

L 3.11

M4

6601

ALBAN THOMAS

**MANUEL ET EXERCICES CORRIGÉS**

# **Économétrie des variables qualitatives**



**DUNOD**

025529164

33

# Économétrie des variables qualitatives

D2

2000 - 61135

# Collection Éco Sup

## Manuels et exercices corrigés

- R. Bourbonnais  
*Économétrie*
- Y. Dodge  
*Analyse de régression appliquée*
- B. Goldfarb, C. Pardoux  
*Introduction à la méthode statistique (Gestion. Économie)*
- O. Jokung-Nguéna  
*Microéconomie de l'incertain (Risque et décisions)*
- J.-P. Lecoutre  
*Statistique et probabilités*
- P. Petauton  
*Théorie de l'assurance dommages  
Théorie et pratique de l'assurance vie*
- J.-L. Viviani  
*Gestion de portefeuille*
- T. de Montbrial, E. Fauchart  
*Introduction à l'économie (Microéconomie. Macroéconomie)*
- A. Thomas  
*Économétrie des variables qualitatives*

## Manuels

- B. Bernier, Y. Simon  
*Initiation à la macroéconomie*
- B. Bernier, H. L. Védie  
*Initiation à la microéconomie*
- J. Fourastié  
*Mathématiques appliquées à l'économie*
- B. Grais  
*Statistique descriptive  
Méthodes statistiques*
- B. Guillochon  
*Économie internationale*
- P. Kauffmann  
*Statistiques. Information, estimation, tests*
- J. de Lagarde  
*Initiation à l'analyse des données*
- A. Planche  
*Mathématiques pour économistes.  
Algèbre  
Mathématiques pour économistes.  
Analyse*

- F. Poulon  
*Économie générale*
- A. Redslob  
*Introduction à la théorie macroéconomique*
- B. et D. Saby  
*Les grandes théories économiques*
- A. de Servigny, I. Zelenko  
*Économie financière*
- J.-L. Sol  
*Mathématiques. Accès à l'université*
- L. Stoléry  
*L'économie  
(Comprendre l'avenir)*
- D. Temam  
*La nouvelle comptabilité nationale*
- A. Varoudakis  
*La politique macroéconomique*

## Exercices corrigés avec rappels de cours

- J. Fourastié  
*Mathématiques appliquées à l'économie*
- B. Grais  
*Statistique descriptive*
- J.-P. Lecoutre, P. Pilibossian  
*Algèbre  
Analyse II*
- J.-P. Lecoutre, S. Maille-Legait, P. Tassi  
*Statistique*
- R. Sandretto  
*Probabilités*

## Travaux dirigés avec rappels de cours

- S. Brana, M.-C. Bergouignan  
*Macroéconomie*
- S. Brana, M. Cazals, P. Kauffmann  
*Économie monétaire et financière*
- B. Guillochon, A. Kawecky  
*Économie internationale*
- J.-P. Lecoutre, P. Pilibossian  
*Analyse I*
- P. Médan  
*Microéconomie*

**MANUEL ET EXERCICES CORRIGÉS**

ALBAN THOMAS

Enseignant à l'université des sciences sociales de Toulouse  
Directeur de recherche à l'INRA

# Économétrie des variables qualitatives

DUNOD



DL- 20.04.2000

17265

Ce pictogramme 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<sup>er</sup> 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 du Centre français d'exploitation du

droit de copie (CFC, 20 rue des Grands-Augustins, 75006 Paris).



© Dunod, Paris, 2000

ISBN 2 10 004665 9

Toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite selon le Code de la propriété intellectuelle (Art L. 122-4) et constitue une contrefaçon réprimée par le Code pénal. • Seules sont autorisées (Art L. 122-5) les copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective, ainsi que les analyses et courtes citations justifiées par le caractère critique, pédagogique ou d'information de l'œuvre à laquelle elles sont incorporées, sous réserve, toutefois, du respect des dispositions des articles L. 122-10 à L. 122-12 du même Code, relatives à la reproduction par reprographie.





