

Économétrie

Régis Bourbonnais

10^e édition

Cours complet

Nombreux exemples

**Applications corrigées sous Excel,
Eviews, Gretl ou Stata**

DUNOD

Le pictogramme qui figure ci-contre mérite une explication. Son objet est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, particulièrement dans le domaine de l'édition technique et universitaire, le développement massif du photocopillage.

Le Code de la propriété intellectuelle du 1^{er} juillet 1992 interdit en effet expressément la photocopie à usage collectif sans autorisation des ayants droit. Or, cette pratique

s'est généralisée dans les établissements

d'enseignement supérieur, provoquant une baisse brutale des achats de livres et de revues, au point que la possibilité même pour

les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée.

Nous rappelons donc que toute reproduction, partielle ou totale, de la présente publication est interdite sans autorisation de l'auteur, de son éditeur ou du

Centre français d'exploitation du droit de copie (CFC, 20, rue des Grands-Augustins, 75006 Paris).



© Dunod, 2018
11, rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-077345-9

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, 2^o et 3^o a), d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (art. L. 122-4).

Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

Table des matières

Avant-propos	XI
1 Qu'est-ce que l'économétrie ?	1
Section 1 La notion de modèle	2
1. Définition	2
2. La construction des modèles en économétrie	2
Section 2 Le rôle de l'économétrie	5
1. L'économétrie comme validation de la théorie	5
2. L'économétrie comme outil d'investigation	5
Section 3 La théorie de la corrélation	6
1. Présentation générale	6
2. Mesure et limite du coefficient de corrélation	8
2 Le modèle de régression simple	13
Section 1 Présentation du modèle	14
1. Exemple introductif	14
2. Rôle du terme aléatoire	15
3. Conséquences du terme aléatoire	17
Section 2 Estimation des paramètres	18
1. Modèle et hypothèses	18
2. Formulation des estimateurs	18
3. Les différentes écritures du modèle : erreur et résidu	22
4. Propriétés des estimateurs	22

Section 3	Conséquences des hypothèses : construction des tests	25
	1. Hypothèse de normalité des erreurs	25
	2. Conséquences de l'hypothèse de normalité des erreurs	25
	3. Test bilatéral, test unilatéral et probabilité critique d'un test	29
Section 4	Équation et tableau d'analyse de la variance	35
	1. Équation d'analyse de la variance	35
	2. Tableau d'analyse de la variance	36
Section 5	La prévision dans le modèle de régression simple	41

3 Le modèle de régression multiple **51**

Section 1	Le modèle linéaire général	52
	1. Présentation	52
	2. Forme matricielle	52
Section 2	Estimation et propriétés des estimateurs	53
	1. Estimation des coefficients de régression	53
	2. Hypothèses et propriétés des estimateurs	56
	3. Équation d'analyse de la variance et qualité d'un ajustement	58
Section 3	Les tests statistiques	63
	1. Le rôle des hypothèses	63
	2. Construction des tests	64
	3. Tests sur les résidus : valeur anormale, effet de levier et point d'influence	66
Section 4	L'analyse de la variance	72
	1. Construction du tableau d'analyse de la variance et test de signification globale d'une régression	72
	2. Autres tests à partir du tableau d'analyse de la variance	74
	3. Généralisation des tests par analyse de la variance	80
Section 5	L'utilisation de variables indicatrices	81
	1. Constitution et finalités des variables indicatrices	81
	2. Exemples d'utilisation	82
Section 6	La prévision à l'aide du modèle linéaire général et la régression récursive	88
	1. Prédiction conditionnelle	88
	2. Fiabilité de la prévision et intervalle de prévision	89
	3. Les tests de stabilité par la régression récursive	92
	4. Le test de spécification de Ramsey	93

Section 7	Exercices récapitulatifs	97
	<i>Annexe</i>	111
	1. Interprétation géométrique de la méthode des moindres carrés	111
	2. Résolution de l'exercice 1 par des logiciels informatiques de régression multiple	112
	3. Estimation de la variance de l'erreur	114
4	Multicolinéarité et sélection du modèle optimal	115
Section 1	Corrélation partielle	116
	1. Exemple introductif	116
	2. Généralisation de la notion de corrélation partielle	116
Section 2	Relation entre coefficients de corrélation simple, partielle et multiple	121
Section 3	Multicolinéarité : conséquences et détection	122
	1. Conséquences de la multicolinéarité	123
	2. Tests de détection d'une multicolinéarité	124
	3. Comment remédier à la multicolinéarité ?	128
Section 4	Sélection du modèle optimal	128
5	Problèmes particuliers : la violation des hypothèses	135
Section 1	L'autocorrélation des erreurs	136
	1. Présentation du problème	136
	2. L'estimateur des Moindres Carrés Généralisés (MCG)	137
	3. Les causes et la détection de l'autocorrélation des erreurs	138
	4. Les procédures d'estimation en cas d'autocorrélation des erreurs	145
Section 2	L'hétéroscédasticité	153
	1. Présentation du problème	153
	2. Correction de l'hétéroscédasticité	155
	3. Tests de détection de l'hétéroscédasticité	159
	4. Autre test d'hétéroscédasticité : le test ARCH	165
Section 3	Modèles à erreurs sur les variables	166
	1. Conséquences lorsque les variables sont entachées d'erreurs	166
	2. La méthode des variables instrumentales	167
	3. Le test d'exogénéité d'Hausman	168
	4. La méthode des moments généralisée	169

