

RÉGIS BOURBONNAIS

Écono- métrie

•
E
C
O
S
U
P
•

12^E ÉDITION

DUNOD

© Dunod, 2024
11, rue Paul Bert, 92240 Malakoff
www.dunod.com
ISBN 978-2-10-087202-2

● Table des matières

Avant-propos	XI
Chapitre 1 Qu'est-ce que l'économétrie ?	1
1. La notion de modèle	1
1.1 Définition	1
1.2 La construction des modèles en économétrie	2
2. Le rôle de l'économétrie	4
2.1 L'économétrie comme validation de la théorie	4
2.2 L'économétrie comme outil d'investigation	5
3. La théorie de la corrélation	5
3.1 Présentation générale	5
3.2 Mesure et limite du coefficient de corrélation	7
L'essentiel	12
Chapitre 2 Le modèle de régression simple	13
1. Présentation du modèle	13
1.1 Exemple introductif	13
1.2 Rôle du terme aléatoire	14
1.3 Conséquences du terme aléatoire	16
2. Estimation des paramètres	17
2.1 Modèle et hypothèses	17
2.2 Formulation des estimateurs	18
2.3 Les différentes écritures du modèle : erreur et résidu	21
2.4 Propriétés des estimateurs	21
3. Conséquences des hypothèses : construction des tests	23
3.1 Hypothèse de normalité des erreurs	23
3.2 Conséquences de l'hypothèse de normalité des erreurs	24
3.3 Test bilatéral, test unilatéral et probabilité critique d'un test	27

4. Équation et tableau d'analyse de la variance	33
4.1 Équation d'analyse de la variance	33
4.2 Tableau d'analyse de la variance	34
5. La prévision dans le modèle de régression simple	39
L'essentiel	48

Chapitre 3 Le modèle de régression multiple	49
1. Le modèle linéaire général	50
1.1 Présentation	50
1.2 Forme matricielle	50
2. Estimation et propriétés des estimateurs	51
2.1 Estimation des coefficients de régression	51
2.2 Hypothèses et propriétés des estimateurs	53
2.3 Équation d'analyse de la variance et qualité d'un ajustement	55
3. Les tests statistiques	60
3.1 Le rôle des hypothèses	60
3.2 Construction des tests	61
3.3 Tests sur les résidus : valeur anormale, effet de levier et point d'influence	63
4. L'analyse de la variance	69
4.1 Construction du tableau d'analyse de la variance et test de signification globale d'une régression	69
4.2 Autres tests à partir du tableau d'analyse de la variance	70
4.3 Généralisation des tests par analyse de la variance	76
5. L'utilisation de variables indicatrices	78
5.1 Constitution et finalités des variables indicatrices	78
5.2 Exemples d'utilisation	78
6. La prévision à l'aide du modèle linéaire général et la régression récursive	87
6.1 Prédiction conditionnelle	87
6.2 Fiabilité de la prévision et intervalle de prévision	87
6.3 Les tests de stabilité par la régression récursive	90
6.4 Le test de spécification de Ramsey	91

7. Exercices récapitulatifs	95
Annexes	113
Interprétation géométrique de la méthode des moindres carrés	113
Résolution de l'exercice 3.1 par des logiciels informatiques de régression multiple	114
Estimation de la variance de l'erreur	116
L'essentiel	117
 Chapitre 4 Multicolinéarité et sélection du modèle optimal	119
1. Corrélation partielle	120
1.1 Exemple introductif	120
1.2 Généralisation de la notion de corrélation partielle	120
2. Relation entre coefficients de corrélation simple, partielle et multiple	125
3. Multicolinéarité: conséquences et détection	127
3.1 Conséquences de la multicolinéarité	127
3.2 Tests de détection d'une multicolinéarité	128
3.3 Comment remédier à la multicolinéarité ?	131
4. Sélection du modèle optimal	132
L'essentiel	139
 Chapitre 5 Problèmes particuliers : la violation des hypothèses	141
1. L'autocorrélation des erreurs	142
1.1 Présentation du problème	142
1.2 L'estimateur des moindres carrés généralisés (MCG)	142
1.3 Les causes et la détection de l'autocorrélation des erreurs	143
1.4 Les procédures d'estimation en cas d'autocorrélation des erreurs	150
2. L'hétéroscédasticité	158
2.1 Présentation du problème	158
2.2 Correction de l'hétéroscédasticité	160
2.3 Tests de détection de l'hétéroscédasticité	163
2.4 Autre test d'hétéroscédasticité : le test ARCH	170