# Table des matières

Avant-propos	IX
<b>1. Introduction aux variables qualitatives</b>	<b>1</b>
I. Classification des variables qualitatives	2
II. Variables indicatrices et régression	4
III. Application	7
IV. Tableaux de contingence	8
V. Application : revenu et nombre d'enfants	10
VI. Modèle log-linéaire	13
VII. Application numérique	15
VIII. Le $\chi^2$ minimum	19
IX. Modèle à probabilité linéaire	22
À retenir	23
Exercices	23
<b>2. Méthodes d'estimation</b>	<b>27</b>
I. Le maximum de vraisemblance	27
A. Définition de la vraisemblance	28
B. Dérivation et propriétés de l'estimateur	29
C. Vraisemblance concentrée	30
D. Inférence statistique	31
E. Exemple : régression hétéroscédastique	34
II. Les modèles de comptage	34
A. Modèle de Poisson	35
B. Extensions	36
C. Application : choix d'irrigation	37
III. Estimateurs des moments	39
A. L'estimateur par variable instrumentale	39
B. L'estimateur GMM	40
C. Exemple : les moindres carrés généralisés	43
D. Inférence statistique	44

E. Application : valeurs foncières	45
IV. Algorithmes d'optimisation	47
A. Position du problème	48
B. Méthodes de gradient	49
À retenir	50
<b>3. Modèles à variable qualitative binaire</b>	<b>51</b>
I. Variable binaire et variable latente	51
II. Modèles binaires Logit et Probit	54
A. Le modèle Probit	55
B. Le modèle Logit	57
C. Comparaison entre le Probit et le Logit	59
III. Inférence dans les modèles binaires	60
A. Interprétation des paramètres estimés	60
B. Calcul des effets marginaux	60
C. Test d'ajustement global	61
D. Critères de choix de modèles	62
E. Le problème des variables omises	64
F. Test de l'hétéroscédasticité	65
IV. Application : choix du mode de climatisation	66
V. Extensions	70
A. Les modèles binaires bivariés	70
B. Application	72
C. Méthodes d'estimation non paramétriques	73
D. Modèles avec données de panel	76
À retenir	78
Exercices	78
<b>4. Les modèles multinomiaux</b>	<b>83</b>
I. Introduction	84
II. Les modèles ordonnés	85
A. Exemple : enquête de consommation	85
B. Formulation courante d'un modèle ordonné	86
III. Application	87
IV. Les modèles séquentiels	90
V. Modèles non ordonnés : le Logit multinomial	91
A. Généralisation du Logit binaire au cas multinomial	91
B. Logit multinomial : spécification et estimation	92
C. Le Logit conditionnel de McFadden	94
D. Un modèle Logit plus général	96
VI. Application : mode de transport	97

VII.	Modèles de choix probabilistes et hypothèse IIA	99
	A. Modèles probabilistes à choix discret	100
	B. L'hypothèse IIA	101
	C. Un test de la propriété IIA	103
	D. Application numérique	104
VIII.	Modèles multinomiaux alternatifs	107
	A. Le Logit multinomial emboîté	107
	B. Le Logit multinomial hiérarchisé	108
	C. Application : transport aérien ou terrestre	111
	D. Le Probit multinomial	113
IX.	Extensions	114
	A. Estimateur MSM de McFadden	114
	B. L'estimateur GHK	116
	<i>À retenir</i>	118
	<i>Exercices</i>	118
<b>5.</b>	<b>Les modèles à variable dépendante limitée</b>	<b>121</b>
I.	Introduction	121
II.	Modèles censurés et tronqués	123
III.	Le modèle Tobit simple	126
	A. Comparaison avec les modèles tronqués	127
	B. Estimation par maximum de vraisemblance	128
	C. Estimation en deux étapes	130
	D. Effets marginaux avec le modèle Tobit	131
IV.	Tests de spécification	132
	A. Hétéroscédasticité	132
	B. Normalité	133
	C. Égalité des paramètres	134
V.	Application : dépenses de restaurant	134
VI.	Extensions du modèle Tobit	137
	A. Modèles Tobit à censure multiple	138
	B. Exemple : les modèles à friction	139
	C. Modèles Tobit généralisés	140
	D. Modèles à seuils stochastiques	142
VII.	Modèles de sélection	144
	A. Mécanisme de sélection	144
	B. Troncature auxiliaire	145
	C. Application : dépenses en eau domestique	146
	D. Application : équation de salaire	148
	E. Modèles à régimes	149
	F. Application : dépenses en logement	151
	<i>À retenir</i>	155
	<i>Exercices</i>	155



Corrigés des exercices	159
Bibliographie	169
Index	177



# Avant-propos

Les variables qualitatives constituent une partie importante de l'économétrie contemporaine. Il existe trois façons de considérer les variables qualitatives en économétrie et en statistique :

- les incorporer comme variables explicatives dans un modèle de régression ;
- étudier leur corrélation ou leur indépendance ;
- les traiter comme variables dépendantes (expliquées).

