COLLECTION « ÉCONOMIE ET STATISTIQUES AVANCÉES »

esa

Série : Ecole Nationale de la Statistique et de l'Administration économique et Centre d'Etudes des Programmes Economiques

Christian GOURIEROUX

Professeur à l'ENSAE

ECONOMETRIE DES VARIABLES QUALITATIVES

2e édition

Economica

Dans la même série

P. ARTUS, M. DELEAU et P. MALGRANGE
Modélisation macroéconomique.

Denis BOSQ et Jean-Pierre LECOUTRE
Théorie de l'estimation fonctionnelle.

Christian GOURIEROUX
Théorie des sondages.
Econométrie des variables qualitatives (2ᵉ éd., juin 89).

Christian GOURIEROUX et Alain MONFORT
Statistique et modèles économétriques.
Vol. 1 - Notions générales.
Vol. 2 - Tests, régions de confiance.

Jean-Michel GRANDMONT
Monnaie et valeur.

Jean-Jacques LAFFONT
Fondements de l'économie publique - Vol. I, Cours de théorie microéconomique, 2ᵉ éd.
Economie de l'incertain et de l'information - Vol. II, Cours de théorie microéconomique.

Gilles MICHEL
L'économie française.

Alain MONFORT
Cours de probabilités, 2ᵉ éd.
Cours de statistique mathématique, 2ᵉ éd.

Pierre-Alain MUET
Théories et modèles de la macroéconomie.
Tome 1 : L'équilibre de courte période, 2ᵉ éd.

Jean-Pierre PATAT
Monnaie, institutions financières et politique monétaire, 4ᵉ éd.

J. PISANI-FERRY, H. STERDYNIAK et P. VILLA
Problèmes de macroéconomie.

Jean TIROLE
Concurrence imparfaite.

Michel VOLLE
Analyse des données, 3ᵉ éd.

COLLECTION « ÉCONOMIE ET STATISTIQUES AVANCÉES »
Série : Ecole Nationale de la Statistique et de l'Administration Economique et Centre d'Etudes des Programmes Economiques

Christian GOURIEROUX
Professeur à l'ENSAE

ECONOMETRIE DES VARIABLES QUALITATIVES

2e édition

ECONOMICA
49, rue Héricart, 75015 Paris

Avant-propos

Cette seconde édition du livre « Économétrie des Variables Qualitatives » comporte divers paragraphes nouveaux. Ils correspondent, soient à des méthodes apparues dans la littérature depuis 1985, soient à des méthodes existant antérieurement, mais dont l'intérêt pour les applications économiques s'est révélé depuis cette date.

Ainsi un chapitre [chapitre XI] a été consacré aux modèles de durée, qui se sont avérés utiles aussi bien pour la description du phénomène de chômage que des problèmes liés à l'attribution de crédit ou à la consommation de biens durables. Les modèles classiques de durée et les façons de les utiliser y sont complètement décrits, en insistant sur les liens existant avec d'autres modèles apparaissant dans le livre [Modèles Tobit, Modèles de Poisson]. Par ailleurs, l'accent a été mis systématiquement dans tous les chapitres sur les problèmes liés aux erreurs de spécification : description des principales procédures de test [chap. III], mesures de qualité d'ajustement [chap. III], construction de graphiques de résidus généralisés [chap. VI, VIII], tests d'exogénéïté [chap. VIII].

Finalement une introduction aux méthodes semi-paramétriques a été donnée dans le cas qualitatif [chap. I] et dans celui des modèles de durée [chap. XI]. Celles-ci ne sont pas encore d'un usage courant, mais devraient se développer rapidement dans les prochaines années.

