Institut National de la Statistique et des Etudes Economiques

# ÉCOLE NATIONALE DE LA STATISTIQUE ET DE L'ADMINISTRATION ÉCONOMIQUE

## MODELES A VARIABLES DEPENDANTES LIMITEES

(C. GOURIEROUX)

—

DIRECTION GÉNÉRALE


3, Avenue Pierre Larousse - 92240 MALAKOFF Téléphone : 540-12-12

# MODELES A VARIABLES DEPENDANTES LIMITEES

C. GOURIEROUX

—

(Modèles qualitatifs, modèles tobit, modèles à changements de régimes endogènes, modèles de Poisson).

980.0692.82.0

Novembre 1981

# 0

# INTRODUCTION

## 0.1 - HISTORIQUE

L'étude de modèles décrivant les modalités prises par une ou plusieurs variables qualitatives date des années 1940-1950 [Berkson (1944), (1951)]. Les premières applications ont essentiellement été menées dans le domaine de la biologie, puis de la sociologie et de la psychologie. Ce n'est que récemment [Mac - Fadden (1974)] que ces modèles ont été utilisés pour décrire des données économiques. Les applications à ce nouveau domaine ont permis le développement des modèles de type qualitatif dans deux directions principales :

- Il a souvent été possible de construire ces modèles à partir d'une théorie économique sous jacente des comportements individuels. Cette approche a permis de mieux comprendre la signification de certains modèles usuels comme le modèle logit [Mac-Fadden (1974)]. D'autre part, il est apparu que divers problèmes économiques (consommation de biens durables, analyse des déséquilibres...) conduisaient à des modèles, qui, s'ils n'étaient pas qualitatifs au sens strict du terme, en étaient mathématiquement proches [Tobin (1958), Fair-Jaffee (1972), Heckman (1976)...].

- Le deuxième apport des applications au domaine économique est l'introduction de variables exogènes. Les modèles sont donc principalement construits dans un but explicatif. Il est alors naturel de comparer ces modèles qualitatifs explicatifs au modèle linéaire habituel.

## 0.2 - QUELQUES RAPPELS SUR LES VARIABLES QUALITATIVES

### 0.2.a - Généralités

Les données statistiques disponibles sont souvent relatives à des caractères qualitatifs comme la catégorie socio-professionnelle, le type d'études suivies, le fait de travailler ou non, d'acheter ou ne pas acheter un certain produit...

Les méthodes d'inférence permettant de traiter de telles données diffèrent sensiblement de celles employées pour étudier des caractères quantitatifs, car elles doivent tenir compte de l'absence de continuité et souvent de l'absence d'ordre naturel entre les modalités que peut prendre le caractère qualitatif.

Dans la suite, nous supposerons que ce caractère y peut prendre K+1 modalités disjointes, notées (k), k=0,...,K. Si K+1=2 [resp. 3] la variable est dite dichotomique [resp : trichotomique] ; dans le cas général, où K est quelconque, elle est dite polytomique.

Lorsque le caractère y considéré est aléatoire, sa loi est définie par la donnée des probabilités que y prenne la modalité (k) ; ces probabilités sont notées $P_k$ k = 0,...,K.

### 0.2.b - Représentations quantitatives d'une variable qualitative

Il est toujours possible d'associer à un caractère qualitatif une variable quantitative (ou codage) apportant la même information. Nous allons examiner ce problème pour la variable qualitative : y = "catégorie socio-professionnelle" pouvant prendre les K+1=3 modalités :

(0) = ouvrier, (1) = employé, (2) = cadre

...

Exemple 0.1 : Définissons la variable quantitative $\tilde{y}$ par :

$$\begin{cases} \tilde{y} = 1 \ , & \text{si } y = \text{"ouvrier"} \ ; \\ \tilde{y} = 2 \ , & \text{si } y = \text{"employé "} \ ; \\ \tilde{y} = 3 \ , & \text{si } y = \text{"cadre"}. \end{cases}$$

La connaissance de la valeur prise par $\tilde{y}$ permet de savoir quelle est la modalité prise par y et inversement.