Ces variables à valeurs discrètes forment souvent la majorité des informations disponibles dans les enquêtes, en raison notamment de leur grande facilité de collecte. D'autre part, les modèles économiques expliquant des événements et non des réalisations de variables continues ont connu un développement important. Les théories du consommateur et du producteur, par exemple, fournissent à présent les bases d'une formalisation rigoureuse des mécanismes de choix des agents économiques : décision de travail, d'achat, choix d'une technologie, etc.

Mais paradoxalement, alors que les variables qualitatives sont de plus en plus utilisées, leur analyse est encore porteuse d'une réputation de complexité, probablement parce que les techniques d'inférence par moindres carrés ne sont pas adaptées. La méthode du *maximum de vraisemblance* devant être utilisée dans la plupart des cas, l'économétrie des variables qualitatives entraînerait donc une grande complexité dans les calculs. Cet ouvrage a pour objectif de montrer que les techniques de modélisation des variables qualitatives possèdent l'avantage d'une grande cohérence méthodologique. Une fois que les principes à la base de la construction de la vraisemblance ont été compris, il est très facile d'estimer des modèles à variable dépendante qualitative. Les modèles les plus courants sont maintenant gérés par la plupart des logiciels économétriques et, pour les modèles plus élaborés, le langage de programmation dédié des logiciels comme SAS, GAUSS, EVIEWS et LIMDEP peut être mis à profit. Ces logiciels modernes permettent en outre de maximiser toute vraisemblance avec des algorithmes de maximisation numériques optimisés, donc fiables et rapides.

Cet ouvrage se limite aux modèles les plus couramment utilisés en économétrie appliquée. Des extensions aux chapitres présentent cependant des résultats récents en économétrie des variables qualitatives. On ne traitera ni

des modèles de durée, ni des modèles multivariés. Le lecteur intéressé trouvera ces thèmes bien traités dans les ouvrages de Gouriéroux (1989), Maddala (1983) et Lancaster (1990). L'optique de l'ouvrage est résolument appliquée ; on illustre les principaux modèles et méthodes au moyen d'applications sur données réelles, avec les commandes pour les logiciels SAS, GAUSS, EVIEWS et LIMDEP. Dans le domaine des variables qualitatives, le logiciel le plus élaboré est certainement LIMDEP, permettant d'estimer quasiment tous les modèles décrits dans cet ouvrage. GAUSS possède des procédures standard pour les modèles de base (essentiellement ceux des chapitres 1, 3 et 4), mais son langage de programmation très intuitif le rend intéressant pour appréhender des modélisations plus avancées. EVIEWS pour sa part est assez comparable à GAUSS : son langage de programmation est moins immédiat, mais la gestion des fichiers de données est plus aisée. Enfin le logiciel SAS, s'il est très utile pour manipuler des bases de données importantes, est peut-être le moins adapté aux méthodes de ce livre : langage de programmation difficile à mettre en œuvre, langage matriciel vite limité, très peu de procédures standard pour les modèles à variable dépendante qualitative.

### Prérequis

Cet ouvrage s'adressant à des étudiants de deuxième et troisième cycles universitaires et de grandes écoles, on demandera au lecteur des connaissances de base en statistique et économétrie : tests, principes de l'estimation paramétrique, algèbre linéaire de base, moindres carrés ordinaires. Le lecteur intéressé par l'acquisition de telles notions pourra se référer utilement aux ouvrages de Régis Bourbonnais, *Économétrie*, et Pascal Kauffmann, *Statistique*, dans la même collection.

### Notations

On utilisera dans la mesure du possible des notations mathématiques cohérentes et invariantes entre les chapitres. L'entier  $N$  dénote la taille de l'échantillon, pour des individus indicés par  $i$ ,  $i = 1, 2, \dots, N$ . Les modalités et choix sont repérés par les indices  $j, k, l$ , et le nombre de modalités par  $m$  ou  $m_i$  selon les cas. La notation  $y_i = \{0, 1\}$  signifie que  $y_i$  peut prendre deux valeurs : 0 ou 1. Les variables expliquées, continues ou à valeurs discrètes sont notées  $y_i$ , les variables explicatives  $x_i$ , et un vecteur ( $1 \times K$ ) ou encore  $z_i$ . Les paramètres sont notés  $\beta$  (un vecteur  $K \times 1$ ) ou  $\gamma$  (un vecteur de dimension  $G \times 1$ ). Ainsi, la combinaison linéaire formée à partir d'un vecteur de variables explicatives pour l'individu  $i$ ,  $x_i$ , et d'un vecteur de paramètres  $\beta$  sera :  $x_i\beta = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \dots + \beta_Kx_{iK}$ .

