OUVERTURES ÉCONOMIQUES

Jeffrey **M. WOOLDRIDGE**

Traduction de la 7e édition américaine par **M. Beine**, **S. Béreau**, **M. de la Rupelle**, **J.-Y. Gnabo**, **C. Heuchenne**, **M. Leturcq** et **M. Petitjean**

3e édition

Introduction à l'économétrie

Une approche moderne

LA RÉFÉRENCE

- Cours complet
- Résumés de chapitres
- Études de cas
- Questions de réflexion et corrigés

OFFERT

+ EN LIGNE

Pour les étudiants
+ de 450 exercices

Pour les enseignants
PPT, corrigés des exercices

deboeck SUPÉRIEUR

Introduction à l'économétrie

3e édition

Une approche moderne

Jeffrey M. **Wooldridge**

Traduction de la 7e édition américaine
par M. Beine, S. Béreau, M. de la Rupelle, J.-Y. Gnabo, C. Heuchenne, M. Leturcq et M. Petitjean

deboeck SUPÉRIEUR

Ouvrage original :
Introductory Econometrics. A Modern Approach, 7th edition by Jeffrey M. Wooldridge

OFFERT : Ressources numériques :

Des ressources complémentaires sont disponibles à l'adresse https://www.deboecksuperieur.com/site/329775 :

Étudiants :
- Exercices et exercices sur ordinateur.
- Le corrigé de certains exercices (en anglais).
- Fichiers de données (en anglais)

Enseignants : Identifiez-vous à l'adresse https://www.deboecksuperieur.com/site/329775 et accédez à des ressources complémentaires en anglais :
- PowerPoint
- Corrigés de tous les exercices

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

© De Boeck Supérieur s.a., 2023 3e édition
Rue du Bosquet, 7 – B-1348 Louvain-la-Neuve
Pour la traduction en langue française.

Dépôt légal :
Bibliothèque nationale, Paris : mai 2023 ISSN : 2030-501X
Bibliothèque royale de Belgique, Bruxelles : 2023/13647/070 ISBN : 978-2-8073-2977-5

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 septième édition anglaise conserve l'organisation globale de la précédente. 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.

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 des 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.

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.

QUOI DE NEUF DANS CETTE ÉDITION ?

J'ai inclus de nouveaux exercices dans de nombreux chapitres, y compris dans les annexes. Plusieurs nouveaux exercices sur ordinateur requièrent l'utilisation de nouvelles bases de données, comme celle sur les performances des équipes masculines de basket-ball universitaire aux États-Unis. J'ai également ajouté des problèmes plus complexes nécessitant des démonstrations.

Il y a plusieurs changements importants dans le texte. La notion de variables explicatives binaires, ou dichotomiques, est introduite dès le chapitre 2 pour montrer que l'estimation par les moindres carrés ordinaires permet d'estimer la différence de moyennes entre deux sous-groupes d'une population. En introduisant les variables binaires plus rapidement dans le livre, l'enseignant est en mesure de recourir à une plus grande diversité d'exemples empiriques. Il peut également introduire de manière formelle l'analyse contrefactuelle

qui est au cœur de la littérature moderne sur l'estimation des effets causaux. Cette approche moderne de l'inférence causale, basée sur l'analyse des résultats potentiels, apparaissait déjà dans les éditions précédentes, mais les chapitres 2, 3, 4 et 7 incluent désormais explicitement de nouvelles sections sur ce sujet. Une illustration concrète de l'utilisation d'une variable explicative binaire consiste à évaluer la mise en place d'un programme de formation par les pouvoirs publics ou une entreprise, puisqu'il y a participation ou pas au programme. Un enseignant peut ignorer cette thématique s'il le souhaite sachant que ces nouveaux éléments sont intégrés dans des sections aisément identifiables. Plusieurs exercices abordent l'analyse contrefactuelle et permettent aux enseignants d'approfondir cette matière.

Le chapitre 3 comprend une nouvelle section sur l'utilisation de la régression multiple dans différents domaines, comme la prévision, l'efficience des marchés, ou l'estimation de l'effet causal du traitement. Après avoir couvert les aspects plus techniques liés à la mécanique des moindres carrés ordinaires (MCO), cette section offre aux étudiants la possibilité de mieux appréhender la portée de la régression multiple. Si nécessaire, cette section peut être ignorée sans perte de continuité. Une nouvelle section dans le chapitre 7 poursuit la discussion sur l'analyse contrefactuelle lorsque l'effet du traitement est non constant. Cette section offre une belle illustration de l'estimation de différentes fonctions de régression pour deux sous-groupes tirés au sein d'une population. Dans ce chapitre, il y a également de nouveaux problèmes qui permettent à l'étudiant d'acquérir une plus grande expérience dans la manière d'ajuster la régression pour estimer l'effet causal.

Dans le chapitre 9, le changement le plus notable porte sur l'utilisation d'indicateurs de données manquantes. Les hypothèses qui sous-tendent cette méthode sont expliquées plus en détail que dans l'édition précédente.

Le chapitre 12 a été réorganisé pour tenir compte de la manière dont est traité aujourd'hui le problème de corrélation sérielle qui affecte les erreurs dans les modèles de régression sur séries chronologiques. Le chapitre traite d'abord de l'ajustement des écarts-types obtenus par les MCO, pour tenir compte de la présence de corrélation sérielle de forme générale. Le plan du chapitre suit désormais celui du chapitre 8. Les deux chapitres mettent l'accent sur l'estimation par les MCO en visant à rendre l'inférence robuste à la violation des hypothèses du modèle de régression linéaire classique. La correction des écarts-types par la méthode des moindres carrés généralisés n'est abordée qu'ensuite, après le traitement des tests de corrélation sérielle.

Les chapitres plus pointus ont également fait l'objet de plusieurs améliorations. En restant très accessible, le chapitre 13 traite dorénavant des extensions que l'on peut apporter à l'analyse par la différence dans les différences, notamment en autorisant l'inclusion de plusieurs groupes de contrôle, de plusieurs périodes de temps et même de tendances spécifiques à un groupe. Le chapitre comprend également une discussion plus approfondie du calcul des écarts-types robustes à la présence de corrélation sérielle lors de l'estimation par différence première dans des modèles de panel.

Le chapitre 14 inclut une discussion plus fine concernant l'estimation des modèles de données de panel à l'aide d'effets fixes, d'effets aléatoires et ou d'effets aléatoires corrélés. L'utilisation d'effets aléatoires corrélés en présence de données manquantes est examinée plus en détail, tout comme la manière de prendre en considération les formes fonctionnelles générales, telles que les formes quadratiques et les interactions entre variables, qui sont déjà abordées dans le chapitre sur les données en coupe transversale dans le chapitre 6. Une section plus étoffée porte également sur l'analyse des modèles de panel visant à évaluer les programmes ou les politiques que peuvent mener des acteurs comme les entreprises ou les États.

Le chapitre 16, qui porte sur les modèles d'équations simultanées, établit désormais un lien explicite entre l'analyse causale basée sur les résultats potentiels et la spécification des modèles à équations simultanées.

Le chapitre 17 inclut une nouvelle discussion sur la méthode qui consiste à ajuster la régression pour estimer les effets causaux (du traitement) lorsque la variable dépendante (de résultat) présente des caractéristiques particulières, par exemple lorsqu'elle est une variable binaire. Le lecteur est ensuite invité à explorer

l'utilisation des modèles logit et probit dans le but d'obtenir des estimations plus fiables de l'effet moyen du traitement en recourant à des estimations séparées pour chaque groupe.

Le chapitre 18 offre une analyse plus détaillée de la manière de calculer l'écart-type de l'intervalle d'une prévision. Cela aidera le lecteur à affiner sa compréhension de l'incertitude dans ce domaine.