Exemple 0.2 : Considérons le vecteur $\varepsilon$ à trois composantes : $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)'$ défini par :

$$\varepsilon_1 = \mathbb{1}_{\{0\}}(y) = \begin{cases} 1, \text{ si } y = \text{"ouvrier"} \\ 0, \text{ sinon} \end{cases}$$

$$\varepsilon_2 = \mathbb{1}_{\{1\}}(y) = \begin{cases} 1, \text{ si } y = \text{"employé"} \\ 0, \text{ sinon} \end{cases}$$

$$\varepsilon_3 = \mathbb{1}_{\{2\}}(y) = \begin{cases} 1, \text{ si } y = \text{"cadre"} \\ 0, \text{ sinon} \end{cases}$$

Il s'agit d'une autre représentation quantitative de y à valeurs cette fois dans $\{0,1\}^3$. Remarquant que $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 = 1$, on voit d'ailleurs qu'un autre codage de y est donné par $(\varepsilon_1, \varepsilon_2)'$.

Exemple 0.3 : Il est facile maintenant de déterminer toutes les représentations quantitatives de y. Elles s'écrivent sous la forme

...

$\psi$ (y), où $\psi$ est une application injective de $\{①, ②, ③\}$ dans un espace $R^p$.

Les exemples précédents se généralisent immédiatement au cas d'une variable qualitative y à K+1 modalités. Ainsi le codage de l'exemple 0.1 serait maintenant: $\tilde{y} = k+1$ si $y = (k)$ et celui de l'exemple 0.2 : $\varepsilon = (\varepsilon_1 \ldots \varepsilon_{K+1})'$ avec $\varepsilon_k = \mathbb{1}_{(k+1)}(y)$. Ici encore $\sum_{k=0}^{K} \varepsilon_k = 1$.

L'intérêt principal d'une représentation quantitative est de pouvoir se ramener à des lois discrètes sur R ou $R^p$. Ainsi la loi de $\varepsilon$ est une loi multinomiale $\mathcal{M}(1 ; P_0, \ldots P_K)$, celle de $\varepsilon_1$ une loi de Bernoulli $\mathcal{B}(1, P_0)$. Il faut cependant utiliser avec prudence la loi d'une telle représentation ; les seules caractèristiques véritablement liées à la variable qualitative y sont celles qui ne dépendent pas de la représentation $\psi$ choisie, et ne sont autres que les valeurs $P_0, \ldots P_K$.

Exemple 0.5 : Les moments (moyenne, variance...) de la représentation $\psi$ (y) ont en général peu de sens. Remarquons cependant que dans le cas du codage $\varepsilon$, l'espérance permet de retrouver le vecteur des probabilités : $P = (P_0, \ldots, P_K)'$.

Exemple 0.6 : Considérons une autre variable x, quantitative, une technique usuelle pour voir, si x est liée à y, consiste à calculer le coefficient de corrélation. Or la valeur et le signe de ce coefficient $\rho [x, \psi (y)]$ dépendent du codage $\psi$ choisi.

Exemple 0.7 : Par contre, la notion d'indépendance peut être étudiée. En effet si $\psi$ et $\psi^*$ sont deux codages, si x et $\psi$ (y) sont indépendantes, x et $\psi^*$ (y) le sont aussi.

...

Exemple 0.8 : Plus importante, car elle justifie ce cours, est l'impossibilité d'effectuer une régression linéaire pour tous les codages. On ne peut généralement avoir simultanément

$$E(\psi(y) / x) = x\,b$$

et $$E(\psi^*(y) / x) = x\,c$$

0.2.C - Vecteur de variables qualitatives

Considérons Q variables qualitatives $y_q$, $q = 1,\ldots Q$ prenant respectivement $K_q + 1$ modalités $k_q$, $k_q = 0,\ldots K_q$. Le vecteur $y = (y_1,\ldots,y_q)'$ peut toujours être considéré comme une variable qualitative unique prenant les $\prod_{q=1}^{Q} (K_q + 1)$ modalités $(k_1,\ldots,k_Q)$. Les probabilités correspondantes seront notées $P_{k_1\ldots k_Q}$.

Inversement une variable qualitative unique peut être considérée comme un vecteur de variables qualitatives dichotomiques. Nous avons en effet remarqué que, si y est une variable à K + 1 modalités, une représentation de y est donnée par $\varepsilon = (\varepsilon_1,\ldots\varepsilon_K)'$ ; or $\varepsilon_k$ décrit le fait que y prenne ou non la modalité $k - 1$

Il n'y a donc fondamentalement aucune différence dans l'étude d'une variable qualitative ou de plusieurs variables qualitatives. Cependant la formulation vectorielle sera utile pour examiner les liaisons pouvant exister entre les variables et calculer les lois marginales et conditionnelles.