Les variables aléatoires  $\varepsilon_i$ ,  $u_i$  ou  $v_i$  sont les résidus (termes d'erreur) et les variables latentes sont notées  $y_i^*$ . La vraisemblance est notée  $L$ , le gradient  $g(\cdot)$ , la matrice hessienne  $H(\cdot)$ . En ce qui concerne les dimensions des vecteurs et matrices, la notation  $A$  ( $N \times K$ ) indique que la matrice  $A$  a  $N$  lignes et  $K$  colonnes. De même,  $B$  ( $N \times 1$ ) signifie que le vecteur  $B$  a  $N$  lignes et 1 colonne : c'est un vecteur colonne.

La transposée d'une matrice ou vecteur  $x$  est notée  $x'$ , et on rappelle ici les règles de dérivation matricielle :  $\partial(x\beta)/\partial\beta = x'$  ;  $\partial(x'Ax)/\partial x = 2Ax$  si  $A$  est symétrique,  $(A + A')x$  sinon. Rappelons de plus l'écriture de la dérivée seconde par rapport à une matrice ou un vecteur :  $\partial^2 f(\beta)/\partial\beta\partial\beta'$ . Par exemple, les dérivées de la fonction  $f = (y_i - x_i\beta)^2$  par rapport au vecteur de paramètres  $\beta$  de dimension  $K \times 1$  sont :  $\partial f/\partial\beta = -2(y_i - x_i\beta) \times x'_i$ , un vecteur  $K \times 1$ , et  $\partial^2 f/\partial\beta\partial\beta' = 2x'_i x_i$ , une matrice  $K \times K$ .

En ce qui concerne les lois statistiques,  $f(\cdot)$  dénote une fonction de densité,  $F(\cdot)$  une fonction de répartition, et dans le cas de la loi normale,  $\phi(\cdot)$  et  $\Phi(\cdot)$  respectivement. Enfin, on utilise dans le texte la notation décimale française avec virgule. En revanche, les exemples d'application étant illustrés de sorties de logiciels anglo-saxons, celles-ci utilisent la convention décimale avec un point.

**N**ous avons écrit ce manuel à l'usage de ceux qui souhaitent découvrir les principes de base de l'économie et de la statistique. Ce manuel est destiné à servir de référence pour les étudiants de première année de licence en économie et de première année de maîtrise en statistique. Il est également destiné à servir de référence pour les étudiants de deuxième année de licence en économie et de deuxième année de maîtrise en statistique. Les méthodes de régression sont le cas d'application principal de ce manuel. Le chapitre s'achève par un exposé sur les résultats obtenus lors de la régression multiple, ainsi que les résultats obtenus lors de l'analyse par moindres carrés de régression des données.

**Contenu du chapitre :** nous verrons d'abord comment estimer les paramètres d'un modèle linéaire ; puis nous verrons comment estimer les paramètres d'un modèle non linéaire ; nous verrons également comment estimer les paramètres d'un modèle à variables multiples ; nous verrons enfin comment estimer les paramètres d'un modèle à variables multiples.

The first part of the paper discusses the importance of the  
second part discusses the importance of the  
third part discusses the importance of the  
fourth part discusses the importance of the  
fifth part discusses the importance of the  
sixth part discusses the importance of the  
seventh part discusses the importance of the  
eighth part discusses the importance of the  
ninth part discusses the importance of the  
tenth part discusses the importance of the



# 1. Introduction aux variables qualitatives

**N**ous introduisons dans ce chapitre l'analyse des variables qualitatives, en discutant d'abord de leur provenance et intérêt. On présente l'étude de modèles de régression avec variables indicatrices en décrivant l'interprétation de leur coefficient. On étudie ensuite les modalités d'analyse de la corrélation entre les variables explicatives, dans un contexte purement descriptif, à l'aide de tableaux de contingence et du modèle log-linéaire. Les méthodes de régression dans le cas d'observations groupées seront ensuite développées autour de la méthode du  $\chi^2$  minimum. Ce chapitre s'achèvera par un aperçu des modèles linéaires avec variable dépendante qualitative, dont les inconvénients motiveront l'analyse par *maximum de vraisemblance* des chapitres 3, 4 et 5.

**Objectifs du chapitre :** proposer une classification des variables qualitatives ; présenter les principales procédures d'analyse et de test utilisées en statistique descriptive des variables qualitatives ; introduire les modèles à variable expliquée qualitative.

**Concepts clés étudiés :** enquêtes, tableaux de contingence, tests du  $\chi^2$ , modèle log-linéaire, méthode du  $\chi^2$  minimum, modèle à probabilité linéaire.