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 [McFadden (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 [McFadden (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 $\boxed{k}$, $k = 0, ..., K$. 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 $\boxed{k}$; 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 :

$$\boxed{0} = \text{ouvrier}, \ \boxed{1} = \text{employé}, \ \boxed{2} = \text{cadre}.$$

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 ε à trois composantes : $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$, défini par :

$$\varepsilon_1 = \mathbf{1}_{\boxed{0}}(y) = \begin{cases} 1, \text{ si } y = \text{« ouvrier »}, \\ \\ 0, \text{ sinon}, \end{cases}$$

$$\varepsilon_2 = \mathbf{1}_{\boxed{1}}(y) = \begin{cases} 1, \text{ si } y = \text{« employé »}, \\ \\ 0, \text{ sinon}, \end{cases}$$

$$\varepsilon_3 = \mathbf{1}_{\boxed{2}}(y) = \begin{cases} 1, \text{ si } y = \text{« cadre »}, \\ \\ 0, \text{ sinon.} \end{cases}$$

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 ϕ (y), où ϕ est une application injective de $\{\boxed{0}, \boxed{1}, \boxed{2}\}$ 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 $y = \boxed{k}$ et celui de l'exemple 0.2 : $\varepsilon = (\varepsilon_1 \ldots \varepsilon_{K+1})'$ avec $\varepsilon_k = \mathbf{1}_{\boxed{k-1}}(y)$. Ici encore

$$\sum_{k=1}^{K+1} \varepsilon_k = 1.$$

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 ε est une loi multinomiale $\mathcal{M}$ (1 ; $P_O, \ldots P_K$), celle de ε_1 une loi de Bernoulli $\mathcal{B}$ (1, P_O). Il faut cependant utiliser avec prudence la loi d'une telle représentation ; les seules caractéristiques véritablement liées à la variable qualitative y sont celles qui ne dépendent pas de la représentation ϕ choisie, et ne sont autres que les valeurs $P_O, \ldots P_K$.

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

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 $\rho\,[x, \phi\,(y)]$ dépendent du codage ϕ choisi.

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

$$E\,(\phi\,(y)/x) = xb,$$

$$\text{et } E\,(\phi^*\,(y)/x) = xc.$$

0.2.c — Vecteur de variables qualitatives

Considérons Q variables qualitatives y_q, $q = 1, \ldots Q$ prenant respectivement $K_q + 1$ modalités $\boxed{k_q}$, $k_q = 0, \ldots K_q$. Le vecteur

$y = (y_1, ..., y_Q)'$ peut toujours être considéré comme une variable qualitative unique prenant les $\pi_{q=1}^{Q} (K_q + 1)$ modalités ($\boxed{k_1}$, ..., $\boxed{k_Q}$). Les probabilités correspondantes seront notées $P_{k_1 ... k_Q}$

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 $\varepsilon = (\varepsilon_1, ... \varepsilon_{K+1})'$, où ε_k décrit le fait que y prenne ou non la modalité $\boxed{k-1}$.

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 (chapitre 6). Les chapitres 7 à 9 sont consacrés à une étude générale des modèles à changement de régimes endogène, où la variable est quantitative, mais a une expression dépendant d'un critère qualitatif. Dans les deux derniers chapitres sont présentés des modèles permettant d'expliquer les valeurs prises par des variables positives discrètes ou continues.

Les modèles les plus simples correspondent au cas où la variable qualitative endogène est dichotomique (chapitre 1). 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 étude 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 chaînes de Markov permet partiellement de répondre à cette question.

Dans le chapitre 6 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 le 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 détaillée dans le chapitre 7.

Il est possible d'inclure tous les modèles précédents, qualitatifs ou tobit, dans une formulation unique (chapitres 8 et 9). 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 10).

Finalement la modélisation de variables discrètes est très liée à la modélisation de durées. Ce problème est étudié dans le 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.

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

Dans le chapitre 6 sont étudiés les modèles où la variable dépendante est quantitative, mais contrainte à appartenir à certains ensembles. Ceux-ci peuvent être des modèles tobit simples ou tobit généralisés. Ces modèles présentent à la fois un aspect qualitatif, dans la mesure où l'on s'intéresse à la variable ou non, 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 (7). Cette dernière application est étudiée de manière détaillée dans le chapitre 8.