6	Les modèles non linéaires	179
Section 1	Les différents types de modèles non linéaires	180
	1. Les fonctions de type exponentiel	180
	2. Les modèles de diffusion	183
Section 2	Méthodes d'estimation des modèles non linéaires	184
	1. Initiation aux méthodes d'estimation non linéaires	184
	2. Exemples d'application	186
7	Les modèles à décalages temporels	191
Section 1	Les modèles linéaires autorégressifs	192
	1. Formulation générale	192
	2. Test d'autocorrélation et méthodes d'estimation	193
Section 2	Les modèles à retards échelonnés	198
	1. Formulation générale	198
	2. Détermination du nombre de retards	199
	3. Distribution finie des retards	203
	4. Distribution infinie des retards	208
Section 3	Deux exemples de modèles dynamiques	214
	1. Le modèle d'ajustement partiel	214
	2. Le modèle d'anticipations adaptatives	215
8	Introduction aux modèles à équations simultanées	235
Section 1	Équations structurelles et équations réduites	236
	1. Exemple introductif	236
	2. Le modèle général	238
Section 2	Le problème de l'identification	239
	1. Restrictions sur les coefficients	239
	2. Conditions d'identification	239
Section 3	Les méthodes d'estimation	241
	1. Les moindres carrés indirects	241
	2. Les doubles moindres carrés	241
	3. Autres méthodes d'estimation	242
	<i>Annexe</i>	255
	<i>Identification : les conditions de rang</i>	255

9	Éléments d'analyse des séries temporelles	257
Section 1	Stationnarité	258
	1. Définition et propriétés	258
	2. Fonctions d'autocorrélation simple et partielle	258
	3. Tests de « bruit blanc » et de stationnarité	260
Section 2	La non-stationnarité et les tests de racine unitaire	263
	1. La non-stationnarité : les processus TS et DS	263
	2. Les tests de racine unitaire et la stratégie séquentielle de test	267
Section 3	Les modèles ARIMA	276
	1. Typologie des modèles AR, MA et ARMA	276
	2. L'extension aux processus ARIMA et SARIMA	279
Section 4	La méthode de Box et Jenkins	280
	1. Recherche de la représentation adéquate : l'identification	280
	2. Estimation des paramètres	281
	3. Tests d'adéquation du modèle et prévision	282
10	La modélisation VAR	297
Section 1	Représentation d'un modèle VAR	298
	1. Exemple introductif	298
	2. La représentation générale	299
	3. La représentation ARMAX	301
Section 2	Estimation des paramètres	301
	1. Méthode d'estimation	301
	2. Détermination du nombre de retards	302
	3. Prévision	302
Section 3	Dynamique d'un modèle VAR	308
	1. Représentation VMA d'un processus VAR	308
	2. Analyse et orthogonalisation des « chocs »	308
	3. Décomposition de la variance	312
	4. Choix de l'ordre de décomposition	312
Section 4	La causalité	316
	1. Causalité au sens de Granger	316
	2. Causalité au sens de Sims	317

11	La cointégration et le modèle à correction d'erreur	321
Section 1	Exemples introductifs	322
	1. Premier exemple	322
	2. Deuxième exemple	323
Section 2	Le concept de cointégration	324
	1. Propriétés de l'ordre d'intégration d'une série	324
	2. Conditions de cointégration	326
	3. Le modèle à correction d'erreur (ECM)	326
Section 3	Cointégration entre deux variables	327
	1. Test de cointégration entre deux variables	328
	2. Estimation du modèle à correction d'erreur	328
Section 4	Généralisation à k variables	331
	1. La cointégration entre k variables	332
	2. Estimation du modèle à correction d'erreur	333
	3. Le modèle à correction d'erreur vectoriel	333
	4. Tests de relation de cointégration	335
	5. Test d'exogénéité faible	338
	6. Synthèse de la procédure d'estimation	339
12	Introduction à l'économétrie des variables qualitatives	345
Section 1	Les problèmes et les conséquences de la spécification binaire	346
Section 2	Les modèles de choix binaires	348
	1. Le modèle linéaire sur variable latente	348
	2. Les modèles Probit et Logit	349
	3. Interprétation des résultats et tests statistiques	351
Section 3	Les modèles à choix multiples	356
	1. Les modèles Probit et Logit ordonnés	357
	2. Le modèle de choix multiples non ordonné : le Logit multinomial	361
Section 4	Les modèles à variable dépendante limitée : le modèle Tobit	363
	1. Le modèle Tobit simple : modèle de régression tronqué ou censuré	364
	2. Estimation et interprétation des résultats	366