# I. Classification des variables qualitatives

Les variables qualitatives sont synonymes de variables à valeurs discrètes (ou plus simplement variables discrètes), c'est-à-dire prenant leurs valeurs dans l'ensemble des entiers naturels. La nature d'une variable qualitative dépend de la façon dont elle a été codée sur un questionnaire ou enregistrée dans une base de données. Dans ce cas, elle apparaît souvent sous une forme aisée à saisir ou à interpréter ultérieurement lors de différents travaux statistiques. Qu'elles proviennent d'une enquête statistique ou d'un fichier constitué par un économètre ayant créé ses propres variables, les variables discrètes représentent souvent la vision qu'a leur créateur du problème à traiter<sup>1</sup>.

On considère deux grandes classes de variables discrètes :

- les variables *binaires*, à deux modalités possibles seulement ;
- les variables *polytomiques*, à plus de deux valeurs possibles.

Dans le premier cas, on parle aussi de variables *dichotomiques*, et dans le second cas, de variables *multinomiales*. Les variables binaires sont souvent sous la forme 0 ou 1. Un exemple est la variable « Sexe de l'individu », égale à 0 ou 1 s'il s'agit d'un homme ou d'une femme, respectivement. On peut créer par exemple cette variable à partir d'une variable notée *SECU* et contenant le numéro de Sécurité sociale à 13 chiffres de l'individu :

$$SEXE = \begin{cases} 1 & \text{si } SECU \text{ commence par 2,} \\ 0 & \text{si } SECU \text{ commence par 1.} \end{cases}$$

La convention est de prendre deux valeurs : 0 ou 1, le 1 indiquant l'existence de la caractéristique, 0 son absence<sup>2</sup>. Une variable binaire, admettant une valeur 0 ou 1, est caractérisée par le fait que son espérance est la proportion sur l'échantillon de cas où la variable vaut 1. En effet, l'espérance d'une variable binaire  $x$  est :

$$E(x) \approx \frac{1}{N} \sum_{i=1}^N x_i = \frac{\text{Nombre de cas où } x \text{ vaut 1}}{\text{Nombre total de cas}} = p(x)$$

où  $N$  est la taille de l'échantillon (nombre d'observations) et  $p(x)$  la fréquence empirique de  $x$  (la proportion). Si la probabilité théorique que  $x = 1$  peut être estimée par la fréquence empirique, alors l'espérance d'une variable binaire

1. Ceci a pour conséquence qu'une variable qualitative peut la plupart du temps être recodifiée de façon à mieux correspondre aux besoins des utilisateurs de la base de données.

2. D'où le nom de variable binaire, l'analogie étant possible avec l'électronique digitale où le 1 représente le passage de courant (« état haut »), le 0 son absence (« état bas »).

est comparable à la probabilité d'occurrence de la catégorie à laquelle  $x$  fait référence.

Dans le second cas, celui des variables polytomiques, la nature des variables dépend du besoin de représenter ou non des catégories. Si une variable décrivant un individu représente les mois passés à la recherche d'un emploi, on a bien affaire à une variable discrète, mais cette variable ne décrit nullement une caractéristique ou une catégorie. Si, au contraire, une variable contient la réponse à une question relative à une caractéristique de l'individu ou à un choix (catégorie socioprofessionnelle avant la perte d'emploi, secteurs dans lesquels la personne souhaiterait trouver un emploi), on parle de *variable catégorielle*. Prenons quelques exemples :

$$REVENU = \begin{cases} 1 & \text{si l'individu gagne moins de 72 000 F par an ;} \\ 2 & \text{si l'individu gagne entre 72 000 F et 96 000 F par an ;} \\ 3 & \text{si l'individu gagne entre 96 000 F et 120 000 F par an ;} \\ 4 & \text{si l'individu gagne entre 120 000 F et 180 000 F par an ;} \\ 5 & \text{si l'individu gagne plus de 180 000 F par an.} \end{cases}$$

$$EMPLOI = \begin{cases} 1 & \text{si exploitant agricole ;} \\ 2 & \text{si ouvrier ;} \\ 3 & \text{si employé ;} \\ 4 & \text{si cadre ;} \\ 5 & \text{si sans profession, étudiant.} \end{cases}$$

$$ÉDUCATION = \begin{cases} 1 & \text{si la personne n'a pas le baccalauréat ;} \\ 2 & \text{si la personne a le baccalauréat mais pas le DEUG ;} \\ 3 & \text{si la personne a le DEUG mais pas la licence ;} \\ 4 & \text{si la personne a la licence mais pas la maîtrise ;} \\ 5 & \text{si l'individu a la maîtrise mais pas de troisième cycle ;} \\ 6 & \text{si l'individu a un troisième cycle.} \end{cases}$$

$$PRÉFÉRENCE = \begin{cases} 1 & \text{si la personne n'aime pas du tout le roquefort ;} \\ 2 & \text{si la personne n'aime pas trop le roquefort ;} \\ 3 & \text{si la personne est indifférente au roquefort ;} \\ 4 & \text{si la personne aime un peu le roquefort ;} \\ 5 & \text{si l'individu aime beaucoup le roquefort.} \end{cases}$$

*REVENU* est catégorielle ordonnée : les niveaux de revenu sont triés par ordre croissant.