Il est possible d'introduire les modèles précédents, qualitatifs ou tobit, dans une formulation unique (chapitres 8 et 9). 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 [illegible] des variables à valeurs entières (chapitre 10).

Finalement la modélisation de variables discrètes est très liée à la modélisation de durées. Ce problème est étudié dans le chapitre 11.

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

— Monfort (1982) : Théorie des Probabilités, Economica
— Malinvaud (1981) : Méthodes Statistiques de l'Économétrie, Dunod.

CHAPITRE I

Le modèle 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 séparée. 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 LINÉAIRE ?

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 linéaires ordinaires ou généralisés) ne sont pas adaptées au problème.

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

$$(1.1.) \qquad \boxed{y_i = x_i b + u_i \qquad i = 1, ..., n,}$$

où b serait un vecteur de K paramètres inconnus et où u_i désignerait la perturbation associée 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 :

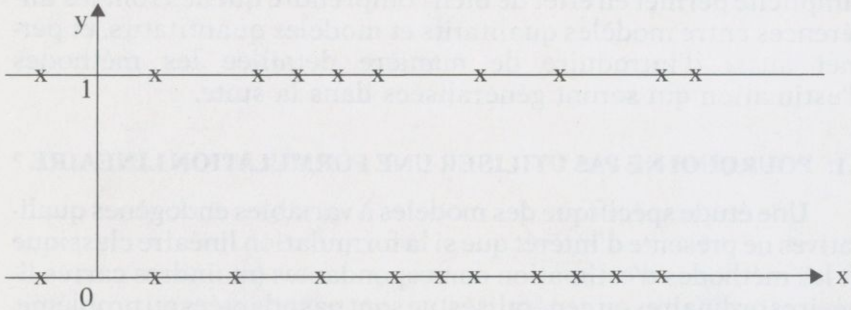
a) Les deux membres de l'égalité (1.1) sont de nature différente : y_i est une variable qualitative et $x_i b + u_i$ une variable quantitative, ce qui a évidemment peu de sens.

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 vableur b_0 de b correspondant à ce codage est différente d'une valeur de b obtenue pour un autre codage. Elle serait par exemple de $2b_0$ si le codage était (0,2). La valeur du 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 :

$$y_i = b_0 + x_i^1 b_1 + u_i \quad , \quad i = 1, ..., n \quad ,$$

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 la 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 déterminée de manière unique, car :

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

$$\Rightarrow p_i = x_i b \quad .$$

Les paramètres ne peuvent alors être quelconques, puisque :

$$0 \leqslant x_i b \leqslant 1, \qquad i = 1, ..., n \quad .$$

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

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

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 $0 \leqslant x_{n+1} b \leqslant 1$ est une conséquence des contraintes $0 \leqslant x_i b \leqslant 1$, $i = 1, ..., n$.

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, puis la matrice de variance-covariance des perturbations dépend du paramètre inconnu b figurant dans l'explication linéaire.

I.2. PRÉSENTATION DU MODÈLE 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 expériences 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 :

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

La modalité prise par y dépend des conditions x_i sous lesquelles est réalisée l'expérience i et de la dose ℓ_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 :

$$\boxed{y_i = \begin{cases} 0, \text{ si } y_i^* > \ell_i \ , \\ 1, \text{ si } y_i^* < \ell_i \ . \end{cases}} \tag{1.2}$$

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

$$\boxed{y_i^* = x_i b + u_i \qquad i = 1, ..., n \ .} \tag{1.3}$$

Les perturbations u_i sont supposées indépendantes, de moyenne nulle et telles que les variables $\frac{u_i}{\sigma}$, où σ 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'un 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 :

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

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

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

$$L\,(y\,;\,b, \sigma) = \prod_{i=1}^{n} \left[p_i^{y_i}\,(1 - p_i)^{1 - y_i}\right]$$

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

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

$$\boxed{L(y; c) = \prod_{i=1}^{n} \{F(z_i c)^{y_i} [1 - F(z_i c)]^{1-y_i}\} \quad ,} \tag{1.4}$$

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

Dans la suite un tel modèle est appelé **dichotomique simple**. La forme (1.4) 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).