13	Introduction à l'économétrie des données de panel	371
Section 1	Présentation des modèles à données de panel	372
	1. Spécificités des données de panel	372
	2. La méthode SUR	373
	3. Le modèle linéaire simple	374
Section 2	Les tests d'homogénéité	375
	1. Procédure séquentielle de tests	375
	2. Construction des tests	376
Section 3	Spécifications et estimations des modèles à effets individuels	381
	1. Le modèle à effets fixes individuels	381
	2. Le modèle à effets aléatoires	383
	3. Effets fixes ou effets aléatoires ? Le test d'Hausman	384
	Liste des exercices	388
	Tables statistiques	391
	Bibliographie	399
	Index	402

Avant-propos

Cette dixième édition, gage que ce livre répond à un besoin constant des étudiants, marque la volonté d'une mise à jour permanente de ce manuel tant sur le plan des concepts de l'économétrie moderne que des applications, tout en lui conservant son aspect très pédagogique. Dans cette nouvelle édition nous avons intégré de manière systématique les logiciels Gretl et Stata dans la correction des exercices à l'aide des fichiers « script » de commandes.

Ce livre couvre tous les champs de l'économétrie : régression simple et multiple, violation des hypothèses (hétéroscédasticité, autocorrélation des erreurs, variables explicatives aléatoires), modèle à décalage, analyse des séries temporelles, tests de racine unitaire, équations multiples, VAR, cointégration, VECM, économétrie des variables qualitatives et des données de panel...

Sur l'ensemble de ces thèmes, ce livre vous propose un cours, des exercices corrigés, et une présentation des logiciels d'économétrie les plus répandus. Souhaitons qu'il corresponde à votre attente.


En effet, nous avons voulu, par une alternance systématique de cours et d'exercices, répondre à un besoin pédagogique qui est de mettre rapidement en pratique les connaissances théoriques et ainsi, d'utiliser de manière opérationnelle les acquis du cours ; les exercices sont repérés grâce à un bandeau grisé. De surcroît, le recours à des logiciels¹, lors de la résolution des exercices, permet une découverte de ces outils et donne une dimension pratique que recherchent l'étudiant et le praticien.

1. Quatre logiciels sont utilisés : EXCEL (copyright Microsoft), Eviews (copyright Quantitative Micro Software), Stata (copyright StataCorp.) et Gretl. Nous recommandons particulièrement le logiciel Gretl (<http://gretl.sourceforge.net/>) qui est un logiciel d'économétrie gratuit, complet et très facile d'apprentissage.

Afin que le lecteur puisse lui-même refaire les exercices, les données utilisées (sous format Excel, Eviews, Gretl et Stata) ainsi que les programmes de traitement de Eviews (extension .prg) ou de Gretl (extension .INP) sont disponibles par téléchargement sur le serveur web.

Les corrigés des exercices et les données sous format Stata ont été réalisés par Dalila Chenaf-Nicet, maître de conférences en économie à l'Université de Bordeaux, et sont disponibles également par téléchargement sur le site web :

<http://regisbourbonnais.dauphine.fr>

Pour chaque exercice faisant appel à un fichier de données, le nom du fichier est cité en tête de l'exercice et repéré par l'icône suivante : .

Nous avons voulu faire de ce manuel un livre d'apprentissage facilement accessible ; c'est pourquoi les démonstrations les plus complexes font l'objet de renvois à une bibliographie plus spécialisée. Cependant, il convient de préciser que l'économétrie fait appel à des notions d'algèbre linéaire et d'induction statistique qu'il est souhaitable de connaître.

Dans le terme « économétrie » figure la racine du mot « économie » car son utilisation est surtout destinée à des fins de traitement de données économiques ; cependant, d'autres domaines tels que la finance, la recherche agronomique, la médecine, etc., font maintenant le plus souvent appel à ces techniques.

Ce livre s'adresse en premier lieu aux étudiants (sciences économiques, gestion, écoles de commerce et d'ingénieurs, etc.) dont la formation requiert une connaissance de l'économétrie. Gageons qu'il sera un support de cours indispensable et un allié précieux pour préparer les séances de travaux dirigés.