*EMPLOI* est catégorielle non ordonnée.

*ÉDUCATION* est catégorielle séquentielle : un niveau est conditionné par l'obtention du précédent diplôme.

*PRÉFÉRENCE* est aussi catégorielle ordonnée quoique un peu différente de *REVENU* : on a trié non pas par rapport à une valeur numérique, mais par rapport à une échelle dans les goûts.

Voir Maddala (1983) pour une présentation intéressante de ces différentes catégories.

Il existe deux utilisations de ces variables qualitatives, selon qu'elles sont *explicatives* (dans un modèle de régression par exemple) ou *dépendantes* (expliquées). Cette distinction conduit à une interprétation de ces variables comme correspondant à un *choix*, que l'on cherchera à expliquer, ou comme un facteur spécifique aux individus, entreprises, etc. Dans le premier cas les variables seront endogènes (engendrées par le modèle économique ou statistique), dans le second cas elles seront prédéterminées (au sens économétrique du terme).

## II. Variables indicatrices et régression

Les variables discrètes sont souvent utilisées dans les modèles de régression pour capter l'impact d'une caractéristique des entités observées sur l'espérance de la variable dépendante. On parle de *variables indicatrices* ou variables muettes (*dummy variables*).

Les mêmes précautions doivent être prises lorsqu'on incorpore des variables indicatrices dans une équation de régression, ainsi qu'avec tout modèle comportant des variables continues. Tout d'abord, à moins de travailler avec des variables centrées par rapport à leur moyenne sur l'échantillon, il importe d'incorporer un terme constant dans le modèle. L'omission de ce dernier revient en effet à spécifier un modèle « passant par l'origine », c'est-à-dire une relation affine, ce qui n'est pas toujours pertinent. De plus, l'inférence basée sur le coefficient de détermination linéaire  $R^2$  est faussée.

Il est également important, dans une régression avec terme constant, de veiller à ce que la somme des variables indicatrices soit différente de 1. Sinon, l'estimateur des *moindres carrés ordinaires* ne pourra être calculé, la matrice des variables explicatives contenant deux colonnes non linéairement indépendantes. La solution usuelle est d'omettre le terme constant, lequel sera reconstitué par les variables indicatrices, ou bien encore de négliger l'une de celles-ci. Concernant les variables polytomiques à plus de deux modalités, une erreur courante consiste à insérer une telle variable qualitative dans une équation de régression. Or, cela n'a que rarement un sens. Considérons par exemple la régression du salaire sur une constante et une variable polytomique représentant la catégorie socioprofessionnelle (CSP). Le coefficient associé à CSP indiquera le gain marginal en termes salariaux de passer d'une catégorie à celle immédiatement supérieure, et ce gain sera constant. Mais puisque CSP est discrète, rien n'empêche de modifier la définition de cette variable dans la base de données, en considérant par exemple un agencement mettant en avant non



plus la progression attendue des salaires, mais la nature du travail demandé (direction, manuel, etc). La régression du salaire sur CSP aboutirait alors à un coefficient estimé dont il serait difficile de fournir une interprétation aisée. La bonne solution, lorsque l'on envisage d'introduire une variable polytomique dans une régression, est de créer une variable indicatrice pour chaque modalité. Une possibilité, pour limiter le nombre d'indicatrices, est de conserver les plus significatives. Pour ce faire, il est utile de disposer d'une mesure de corrélation entre la variable dépendante continue et une variable discrète. Le résultat suivant donne la mesure de corrélation entre une variable continue et une variable polytomique.

### Définition

**Corrélation entre une variable continue et une variable polytomique** : pour un échantillon de taille  $N$ , la corrélation entre une variable continue  $z$  et une variable polytomique  $y$  admettant  $m$  modalités est calculée par :

$$\rho^2 = \frac{1/N \sum_{j=1}^m n_j (\bar{z}_j - \bar{z})^2}{s_z^2}$$

où  $n_j$  est l'effectif pour la modalité  $j$ ,  $\bar{z}_j$  la moyenne empirique de  $z$  pour la catégorie  $j$ ,  $\bar{z}$  la moyenne empirique de  $z$  sur l'échantillon, et  $s_z^2$  la variance de  $z$ .