I.3. EXEMPLES

Ces modèles ont de nombreuses applications. Nous en donnons ci-dessous deux exemples:

1.3.a — Choix des collèges

Parmi les premières études utilisant les modèles à réponses qualitatives, plusieurs s'intéressaient aux comportements des étudiants et en particulier aux motivations qui leur faisaient choisir tel établissement d'enseignement supérieur plutôt que tel autre. Les études de ce type ont essentiellement été développées aux Etats-Unis du fait de l'organisation particulière de l'enseignement supérieur (voir Radner-Miller [1970], Kohn-Manski-Mundel [1976]...).

Les établissements de ce pays peuvent être classés en deux catégories principales: ceux situés dans les villes et ceux munis d'un campus.

Le choix par un étudiant d'un établissement de l'une de ces catégories dépend de plusieurs facteurs tels: la distance de l'établissement au lieu de domicile habituel, le revenu, le sexe de l'étudiant, le fait qu'il préfère vivre sur un campus.

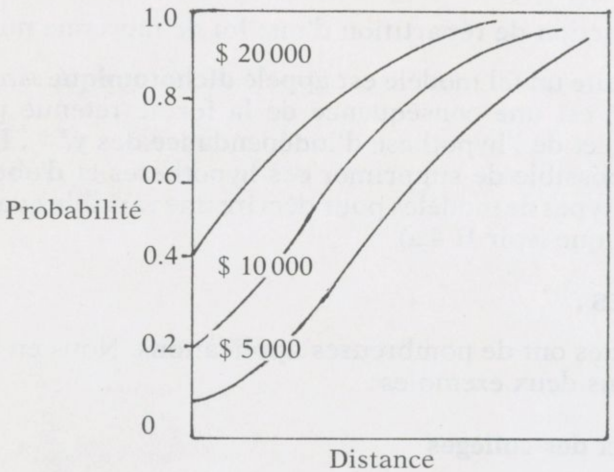
La probabilité de choisir de résider sur un campus peut être modélisée sous la forme:

$$P(y_i = 1) = F(x_i b) \quad ,$$

où x_i est l'ensemble des caractéristiques de l'étudiant i.

L'estimation d'un tel modèle à partir de résultats d'enquête montre (voir figure) que, toutes choses égales par ailleurs, cette probabilité croît avec la distance et croît avec le revenu.

Probabilité de vivre sur le campus (pour les garçons, préférant vivre sur un campus)



On constate également que cette probabilité est plus grande pour ceux qui préfèrent ce type de résidence ; il n'est évidemment pas surprenant que les choix effectués reflètent les préférences pures des étudiants, cependant ce facteur apparaît secondaire par rapport aux revenus et à la distance.

La probabilité de résider sur un campus est d'autre part plus grande pour les garçons que pour les filles.

1.3.b — Non-réponses

Il existe toujours lors d'une enquête une certaine proportion de non-réponses qui peuvent être dues à l'absence de l'enquêté de son domicile (lors d'enquêtes par interview), au refus de l'enquêté de répondre, à la non-compréhension des questions. Le taux de non-réponse dépend de la méthode de collecte employée, mais aussi des caractéristiques des enquêteurs (âge, niveau d'éducation, sexe) et de celles des répondants. Savoir quelle est la forme de cette dépendance pour un certain type de questions est évidemment important, si on souhaite diminuer le taux de non-réponse.

Ce problème revient à décrire la variable dichotomique « fait de ne pas répondre » en fonction d'autres variables et peut être résolu au moyen d'un modèle de la forme (1.4).