N'oublions pas cependant le praticien de l'économétrie (économiste d'entreprise, chercheur, etc.) qui, confronté à des problèmes d'estimation statistique, trouvera dans ce livre les réponses pratiques aux différentes questions qu'il peut se poser.

Enfin, j'exprime toute ma gratitude à toutes les personnes – collègues et étudiants – qui ont eu la gentillesse de me faire des commentaires et dont les conseils et suggestions contribuent à la qualité pédagogique de ce livre. Je reste, bien entendu, le seul responsable des erreurs qui subsisteraient¹.

1. Les lecteurs souhaitant faire des commentaires ou des remarques peuvent me contacter : Régis Bourbonnais, université de Paris-Dauphine, place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, E-mail : regis.bourbonnais@dauphine.fr

Chapitre

1

Qu'est-ce que l'économétrie ?

SOMMAIRE

- SECTION 1** La notion de modèle
- SECTION 2** Le rôle de l'économétrie
- SECTION 3** La théorie de la corrélation

Ce premier chapitre est consacré à la présentation de l'économétrie et à sa liaison avec la théorie économique.

Section 1 LA NOTION DE MODÈLE

1 Définition

Il est délicat de fournir une définition unique de la notion de modèle¹. Dans le cadre de l'économétrie, nous pouvons considérer qu'un modèle consiste en une *présentation formalisée d'un phénomène* sous forme d'équations dont les variables sont des grandeurs économiques. L'objectif du modèle est de représenter les traits les plus marquants d'une réalité qu'il cherche à styliser. Le modèle est donc l'outil que le modélisateur utilise lorsqu'il cherche à comprendre et à expliquer des phénomènes. Pour ce faire, il émet des hypothèses et explicite des relations.

- Pourquoi des modèles ?
- Nombreux sont ceux – sociologues, économistes ou physiciens – qui fondent leurs analyses ou leurs jugements sur des raisonnements construits et élaborés. Ces constructions réfèrent implicitement à des modèles ; alors pourquoi ne pas expliciter clairement les hypothèses et les relations au sein d'un modèle ?

Le modèle est donc une présentation schématique et partielle d'une réalité naturellement plus complexe. Toute la difficulté de la modélisation consiste à ne retenir que la ou les représentations intéressantes pour le problème que le modélisateur cherche à expliciter. Ce choix dépend de la nature du problème, du type de décision ou de l'étude à effectuer. La même réalité peut ainsi être formalisée de diverses manières en fonction des objectifs.

2 La construction des modèles en économétrie

Dans les sciences sociales, et particulièrement en économie, les phénomènes étudiés concernent le plus souvent des comportements afin de mieux comprendre la nature et le fonctionnement des systèmes économiques. L'objectif du modélisateur est, dans le cadre de l'économétrie et au travers d'une mesure statistique, de permettre aux agents

1. La notion de modèle est relative au point de vue auquel nous nous plaçons : la physique, l'épistémologie...

économiques (ménages, entreprises, État...) d'intervenir de manière plus efficace. La construction d'un modèle comporte un certain nombre d'étapes qui sont toutes importantes. En effet, en cas de faiblesse d'un des « maillons », le modèle peut se trouver invalidé pour cause d'hypothèses manquantes, de données non représentatives ou observées avec des erreurs, etc. Examinons les différentes étapes à suivre lors de la construction d'un modèle, ceci à partir de l'exemple du modèle keynésien simplifié.

2.1 Référence à une théorie

Une théorie s'exprime au travers d'hypothèses auxquelles le modèle fait référence. Dans la théorie keynésienne, quatre propositions sont fondamentales :

1. la consommation et le revenu sont liés ;
2. le niveau d'investissement privé et le taux d'intérêt sont également liés ;
3. il existe un investissement autonome public ;
4. enfin, le produit national est égal à la consommation plus l'investissement privé et public.

2.2 Formalisation des relations et choix de la forme des fonctions

À partir des propositions précédentes, nous pouvons construire des relations :

1. la consommation est fonction du revenu : $C = f(Y)$ avec $f' > 0$;
2. l'investissement privé dépend du taux d'intérêt : $I = g(r)$ avec $g' < 0$;
3. il existe un investissement autonome public : \bar{I} ;
4. enfin, le produit national (ou le revenu national) est égal à la consommation plus l'investissement : $Y \equiv C + I + \bar{I}$.