CARACTÉRISTIQUES DE L'OUVRAGE

De nombreuses questions de réflexion « pour aller plus loin », assez brèves, sont insérées dans le corps même des chapitres de ce manuel. Ces questions, dont les réponses sont reprises en fin d'ouvrage, 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 (ou problèmes) et exercices sur ordinateur, disponibles en ligne, 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.

1 Ces bases de données sont disponibles uniquement en anglais.

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

À propos de l'auteur

Jeffrey M. Wooldridge est professeur émérite 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

SOMMAIRE

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 utilisation courante 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 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 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 :

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

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 :

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

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 β_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 :

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

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.

1.3a 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

1.3b 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.

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

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

obsno	year	avgmin	avgcov	prunemp	prgnp
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.

1.3c 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 On rencontre parfois les termes « coupes transversales empilées », « coupes transversales regroupées » ou « coupes transversales agrégées » (note de la traduction).

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
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.

1.3d 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, ou observations. 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.

1.3e 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 CAUSALITÉ, *CETERIS PARIBUS*, ET LE RAISONNEMENT CONTREFACTUEL

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 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.

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.

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.

La notion de *ceteris paribus* peut également être appréhendée par un **raisonnement en termes de contrefactuels**, qui est un cadre d'analyse devenu structurant pour étudier différentes formes d'interventions, par exemple l'impact de changements de politiques. L'idée est d'imaginer une unité économique, comme un individu ou une entreprise, dans deux états du monde. Par exemple, imaginons que l'on souhaite étudier l'impact d'une formation professionnelle sur les revenus d'un salarié. Pour chaque salarié de la population étudiée, on peut imaginer ce que seraient ses revenus à venir dans deux états du monde : s'il ou elle a participé au programme de formation professionnelle ou s'il ou elle n'y a pas participé. Considérer les **résultats contrefactuels** (appelés également *résultats potentiels*, ou *résultats hypothétiques*) permet de se placer facilement dans un cadre « toutes choses égales par ailleurs », puisque l'expérience de pensée en termes de contrefactuel s'applique à chaque individu séparément. On peut alors penser la causalité comme le fait que le résultat (ici, les revenus du travail) n'est pas le même dans les deux états du monde, au moins pour certains salariés. Dans les faits, on n'observe chaque salarié que dans un seul état du monde, ce qui soulève d'importants problèmes d'estimation, mais il s'agit là d'un problème différent de la question de la définition de la causalité. Nous introduirons formellement un dispositif pour discuter des résultats contrefactuels dans le chapitre 2.

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. En revanche, pour chaque exemple, nous discuterons quels autres facteurs nous aimerions garder constants, et nous agrémenterons la discussion d'un raisonnement contrefactuel. Pour chaque exemple, nous allons montrer qu'inférer la causalité serait relativement aisé s'il était possible de 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.

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).

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.

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. Voyez ici l'élément de raisonnement contrefactuel : on imagine le salaire d'un individu pour différents niveaux du nombre d'années d'étude, c'est-à-dire dans différents états de la nature. Au final, les données ne nous permettent d'observer qu'un seul état de la nature, c'est-à-dire le nombre d'années d'étude avec lequel l'individu va se retrouver, qui est le résultat d'un mélange complexe de capacités intellectuelles, de contribution parentale et d'influence de la société.

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 expérience de pensée similaire est d'expliciter la question en termes de résultats contrefactuels : que serait le taux de criminalité d'une ville donnée si l'on faisait varier ses effectifs de police ? 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 discuté les notions de causalité, *ceteris paribus*, et de contrefactuels. 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 nous l'avons vu, une façon de comprendre les exigences de l'analyse *ceteris paribus* est d'opérer une expérience de pensée dans laquelle une même unité économique se trouve dans deux états du monde différents, par exemple sujette à deux types de politiques différentes. 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. 17
Ceteris paribus p. 27
Données empilées (ou coupes transversales empilées) p. 24
Données en coupe transversale p. 20
Données expérimentales p. 16
Données non expérimentales p. 16
Données observationnelles (ou données non expérimentales) p. 16
Données de panel (ou données longitudinales) p. 25
Échantillonnage aléatoire p. 20
Effet causal p. 27
Fréquence des données p. 23
Modèle économétrique p. 19
Modèle économique p. 17
Raisonnement contrefactuel p. 28
Résultats contrefactuels p. 28
Séries chronologiques (ou série temporelle) p. 23

PARTIE

1

L'ANALYSE DE RÉGRESSION SUR DONNÉES EN COUPE TRANSVERSALE

SOMMAIRE

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

SOMMAIRE

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 :

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

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 :

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

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 :

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

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 de l'annexe B 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

$$\mathrm{E}(u|x) = \mathrm{E}(u). \qquad \textbf{[2.6]}$$

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**, $\mathrm{E}(u|x) = 0$. 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 $\mathrm{E}(apt|8)$ désigne l'aptitude innée moyenne du groupe de personnes qui ont suivi un enseignement pendant huit ans, et que $\mathrm{E}(apt|16)$ 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,

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

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 $\mathrm{E}(u|x) = 0$, on obtient

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

L'équation (2.8) correspond à la **fonction de régression de la population (FRP)**, $\mathrm{E}(y|x)$, 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 $\mathrm{E}(y|x)$, comme représenté sur la figure 2.1.

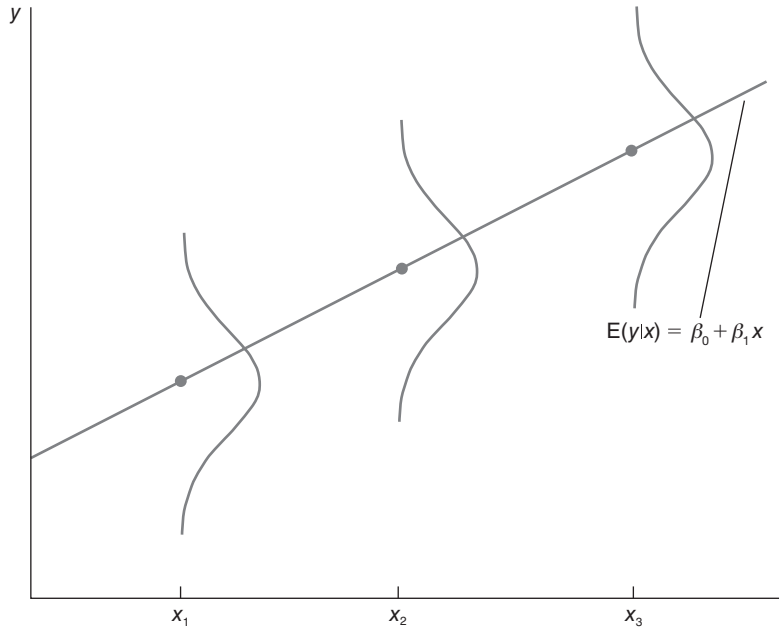


Figure 2.1 *E(y|x) en tant que fonction linéaire de x.*

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

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

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.

