.25), nous obtenons :

$$\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}})} = [\mathrm{n}/(\mathrm{n} - \mathrm{k} - 1)](\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{t=1}^{n} \hat{\mathrm{u}}^2_t \mathbf{x}'_t \mathbf{x}_t\right) (\mathbf{X}'\mathbf{X})^{-1} \qquad \textbf{[E.26]}$$

Les racines carrées des éléments diagonaux de cette matrice sont les mêmes que les écarts-types estimés robustes à l'hétéroscédasticité dérivés à la section 8.2 dans le cas d'un modèle en coupe instantanée. Il est également possible de proposer une extension matricielle de l'expression des écarts-types estimés robustes à l'autocorrélation (et à l'hétéroscédasticité) tels qu'obtenus à la section 12.5. Dans ce cas, la matrice devant être introduite dans l'équation (E.25) est rendue plus complexe en raison de la présence de corrélation sérielle. Voir par exemple, Hamilton (1994, section 10.5).

Statistiques de Wald pour tester des hypothèses multiples

Des démonstrations similaires peuvent être utilisées pour dériver la distribution asymptotique de la **statistique de Wald** dans le cadre de tests d'hypothèses multiples. Soit $\mathbf{R}$ une matrice de dimensions $q \times (k + 1)$ avec $q \leq (k + 1)$. On suppose q restrictions appliquées au vecteur de paramètres $(k + 1) \times 1$, $\boldsymbol{\beta}$, exprimées comme suit : H_0 : $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, avec $\mathbf{r}$ un vecteur de dimension q × 1 de constantes inconnues. Sous les hypothèses SC.1′ à SC.5′ on peut montrer que sous H_0,

$$[\sqrt{n}\ (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})]'(\sigma^2 \mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}[\sqrt{n}\ (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})] \overset{a}{\sim} \chi^2_q, \qquad \textbf{[E.27]}$$

où $\mathrm{A} = \mathrm{E}(\mathbf{x}'_t \mathbf{x}_t)$, comme dans les démonstrations des théorèmes 11.1 et 11.2. L'intuition sous-jacente de l'expression (E.25) est relativement simple. Du fait que $\sqrt{n}\ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ suit approximativement une distribution Normale$(\mathbf{0}, \sigma^2 \mathbf{A}^{-1})$, $\mathbf{R}[\sqrt{n}\ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sqrt{n}\ \mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ suit approximativement une Normale$(\mathbf{0}, \sigma^2 \mathrm{RA}^{-1}\mathrm{R}')$ par application de la propriété 3 de la distribution normale multivariée rappelée en annexe D. Sous H_0, $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, de sorte que $\sqrt{n}\ (\mathbf{R}\boldsymbol{\beta} - \mathbf{r}) \sim$ Normale$(0, \sigma^2 \mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ sous H_0. D'après la propriété 3 de la distribution du chi-deux, $\mathrm{z}'(\sigma^2 \mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}\ \mathrm{z} \sim \chi^2_q$ si z ~ Normale$(0, \sigma^2 \mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$. Pour dériver le résultat final

formellement, il faut avoir recours à la version asymptotique de cette propriété, qui peut être trouvée dans Wooldridge (2010, chapitre 3).

À partir des résultats obtenus en (E.25), nous sommes en mesure de calculer une statistique de test en remplaçant $\mathbf{A}$ et σ^2 par leurs estimateurs convergents ; ce faisant, nous ne modifions en rien la distribution asymptotique. La statistique résultant de cette transformation est appelée statistique de Wald, et peut, après avoir éliminé les tailles d'échantillon et procédé à quelques manipulations algébriques, s'écrire comme suit :

$$\mathbf{W} = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/\hat{\sigma}^2. \qquad \textbf{[E.28]}$$

Sous H_0, $W \overset{a}{\sim} \chi^2_q$, avec q le nombre de restrictions supposées. Si $\hat{\sigma}^2 = SCR/(n - k - 1)$, on peut montrer que W/q se comporte exactement comme une statistique F décrite au chapitre 4 dans le cadre de tests de restrictions multiples. [Voir, par exemple, Greene (1997, chapitre 7).] Dès lors, sous les hypothèses classiques du modèle de régression linéaire SC.1 à SC.6 dans le chapitre 10, W/q suit une distribution exacte de $F_{q,\,n-k-1}$. Sous les hypothèses SC.1′ à SC.5′, nous n'aboutissons qu'au résultat asymptotique décrit en équation (E.26). Néanmoins, il est d'usage de considérer que la statistique de Fisher habituelle suit approximativement une distribution $F_{q,\,n-k-1}$.

Une statistique de Wald robuste à l'hétéroscédasticité de forme inconnue peut être dérivée en faisant usage de la matrice explicitée en E.26 en lieu et place de $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, il en va de même pour l'expression d'une statistique de test robuste à l'hétéroscédasticité et l'autocorrélation. Il est à noter que ces versions robustes de la statistique de Wald ne peuvent être calculées à partir des sommes de carrés de résidus ou des R-carrés des modèles contraints et non contraints.

RÉSUMÉ

Dans cette annexe nous avons passé en revue brièvement l'analyse du modèle de régression linéaire sous forme matricielle. Si ces développements peuvent être inclus dans le cadre de cours plus avancés requérant l'usage de l'algèbre linéaire, ils ne sont pas nécessaires à la lecture du contenu de ce manuel. En effet, l'annexe démontre un ensemble de résultats que nous avions admis sans preuve ou démontré dans des cas particuliers ou via des méthodes moins élégantes. D'autres thèmes – tels que les propriétés asymptotiques, l'estimation des variables instrumentales, et les modèles de données de panel – peuvent être introduits sans le recours aux notations matricielles. Parmi les manuels plus avancés d'économétrie, les ouvrages de Davidson et MacKinnon (1993), Greene (1997), Hayashi (2000), et Wooldridge (2010), pourront être consultés avec profit.

MOTS-CLÉS

Condition du premier ordre p. 915
Estimateur du Quasi-Maximum de Vraisemblance (EQMV) p. 923
Estimateur sans biais à variance minimale p. 923
Matrice de variance-covariance de l'estimateur des MCO p. 919
Matrice de variance-covariance sphérique p. 919
Notations matricielles p. 914
Statistique de Wald p. 926
Théorème de Frisch-Waugh p. 917

EXERCICES

1. Soit $\mathbf{x}_t$ le vecteur de variables explicatives de dimension $1 \times (k + 1)$ pour l'observation t. Montrez que l'estimateur des MCO, $\hat{\boldsymbol{\beta}}$, peut être écrit comme suit :

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{n} \mathbf{x}_t' \mathbf{x}_t \right)^{-1} \left(\sum_{t=1}^{n} \mathbf{x}_t' y_t \right)$$

En divisant chacune des sommes par n, $\hat{\boldsymbol{\beta}}$ est alors exprimé comme une fonction des moyennes empiriques.

2. Soit $\hat{\boldsymbol{\beta}}$ le vecteur de taille $(k + 1) \times 1$ des paramètres estimés par les MCO.

i. Montrez que pour tout vecteur $\mathbf{b}$ de taille $(k + 1) \times 1$, la somme des carrés des résidus s'écrit :

$$\mathbf{SCR(b) = \hat{u}'\hat{u} + (\hat{\beta} - b)'X'X(\hat{\beta} - b).}$$

{*Indice* : Écrivez $(\mathbf{y - Xb})'(\mathbf{y - Xb}) = [\mathbf{\hat{u} + X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]'[\mathbf{\hat{u} + X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]$ et utilisez le fait que $\mathbf{X'\hat{u} = 0}$}

ii. Expliquez comment l'expression de SCR($\mathbf{b}$) de la question (i) prouve que $\hat{\boldsymbol{\beta}}$ est la solution unique de la minimisation de SCR($\mathbf{b}$) parmi toutes les valeurs possibles de $\mathbf{b}$, en supposant que $\mathbf{X}$ est de rang $k + 1$.

3. Soit $\hat{\boldsymbol{\beta}}$ le vecteur de paramètres estimés par les MCO issus de la régression de $\mathbf{y}$ sur $\mathbf{X}$. Soit $\mathbf{A}$ une matrice non singulière de dimension $(k + 1) \times (k + 1)$ et $\mathbf{z}_t \equiv \mathbf{x}_t\mathbf{A}$, $t = 1, \ldots, n$. Dès lors, $\mathbf{z}_t$ est une combinaison linéaire non singulière de $\mathbf{x}_t$ de dimension $1 \times (k + 1)$. Soit $\mathbf{Z}$ la matrice de taille $n \times (k + 1)$ contenant les vecteurs $\mathbf{z}_t$ en ligne et $\tilde{\boldsymbol{\beta}}$ le vecteur de paramètres estimés par les MCO issus de la régression de $\mathbf{y}$ sur $\mathbf{Z}$.

i. Montrez que $\tilde{\boldsymbol{\beta}} = \mathbf{A}^{-1}\tilde{\boldsymbol{\beta}}$

ii. Soit $\hat{y}_t$ les valeurs ajustées issues de la régression de départ et $\tilde{y}_t$ les valeurs ajustées issues de la régression de $\mathbf{y}$ sur $\mathbf{Z}$. Montrez que $\tilde{y}_t = \hat{y}_t$, pour tout $t = 1, 2, \ldots, n$. Comparez les résidus issus des deux régressions.

iii. Montrez que la variance estimée de $\tilde{\boldsymbol{\beta}}$ est donnée par $\hat{\sigma}^2\mathbf{A}^{-1}(\mathbf{X'X})^{-1}\mathbf{A}^{-1}$, avec $\hat{\sigma}^2$ la variance estimée issue de la régression de $\mathbf{y}$ sur $\mathbf{X}$.

iv. Soit $\hat{\beta}_j$ les paramètres estimés par les MCO issus de la régression de y_t sur $1, x_{t1}, \ldots, x_{tk}$, et $\tilde{\beta}_j$ les paramètres estimés issus de la régression de y_t sur $1, a_1x_{t1}, \ldots, a_k'x_{tk}$, avec $a_j \neq 0$, $j = 1, \ldots, k$. Utilisez les résultats de la question (i) pour identifier la relation entre $\tilde{\beta}_j$ et $\hat{\beta}_j$.

v. À supposer le cadre d'analyse décrit à la question (iv), utilisez les résultats de la question (iii) pour montrer que $\hat{\sigma}(\tilde{\beta}_j) = \hat{\sigma}(\hat{\beta}_j)/|a_j|$.

vi. À supposer le cadre d'analyse décrit à la question (iv), montrez que les valeurs absolues des statistiques t de β_j et $\hat{\beta}_j$ sont rigoureusement identiques.

4. On suppose que le modèle $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfait les hypothèses de Gauss-Markov, soit $\mathbf{G}$ une matrice non singulière et non aléatoire de rang $(k + 1) \times (k + 1)$, on définit $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$, de sorte que $\boldsymbol{\delta}$ est un vecteur de taille $(k + 1) \times 1$. Soit $\hat{\boldsymbol{\beta}}$ le vecteur des estimateurs des paramètres du modèle par les MCO, de taille $(k + 1) \times 1$. On définit $\boldsymbol{\delta} = \mathbf{G}\hat{\boldsymbol{\beta}}$ comme l'estimateur des MCO de $\boldsymbol{\delta}$.

i. Montrez que $E(\hat{\delta}|X) = \delta$.

ii. Dérivez l'expression de $Var(\hat{\boldsymbol{\delta}}|\mathbf{X})$ en fonction de σ^2, $\mathbf{X}$, et $\mathbf{G}$.

iii. À partir des éléments de réponses de l'exercice E.3 vérifiez que $\hat{\boldsymbol{\delta}}$ et l'estimation de $Var(\hat{\boldsymbol{\delta}}|\mathbf{X})$ peuvent être obtenus à partir de la régression de $\mathbf{y}$ sur $\mathbf{XG}^{-1}$.

iv. Soit $\mathbf{c}$ un vecteur de taille $(k + 1) \times 1$ contenant des éléments non nuls. Par exemple, supposons que $c_k \neq 0$. On définit $\theta = \mathbf{c}'\boldsymbol{\beta}$, de sorte que θ est un scalaire. Définissez $\delta_j = \beta_j$, $j = 1, \ldots, k - 1$ et $\delta_k = \theta$. Montrez comment définir une matrice non singulière $\mathbf{G}$ de dimension $(k + 1) \times (k + 1)$ de sorte que $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$. (*Astuce* : Chacune des k lignes de $\mathbf{G}$ devrait contenir k zéros et un élément unitaire. Quelle est la valeur de la dernière ligne ?)

v. Montrez qu'étant donné le choix de $\mathbf{G}$ à la question (iv),

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & & & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots \mathbf{1} & 0 \\ -C_0/C_K & -C_1/C_K & & \ldots & 1/C_K \end{bmatrix}$$

Utilisez cette expression pour $\mathbf{G}^{-1}$ et les résultats de la question (iii) pour conclure que θ et son écart-type estimé sont obtenus comme les coefficients de x_{tk}/c_k à l'issue de la régression de

$$y_t \text{ sur } [1 - (c_0/c_k)x_{tk}], [x_{t1} - (c_1/c_k)x_{tk}], \ldots, [x_{t,k-1} - (c_{k-1}/c_k)x_{tk}], x_{tk}/c_k, t = 1, \ldots, n.$$

Cette régression est exactement celle obtenue en exprimant β_k en fonction de θ et β_0, β_1, ..., β_{k-1}, puis en introduisant les expressions dans le modèle original et en réarrangeant l'expression. Dès lors, nous pouvons formellement justifier l'astuce dont nous avons usé tout au long de nos développements dans ce manuel, pour obtenir les écarts-types estimés de combinaisons linéaires de paramètres.

5. On suppose que le modèle $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfait les hypothèses de Gauss-Markov. Soit $\hat{\boldsymbol{\beta}}$ l'estimateur par les MCO de $\boldsymbol{\beta}$. Soit $\mathbf{Z} = \mathbf{G}(\mathbf{X})$ une matrice de taille $n \times (k + 1)$ fonction de $\mathbf{X}$ et supposons que $\mathbf{Z}'\mathbf{X}$ [une matrice de taille $(k + 1) \times (k + 1)$] est non singulière. Définissez une nouvel estimateur de $\boldsymbol{\beta}$ comme étant $\tilde{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$.

i. Montrez que $E(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$, de sorte que $\tilde{\boldsymbol{\beta}}$ est également un estimateur sans biais des paramètres conditionnellement à $\mathbf{X}$.

ii. Identifiez $Var(\tilde{\boldsymbol{\beta}}|\mathbf{X})$. Assurez-vous que la matrice est bien symétrique et de dimension $(k + 1) \times (k + 1)$ dépendant de $\mathbf{Z}$, $\mathbf{X}$, et σ^2.

iii. Quel estimateur a votre préférence entre $\hat{\boldsymbol{\beta}}$ et $\tilde{\boldsymbol{\beta}}$? Justifiez.

6. Considérez le cadre du théorème de Frisch-Waugh.

(i) Montrez en utilisant les matrices partitionnées que les conditions du premier ordre, $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ peuvent s'écrire comme suit :

$$\mathbf{X}_1'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_1'\mathbf{y}$$
$$\mathbf{X}_2'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{y}$$

Multipliez ensuite le premier système d'équations par $X_2'X_1(X_1'X_1)^{-1}$, puis retranchez le résultat obtenu du second système d'équations afin de montrer que :

$$\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 = \mathbf{X}_2'\mathbf{M}_1\mathbf{y},$$

Où $\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. Concluez que :

$$\hat{\beta}_2 = (\ddot{\mathbf{X}}_2'\ \ddot{\mathbf{X}}_2)^{-1}\ddot{\mathbf{X}}_2'\mathbf{y},$$

(iii) Utilisez le résultat de la question (ii) afin de montrer que :

$$\hat{\beta}_2 = (\ddot{\mathbf{X}}_2'\ \ddot{\mathbf{X}}_2)^{-1}\ddot{\mathbf{X}}_2'\ddot{\mathbf{y}}.$$

(iv) Utilisez l'égalité suivante $\mathbf{M}_1\mathbf{X}_1 = 0$ afin de montrer que les résidus $\ddot{\mathbf{u}}$ de la régression de $\ddot{\mathbf{y}}$ sur $\ddot{\mathbf{X}}_2$ sont identiques aux résidus $\hat{\mathbf{u}}$ issus de la regression auxiliaire de $\mathbf{y}$ sur $\mathbf{X}_1$, $\mathbf{X}_2$ [*Astuce* : Par définition et à partir du théorème FW, on a :

$$\ddot{\mathbf{u}} = \ddot{\mathbf{y}} - \ddot{\mathbf{X}}_2\hat{\beta}_2 = \mathbf{M}_1(\mathbf{y} - \mathbf{X}_2\hat{\beta}_2) = \mathbf{M}_1(\mathbf{y} - \mathbf{X}_1\hat{\beta}_1 - \mathbf{X}_2\hat{\beta}_2)$$

Le reste du développement vous incombe...]

7. On suppose que le modèle linéaire écrit sous forme matricielle de la manière suivante :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

satisfait les hypothèses E.1, E.2 et E.3. Partitionnez le modèle de sorte que :

$$\mathbf{y} = \mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + u,$$

où $\mathbf{X}_1$est de dimensions $n \times (k_1 + 1)$ et $\mathbf{X}_2$ de dimensions $n \times k_2$.

i. Considérez la proposition suivante pour estimer β_2. Premièrement, régressez $\mathbf{y}$ sur $\mathbf{X}_1$ et extrayez le résidu, que vous noterez $\ddot{\mathbf{y}}$. Puis, regressez $\ddot{\mathbf{y}}$ sur $\ddot{\mathbf{X}}_2$ afin d'obtenir $\tilde{\beta}_2$. Montrez que $\tilde{\beta}_2$ est généralement biaisé et quantifiez ce biais. [Vous devez exprimer $E(\tilde{\beta}_2|\mathbf{X})$ en fonction de β_2, $\mathbf{X}_2$et de la matrice $\mathbf{M}_1$]

ii. En particulier, écrivez :

$$y = \mathbf{X}_1\beta_1 + \mathbf{X}_k\beta_k + \boldsymbol{u},$$

où $\mathbf{X}_k$ est un vecteur de dimension $n \times 1$ de la variable x_{tk}. Montrez que :

$$E(\tilde{\beta}_k|\mathbf{X}) = \left(\frac{SCR_k}{\sum_{t=1}^{n} x_{tk}^2}\right)\beta_k,$$

avec SCR_k la somme des carrés des résidus de la régression auxiliaire de x_{tk} sur 1, x_{t1}, x_{t2}, ..., $x_{t,k-1}$. Pour quelle raison le facteur multiplicatif, β_k, n'est-il jamais supérieur à un ?

(iii) On suppose β_1 connu. Montrez que la régression $\mathbf{y} - \mathbf{X}_1\beta_1$ sur $\mathbf{X}_2$ donne un estimateur sans biais de β_2 (conditionnellement à $\mathbf{X}$).

ANNEXE

F

RÉPONSES AUX QUESTIONS INTITULÉES « POUR ALLER PLUS LOIN »

Traduction de Pierre André et Mikael Petitjean

F.1 CHAPITRE 2

Question 2.1. L'équation (2.6) est appropriée à la condition qu'il n'y ait pas de corrélation entre la variable « présence aux cours » et les autres facteurs compris dans u, tels que l'aptitude de l'étudiant, sa motivation, son âge, etc. Cela semble improbable.

Question 2.2. Environ 11,05 dollars. Pour parvenir à cette réponse, nous devons calculer le déflateur des prix en nous basant sur les salaires moyens mesurés en 1976 et 2003, soit 19,06/5,90 ≈ 3,23. Nous obtenons 11,05 en multipliant 3,42 dollars par 3,23.

Question 2.3. Lorsque *shareA* = 60 dans (2.28), la réponse est 54,65. Ce n'est pas absurde : si le candidat A effectue 60 % du total des dépenses électorales, on s'attend à ce qu'il capte environ 55 % du total des votes.

Question 2.4. L'équation sera $\widehat{salaryhun} = 9\,631{,}91 + 185{,}01roe$; il suffit de multiplier l'équation (2.39) par 10.

Question 2.5. L'équation 2.58 est : $\text{Var}(\hat{\beta}_0) = (\sigma^2 n^{-1})(\sum_{i=1}^{n} x_i^2) / \sum_{i=1}^{n}(x_i - \bar{x})^2$. Or, lorsque $\bar{x} = 0$, $\text{Var}(\hat{\beta}_0) = \sigma^2 n^{-1}$, car le second terme (celui qui multiplie σ^2/n) est égal à 1. Dans ce cas, la variance est minimisée.

F.2 CHAPITRE 3

Question 3.1. Les facteurs dans u peuvent inclure l'âge, la répartition hommes-femmes, l'effectif des forces de police (ou, plus généralement, les ressources consacrées à la lutte contre la criminalité), ainsi que d'autres facteurs historiques et socio-économiques (comme le taux de chômage ou le revenu par habitant). La plupart de ces facteurs sont très vraisemblablement corrélés à *prbconv* et *avgsen*, ce qui implique que les estimateurs de (3.5) sont biaisés. Par exemple, l'effectif des forces de police est sans doute corrélé avec les variables *prbcon* et *avgsen*, étant donné que les missions de la police portent à la fois sur la prévention et sur la répression de la criminalité. Nous devons veiller à inclure le maximum de variables susceptibles d'expliquer le taux d'homicide.

Question 3.2. Nous pouvons utiliser la troisième propriété des MCO, qui porte sur les valeurs estimées et les résidus : lorsque nous utilisons les valeurs moyennes des variables indépendantes dans la droite de régression des MCO, nous obtenons la moyenne de la variable dépendante. Donc, $\overline{colGPA} = 1{,}29 + 0{,}453\,\overline{hsGPA} + 0{,}0094\,\overline{ACT} = 1{,}29 + 0{,}453(3{,}4) + 0{,}0094(24{,}2) \approx 3{,}06$. Vous pouvez le vérifier en utilisant la base de données GPA1.

Question 3.3. Non, car la variable *shareA* n'est pas une fonction parfaitement linéaire de *expendA* et *expendB*. Il n'y a donc pas de colinéarité parfaite. Il s'agit plutôt d'une fonction non linéaire puisque $shareA = 100 \cdot [expendA/(expendA + expendB)]$. Il est donc légitime d'avoir *expendA expendB* et *shareA* comme variables explicatives dans le même modèle.

Question 3.4. Si nous cherchons à expliquer l'effet de la présence au cours (soit x_1) sur y, la corrélation *entre* les *autres* variables explicatives (x_2, x_3, etc.) n'intervient pas dans le calcul de $\hat{\beta}_1$ ou de $\text{Var}(\hat{\beta}_1)$. Nous en avons discuté dans la section 3.4. Ces *autres* variables jouent le rôle de variables de contrôle et nous n'avons pas à nous préoccuper du degré de colinéarité *entre elles* si notre intérêt porte sur x_1. Nous incluons les variables de contrôle dans le modèle en raison de leur corrélation potentielle avec *la variable* x_1. L'objectif est d'effectuer une analyse *ceteris paribus* de l'effet de x_1 sur y qui tienne la route.

F.3 CHAPITRE 4

Question 4.1. Les hypothèses de Gauss-Markov restent satisfaites : si u est indépendante des variables explicatives, alors $\text{E}(u|x_1, \ldots, x_k) = \text{E}(u)$, et $\text{Var}(u|x_1, \ldots, x_k) = \text{Var}(u)$. Par ailleurs, on peut constater que $\text{E}(u) = 0$.