I.4. ESTIMATION PAR LA MÉTHODE DU MAXIMUM DE VRAISEMBLANCE

1.4.a — Le modèle

Une fois mis en évidence les paramètres identifiables, nous avons vu que le modèle dichotomique simple correspond à une vraisemblance de la forme :

$$\boxed{L\,(y\,;\,b) = \prod_{i=1}^{n} \{F\,(x_i b)^{y_i}\,[1 - F\,(x_i b)]^{1-y_i}\}} \,, \tag{1.5}$$

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

Il reste pour compléter le modèle à choisir la forme de F. On retient habituellement pour cette distribution soit une loi normale centrée réduite (le modèle est alors appelé modèle **probit**), soit une loi logistique (modèle **logit**).

Dans le premier cas, on a :

$$F\,(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\Pi}}\, e^{-\frac{t^2}{2}}\, dt = \Phi\,(x) \text{ (par définition)} \quad ,$$

et dans le second :

$$F\,(x) = \frac{1}{1 + e^{-x}} \quad .$$

Ces deux distributions étant symétriques ($F(-x) = 1 - F(x)$), la moyenne est bien égale à 0.

1.4.b — Unicité du maximum de vraisemblance

Le logarithme de la vraisemblance est donné par :

$$\begin{aligned}\text{Log}\, L &= \sum_{i=1}^{n} \{y_i \,\text{Log}\, F\,(x_i b) + (1 - y_i)\,\text{Log}\,[1 - F\,(x_i b)]\} \\ &= \sum_{i:y_i=1} \text{Log}\, F\,(x_i b) + \sum_{i:y_i=0} \text{Log}\,[1 - F\,(x_i b)] \quad .\end{aligned}$$

Une condition suffisante pour que le maximum global de Log L soit unique (s'il existe) est que cette fonction soit strictement concave, ou de façon équivalente que Log F et Log (1 − F) soient strictement concaves.

Cette condition est en particulier satisfaite pour les modèles probit et logit. Vérifions-le pour le modèle logit.

On a :

$$\frac{d}{dx}[\text{Log } F(x)] = \frac{1}{F(x)}\frac{d}{dx}F(x) = \frac{e^{-x}}{1+e^{-x}} = 1 - F(x) \ ,$$

$$\Rightarrow \frac{d^2}{dx^2}[\text{Log } F(x)] = \frac{-e^{-x}}{(1+e^{-x})^2} = -F(x)[1-F(x)] < 0 \ .$$

Log F est donc strictement concave.

Comme $\text{Log}[1 - F(x)] = \text{Log}\,\frac{e^{-x}}{1+e^{-x}} = -x + \text{Log } F(x)$, cette fonction est aussi strictement concave.

Le cas de la loi normale est laissé en exercice au lecteur (voir exercice 4).

Dans les cas usuels, il suffit donc d'écrire les conditions du premier ordre pour obtenir l'estimateur du maximum de vraisemblance de b.

1.4.c — Equation de vraisemblance

Dérivons la log-vraisemblance par rapport au vecteur b des paramètres et annulons cette dérivée ; on a :

$$\frac{\partial \text{ Log } L}{\partial b} = 0$$

$$\Leftrightarrow (1.6) \quad \boxed{\sum_{i:y_i=1} \frac{f(x_i b)}{F(x_i b)} x_i' - \sum_{i:y_i=0} \frac{f(x_i b)}{1-F(x_i b)} x_i' = 0} \ ,$$

où f est la densité associée à F et où x_i' désigne le transposé du vecteur x_i.

$$\Leftrightarrow \quad \sum_{i=1}^{n}\left[\frac{y_i}{F(x_i b)} - \frac{1-y_i}{1-F(x_i b)}\right] f(x_i b)\, x_i' = 0$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} \frac{y_i - F(x_i b)}{F(x_i b)[1-F(x_i b)]} f(x_i b)\, x_i' = 0 \ .$$

Dans le cas du modèle logit, cette dernière relation se simplifie car :

$$f(x) = F(x)\,[1 - F(x)] = \frac{e^{-x}}{(1 + e^{-x})^2}\ ;$$

on obtient :

$$\sum_{i=1}^{n} [y_i - F(x_i b)]\, x_i' = 0$$

$$\Leftrightarrow \quad (1.7) \qquad \boxed{\sum_{i=1}^{n} y_i\, x_i' = \sum_{i=1}^{n} F(x_i b)\, x_i'}\ .$$

Les équations de vraisemblance associées aux modèles dichotomiques simples, probit ou logit, sont non linéaires dans les paramètres ; il n'est pas possible d'exprimer les estimateurs comme fonctions simples des observations et les équations devront être résolues au moyen d'algorithmes (voir 1.5).