3. Modèles à erreurs sur les variables	171
3.1 Conséquences lorsque les variables sont entachées d'erreurs	171
3.2 La méthode des variables instrumentales	172
3.3 Le test d'exogénéité d'Hausman	173
3.4 La méthode des moments généralisée	173
L'essentiel	182
 Chapitre 6 Les modèles non linéaires	183
1. Les différents types de modèles non linéaires	183
1.1 Les fonctions de type exponentiel	183
1.2 Les modèles de diffusion	186
2. Méthodes d'estimation des modèles non linéaires	188
2.1 Initiation aux méthodes d'estimation non linéaires	188
2.2 Exemples d'application	190
L'essentiel	195
 Chapitre 7 Les modèles à décalages temporels	197
1. Les modèles linéaires autorégressifs	197
1.1 Formulation générale	197
1.2 Test d'autocorrélation et méthodes d'estimation	198
2. Les modèles à retards échelonnés	203
2.1 Formulation générale	203
2.2 Détermination du nombre de retards	204
2.3 Distribution finie des retards	208
2.4 Distribution infinie des retards	213
3. Deux exemples de modèles dynamiques	218
3.1 Le modèle d'ajustement partiel	218
3.2 Le modèle d'anticipations adaptatives	219
L'essentiel	239

Chapitre 8	Introduction aux modèles à équations simultanées	241
1.	Équations structurelles et équations réduites	242
1.1	Exemple introductif	242
1.2	Le modèle général	243
2.	Le problème de l'identification	244
2.1	Restrictions sur les coefficients	244
2.2	Conditions d'identification	245
3.	Les méthodes d'estimation	246
3.1	Les moindres carrés indirects	246
3.2	Les doubles moindres carrés	247
3.3	Autres méthodes d'estimation	248
	Annexes	260
	Identification : les conditions de rang	260
	L'essentiel	263
Chapitre 9	Éléments d'analyse des séries temporelles	265
1.	Stationnarité	266
1.1	Définition et propriétés	266
1.2	Fonctions d'autocorrélation simple et partielle	266
1.3	Tests de « bruit blanc » et de stationnarité	268
2.	La non-stationnarité et les tests de racine unitaire	271
2.1	La non-stationnarité : les processus TS et DS	271
2.2	Les tests de racine unitaire et la stratégie séquentielle de test	275
3.	Les modèles ARIMA	283
3.1	Typologie des modèles AR, MA et ARMA	283
3.2	L'extension aux processus ARIMA et SARIMA	286
4.	La méthode de Box et Jenkins	287
4.1	Recherche de la représentation adéquate : l'identification	287
4.2	Estimation des paramètres	288
4.3	Tests d'adéquation du modèle et prévision	289
	L'essentiel	305

Chapitre 10	La modélisation VAR	307
1.	Représentation d'un modèle VAR	308
1.1	Exemple introductif	308
1.2	La représentation générale	309
1.3	La représentation ARMAX	310
2.	Estimation des paramètres	310
2.1	Méthode d'estimation	311
2.2	Détermination du nombre de retards	311
2.3	Prévision	312
3.	Dynamique d'un modèle VAR	317
3.1	Représentation VMA d'un processus VAR	317
3.2	Analyse et orthogonalisation des « chocs »	318
3.3	Décomposition de la variance	321
3.4	Choix de l'ordre de décomposition	321
4.	La causalité	325
4.1	Causalité au sens de Granger	325
4.2	Causalité au sens de Sims	326
	L'essentiel	330
Chapitre 11	La cointégration et le modèle à correction d'erreur	331
1.	Exemples introductifs	331
1.1	Premier exemple	331
1.2	Deuxième exemple	332
2.	Le concept de cointégration	333
2.1	Propriétés de l'ordre d'intégration d'une série	333
2.2	Conditions de cointégration	335
2.3	Le modèle à correction d'erreur (ECM)	336
3.	Cointégration entre deux variables	337
3.1	Test de cointégration entre deux variables	337
3.2	Estimation du modèle à correction d'erreur	337
4.	Généralisation à k variables	340
4.1	La cointégration entre k variables	341
4.2	Estimation du modèle à correction d'erreur	342