Dès lors, RLM.4 et RLM.5 sont vérifiées. Par contre, les hypothèses du MRLC ne sont pas respectées puisque u n'est pas normalement distribuée (ce qui constitue une violation de RLM.6).

Question 4.2. $H_0 : \beta_1 = 0$, $H_1 : \beta_1 < 0$.

Question 4.3. Étant donné que $\hat{\beta}_1 = 0{,}56 > 0$ et que nous testons l'hypothèse $H_1 : \beta_1 > 0$, la valeur p du test unilatéral est égale à la moitié de la valeur p du test bilatéral, soit 0,043.

Question 4.4. $H_0 : \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. $k = 8$ et $q = 4$. La version restreinte du modèle, valide sous l'hypothèse nulle, est $score = \beta_0 + \beta_1 classize + \beta_2 expend + \beta_3 tchcomp + \beta_4 enroll + u$.

Question 4.5. La statistique F du test de redondance de ACT est $[(0{,}291 - 0{,}183)/(1 - 0{,}291)](680 - 3) \approx 103{,}13$. Par conséquent, la valeur absolue de la statistique t est environ 10,16, correspondant à la racine carrée de 103,13. Comme $\hat{\beta}_{ACT}$ est négatif, la valeur de la statistique t doit l'être également, soit $t_{ACT} = -10{,}16$.

Question 4.6. Cela n'a pas beaucoup d'impact. Nous pouvons réaliser le test F de signification conjointe de *droprate* et *gradrate* à partir des R carrés du tableau 4.1 : $F = [(0{,}361 - 0{,}353)/(1 - 0{,}361)](402/2) \approx 2{,}52$. Dans le tableau G.3a, la valeur critique du test à un seuil de 10 % est 2,30 ; à un seuil de 5 %, le tableau G.3b indique une valeur égale à 3. La valeur p est égale à environ 0,082. Donc, *droprate* et *gradrate* ne sont conjointement significatives qu'à un seuil de 10 %. De toutes manières, l'ajout de ces deux variables n'a qu'un effet marginal sur le coefficient de *b/s*.

F.4 CHAPITRE 5

Question 5.1. Cela exige la formulation de plusieurs hypothèses. Il semble tout d'abord raisonnable de penser que *score* dépend positivement de *priGPA* si bien que $\beta_2 > 0$. Ensuite, la corrélation entre *skipped* et *priGPA* est très probablement négative si bien que $\text{Cov}(skipped, priGPA) < 0$. Par conséquent, $\beta_2\delta_1 < 0$, ce qui implique que plim $\tilde{\beta}_1 < \beta_1$. En conclusion, une régression linéaire simple, qui ignorerait la variable *priGPA*, est donc susceptible de conduire à une surestimation de l'importance de l'absentéisme au cours, en se rappelant au passage que β_1 est probablement négatif (au pire, non positif).

Question 5.2. $\hat{\beta}_j \pm 1{,}96\,\hat{\sigma}(\hat{\beta}_j)$ est l'intervalle de confiance asymptotique à 95 %. Nous pouvons éventuellement remplacer 1,96 par 2 pour faciliter le calcul.

F.5 CHAPITRE 6

Question 6.1. Comme $fincdol = 1\,000 \cdot faminc$, le coefficient de *fincdol* est égal au coefficient de *faminc* divisé par 1 000, soit 0,0927/1 000 = 0,0000927. L'écart-type estimé est également divisée par 1 000 de telle sorte que les statistiques t et F ne changent pas. Pour plus de lisibilité, il est préférable de mesurer le revenu familial en milliers de dollars.

Question 6.2. D'une manière plus général, l'équation peut s'écrire

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \ldots,$$

dans laquelle x_2 est une proportion plutôt qu'un pourcentage. *Ceteris paribus*, $\Delta\log(y) = \beta_2 \Delta x_2$, $100 \cdot \Delta\log(y) = \beta_2(100 \cdot \Delta x_2)$, ou $\%\Delta y \approx \beta_2(100 \cdot \Delta x_2)$. Étant donné que Δx_2 est une variation en proportion, $100 \cdot \Delta x_2$ représente une variation en pourcentage [*de la variable y*]. Plus précisément, si $\Delta x_2 = 0{,}01$ [soit une augmentation de 1 point de pourcentage pour la variable x_2], alors $100 \cdot \Delta x_2 = 1$, ce qui correspond bien à une variation en pourcentage [de la variable y]. En conclusion, β_2 mesure bien la variation en pourcentage de la variable y lorsque x_2 augmente d'un point de pourcentage.

Question 6.3. Le nouveau modèle est $stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 + \beta_5 ACT^2 + \beta_6 priGPA.atndrte + \beta_7 ACT.atndrte + u$. Par conséquent, l'effet marginal (ou partiel) de *atndrte* sur *stndfnl* est $\beta_1 + \beta_6 priGPA + \beta_7 ACT$. Cela correspond aux termes que nous devons multiplier par $\Delta atndrte$ pour obtenir la variation de *stndfnl*, *ceteris paribus*.

Question 6.4. L'équation (6.21) indique que $\bar{R}_2 = 1 - \hat{\sigma}^2/[SCT/(n-1)]$. Pour un échantillon et une variable dépendante donnés, SCT/(n – 1) est fixe. Quand nous utilisons différentes variables explicatives, seul $\hat{\sigma}^2$ change. Lorsque $\hat{\sigma}^2$ diminue, $\bar{R}_2$ augmente. Si nous minimisons $\hat{\sigma}$ (et donc $\hat{\sigma}^2$), nous maximisons le R^2 par la même occasion.

Question 6.5. Une possibilité est de recueillir des données sur les rémunérations d'acteurs de cinéma au sein d'un échantillon, en reprenant également la rentabilité des films dans lesquels ils sont intervenus. La fréquence peut être annuelle. Dans une régression linéaire simple, nous pouvons expliquer les rémunérations par la rentabilité. Nous devons néanmoins penser à ajouter des variables de contrôle qui influencent la rémunération, comme l'âge, le sexe et la catégorie du film. Les méthodes qui permettent l'inclusion de variables qualitatives sont abordées dans le chapitre 7.

F.6 CHAPITRE 7

Question 7.1. Non, car nous ne savons pas quel parti politique est représenté par *party* : quel sera le parti représenté par la valeur 1 ? Un nom plus approprié serait *Dem*, égal à 1 lorsqu'il s'agit d'un candidat démocrate et 0 autrement. Nous pourrions également choisir *Rep*, égal à 1 pour un républicain et 0 pour un démocrate.

Question 7.2. En considérant *outfield* comme le groupe de référence, nous pouvons inclure *frstbase*, *scndbase*, *thrdbase*, *shrtstop*, et *catcher*.

Question 7.3. L'hypothèse nulle est $H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$, si bien qu'il y a 4 restrictions. Comme d'habitude, nous pouvons recourir à un test F (avec $q = 4$ et k égal au nombre de variables explicatives).

Question 7.4. Comme *tenure* apparaît sous une forme quadratique, nous devrions l'appliquer à la fois aux femmes et aux hommes. Par exemple, nous devrions ajouter les variables *female·tenure* et *female·tenure*^{2} aux variables explicatives déjà présentes dans le modèle.

Question 7.5. Dans (7.31), nous devons fixer *pcnv* = 0, *avgsen* = 0, *tottime* = 0, *ptime86* = 0, *qemp86* = 4, *black* = 1, et *hispan* = 0. Cela donne : $\widehat{arr86} = 0{,}380 - 0{,}038(4) + 0{,}170 = 0{,}398$, soit presque 0,4. Il est difficile de déterminer si cette estimation est « raisonnable ». Pour une personne sans condamnation antérieure, qui a travaillé toute l'année, cette estimation pourrait sembler élevée. Rappelez-vous néanmoins que la population concerne des hommes qui ont déjà été incarcérés avant 1986.

F.7 CHAPITRE 8

Question 8.1. Cette affirmation est manifestement fausse. Par exemple, dans l'équation (8.7), l'écart-type estimé standard pour *black* est 0,147, alors que l'écart-type estimé robuste à la présence d'hétéroscédasticité est 0,118.

Question 8.2. Le test F peut être effectué en régressant $\hat{u}^2$ sur *marrmale*, *marrfem* et *singfem* (*singmale* est le groupe de référence). Comme il y a trois variables explicatives dans la régression et que $n = 526$, les *ddl* sont 3 et 522.

Question 8.3. Une statistique t égale à 2,96 indique que l'hétéroscédasticité est clairement un problème dans l'équation de la richesse. Sur un plan plus pratique, nous savons que l'écart-type estimé des MCP est 0,063

alors que l'écart-type estimé robuste à la présence d'hétéroscédasticité est sensiblement plus élevée, soit 0,104. Sachant, par ailleurs, que l'écart-type estimé standard des MCO est 0,061, on peut constater que l'ajustement des écarts-types estimés en présence d'une forme inconnue d'hétéroscédasticité n'est pas marginal.

Question 8.4. Dans la distribution F, si $ddl = (2, \infty)$, la valeur critique à un seuil de 1 % est 4,61. Par conséquent, une statistique F égale à 11,15 indique clairement que nous devons rejeter l'hypothèse nulle selon laquelle les erreurs transformées, $u_i / \sqrt{h_i}$, sont homoscédastiques. (En fait, la valeur p est inférieure à 0,00002, sur base de la distribution $F_{2,804}$.) En conclusion, notre modèle pour Var(u|**x**) est inadéquat car il ne permet pas d'éliminer l'hétéroscédasticité dans u.

F.8 CHAPITRE 9

Question 9.1. Comme il s'agit de variables binaires, les mettre au carré ne sert à rien : $black^2 = black$ et $hispan^2 = hispan$.

Question 9.2. Lorsque le terme d'interaction $educ \cdot IQ$ est inclus dans l'équation, le coefficient de *educ*, soit β_1, mesure l'effet de *educ* sur log(*wage*) lorsque $IQ = 0$. Or, il n'y a aucune personne dans cette population dont le QI est proche de zéro. En fait, l'effet marginal du niveau d'instruction sur log(*wage*) est égal à $\beta_1 + \beta_9 IQ$. Dans la colonne 3, si nous considérons la moyenne du QI dans la population (égale à 100), l'estimation du rendement d'une année d'instruction supplémentaire est égale à 0,018 + 0,00034(100), soit 0,052. Cette estimation est presque identique à celle du coefficient de *educ* dans la colonne (2).

Question 9.3. Non. Si $educ^*$ est un nombre entier, l'erreur de mesure est nulle car $educ^* = educ$: la personne n'a pas reçu d'instruction au-delà de son plus haut niveau de scolarité atteint. Si $educ^*$ n'est pas un nombre entier, $educ < educ^*$: l'erreur de mesure est donc négative. À tout le moins, la moyenne de e_1 n'est pas nulle ; il est également probable que le terme d'erreur e_1 et la variable $educ^*$ soient corrélés.

Question 9.4. La décision d'un responsable politique en exercice de ne pas se représenter aux élections peut être systématiquement liée à son évaluation de la probabilité de remporter la prochaine élection. Par conséquent, si nous ne tenons pas compte des responsables politiques en exercice qui décident de ne pas se représenter, l'échantillon ne contiendra que ceux qui sont en position de force et il ne sera pas représentatif de l'ensemble de la population des responsables politiques en exercice. Cela conduit à un problème de sélection de l'échantillon. Par contre, si nous désirons analyser l'effet des dépenses électorales uniquement pour les élus qui décident de se représenter aux élections, ce problème de sélection de l'échantillon ne se pose pas.

F.9 CHAPITRE 10

Question 10.1. Le multiplicateur d'impact est égal à 0,48 tandis que le multiplicateur de long-terme vaut 0,48 – 0,15 + 0,32 = 0,65.

Question 10.2. Les variables explicatives sont $x_{t1} = z_t$ et $x_{t2} = z_{t-1}$. L'absence de colinéarité parfaite signifie que ces variables ne peuvent pas être constantes et qu'il ne peut pas exister de relation exacte sur le plan linéaire entre ces variables au sein de l'échantillon. Cela exclut la possibilité que les observations $z_1, \ldots, z_n$ prennent toutes la même valeur ou que les $z_0, z_1, \ldots, z_{n-1}$ prennent toutes la même valeur. Cela exclut également d'autres types de relation. Par exemple, si $z_t = a + bt$ (a et b étant des constantes), alors $z_{t-1} = a + b(t - 1) = (a + bt) - b = z_t - b$, ce qui correspond à une fonction parfaitement linéaire de z_t.

Question 10.3. Si $\{z_t\}$ évolue lentement au cours du temps (comme c'est le cas de la plupart des séries temporelles en économie, qu'elles soient en niveau ou en log), alors z_t et z_{t-1} peuvent être fortement corrélées. Par exemple, la corrélation entre $unem_t$ et $unem_{t-1}$ dans PHILLIPS est égale à 0,75.

Question 10.4. Non, car une tendance temporelle linéaire avec $\alpha_1 < 0$ devient de plus en plus négative au fur et à mesure que t augmente. Comme le taux général de fertilité (*gfr*) ne peut pas être négatif, un trend linéaire dont le coefficient est négatif conduira à des valeurs absurdes de *gfr* à un moment donné dans le futur.

Question 10.5. Le point d'intersection pour le mois de mars est $\beta_0 + \delta_2$. Les variables binaires saisonnières sont strictement exogènes puisqu'elles suivent une évolution déterministe. En effet, les mois de l'année ne changent pas en fonction des variations enregistrées par les variables explicatives ou par la variable dépendante.

F.10 CHAPITRE 11

Question 11.1. i. Non, car $E(y_t) = \delta_0 + \delta_1 t$ dépend effectivement de t. ii. Oui, car $y_t - E(y_t) = e_t$ est distribuée indépendamment et identiquement.

Question 11.2. Si nous utilisons $inf_t^e = (1/2)inf_{t-1} + (1/2)inf_{t-2}$ dans $inf_t - inf_t^e = \beta_1(unem_t - \mu_0) + e_t$, alors nous pouvons écrire $inf_t - (1/2)(inf_{t-1} + inf_{t-2}) = \beta_0 + \beta_1 unem_t + e_t$, où $\beta_0 = -\beta_1\mu_0$, comme auparavant. Ensuite, nous pouvons régresser y_t sur $unem_t$, sachant que $y_t = inf_t - (1/2)(inf_{t-1} + inf_{t-2})$. En construisant y_t de la sorte, nous perdons néanmoins les deux premières observations.

Question 11.3. Non, car les erreurs u_t et u_{t-1} sont corrélées. Plus précisément, $\text{Cov}(u_t,u_{t-1}) = E[(e_t + \alpha_1 e_{t-1})(e_{t-1} + \alpha_1 e_{t-2})] = \alpha_1 E(e^2_{t-1}) = \alpha_1\sigma_e^2 \neq 0$ si $\alpha_1 \neq 0$. Si les erreurs sont autocorrélées, la dynamique du modèle n'a pas été complètement capturée.

F.11 CHAPITRE 12

Question 12.1. Partons de l'équation (12.4). Dans le cas d'un processus MA(1), seuls les termes adjacents sont corrélés, ce qui veut dire que la covariance entre $x_t u_t$ et $x_{t+1}u_{t+1}$ est $x_t x_{t+1}\text{Cov}(u_t,u_{t+1}) = x_t x_{t+1}\alpha\sigma_e^2$. Par conséquent, la formule peut s'écrire :

$$\text{Var}(\hat{\beta}_1) = SCT_x^{-2}\left(\sum_{t=1}^{n} x_t^2\text{Var}(u_t) + 2\sum_{t=1}^{n-1} x_t x_{t+1}E(u_t u_{t+1})\right)$$

$$= \sigma^2/SCT_x + (2/SCT_x^2)\sum_{t=1}^{n-1}\alpha\sigma_e^2 x_t x_{t+1}$$

$$= \sigma^2/SCT_x + \alpha\sigma_e^2(2/SCT_x^2)\sum_{t=1}^{n-1} x_t x_{t+1},$$

où $\sigma^2 = \text{Var}(u_t) = \sigma_e^2 + \alpha_1^2\sigma_e^2 = \sigma_e^2(1 + \alpha_1^2)$. Si x_t et x_{t+1} sont corrélées au sein de l'échantillon, le second terme sera non nulle lorsque $\alpha \neq 0$. Notez que si x_t et x_{t+1} sont positivement corrélées et que $\alpha < 0$, la variance standard de l'estimateur surestime la variance vraie. Il s'agit du cas typique lorsque les erreurs suivent un processus MA(1), sachant que l'équation est en niveau et non en différence première.

Question 12.2. $\hat{\rho} \pm 1{,}96\ \hat{\sigma}(\hat{\rho})$, où $\hat{\sigma}(\hat{\rho})$ est l'écart-type estimé de $\hat{\rho}$ dans la régression (12.14). Si nécessaire, nous pouvons utiliser l'écart-type estimé des MCO robuste à la présence de l'hétéroscédasticité. Il est néanmoins plus difficile de prouver la validité de l'intervalle sur le plan asymptotique dans ce cas, car les résidus des MCO dépendent de $\hat{\beta}_j$.

Question 12.3. Le modèle qu'il faut envisager est : $u_t = \rho_1 u_{t-1} + \rho_4 u_{t-4} + e_t$. L'hypothèse nulle est $H_0 : \rho_1 = 0$, $\rho_4 = 0$. Sous l'hypothèse alternative, au moins une de ces deux conditions n'est pas respectée. Il convient

donc de régresser $\hat{u}_t$ sur $\hat{u}_{t-1}$ et $\hat{u}_{t-4}$ afin d'obtenir la statistique F de signification conjointe (puisque nous testons deux contraintes).

Question 12.4. Nous serions probablement amenés à utiliser une équation en différence première. En effet, $\hat{\rho} = 0{,}92$ est suffisamment proche de 1 pour mettre en doute la validité de l'équation en niveau. Voir le chapitre 18 pour une discussion plus poussée.

Question 12.5. Comme l'équation ne contient qu'une seule variable explicative, le test de White est facile à réaliser. Il suffit de régresser $\hat{u}_t^2$ sur $return_{t-1}$ et $return_{t-1}^2$, en n'oubliant pas le point d'intersection (comme toujours), puis de calculer le test F de signification conjointe de $return_{t-1}$ et $return_{t-1}^2$. Si ces deux variables sont conjointement significatives (à un seuil de 5 %, par exemple), alors l'hypothèse nulle d'homoscédasticité est rejetée.

F.12 CHAPITRE 13

Question 13.1. Oui, en supposant que nous avons bien pris en compte toutes les variables pertinentes. Le coefficient de *black* est 1,076 ; avec un écart-type de 0,174, il n'est pas statistiquement différent de 1. L'intervalle de confiance à 95 % est environ de 0,735 à 1,417.

Question 13.2. Le coefficient de *highearn* montre que si le salaire maximum ne change pas, les gros revenus passent nettement plus de temps en arrêt-maladie – environ 29,2 % en moyenne [puisque $\exp(0{,}256) - 1 \approx 0{,}292$].

Question 13.3. Tout d'abord, $E(v_{i1}) = E(a_i + u_{i1}) = E(a_i) + E(u_{i1}) = 0$. De la même manière, $E(v_{i2}) = 0$. La covariance entre v_{i1} et v_{i2} est donc simplement $E(v_{i1}v_{i2}) = E[(a_i + u_{i1})(a_i + u_{i2})] = E(a_i^2) + E(a_iu_{i1}) + E(a_iu_{i2}) + E(u_{i1}u_{i2}) = E(a_i^2)$, puisque toutes les covariances sont nulles par hypothèse. Mais $E(a_i^2) = \text{Var}(a_i)$, puisque $E(a_i) = 0$. Cela implique une corrélation sérielle positive entre les erreurs du même i, ce qui biaise les écarts-types habituels des MCO dans une régression sur données empilées.

Question 13.4. Comme $\Delta admn = admn_{90} - admn_{85}$ est la différence entre deux variables indicatrices, elle vaut – 1 quand $admn_{90} = 0$ et $admn_{85} = 1$. En d'autres termes, l'État de Washington avait une loi de suspension immédiate du permis de conduire qui a été annulée en 1990.

Question 13.5. Non, tout comme cela ne cause pas de biais ou d'absence de convergence dans une régression de séries temporelles avec des variables explicatives strictement exogènes. Il y a deux raisons à cette préoccupation. Tout d'abord, dans le cas général et quelle que soit l'équation, la corrélation sérielle des erreurs biaise les écarts-types et les statistiques de test des MCO. Ensuite, elle veut dire que les MCO sur données empilées ne sont pas aussi efficaces que les estimateurs qui prennent en compte la corrélation sérielle (comme dans le chapitre 12).

F.13 CHAPITRE 14

Question 14.1. Que nous estimions des différences premières ou que nous faisions la transformation *within*, il sera difficile d'estimer le coefficient de $kids_{it}$. Par exemple, avec la transformation *within*, si $kids_{it}$ ne varie pas pour la famille i, $\ddot{kids}_{it} = kids_{it} - \overline{kids}_i = 0$ pour $t = 1, 2, 3$. Tant que certaines familles ont des variations de $kids_{it}$, nous pouvons calculer un estimateur à effets fixes, mais le coefficient de *kids* risque d'être estimé de manière très imprécise. Il y a une forme de multicolinéarité dans la transformation à effets fixes (ou dans l'estimation en différences premières).

Question 14.2. Si une entreprise ne recevait pas de subvention la première année, elle pouvait éventuellement en recevoir une la seconde année. Mais si elle avait reçu une subvention la première année, elle ne pouvait pas la recevoir la seconde année. Donc si $grant_{-1} = 1$, alors $grant = 0$. Cela induit une corrélation négative

entre *grant* et $grant_{-1}$. Nous pouvons le vérifier en régressant *grant* sur $grant_{-1}$, en utilisant les données de JTRAIN pour 1989. Utilisant toutes les entreprises, nous trouvons

$$\widehat{grant} = 0{,}248 - 0{,}248\, grant_{-1}$$
$$(0{,}035)\ (0{,}072)$$
$$n = 157, R^2 = 0{,}070.$$

Le coefficient de $grant_{-1}$ doit être l'opposé de la constante puisque $\widehat{grant} = 0$ quand $grant_{-1} = 1$.

Question 14.3. Cela suggère que l'effet inobservé a_i est corrélé positivement à $union_{it}$. Souvenez-vous que les MCO sur données empilées laissent a_i dans le terme d'erreur alors que les effets fixes l'en extraient. Par définition, a_i a un effet positif sur log(*wage*). Par l'analyse standard des variables omises (voir le chapitre 3), les MCO sont biaisés vers le haut quand la variable explicative (*union*) est positivement corrélée à la variable omise (a_i). Appartenir à un syndicat semble donc lié positivement aux facteurs inobservés et constants au cours du temps qui affectent le salaire.

Question 14.4. Cela n'a pas de sens si toutes les sœurs dans une famille ont le même père et la même mère. Dans ce cas, l'origine ethnique des parents est la même pour toutes les sœurs, et s'annule donc dans (14.13).

F.14 CHAPITRE 15

Question 15.1. Probablement pas. Dans l'équation (15.18), la durée de la scolarisation fait partie du terme d'erreur. Si certains hommes qui ont eu un petit numéro à la loterie prolongent leurs études, le numéro à la loterie et la durée de scolarité sont corrélées, ce qui viole la première condition des variables instrumentales, dans l'équation (15.4).