À ce stade, nous n'avons postulé aucune forme particulière en ce qui concerne les fonctions f et g . Ainsi, bien que des considérations d'ordre théorique nous renseignent sur le signe des dérivées, il existe une multitude de fonctions de formes très différentes et ayant des signes de dérivées identiques, par exemple $C = a_0 + a_1 Y$ et $C = a_0 Y^{a_1}$. Cependant, ces deux relations ne reflètent pas le même comportement ; une augmentation du revenu provoque un accroissement proportionnel pour la première relation, alors que, dans la seconde, l'effet s'estompe avec l'augmentation du revenu (si $0 < a_1 < 1$). Nous appelons « forme fonctionnelle » ce choix (arbitraire ou fondé) de spécification précise du modèle. Dans notre exemple, le modèle explicité s'écrit :

$$C = a_0 + a_1 Y \quad \text{avec } a_0 > 0 \text{ et } 0 < a_1 < 1$$

$$a_1 = \text{propension marginale à consommer}$$

$$\text{et } a_0 = \text{consommation incompressible ;}$$

$$I = b_0 + b_1 r \quad \text{avec } b_0 > 0 \text{ et } b_1 < 0 ;$$

$$Y \equiv C + I + \bar{I}$$

Les deux premières équations reflètent des relations de comportements alors que la troisième est une identité (aucun paramètre n'est à estimer).

2.3 Sélection et mesure des variables

Le modèle étant spécifié, il convient de collecter les variables représentatives des phénomènes économiques. Ce choix n'est pas neutre et peut conduire à des résultats différents, les questions qu'il convient de se poser sont par exemple :

- *Faut-il raisonner en euros constants ou en euros courants ?*
- *Les données sont-elles brutes ou CVS¹ ?*
- *Quel taux d'intérêt faut-il retenir (taux au jour le jour; taux directeur de la Banque centrale européenne...) ? etc.*

- Nous distinguons plusieurs types de données selon que le modèle est spécifié en :
- *série temporelle* : c'est le cas le plus fréquent en économétrie, il s'agit de variables observées à intervalles de temps réguliers (la consommation annuelle, totale France, exprimée en euros courants sur 20 ans) ;
 - *coupe instantanée* : les données sont observées au même instant et concernent les valeurs prises par la variable pour un groupe d'individus² spécifiques (consommation observée des agriculteurs pour une année donnée) ;
 - *panel* : la variable représente les valeurs prises par un échantillon d'individus à intervalles réguliers (la consommation d'un échantillon de ménages de la région parisienne sur 20 ans) ;
 - *cohorte* : très proches des données de panel, les données de cohorte se distinguent de la précédente par la constance de l'échantillon, les individus sondés sont les mêmes d'une période sur l'autre.

2.4 Décalages temporels

Dans le cadre de modèle spécifié en séries temporelles, les relations entre les variables ne sont pas toujours synchrones mais peuvent être décalées dans le temps. Nous pouvons concevoir que la consommation de l'année t est expliquée par le revenu de l'année $t - 1$ et non celui de l'année t . Pour lever cette ambiguïté, il est d'usage d'écrire le modèle en le spécifiant à l'aide d'un indice de temps : $C_t = a_0 + a_1 Y_{t-1}$. La variable Y_{t-1} est appelée « variable exogène retardée ».

On appelle « variable exogène » une variable dont les valeurs sont prédéterminées, et « variable endogène » une variable dont les valeurs dépendent des variables exogènes.

1. Corrigées des Variations Saisonnières.

2. Le terme d'individu est employé au sens statistique, c'est-à-dire comme un élément d'une population : une personne, une parcelle de terre...

2.5 Validation du modèle

La dernière étape est celle de la validation¹ du modèle :

- *Les relations spécifiées sont-elles valides ?*
- *Peut-on estimer avec suffisamment de précision les coefficients ?*
- *Le modèle est-il vérifié sur la totalité de la période ?*
- *Les coefficients sont-ils stables ? Etc.*

À toutes ces questions, les techniques économétriques s'efforcent d'apporter des réponses.

Section 2 LE RÔLE DE L'ÉCONOMÉTRIE

1 L'économétrie comme validation de la théorie

L'économétrie est un outil à la disposition de l'économiste qui lui permet d'infirmer ou de confirmer les théories qu'il construit. Le théoricien postule des relations ; l'application de méthodes économétriques fournit des estimations sur la valeur des coefficients ainsi que la précision attendue.

Une question se pose alors : pourquoi estimer ces relations, et les tester statistiquement ? Plusieurs raisons incitent à cette démarche : tout d'abord cela force l'individu à établir clairement et à estimer les interrelations sous-jacentes. Ensuite, la confiance aveugle dans l'intuition peut mener à l'ignorance de liaisons importantes ou à leur mauvaise utilisation. De plus, des relations marginales mais néanmoins explicatives, qui ne sont qu'un élément d'un modèle global, doivent être testées et validées afin de les mettre à leur véritable place. Enfin, il est nécessaire de fournir, en même temps que l'estimation des relations, une mesure de la confiance que l'économiste peut avoir en celles-ci, c'est-à-dire la précision que l'on peut en attendre. Là encore, l'utilisation de méthodes purement qualitatives exclut toute mesure quantitative de la fiabilité d'une relation.