0.3 - PLAN DU COURS

Les chapitres sont écrits de façon à introduire progressivement l'aspect quantitatif dans les modèles. Les modèles où les variables endogènes sont qualitatives sont présentés dans les cinq premiers chapitres. On considère ensuite le cas où la variable endogène est parfois qualitative, parfois quantitative (chapitres 6 et 7). Les chapitres 8 à 10 sont consacrés à une étude

...

générale des modèles à changement de régimes endogène , où la variable est quantitative, mais à une expression dépendant d'un critère qualitatif.

Dans le dernier chapitre sont présentés des modèles permettant d'expliquer les valeurs prises par une variable discrète.

Les modèles les plus simples correspondent au cas où la variable qualitative endogène est dichotomique (chapitre I). L'étude de ce cas permet de bien comprendre les différences entre modèles qualitatifs et modèles quantitatifs, et permet de présenter de manière approfondie les principales méthodes d'estimation.

La phase de modélisation prend une importance particulière dès que l'on étudie des variables qualitatives à plus de deux modalités. Contrairement à ce qui se passe dans le cas dichotomique, les modèles peuvent en effet avoir des formes sensiblement différentes. La détermination d'une forme appropriée doit alors souvent s'appuyer sur des raisonnements de type économique. Des exemples de tels raisonnements et les modèles qui en résultent sont donnés au chapitre 2. Les méthodes pour estimer ces modèles et les propriétés des estimateurs obtenus sont décrites dans le chapitre 3.

Le chapitre 4 est consacré à l'utilisation descriptive des modèles qualitatifs. La formulation log-linéaire, qui y est présentée, se révèle adaptée à l'étude des problèmes d'indépendance.

L'explication de données spatio-temporelles qualitatives introduit une difficulté nouvelle : il faut pouvoir prendre en compte les corrélations éventuelles entre les observations. Nous regardons dans le chapitre 5 comment l'utilisation des chaines de Markov permet partiellement de répondre à cette question.

Dans les chapitres 6 et 7 sont étudiés les modèles où la variable dépendante est quantitative, mais contrainte à dépasser un certain seuil. Ce seuil peut être fixe (modèle tobit simple) ou aléatoire (modèle tobit généralisé).

Ces modèles présentent à la fois un aspect qualitatif, dans l'observation du fait que la variable touche ou non le seuil, et un aspect quantitatif. Ils ont une importance particulière pour la modélisation des phénomènes économiques et servent par exemple à décrire les consommations de biens durables ou les marchés en déséquilibre. Ce dernier type d'applications est étudié de manière détaillé dans le chapitre 8.

Il est possible d'inclure tous les modèles précédents, qualitatifs ou tobit, dans une formulation unique (chapitres 9 et 10). Une telle présentation n'a pas pour seul but d'unifier la théorie des modèles à variables dépendantes limitées. Elle permet en effet d'introduire des modèles comportant plusieurs variables endogènes limitées ou non, en particulier de prendre en compte des phénomènes de simultanéité ; d'autre part, elle fait apparaître certaines difficultés dans la construction de tels modèles, par exemple le problème de l'existence d'une forme réduite (cohérence).

Les modèles tobit peuvent être considérés comme intermédiaires entre les modèles qualitatifs et le modèle linéaire habituel. D'autres modèles intermédiaires sont obtenus en décrivant des variables à valeurs entières (chapitre 11).

L'ensemble du cours suppose connus les principaux résultats de statistique et d'économétrie. Ceux ci peuvent être trouvés respectivement dans :

Monfort (1981) : Théorie des Probabilités, Economica.
Malinvaud (1969) : Méthodes Statistiques de l'Econométrie, Dunod.

# I

# LE MODELE DICHOTOMIQUE SIMPLE

Dans cette section nous supposons que la variable endogène qualitative y est dichotomique. Les deux modalités qu'elle peut prendre sont par convention codées 0 et 1. Le modèle que nous obtiendrons est un cas particulier de ceux étudiés en II et III. Il est cependant intéressant d'en faire une présentation separée. Sa simplicité permet en effet de bien comprendre quelles sont les différences entre modèles qualitatifs et modèles quantitatifs, et permet aussi d'introduire de manière détaillée les méthodes d'estimation qui seront généralisées dans la suite.

## I.1 - POURQUOI NE PAS UTILISER UNE FORMULATION LINEAIRE ?

Une étude spécifique des modèles à variables endogènes qualitatives ne présente d'intérêt que si la formulation linéaire classique et les méthodes d'estimation correspondantes (moindres carrés ordinaires ou généralisés) ne sont pas adaptées au problème.