4.3	Le modèle à correction d'erreur vectoriel	342
4.4	Tests de relation de cointégration	344
4.5	Test d'exogénéité faible	347
4.6	Synthèse de la procédure d'estimation	348
	L'essentiel	354
Chapitre 12	Introduction à l'économétrie des variables qualitatives	355
1.	Les problèmes et les conséquences de la spécification binaire	356
2.	Les modèles de choix binaires	358
2.1	Le modèle linéaire sur variable latente	358
2.2	Les modèles Probit et Logit	359
2.3	Interprétation des résultats et tests statistiques	360
3.	Les modèles à choix multiples	365
3.1	Les modèles Probit et Logit ordonnés	366
3.2	Le modèle de choix multiples non ordonné : le Logit multinomial	371
4.	Les modèles à variable dépendante limitée : le modèle Tobit	372
4.1	Le modèle Tobit simple : modèle de régression tronqué ou censuré	373
4.2	Estimation et interprétation des résultats	375
	L'essentiel	381
Chapitre 13	Introduction à l'économétrie des données de panel	383
1.	Présentation des modèles à données de panel	384
1.1	Spécificités des données de panel	384
1.2	La méthode SUR	385
1.3	Le modèle linéaire simple	386
2.	Les tests d'homogénéité	387
2.1	Procédure séquentielle de tests	387
2.2	Construction des tests	387

3. Spécifications et estimations des modèles à effets individuels	393
3.1 Le modèle à effets fixes individuels	393
3.2 Le modèle à effets aléatoires	394
3.3 Effets fixes ou effets aléatoires ? Le test d'Hausman	395
L'essentiel	400
Testez-vous - Testez vos connaissances sur les chapitres 1 à 7	401
Liste des exercices	411
Tables statistiques	415
Bibliographie	423
Index	427

Avant-propos

Cette douzième édition 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. La mise en page en facilite la lecture et met l'accent sur les parties importantes du cours. À la fin de chaque chapitre une synthèse est proposée. Un QCM permet aux étudiants de vérifier la connaissance des fondamentaux du cours.

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. 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.

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 :

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.

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.

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 Paris Dauphine-PSL, place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16.

E-mail : regis.bourbonnais@dauphine.psl.eu

Qu'est-ce que l'économétrie ?

Introduction

Ce premier chapitre est consacré à la présentation de l'économétrie et à sa liaison avec la théorie économique. Nous abordons tout d'abord en 1 la notion de modèle ainsi que les différentes étapes de la modélisation. L'apport de l'économétrie en tant qu'outil de validation est étudié en 2. Enfin, la théorie de la corrélation – fondement de l'économétrie – fait l'objet du 3.

Objectifs

- Présenter** la notion de modèle en économétrie.
- Définir** le rôle de l'économétrie.
- Savoir** calculer et interpréter un coefficient de corrélation.

Concepts clés

Construction des modèles en économétrie
Différents types de variables : série temporelle, coupe instantanée, panel, cohorte
Théorie de la corrélation

1 La notion de modèle

1.1 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 ?

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

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.

1.2 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 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.

a) 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.

b) 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 = propension marginale à consommer
et a_0 = 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).

c) 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.

d) 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

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...

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 ».

DÉFINITIONS

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

e) 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.

2 Le rôle de l'économétrie

2.1 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.

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

2.2 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 :

- 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.

3 La théorie de la corrélation

3.1 Présentation générale

Lorsque deux phénomènes ont une évolution commune, nous disons qu'ils sont « corrélés ». La corrélation 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 %).

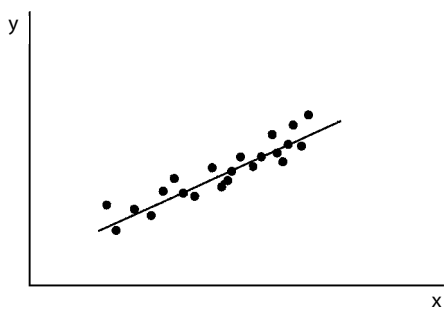
2. Pour découvrir l'utilisation de l'économétrie à des fins de prévision de ventes, voir BOURBONNAIS R. et USUNIER J.-C. (2017).

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

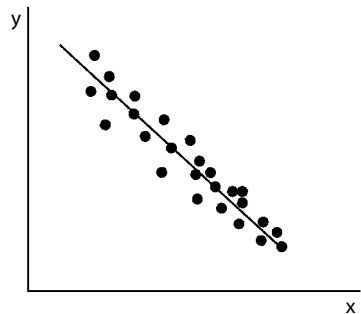
Le tableau 1.1, en croisant les critères de linéarité et de corrélation, renvoie à une représentation graphique.