1.4.d — Existence et propriétés asymptotiques des estimateurs du maximum de vraisemblance

Nous venons de montrer que la solution des équations de vraisemblance, si elle existait, était unique pour les modèles probit et logit et correspondait bien au maximum de la fonction Log L. En fait cette solution n'existe pas toujours. Considérons, par exemple, le cas d'un modèle logit comportant une seule variable explicative à valeurs positives ; l'équation (1.7) s'écrit :

$$\sum_{i=1}^{n} y_i\, x_i = \sum_{i=1}^{n} F(x_i b)\, x_i\ .$$

Le premier membre prend ses valeurs entre 0 (si tous les y_i valent 0) et $\sum_{i=1}^{n} x_i$ (si tous les y_i valent 1). Lorsque b varie de $-\infty$ à $+\infty$, le second membre prend toutes les valeurs de $]0, \sum_{i=1}^{n} x_i[$. La solution de l'équation de vraisemblance existe donc, si et seulement si certaines observations de y correspondent à la valeur 0, d'autres à la valeur 1.

Même dans le cas où les équations de vraisemblance admettent une solution, celle-ci ne possède pas forcément de bonnes propriétés asymptotiques. La convergence et la normalité asymptotique ne seront assurées que si certaines conditions sont imposées à la façon dont évoluent les variables explicatives. Deux approches sont envisageables.

Supposer que les variables explicatives sont des variables aléatoires ; les conditions sont alors du type : « les x_i sont indépendantes, de même loi, admettant des moments d'ordre suffisant » [Amemiya (1976), McFadden (1974)] ou supposer que les valeurs des variables explicatives sont déterministes ; les conditions imposent alors à ces valeurs d'être bornées : $\exists\ m > 0$ et $M < \infty$ tels que $m < |x_i^k| < M \quad \forall i, k$ et d'être telles que la matrice de variance-covariance asymptotique existe [Gouriéroux-Monfort (1981)].

Sous ces conditions que nous supposons satisfaites dans la suite, l'estimateur du maximum de vraisemblance $\hat{b}$ existe asymptotiquement, converge vers la vraie valeur du paramètre et suit asymptotiquement une loi normale, de moyenne la vraie valeur b, de matrice de variance-covariance l'inverse de la matrice d'information de Fisher :

$$I^{-1} = \left[- E \frac{\partial^2 \text{Log } L}{\partial b\, \partial b'}\right]^{-1} ,$$

où l'espérance est relative à la loi conditionnelle de y sachant x.

$$\hat{b} \approx N \left\{b, \left[- E \frac{\partial^2 \text{Log } L}{\partial b\, \partial b'}\right]^{-1}\right\}$$

(Cette notation qui n'est pas parfaitement exacte du point de vue mathématique signifie que :

$$\left[- E \frac{\partial^2 \text{Log } L}{\partial b\, \partial b'}\right]^{1/2} (\hat{b} - b) ,$$

converge en loi vers une variable aléatoire admettant une loi normale centrée réduite).