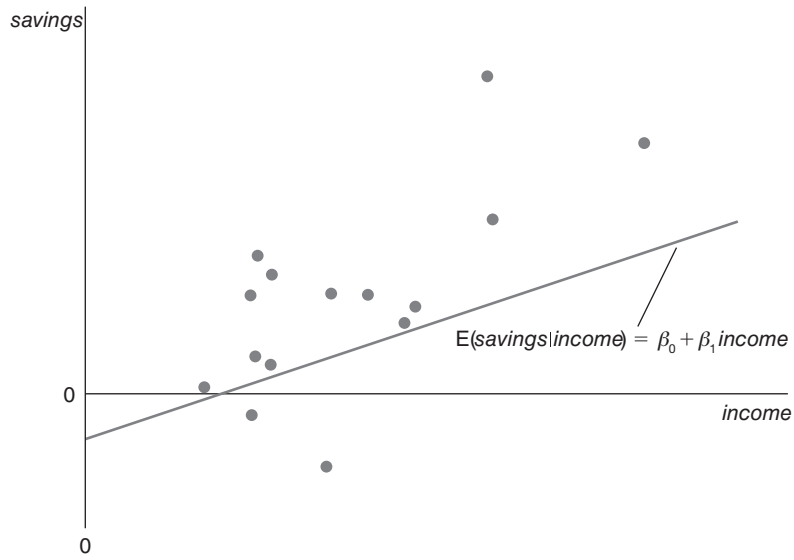


Figure 2.2 *Diagramme de dispersion de l'épargne (savings) et du revenu (income) pour 15 familles, et fonction de régression de la population* $E(\text{saving}\,|\,\text{income}) = \beta_0 + \beta_1\,\text{income}$.

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 de l'annexe B 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, les équations (2.12) et (2.13) doivent nous permettre d'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 de l'annexe C 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

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

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

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

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

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

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

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) dans l'annexe A],

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

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.

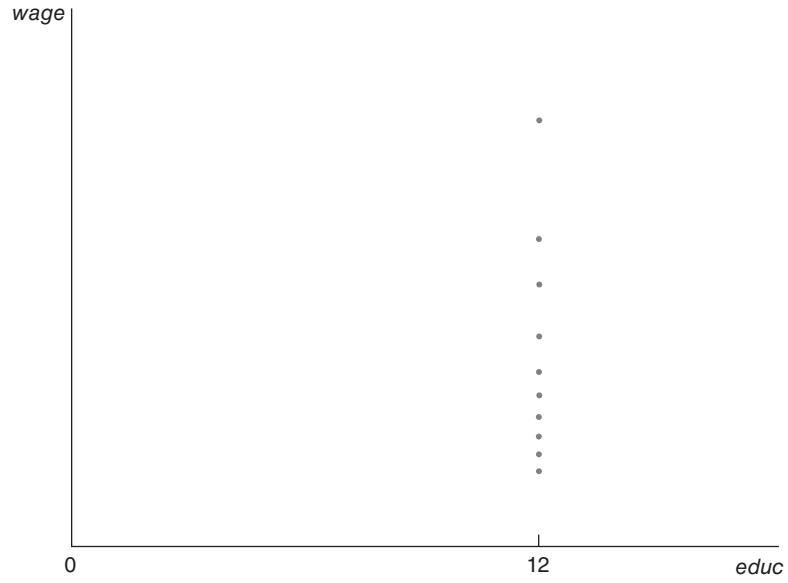


Figure 2.3 *Diagramme de dispersion du salaire et du niveau d'instruction lorsque* $\text{educ}_i = 12$ *pour tout* i.

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.

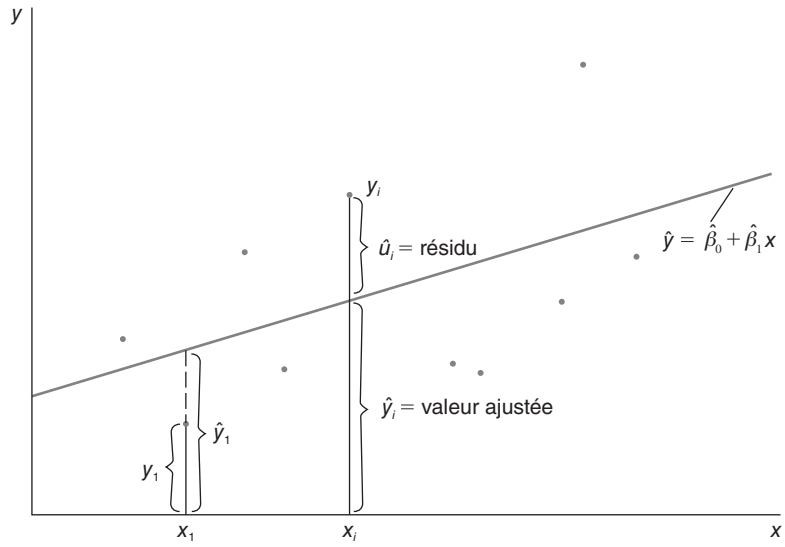


Figure 2.4 *Valeurs ajustées et résidus.*

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 9.6. 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 d'un grand nombre de 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.

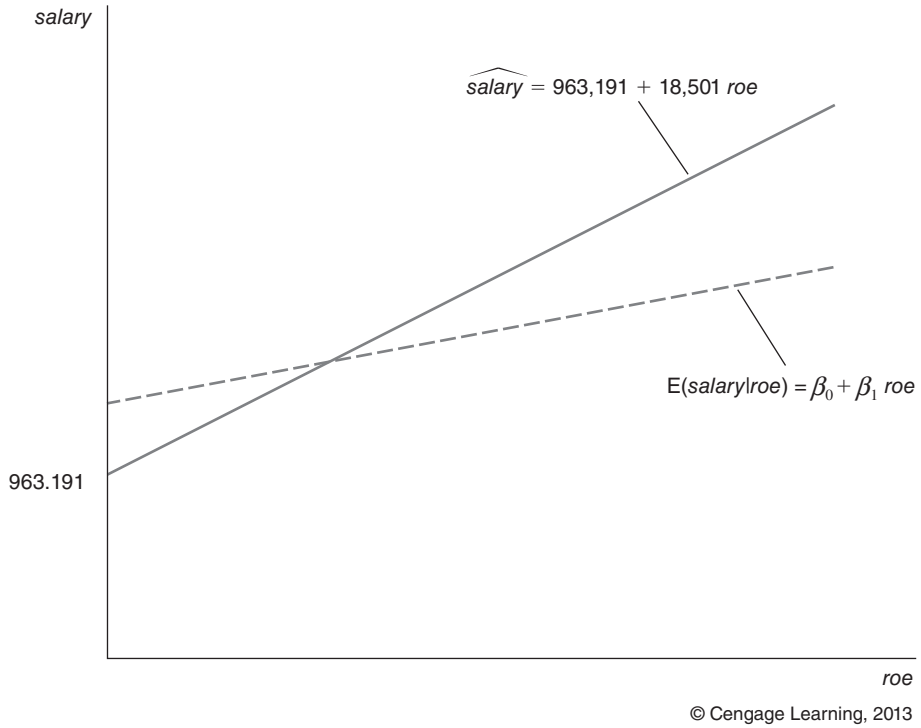


Figure 2.5 *Droite de régression des MCO,* $\widehat{salary} = 963{,}191 + 18{,}501\ roe$, *et fonction (inconnue) de régression de la population.*

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 24,90 USD en 2016, 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 :

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

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).

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.

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 2016 ? (*Astuce* : il y a suffisamment d'information disponible dans l'exemple 2.4 pour répondre à cette question).

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 ?

2.2a 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.

2.3a 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	$\widehat{salary}$	$\hat{u}$
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	$\widehat{salary}$	$\hat{u}$
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

2.3b 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 $(\overline{x}, \overline{y})$ 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 $\overline{x}$, la valeur ajustée de y sera égale à $\overline{y}$, 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 $\overline{y} = \overline{\hat{y}}$. 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 :

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

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

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

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 $\overline{\hat{y}} = \overline{y}$. 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 :

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

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

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

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.

2.3c 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.

2.4a 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 $\Delta roe = 1$ 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.

2.4b 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.

