Introduction à l'économétrie

Une approche moderne

Jeffrey M. Wooldridge

Traduction de la 5e 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

Introduction à l'économétrie

OUVERTURES ◂▸ ÉCONOMIQUES

Introduction à l'économétrie

Une approche moderne

Jeffrey M. Wooldridge

Traduction de la 5e 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

OUVERTURES ÉCONOMIQUES

de boeck

Ouvrage original :
Introductory Econometrics. A modern Approach, 5th edition, by Jeffrey M. Wooldridge

1re édition

Imprimé en Belgique

Dépôt légal :
Bibliothèque nationale, Paris : juin 2015 ISSN 2030-501X
Bibliothèque royale de Belgique, Bruxelles : 2015/13647/007 ISBN : 978-2-8041-7131-5

SOMMAIRE

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 ?

De nouveaux exercices ont été ajoutés dans la quasi-totalité des chapitres. Il s'agit d'exercices informatiques qui requièrent la manipulation de base de données, dont plusieurs sont inédits, de simulations informatiques qui permettent d'étudier les propriétés de l'estimateur des MCO, ou encore de problèmes plus exigeants sur le plan formel, pour lesquels une démonstration est nécessaire.

J'ai effectué plusieurs changements dans différents chapitres. Dans le chapitre 3, j'ai approfondi la discussion sur la multicolinéarité et les facteurs d'inflation de variance, que j'avais introduite dans la quatrième édition. Le chapitre 3 contient également une nouvelle section sur la terminologie que devraient utiliser les chercheurs lorsqu'ils discutent des équations estimées par les MCO. Pour les étudiants qui découvrent l'économétrie, il est important de bien comprendre la différence entre un modèle et une méthode d'estimation ; les étudiants devront s'en souvenir lorsqu'ils étudieront des procédures plus sophistiquées et réaliseront des travaux empiriques.

Le chapitre 5 inclut une discussion plus intuitive des échantillons de grande taille. Un point sur lequel j'insiste est que la taille de l'échantillon influence la distribution des moyennes d'échantillon ; elle n'a pas d'impact sur les distributions issues de la population, par définition. Le chapitre 6 offre un examen plus approfondi des transformations logarithmiques appliquées aux proportions. Ce chapitre contient également une liste exhaustive des éléments à prendre en considération lorsqu'il s'agit de choisir la forme fonctionnelle la plus appropriée parmi les plus courantes, c'est-à-dire la fonction logarithmique, la fonction quadratique et les termes d'interaction.

Le chapitre 7 inclut deux nouveautés. En premier lieu, je montre que le test de Chow peut être construit à partir de la somme des carrés des résidus lorsque l'hypothèse nulle autorise une variation de la constante entre groupes. En second lieu, dans la section 7.7, j'offre une interprétation simple et plus réaliste des modèles linéaires dont la variable dépendante est binaire.

Le chapitre 9 propose un examen plus poussé de l'utilité des variables de substitution dont l'inclusion dans une régression multiple peut permettre de remédier à l'omission de variables explicatives importantes. Mon espoir est de dissiper les doutes quant à la pertinence d'inclure une variable de substitution alors qu'elle peut introduire de la multicolinéarité dans la régression. Dans ce chapitre, j'ai également approfondi l'analyse des estimations obtenues par les moindres déviations absolues. Deux nouveaux problèmes sont abordés : l'un sur la détection du biais d'omission, l'autre sur l'hétéroscédasticité et l'estimation par les moindres déviations absolues. Ces deux sujets stimuleront la réflexion des étudiants consciencieux.

À la fin du chapitre 13, l'annexe comprend dorénavant une discussion des écarts-types robustes à l'autocorrélation et à l'hétéroscédasticité dans le cadre des modèles de données de panel estimés en différence première. Le calcul de ces écarts-types est fréquent dans les études appliquées en microéconomie, qui reposent souvent sur l'utilisation de données de panel. L'aspect théorique de cette analyse dépasse la portée de ce manuel d'introduction ; il est néanmoins facile d'en saisir l'idée de base. Une discussion similaire est disponible dans l'annexe du chapitre 14, dans le contexte des modèles de panel estimés avec effets fixes et aléatoires. Le chapitre 14 contient également une nouvelle section (14.3) sur l'estimation des modèles de données de panel avec « effets aléatoires corrélés », qui permettent également de tenir compte de l'hétérogénéité non observée. Bien que ce sujet soit plus difficile à

comprendre, il permet de synthétiser l'analyse des méthodes à effets fixes ou aléatoires, et de mieux comprendre l'importance des tests de spécification qui sont souvent utilisés dans les travaux empiriques.

Le chapitre 15 sur l'estimation des variables instrumentales a, lui aussi, été enrichi. Il souligne l'importance de bien vérifier, dans les équations de forme réduite, les signes des coefficients propres aux variables instrumentales ; il inclut une discussion de l'interprétation que la forme réduite peut avoir à l'égard de la variable dépendante. Comme au chapitre 3 (dans le cadre des MCO), j'insiste sur le fait que les variables instrumentales ne sont pas un « modèle », mais une méthode d'estimation.

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

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

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.

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 première 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.

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 :

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

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 certains 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 :

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

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 :

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

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 :

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

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

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

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 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 chercher 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.RAW). 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 de la table 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
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

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

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.

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.

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

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

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.

TABLE DES MATIÈRES

Partie 3 Thèmes avancés

La nouvelle référence en économétrie !

Introduction à l'économétrie

En recourant à de nombreuses applications empiriques, ce **manuel d'introduction** 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.

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 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'Université catholique de Louvain.

WOOLDRIDGE
ISBN 978-2-8041-7131-5
ISSN 2030-501X

www.deboecksuperieur.com

9 782804 171315