2 L'économétrie comme outil d'investigation

L'économétrie n'est pas seulement un système de validation, mais également un outil d'analyse. Nous pouvons citer quelques domaines où l'économétrie apporte une aide à la modélisation, à la réflexion théorique ou à l'action économique par :

1. Validation, c'est-à-dire en conformité avec les données disponibles.

- la mise en évidence de relations entre des variables économiques qui n'étaient pas *a priori* évidentes ou pressenties ;
- l'induction statistique ou l'inférence statistique, qui consiste à inférer, à partir des caractéristiques d'un échantillon, les caractéristiques d'une population. Elle permet de déterminer des intervalles de confiance pour des paramètres du modèle ou de tester si un paramètre est significativement¹ inférieur, supérieur ou simplement différent d'une valeur fixée ;
- la simulation qui mesure l'impact de la modification de la valeur d'une variable sur une autre ($\Delta C_t = a_1 \Delta Y_t$) ;
- la prévision², par l'utilisation de modèles économétriques, qui est utilisée par les pouvoirs publics ou l'entreprise afin d'anticiper et éventuellement de réagir à l'environnement économique.

Dans cet ouvrage, nous nous efforcerons de montrer, à l'aide d'exemples, les différentes facettes de l'utilisation des techniques économétriques dans des contextes et pour des objectifs différents.

Section 3 LA THÉORIE DE LA CORRÉLATION

1 Présentation générale

Lorsque deux phénomènes ont une évolution commune, nous disons qu'ils sont « corrélés ». La corrélation simple mesure le degré de liaison existant entre ces deux phénomènes représentés par des variables. Si nous cherchons une relation entre trois variables ou plus, nous ferons appel alors à la notion de corrélation multiple.

Nous pouvons distinguer la corrélation linéaire, lorsque tous les points du couple de valeurs (x,y) des deux variables semblent alignés sur une droite, de la corrélation non linéaire lorsque le couple de valeurs se trouve sur une même courbe d'allure quelconque.

Deux variables peuvent être :

- en corrélation positive ; on constate alors une augmentation (ou diminution, ou constance) simultanée des valeurs des deux variables ;
- en corrélation négative, lorsque les valeurs de l'une augmentent, les valeurs de l'autre diminuent ;

1. Au sens statistique, c'est-à-dire avec un seuil (risque d'erreur à ne pas dépasser, souvent 5 %).

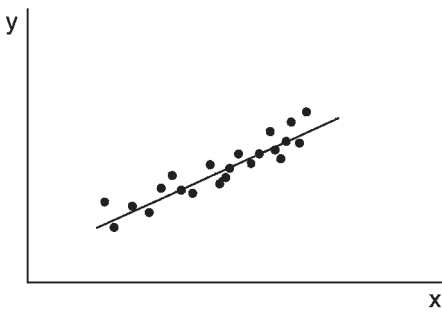
2. Pour découvrir l'utilisation de l'économétrie à des fins de prévision de ventes, voir Bourbonnais R. et Usunier J.-C. (2017).

– non corrélées, il n'y a aucune relation entre les variations des valeurs de l'une des variables et les valeurs de l'autre.

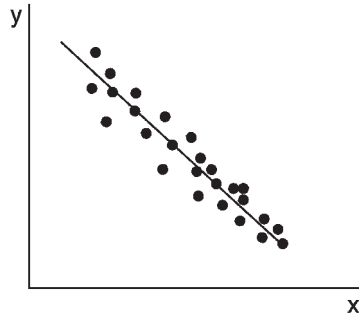
Le tableau 1, en croisant les critères de linéarité et de corrélation, renvoie à une représentation graphique.

Tableau 1 – Linéarité et corrélation

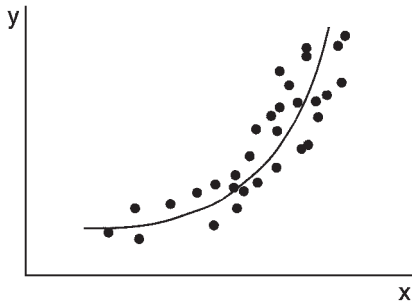
	Corrélation positive	Corrélation négative	Absence de corrélation
Relation linéaire	Graphe 1	Graphe 2	Graphe 5
Relation non linéaire	Graphe 3	Graphe 4	Graphe 5