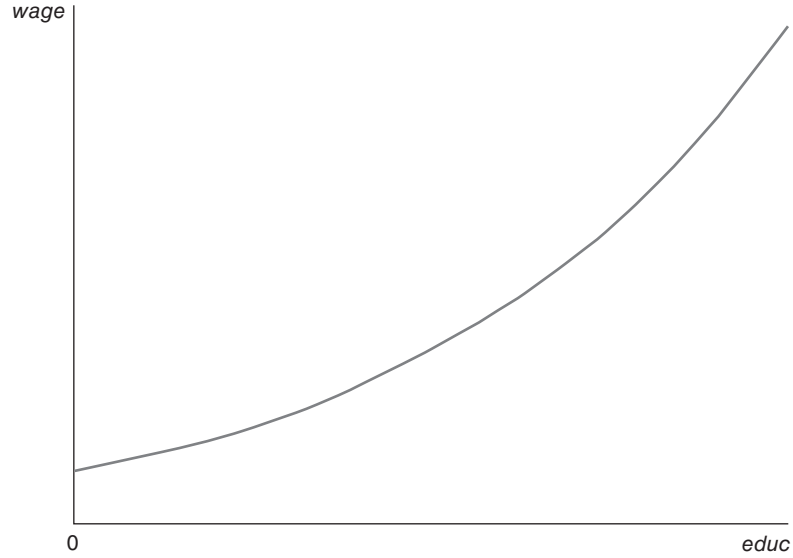


Figure 2.6 $wage = exp(\beta_0 + \beta_1 educ)$, *avec* $\beta_1 > 0$.

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

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

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

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

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 en ligne, 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 réfèrerons souvent dans les autres chapitres.

2.4c 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.)

2.5a 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 :

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

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.

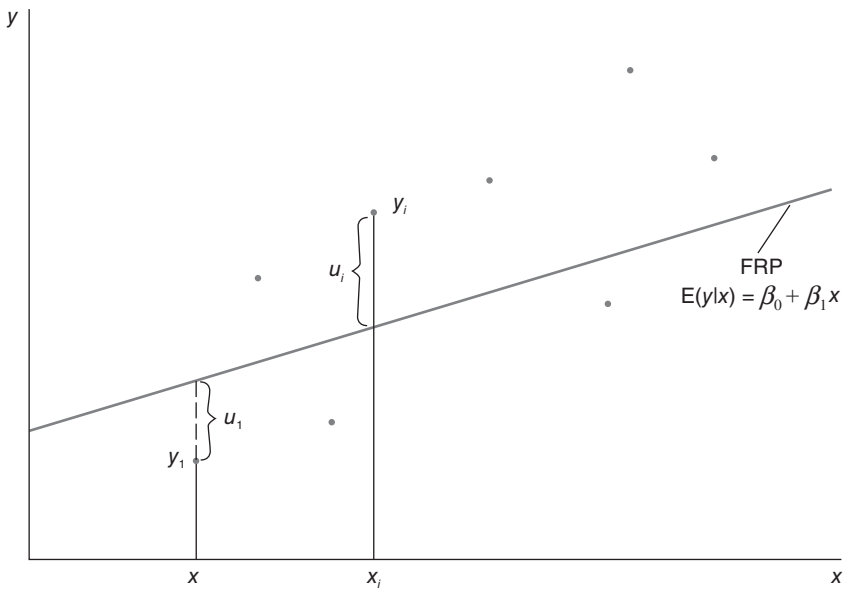


Figure 2.7 *Graphique de* $y = \beta_0 + \beta_1 x_i + u_i$.

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, …, 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, …, à y_n. Ensuite, un autre échantillon de y peut être constitué sur la base des *mêmes* valeurs de x_1, x_2, …, x_n (mais différentes valeurs de u_1, u_2, …, u_n). L'étape précédente peut à nouveau être répétée en utilisant les mêmes x_1, x_2, …, x_n (mais, à nouveau, différentes valeurs pour u_1, u_2, …, 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 $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i$ (voir l'annexe A), l'estimateur de la pente de l'équation (2.19) peut s'écrire sous la forme suivante :

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

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

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

où la variation totale de x_i est égale à $\text{SCT}_x = \sum_{i=1}^{n}(x_i - \overline{x})^2$. (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

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

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

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

où $d_i = (x_i - \overline{x})$. 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,

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

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

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

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.

2.5b 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 $\beta_1 > 0$.

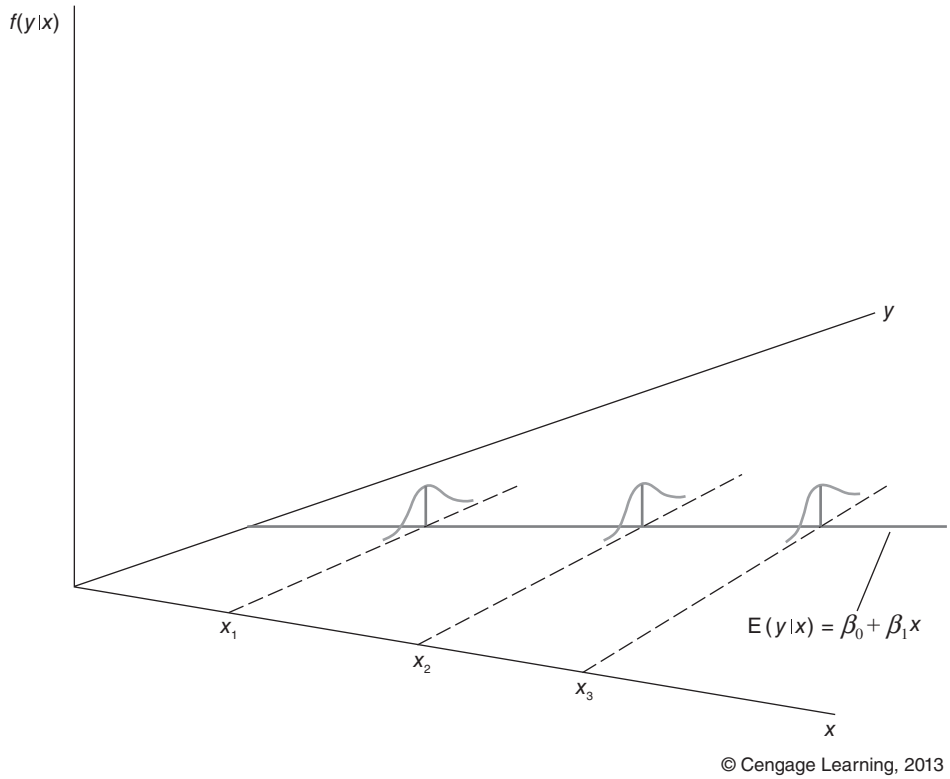


Figure 2.8 *Le modèle de régression simple sous l'hypothèse d'homoscédasticité.*

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.

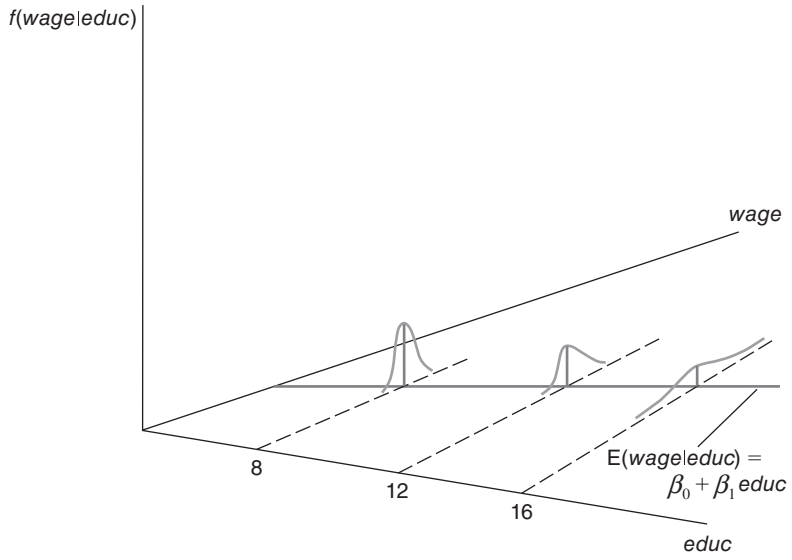


Figure 2.9 *Var(*wage|educ*) est une fonction croissante de* educ.