On voit que  $\rho^2$  est nul si  $\bar{z}_1 = \bar{z}_2 = \dots = \bar{z}_m$ , c'est-à-dire si l'on a absence de dépendance en moyenne. Dans le cas opposé,  $\rho^2 = 1$  si tous les individus de la catégorie  $j$  ont la même valeur de  $z$ , quel que soit  $j$ . Dans le cas de deux catégories seulement, le coefficient de corrélation est :

$$\rho^2 = (s_z^2)^{-1} \frac{n_1 n_2}{N^2} (\bar{z}_1 - \bar{z}_2)^2$$

Pour tester l'absence de corrélation ( $\rho^2 = 0$ ), on fait l'hypothèse que la loi de  $z$  est normale dans chaque modalité. Sous l'hypothèse nulle que les  $z$  dans chaque modalité ont des moyennes et des variances identiques, on calcule la statistique de test suivante :

$$\frac{\rho^2 / (m - 1)}{(1 - \rho^2) / (N - m)} \sim F(m - 1, N - m)$$

qui suit une distribution de Fisher. En comparant la valeur calculée aux valeurs tabulées de la loi de Fisher à  $\nu_1 = m - 1$  et  $\nu_2 = N - m$  degrés de liberté, on rejettera l'hypothèse de non-corrélation si la statistique de test est supérieure à la valeur théorique pour un niveau de confiance donné (95 % par exemple).



Considérons un échantillon constitué de  $y$ , une variable continue, et d'une variable indicatrice notée  $x$ . Pour l'échantillon complet, l'estimateur de l'espérance de  $y_i$  est  $1/N \sum_{i=1}^N y_i$ , la moyenne arithmétique de la variable dépendante, qui coïncide avec l'estimateur par moindres carrés ordinaires (MCO) obtenu par régression de  $y_i$  sur une constante. Considérons le partage de cet échantillon en deux sous-échantillons de taille  $N_1$  et  $N_2$  respectivement, d'après la règle suivante : les  $N_1$  observations correspondent à  $x_i = 0$ , les  $N_2$  observations à  $x_i = 1$ . Il est clair que la moyenne de  $y_i$  sur ces deux sous-échantillons correspondra à l'espérance conditionnelle  $E(y_i|x_i = 0)$  et  $E(y_i|x_i = 1)$  respectivement. On a par exemple :

$$E(y_i|x_i = 1) = \frac{E(y_i x_i)}{\text{Prob}(x_i)} = \frac{E(y_i x_i)}{E(x_i)}$$

qui peut être estimé par :

$$E(y_i|x_i = 1) \approx \frac{1/N \sum_{i=1}^N y_i x_i}{1/N \sum_{i=1}^N x_i}$$

Considérons maintenant le modèle de régression suivant :

$$y_i = \beta_0 + x_i \beta_1 + \varepsilon_i$$

où  $\beta_0$  et  $\beta_1$  sont deux paramètres,  $\varepsilon_i$  un terme d'erreur de moyenne 0. Pour le premier sous-échantillon, on a  $x_i = 0$  et donc seule la valeur de  $\beta_0$  devrait intervenir dans l'espérance conditionnelle. Pour l'autre sous-échantillon, cette dernière est calculée à partir des deux paramètres. Les estimateurs par MCO de cette équation correspondent aux quantités recherchées :

$$E(\hat{\beta}_0) = E(y_i|x_i = 0)$$

et :

$$E(\hat{\beta}_0 + \hat{\beta}_1) = E(y_i|x_i = 1)$$

Il est important de garder à l'esprit cette notion de conditionnement : une variable indicatrice fait référence à une certaine catégorie, un sous-échantillon des données d'origine. L'espérance de la variable dépendante n'est pas l'espérance du produit des deux variables quand la variable indicatrice vaut 1. Il faut diviser ce produit par la proportion de l'échantillon telle que l'indicatrice vaut 1, c'est-à-dire par son espérance.

On peut bien sûr croiser des variables indicatrices entre elles et les utiliser dans une régression, de la même façon que pour des variables continues. Les considérations relatives à l'espérance conditionnelle de  $y$  (cf. plus haut) restent valables (McCullagh et Nelder, 1983).