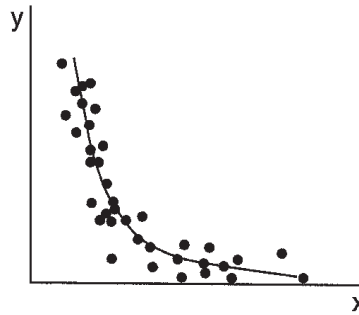
Graphe 1



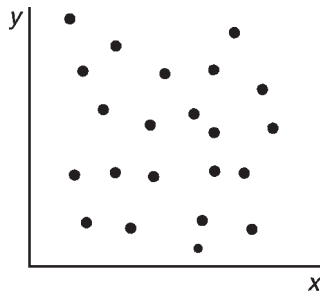
Graphe 2



Graphe 3



Graphe 4



Graphe 5

2 Mesure et limite du coefficient de corrélation

2.1 Le coefficient de corrélation linéaire

La représentation graphique ne donne qu'une « impression » de la corrélation entre deux variables sans donner une idée précise de l'intensité de la liaison, c'est pourquoi nous calculons une statistique appelée *coefficient de corrélation linéaire simple*, noté $r_{x,y}$. Il est égal à :

$$r_{x,y} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad [1]$$

avec :

$\text{Cov}(x,y)$ = covariance entre x et y ;

σ_x et σ_y = écart type de x et écart type de y ;

n = nombre d'observations.

En développant la formule [1], il vient :

$$r_{x,y} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}} \quad [2]$$

- On peut démontrer que, par construction, ce coefficient reste compris entre -1 et 1 :
- proche de 1 , les variables sont corrélées positivement ;
 - proche de -1 , les variables sont corrélées négativement ;
 - proche de 0 , les variables ne sont pas corrélées.

Dans la pratique, ce coefficient est rarement très proche de l'une de ces trois bornes et il est donc difficile de proposer une interprétation fiable à la simple lecture de ce coefficient. Ceci est surtout vrai en économie où les variables sont toutes plus au moins liées entre elles. De plus, il n'est calculé qu'à partir d'un échantillon d'observations et non pas sur l'ensemble des valeurs. On appelle $\rho_{x,y}$ ce coefficient empirique qui est une estimation du coefficient vrai $r_{x,y}$. La théorie des tests statistiques nous permet de lever cette indétermination.

Soit à tester l'hypothèse $H_0 : r_{x,y} = 0$, contre l'hypothèse $H_1 : r_{x,y} \neq 0$.

Sous l'hypothèse H_0 , nous pouvons démontrer que $\frac{\rho_{x,y}}{\sqrt{\frac{(1-\rho_{x,y}^2)}{n-2}}}$ suit une loi de

Student à $n - 2$ degrés de liberté¹. Nous calculons alors une statistique, appelé le t de Student empirique :

$$t^* = \frac{|\rho_{x,y}|}{\sqrt{\frac{(1-\rho_{x,y}^2)}{n-2}}} \quad [3]$$

Si $t^* > t_{n-2}^{\alpha/2}$ valeur lue dans une table de Student² au seuil $\alpha = 0,05$ (5 %) à $n - 2$ degrés de liberté³, nous rejetons l'hypothèse H_0 , le coefficient de corrélation est donc significativement différent de 0 ; dans le cas contraire, l'hypothèse d'un coefficient de corrélation nul est acceptée. La loi de Student étant symétrique, nous calculons la valeur absolue du t empirique et nous procédons au test par comparaison avec la valeur lue directement dans la table.

EXERCICE n° 1

↓ Fichier C1EX1

Calcul d'un coefficient de corrélation

Un agronome s'intéresse à la liaison pouvant exister entre le rendement de maïs x (en quintal) d'une parcelle de terre et la quantité d'engrais y (en kilo). Il relève 10 couples de données consignés dans le tableau 2.

Tableau 2 – Rendement de maïs et quantité d'engrais

Rendement x	16	18	23	24	28	29	26	31	32	34
Engrais y	20	24	28	22	32	28	32	36	41	41

- Tracer le nuage de points et le commenter.
- Calculer le coefficient de corrélation simple et tester sa signification par rapport à 0 pour un seuil $\alpha = 0,05$.

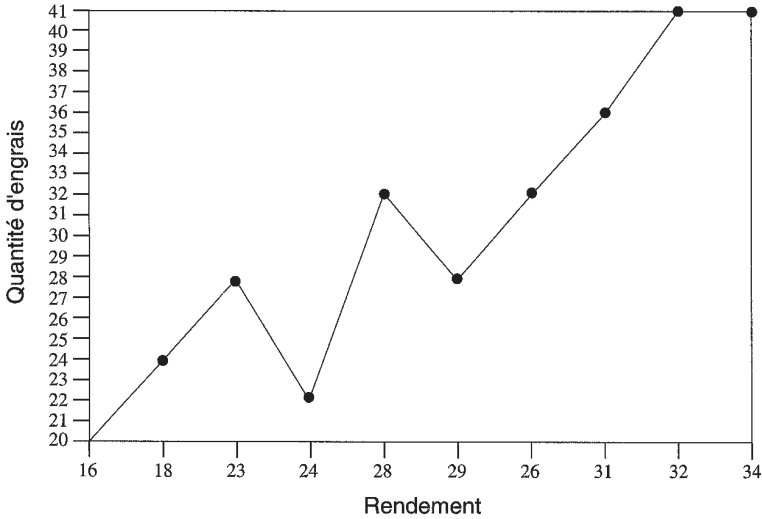
1. La notion de degrés de liberté est explicitée au chapitre 2.