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,

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

et

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

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 :

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

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

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

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.

2.5c 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 : $0 = \overline{u} - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)\overline{x}$. Si cette égalité est soustraite de (2.59), cela donne $\hat{u}_i = (u_i - \overline{u}) - (\hat{\beta}_1 - \beta_1)(x_i - \overline{x})$. Par conséquent, $\hat{u}_i^2 = (u_i - \overline{u})^2 + (\hat{\beta}_1 - \beta_1)^2(x_i - \overline{x})^2 - 2(u_i - \overline{u})(\hat{\beta}_1 - \beta_1)(x_i - \overline{x})$. En sommant par rapport à i,

$\sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i=1}^{n}(u_i - \overline{u})^2 + (\hat{\beta}_1 - \beta_1)^2\sum_{i=1}^{n}(x_i - \overline{x})^2 - 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}u_i(x_i - \overline{x})$. 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 $\text{E}[\text{SCR}/(\text{n} - 2)] = \sigma^2$

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 $\hat{\sigma}_{\hat{u}}$. 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

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

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 :

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

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, $\overline{x} = 0$. 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 $\overline{y}$, 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 $\overline{y}$ vaut zéro. Dans un tel cas,

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

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,

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

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 $\overline{y}$ 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 $\overline{y}$. 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.

2.7 RÉGRESSION SUR VARIABLE EXPLICATIVE BINAIRE

Notre discussion a porté jusqu'ici sur des régressions dans lesquelles la variable explicative, x, avait une interprétation quantitative. Citons par exemple le niveau d'instruction, le rendement des capitaux propres d'une entreprise ou le pourcentage d'élèves pouvant bénéficier de repas scolaires subventionnés par l'État. Nous pouvons désormais interpréter le coefficient de la pente dans chacun de ces cas. Nous avons également discuté de l'interprétation du coefficient de la pente lorsque nous utilisons la transformation logarithmique de la variable dépendante, de la variable explicative, ou des deux.

La régression simple peut être également appliquée dans le cas où x est une **variable binaire**, souvent appelée **variable dichotomique** ou **variable « dummy »** en anglais. Comme son appellation le suggère, x ne prend alors que deux valeurs, zéro et un. Ces deux valeurs permettent de répartir chaque unité de la population dans l'un des deux groupes représentés par $x = 0$ et $x = 1$. Par exemple, nous pouvons utiliser une variable binaire pour décrire la participation d'un travailleur à un programme de formation professionnelle. Nous pouvons ensuite donner un nom descriptif à cette variable binaire, soit $train = 1$ si la personne participe à la formation professionnelle (ou « training » en anglais), soit $train = 0$ si elle n'y participe pas. Comme d'habitude, nous ajoutons un indice i de sorte que $train_i$ indique le statut de formation professionnelle d'une personne tirée au sort dans un échantillon de données.

Que signifie une équation de régression dont la variable dépendante est y, comme auparavant, mais dont la variable explicative x est binaire ? Considérons à nouveau l'équation

$$y = \beta_0 + \beta_1 x + u$$

où la variable x est désormais une variable binaire. Si nous appliquons l'hypothèse de l'espérance conditionnelle nulle de l'erreur RLS.4, alors nous obtenons

$$E(y|x) = \beta_0 + \beta_1 x + E(x) = \beta_0 + \beta_1 x, \qquad \textbf{[2.69]}$$

comme dans l'équation (2.8). La seule différence est que x ne prend que deux valeurs. En introduisant les valeurs 0 et 1 dans (2.69), on constate aisément que

$$E(y|x = 0) = \beta_0, \qquad \textbf{[2.70]}$$

$$E(y|x = 1) = \beta_0 + \beta_1. \qquad \textbf{[2.71]}$$

Il s'ensuit immédiatement que

$$\beta_1 = E(y|x = 1) - E(x = 0). \qquad \textbf{[2.72]}$$

En d'autres termes, β_1 est la différence entre la valeur moyenne de y dans la sous-population pour laquelle $x = 1$ et la valeur moyenne de y dans la sous-population pour laquelle $x = 0$. Comme dans toutes les analyses de régression simple, cette différence mesurée par β_1 est soit tout simplement descriptive, soit utilisée pour estimer l'effet causal d'une intervention, comme discuté dans la sous-section suivante.

À titre d'exemple, supposons que chaque travailleur payé à l'heure dans une industrie soit réparti dans l'une des deux catégories ethniques suivantes : les « Caucasiens » et les « autres ». Il s'agit évidemment d'une manière simpliste de catégoriser les travailleurs, mais elle a été utilisée dans plusieurs études de cas. Définissons la variable $white = 1$ lorsqu'une personne est classée comme caucasienne et 0 sinon. La variable *wage* représente le salaire horaire. Dans ce cas,

$$\beta_1 = E(wage|white = 1) - E(white = 0).$$

Le coefficient β_1 correspond à la différence de salaire horaire moyen entre les travailleurs blancs et non blancs. De manière équivalente,

$$E(wage|white) = \beta_0 + \beta_1 white.$$

Notons que β_1 correspond toujours à la différence de salaire moyen entre les « blancs » et les « non-blancs ». Soyons prudents néanmoins. Ce coefficient ne mesure pas nécessairement la discrimination salariale entre ces deux groupes puisqu'il existe de nombreux facteurs qui peuvent expliquer cette différence, comme le niveau d'éducation qui pourrait être différent, en moyenne, entre ces deux groupes.

Que la variable x soit binaire ou pas, la mécanique des MCO est toujours la même. Considérons un échantillon de taille n, $\{(x_i,y_i): i = 1, \ldots, n\}$. Les estimateurs de l'ordonnée à l'origine et de la pente des MCO sont toujours donnés par (2.16) et (2.19) respectivement ; les résidus ont toujours une moyenne nulle et ne sont pas corrélés avec les x_i de l'échantillon ; la définition du R-carré reste inchangée ; et ainsi de suite. Parce que x_i est binaire, les estimations des MCO ont néanmoins une interprétation à la fois simple et instructive. Soit $\bar{y}_0$, la moyenne des y_i quand $x_i = 0$ et $\bar{y}_1$, la moyenne quand $x_i = 1$. Le problème 2.13 vous demande d'ailleurs de montrer que

$$\hat{\beta}_0 = \bar{y}_0 \qquad \textbf{[2.73]}$$

$$\hat{\beta}_0 = \bar{y}_1 - \bar{y}_0. \qquad \textbf{[2.74]}$$

Dans l'exemple ci-dessus portant sur *wage* et *white*, si nous exécutons la régression suivante

$$wage_i \text{ sur } white_i,\ i = 1, \ldots, n$$

alors $\hat{\beta}_0 = \overline{wage}_0$, soit le salaire horaire moyen des « non-Caucasiens » ; $\hat{\beta}_1 = \overline{wage}_1 - \overline{wage}_0$, soit la différence entre le salaire horaire moyen des « Caucasiens » et des « non-Caucasiens ». De manière générale, l'équation (2.74) indique que la pente de la régression correspond à la différence entre deux moyennes ; il s'agit d'un estimateur classique en statistique inférentielle lorsque l'on compare deux groupes.

Les propriétés statistiques des MCO restent également identiques lorsque x est binaire pour la raison simple que les hypothèses n'excluent pas la possibilité que x soit binaire. L'hypothèse RLS.3 est respectée à condition que nous obtenions dans notre échantillon des zéros et des uns pour x_i. Dans l'exemple ci-dessus portant sur *wage* et *white*, nous devons disposer de données sur des « Caucasiens » et des « non-Caucasiens » afin d'estimer $\hat{\beta}_1$. C'est une évidence.