### III. Application

Pour un échantillon de 3 177 ménages, on observe le revenu annuel  $REV$ , et une indicatrice  $SEX$  égale à 1 si le chef de famille est un homme, 0 s'il s'agit d'une femme. Les moyennes sur l'échantillon complet sont calculées avec les commandes GAUSS :

```
meanc(rev) ;meanc(sex ;meanc(rev.*sex) ;meanc((rev.*(1-sex)) ;
meanc((rev.*sex)/meanc(sex)) ; meanc((rev.*(1-sex))/meanc(1-sex)) ;
```

	Moyenne
$REV$	19 954,576
$SEX$	0,803 588 29
$REV \times SEX$	17 595,703
$REV \times (1 - SEX)$	2 358,873 2
$REV \times SEX / \overline{SEX}$	21 896,415
$REV \times (1 - SEX) / \overline{1 - SEX}$	12 009,840

Comparons maintenant avec la régression de  $REV$  sur  $SEX$  et une constante. Les résultats avec le logiciel GAUSS sont :

Valid cases : [1]	3177	Dependent variable : [7]	Y
Missing cases : [2]	0	Deletion method : [8]	None
Total SS : [3]	739534945367.892	Degrees of freedom : [9]	3175
R-squared : [4]	0.066	Rbar-squared : [10]	0.066
Residual SS : [5]	690522094969.865	Std error of est : [11]	14747.450
F(1,3175) : [6]	225.360	Probability of F : [12]	0.000

Variable	Estimate	Standard Error	t-value	Prob > t	Standardized Estimate
	[13]	[14]	[15]	[16]	[17]
CONSTANT	12009.839744	590.370475	20.342887	0.000	—
X1	9886.575454	658.578927	15.011983	0.000	0.257440

Les indications fournies par le logiciel sont :

[1] : nombre d'observations utilisées ; [2] : valeurs manquantes ; [3] : somme des carrés totale (*Sum of Squares*) ; [4] : coefficient  $R^2$  ; [5] : somme des carrés des résidus (*Residual Sum of Squares*) ; [6] : statistique de significativité globale de Fisher ; [7] : variable dépendante (expliquée) ; [8] : méthode de sélection des observations ; [9] : degrés de liberté ; [10] : coefficient de détermination ajusté  $\bar{R}^2$  ; [11] : écart type de la variable dépendante estimée ; [12] : valeur critique

ÉCO SUP



Alban Thomas

## ÉCONOMÉTRIE DES VARIABLES QUALITATIVES

MANUEL  
ET EXERCICES  
CORRIGÉS

Réputés complexes, les modèles à variables qualitatives sont en réalité de plus en plus utilisés parmi l'éventail des outils d'inférence statistique. Leurs applications se révèlent fort diverses, des études de marketing aux bilans commerciaux, en passant par le marché du travail.

Dès lors, cet ouvrage offre au lecteur une classification des variables qualitatives, avant d'en présenter la plupart des utilisations (modèles pour variables binaires, modèles multinomiaux, de régression...).

Il propose en outre :

- de nombreuses applications, sur données réelles, illustrant de manière claire les méthodes exposées et leur mise en œuvre avec des logiciels usuels (SAS, LIMDEP, GAUSS...);
- des échantillons que le lecteur peut utiliser pour vérifier les résultats et développer de nouveaux modèles à sa convenance;
- des exercices corrigés permettant de consolider les acquis au fur et à mesure de la lecture.

ALBAN THOMAS est directeur de recherche à l'INRA et enseigne à l'université des sciences sociales de Toulouse.

- Deuxième et troisième cycles de sciences économiques et de gestion
- IUP



ISBN 2 10 004665 9  
Code 044665



<http://www.dunod.com>



DUNOD

Participant d'une démarche de transmission de fictions ou de savoirs rendus difficiles d'accès par le temps, cette édition numérique redonne vie à une œuvre existant jusqu'alors uniquement sur un support imprimé, conformément à la loi n° 2012-287 du 1<sup>er</sup> mars 2012 relative à l'exploitation des Livres Indisponibles du XX<sup>e</sup> siècle.

Cette édition numérique a été réalisée à partir d'un support physique parfois ancien conservé au sein des collections de la Bibliothèque nationale de France, notamment au titre du dépôt légal. Elle peut donc reproduire, au-delà du texte lui-même, des éléments propres à l'exemplaire qui a servi à la numérisation.

Cette édition numérique a été fabriquée par la société FeniXX au format PDF.

La couverture reproduit celle du livre original conservé au sein des collections de la Bibliothèque nationale de France, notamment au titre du dépôt légal.

\*

La société FeniXX diffuse cette édition numérique en accord avec l'éditeur du livre original, qui dispose d'une licence exclusive confiée par la Sofia – Société Française des Intérêts des Auteurs de l'Écrit – dans le cadre de la loi n° 2012-287 du 1<sup>er</sup> mars 2012.

Avec le soutien du