Question 15.2. i. Dans (15.27), nous avons besoin que les effets de pairs au sein des groupes de lycée continuent à agir à l'université. C'est-à-dire : pour le même résultat aux tests d'admission, un étudiant qui allait dans une école où l'on fumait beaucoup de cannabis aurait plus de chances de fumer à l'université. Même si la condition d'identification (15.27) peut être valide, le lien risque d'être faible.

ii. Nous devons supposer que le pourcentage de lycéens fumant du cannabis dans un lycée donné n'est pas corrélé avec les caractéristiques inobservées affectant les notes à l'université. Bien que la qualité du lycée soit partiellement prise en compte en incluant *SAT* dans l'équation, cela risque de ne pas suffire. Il se peut que les lycées qui préparent le mieux leurs étudiants pour l'université aient moins de fumeurs de cannabis. Ou que l'utilisation de cannabis soit corrélée avec les revenus moyens des familles. Ce sont bien entendu des questions empiriques auxquelles il est possible de répondre dans certains cas.

Question 15.3. Le nombre de membres de la National Rifle Association et le nombre d'abonnés aux magazines d'armes sont probablement corrélés avec l'existence de lois sur le contrôle des armes. Cependant, il est possible que ces variables instrumentales soient corrélées avec des facteurs inobservés qui affectent les agressions à main armée. En effet, on pourrait penser que la population s'intéresse aux armes en réaction à une criminalité élevée, et que prendre en compte les variables économiques et démographiques ne soit pas suffisant pour capturer ce mécanisme. Il sera difficile d'affirmer de manière indiscutable que ces variables sont exogènes dans l'équation de la criminalité.

Question 15.4. Comme d'habitude, il y a deux conditions. Tout d'abord, il faut que les dépenses publiques soient liées au parti du président, après avoir contrôlé pour le taux d'investissement et la croissance de la population active. En d'autres termes, l'instrument doit avoir une corrélation partielle avec la variable explicative endogène. On pourrait penser que les dépenses publiques croissent moins vite avec un président républicain, mais cela n'a manifestement pas toujours été le cas aux États-Unis et il faudrait le tester en

utilisant la statistique t de REP_{t-1} dans la forme réduite $gGOV_t = \pi_0 + \pi_1 REP_{t-1} + \pi_2 INVRAT_t + \pi_3 gLAB_t + v_t$. Nous devons aussi faire l'hypothèse que le parti du président n'a pas d'effet direct sur $gGDP$. Ce serait faux si, par exemple, les politiques monétaires diffèrent entre les partis et ont un effet direct sur la croissance du PIB.

F.15 CHAPITRE 16

Question 16.1. Probablement pas. Comme les entreprises choisissent leurs prix de vente et les dépenses publicitaires de manière jointe, nous ne sommes pas intéressés par une expérience où, par exemple, la publicité changerait de manière exogène pour en déterminer l'effet sur le prix de vente. Au lieu de cela, nous devrions modéliser les prix et les dépenses de publicité en fonction de la demande et des coûts. C'est ce que la théorie économique encourage.

Question 16.2. Nous devons faire deux hypothèses. Premièrement, la croissance de l'offre monétaire doit intégrer l'équation (16.22) et avoir une corrélation partielle avec *inf*. Ensuite, nous devons supposer qu'il est inutile d'inclure la croissance de l'offre monétaire dans l'équation (16.23). Si nous pensons que l'offre monétaire doit être incluse dans (16.23), nous devons encore trouver un instrument pour *inf*. Bien entendu, l'hypothèse que la croissance de l'offre monétaire est exogène peut aussi être défendue.

Question 16.3. Il faut utiliser le test d'Hausman du Chapitre 15. En particulier, soit $\hat{v}_2$ le résidu des MCO de la régression de forme réduite de *open* sur log(*pcinc*) et log(*land*). Ensuite, il faut régresser *inf* par les MCO sur *open*, log(*pcinc*), et $\hat{v}_2$; puis calculer la statistique t de la significativité de $\hat{v}_2$. Si $\hat{v}_2$ est significatif, les estimateurs des DMC et des MCO sont statistiquement différents.

Question 16.4. L'équation de demande ressemble à

$$\log(fish_t) = \beta_0 + \beta_1 \log(prcfish_t) + \beta_2 \log(inc_t) + \beta_3 \log(prcchick_t) + \beta_4 \log(prcbeef_t) + u_{t1},$$

où les logarithmes permettent de rendre toutes les élasticités constantes. Par hypothèse, la fonction de demande n'inclut pas de saisonalité, l'équation ne contient donc pas les variables indicatrices liées aux mois de l'année (disons fev_t, mar_t, ..., dec_t, janvier étant le mois de référence). De plus, par hypothèse, l'offre de poisson est saisonnière, ce qui veut dire que l'offre dépend d'au moins une de ces variables indicatrices mensuelles. Même sans écrire la forme réduite de log(*prcfish*), nous pouvons savoir qu'elle dépend des variables indicatrices mensuelles. Comme ces variables sont exogènes, elles peuvent être utilisées comme instrument de log(*prcfish*) dans l'équation de demande. Nous pouvons donc estimer une équation de demande de poisson en utilisant les variables indicatrices mensuelles comme instrument de log(*prcfish*). Pour l'identification, il faut qu'au moins une variable indicatrice mensuelle ait un coefficient différent de zéro dans la forme réduite de log(*prcfish*).

F.16 CHAPITRE 17

Question 17.1. $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$, il y a trois restrictions donc trois degrés de liberté dans le test du RV ou de Wald.

Question 17.2. Il nous faut la dérivée partielle de $\Phi(\hat{\beta}_0 + \hat{\beta}_1 nwifeinc + \hat{\beta}_2 educ + \hat{\beta}_3 exper + \hat{\beta}_4 exper^2 + ...)$ par rapport à *exper*, qui est $\phi(\cdot)(\hat{\beta}_3 + 2\hat{\beta}_4 exper)$, où $\phi(\cdot)$ est évaluée aux valeurs données et au niveau d'expérience initial. Il nous faut donc calculer la densité de la loi normale standardisée en $0{,}270 - 0{,}012(20{,}13) + 0{,}131(12{,}3) + 0{,}123(10) - 0{,}0019(10^2) - 0{,}053(42{,}5) - 0{,}868(0) + 0{,}036(1) \approx 0{,}463$, en fonction du niveau initial d'expérience (10). Or $\phi(0{,}463) = (2\pi)^{-1/2} \exp[-(0{,}463^2)/2] \approx 0{,}358$. Il nous faut ensuite multiplier ceci

par $\hat{\beta}_3 + 2\hat{\beta}_4 exper$ qui doit être évalué en $exper = 10$. L'effet partiel est donc approximativement 0,358[0,123 – 2(0,0019)(10)] ≈ 0,030. En d'autres termes, aux valeurs des variables explicatives données et en partant d'$exper = 10$, une année supplémentaire d'expérience augmente la probabilité de participation à la force de travail d'environ 0,03.

Question 17.3. Non. Le nombre de relations extraconjugales est un nombre entier positif, qui prend probablement des valeurs nulles ou de petites valeurs pour une grande partie de la population. Il n'est pas réaliste d'utiliser un modèle Tobit, qui permet d'avoir un point de masse en zéro, mais où y a une distribution continue sur les valeurs positives. Mathématiquement, la nature discrète du nombre de relations extraconjugales quand $y > 0$ contredit l'hypothèse selon laquelle $y = \max(0, y^*)$ où y^* suit une distribution normale.

Question 17.4. Les écarts-types estimés ajustés sont les écarts-types estimés habituels de l'EMV de Poisson standard multipliés par $\hat{\sigma} = \sqrt{2} \approx 1{,}41$; les écarts-types ajustés seront donc environ 41 % plus grands. La statistique du quasi-*RV* est la statistique du *RV* divisée par $\hat{\sigma}^2$; ce sera donc la moitié de la statistique du *RV* habituelle.

Question 17.5. Par hypothèse, $mvp_i = \beta_0 + \mathbf{x}_i\beta + u_i$, où $\mathbf{x}_i\beta$ est une fonction linéaire des variables exogènes comme d'habitude. Le salaire observé est maintenant la plus grande valeur entre le salaire minimum et la valeur de la productivité marginale : $wage_i = \max(minwage_i, mvp_i)$, ce qui ressemble beaucoup à l'équation (17.34), sauf que l'opérateur du max a replacé l'opérateur du min.

F.17 CHAPITRE 18

Question 18.1. Nous pouvons mettre ces valeurs dans l'équation (18.1) et calculer leurs espérances. Tout d'abord, comme $z_s = 0$ pour tout $s < 0$, $y_{-1} = \alpha + u_{-1}$. De plus, $z_0 = 1$, donc $y_0 = \alpha + \delta_0 + u_0$. Pour $h \geq 1$, $y_h = \alpha + \delta_{h-1} + \delta_h + u_h$. Comme les erreurs sont d'espérance nulle, $E(y_{-1}) = \alpha$, $E(y_0) = \alpha + \delta_0$, et $E(y_h) = \alpha + \delta_{h-1} + \delta$ pour tout $h \geq 1$. Quand $h \to \infty$, $\delta_h \to 0$. Donc, $E(y_h) \to \alpha$ quand $h \to \infty$, c'est-à-dire l'espérance de y_h revient à l'espérance avant l'augmentation de z, à date zéro. Cela semble logique : bien que l'augmentation de z ait duré deux périodes, c'est toujours une augmentation temporaire.

Question 18.2. Dans le cas décrit, Δy_t et Δx_t sont des suites i.i.d., indépendantes l'une de l'autre. En particulier, Δy_t et Δx_t ne sont pas corrélées. Si $\hat{\gamma}_1$ est le paramètre de pente de la régression de Δy_t sur Δx_t, $t = 1, 2, \ldots, n$, alors plim $\hat{\gamma}_1 = 0$. C'est ce que nous voulons, puisque nous régressons un processus I(0) sur un autre processus I(0), et qu'ils ne sont pas corrélés. Écrivons l'équation $\Delta y_t = \gamma_0 + \gamma_1 \Delta x_t + e_t$, où $\gamma_0 = \gamma_1 = 0$. Comme $\{e_t\}$ est indépendant de $\{\Delta x_t\}$, l'hypothèse d'exogénéité stricte est valide. De plus, $\{e_t\}$ a une corrélation sérielle et est homoscédastique. Le théorème 11.2 du chapitre 11 implique donc que la statistique t de $\hat{\gamma}_1$ suit approximativement une loi normale standardisée. Si e_t suit une distribution normale, les hypothèses du modèle linéaire classique sont validées et la statistique t suit exactement une distribution de Student.

Question 18.3. Écrivons $x_t = x_{t-1} + a_t$ où $\{a_t\}$ est I(0). Par hypothèse, il y a une combinaison linéaire qui est I(0) ; appelons-la $s_t = y_t - \beta x_t$. Donc, $y_t - \beta x_{t-1} = y_t - \beta(x_t - a_t) = s_t + \beta a_t$. Comme s_t et a_t sont I(0) par hypothèse, $s_t + \beta a_t$ l'est aussi.

Question 18.4. Supposons l'homoscédasticité et écrivons la forme en somme des carrés des résidus du test *F*. La SCR restreinte s'obtient en régressant $\Delta hy6_t - \Delta hy3_{t-1} + (hy6_{t-1} - hy3_{t-2})$ sur une constante. Remarquez qu'α_0 est le seul paramètre à estimer dans $\Delta hy6_t = \alpha_0 + \gamma_0 \Delta hy3_{t-1} + \delta(hy6_{t-1} - hy3_{t-2})$ quand on a imposé toutes les restrictions. La somme des carrés des résidus s'obtient à partir de l'équation (18.39).

Question 18.5. Nous estimons deux équations : $\hat{y}_t = \hat{\alpha} + \hat{\beta}t$ et $\hat{y}_t = \hat{\gamma} + \hat{\delta} year_t$. Nous pouvons obtenir la relation entre les paramètres en remarquant que $year_t = t + 49$. En substituant ceci dans la seconde équation, on trouve $\hat{y}_t = \hat{\gamma} + \hat{\delta}(t + 49) = (\hat{\gamma} + 49\hat{\delta}) + \hat{\delta}t$. En comparant avec la première équation, on trouve $\hat{\delta} = \hat{\beta}$

(c'est-à-dire que les pentes de t et $year_t$ sont les mêmes) et $\hat{\alpha} = \hat{\gamma} + 49\hat{\delta}$. D'une manière générale, si on utilise *year* plutôt que t, la constante changera mais pas la pente (vous pouvez le vérifier en utilisant une des bases de données de séries temporelles, comme HSEINV ou INVEN). Que nous utilisions t ou une mesure de *year* ne change pas les valeurs prédites et ne change bien entendu pas les prédictions des valeurs futures. La constante s'ajuste simplement aux différentes manières d'ajouter une tendance dans la régression.

ANNEXE

G

TABLES STATISTIQUES

Tableau G.1 *Surface sous la distribution normale standardisée*

z	0	1	2	3	4	5	6	7	8	9
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359

(*suite*)

Tableau G.1 (*suite*)

z	0	1	2	3	4	5	6	7	8	9
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Exemples : Si $Z \sim N(0,1)$, alors $P(2 \leq -1,32) = 0,0934$ et $P(2 \leq 1,84) = 0,9671$.

Source : Cette table a été produite à l'aide de la fonction normprob du logiciel Stata®.

Tableau G.2 *Valeurs critiques de la distribution* t

		Niveau de signification				
Unilatéral :		**.10**	**.05**	**.025**	**.01**	**.005**
Bilatéral :		**.20**	**.10**	**.05**	**.02**	**.01**
Degrés de liberté *(ddl)*	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	21	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763

Tableau G.2 (*suite*)

		Niveau de signification				
Unilatéral :		**.10**	**.05**	**.025**	**.01**	**.005**
Bilatéral :		**.20**	**.10**	**.05**	**.02**	**.01**
	29	1.311	1.699	2.045	2.462	2.756
	30	1.310	1.697	2.042	2.457	2.750
	40	1.303	1.684	2.021	2.423	2.704
	60	1.296	1.671	2.000	2.390	2.660
	90	1.291	1.662	1.987	2.368	2.632
	120	1.289	1.658	1.980	2.358	2.617
	∞	1.282	1.645	1.960	2.326	2.576

Exemples : La valeur critique à un niveau de signification égal à 1 % dans un test unilatéral (à droite) avec 25 degrés de liberté est égale à 2,485. La valeur critique à 5 % dans un test bilatéral dont le nombre de degrés de liberté est élevé (*ddl* > 120) est égale à 1,96.

Source : Cette table a été produite à l'aide de la fonction invttail du logiciel Stata®.

Tableau G.3a *Valeurs critiques de la distribution F pour un niveau de signification égal à 10 %*

Degrés de liberté (*ddl*) du dénominateur	Degrés de liberté (*ddl*) du numérateur									
	1	2	3	4	5	6	7	8	9	10
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60

Exemple : la valeur critique à 10 % avec des *ddl* du numérateur = 2 et des *ddl* du dénominateur = 40, est égale à 2,44.

Source : cette table a été produite à l'aide de la fonction invFtail de logiciel Stata®.

Tableau G.3b *Valeurs critiques de la distribution F pour un niveau de signification égal à 5 %*

Degrés de liberté (*ddl*) du dénominateur	Degrés de liberté (*ddl*) du numérateur									
	1	2	3	4	5	6	7	8	9	10
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Exemple : la valeur critique à 5 % avec des *ddl* du numérateur = 4 & des *ddl* du dénominateur très élevés (∞), est égale à 2,37.

Source : cette table a été produite à l'aide de la fonction invFtail du logiciel Stata®.

Tableau G.3c *Valeurs critiques de la distribution F pour un seuil de signification égal à 1 %*

Degrés de liberté (*ddl*) du dénominateur	Degrés de liberté (*ddl*) du numérateur									
	1	2	3	4	5	6	7	8	9	10
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61	2.52
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Exemple : la valeur critique à 1 % pour des *ddl* du numérateur = 3 et des *ddl* du dénominateur = 60, est égale à 4,13.

Source : cette table a été produite à l'aide de la fonction invFtail du logiciel Stata®.

Tableau G.4 *Valeurs critiques de la distribution du χ^2*

Degrés de liberté (ddl)	Niveau de signification		
	.10	.05	.01
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.42	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

Exemple : la valeur critique à 5 %, pour laquelle *ddl* = 8, est égale à 15,51.

Source : cette table a été produite à l'aide de la fonction invchi2tail du logiciel Stata®.

RÉFÉRENCES

Angrist, J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review* 80, 313–336.

Angrist, J. D., and A. B. Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106, 979–1014.

Ashenfelter, O., and A. B. Krueger (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review* 84, 1157–1173.

Averett, S., and S. Korenman (1996), "The Economic Reality of the Beauty Myth," *Journal of Human Resources* 31, 304–330.

Ayres, I., and S. D. Levitt (1998), "Measuring Positive Externalities from Unobservable Victim Precaution: An Empirical Analysis of Lojack," *Quarterly Journal of Economics* 108, 43–77.

Banerjee, A., J. Dolado, J. W. Galbraith, and D. F. Hendry (1993), *Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data*. Oxford: Oxford University Press.

Bartik, T. J. (1991), "The Effects of Property Taxes and Other Local Public Policies on the Intrametropolitan Pattern of Business Location," in *Industry Location and Public Policy*, ed. H. W. Herzog and A. M. Schlottmann, 57–80. Knoxville: University of Tennessee Press.

Becker, G. S. (1968), "Crime and Punishment: An Economic Approach," *Journal of Political Economy* 76, 169–217.

Belsley, D., E. Kuh, and R. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Berk, R. A. (1990), "A Primer on Robust Regression," in *Modern Methods of Data Analysis*, ed. J. Fox and J. S. Long, 292–324. Newbury Park, CA: Sage Publications.

Betts, J. R. (1995), "Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth," *Review of Economics and Statistics* 77, 231–250.

Biddle, J. E., and D. S. Hamermesh (1990), "Sleep and the Allocation of Time," *Journal of Political Economy* 98, 922–943.

Biddle, J. E., and D. S. Hamermesh (1998), "Beauty, Productivity, and Discrimination: Lawyers' Looks and Lucre," *Journal of Labor Economics* 16, 172–201.

Blackburn, M., and D. Neumark (1992), "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials," *Quarterly Journal of Economics* 107, 1421–1436.

Blinder, A. S. and M. W. Watson (2014), "Presidents and the U.S. Economy: An Econometric Exploration," National Bureau of Economic Research Working Paper No. 20324.

Blomström, M., R. E. Lipsey, and M. Zejan (1996), "Is Fixed Investment the Key to Economic Growth?" *Quarterly Journal of Economics* 111, 269–276.

Blundell, R., A. Duncan, and K. Pendakur (1998), "Semiparametric Estimation and Consumer Demand," *Journal of Applied Econometrics* 13, 435–461.

Bollerslev, T., R. Y. Chou, and K. F. Kroner (1992), "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence," *Journal of Econometrics* 52, 5–59.

Bollerslev, T., R. F. Engle, and D. B. Nelson (1994), "ARCH Models," in *Handbook of Econometrics*, volume 4, chapter 49, ed. R. F. Engle and D. L. McFadden, 2959–3038. Amsterdam: North-Holland.

Bound, J., D. A. Jaeger, and R. M. Baker (1995), "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and Endogenous Explanatory Variables Is Weak," *Journal of the American Statistical Association* 90, 443–450.

Breusch, T. S., and A. R. Pagan (1979), "A Simple Test for Heteroskedasticity and Random Coefficient Variation," *Econometrica* 47, 987–1007.

Cameron, A. C., and P. K. Trivedi (1998), *Regression Analysis of Count Data.* Cambridge: Cambridge University Press.

Campbell, J. Y., and N. G. Mankiw (1990), "Permanent Income, Current Income, and Consumption," *Journal of Business and Economic Statistics* 8, 265–279.

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, ed. L. N. Christophides, E. K. Grant, and R. Swidinsky, 201–222. Toronto: University of Toronto Press.

Card, D., and A. Krueger (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy* 100, 1–40.

Castillo-Freeman, A. J., and R. B. Freeman (1992), "When the Minimum Wage Really Bites: The Effect of the U.S.-Level Minimum on Puerto Rico," in *Immigration and the Work Force*, ed. G. J. Borjas and R. B. Freeman, 177–211. Chicago: University of Chicago Press.

Clark, K. B. (1984), "Unionization and Firm Performance: The Impact on Profits, Growth, and Productivity," *American Economic Review* 74, 893–919.

Cloninger, D. O. (1991), "Lethal Police Response as a Crime Deterrent: 57-City Study Suggests a Decrease in Certain Crimes," *American Journal of Economics and Sociology* 50, 59–69.

Cloninger, D. O., and L. C. Sartorius (1979), "Crime Rates, Clearance Rates and Enforcement Effort: The Case of Houston, Texas," *American Journal of Economics and Sociology* 38, 389–402.

Cochrane, J. H. (1997), "Where Is the Market Going? Uncertain Facts and Novel Theories," *Economic Perspectives* 21, Federal Reserve Bank of Chicago, 3–37.

Cornwell, C., and W. N. Trumbull (1994), "Estimating the Economic Model of Crime Using Panel Data," *Review of Economics and Statistics* 76, 360–366.

Craig, B. R., W. E. Jackson III, and J. B. Thomson (2007), "Small Firm Finance, Credit Rationing, and the Impact of SBA-Guaranteed Lending on Local Economic Growth," *Journal of Small Business Management* 45, 116–132.

Currie, J. (1995), *Welfare and the Well-Being of Children*. Chur, Switzerland: Harwood Academic Publishers.

Currie, J., and N. Cole (1993), "Welfare and Child Health: The Link between AFDC Participation and Birth Weight," *American Economic Review* 83, 971–983.

Currie, J., and D. Thomas (1995), "Does Head Start Make a Difference?" *American Economic Review* 85, 341–364.

Davidson, R., and J. G. MacKinnon (1981), "Several Tests of Model Specification in the Presence of Alternative Hypotheses," *Econometrica* 49, 781–793.

Davidson, R., and J. G. MacKinnon (1993), *Estimation and Inference in Econometrics*. New York: Oxford University Press.

De Long, J. B., and L. H. Summers (1991), "Equipment Investment and Economic Growth," *Quarterly Journal of Economics* 106, 445–502.

Dickey, D. A., and W. A. Fuller (1979), "Distributions of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association* 74, 427–431.

Diebold, F. X. (2001), *Elements of Forecasting*. 2nd ed. Cincinnati: South-Western.

Downes, T. A., and S. M. Greenstein (1996), "Understanding the Supply Decisions of Nonprofits: Modeling the Location of Private Schools," *Rand Journal of Economics* 27, 365–390.

Draper, N., and H. Smith (1981), *Applied Regression Analysis*. 2nd ed. New York: Wiley.

Duan, N. (1983), "Smearing Estimate: A Nonparametric Retransformation Method," *Journal of the American Statistical Association* 78, 605–610.

Durbin, J. (1970), "Testing for Serial Correlation in Least Squares Regressions when Some of the Regressors Are Lagged Dependent Variables," *Econometrica* 38, 410–421.

Durbin, J., and G. S. Watson (1950), "Testing for Serial Correlation in Least Squares Regressions I," *Biometrika* 37, 409–428.

Eicker, F. (1967), "Limit Theorems for Regressions with Unequal and Dependent Errors," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 59–82. Berkeley: University of California Press.

Eide, E. (1994), *Economics of Crime: Deterrence and the Rational Offender*. Amsterdam: North-Holland.

Engle, R. F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica* 50, 987–1007.

Engle, R. F., and C. W. J. Granger (1987), "Cointegration and Error Correction: Representation, Estimation, and Testing," *Econometrica* 55, 251–276.

Evans, W. N., and R. M. Schwab (1995), "Finishing High School and Starting College: Do Catholic Schools Make a Difference?" *Quarterly Journal of Economics* 110, 941–974.

Fair, R. C. (1996), "Econometrics and Presidential Elections," *Journal of Economic Perspectives* 10, 89–102.