2. Les lois de probabilité sont en fin d'ouvrage.

3. Si le nombre d'observations n est supérieur à 30, on peut approximer la loi de Student par une loi normale, soit $t^{\alpha/2} \approx 1,96$.

Solution

1 ■ Le nuage de points (graphique 6) indique que les couples de valeurs sont approximativement alignés : les deux variables semblent corrélées positivement.



Graphique 6 – Nuage du couple de valeurs : rendement-quantité d’engrais

2 ■ Afin d’appliquer la formule [2], nous dressons le tableau de calcul 3.

Tableau 3 – Calcul d’un coefficient de corrélation

	x	y	x ²	y ²	xy
	16	20	256	400	320
	18	24	324	576	432
	23	28	529	784	644
	24	22	576	484	528
	28	32	784	1 024	896
	29	28	841	784	812
	26	32	676	1 024	832
	31	36	961	1 296	1 116
	32	41	1 024	1 681	1 312
	34	41	1 156	1 681	1 394
Somme	261	304	7 127	9 734	8 286

$$\rho_{x,y} = \frac{(10)(8\ 286) - (261)(304)}{\sqrt{(10)(7\ 127) - 261^2} \sqrt{(10)(9\ 734) - 304^2}} = \frac{3\ 516}{(56,11)(70,17)}$$

soit $\rho_{x,y} = 0,89$ et $\rho_{x,y}^2 = 0,79$

Le t de Student empirique (d'après [3]) est égal à :

$$t^* = \frac{|\rho_{x,y}|}{\sqrt{\frac{(1 - \rho_{x,y}^2)}{n - 2}}} = \frac{0,89}{0,1620} = 5,49 > t_8^{0,025} = 2,306$$

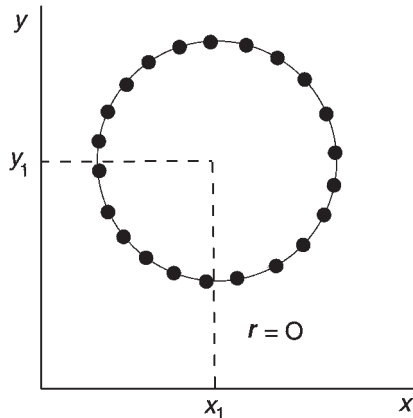
le coefficient de corrélation entre x et y est significativement différent de 0.

2.2 Limites de la notion de corrélation

■ La relation testée est linéaire

L'application de la formule [1] ou [2] ne permet de déterminer que des corrélations linéaires entre variables. Un coefficient de corrélation nul indique que la covariance entre la variable x et la variable y est égale à 0. C'est ainsi que deux variables en totale dépendance peuvent avoir un coefficient de corrélation nul, comme l'illustre l'exemple suivant : l'équation d'un cercle nous est donnée par $(x - x_1)^2 + (y - y_1)^2 = R^2$, les variables x et y sont bien liées entre elles fonctionnellement (graphique 7) et pourtant leur covariance est nulle et donc leur coefficient de corrélation égal à 0.

Pour pallier cette limite, il convient éventuellement de transformer les variables, préalablement au calcul du coefficient de corrélation, afin de linéariser leur relation, par exemple au moyen d'une transformation de type logarithmique.



Graphique 7 – La relation fonctionnelle n'est pas corrélation linéaire

■ **Corrélation n'est pas causalité**

Le fait d'avoir un coefficient de corrélation élevé entre deux variables ne signifie pas qu'il existe un autre lien que statistique. En d'autres termes, une covariance significativement différente de 0 n'implique pas une liaison d'ordre économique, physique ou autre. Nous appelons *corrélation fortuite* ce type de corrélation que rien ne peut expliquer.

L'exemple le plus fameux concerne la forte corrélation existante entre le nombre de taches solaires observées et le taux de criminalité aux États-Unis. Cela ne signifie pas qu'il existe une relation entre les deux variables, mais qu'une troisième variable, l'évolution de long terme (la tendance) ici, explique conjointement les deux phénomènes. La théorie de la cointégration traite de ce problème (*cf.* chapitre 11).