Tableau 1.1 – Linéarité et corrélation

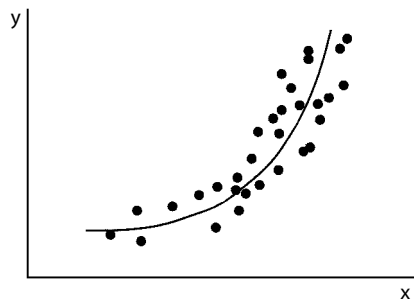
	Corrélation positive	Corrélation négative	Absence de corrélation
Relation linéaire	Graphe 1.1	Graphe 1.2	Graphe 1.5
Relation non linéaire	Graphe 1.3	Graphe 1.4	Graphe 1.5



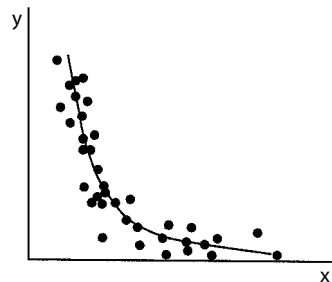
Graphe 1.1



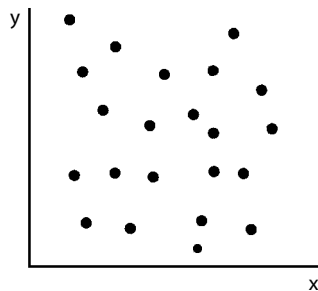
Graphe 1.2



Graphe 1.3



Graphe 1.4



Graphe 1.5

3.2 Mesure et limite du coefficient de corrélation

a) 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 ou 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 :

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

[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 D'APPLICATION 1.1

 **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 1.2.

Tableau 1.2 – 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

1 ■ Tracez le nuage de points et commentez-le.

2 ■ Calculez le coefficient de corrélation simple et testez sa significativité par rapport à 0 pour un seuil $\alpha = 0,05$.

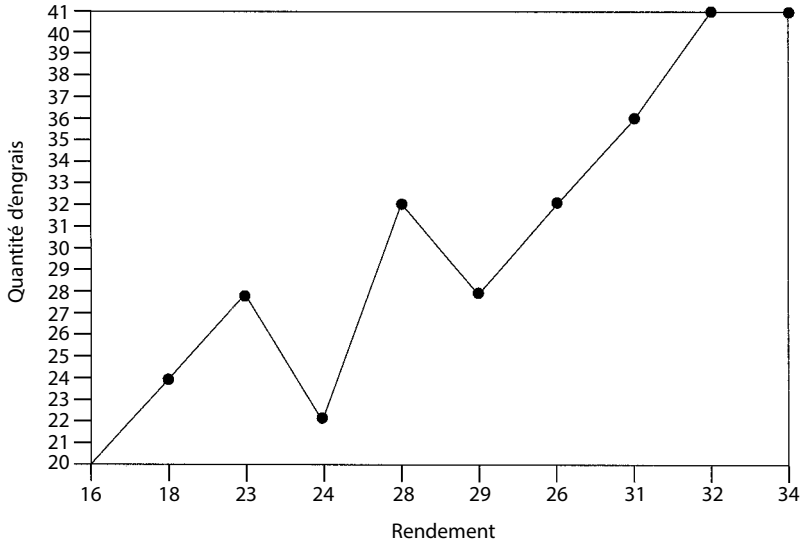
▶

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 1.6) indique que les couples de valeurs sont approximativement alignés : les deux variables semblent corrélées positivement.



Graphique 1.6 – Nuage du couple de valeurs : rendement-quantité d'engrais

2 ■ Afin d'appliquer la formule [2], nous dressons le tableau de calcul 1.3.

Tableau 1.3 – 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 :

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

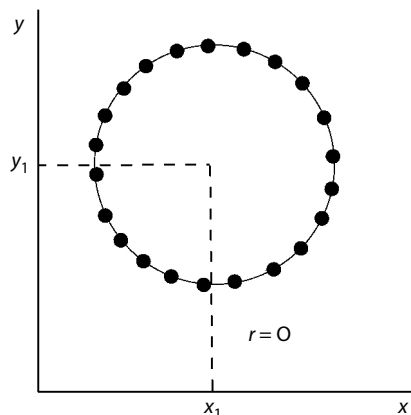
Le coefficient de corrélation entre x et y est significativement différent de 0.

b) 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 1.7) 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 1.7 – La relation fonctionnelle n'est pas corrélation linéaire