La principale préoccupation est l'hypothèse de l'espérance conditionnelle nulle de l'erreur, RLS.4, comme c'est toujours le cas pour les régressions simples. Dans de nombreux cas, cette condition ne sera pas respectée parce que la variable x est liée à d'autres facteurs qui affectent y, et que ces autres facteurs sont nécessairement inclus dans u dans une régression simple. Nous y avons fait allusion dans la discussion précédente portant sur les différences de salaire horaire moyen selon l'appartenance ethnique. Par exemple, l'éducation et l'expérience professionnelle sont deux variables qui affectent le salaire horaire et qui pourraient diverger selon les groupes. Considérons un autre exemple et supposons que nous disposons de données sur les résultats obtenus en mathématiques et en vocabulaire à un test, appelé SAT (« Scholastic Assessment Test »), que les étudiants doivent passer aux États-Unis avant d'entrer à l'université. Soit *sat*, la variable dépendante correspondant au résultat obtenu à ce test. La variables binaire x, que nous appellerons *cours*, est une variable binaire égale à 1 lorsque ces étudiants ont décidé de suivre un cours de préparation au SAT, 0 autrement. Très vraisemblablement, cette décision de suivre un cours est liée à d'autres facteurs qui expliquent le résultat obtenu au SAT, tels que le revenu familial et le niveau d'éducation des parents. Il est donc peu probable qu'une comparaison des résultats obtenus en moyenne au SAT entre ces deux groupes permette d'estimer correctement l'effet causal du cours de préparation sur le résultat obtenu au SAT. L'objectif de la sous-section suivante est précisément d'identifier les conditions particulières sous lesquelles une régression simple peut nous permettre d'estimer un effet causal.

2.7a Résultats potentiels, causalité et analyse des politiques

Après avoir introduit la notion de variable explicative binaire, il est temps de fournir un cadre formel pour l'étude des résultats potentiels, appelés également résultats contrefactuels ou hypothétiques, que nous avons brièvement expliqués au chapitre 1. Notre intérêt porte en particulier sur **l'effet causal**, appelé également **l'effet du traitement**.

Considérons le cas le plus simple où il n'existe que deux états du monde : soit une unité d'observation est soumise à une intervention, soit elle ne l'est pas. Les unités qui ne sont pas soumises à l'intervention constituent **le groupe de contrôle**, parfois appelé **groupe témoin**, et celles qui sont soumises à l'intervention constituent **le groupe de traitement**. En repartant du cadre d'analyse introduit dans le chapitre 1, supposons qu'il existe des résultats dans les deux états du monde pour chaque unité i de la population, soit $y_i(0)$ et $y_i(1)$. Il n'y a aucune unité qui existe dans les deux états du monde, mais nous pouvons néanmoins conceptualiser ces deux états. Par exemple, une personne peut décider de participer ou non à un programme de formation professionnelle. Le revenu de cette personne i est alors $y_i(1)$ si elle y participe et $y_i(0)$ si elle n'y participe pas.

L'effet causal de l'intervention pour l'unité i, plus communément appelé l'effet du traitement, correspond simplement à l'effet de cette intervention sur le revenu de l'unité i, soit

$$te_i = y_i(1) - y_i(0). \qquad \textbf{[2.75]}$$

L'effet du traitement, te_i, correspond à la différence entre ces deux résultats dont l'un des deux sera, par définition, toujours potentiel, hypothétique ou contrefactuel. Il est donc impossible d'observer cet effet pour une unité i bien précise. Par ailleurs, te_i peut être négatif, nul ou positif ; il peut être négatif pour certaines unités, positif ou nul pour d'autres.

Il est donc impossible d'estimer te_i pour chaque unité i. C'est la raison pour laquelle l'attention se porte sur l'estimation de l'**effet moyen du traitement** (EMT), également appelé l'**effet causal moyen** (ECM). Il s'agit tout simplement de la moyenne des effets du traitement sur l'ensemble de la population. L'EMT est d'ailleurs parfois appelé l'effet moyen du traitement de la population. Nous pouvons écrire le paramètre EMT comme suit :

$$\tau_{emt} = E[te_i] = E[y_i(1)] - E[y_i(0)], \qquad \textbf{[2.76]}$$

qui repose sur la linéarité de la valeur attendue. En considérant la population, nous pouvons aussi écrire $\tau_{emt} = E[y(1) - y(0)]$, où $y(1)$ et $y(0)$ sont deux variables aléatoires représentant les résultats potentiels dans la population.

Définissons maintenant x_i, une variable binaire caractérisant la participation au programme de formation professionnelle pour chaque personne i. Lorsque la personne participe à la formation, $x_i = 1$; lorsqu'elle n'y participe pas, $x_i = 0$. Le revenu y_i peut alors s'écrire comme suit :

$$y_i = (1 - x_i)y_i(0) + x_i y_i(1), \qquad \textbf{[2.77]}$$

qui sert de raccourci pour écrire $y_i = y_i(0)$ si $x_i = 0$ et $y_i = y_i(1)$ si $x_i = 1$. Étant donné un échantillon aléatoire de la population, cette équation montre que nous ne pouvons observer qu'un seul des $y_i(0)$ et $y_i(1)$.

Pour estimer l'EMT, il est utile de réarranger (2.77) comme suit :

$$y_i = y_i(0) + [y_i(1) - y_i(0)]x_i. \qquad \textbf{[2.78]}$$

Supposons un effet de traitement constant, soit $te_i = \tau$, une hypothèse généralement irréaliste mais très utile à ce stade de l'analyse. Nous la relâchons plus tard. Pour toutes les unités i, en fonction de (2.75), nous obtenons

$$\tau = y_i(1) - y_i(0), \qquad \textbf{[2.79]}$$

où $y_i(1) = y_i(0) + \tau$. En appliquant ce résultat à (2.78), on obtient :

$$y_i = y_i(0) + \tau x_i.$$

Écrivons maintenant $y_i(0) = \alpha_0 + u_i(0)$ où, par définition, $\alpha_0 = E[y_i(0)]$ et $E[u_i(0)] = 0$. En insérant ces égalités, on obtient :

$$y_i = \alpha_0 + \tau x_i + u_i(0). \qquad \textbf{[2.80]}$$

Si nous définissons $\beta_0 = \alpha_0$, $\beta_1 = \tau$, et $u_i = u_i(0)$, alors l'équation (2.80) correspond à l'équation (2.48) :

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

où $\beta_1 = \tau$ est l'effet du traitement (ou l'effet causal).

Dans cette régression simple, l'estimateur de β_1, qui correspond à l'estimateur de la différence entre deux moyennes, est un estimateur sans biais de l'effet du traitement, τ, si x_i est indépendant de $u_i(0)$. Dans ce cas, l'hypothèse RLS.4 est respectée et nous avons

$$E[u_i(0)|x_i] = 0.$$

L'hypothèse RLS.1 de linéarité du modèle dans ses paramètres a également été respectée dans la dérivation de l'équation (2.80). Comme c'est toujours le cas, nous supposons que l'échantillonnage est aléatoire (RLS.2). Nous avons également de la variabilité au sein de l'échantillon (RLS.3) puisque l'échantillon contient par définition des unités traitées et des unités de contrôle, une condition requise pour estimer l'effet du traitement. Comment pourrions-nous effectivement estimer l'effet du traitement si toutes les unités échantillonnées étaient des unités de contrôle ?