Supposons que nous disposions de n observations $y_i$, $i=1,\ldots,n$ de la variable endogène, faites lorsque les valeurs de K variables exogènes sont respectivement : $x_i = (x_i^1 \ldots x_i^K)$ $i = 1,\ldots,n$.
Le modèle linéaire s'écrirait :

$$y_i = x_i b + u_i \qquad i = 1,\ldots,n \tag{1.1}$$

où b serait un vecteur de K paramètres inconnus et où $u_i$ désignerait la perturbation associé à la $i^e$ observation.

...

L'inadéquation d'une telle formulation peut facilement être mise en évidence par des arguments intuitifs et par des arguments mathématiques. Donnons en quelques uns :

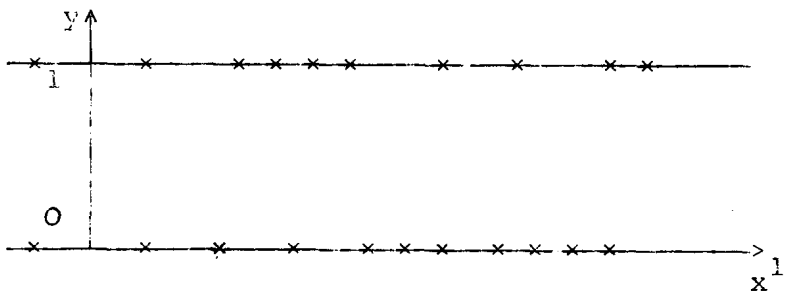
a) Les deux membres de l'égalité (1.1) sont de nature différente : $y_i$ est une variable qualitative et $x_i b + u_i$ une variable quantitative, ce qui a évidemment peu de sens.

b) On peut répondre à ceci que le membre de gauche est en fait la valeur du codage : 0 ou 1. Mais ce codage est évidemment arbitraire ; la valeur $b_o$ de b correspondant à ce codage est différente d'une valeur de b obtenue pour un autre codage. Elle serait par exemple de $2b_o$ si le codage était (0,2). Le paramètre b est donc non interprétable.

c) Une étude graphique des observations montre également que l'approximation linéaire est peu adaptée au problème. Considérons, pour pouvoir faire un dessin, le cas du modèle de régression simple :

$$y_i = b_o + x_i^1 b_1 + u_i \qquad i = 1, \ldots, n$$

et reportons les observations $(x_i^1, y_i)$ dans un système d'axes orthonormés. Le nuage des points observations, qui se trouve sur les deux droites parallèles $y = 0$ et $y = 1$ peut difficilement être bien approché par une seule droite



Ces arguments intuitifs suffiraient à rejeter la formulation linéaire ; ils sont cependant renforcés par certains problèmes mathématiques que poserait une écriture telle que (1.1).

d) Comme $y_i$ ne peut prendre que deux valeurs (0 et 1), il en est de même de la perturbation $u_i$ :

$u_i$ prend la valeur $1 - x_i b$ avec probabilité $p_i$

la valeur $- x_i b$ avec probabilité $1 - p_i$

La perturbation admet obligatoirement une loi discrète, ce qui interdit de faire l'hypothèse de normalité.

e) Si nous imposons aux perturbations d'être de moyenne nulle, $p_i$ est déterminé de manière unique, car :

$$Eu_i = p_i(1 - x_i b) - (1 - p_i)\, x_i b = 0$$

$$\Rightarrow p_i = x_i b$$

Les paramètres ne peuvent alors être quelconques puisque

$$0 \leq x_i b \leq 1 \qquad i = 1, \ldots, n$$

Ces contraintes peuvent être non compatibles ; dans ce cas le modèle

$$y_i = x_i b + u_i, \quad Eu_i = 0, \quad i = 1, \ldots, n$$

n'a pas de sens.

f) Si les contraintes sont compatibles, ceci crée au moins deux difficultés :

- le paramètre b doit être estimé sous contraintes à l'inégalité

- la prévision de y correspondant aux valeurs $x_{n+1}$ des variables explicatives ne peut être faite que si la contrainte $0 \leq x_{n+1}b \leq 1$ est une conséquence des contraintes $0 \leq x_i b \leq 1 \quad i = 1,\ldots, n$.

g) Remarquons finalement que la variance des perturbations vaut $Vu_i = x_i b(1-x_i b)$ ; il y a hétéroscedasticité ; la méthode des moindres carrés généralisés (contrainte ) n'est cependant pas applicable, puisque la matrice de variance - covariance des perturbations dépend du paramètre inconnu b figurant dans l'explication linéaire.