Franses, P. H., and R. Paap (2001), *Quantitative Models in Marketing Research*. Cambridge: Cambridge University Press.

Freeman, D. G. (2007), "Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws," *Contemporary Economic Policy* 25, 293–308.

Friedman, B. M., and K. N. Kuttner (1992), "Money, Income, Prices, and Interest Rates," *American Economic Review* 82, 472–492.

Geronimus, A. T., and S. Korenman (1992), "The Socioeconomic Consequences of Teen Childbearing Reconsidered," *Quarterly Journal of Economics* 107, 1187–1214.

Goldberger, A. S. (1991), *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Graddy, K. (1995), "Testing for Imperfect Competition at the Fulton Fish Market," *Rand Journal of Economics* 26, 75–92.

Graddy, K. (1997), "Do Fast-Food Chains Price Discriminate on the Race and Income Characteristics of an Area?" *Journal of Business and Economic Statistics* 15, 391–401.

Granger, C. W. J., and P. Newbold (1974), "Spurious Regressions in Econometrics," *Journal of Econometrics* 2, 111–120.

Greene, W. (1997), *Econometric Analysis*. 3rd ed. New York: MacMillan.

Griliches, Z. (1957), "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics* 39, 8–20.

Grogger, J. (1990), "The Deterrent Effect of Capital Punishment: An Analysis of Daily Homicide Counts," *Journal of the American Statistical Association* 410, 295–303.

Grogger, J. (1991), "Certainty vs. Severity of Punishment," *Economic Inquiry* 29, 297–309.

Hall, R. E. (1988), "The Relation between Price and Marginal Cost in U.S. Industry," *Journal of Political Economy* 96, 921–948.

Hamermesh, D. S., and J. E. Biddle (1994), "Beauty and the Labor Market," *American Economic Review* 84, 1174–1194.

Hamermesh, D. H., and A. Parker (2005), "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity," *Economics of Education Review* 24, 369–376.

Hamilton, J. D. (1994), *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hansen, C.B. (2007), "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T Is Large," Journal of Econometrics 141, 597–620.

Hanushek, E. (1986), "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature* 24, 1141–1177.

Harvey, A. (1990), *The Econometric Analysis of Economic Time Series*. 2nd ed. Cambridge, MA: MIT Press.

Hausman, J. A. (1978), "Specification Tests in Econometrics," *Econometrica* 46, 1251–1271.

Hausman, J. A., and D. A. Wise (1977), "Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica* 45, 319–339.

Hayasyi, F. (2000), *Econometrics*. Princeton, NJ: Princeton University Press.

Heckman, J. J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5, 475–492.

Herrnstein, R. J., and C. Murray (1994), *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Hersch, J., and L. S. Stratton (1997), "Housework, Fixed Effects, and Wages of Married Workers," *Journal of Human Resources* 32, 285–307.

Hines, J. R. (1996), "Altered States: Taxes and the Location of Foreign Direct Investment in America," *American Economic Review* 86, 1076–1094.

Holzer, H. (1991), "The Spatial Mismatch Hypothesis: What Has the Evidence Shown?" *Urban Studies* 28, 105–122.

Holzer, H., R. Block, M. Cheatham, and J. Knott (1993), "Are Training Subsidies Effective? The Michigan Experience," *Industrial and Labor Relations Review* 46, 625–636.

Horowitz, J. (2001), "The Bootstrap," in *Handbook of Econometrics*, volume 5, chapter 52, ed. E. Leamer and J. L. Heckman, 3159–3228. Amsterdam: North Holland.

Hoxby, C. M. (1994), "Do Private Schools Provide Competition for Public Schools?" National Bureau of Economic Research Working Paper Number 4978.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 221–233. Berkeley: University of California Press.

Hunter, W. C., and M. B. Walker (1996), "The Cultural Affinity Hypothesis and Mortgage Lending Decisions," *Journal of Real Estate Finance and Economics* 13, 57–70.

Hylleberg, S. (1992), *Modelling Seasonality*. Oxford: Oxford University Press.

Kane, T. J., and C. E. Rouse (1995), "Labor-Market Returns to Two- and Four-Year Colleges," *American Economic Review* 85, 600–614.

Kiefer, N. M., and T. J. Vogelsang (2005), "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory* 21, 1130–1164.

Kiel, K. A., and K. T. McClain (1995), "House Prices during Siting Decision Stages: The Case of an Incinerator from Rumor through Operation," *Journal of Environmental Economics and Management* 28, 241–255.

Kleck, G., and E. B. Patterson (1993), "The Impact of Gun Control and Gun Ownership Levels on Violence Rates," *Journal of Quantitative Criminology* 9, 249–287.

Koenker, R. (1981), "A Note on Studentizing a Test for Heteroskedasticity," *Journal of Econometrics* 17, 107–112.

Koenker, R. (2005), *Quantile Regression*. Cambridge: Cambridge University Press.

Korenman, S., and D. Neumark (1991), "Does Marriage Really Make Men More Productive?" *Journal of Human Resources* 26, 282–307.

Korenman, S., and D. Neumark (1992), "Marriage, Motherhood, and Wages," *Journal of Human Resources* 27, 233–255.

Krueger, A. B. (1993), "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984–1989," *Quarterly Journal of Economics* 108, 33–60.

Krupp, C. M., and P. S. Pollard (1996), "Market Responses to Antidumping Laws: Some Evidence

from the U.S. Chemical Industry," *Canadian Journal of Economics* 29, 199–227.

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin (1992), "Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?" *Journal of Econometrics* 54, 159–178.

Lalonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76, 604–620.

Larsen, R. J., and M. L. Marx (1986), *An Introduction to Mathematical Statistics and Its Applications.* 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.

Leamer, E. E. (1983), "Let's Take the Con Out of Econometrics," *American Economic Review* 73, 31–43.

Levine, P. B., A. B. Trainor, and D. J. Zimmerman (1996), "The Effect of Medicaid Abortion Funding Restrictions on Abortions, Pregnancies, and Births," *Journal of Health Economics* 15, 555–578.

Levine, P. B., and D. J. Zimmerman (1995), "The Benefit of Additional High-School Math and Science Classes for Young Men and Women," *Journal of Business and Economics Statistics* 13, 137–149.

Levitt, S. D. (1994), "Using Repeat Challengers to Estimate the Effect of Campaign Spending on Election Outcomes in the U.S. House," *Journal of Political Economy* 102, 777–798.

Levitt, S. D. (1996), "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Legislation," *Quarterly Journal of Economics* 111, 319–351.

Little, R. J. A. and D. B. Rubin (2002), Statistical Analysis with Missing Data. 2nd ed. Wiley: New York.

Low, S. A., and L. R. McPheters (1983), "Wage Differentials and the Risk of Death: An Empirical Analysis," *Economic Inquiry* 21, 271–280.

Lynch, L. M. (1992), "Private Sector Training and the Earnings of Young Workers," *American Economic Review* 82, 299–312.

MacKinnon, J. G., and H. White (1985), "Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics* 29, 305–325.

Maloney, M. T., and R. E. McCormick (1993), "An Examination of the Role that Intercollegiate Athletic Participation Plays in Academic Achievement: Athletes' Feats in the Classroom," *Journal of Human Resources* 28, 555–570.

Mankiw, N. G. (1994), *Macroeconomics.* 2nd ed. New York: Worth.

Mark, S. T., T. J. McGuire, and L. E. Papke (2000), "The Influence of Taxes on Employment and Population Growth: Evidence from the Washington, D.C. Metropolitan Area," *National Tax Journal* 53, 105–123.

McCarthy, P. S. (1994), "Relaxed Speed Limits and Highway Safety: New Evidence from California," *Economics Letters* 46, 173–179.

McClain, K. T., and J. M. Wooldridge (1995), "A Simple Test for the Consistency of Dynamic Linear Regression in Rational Distributed Lag Models," *Economics Letters* 48, 235–240.

McCormick, R. E., and M. Tinsley (1987), "Athletics versus Academics: Evidence from SAT Scores," *Journal of Political Economy* 95, 1103–1116.

McFadden, D. L. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. P. Zarembka, 105–142. New York: Academic Press.

Meyer, B. D. (1995), "Natural and Quasi-Experiments in Economics," *Journal of Business and Economic Statistics* 13, 151–161.

Meyer, B. D., W. K. Viscusi, and D. L. Durbin (1995), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review* 85, 322–340.

Mizon, G. E., and J. F. Richard (1986), "The Encompassing Principle and Its Application to Testing Nonnested Hypotheses," *Econometrica* 54, 657–678.

Mroz, T. A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica* 55, 765–799.

Mullahy, J., and P. R. Portney (1990), "Air Pollution, Cigarette Smoking, and the Production of Respiratory Health," *Journal of Health Economics* 9, 193–205.

Mullahy, J., and J. L. Sindelar (1994), "Do Drinkers Know When to Say When? An Empirical Analysis of Drunk Driving," *Economic Inquiry* 32, 383–394.

Netzer, D. (1992), "Differences in Reliance on User Charges by American State and Local Governments," *Public Finance Quarterly* 20, 499–511.

Neumark, D. (1996), "Sex Discrimination in Restaurant Hiring: An Audit Study," *Quarterly Journal of Economics* 111, 915–941.

Neumark, D., and W. Wascher (1995), "Minimum Wage Effects on Employment and School Enrollment," *Journal of Business and Economic Statistics* 13, 199–206.

Newey, W. K., and K. D. West (1987), "A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55, 703–708.

Papke, L. E. (1987), "Subnational Taxation and Capital Mobility: Estimates of Tax-Price Elasticities," *National Tax Journal* 40, 191–203.

Papke, L. E. (1994), "Tax Policy and Urban Development: Evidence from the Indiana Enterprise Zone Program," *Journal of Public Economics* 54, 37–49.

Papke, L. E. (1995), "Participation in and Contributions to 401(k) Pension Plans: Evidence from Plan Data," *Journal of Human Resources* 30, 311–325.

Papke, L. E. (1999), "Are 401(k) Plans Replacing Other Employer-Provided Pensions? Evidence from Panel Data," *Journal of Human Resources*, 34, 346–368.

Papke, L. E. (2005), "The Effects of Spending on Test Pass Rates: Evidence from Michigan," *Journal of Public Economics* 89, 821–839.

Papke, L. E. and J. M. Wooldridge (1996), "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates," Journal of Applied Econometrics 11, 619–632.

Park, R. (1966), "Estimation with Heteroskedastic Error Terms," *Econometrica* 34, 888.

Peek, J. (1982), "Interest Rates, Income Taxes, and Anticipated Inflation," *American Economic Review* 72, 980–991.

Pindyck, R. S., and D. L. Rubinfeld (1992), *Microeconomics*. 2nd ed. New York: Macmillan.

Ram, R. (1986), "Government Size and Economic Growth: A New Framework and Some Evidence from Cross-Section and Time-Series Data," *American Economic Review* 76, 191–203.

Ramanathan, R. (1995), *Introductory Econometrics with Applications*. 3rd ed. Fort Worth: Dryden Press.

Ramey, V. (1991), "Nonconvex Costs and the Behavior of Inventories," *Journal of Political Economy* 99, 306–334.

Ramsey, J. B. (1969), "Tests for Specification Errors in Classical Linear Least-Squares Analysis," *Journal of the Royal Statistical Association*, Series B, 71, 350–371.

Romer, D. (1993), "Openness and Inflation: Theory and Evidence," *Quarterly Journal of Economics* 108, 869–903.

Rose, N. L. (1985), "The Incidence of Regulatory Rents in the Motor Carrier Industry," *Rand Journal of Economics* 16, 299–318.

Rose, N. L., and A. Shepard (1997), "Firm Diversification and CEO Compensation: Managerial Ability or Executive Entrenchment?" *Rand Journal of Economics* 28, 489–514.

Rouse, C. E. (1998), "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *Quarterly Journal of Economics* 113, 553–602.

Sander, W. (1992), "The Effect of Women's Schooling on Fertility," *Economic Letters* 40, 229–233.

Savin, N. E., and K. J. White (1977), "The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors," *Econometrica* 45, 1989–1996.

Shea, J. (1993), "The Input-Output Approach to Instrument Selection," *Journal of Business and Economic Statistics* 11, 145–155.

Shughart, W. F., and R. D. Tollison (1984), "The Random Character of Merger Activity," *Rand Journal of Economics* 15, 500–509.

Solon, G. (1985), "The Minimum Wage and Teenage Employment: A Re-analysis with Attention to Serial Correlation and Seasonality," *Journal of Human Resources* 20, 292–297.

Staiger, D., and J. H. Stock (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica* 65, 557–586.

Stigler, S. M. (1986), *The History of Statistics*. Cambridge, MA: Harvard University Press.

Stock, J. H., and M. W. Watson (1989), "Interpreting the Evidence on Money-Income Causality," *Journal of Econometrics* 40, 161–181.

Stock, J. H., and M. W. Watson (1993), "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica* 61, 783–820.

Stock, J. H. and M. Yogo (2005), "Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments," in Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg, ed. D. W. K. Andrews and J. H. Stock, 109–120. Cambridge: Cambridge University Press.

Stock, J. W. and M. W. Watson (2008), "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression," Econometrica 76, 155–174.

Sydsaeter, K., and P. J. Hammond (1995), *Mathematics for Economic Analysis*. Englewood Cliffs, NJ: Prentice Hall.

Terza, J. V. (2002), "Alcohol Abuse and Employment: A Second Look," *Journal of Applied Econometrics* 17, 393–404.

Tucker, I. B. (2004), "A Reexamination of the Effect of Big-time Football and Basketball Success on Graduation Rates and Alumni Giving Rates," *Economics of Education Review* 23, 655–661.

Vella, F., and M. Verbeek (1998), "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men," *Journal of Applied Econometrics* 13, 163–183.

Wald, A. (1940), "The Fitting of Straight Lines if Both Variables Are Subject to Error," *Annals of Mathematical Statistics* 11, 284–300.

Wallis, K. F. (1972), "Testing for Fourth-Order Autocorrelation in Quarterly Regression Equations," *Econometrica* 40, 617–636.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48, 817–838.

White, H. (1984), *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

White, M. J. (1986), "Property Taxes and Firm Location: Evidence from Proposition 13," in *Studies in State and Local Public Finance*, ed. H. S. Rosen, 83–112. Chicago: University of Chicago Press.

Whittington, L. A., J. Alm, and H. E. Peters (1990), "Fertility and the Personal Exemption: Implicit Pronatalist Policy in the United States," *American Economic Review* 80, 545–556.

Wooldridge, J. M. (1989), "A Computationally Simple Heteroskedasticity and Serial Correlation-Robust Standard Error for the Linear Regression Model," *Economics Letters* 31, 239–243.

Wooldridge, J. M. (1991a), "A Note on Computing *R*-Squared and Adjusted *R*-Squared for Trending and Seasonal Data," *Economics Letters* 36, 49–54.

Wooldridge, J. M. (1991b), "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances," *Journal of Econometrics* 47, 5–46.

Wooldridge, J. M. (1994a), "A Simple Specification Test for the Predictive Ability of Transformation Models," *Review of Economics and Statistics* 76, 59–65.

Wooldridge, J. M. (1994b), "Estimation and Inference for Dependent Processes," in *Handbook of Econometrics*, volume 4, chapter 45, ed. R. F. Engle and D. L. McFadden, 2639–2738. Amsterdam: North-Holland.

Wooldridge, J. M. (1995), "Score Diagnostics for Linear Models Estimated by Two Stage Least Squares," in *Advances in Econometrics and Quantitative Economics*, ed. G. S. Maddala, P. C. B. Phillips, and T. N. Srinivasan, 66–87. Oxford: Blackwell.

Wooldridge, J.M. (2001), "Diagnostic Testing," in Companion to Theoretical Econometrics, ed. B. H. Baltagi, 180–200. Oxford: Blackwell.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

GLOSSAIRE

A

Absence de convergence. *Voir* non-convergence.

Aléa. Perturbation aléatoire. *Voir* terme d'erreur.

Analyse de durée. *Voir* analyse de survie.

Analyse de la spécification d'un modèle. Processus d'évaluation de la spécification d'un modèle, dont l'objectif est d'identifier la présence d'estimateurs biaisés induite par l'omission de variables importantes, la présence d'erreur de mesure, la détermination simultanée des variables, ou toutes autres sources de mauvaise spécification.

Analyse de politiques (économique). Analyse empirique s'appuyant sur des méthodes économétriques, dont l'objectif est d'évaluer les effets de prises de décision discrétionnaires, que ce soit dans la sphère publique ou dans celle de l'entreprise, par exemple.

Analyse de Régression Linéaire Multiple (régression multiple). Analyse visant à étudier les résultats de l'estimation d'un modèle de régression linéaire multiple, que ce soit sur le plan statistique ou économique. *Voir aussi* Régression Linéaire Multiple.

Analyse de sensibilité. Processus visant à vérifier si les effets estimés et la signification statistique des variables explicatives d'intérêt sont sensibles à différents changements, tels que l'inclusion d'une variable explicative supplémentaire, l'utilisation d'une autre forme fonctionnelle, la suppression d'observations isolées, ou l'utilisation de différentes méthodes d'estimation.

Analyse des résidus. Analyse effectuée après l'estimation d'un modèle, dont l'objectif est d'étudier le signe et la taille des résidus, souvent pour des observations particulières.

Analyse de survie. Application d'un modèle de régression sur variables censurées, dans lequel la variable dépendante mesure le temps écoulé précédant l'occurrence d'un événement, tel que le temps qu'une personne en recherche d'emploi doit attendre avant de trouver du travail.

Analyse empirique. Étude reposant sur une analyse économétrique formelle visant à tester une théorie, estimer une relation, ou déterminer l'efficacité d'une décision économique.

Asymptotiquement efficace. Se dit de l'estimateur dont la variance est la plus faible parmi tous les estimateurs convergents qui suivent asymptotiquement une loi normale.

Asymptotiquement non corrélée. *Voir* faiblement dépendante.

Autocorrélation. *Voir* corrélation sérielle.

Autocorrélation AR(1). Autocorrélation qui correspond au coefficient de corrélation pour les paires d'observations consécutives dans une série chronologique.

Autocorrélation d'ordre 1 (*ou* du premier ordre). *Voir* autocorrélation AR(1).

Auto-sélection. Décision d'effectuer (ou pas) une action basée sur les bénéfices (ou inconvénients) attendus de la (non-) réalisation de cette action.

B

Bases de données en ligne. Bases de données accessibles sur un réseau informatique, comme internet.

Biais. Différence entre la valeur espérée d'un estimateur et la valeur (vraie) dans la population que cet estimateur est censé estimer.

Biais asymptotique. *Voir* non-convergence.

Biais d'atténuation. *Voir* biais vers zéro.

Biais de sélection de l'échantillon. Biais de l'estimateur des MCO lorsque les données proviennent d'une sélection endogène de l'échantillon.

Biais de simultanéité. Biais de l'estimateur des MCO lorsqu'il s'agit d'estimer une équation appartenant à un système d'équations simultanées.

Biais d'hétérogénéité. Biais de l'estimateur des MCO lorsque des données empilées sont utilisées et qu'il existe une hétérogénéité en coupe transversale (autrement dit, lorsque des variables explicatives constantes dans le temps sont omises dans la régression).

Biais d'omission de variable. Biais de l'estimateur des MCO lorsqu'une variable pertinente n'est pas incluse dans la régression.

Biais vers le bas. Biais d'un estimateur lorsque sa valeur espérée est inférieure à la valeur (vraie) du paramètre dans la population.

Biais vers le haut. Biais d'un estimateur lorsque sa valeur espérée est supérieure à la valeur (vraie) du paramètre dans la population.

Biais vers zéro (ou vers l'origine). Biais d'un estimateur lorsque l'espérance de la valeur absolue de cet estimateur est plus petite que la valeur absolue du paramètre dans la population.

BLUE. *Voir* meilleur estimateur linéaire et sans biais.

Bootstrap. Méthode de ré-échantillonnage qui permet de construire des échantillons, tirés de manière aléatoire et avec remise, à partir de l'échantillon initial contenant les données originales.

C

Catégorie de référence. Catégorie représentée par la constante dans un modèle de régression multiple où des variables indicatrices sont utilisées comme variables explicatives.

Causalité de Granger. Notion de causalité reposant sur l'idée que les valeurs passées d'une série (x_t) sont utiles pour prédire les valeurs futures d'une autre série (y_t), une fois prises en compte les valeurs passées de cette série (y_t).

Ceteris paribus. Locution latine se traduisant par « toutes choses (pertinentes) (étant) égales par ailleurs », « étant donné toutes les autres variables explicatives », ou encore « tous les autres facteurs pertinents étant fixes ».

Coefficient d'aplatissement. Mesure de l'épaisseur des extrémités de la distribution d'une variable aléatoire, qui est basée sur le moment d'ordre quatre de la variable centrée et réduite et qui est habituellement comparée au coefficient d'aplatissement (égal à trois) de la distribution normale standard.

Coefficient de corrélation. Mesure de dépendance linéaire entre deux variables aléatoires, qui ne dépend pas des unités de mesure et qui est bornée entre – 1 et 1.

Coefficient de corrélation de l'échantillon. Estimation du coefficient de corrélation de la population, obtenue à partir d'un échantillon de données.

Coefficient de détermination. *Voir R* carré.

Coefficient d'asymétrie. Mesure de l'écart d'une distribution par rapport à la symétrie, reposant sur le troisième moment de la variable aléatoire standardisée.

Coefficients bêta. *Voir* coefficients standardisés.

Coefficients standardisés. Coefficient de la régression qui mesure la variation de la variable dépendante en nombre d'écart-type, suite à l'augmentation d'un écart-type d'une variable indépendante.

Cointégration. Cas dans lequel une combinaison linéaire de deux séries chronologiques, chacune intégrées d'ordre un, est intégrée d'ordre zéro.

Colinéarité parfaite. Cas dans lequel une variable indépendante est une fonction linéaire exacte d'une ou plusieurs autres variables explicatives.

Condition de rang. Condition suffisante pour l'identification d'un modèle dont une ou plusieurs variables explicatives sont endogènes.

Condition d'ordre. Condition nécessaire pour l'identification d'un modèle dont une ou plusieurs variables explicatives sont endogènes. Le nombre total de variables exogènes doit être au moins aussi grand que le nombre total de variables explicatives.

Conditionnellement aux variables. Étant donné, et quelles que soient, les valeurs prises par les variables.

Conditions du premier ordre. Ensemble d'équations linéaires utilisées pour résoudre le problème d'optimisation et identifier l'estimateur des MCO.

Conjointement significatifs. Se dit de plusieurs variables explicatives lorsque l'hypothèse nulle, selon laquelle leurs coefficients de population sont tous égaux à zéro, est rejetée à un niveau de signification donné.

Consistance. Anglicisme désignant la propriété d'un estimateur qui converge en probabilité *vers la valeur (vraie) du paramètre dans la population* lorsque la taille de l'échantillon grandit. *Voir aussi* convergence.

Constante. Dans l'équation d'une droite, valeur de la variable y quand la variable X est égale à zéro.

Contraintes d'exclusion. Contraintes sous lesquelles certaines variables sont exclues du modèle (ou sont associées à des coefficients dont la valeur vraie au niveau de la population est nulle).

Contraintes multiples. Plus d'une restriction sur les paramètres d'un modèle économétrique.

Convergence. Propriété d'un estimateur qui, lorsque la taille de l'échantillon grandit, converge en probabilité vers une valeur égale à la vraie valeur du paramètre dans la population lorsqu'il s'agit d'un estimateur sans biais asymptotique.