La **répartition aléatoire** des unités d'observation permet de garantir le respect de l'hypothèse selon laquelle x_i est indépendant de $u_i(0)$, hypothèse identique à celle selon laquelle x_i est indépendant de $y_i(0)$. Cela implique que les unités d'observation sont affectées aux groupes de traitement et de contrôle de manière aléatoire, en veillant à ne pas les filtrer en fonction de leur caractéristique individuelle. On parle aussi de **randomisation**. Par exemple, si l'on vise à évaluer un programme de formation professionnelle, la répartition aléatoire des travailleurs est respectée lorsque l'on tire à pile ou face l'affectation de chaque travailleur au groupe de contrôle ou au groupe de traitement. (Pour autant que le tirage est aléatoire, la probabilité de tirer pile ou face ne doit pas nécessairement être égale à 0,5). Si les unités ne se conforment pas à ce tirage aléatoire, alors la randomisation est compromise.

Le principe de répartition aléatoire est au cœur des **expériences aléatoires contrôlées (EAC)**, qui sont considérées depuis longtemps comme la méthode la plus appropriée pour estimer l'effet causal des interventions médicales. Les EAC, qui sont également appelées **essais contrôlés randomisés**, permettent d'obtenir des données expérimentales telles que celles décrites au chapitre 1. Au cours de ces dernières années, les EAC ont gagné en popularité en économie, comme dans les domaines de l'économie du développement et de l'économie comportementale. La mise en œuvre d'une EAC peut néanmoins être très coûteuse et, dans de nombreux cas, la randomisation des sujets dans les groupes de contrôle et de traitement pose des problèmes éthiques. (Par exemple, si une EAC permet de démontrer que l'accès gratuit accordé aux familles à faibles revenus améliore la santé de leurs enfants en moyenne, alors les enfants de familles à faibles revenus, inclus dans le groupe de contrôle, sont pénalisés).

Même s'il n'est pas toujours possible de recourir à une EAC, il est très instructif de réfléchir à l'expérience que l'on mènerait *si la répartition aléatoire était possible*. En règle générale, cet exercice de réflexion permet de s'assurer que l'on se pose les bonnes questions avant de recueillir des données qui sont alors non expérimentales. Par exemple, si nous souhaitions étudier l'effet de l'accès à l'Internet sur les performances des élèves dans les zones rurales, il nous serait très probablement impossible sur le plan éthique ou matériel d'octroyer de manière aléatoire l'accès à l'Internet à certains élèves et pas à d'autres. Néanmoins, réfléchir

à la manière dont une telle répartition aléatoire est mise en œuvre, affine notre compréhension des résultats contrefactuels et de l'effet du traitement.

Dans l'analyse du principe de répartition aléatoire que nous avons menée jusqu'à présent, nous avons considéré un effet du traitement constant et nous avons montré que, dans un tel cas de figure, l'estimateur de la différence entre moyennes, $\bar{y}(1) - \bar{y}(0)$, est un estimateur sans biais de l'effet de traitement, τ. Nous pouvons néanmoins relâcher très facilement l'hypothèse d'un effet du traitement constant. Dans ce cas, l'effet du traitement pour chaque unité i peut être écrit comme suit

$$te_i = y_i(1) - y_i(0) = \tau_{emt} + [u_i(1) - u_i(0)], \qquad \textbf{[2.81]}$$

où $y_i(1) = \alpha_1 + u_i(1)$ et $\tau_{emt} = \alpha_1 - \alpha_0$. Il est utile de se rappeler que τ_{emt} correspond à la moyenne de l'effet du traitement dans toute la population et de considérer $u_i(1) - u_i(0)$ comme l'écart par rapport à cette moyenne pour l'unité i spécifiquement. En introduisant (2.81) dans (2.78), on obtient

$$y_i = \alpha_0 + \tau_{emt} x_i + u_i(0) + [u_i(1) - u_i(0)]x_i \equiv \alpha_0 + \tau_{emt} x_i + u_i, \qquad \textbf{[2.82]}$$

où le terme d'erreur est désormais

$$u_i = u_i(0) + [u_i(1) - u_i(0)]x_i.$$

L'hypothèse de répartition aléatoire requiert désormais que x_i est indépendant de $u_i(0)$ *et* de $u_i(1)$. Cela signifie que la répartition des unités d'observations ne doit pas dépendre de leurs caractéristiques non observées, quel que soit le groupe dans lequel elles sont affectées. Dans un tel cas de figure, alors que u_i dépend de x_i, l'espérance conditionnelle de l'erreur reste nulle car

$$E(x_i) = E[u_i(0)|x_i] + E[u_i(1)|x_i - u_i(0)|x_i]x_i = 0 + 0 \cdot x_i = 0.$$

Comme l'hypothèse RLS.4 est respectée, nous pouvons conclure à nouveau que l'estimateur des MCO est sans biais à la fois pour α_0 et τ_{emt}, sachant que τ_{emt} est l'estimateur de la différence entre moyennes, soit $\tau_{emt} = \beta_1 = \bar{y}(1) - \bar{y}(0)$. [Lorsque l'erreur u_i n'est pas indépendante de x_i, Var($u_i|x_i$) diffère en fonction de x_i car les variances entre résultats potentiels diffèrent également, comme le montre le problème 2.17. N'oubliez pas néanmoins que l'hypothèse RLS.5 d'homoscédasticité n'est pas requise pour démontrer que les estimateurs des MCO sont sans biais.

Le fait que l'estimateur des MCO dans la régression linéaire simple produise un estimateur sans biais lorsque le traitement est appliqué « au hasard » entre unités d'observation est un résultat très puissant. Gardez à l'esprit que ce résultat dépend fortement de l'hypothèse de répartition aléatoire. À partir du chapitre 3, nous verrons comment l'analyse de régression multiple peut être utilisée lorsque ce principe de randomisation pure ne peut pas être respecté.

EXEMPLE 2.14
Évaluation d'un programme de formation professionnelle

La base de données JTRAIN2 contient des données expérimentales sur un programme de formation professionnelle qui s'est déroulé dans les années 1970. La répartition aléatoire entre groupes de traitement et de contrôle avait concerné des hommes dont l'insertion sur le marché du travail était problématique. Ces données ont été fréquemment utilisées dans la littérature par la suite pour comparer les estimations de l'effet de programmes de formation dans des études qui ne reposaient pas sur des données expérimentales. L'indicateur de répartition pour la formation professionnelle est *train*. La variable dépendante est *re78*, qui correspond au revenu annuel (réel) observé en 1978 et mesuré en milliers de dollars. Sur les 445 hommes de l'échantillon, 185 ont été inclus dans le groupe de traitement et ont donc participé au programme avant 1978. Les 260 autres personnes constituent le groupe de contrôle.

La régression simple donne

$$\widehat{re78} = 4.55 + 1.79 train = 445, R^2 = 0.018$$

L'estimation du coefficient de la pente est égale à 1,79 et correspond à l'estimation de la différence de revenu *en moyenne* entre le groupe traité et le groupe témoin. Cela signifie que les hommes qui ont participé au programme ont gagné en moyenne 1 790 dollars de plus que les hommes qui n'y ont pas participé. Il s'agit d'un effet économiquement important car non seulement les dollars sont ceux de 1978 mais le revenu moyen des hommes qui n'ont pas participé à la formation était de 4 550 dollars. En pourcentage, le gain moyen en revenu s'élève donc à 39,3 % environ. (Nous aurions besoin de connaître le coût par individu du programme de formation pour faire une analyse « coût-bénéfice »).

Rappelez-vous que le problème fondamental dans cet exercice d'évaluation est que nous n'observons aucune des unités d'observation dans les deux états du monde. Nous n'observons qu'un seul des deux résultats en termes de gains pour chaque personne. Néanmoins, la répartition aléatoire de ces personnes entre le groupe de traitement et le groupe de contrôle nous permet d'obtenir un estimateur sans biais de l'effet moyen du traitement.