## I.2 - PRESENTATION DU MODELE DICHOTOMIQUE SIMPLE

Ces modèles ont été initialement utilisés pour les études biologiques, mais ont un champ d'application très vaste. Ils sont notamment employés pour déterminer la façon dont des individus (insectes, herbes, personnes) tolèrent un certain produit (insecticide, désherbant, médicament). Pour cela on effectue plusieurs expérience où des individus de caractéristiques différentes, placés dans des conditions différentes, sont soumis à diverses doses du produit. On observe à chaque fois si l'individu a ou non bien supporté l'expérience. Pour chacune des expériences i = 1,...,n la variable endogène observée $y_i$ est dichotomique :

$$y_i = \begin{cases} 0 & \text{si l'individu a bien supporté l'expérience} \\ 1 & \text{si l'individu a mal supporté l'expérience} \end{cases}$$

La modalité prise par y dépend des conditions $x_i$ dans lesquelles est réalisée l'expérience i et de la dose $\ell_i$ à laquelle l'individu est soumis. On introduit habituellement pour compléter ce modèle une variable quantitative auxiliaire : le seuil de tolérance $y_i^*$. $y_i^*$ est la dose maximale que peut supporter

...

l'individu au cours d'une expérience de type i. Cette variable dépend des conditions $x_i$ et peut être considérée comme aléatoire, deux individus de mêmes caractéristiques, placés sous les mêmes conditions, n'ayant pas forcément les mêmes réactions.

La variable qualitative observée est définie à partir de cette variable auxiliaire par :

$$(1.2) \qquad y_i = \begin{cases} 0 & \text{si } y_i^* > \ell_i \\ 1 & \text{si } y_i^* < \ell_i \end{cases}$$

Il reste à spécifier la façon dont le seuil de tolérance dépend des conditions de l'expérience. On utilise habituellement pour cela un modèle linéaire :

$$(1.3) \qquad y_i^* = x_i b + u_i \quad i = 1,\ldots,n$$

Les perturbations $u_i$ sont supposées indépendantes, de moyenne nulle et telles que les variables $\frac{u_i}{\sigma}$, où $\sigma$ est un paramètre positif, suivent une même loi de fonction de répartition F.

L'hypothèse d'indépendance des perturbations traduit certaines conditions que doit satisfaire l'expérience. Ainsi dans notre exemple, les observations doivent être faites sur des individus différents, sinon les résultats d'une expérience pourraient dépendre des résultats de la précédente.

Remarquons que la formulation (1.3) a bien un sens, les deux membres de l'égalité étant quantitatifs.

...

On déduit facilement de (1.2) et (1.3) la loi de y :

$$P[y_i = 1] = P[y_i^* < \ell_i] = P[x_i b + u_i < \ell_i]$$

$$= P\left[\frac{u_i}{\sigma} < \frac{\ell_i}{\sigma} - \frac{x_i b}{\sigma}\right] = F\left[\frac{\ell_i}{\sigma} - \frac{x_i b}{\sigma}\right] = p_i \text{ (par définition)}$$

Appliquant l'hypothèse d'indépendance, on obtient la vraisemblance :

$$L(y_1,\ldots,y_n) = \prod_{i=1}^{n} \left[p_i^{y_i}(1-p_i)^{1-y_i}\right]$$

$$= \prod_{i=1}^{n} \left\{F\left(\frac{\ell_i}{\sigma} - \frac{x_i b}{\sigma}\right)^{y_i}\left[1 - F\left(\frac{\ell_i}{\sigma} - \frac{x_i b}{\sigma}\right)\right]^{1-y_i}\right\}$$

Quitte à appeler différemment les variables exogènes : $z_i = (\ell_i, -x_i)$ et les paramètres $c = (\frac{1}{\sigma}, \frac{b}{\sigma})$, ce modèle est de la forme

$$(1.4) \qquad \boxed{L(y_1,\ldots,y_n) = \prod_{i=1}^{n} \{F(z_i c)^{y_i}[1-F(z_i c)]^{1-y_i}}$$

où F est la fonction de répartition d'une loi de moyenne nulle.

Dans la suite un tel modèle appelé <u>dichotomique simple</u>. La forme (2.5) est une conséquence de la forme retenue pour $p_i = P[y_i = 1]$ et de l'hypothèse d'indépendance des $y_i^*$. Il est évidemment possible de supprimer ces hypothèses et d'obtenir alors d'autres types de modèles pour décrire une variable qualitative dichotomique (voir II 4.a).

...