Correction de l'effet de grappe. Dans le cas des données de panel, action de corriger les écarts-types et statistiques de test en les rendant robustes à toute forme de corrélations sérielle (et d'hétéroscédasticité).

Corrélation de l'échantillon. Est égale à la covariance de l'échantillon divisée par le produit des écarts-types de l'échantillon, s'agissant des valeurs de deux variables aléatoires.

Corrélation fallacieuse. Corrélation entre deux variables qui ne s'explique pas par un lien de causalité, mais par leur dépendance éventuelle à l'égard d'un autre facteur non observé.

Corrélation sérielle AR(1). Corrélation de type AR(1) dans les erreurs d'un modèle de régression sur séries chronologiques.

Corrélation sérielle des erreurs. Corrélation entre les erreurs relatives à des périodes de temps différentes, dans un modèle sur séries chronologiques ou données de panel.

Coupes transversales indépendantes empilées. *Voir* données empilées.

Coupure à droite. Forme de censure de l'échantillon qui intervient lorsque la valeur d'une variable n'est reportée que jusqu'à un certain seuil, au-delà duquel nous savons seulement que la valeur est au moins aussi grande que le seuil.

Covariable. Variable qui joue un rôle explicatif dans un modèle, mais dont la variation n'est pas étudiée en tant que telle. *Voir* variable de contrôle.

Covariance. Mesure de dépendance linéaire entre deux variables aléatoires (dans la population).

Covariance de l'échantillon. Estimateur sans biais de la covariance entre deux variables aléatoires dans la population.

Covariance stationnaire. Anglicisme. *Voir* Stationnaire dans la covariance.

Covariate. Anglicisme. *Voir* covariable.

Critère de sélection hors-échantillon (ou à l'extérieur de l'échantillon). Critère utilisé pour sélectionner un modèle en fonction de la qualité de ses prévisions obtenues sur la partie de l'échantillon n'ayant pas servi à estimer les paramètres.

Critère propre à l'échantillon (ou à l'intérieur de l'échantillon). Critère utilisé pour sélectionner un modèle en fonction de la qualité de son ajustement aux données de l'échantillon servant à estimer les paramètres.

D

Data mining. Anglicisme. *Voir* fouille de données.

Date de référence. Date par rapport à laquelle les indices (comme les indices de prix ou de production) sont évalués aux autres dates.

Définie positive. Se dit d'une matrice symétrique lorsque toutes les formes quadratiques (ou bilinéaires) sont strictement positives (à l'exception de la forme triviale qui doit être nulle).

Degrés de liberté (*ddl*). Différence entre le nombre d'observations et le nombre de paramètres estimés.

Degrés de liberté au dénominateur. Nombre de degrés de liberté du modèle non contraint dans un test F.

Degrés de liberté du numérateur. Nombre de contraintes testées dans un test F.

Densité de probabilité. Fonction qui donne la probabilité associée à chaque valeur prise par une variable aléatoire discrète ; pour une variable aléatoire continue, la surface se trouvant sous la fonction donne la probabilité associée à différents événements.

Dépendance faible. Propriété d'un processus temporel dont la dépendance entre des variables aléatoires à deux moments dans le temps (mesurée par la corrélation, par exemple) diminue lorsque l'intervalle entre ces deux moments dans le temps augmente.

Dépendance forte. *Voir* hautement persistant.

Déplacement de l'ordonnée à l'origine. S'observe lorsque la constante d'un modèle de régression varie d'un groupe à l'autre ou d'une période à l'autre.

Dérivée. Pente d'une fonction continue, définie à l'aide du calcul différentiel.

Dérivée partielle. Pente d'une fonction continue de plusieurs variables, dans une direction donnée.

Désaisonnalisation. Traitement statistique éliminant l'influence des composantes saisonnières dans les séries chronologiques.

Detrending. Anglicisme. *Voir* filtrage de la tendance.

Différence de martingale : Première différence d'une martingale, qui est imprévisible (car sa moyenne est nulle, quelles que soient les valeurs passées).

Différence de pentes (ou entre pentes). Variation du paramètre de la pente en fonction d'un groupe ou d'une période temporelle.

Différences premières. Transformation d'une série chronologique qui consiste à calculer la différence entre des observations adjacentes, c'est-à-dire à soustraire la valeur la plus ancienne de la plus récente.

Distribution conditionnelle. Loi de probabilités suivie par une variable aléatoire, étant donné les valeurs prises par une ou plusieurs autres variables aléatoires.

Distribution conjointe. Distribution conjointe de deux variables (ou plus), qui dépend du degré de covariance entre les distributions de chacune de ces variables.

Distribution d'échantillonnage. Distribution de probabilité d'un estimateur pour toutes les valeurs possibles de l'échantillon.

Distribution de Dickey-Fuller. Distribution limite de la statistique t dans le cadre d'un test dont l'hypothèse nulle est celle de racine unitaire.

Distribution de Poisson. Distribution de probabilité d'une variable de comptage.

Distribution des retards. Dans un modèle à retards échelonnés finis ou infinis, représentation graphique des coefficients des variables retardées en fonction de la longueur des retards.

Distribution du chi(-)deux. Loi de probabilité obtenue en additionnant le carré de variables distribuées selon une loi normale centrée réduite, en notant que le nombre de termes dans la somme est égal au nombre de degrés de libertés de la distribution.

Distribution jointe. Distribution qui permet d'obtenir les probabilités des réalisations impliquant deux variables aléatoires (ou plus).

Distribution *F*. Distribution de probabilité obtenue en formant le ratio de deux variables indépendantes, chacune distribuées selon une loi chi-deux et divisées par leur nombre de degrés de libertés respectif.

Distribution normale centrée réduite. Distribution normale de moyenne nulle et de variance unitaire.

Distribution normale standard. Distribution normale dont la moyenne et la variance valent respectivement zéro et un.

Distribution symétrique. Distribution de probabilité caractérisée par une fonction de densité de probabilité symétrique autour de la valeur médiane correspondant également à la valeur moyenne (pour autant qu'elle existe).

Distribution *t* (ou distribution de Student). Distribution de probabilité d'une variable aléatoire correspondant au quotient entre une variable aléatoire normale centrée réduite et la racine carrée d'une variable aléatoire indépendante distribuée selon une loi du khi-carré et divisée par son nombre de degré de liberté.

Document texte (ASCII). Format universel de document qui peut être lu par différentes plateformes informatiques.

Données censurées. Situation dans laquelle les réalisations de la variable dépendante ne sont pas toutes directement observables en raison de l'existence d'un seuil au-delà ou en deçà duquel nous ne connaissons pas les valeurs précises que peut prendre cette variable, hormis l'information que ce seuil a été effectivement franchi.

Données (de panel) centrées sur (en écart à) leur moyenne temporelle. Données de panel avec moyenne temporelle retranchée. Plus précisément, données de panel pour lesquelles la moyenne temporelle (calculée sur l'ensemble des périodes) de chaque unité de coupe transversale a été soustraite de la valeur observée à chaque période.

Données de panel. Données construites à partir de coupes transversales observées à différents moments dans le temps. Le panel sera dit *cylindré* (ou *équilibré*) si le nombre d'unités d'observation dans la coupe transversale est inchangé au cours du temps. Dans le cas contraire, si certaines unités ne peuvent pas être observées à certaines périodes, on parlera de panel non *cylindré* (ou *non équilibré*).

Données de séries chronologiques (ou temporelles). Données recueillies pour une ou plusieurs variables à différents moments dans le temps.

Données empilées (ou regroupées). Données construites à partir de coupes transversales indépendantes les unes des autres, qui sont observées généralement à différents moments dans le temps et ensuite regroupées (ou empilées) dans une seule et même base de données. Les unités d'observation peuvent donc varier au cours du temps.

Données en coupe transversale. Données tirées d'une population à un moment donné dans le temps.

Données en quasi-écart à leur moyenne. Données de panel avec moyenne quasi-retranchée. Plus précisément, données de panel utilisées dans un modèle à

effets aléatoires, pour lesquelles on retranche, pour chaque unité d'observation et à chaque période, une fraction de la moyenne temporelle.

Données expérimentales. Données obtenues à l'issue d'une expérience dite « contrôlée ».

Données longitudinales. *Voir* données de panel.

Données manquantes. Problème de données qui apparaît lorsque nous n'observons pas de valeur pour certaines variables d'observations dans l'échantillon (qu'il s'agisse d'individus, villes, périodes de temps, etc.).

Données non-expérimentales. Données qui n'ont pas été obtenues à l'aide d'une expérience dite « contrôlée ».

Données observationnelles. *Voir* données non-expérimentales.

Données quasi-différenciées. Données obtenues en calculant la différence entre la valeur actuelle et la valeur observée à la période précédente, sachant que cette dernière est multipliée par le paramètre autorégressif du processus AR(1) caractérisant les erreurs du modèle de séries chronologiques utilisé sur les données brutes.

Données rétrospectives. Données reposant sur des informations passées, plutôt que sur des informations actuelles.

Droite de régression des MCO. Équation reliant les valeurs ajustées (ou prédites) de la variable dépendante à celles qui sont observées pour les variables indépendantes, sachant que les estimations des paramètres sont obtenues par les MCO.

E

Écart-type. Mesure traditionnelle de la dispersion de la distribution d'une variable aléatoire.

Écart-type asymptotique. Écart-type valable pour de grands échantillons.

Écart-type de $\hat{\beta}_j$. Mesure théorique classique de la dispersion de la distribution d'échantillonnage de $\hat{\beta}_j$.

Écart-type estimé par « bootstrap(ing) ». Écart-type estimé sur base d'un ensemble de « nouveaux échantillons » obtenus par tirage avec remise à partir de l'échantillon initial. *Voir* bootstrap.

Écart-type d'échantillonnage. Écart-type d'un estimateur, c'est-à-dire écart-type de la distribution d'échantillonnage.

Écart-type de la régression (ETR). Estimation de l'écart-type de l'erreur de la population, obtenue en calculant la racine carrée de la somme des carrés des résidus divisée par le nombre de degrés de liberté.

Écart-type de l'échantillon. Estimateur convergent de l'écart-type de la population.

Écart-type de l'estimation. *Voir* écart-type de la régression.

Écart-type estimé. Estimation de l'écart-type théorique d'un estimateur (en termes généraux).

Écart-type estimé de $\hat{\beta}_j$. Estimation de l'écart-type théorique de la distribution d'échantillonnage de $\hat{\beta}_j$.

Écart-type robuste à l'autocorrélation. Écart-type estimé qui est (asymptotiquement) valide même en présence de corrélation sérielle dans les erreurs.

Écart-type robuste à l'hétéroscédasticité. Écart-type estimé d'un estimateur qui est (asymptotiquement) valide même en présence d'hétéroscédasticité de forme inconnue dans les erreurs.

Écart-type théorique. *Voir* Écart-type de $\hat{\beta}_j$.

Échantillon aléatoire. Échantillon obtenu par échantillonnage (ou tirage) aléatoire à partir de la population d'origine.

Échantillon apparié. Échantillon dans lequel chaque observation est jumelée avec une autre observation présente dans un second échantillon, comme dans le cas d'un échantillon incluant les époux et apparié à celui incluant les épouses.

Échantillonnage aléatoire. Procédure d'échantillonnage selon laquelle chaque observation est issue d'un tirage aléatoire au sein de la population. Chaque observation a la même probabilité d'être sélectionnée lors d'un tirage, et chaque tirage est réalisé de manière indépendante.

Échantillonnage stratifié. Procédure d'échantillonnage non aléatoire qui consiste à diviser la population en plusieurs strates exhaustives qui ne se chevauchent pas, puis à tirer des échantillons aléatoires pour chaque strate.

Échantillon non aléatoire. Échantillon obtenu d'une autre façon que par échantillonnage aléatoire de la population d'intérêt.

Échantillon sélectionné. Échantillon non aléatoire de données obtenu par une sélection effectuée sur base de certaines caractéristiques (non) observées.

Échantillon par grappes. Échantillon obtenu en découpant la population en sous-populations (ou grappes) hétérogènes, puis en sélectionnant les grappes à retenir de manière aléatoire. Les grappes contiennent souvent des individus.

Éditeur de texte. Programme informatique utilisé pour éditer des documents de texte.

Effets aléatoires corrélés. Approche de l'analyse des données de panel dans laquelle la corrélation entre l'effet non observé et les variables explicatives est modélisée sous la forme d'une relation linéaire.

Effet causal. Variation *ceteris paribus* d'une variable qui entraîne la variation d'une autre. *Voir également ceteris paribus.*

Effet *ceteris paribus*. *Voir* effet causal.

Effet cumulé. En tout point du temps, variation d'une variable de réponse suite à une augmentation permanente d'une variable explicative – habituellement dans le contexte des modèles à retards distribués.

Effet de clustering. Anglicisme. *Voir* effet de grappe.

Effet de court terme. Dans un modèle à retards distribués, variation immédiate de la variable dépendante étant donné une augmentation d'une unité de la variable indépendante.

Effet de grappe. Effet non observé commun à toutes les unités (souvent des individus) de la même grappe.

Effet d'interaction. Dans la régression multiple, effet *ceteris paribus* (ou marginal) d'une variable explicative, qui dépend des valeurs prises par une autre variable explicative.

Effet fixe. *Voir* effet non observé.

Effet inobservé. *Voir* effet non observé.

Effet non observé. Dans le cadre d'un modèle de données de panel, effet lié à une variable non observée, incluse dans le terme d'erreur, qui ne varie pas au cours du temps. Dans le cadre d'échantillons en grappes, effet lié à une variable non observée commune à l'ensemble des unités de la grappe.

Effet marginal. Effet *ceteris paribus* sur la variable dépendante d'une petite variation de la variable indépendante.

Effet marginal décroissant. Effet marginal d'une variable explicative, qui décroît à mesure que la valeur de la variable en question augmente.

Effet marginal au point moyen (EMPM). Effet partiel qui est calculé dans le cadre d'un modèle dont les effets marginaux ne sont pas constants et qui est évalué en gardant fixes les variables explicatives au niveau de leur valeur moyenne.

Effet marginal moyen (EMM). Dans les modèles à effets partiels non constants, effet partiel qui est calculé sur base de la moyenne de l'effet partiel dans la population.

Effet partiel. Effet d'une variable explicative sur la variable dépendante, toutes choses égales par ailleurs. *Voir* effet causal.

Effet partiel moyen. *Voir* effet marginal moyen.

Effet propre. *Voir* effet *ceteris paribus*.

Élasticité. Variation en pourcentage d'une variable donnée, suite à l'augmentation de 1 pourcent d'une autre variable, toutes choses étant égales par ailleurs.

Élasticité à court terme. Incidence immédiate (de la variable dépendante suite à une augmentation d'une unité de la variable indépendante) dans un modèle à retards échelonnés dont les variables dépendante et indépendantes sont sous forme logarithmique.

Élasticité à long terme. Impact (ou propension) de long terme mesurée par la variation finale en pourcentage de la variable expliquée, étant donné une variation permanente de la variable explicative de 1 %, dans un modèle à retards échelonnés dont les variables dépendante et indépendantes sont sous forme logarithmique.

Endogénéité. Terme usité pour décrire la présence d'une variable explicative endogène.

Ensemble d'information. Ensemble des variables que nous pouvons observer avant de faire une prévision.

Équation (en forme) réduite. Équation linéaire pour laquelle une variable endogène est fonction d'un ensemble de variables explicatives et d'erreurs non observées.

Équation en différence première. Dans les modèles de séries chronologiques ou de données de panel, équation dont les variables dépendante et indépendantes sont toutes exprimées en différence première.

Équation identifiée. Équation dont les paramètres peuvent être estimés de manière consistante, en particulier pour les modèles avec des variables explicatives endogènes.

Équation juste identifiée. Équation dont au moins une variable explicative est endogène et dont les paramètres ne seraient pas identifiés s'il y avait une variable instrumentale en moins.

Équation non identifiée. Équation dont au moins une variable explicative est endogène et dont le nombre de variables instrumentales n'est pas suffisant pour en identifier les paramètres.

Équation structurelle. Équation provenant de la théorie économique ou d'un autre type de raisonnement économique moins formel.

Équation sur-identifiée. Équation incluant une ou plusieurs variables explicatives endogènes, dans laquelle le nombre de variables instrumentales est

strictement supérieur au nombre de variables explicatives endogènes.

Erreur Absolue Moyenne (EAM). Mesure de la performance prévisionnelle, calculée comme la moyenne des valeurs absolues des erreurs de prévision. De l'anglais, « Mean Absolute Error » (MAE).

Erreur classique dans les variables (ECV). Modèle d'erreur de mesure dans lequel l'erreur correspond à la somme de la vraie valeur et d'une erreur de mesure indépendante (ou au moins non corrélée).

Erreur de la forme réduite. Terme d'erreur apparaissant dans une équation de forme réduite.

Erreur de mesure. Différence entre la mesure observée d'une variable et sa mesure réelle.

Erreur de mesure multiplicative. Erreur de mesure qui provient du fait que la variable observée est le produit de la vraie variable non observée et d'une erreur de mesure positive.

Erreur de prévision. Différence entre le résultat observé et le résultat prédit.

Erreur de type I. Rejet de l'hypothèse nulle alors qu'elle est vraie.

Erreur de mesure sur les régresseurs. Situation dans laquelle la variable dépendante ou certaines des variables indépendantes sont sujettes à des erreurs de mesure.

Erreur de type II. Non-rejet de l'hypothèse nulle alors qu'elle est fausse.

Erreur idiosyncratique. Dans les modèles de données de panel, erreur qui varie à la fois dans le temps et entre les unités d'observation (individus, firmes, villes, etc.).

Erreur quadratique moyenne (MSE). Espérance de la distance au carré entre l'estimateur du paramètre et sa valeur vraie dans la population cible. Est égale à la variance de l'estimateur augmentée du carré du biais de l'estimateur.

Erreur structurelle. Erreur d'une équation structurelle qui peut appartenir à un modèle d'équations simultanées.

Erreurs non autocorrélées. Erreurs qui *ne* sont *pas* deux à deux corrélées à travers le temps, dans un modèle de séries chronologiques ou de données de panel.

Erreur type. Anglicisme. *Voir* écart-type estimé.

Espérance. *Voir* valeur espérée.

Espérance conditionnelle. Valeur espérée (ou attendue) d'une variable aléatoire, soit la variable expliquée, qui dépend des valeurs d'une ou de plusieurs autres variables, soit les variables explicatives.

Estimateur. Règle de calcul, applicable à tout échantillon de données, qui permet d'obtenir une valeur numérique pour le paramètre d'une population et dont la forme ne dépend pas de l'échantillon considéré.

Estimateur à effets aléatoires. Estimateur des moindres carrés quasi-généralisés, pour lequel l'effet non observé est supposé être non corrélé avec les variables explicatives à chaque période de temps.

Estimateur à effets fixes. Dans un modèle de données de panel à effets non observés, estimateur obtenu en empilant les données et en appliquant les MCO sur l'équation exprimée en écart aux valeurs moyennes temporelles.

Estimateur avance/retard. Estimateur d'un paramètre de cointégration dans une régression comprenant des variables I(1), dans laquelle sont inclus, parmi les régresseurs, les différences premières de la variable explicative à différentes périodes (passées, présentes, et futures.

Estimateur biaisé. Estimateur dont la moyenne d'échantillonnage n'est pas égale à la valeur qu'il est censé estimer dans la population.

Estimateur convergent. Estimateur qui converge en probabilité vers le paramètre de population lorsque la taille de l'échantillon tend vers l'infini. (*Voir* Convergence.)

Estimateur par la méthode des moments. Estimateur obtenu en utilisant l'échantillon analogue aux moments d'une population. Par exemple, l'estimateur par MCO et l'estimateur par DMC sont des estimateurs obtenus par la méthode des moments.

Estimateur par doubles moindres carrés (DMC). Estimateur basé sur l'emploi d'une variable instrumentale correspondant à la valeur prédite (ou ajustée) de la variable explicative endogène, sachant que cette valeur prédite est obtenue en régressant la variable endogène sur l'ensemble des variables exogènes.

Estimateur par différence de différences. Estimateur utilisé sur des données couvrant deux périodes temporelles, souvent dans le cadre de l'évaluation de politique. Une version de cet estimateur s'applique aux données empilées et une autre aux données de panel.

Estimateur par moindres carrés. Estimateur qui minimise la somme des carrés des résidus.

Estimateur par Moindres Carrés Généralisés (MCG). Estimateur qui implique une transformation

du modèle original afin de tenir compte de la structure supposée connue de la variance et/ou de la corrélation sérielle du terme d'erreur.

Estimateur par Moindres Carrées Quasi-Généralisés (MCQG). Procédure qui s'applique lorsque les paramètres de variance ou de corrélation [du terme d'erreur] sont inconnus et, par conséquent, doivent être estimés.

Estimateur par variables instrumentales (VI). Estimateur basé sur l'utilisation de variables instrumentales lorsqu'elles sont disponibles pour une ou plusieurs variables explicatives endogènes.

Estimateur par maximum de vraisemblance. Estimateur qui maximise (le log de) la fonction de vraisemblance.

Estimateur par différence première. Dans le cadre des modèles sur données de panel, estimateur correspondant à l'estimateur des MCO sur données empilées appliqué aux données prises en différences premières.

Estimateur linéaire sans biais. Estimateur non biaisé, qui est une fonction linéaire des réalisations de la variable dépendante.

Estimation par Maximum de Vraisemblance (EMV). Méthode d'estimation très flexible, dont l'objectif est de choisir les estimations des paramètres qui maximisent la fonction log-vraisemblance.

Estimateur sans biais de variance minimale. Un estimateur ayant la plus petite variance dans la catégorie de tous les estimateurs non biaisés.

Estimateur sans biais. Estimateur dont la valeur espérée (ou moyenne de sa distribution d'échantillon) est égale à la valeur de la population (quelle que soit la valeur de la population)

Estimateur *within*. *Voir* estimateur à effets fixes.

Estimation. Valeur numérique de l'estimateur pour un échantillon particulier de données.

Estimation de Cochrane-Orcutt (CO). Méthode d'estimation d'un modèle de régression multiples dont les variables explicatives sont strictement exogènes et les erreurs sont AR(1). Contrairement à l'approche de Prais-Winsten, Cochrane-Orcutt n'applique pas l'équation à la première observation.

Estimation de la pente par MCO. Valeur de la pente de la droite de régression des MCO.

Estimation de l'ordonnée à l'origine par MCO. Valeur de l'ordonnée à l'origine de la droite de régression des MCO.

Estimation de Prais-Winsten (PW). Méthode d'estimation destinée aux modèles de régressions linéaires multiples possédant des termes d'erreur de type AR(1) et des variables explicatives strictement exogènes. À la différence de l'approche de Cochrane-Orcutt, Prais-Winsten conserve la première observation temporelle pour l'estimation de l'équation de référence.

Estimation des MCO sur données empilées. Estimation par application des MCO sur des observations empilées suivant les dimensions temporelle et transversale.

Estimation par quasi-maximum de vraisemblance. Estimation par maximum de vraisemblance lorsque la fonction de log-vraisemblance peut ne pas correspondre à la véritable distribution conditionnelle de la variable dépendante.

Estimation d'intervalle. Estimation consistant à obtenir les bornes inférieures et supérieures d'un intervalle de confiance pour un paramètre de population. *Voir* intervalle de confiance.

Estimation *smearing*. Estimation résultant d'une méthode particulièrement utile lorsqu'il s'agit de prédire le niveau d'une variable de réponse qui est inclus sous forme logarithmique.

Étude d'événement. Analyse économétrique des effets que peut avoir un événement particulier (tel qu'un changement de politique économique ou une modification de la législation) sur une variable d'intérêt.