Ensuite, remarquez la valeur du *R*-carré. L'indicateur de participation à la formation explique moins de 2 % de la variation des revenus entre individus dans l'échantillon. En réalité, ce n'est pas surprenant : les revenus sur le marché du travail sont influencés par de nombreux autres facteurs, dont l'éducation, l'expérience, l'intelligence, l'âge, la motivation, etc. Cela montre à quel point le fait de se concentrer sur le *R*-carré est souvent improductif et parfois même nuisible. Les étudiants qui débutent pensent parfois qu'un faible *R*-carré indique un « biais » dans les estimateurs des MCO. Ce n'est pas le cas. Cela signifie simplement que la variance des variables non observables, Var(u), est importante par rapport à Var(y). Dans cet exemple, nous savons que les hypothèses SLR.1 à SLR.4 tiennent en raison de la randomisation. Aucune de ces hypothèses ne mentionne le niveau que devrait avoir le *R*-carré. L'absence de biais n'en dépend pas.

Bien que l'effet du traitement, estimé à 1 790 dollars, soit important sur le plan économique, notez que nous ne savons pas si cette estimation est statistiquement significative. Nous reviendrons sur ce sujet au chapitre 4.

Avant de conclure ce chapitre, il est important de dissiper toute confusion possible concernant l'adjectif « aléatoire » que nous avons utilisé à deux reprises dans cette sous-section. Il y a tout d'abord la notion d'échantillonnage aléatoire qui est introduite dans l'hypothèse RLS.2. (Elle est également abordée dans l'annexe C). Un échantillonnage aléatoire implique que les données que nous obtenons proviennent de tirages indépendants et identiquement distribués à partir de la population représentée par les variables aléatoires (x, y). L'échantillonnage aléatoire est un concept séparé et distinct de celui de la répartition aléatoire dont dépend le respect de l'hypothèse RLS.4. La répartition aléatoire signifie que la valeur de x_i est déterminée indépendamment des résultats contrefactuels $[y_i(0), y_i(1)]$. Dans l'exemple 2.14, nous avons à la fois un échantillon aléatoire issu de la population concernée et une répartition aléatoire qui nous permet de constituer des groupes de traitement et de contrôle. Dans d'autres cas, l'échantillonnage est aléatoire alors que la répartition ne l'est pas. Par exemple, il est relativement facile de tirer un échantillon aléatoire au sein d'une population d'étudiants qui ont passé le test SAT et suivi un cours de préparation pour certains d'entre eux. Cela ne signifie pas que la participation à ce cours de préparation est indépendante des résultats contrefactuels obtenus au test SAT. Pour garantir l'indépendance entre la participation au cours de préparation et les résultats potentiels obtenus à ce test, il est nécessaire de répartir les étudiants de manière aléatoire entre un groupe qui peut suivre le cours de préparation et un groupe qui ne le peut pas. (Il faut ensuite s'assurer que les étudiants respectent cette répartition). Si nous ne pouvons pas compter sur cette répartition aléatoire, les données ne sont que rétrospectives, c'est-à-dire que nous enregistrons la participation des étudiants à ce cours de préparation *après* qu'il a eu lieu. Il est alors improbable que l'hypothèse d'indépendance soit respectée. Dans les deux cas, cela n'a rien à voir avec le fait que nous ayons obtenu

un échantillon aléatoire d'étudiants tiré de la population. Cela confirme bien que les hypothèses RLS.2 et RLS.4 sont fondamentalement différentes.

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.

Dans la section 2.7, nous avons abordé le cas important où x est une variable binaire qui capture la présence ou l'absence d'une intervention. Nous avons montré que l'estimateur des MCO de la « pente » est égale à $\hat{\beta}_1 = \bar{y}(1) - \bar{y}(0)$, soit la différence entre les moyennes de y_i au lorsque $x_i = 1$ et lorsque $x_i = 0$. Nous avons également discuté des conditions sous lesquelles l'estimateur $\hat{\beta}_1$ est un estimateur sans biais de l'effet moyen du traitement, en particulier de l'exigence de répartition aléatoire des unités d'observations entre groupes de contrôle et de traitement. Dans le chapitre 3 et au-delà, nous étudierons le cas où l'intervention ou le traitement n'est pas randomisé, mais dépend de facteurs observés et même non observés.

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 :

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

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,

$$\mathrm{E}(u|x) = 0.$$

Hypothèse RLS.5 (Homoscédasticité)

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

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

MOTS-CLÉS

Essais contrôlés randomisés p. 79
Fonction de régression de l'échantillon (FRE) p. 47
Fonction de régression de la population (FRP) p. 41
Groupe de contrôle p. 78
Groupe de traitement p. 78
Groupe témoin p. 78
Hétéroscédasticité p. 69
Homoscédasticité p. 68
Hypothèse d'espérance conditionnelle de l'erreur égale à zéro p. 41
Hypothèse de l'espérance indépendante de *x* p. 41
Hypothèses de Gauss-Markov p. 83
Modèle à élasticité constante p. 60
Modèle de régression linéaire simple p. 38
Moindres carrés ordinaires (MCO) p. 46
R carré p. 55
Randomisation p. 79
Régresseur p. 38
Régression de *y* sur une constante p. 75
Régression passant par l'origine p. 74
Répartition aléatoire p. 79
Résidu p. 46
Semi-élasticité p. 61
Somme des carrés des résidus (SCR) p. 46, 54
Somme des carrés expliqués (SCE) p. 54
Somme des carrés totaux (SCT) p. 54
Terme d'erreur (perturbation) p. 39
Valeur ajustée p. 46
Variable binaire p. 76
Variable de contrôle p. 38
Variable dépendante p. 38
Variance de l'erreur p. 69
Variable de réponse p. 38
Variable dichotomique p. 76
Variable dummy p. 76
Variable endogène p. 38
Variable exogène p. 38
Variable explicative p. 38
Variable expliquée p. 38
Variable indépendante p. 38
Variable prédictive p. 38
Variable prédite p. 38
Variable résultat p. 38
Variable stimulus p. 38

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 :

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

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

La référence en économétrie !

Jeffrey M. Wooldridge est professeur émérite d'économie à l'Université d'état du Michigan (MSU) où il enseigne depuis 1991. 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.

Traducteurs :

Michel Beine est professeur à l'Université du Luxembourg.

Sophie Béreau est professeur à l'Université de Lorraine et à l'Université de Namur.

Maëlys de la Rupelle est maître de conférences à l'Université de Cergy-Pontoise.

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.

Ce livre d'introduction réussit l'exploit de **simplifier la présentation de l'économétrie**. 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.

Ce manuel de référence :

- permet de comprendre et d'interpréter les hypothèses d'un modèle à la lumière de nombreuses applications empiriques et distingue clairement le type de données analysées.
- couvre les données en coupe transversale et les séries chronologiques.
- aborde les données de panel.
- offre également une introduction aux modèles à variable dépendante limitée, essentiels en économie appliquée et en gestion.

Chaque chapitre propose le cours, un résumé, des mots clés, ainsi qu'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 exemples empiriques développés dans les chapitres de l'ouvrage et maîtriser toutes les étapes de la modélisation économétrique.

Conception graphique : Primo&Primo®

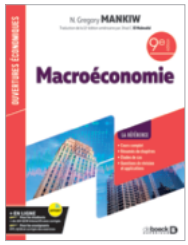
RESSOURCES NUMÉRIQUES

Étudiants : + de **450** exercices à télécharger sur www/deboecksuperieur.com/site/329775

Enseignants : identifiez-vous à la même adresse et accédez à des ressources en anglais (PPT, corrigés des exercices).

- Résumés de chapitres
- Études de cas et exemples
- Questions de réflexion avec corrigés
- Mots-clés et glossaire

Dans la même collection



ISSN : 2030-501X
ISBN : 978-2-8073-2977-5
9 782807 329775
64,90 €

deboeck SUPÉRIEUR
www.deboecksuperieur.com