Évaluation de politique publique. Analyse d'un programme privé ou d'une politique publique, dont l'impact causal est évalué à l'aide de méthodes économétriques.

Exclusion d'une variable pertinente. Situation se produisant lorsqu'une variable n'est pas incluse dans un modèle, alors que son effet *ceteris paribus* sur la variable dépendante est significatif.

Exogénéité d'un instrument. Condition nécessaire à l'estimation par variables instrumentales, sous laquelle une variable instrumentale ne doit pas être corrélée avec le terme d'erreur.

Exogénéité séquentielle. Propriété d'une variable explicative dans un modèle de séries temporelles (ou de données de panel) dont le terme d'erreur au temps présent a une moyenne nulle conditionnellement à toutes les variables explicatives présentes et passées ; une définition moins stricte fait appel aux corrélations nulles.

Exogénéité stricte. Hypothèse stipulant que toutes les variables explicatives d'un modèle de série temporelle ou de données de panel sont exogènes.

Expérience. En théorie des probabilités, ce terme désigne un événement dont l'issue est incertaine. Dans le cadre de l'analyse économétrique, il indique une situation où les données sont collectées après avoir assigné les individus, de manière aléatoire, aux groupes de contrôle et de traitement.

Expérience naturelle. Situation dans laquelle l'environnement économique (qui est parfois caractérisé par une variable explicative) change de manière exogène, que ce soit dû au hasard ou à un changement institutionnel.

F

Faiblement dépendante. Se dit d'une série temporelle dont la corrélation entre variables aléatoires à deux moments dans le temps tend vers zéro lorsque l'intervalle de temps entre ces deux périodes tend vers l'infini.

Facteur d'inflation de la variance (VIF). Composante de la variance d'échantillon, qui est affectée par le degré de corrélation entre les variables explicatives.

Filtrage de la tendance. Pratique consistante à retirer la composante tendancielle d'une série chronologique.

Fonction de densité de probabilité (FDP). *Voir* densité de probabilité.

Fonction de distribution cumulative (FDC). Anglicisme. *Voir* fonction de répartition.

Fonction de log-vraisemblance. Somme des log-vraisemblances, sachant que la log-vraisemblance de chaque observation est égale au log de la densité de la variable dépendante conditionnellement aux variables explicatives ; la fonction de log-vraisemblance est une fonction des paramètres à estimer.

Fonction de perte. Fonction mesurant la « perte » (de précision) liée à la différence entre la valeur prédite et la valeur observée ; les deux fonctions de perte les plus utilisées sont la perte mesurée par la valeur absolue de l'écart et la perte mesurée par le carré de l'écart.

Fonction de régression de la population (FRP). *Voir* espérance conditionnelle.

Fonction de régression de l'échantillon (FRE). *Voir* droite de régression des MCO.

Fonction de répartition. Fonction qui donne la probabilité qu'une variable aléatoire soit inférieure ou égale à n'importe quel nombre réel particulier.

Fonction exponentielle. Fonction mathématique définie pour tout réel, dont la pente (ou dérivée première) est croissante et la variation proportionnelle (ou dérivée première relative) est constante.

Fonction linéaire. Fonction pour laquelle, si y est une fonction linéaire de x, une variation de x d'une unité entraîne une variation constante de y.

Fonction log (ou fonction logarithmique). Fonction mathématique définie seulement pour des arguments strictement positifs, dont la dérivée est positive, mais décroissante.

Fonction non linéaire. Fonction dont la pente n'est pas constante.

Fonction quadratique. Fonction qui comprend le terme au carré d'une ou plusieurs variables explicatives et qui permet d'en capturer les effets croissants ou décroissants sur la variable dépendante.

Forme quadratique. Fonction mathématique d'un vecteur ($\mathbf{x}$) qui consiste à multiplier une matrice carrée symétrique ($\mathbf{A}$), à droite par ce vecteur et à gauche par la transposée de ce vecteur.

Forme en *R* carré de la statistique *F*. Formule de la statistique de Fisher basée sur les valeurs du R carré des modèles contraint et non contraint.

Fortement dépendant. *Voir* fortement persistant.

Fortement persistant. Se dit d'un processus de séries temporelles dont les réalisations dans un futur lointain sont fortement corrélées avec les réalisations contemporaines.

Fouille de données. Pratique visant à estimer, sur la même base de données, un grand nombre de modèles afin d'en identifier le « meilleur ».

Fréquence des données. Intervalle de temps auquel les séries chronologiques sont collectées. Les fréquences d'observation les plus courantes sont annuelles, trimestrielles, mensuelles ou journalières.

G

Groupe de contrôle. Groupe qui ne participe pas au programme, à l'expérience ou qui ne reçoit pas le traitement.

Groupe de référence. *Voir* catégorie de référence.

Groupe d'expérience. *Voir* groupe de traitement.

Groupe de traitement. Groupe qui participe au programme, à l'expérience ou qui reçoit le traitement.

H

Hautement persistant. *Voir* fortement persistant.

Homoscédasticité. Propriété dont dispose le terme d'erreur lorsque sa variance, conditionnelle aux variables explicatives, est constante.

Hétérogénéité non observée. *Voir* effet non observé.

Hétéroscédasticité. Caractéristique du terme d'erreur lorsque sa variance, étant donné les variables explicatives, n'est pas constante.

Hétéroscédasticité conditionnelle autorégressive (ARCH). Modèle d'hétéroscédasticité dynamique dans lequel la variance du terme d'erreur – étant donnée l'information passée – dépend linéairement du carré des erreurs passées.

Hétéroscédasticité de forme inconnue. Hétéroscédasticité qui s'explique par une dépendance, dont la nature n'est pas connue, à l'égard des variables explicatives.

Homoscédastique sur le plan contemporain. Dans un modèle de séries temporelles ou de données de panel, se dit du terme d'erreur lorsque sa variance, conditionnelle aux régresseurs *durant la même période de temps*, est constante.

Hypothèse alternative. Hypothèse contre laquelle l'hypothèse nulle est testée.

Hypothèse alternative bilatérale. Hypothèse alternative sous laquelle le paramètre de population peut être plus petit *ou* plus grand que la valeur énoncée sous l'hypothèse nulle.

Hypothèse alternative unilatérale. Hypothèse alternative sous laquelle le paramètre est *soit* plus grand, *soit* plus petit que la valeur précisée sous l'hypothèse nulle. Dans le premier cas, l'hypothèse alternative unilatérale sera « à droite » ; dans le second, elle sera « à gauche ».

Hypothèse de normalité. Hypothèse du Modèle Linéaire Classique sous laquelle l'erreur (ou la variable dépendante) suit une loi normale, conditionnellement aux variables explicatives.

Hypothèses de Gauss-Markov. Ensemble des hypothèses (Hypothèses RLM.1 à RLM.5 ou SC.1 à SC.5) selon lesquelles l'estimateur des MCO est BLUE.

Hypothèse de moyenne conditionnelle nulle. *Voir* hypothèse d'espérance conditionnelle nulle.

Hypothèse d'espérance conditionnelle nulle. Hypothèse fondamentale, utilisée dans le cadre de l'analyse de régression, sous laquelle la valeur espérée (ou moyenne) de l'erreur est nulle, quelles que soient (et étant donné) les valeurs des variables explicatives. *Voir* les hypothèses RLM.4, ST.3 et ST.3′ dans l'ouvrage.

Hypothèses du modèle linéaire classiques (MLC). Ensemble des hypothèses idéales du modèle de régression multiple. Pour les données en coupe transversale, il s'agit des hypothèses RLM.1 à RLM.6. Pour l'analyse des séries temporelles, il s'agit des hypothèses ST.1 à ST.6. Les hypothèses incluent : la linéarité en fonction des paramètres, l'absence de colinéarité stricte, l'hypothèse de moyenne conditionnelle nulle, l'homoscédasticité, l'absence de corrélation sérielle, et la normalité des erreurs.

Hypothèse nulle. Hypothèse de travail qui n'est pas rejetée si le test d'hypothèse classique, reposant sur un échantillon de données, ne conduit pas à son rejet en faveur de l'hypothèse alternative. Lorsqu'une hypothèse n'est pas rejetée, cela ne veut pas dire qu'elle doit être « acceptée » pour autant (c'est-à-dire considérée comme « vraie une fois pour toutes »). L'absence de « preuves suffisantes », à l'encontre d'une hypothèse, dans un échantillon, ne constitue pas la preuve que cette hypothèse est vraie dans la population. *Voir également* erreur de type I.

I

i.i.d. Se dit d'une séquence de variables aléatoires « indépendantes et identiquement distribuées », qui ont toutes la même distribution de probabilité et sont mutuellement indépendantes.

Impact de long terme. Variation finale de la variable dépendante dans un modèle à retards échelonnés, étant donné une augmentation permanente d'une unité de la variable explicative.

Inclusion d'une variable superflue (ou non pertinente). Dans un modèle de régression, inclusion d'une variable explicative dont le paramètre de population est égal à zéro.

Inconsistance. Anglicisme. *Voir* absence de convergence ou non-convergence.

Inconsistant. Anglicisme. Décrit un estimateur qui ne converge pas (en probabilité) vers le paramètre de population lorsque la taille de l'échantillon s'accroît. *Voir* non convergent.

Identification d'un paramètre. Cas où le paramètre de population peut être exprimé en fonction des moments de la population et être estimé de manière convergente à partir d'un échantillon de données.

Indépendance de l'espérance (de l'erreur). Propriété grâce à laquelle l'erreur a une moyenne constante, quels que soient les sous-ensembles de la population

définis par des valeurs différentes prises par les variables explicatives.

Indice. Nombre permettant de résumer et de caractériser la variation relative d'une grandeur (comme la production ou les prix) entre deux situations, dont l'une sert de base.

Inférence statistique. Ensemble des méthodes permettant de tester des hypothèses portant sur une population à partir des résultats observés sur un échantillon extrait de cette population.

Instruments faibles. Variables instrumentales qui ne sont que faiblement corrélées à la (ou les) variable(s) explicative(s) endogène(s).

Intégré d'ordre 1 [I(1)]. Se dit d'un processus de séries temporelles qui se transforme en un processus I(0) (intégré d'ordre 0) après avoir été différencié une fois.

Intégré d'ordre 0 [I(0)]. Se dit d'un processus de séries temporelles stationnaire, faiblement dépendant, qui satisfait la loi des grands nombres et le théorème central limite lorsqu'il est utilisé dans une analyse par régression.

Intervalle de confiance (IC). Intervalle $[t_1, t_2]$ qui comprend les valeurs possibles d'un paramètre t à estimer et qui est construit de telle sorte que la probabilité d'inclure la valeur de la population t dans cet intervalle soit égale à $1 - \alpha$ ($1 - \alpha$ étant un nombre fixé positif et inférieur à 1, appelé niveau de confiance).

Intervalle de confiance asymptotique. Intervalle de confiance approximativement valable pour de grandes tailles d'échantillon.

Intervalle de prévision (ou de prédiction). Intervalle de confiance portant sur la réalisation future d'une variable (de série chronologique, le plus souvent).

Inverse d'une matrice. Pour une matrice $n \times n$, son inverse correspond à la matrice $n \times n$ qui, multipliée par la matrice originale, aboutit à la matrice identité (si elle existe).

Inverse du rapport de Mills. Terme qui peut être ajouté à un modèle de régression multiple lorsqu'il s'agit d'éliminer le biais de sélection de l'échantillon.

K

Kurtosis. Anglicisme. *Voir* coefficient d'aplatissement.

L

Limite en probabilité. Valeur vers laquelle un estimateur converge lorsque la taille de l'échantillon tend vers l'infini.

Linéaire. *Voir* fonction linéaire.

Lissage exponentiel. Méthode simple de prévision univariée qui consiste à pondérer toutes les valeurs passées relatives à cette même variable.

Logarithme naturel. *Voir* fonction logarithmique.

Loi binomiale. Loi de probabilités portant sur le nombre de réussites de n tirages indépendants de Bernoulli, dont chaque tirage a la même probabilité de réussite.

Loi des espérances itérées. Loi de probabilité qui permet de relier les espérances non conditionnelles et conditionnelles.

Loi des grands nombres (LGN). Théorème selon lequel une moyenne, calculée à partir d'un échantillon aléatoire, converge en probabilité vers la moyenne de la population. Ce théorème s'applique également aux séries temporelles stationnaires et faiblement dépendantes.

Loi normale. Loi de distribution communément utilisée en statistique et en économétrie pour modéliser une population. La densité de probabilité de la loi normale forme une courbe en cloche.

Loi normale multivariée. Distribution de plusieurs variables aléatoires où chaque combinaison linéaire de variables suit une distribution normale univariée (ou unidimensionnel).

M

Marche aléatoire. Processus de séries chronologiques, dont la valeur à la période suivante s'obtient à partir de la valeur courante, à laquelle vient s'ajouter un terme d'erreur aléatoire indépendant (ou du moins non corrélé).

Marche aléatoire avec dérive (ou avec tendance). Variable qui suit une marche aléatoire à laquelle on ajoute une constante qui permet d'incorporer une tendance déterministe (évoluant en fonction du temps qui passe).

Martingale. Processus de séries chronologiques dont la valeur espérée, étant donné toutes les informations passées sur la série, est égale à la valeur observée la plus récente. Autrement dit, la connaissance des événements passés n'aide pas à prédire la moyenne des « gains » futurs.

Matrice. Arrangement ordonné d'un ensemble d'éléments, sous forme d'un tableau à double entrée comportant des lignes et des colonnes.

Matrice carrée. Matrice dont le nombre de lignes est égal au nombre de colonnes.

Matrice de variance-covariance. Matrice semi-définie positive, associée à un vecteur aléatoire et obtenue en indiquant les variances des variables aléatoires sur la diagonale et les covariances hors de la diagonale, dans les entrées appropriées.

Matrice de variance-covariance de l'estimateur MCO. Matrice des variances et covariances d'échantillon, associées au vecteur des coefficients MCO.

Matrice scalaire des variances-covariances. Matrice des variances-covariances dans laquelle tous les éléments hors de la diagonale sont nuls et tous les éléments diagonaux sont égaux à la même constante positive.

Matrice diagonale. Matrice présentant des valeurs nulles, sauf sur la diagonale principale.

Matrice idempotente. Matrice (carrée) dont la multiplication par elle-même donne la même matrice.

Matrice identité. Matrice carrée dont tous les éléments diagonaux sont égaux à 1 et tous les éléments en dehors de la diagonale sont égaux à 0.

Matrice symétrique. Matrice (carrée) qui est égale à sa transposée.

Matrice zéro. Matrice dont toutes les entrées sont égales à zéro.

Mauvaise spécification de la forme fonctionnelle. Problème qui survient lorsque le modèle de régression n'incorpore pas la fonction adéquate d'une ou plusieurs variables explicatives (en ignorant une fonction quadratique, par exemple) ou qu'il se base sur une mauvaise spécification pour les variables dépendante et/ou indépendantes.

MCO. *Voir* Moindres Carrés Ordinaires.

Médiane. Dans une distribution de probabilité, valeur pour laquelle il y a 50 % de chance de se situer en dessous et 50 % de se situer au-delà de cette valeur. Pour une liste ordonnée de nombres, c'est la valeur qui se trouve au milieu de l'échantillon.

Médiane conditionnelle. Médiane d'une variable de réponse conditionnelle à certaines variables explicatives.

Meilleur estimateur linéaire sans biais (BLUE). Estimateur ayant la plus faible variance parmi les estimateurs linéaires et sans biais. Sous les hypothèses de Gauss-Markov, les MCO sont *BLUE*, conditionnellement aux valeurs des variables explicatives. L'acronyme *BLUE* signifie en anglais « *Best Linear Unbiased Estimator* ».

Méthode de ré-échantillonnage. Méthode dont l'objectif est d'approximer les écarts-types estimés (ainsi que les distributions des statistiques de test) en générant des échantillons à partir d'un tirage aléatoire sur les données originales et en calculant les estimations désirées pour chacun des nouveaux échantillons.

Méthode Heckit. Procédure économétrique utilisée pour corriger le biais de sélection d'un échantillon dû à la troncature fortuite ou pour corriger toute autre forme de données manquantes pour des raisons non aléatoires.

Micronumérosité. Terme introduit par Arthur Goldberger pour décrire les propriétés des estimateurs économétriques lorsque l'échantillon est de petite taille.

Modèle contraint. Modèle valide sous l'hypothèse nulle, c'est-à-dire modèle incorporant toutes les restrictions imposées par l'hypothèse nulle.

Modèle à coefficient (de pente) aléatoire. Régression multiple où les paramètres de pente peuvent dépendre de variables non observées au niveau des unités d'observation.

Modèle à correction d'erreur. Modèle de séries temporelles en différence première ; ce modèle inclut un terme de correction d'erreur (ou terme de cointégration) qui permet à deux séries I(1) de converger vers leurs niveaux d'équilibre de long terme.

Modèle à effets aléatoires. Modèle de panel dans lequel l'effet non observé est (supposé) non corrélé avec les variables explicatives, quelle que soit la période de temps.

Modèle à effets fixes. Modèle de panel dans lequel l'effet non observé peut afficher une corrélation quelconque avec les variables explicatives, quelle que soit la période de temps.

Modèle à effets inobservés. Modèle de données de panel ou sur échantillons par grappes, dans lequel le terme d'erreur contient un effet non observé. *Voir également* effet non observé.

Modèle à élasticité constante. Modèle dans lequel l'élasticité de la variable dépendante par rapport à la variable explicative est constante ; les deux variables apparaissent sous forme logarithmique dans une régression multiple.

Modèle à équations simultanées (MES). Modèle qui détermine conjointement deux variables endogènes au minimum et dans lequel chaque variable endogène est expliquée par d'autres variables (endogènes ou exogènes) et par un terme d'erreur.

Modèle à réponse binaire. *Voir* modèle de choix binaire.

Modèle de choix binaire. Modèle dont la variable dépendante est une variable binaire (ou indicatrice).

Modèle à probabilité linéaire (MPL). Modèle de choix binaire, dont la probabilité de réponse est une fonction linéaire par rapport à ses paramètres.

Modèle à retards échelonnés. Modèle de séries chronologiques reliant la variable dépendante aux valeurs courantes et passées des variables explicatives.

Modèle à retards échelonnés de Koyck. Modèle à retards échelonnés infinis où les coefficients des variables retardées suivent une progression géométrique décroissante.

Modèle à retards échelonnés finis. Modèle dynamique dans lequel une ou plusieurs variables explicatives n'influencent la variable dépendante que de façon non contemporaine, ce qui explique leur inclusion sous forme retardée (d'une période, au minimum) dans le modèle de régression.

Modèle à retards échelonnés infinis. Modèle à retards échelonnés où une variation de la variable explicative est susceptible d'avoir un effet sur la variable dépendante pendant une période illimitée.

Modèle à retards échelonnés rationnels (MRER). Modèle à retards échelonnés où la distribution des retards dépend d'un nombre relativement restreint de paramètres.

Modèle autorégressif vectoriel. Modèle incluant au minimum deux séries temporelles, dans lequel chaque variable est une fonction linéaire des valeurs passées de l'ensemble des variables auxquelles sont ajoutées un terme d'erreur de moyenne nulle, étant donné toutes les valeurs passées des variables observées.

Modèle à variable latente. Modèle dont la variable dépendante observée est supposée être une fonction d'une variable sous-jacente latente, ou non observée.

Modèle de régression bivarié. *Voir* modèle de régression linéaire simple.

Modèle de régression censurée. Modèle de régression multiple dans lequel la variable dépendante a été censurée au-dessus ou au-dessous d'un seuil connu. *Voir aussi* données censurées.

Modèle de régression de Poisson. Modèle adapté à une variable dépendante de comptage qui, conditionnellement aux variables explicatives, suit une distribution de Poisson.

Modèle de régression linéaire multiple (régression multiple). Modèle linéaire dans ses paramètres, dont la variable dépendante est une fonction de variables indépendantes et d'un terme d'erreur.

Modèle de régression linéaire simple (régression simple). Modèle linéaire dans ses paramètres, dont la variable dépendante est une fonction linéaire d'une seule variable indépendante et d'un terme d'erreur.

Modèle de régression normale censurée. Cas particulier d'un modèle de régression censurée, pour lequel le modèle sous-jacent de la population satisfait les hypothèses du modèle linéaire classique.

Modèle de régression normale tronquée. Cas particulier d'un modèle de régression tronquée pour lequel le modèle sous-jacent de la population satisfait les hypothèses du modèle linéaire classique.

Modèle de régression tronquée. Modèle de régression linéaire sur données en coupe transversale que l'on obtient à partir d'un plan de sondage, en excluant complètement une partie de la population. Le critère d'exclusion s'appuie sur les valeurs observées de la variable dépendante. La distribution observée de la variable dépendante est alors tronquée.

Modèle VAR. *Voir* modèle autorégressif vectoriel.

Modèle de population. Modèle, en particulier un modèle de régression linéaire, décrivant une population.

Modèle dynamique complet. Modèle de séries chronologiques, qui comprend *tous* les retards et variables nécessaires à l'explication de l'espérance de la variable dépendante.

Modèle économétrique. Équation reliant la variable dépendante à un ensemble de variables indépendantes et d'erreurs non observées et dont les paramètres de la population sous-jacente déterminent l'effet *ceteris paribus* de chacune de ces variables sur la variable dépendante.

Modèle économique. Relation dérivée de la théorie économique ou d'un raisonnement économique moins formel.

Modèle linéaire classique (MLC). Modèle de régression linéaire multiple, qui vérifie toutes les hypothèses requises du modèle linéaire classique.

Modèle logit. Modèle probabiliste non linéaire utilisé lorsque la variable à expliquer est de type binaire et dont la probabilité de réponse est donnée par la fonction logit.

Modèle log-log. Modèle de régression dont la variable dépendante et certaines variables explicatives sont sous forme logarithmique.

Modèle log-niveau. Modèle de régression dont la variable dépendante est sous forme logarithmique et dont les variables indépendantes sont « en niveau » (sous leur forme originelle).

Modèle niveau-log. Modèle de régression dont la variable dépendante est « en niveau » (sous sa forme originelle) et dont certaines variables indépendantes sont sous forme logarithmique.

Modèle niveau-niveau. Modèle de régression dont la variable dépendante et les variables indépendantes sont exprimées « en niveau » (sous leur forme originelle).

Modèles non emboîtés. Deux modèles (ou plus) dont aucun ne peut être écrit comme un cas particulier de l'autre, quelles que soient les contraintes imposées sur les paramètres.

Modèle non restreint. Dans le cadre du test d'hypothèse, modèle pour lequel aucune restriction n'a été placée sur ses paramètres.

Modèle parcimonieux. Modèle qui consume le moins de paramètres possible, tout en parvenant à capturer les caractéristiques désirées.

Modèle probit. Modèle probabiliste non linéaire utilisé lorsque la variable à expliquer est de type binaire et dont la probabilité de réponse est donnée par la fonction de répartition de la loi normale.

Modèle restreint. En tests d'hypothèses, modèle obtenu après avoir imposé toutes les restrictions requises sous l'hypothèse nulle.

Modèle statique. Modèle de séries temporelles, dans lequel seules des variables explicatives contemporaines affectent la variable dépendante.

Modèle Tobit. Modèle adapté à l'étude d'une variable dépendante qui peut prendre la valeur zéro avec une probabilité positive et être distribuée de façon continue pour les valeurs strictement supérieures à zéro. *Voir aussi* solution de coin.

Modèle vrai. Modèle de la population qui relie la variable dépendante aux variables indépendantes pertinentes et à un terme d'erreur pour lequel l'hypothèse de moyenne conditionnelle nulle est vérifiée.

Moindres Carrés Ordinaires (MCO). Méthode utilisée pour estimer les paramètres d'un modèle de régression linéaire multiple. Les estimateurs MCO sont obtenus en minimisant la somme des carrés des résidus.

Moindres déviations absolues (MDA). Méthode utilisée pour estimer les paramètres d'un modèle de régression multiple et basée sur la minimisation de la somme des valeurs absolues des résidus.

Moyenne. *Voir* valeur espérée.

Moyenne de l'échantillon. Mesure de tendance centrale, égale à la somme de n nombres divisée par n.

Multicolinéarité. Corrélation élevée entre les variables indépendantes d'un modèle de régression multiple, bien que l'ordre de grandeur de la corrélation ne soit pas clairement défini dans les faits.

Multiplicateur de court terme. *Voir* effet de court terme.

Multiplicateur de Lagrange. *Voir* statistique du multiplicateur de Lagrange.

Multiplicateur de long terme. *Voir* impact de long terme.

Multiplication scalaire. Méthode de calcul permettant de multiplier un scalaire (nombre) par une matrice ou un vecteur.

N

Niveau de confiance. Pourcentage des échantillons pour lesquels l'intervalle de confiance devrait contenir la valeur de la population ; bien que 95 % est le niveau de confiance le plus communément utilisé, 90 % et 99 % sont également utilisés.

Niveau de signification. Risque de commettre une erreur de type I, que l'on détermine avant d'effectuer un test d'hypothèse. Plus on désire minimiser ce risque, plus la probabilité choisie doit être proche de zéro. En sciences sociales, la probabilité la plus fréquemment utilisée est 5 % ; les autres probabilités les plus courantes sont 1 % et 10 %.

Niveau de significativité. *Voir* niveau de signification.

Niveau de test. *Voir* niveau de signification.

Non conjointement significatifs. Se dit des coefficients d'un groupe de variables explicatives lorsque l'hypothèse nulle d'un test F, selon laquelle la valeur vraie de chacun de ces coefficients est nulle, n'est pas rejetée à un niveau de signification donné.

Non convergence. Différence entre la limite de probabilité d'un estimateur et la valeur (vraie) du paramètre dans la population.

Non convergent. Se dit d'un estimateur qui ne converge pas (en probabilité) vers la valeur vraie du paramètre de population lorsque la taille de l'échantillon augmente.

Normalité asymptotique. Propriété d'un estimateur centré réduit dont la distribution converge vers une loi normale standard (centrée et réduite).

Notation matricielle. Notation mathématique fondée sur l'algèbre matricielle et utilisée pour présenter et analyser le modèle de régression multiple.

O

Observations aberrantes. *Voir* observations isolées.

Observations isolées. Observations dans un ensemble de données, qui diffèrent considérablement de la plupart des données, en raison d'erreurs ou parce que certaines données sont générées par un modèle différent du reste de l'ensemble de données.

Outliers. Anglicisme. *Voir* observations isolées.

Opérateur somme. Notation, désignée par Σ, qui est utilisée pour calculer la somme d'un ensemble de nombres.

Over Controlling. Anglicisme indiquant que trop de facteurs de contrôle ont été pris en compte dans un modèle de régression multiple lors de l'étude *ceteris paribus* d'une ou plusieurs autres variables explicatives. *Voir aussi* inclusion d'une variable superflue.

P

Panel cylindré. Base de données de panel dans laquelle chaque année (ou chaque date) est disponible pour toutes les unités d'observation.

Panel non cylindré. Base de données de panel dans laquelle les données pour certaines années (ou périodes) sont manquantes pour certaines unités d'observation.

Panel équilibré. Traduction littérale de « *balanced panel* ». *Voir* panel cylindré.

Paramètre. Nom donné à certains coefficients, autres que la variable ou l'inconnue, en fonction desquels on peut exprimer une proposition ou les solutions d'un problème au sein d'une population.

Paramètre de la constante. Paramètre qui, dans un modèle de régression linéaire, donne la valeur espérée de la variable dépendante lorsque toutes les variables indépendantes sont égales à zéro.

Paramètre de pente. Coefficient d'une variable indépendante dans un modèle de régression linéaire multiple, dont l'estimation permet de mesurer son effet *ceteris paribus* sur la variable dépendante.

Paramètres de la forme réduite. Paramètres apparaissant dans une équation de la forme réduite. *Voir également* équation de la forme réduite.

Paramètres structurels. Paramètres qui apparaissent dans l'équation structurelle. *Voir également* équation structurelle.

Pente. Dans l'équation d'une droite, changement de la variable y quand la variable x augmente.

Périodicité. Propriété d'une série chronologique, qui se traduit par une valeur moyenne différente en fonction de certains moments de l'année (au niveau journalier, hebdomadaire, mensuel ou trimestriel).

Pertinence d'un instrument. Condition nécessaire à l'estimation par variables instrumentales, sous laquelle une variable instrumentale doit contribuer à l'explication de la variation de la variable endogène.

Prévision ponctuelle. Valeur prédite d'une réalisation future.

Population. Ensemble d'individus au périmètre clairement défini (personnes, entreprises, ménages, etc.), qui constitue l'objet d'une analyse statistique ou économétrique.

Pourcentage de prédictions correctes. Pourcentage de prédiction correcte d'occurrence des valeurs 0 ou 1 de la variable endogène, dans le cadre d'un modèle de choix binaire.

Prévision conditionnelle. Prévision qui suppose que les valeurs futures de certaines variables explicatives sont connues avec certitude.

Prévision non conditionnelle. Prévision qui ne s'appuie pas sur les valeurs (supposées) connues des variables explicatives futures.

Prévision sur une période. Prévision d'une série temporelle sur une seule période.

Prévisions sur plusieurs périodes. Prévision d'une série temporelle couvrant plusieurs périodes.

Probabilité critique. *Voir* p-valeur.

Probabilité de réponse. Probabilité que la variable dépendante d'un modèle de choix binaire prenne la valeur 1, étant donné les valeurs des variables explicatives.

Procédure d'Engle-Granger en deux étapes. Méthode en deux étapes visant à estimer un modèle à correction d'erreur. Les paramètres de l'équation de cointégration sont estimés lors de la première étape, alors que ceux du modèle à correction d'erreur le sont lors de la seconde.

Processus AR(1) stable. Processus AR(1) pour lequel le paramètre du retard (ou coefficient autorégressif d'ordre 1) est inférieur à 1 en valeur absolue. Par ailleurs, la corrélation entre deux variables

aléatoires diminue vers zéro à un taux géométrique à mesure que la distance entre les deux variables augmente. Un processus AR(1) stable est donc faiblement dépendant.

Processus à tendance stationnaire. Processus dont la stationnarité n'est obtenue qu'après en avoir soustrait la tendance temporelle. En règle générale, il est implicitement admis que cette série « épurée » est faiblement dépendante.

Processus autorégressif d'ordre 1 [AR(1)]. Modèle de séries temporelles, dans lequel la valeur actuelle d'une variable dépend linéairement de sa valeur la plus récente et d'un terme d'erreur impossible à prédire.

Processus à racine unitaire. Processus temporel fortement persistant, pour lequel la valeur courante est égale à la valeur de la période précédente, à laquelle s'ajoute une perturbation faiblement dépendante.

Processus à tendance. Processus temporel, dont la valeur espérée est une fonction croissante ou décroissante du temps.

Processus de moyenne mobile d'ordre un [MA(1)]. Modèle de séries temporelles, dans lequel la valeur actuelle d'une variable dépend linéairement des valeurs actuelle et retardée d'un processus stochastique i.i.d., de moyenne nulle et de variance constante.

Processus non stationnaire. Processus temporel dont la distribution conjointe n'est pas constante au cours du temps.

Processus stationnaire. Processus temporel dont toutes les distributions marginales et conjointes ne varient pas au cours du temps.

Processus stationnaire en différence première. Processus temporel qui devient I(0) lorsqu'il est exprimé en différence première.

Processus stochastique. Séquence de variables aléatoires indexées par le temps.

Processus temporel. *Voir* processus stochastique.

Produit matriciel : Méthode de calcul permettant de multiplier deux matrices compatibles.

Propriétés asymptotiques. Propriétés des estimateurs et statistiques lorsque la taille de l'échantillon tend vers l'infini.

Propriétés en grand échantillon. *Voir* propriétés asymptotiques.

Pseudo *R* carré. Mesure de qualité d'ajustement adaptée aux modèles à variable dépendante limitée.

Puissance d'un test. Probabilité de rejeter l'hypothèse nulle lorsque celle-ci est fausse ; cette probabilité dépend des valeurs des paramètres de la population déterminées sous l'hypothèse alternative.

***p*-valeur.** La plus petite valeur de probabilité pour laquelle l'hypothèse nulle peut encore être rejetée. Plus cette probabilité est faible, plus le risque de commettre une erreur de type I est faible. En d'autres termes, la *p*-valeur mesure le niveau de signification [« ex post » ou « exact »] en dessous duquel l'hypothèse nulle ne peut plus être rejetée.

Q

Quasi-expérimentation. Méthodologie qui s'inspire du schéma expérimental, mais qui ne requiert pas d'attribution aléatoire des groupes de contrôle et de traitement. Conception moins rigoureuse que l'expérience naturelle (car elle ne permet pas le contrôle intégral de tous les facteurs explicatifs), mais très utilisée en sciences sociales où de nombreux problèmes éthiques peuvent se poser lors de l'attribution aléatoire des groupes de contrôle et de traitement.

R

Racine (carrée) de l'erreur quadratique moyenne (REQM). *Voir* écart-type de la régression (ETR).

Rang d'une matrice. Nombre de colonnes linéairement indépendantes dans une matrice.

Ratio *t* (de Student). Statistique *t* obtenue en considérant que la valeur vraie du paramètre (dans la population) est égale à zéro. *Voir* statistique *t*.

***R* barre carré.** *Voir R* carré ajusté.

***R* carré.** Proportion des variations totales de la variable dépendante dans l'échantillon, qui est expliquée par les variables indépendantes.

***R* carré ajusté.** Mesure de la qualité d'ajustement d'une régression multiple, qui introduit une pénalité, lors de l'ajout de variables explicatives, en ajustant l'estimation de la variance de l'erreur en fonction du nombre de degrés de liberté.

***R* carré non centré.** *R* carré dont la somme des carrés totaux (SCT) est calculée sans soustraire la moyenne d'échantillon de la variable dépendante.

***R* carré de la population.** Part des variations totales de la variable dépendante dans la population, qui est expliquée par les variables explicatives.

Région de rejet. *Voir* Zone de rejet.

Règle de décision. Règle utilisée dans le cadre d'un test d'hypothèse, qui permet de déterminer à quel moment l'hypothèse nulle doit être rejetée en faveur de l'hypothèse alternative.

Régresseur. *Voir* variable explicative.

Régression à l'origine. Analyse de régression pour laquelle la valeur de l'ordonnée à l'origine est fixée à zéro ; les paramètres de pentes sont obtenus comme d'habitude, en minimisant la somme des carrés des résidus.

Régression auxiliaire. Régression utilisée pour calculer une statistique de test, comme dans les tests d'hétéroscédasticité et de corrélation sérielle par exemple. En règle générale, toute régression qui n'estime pas directement le modèle d'intérêt.

Régression fallacieuse. Régression qui identifie une relation significative entre deux séries chronologiques alors qu'en réalité elles ne sont pas liées entre elles dans la population ; cela peut arriver lorsque ces deux séries suivent une tendance et/ou sont intégrées d'ordre 1 (comme dans le cas d'une marche aléatoire).

Régression logistique. *Voir* modèle logit.

Régression sur variables indicatrices. Dans le cadre des modèles de panel, régression qui inclut une variable indicatrice pour chaque unité de coupe transversale, en sus des autres variables explicatives. Cela correspond à l'estimateur à effets fixes.

RESET. *Voir* Test d'erreur de spécification de la régression.

Résidu. Différence entre la valeur observée (ou réelle) et la valeur ajustée (ou prédite) de la variable dépendante ; il existe un résidu pour chaque observation de l'échantillon qui sert à obtenir la droite de régression des MCO.

Résidus de Student (ou « résidus studentisés ». Résidus calculés en excluant de l'estimation chaque observation, l'une après l'autre, et en divisant par l'écart-type estimé du terme d'erreur.

Restriction d'exclusion. *Voir* Contrainte d'exclusion.

Restrictions de suridentification. Conditions de moments additionnelles imposées au modèle, en raison d'un nombre d'instruments supérieur au nombre de variables explicatives endogènes.

S

Saisonnalité. *Voir* périodicité.

Sélection endogène de l'échantillon. Échantillon non aléatoire dont la constitution dépend de la variable dépendante, que ce soit de manière directe ou indirecte (via le terme d'erreur de l'équation), ce qui introduit un biais dans l'estimateur des MCO.

Sélection exogène de l'échantillon. Échantillon non aléatoire dont la constitution dépend des variables explicatives exogènes ou est indépendante du terme d'erreur de l'équation d'intérêt.

Semi-définie positive (ou autoadjointe positive). Se dit d'une matrice symétrique dont toutes les formes quadratiques (ou bilinéaires) sont non négatives.

Semi-élasticité. Changement en pourcentage de la variable dépendante, étant donné une augmentation d'une unité de la variable indépendante considérée.

Séries chronologiques (ou temporelles). *Voir* données de séries chronologiques (ou temporelles).

Série temporelle désaisonnalisée. Données de série temporelle dont la composante saisonnière a été éliminée à l'aide d'une procédure statistique (comme l'emploi d'une régression sur des variables saisonnières binaires).

Signification économique. *Voir* signification pratique.

Signification globale d'une régression. Test de significativité jointe de toutes les variables explicatives qui apparaissent dans l'équation de régression multiple.

Signification pratique. Importance pratique ou économique d'une estimation, qui s'évalue en fonction du signe et de l'ampleur de l'estimation (et non en fonction de sa signification statistique).

Signification statistique. Importance de l'estimation d'un paramètre sur le plan statistique, évaluée habituellement à l'aide d'un test dont l'hypothèse nulle stipule que la valeur vraie du paramètre est égale à zéro. Plus la *p*-valeur de ce test est proche de zéro, plus la signification statistique de ce paramètre est élevée. *Voir également p*-valeur.

Significativité statistique. *Voir* signification statistique.

Simultanéité. Terme signifiant qu'au moins une variable explicative du modèle de régression linéaire multiple est déterminée conjointement avec la variable dépendante.

Smearing estimate. *Voir* estimation *smearing*.

Solution de coin. Se dit d'une variable dépendante non négative qui est plus ou moins continue sur des valeurs strictement positives, mais qui prend la valeur zéro de manière régulière.

Solution de remplacement pour répondre au biais de variable omise. Utilisation d'une variable de

substitution, en remplacement d'une variable omise non observée, dans un modèle de régression estimé par les MCO.

Somme des carrés de résidus (SCR). Somme des carrés des résidus obtenus par les MCO et calculés sur l'ensemble des observations de l'échantillon.

Somme des carrés expliqués (SCE). Variation totale des valeurs ajustées (ou prédites) de la variable dépendante autour de la moyenne d'échantillon.

Somme des carrés totaux (SCT). Variance totale des valeurs observées (ou réelles) de la variable dépendante autour de la moyenne d'échantillon.

Stationnaire dans la covariance. Se dit d'une série temporelle (dont la moyenne et la variance sont constantes) quand la covariance entre deux variables aléatoires quelconques dans la séquence dépend seulement de la distance entre elles.

Statistique de Durbin-Watson (DW). Statistique utilisée pour tester la présence d'autocorrélation d'ordre 1 dans les erreurs d'un modèle en séries temporelles sous les hypothèses classiques du modèle de régression linéaire.

Statistiques descriptives. Statistiques utilisées pour résumer l'information relative à un ensemble d'observations. La moyenne, la médiane et l'écart-type empiriques sont les éléments de statistique descriptives les plus courants.

Statistique de Student. Voir Statistique *t*.

Statistique de test. Règle utilisée pour tester des hypothèses lorsque chaque échantillon aboutit à une valeur numérique différente.

Statistique de Wald. Statistique utilisée pour tester des hypothèses dans des circonstances assez diverses et qui suit typiquement une distribution asymptotique du chi-deux.

Statistique du multiplicateur de Lagrange (Statistique ML). Statistique de test qui repose sur les propriétés des grands échantillons et qui est souvent utilisée pour tester des problèmes de spécification (tels que l'omission de variables, la présence d'hétéroscédasticité, la présence de corrélation sérielle, etc.).

Statistique du rapport de vraisemblance. Statistique qui est utilisée pour tester une ou plusieurs hypothèses lorsque les modèles contraints et non contraints ont été estimés par maximum de vraisemblance. La statistique est égale au double de la différence entre la log-vraisemblance non contrainte et la log-vraisemblance contrainte.

Statistique du rapport de quasi-maximum de vraisemblance. Modification de la statistique du rapport de vraisemblance permettant de prendre en compte une éventuelle erreur de spécification dans la distribution, comme dans le cas des modèles de régression de Poisson.

Statistique *n*-*R*-carré. *Voir* statistique du multiplicateur de Lagrange.

Statistique du score. *Voir* statistique du multiplicateur de Lagrange.

Statistique *F*. Statistique qui suit une distribution de Fisher et qui est utilisée pour tester des hypothèses multiples relatives aux paramètres d'un modèle de régression multiple.

Statistique *F* robuste à l'hétéroscédasticité. Statistique *F* qui est (asymptotiquement) robuste à l'hétéroscédasticité de forme inconnue.

Statistique *ML* robuste à l'hétéroscédasticité. Statistique *ML* qui est robuste à l'hétéroscédasticité de forme inconnue. Statistiquement différent de zéro. *Voir* Statistiquement significatif.

Statistiquement non significative. Se dit d'une variable lorsqu'à un niveau de signification donné, il y a non-rejet de l'hypothèse nulle stipulant que la valeur de son paramètre est égale à zéro dans la population.

Statistiquement significative. Se dit d'une variable lorsqu'à un niveau de signification donné, il y a rejet de l'hypothèse nulle stipulant que la valeur de son paramètre est égale à zéro dans la population.

Statistique *t*. Statistique qui suit une distribution de Student et qui est utilisée pour tester une hypothèse simple sur les paramètres d'un modèle économétrique.

Statistique *t* asymptotique. Statistique *t* qui suit approximativement une loi normale dans des échantillons de grande taille.

Statistique *t* robuste à l'hétéroscédasticité. Statistique *t* qui est (asymptotiquement) robuste à l'hétéroscédasticité de forme inconnue.

Strictement exogène. Caractéristique d'une variable explicative, dans les modèles de séries temporelles ou de données de panel, qui signifie que la valeur espérée du terme d'erreur est nulle à chaque période de temps considérée, quelles que soient les valeurs que prend cette variable explicative dans le temps (passé, présent ou futur). Une version moins restrictive existe et est exprimée en termes de corrélation nulle.

Suridentification d'un modèle. *Voir* inclusion d'une variable superflue.

Surdispersion. S'observe dans le cadre de la modélisation d'une variable de comptage, lorsque la variance est supérieure à la moyenne.

T

Tableur. Logiciel utilisé pour encoder et manipuler des données.

Taux de croissance. Variation proportionnelle d'une série temporelle, qui est parfois approximée par la différence des logs. Ce taux est souvent reporté en pourcentage : on parle alors de pourcentage de variation ou de taux de variation (en pourcentage).

Taux de variation (en pourcentage). Variation proportionnelle, multipliée par 100, d'une variable exprimée dans *une unité de mesure* donnée.

Taux de variation en points de pourcentage. Variation absolue d'une variable exprimée en *pourcentage*.

Tendance exponentielle. Tendance caractérisée par un taux de croissance constant.

Tendance temporelle. Composante d'une série chronologique, qui est liée au temps.

Tendance temporelle linéaire. Composante d'une série chronologique, qui est liée linéairement au temps.

Terme d'erreur. Variable aléatoire dans un modèle de régression, qui contient les facteurs non observés affectant la variable dépendante. Le terme d'erreur peut aussi inclure les erreurs de mesure des variables dépendantes ou indépendantes observées.

Terme d'erreur composite. Dans des données de panel, somme d'un effet inobservé constant au cours du temps et d'une erreur idiosyncratique.

Terme d'interaction. Dans un modèle de régression, variable indépendante qui est le produit de deux variables explicatives différentes.

Test bilatéral. Test d'une hypothèse nulle contre une hypothèse alternative bilatérale. *Voir aussi* test d'hypothèse.

Test convergent. Test pour lequel, sous l'hypothèse alternative, la probabilité de rejeter l'hypothèse nulle converge vers un lorsque la taille de l'échantillon grandit indéfiniment.

Test de Breusch-Godfrey. Test de corrélation sérielle d'ordre q, qui est asymptotiquement valide même en présence de variables dépendantes retardées ou d'autres régresseurs non-exogènes.

Test de Breusch-Pagan. Test d'hétéroscédasticité, qui consiste à régresser le carré des résidus des MCO sur les variables explicatives du modèle d'intérêt.

Test de Chow. Statistique F qui teste l'égalité des paramètres d'une régression entre différents groupes (par exemple, entre les hommes et les femmes) ou entre différents intervalles de temps (par exemple, avant et après un changement politique).

Test de Davidson-MacKinnon. Test utilisé pour évaluer la validité d'un modèle particulier contre une alternative non imbriquée. Il peut être mis en œuvre au moyen d'un test de Student portant sur les valeurs ajustées du modèle alternatif.

Test de Dickey-Fuller (DF). Test t dont l'hypothèse nulle caractérise la présence d'une racine unitaire dans un modèle AR(1). *Voir également* test de Dickey-Fuller augmenté.

Test de Dickey-Fuller augmenté. Test de racine unitaire qui inclut, comme régresseurs, les variations retardées de la variable d'intérêt.

Test d'Engle-Granger. Test dont l'hypothèse nulle est l'absence de cointégration entre deux séries temporelles ; la statistique du test est obtenue à partir des résidus (en différence première) obtenus par les MCO, comme pour la statistique de Dickey-Fuller.

Test d'erreur de spécification de la régression (RESET). Test général visant à évaluer la forme fonctionnelle d'un modèle de régression multiple. Il s'agit d'un test joint de Fisher sur les valeurs ajustées du modèle des MCO élevés au carré, au cube et parfois même à des puissances supérieures.

Test de White. Test visant à évaluer la présence d'hétéroscédasticité dans les erreurs d'un modèle de régression et qui consiste à régresser les résidus des MCO au carré sur les valeurs prédites de la variable dépendante ainsi que sur leur carré. Dans sa forme la plus générale, les résidus MCO au carré sont régressés sur les variables explicatives, le carré des variables explicatives et tous les termes d'interactions des variables explicatives (quand ils ne sont pas redondants).

Test d'hypothèse. Test visant à évaluer la vraisemblance d'une hypothèse, appelée hypothèse nulle, relativement à une autre, appelée hypothèse alternative.

Test d'hypothèses jointes. Test impliquant au moins deux restrictions sur les paramètres d'un modèle.

Test d'hypothèses multiples. *Voir* test d'hypothèses jointes.

Test unilatéral. Test d'une hypothèse nulle contre une hypothèse alternative unilatérale. *Voir aussi* test d'hypothèse.

Théorème central limite (TCL). Résultat important en théorie des probabilités qui implique que le rapport entre une somme de variables aléatoires indépendantes (ou même faiblement dépendantes) et son écart-type suit une distribution qui tend vers la loi normale centrée réduite quand la taille de l'échantillon augmente.

Théorème de Gauss-Markov. Théorème qui, sous les cinq hypothèses de Gauss-Markov, établit que l'estimateur des MCO est *BLUE* (conditionnellement aux valeurs prises par les variables explicatives sur l'échantillon de données considéré).

Trace d'une matrice. Somme des éléments de la diagonale d'une matrice carrée.

Transformation à effets fixes. Dans le cadre des modèles de panel, approche consistant à exprimer les données en écart à leur moyenne temporelle.

Transformation *within*. *Voir* Transformation par effets fixes.

Transposée. Matrice construite à partir d'une autre matrice dont on a permuté les colonnes et les lignes.

« Trappe à variables indicatrices ». Erreur liée à l'introduction d'un trop grand nombre de variables indicatrices parmi les variables indépendantes du modèle. Par exemple, dans le modèle de régression en données de panel, elle survient lorsqu'une constante est présente à côté de variables indicatrices représentant *chaque* unité ou groupe d'observation.

Troncature auxiliaire. Problème de sélection d'échantillons dans lequel une variable, habituellement la variable dépendante, est observée seulement pour certaines réalisations d'une autre variable.

Troncature accessoire. Troncature fortuite. *Voir* troncature auxiliaire.

V

Valeur critique. Valeur à laquelle la statistique de test est comparée dans le but de déterminer si l'hypothèse nulle est rejetée ou non.

Valeur de référence. Valeur donnée à un indice à une date de référence. Elle est généralement égale à 1 ou 100.

Valeur espérée. Mesure de tendance centrale associée à la distribution d'une variable aléatoire, consistant à pondérer toutes les réalisations possibles par leurs probabilités associées.

Valeurs ajustées. Valeurs de la variable dépendante, obtenues après estimation du modèle par les MCO et lorsque les variables indépendantes prennent des valeurs spécifiques.

Valeur *p*. *Voir* *p*-valeur.

Valeur prédite. *Voir* valeur ajustée.

Variable aléatoire. Tout nombre réel aléatoire dont la valeur dépend du résultat d'une expérience probabiliste.

Variable aléatoire binaire. *Voir* variable de Bernoulli.

Variable aléatoire centrée réduite. Variable aléatoire transformée en lui soustrayant sa valeur espérée et en divisant le résultat obtenu par son écart-type. Cette variable centrée et réduite est de moyenne nulle et de variance unitaire.

Variable aléatoire chi-deux. Variable aléatoire suivant une loi du chi-deux.

Variable aléatoire continue. Variable dont les valeurs possibles sont réparties de façon continue sur un intervalle et pour laquelle la probabilité d'observer une valeur spécifique est égale à zéro. C'est la raison pour laquelle on considère plutôt la probabilité que cette valeur soit comprise dans un intervalle donné.

Variable aléatoire discrète. Variable aléatoire dont le support est un ensemble fini ou infini dénombrable.

Variable aléatoire *F*. Variable aléatoire caractérisée par une distribution de Fisher.

Variable aléatoire standardisée. *Voir* variable aléatoire centrée réduite.

Variables aléatoires indépendantes. Variables aléatoires dont la distribution jointe est égale au produit de leurs distributions marginales.

Variables aléatoires non corrélées. Variables aléatoires qui ne sont pas liées l'une à l'autre sur le plan linéaire.

Variables aléatoires non corrélées deux à deux. Un ensemble de deux ou plusieurs variables aléatoires, dans lequel aucun couple de variables n'est corrélé.

Variable de Bernoulli. Variable aléatoire qui prend deux valeurs : zéro ou un.

Variable binaire. *Voir* variable indicatrice.

Variable de comptage. Variable dont les valeurs sont entières et non négatives.

Variable de contrôle. *Voir* variable explicative.

Variable dépendante. Variable que le modèle (de régression multiple, entre autres) cherche à expliquer.

Variable dépendante limitée (VDL). Variable dépendante, ou variable de réponse, dont l'étendue des valeurs possibles est fortement réduite.

Variable dépendante retardée. Valeur antérieure de la variable dépendante, qui sert souvent de variable explicative dans les modèles de séries temporelles.

Variable de réponse. *Voir* variable dépendante.

Variable de substitution. Variable observée qui est liée à une variable explicative non observable, sans lui être identique pour autant.

Variable dichotomique. *Voir* variable indicatrice.

Variable endogène retardée. Valeur retardée d'une des variables endogènes dans un modèle à équations simultanées.

Variable exogène. Variable non corrélée avec le terme d'erreur du modèle.

Variable explicative. Variable qui est utilisée pour expliquer les variations de la variable dépendante dans un modèle de régression.

Variable explicative endogène. Variable explicative corrélée avec le terme d'erreur en raison de l'omission d'une variable pertinente, d'une erreur de mesure ou d'un problème de simultanéité.

Variable explicative exogène. Variable explicative qui n'est pas corrélée avec le terme d'erreur du modèle.

Variable expliquée. *Voir* variable dépendante.

Variable indépendante. *Voir* variable explicative.

Variable indicatrice. Variable prenant, comme valeurs possibles, zéro ou un.

Variable indicatrice dépendante. *Voir* modèle de choix binaire.

Variable instrumentale (VI). Variable qui permet d'obtenir un estimateur convergent du coefficient d'une variable explicative endogène, à condition que la VI soit corrélée avec la variable endogène, mais pas avec le terme d'erreur du modèle de régression. La VI intervient dans le calcul de l'estimateur, mais ne sert pas de variable explicative dans le modèle.

Variable nominale. Variable mesurée en valeur, à prix courants.

Variable omise. Variable pertinente qui n'est pas incluse dans le modèle comme variable explicative, mais dont il aurait fallu contrôler l'effet sur la variable dépendante.

Variable ordinale. Variable dont l'ensemble des modalités est ordonnée (ou hiérarchisée). Les modalités sont souvent désignées par un nombre qui ne représente pas une quantité absolue objectivement mesurable, mais bien un rang, un degré, ou un niveau sur une échelle ou gradation donnée.

Variable prédéterminée. Variable endogène ou exogène retardée, dans un modèle d'équations simultanées.

Variable prédictive. *Voir* variable explicative.

Variable prédite. *Voir* variable dépendante.

Variable qualitative. Variable relative aux caractéristiques non-quantitatives d'une unité d'observation (comme un individu, une entreprise, une ville, etc.).

Variable réelle. Valeur monétaire mesurée en fonction d'une valeur de référence.

Variable zéro-un. Variable binaire ou variable muette. *Voir* variable indicatrice.

Variables endogènes. Variables déterminées par les équations simultanées du modèle.

Variables indicatrices par année. Dans les bases de données temporelles, variable indicatrice qui vaut 1 pour l'année pertinente, et 0 sinon.

Variables saisonnières binaires. Ensemble de variables binaires utilisées pour tenir compte d'une saisonnalité dans des données temporelles.

Variance. Mesure de dispersion de la distribution d'une variable aléatoire.

Variance asymptotique. Variance qui, en divisant l'estimateur, permet d'obtenir une distribution normale standard asymptotique.

Variance conditionnelle. Variance d'une variable aléatoire, étant donné une ou plusieurs variables aléatoires.

Variance d'échantillonnage. Variance de la distribution d'échantillonnage d'un estimateur. Elle mesure l'étalement de la distribution d'échantillonnage.

Variance de l'échantillon. Estimateur sans biais et convergent de la variance de la population.

Variance de l'erreur. Variance du terme d'erreur dans un modèle de régression multiple.

Variance de l'erreur de prédiction. Variance de l'erreur provenant de la prédiction d'une valeur future de la variable dépendante à partir de l'estimation d'une équation de régression multiple.

Variation absolue. Variation entre une valeur finale et une valeur initiale.

Variation proportionnelle. Variation absolue d'une variable, divisée par sa valeur initiale.

Vecteur aléatoire. Vecteur composé de variables aléatoires.

Vecteur colonne. Vecteur de n lignes et 1 colonne.

Vecteur ligne. Vecteur de 1 ligne et n colonnes.

Vecteurs linéairement indépendants. Ensemble de vecteurs dont aucun vecteur n'est la combinaison linéaire des autres.

Z

Zone de rejet. Ensemble des valeurs d'un test statistique, qui conduisent au rejet de l'hypothèse nulle.

TABLE DES MATIÈRES

OUVERTURES ◂▸ ÉCONOMIQUES

ALLEGRET J.-P., LE MERRER P., *Économie de la mondialisation. Vers une rupture durable ?* 2e édition

AMELON J.-L., CARDEBAT J.-M., *Les nouveaux défis de l'internationalisation. Quel développement international pour les entreprises après la crise ?*

ANDERSON R. D., SWEENEY J. D., WILLIAMS A. TH., CAMM J. D., COCHRAN J. J., *Statistiques pour l'économie et la gestion.* 5e édition. Traduction de la 7e édition américaine par Cl. Borsenberger

AUREZ V., GEORGEAULT L., *Économie circulaire. Système économique et finitude des ressources*

BÉNASSY-QUÉRÉ A., COEURÉ B., JACQUET P., PISANI-FERRY J., *Politique économique.* 4e édition

BEREND IVAN T., *Histoire économique de l'Europe du XXe siècle*
traduction de la 1re édition anglaise par Amandine Nguyen

BERGSTROM T., VARIAN H., *Exercices de microéconomie – 1. Premier cycle. Notions fondamentales.* 3e édition
traduction de la 5e édition américaine par A. Marciano

BERGSTROM T., VARIAN H., *Exercices de microéconomie – 2. Premier cycle et spécialisation.* 2e édition française
traduction de la 5e édition américaine par J.-M. Baland, S. Labenne et Ph. Van Kerm
avec la collaboration scientifique d'A. Marciano.

BESANKO, DRANOVE, SHANLEY, SCHAEFER, *Principes économiques de stratégie*

BILEK A., HENRIOT A., *Analyse conjoncturelle pour l'entreprise. Observer, comprendre, prévoir*

BISMANS F., *Mathématiques pour l'économie – Volume 1. Fonctions d'une variable réelle*

BOUTHEVILLAIN C., DUFRÉNOT G., FROUTÉ PH., PAUL L., *Les politiques budgétaires dans la crise. Comprendre les enjeux actuels et les défis futurs*

BOUTILLIER S., PEAUCELLE I., UZUNIDIS D., *L'économie russe depuis 1990*

BURDA M., WYPLOSZ C., *Macroéconomie. À l'échelle européenne.* 6e édition
traduction de la 6e édition anglaise par Stanislas Standaert

BRIEC W., PEYPOCH N., *Microéconomie de la production. La mesure de l'efficacité et de la productivité*

CADORET I., BENJAMIN C., MARTIN F., HERRARD N., TANGUY S., *Économétrie appliquée.* 2e édition
Méthodes, Applications, Corrigés

CAHUC P., ZYLBERBERG A., *Le marché du travail*

CAHUC P., ZYLBERBERG A., *Économie du travail. La formation des salaires et les déterminants du chômage*

CARLTON D. W., PERLOFF J. M., *Économie industrielle*, traduction de la 2e édition américaine par F. Mazerolle. 2e édition

CARTELIER J., *L'économie de Keynes*

CAVES R. E., FRANKEL J. A., JONES R. W., *Commerce international et paiements,*
traduction de la 9e édition américaine par M. Chiroleu-Assouline

CAYATTE J.-L., *Introduction à l'économie de l'incertitude*

COLLECTIF, *Économie sociale. Enjeux conceptuels, insertion par le travail et services de proximité*

COMMISSARIAT GÉNÉRAL DU PLAN, *L'intégration régionale. Une nouvelle voie pour l'organisation de l'économie mondiale ?*

CORNET B. et TULKENS H. (Éds), *Modélisation et décisions économiques*

CORNUEL D., *Économie immobilière et des politiques du logement*

CÔTÉ D., *Les holdings coopératifs. Évolution ou transformation définitive ?*

CRÉPON B., JACQUEMET N., *Économétrie : méthode et applications*

CUTHBERTSON K., *Économie financière quantitative. Actions, obligations et taux de change,*
traduction de la 1re édition anglaise par C. Puibasset

DARREAU PH., *Croissance et politique économique*

DE CROMBRUGGHE A., *Introduction aux principes de l'économie. Choix et décisions économiques.* 2e édition

DE BANDT O., DRUMETZ FR., PFISTER CHR., *Stabilité financière*

DEFFAINS B., LANGLAIS É., *Analyse économique du droit. Principes, méthodes, résultats*

DEFOURNY J., *Démocratie coopérative et efficacité économique. La performance comparée des SCOP françaises*

DEFOURNY J., NYSSENS M. (sous la direction de), *Économie sociale et solidaire. Socioéconomie du 3e secteur*

DEFOURNY J., DEVELTERE P., FONTENEAU B. (Éds), *L'économie sociale au Nord et au Sud*

DEFOURNY J., MONZON CAMPOS J.L. (Éds), *Économie sociale/The Third Sector. Entre économie capitaliste et économie publique/Cooperative Mutual and Non-profit Organizations*

DEFRAIGNE J. CHR., NOUVEAU P., *Introduction à l'économie européenne*. 2e édition

DE GRAUWE P., *Économie de l'intégration monétaire*, traduction de la 3e édition anglaise par M. Donnay

DE GRAUWE P., *La monnaie internationale. Théories et perspectives*, traduction de la 2e édition anglaise par M.-A. Sénégas

DEISS J., GUGLER PH., *Politique économique et sociale*

DE KERCHOVE A.-M., GEELS TH., VAN STEENBERGHE V., *Questions à choix multiple d'économie politique*. 3e édition

DE MELO J., GRETHER J.-M., *Commerce international. Théories et applications*

DEVELTERE P., *Économie sociale et développement. Les coopératives, mutuelles et associations dans les pays en voie de développement*

DRÈZE J., *Pour l'emploi, la croissance et l'Europe*

DRUMETZ F., PFISTER C., SAHUC J.-G., *Politique monétaire*. 2e édition

DUPRIEZ P., OST C., HAMAIDE C., VAN DROOGENBROECK N., *L'économie en mouvement. Outils d'analyse de la conjoncture*. 2e édition

ESCH L., *Mathématique pour économistes et gestionnaires*. 4e édition

ESSAMA-NSSAH B., *Inégalité, pauvreté et bien-être social. Fondements analytiques et normatifs*

GAZON J., *Politique industrielle et industrie*
Volume 1. Controverses théoriques. Aspects légaux et méthodologie

GILLIS M. *et al.*, *Économie du développement*, traduction de la 4e édition américaine par B. Baron-Renault

GODARD O., *Environnement et développement durable. Une approche méta-économique*

GOMEZ P.-Y., KORINE HARRY, *L'entreprise dans la démocratie, Une théorie politique du gouvernement des entreprises*

GUJARATI D. N., *Économétrie*, traduction de la 4e édition américaine par B. Bernier

HANSEN J.-P. – PERCEBOIS J., *Énergie. Économie et politiques*. 2e édition

HARRISON A., DALKIRAN E., ELSEY E., *Business international et mondialisation. Vers une nouvelle Europe*

HEERTJE A., PIERETTI P., BARTHÉLEMY PH., *Principes Analyse conjoncturelle pour l'entreprise. Observer, comprendre, prévoir d'économie politique*. 4e édition

HINDRIKS J., *Gestion publique. Théorie et pratique*

HIRSHLEIFER J., GLAZER A., HIRSHLEIFER D., *Microéconomie : théories et applications. Décision, marché, formation des prix et répartition des revenus*

JACQUEMIN A., TULKENS H., MERCIER P., *Fondements d'économie politique*. 3e édition

JACQUEMIN A., PENCH L. R. (Éds), *Pour une compétitivité européenne. Rapports du Groupe Consultatif sur la Compétitivité*

JALLADEAU J., *Introduction à la macroéconomie. Modélisations de base et redéploiements théoriques contemporains*. 2e édition

JALLADEAU J., DORBAIRE P., *Initiation pratique à la macroéconomie. Études de cas, exercices et QCM*. 2e édition

JASKOLD GABSZEWICZ J., *Théorie microéconomique*. 2e édition

JAUMOTTE Ch., *Les mécanismes de l'économie*

JONES Ch. I., *Théorie de la croissance endogène*, traduction de la 1re édition américaine par F. Mazerolle

JURION B., *Économie politique*. 4e édition

JURION B., LECLERCQ A., *Exercices d'économie politique*

KOHLI U., *Analyse macroéconomique*

KRUGMAN P. R. et OBSTFELD M., *Économie internationale*. 4e édition
traduction de la 6e édition américaine par A. Hannequart et F. Leloup

KRUGMAN P., *L'économie auto-organisatrice*, traduction de la 1re édition américaine par F. Leloup. 2e édition

KRUGMAN P., WELLS R., *Macroéconomie*, traduction de la 4e édition américaine par L. Baechler. 3e édition

KRUGMAN P., WELLS R., *Microéconomie*, traduction de la 4e édition américaine par L. Baechler. 3e édition

LANDAIS B., *Leçons de politique budgétaire*

LANDAIS B., *Leçons de politique monétaire*

LECAILLON J.-D., LE PAGE J.-M., *Économie politique contemporaine*. 5e édition

LEHMANN P.-J., *Économie des marchés financiers*. 2e édition

LEMOINE M., MADIÈS P., MADIÈS T., *Les grandes questions d'économie et finance internationales. Décoder l'actualité*. 3e édition

LEROUX A., MARCIANO A., *Traité de philosophie économique*

LESUEUR J.-Y., SABATIER M., *Microéconomie de l'emploi. Théories et applications*

LÖWENTHAL P., *Une économie politique*

MANKIW G. N., *Macroéconomie*, traduction de la 9e édition américaine par Jihad C. El Naboulsi. 7e édition

MANKIW G N., TAYLOR M. P., *Principes de l'économie*, traduction d'Élise Tosi. 4e édition

MANSFIELD E., *Économie managériale. Théorie et applications*, traduction et adaptation de la 4e édition américaine par B. Jérôme

MASSÉ G., THIBAUT Fr., *Intelligence économique. Un guide pour une économie de l'intelligence*

MARCIANO A., *Éthiques de l'économie. Introduction à l'étude des idées économiques*

MILGROM P., ROBERTS J., *Économie, organisation et management*

MONNIER L., THIRY B. (Éds), *Mutations structurelles et intérêt général. Vers quels nouveaux paradigmes pour l'économie publique, sociale et coopérative ?*

MUELLER C. D., FACCHINI F., FOUCAULT M., FRANÇOIS A., MAGNI-BERTON R., MELKI M., *Choix publics. Analyse économique des décisions publiques*

NORRO M., *Économies africaines. Analyse économique de l'Afrique subsaharienne*. 2e édition

PERKINS D. H., RADELET S., LINDAUER D. L., *Économie du développement*. 3e édition

PROMEURO, *L'Euro pour l'Europe. Des monnaies nationales à la monnaie européenne*. 2e édition

RASMUSEN E., *Jeux et information. Introduction à la théorie des jeux*, traduction de la 3e édition anglaise par F. Bismans

SALVATORE D. C., *Économie internationale*, traduction de la de la 9e édition américaine par Fabienne Leloup et Achille Hannequart

SHAPIRO C., VARIAN H. R., *Économie de l'information. Guide stratégique de l'économie des réseaux*, traduction de la 1re édition américaine par F. Mazerolle

SHILLER J. R., *Le nouvel ordre financier. La finance moderne au service des nouveaux risques économiques*, traduction de la 1re édition américaine par Paul-Jacques Lehmann

SIMON C. P., BLUME L., *Mathématiques pour économistes*, traduction de la 1re édition américaine par G. Dufrenot, O. Ferrier, M. Paul, A. Pirotte, B. Planes et M. Seris

SINN G., SINN H. W., *Démarrage à froid. Une analyse des aspects économiques de l'unification allemande*, traduction de la 3e édition allemande par C. Laurent

STIGLITZ J. E., LAFAY J.-D., ROSENGARD J. K., *Économie du secteur public*

STIGLITZ J. E., WALSH C. E., LAFAY J.-D., *Principes d'économie moderne*. 3e édition, traduction de la 3e édition américaine par F. Mayer

SZPIRO D., *Économie monétaire et financière*.

VARIAN H., *Introduction à la microéconomie*. 8e édition, traduction de la 9e édition américaine par B. Thiry

VARIAN H., *Analyse microéconomique*, traduction de la 3e édition américaine par J.-M. Hommet. 2e édition

VAN DER LINDEN B. (Éd.), *Chômage. Réduire la fracture*

WICKENS M., *Analyse macroéconomique approfondie. Une approche par l'équilibre général dynamique*

WOOLDRIDGE J., *Introduction à l'économétrie. Une approche moderne*. 2e édition

ZÉVI A., MONZÓN CAMPOS J.-L., *Coopératives, marchés, principes coopératifs*

NOTES

NOTES

NOTES

NOTES

NOTES

Achevé d'imprimer en juillet par Unigraf en Espagne
N° éditeur : 2018_0048

La référence en économétrie !

Jeffrey M. Wooldridge est professeur d'économie à l'Université d'État du Michigan (MSU) où il enseigne depuis 1991. De 1986 à 1991, il a été professeur d'économie au Massachusetts Institute of Technology (MIT). Il a obtenu sa licence en économie et informatique à l'Université de Californie à Berkeley en 1982, et sa thèse de doctorat en économie à l'Université de Californie à San Diego en 1986. Le professeur Wooldridge a publié de nombreux articles dans des revues de renommée internationale, ainsi que plusieurs chapitres de livres.

Pierre André est maître de conférences à l'Université de Cergy-Pontoise.

Michel Beine est professeur à l'Université du Luxembourg.

Sophie Béreau est professeur à l'Université de Namur et à l'Université catholique de Louvain.

Maëlys de la Rupelle est maître de conférences à l'Université de Cergy-Pontoise.

Alain Durré est professeur à l'IESEG School of Management.

Jean-Yves Gnabo est professeur à l'Université de Namur.

Cédric Heuchenne est professeur à l'Université de Liège.

Marion Leturcq est chercheur à l'Institut National d'Études Démographiques.

Mikael Petitjean est professeur à l'IESEG School of Management et l'Université catholique de Louvain.

En recourant à de nombreuses applications empiriques, ce **manuel d'introduction, dans sa seconde édition**, réussit l'exploit de **simplifier la présentation de l'économétrie** sans renoncer aux exigences de rigueur et de cohérence requises au niveau universitaire. Les méthodes économétriques sont présentées avec l'objectif de répondre à des **questions pratiques** liées à l'analyse du comportement des agents économiques, l'évaluation de politiques publiques ou la réalisation de prévisions.

Devenu une référence dans le monde anglo-saxon, cet ouvrage permet de comprendre et d'interpréter les hypothèses d'un modèle à la lumière de **nombreuses applications empiriques**. L'ouvrage distingue clairement le type de données analysées. Non seulement, il couvre les données en coupe transversale et les séries chronologiques, mais il aborde également les **données de panel** dont l'utilisation est devenue très fréquente aujourd'hui. Ce livre offre également une introduction aux **modèles à variable dépendante limitée** qui sont d'une grande utilité en économie appliquée et en gestion.

Chaque chapitre contient un **large éventail d'exercices**, dont un grand nombre repose sur l'utilisation de **bases de données économiques** disponibles sur le web. Le lecteur peut ainsi reproduire les nombreux exemples empiriques développés dans les chapitres de l'ouvrage et maîtriser toutes les étapes de la modélisation économétrique.

Cet ouvrage intéressera non seulement les **étudiants et professeurs de premier cycle universitaire**, mais également les **étudiants de Master** et les **praticiens** de l'économie.

Conception graphique : Primo&Primo®

ISSN : 2030-501X
ISBN : 978-2-8073-0683-7

deboeck SUPÉRIEUR

www.deboecksuperieur.com