En pratique, la matrice de variance-covariance asymptotique de $\hat{b}$, qui dépend du paramètre inconnu devra être estimée. Nous le ferons en remplaçant dans son expression b par $\hat{b}$.

$$\boxed{\hat{V}_{as}\, \hat{b} = \left[- E \frac{\partial^2 \text{Log } L}{\partial b\, \partial b'}\right]^{-1}_{b = \hat{b}} .} \tag{1.8}$$

1.4.e — Calcul de la matrice de variance-covariance asymptotique

La dérivée seconde de la log-vraisemblance est donnée par :

$$\frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'} = \frac{\partial}{\partial b}\left[\frac{\partial \operatorname{Log} L}{\partial b}\right]'$$

$$= \sum_{i : y_i = 1} \frac{\Delta f (x_i b)\, F (x_i b) - f^2 (x_i b)}{F^2 (x_i b)}\, x_i'\, x_i$$

$$- \sum_{i : y_i = 0} \frac{\Delta f (x_i b)\, [1 - F (x_i b)] + f^2 (x_i b)}{[1 - F (x_i b)]^2}\, x_i'\, x_i \quad ,$$

où Δ f désigne la dérivée de la fonction f.

$$\frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'} = \sum_{i=1}^{n} \left\{\frac{\Delta f (x_i b)\, F (x_i b) - f^2 (x_i b)}{F^2 (x_i b)}\, x_i'\, x_i\, y_i\right.$$

$$\left. - \frac{\Delta f (x_i b)\, [1 - F (x_i b)] + f^2 (x_i b)}{[1 - F (x_i b)]^2}\, x_i'\, x_i\, (1 - y_i)\right\} \quad .$$

Prenons l'espérance de l'opposé de cette dérivée seconde en remarquant que E y_i = F $(x_i b)$; la matrice d'information de Fisher est donnée par :

$$(1.9) \qquad \boxed{I = E\left[\frac{- \partial^2 \operatorname{Log} L}{\partial b\, \partial b'}\right] = \sum_{i=1}^{n} \frac{f^2 (x_i b)}{F (x_i b)\, [1 - F (x_i b)]}\, x_i'\, x_i}$$

La matrice de variance-covariance asymptotique de l'estimateur $\hat{b}$ est donc :

$$V_{as}\, \hat{b} = I^{-1} = \left\{\sum_{i=1}^{n} \frac{f^2 (x_i b)}{F (x_i b)\, [1 - F (x_i b)]}\, x_i'\, x_i\right\}^{-1} \quad .$$

Elle est estimée par :

$$\hat{V}_{as}\, \hat{b} = \left[\sum_{i=1}^{n} \frac{f^2 (x_i \hat{b})}{F (x_i \hat{b})\, [1 - F (x_i \hat{b})]}\, x_i'\, x_i\right]^{-1} \quad .$$

Remarque 1.10

Dans le cas particulier du modèle logit, on a :

$$f = F\, (1 - F) \quad ,$$

et donc :

$$E\left[\frac{- \partial^2 \operatorname{Log} L}{\partial b\, \partial b'}\right] = \sum_{i=1}^{n} F (x_i b)\, [1 - F (x_i b)]\, x_i'\, x_i \quad .$$

Remarque 1.11

On peut donner de la matrice des dérivées secondes une expression plus compacte. Notons X la matrice des observations des variables explicatives ; X est de taille (n, K) et a pour i^e ligne x_i. Appelons d'autre part Λ, la matrice diagonale de taille n, dont le i^e élément de la diagonale est :

$$\frac{f^2(x_i b)}{F(x_i b)\,[1 - F(x_i b)]} \quad .$$

On a :

$$E\left[\frac{-\partial^2 \operatorname{Log} L}{\partial b\, \partial b'}\right] = X' \Lambda X$$

$$\Rightarrow V_{as}\, \hat{b} = (X' \Lambda X)^{-1} \quad .$$

Cette forme de matrice de covariance est semblable à celle d'un estimateur des moindres carrés généralisés.

I.5. RÉSOLUTION NUMÉRIQUE DES ÉQUATIONS DE VRAISEMBLANCE

Les méthodes classiques de résolution numérique des équations de vraisemblance sont toutes fondées sur la méthode de Newton. Son application directe conduit à l'algorithme de Newton-Raphson.

1.5.a — Algorithme de Newton-Raphson

Le but de l'algorithme est de trouver une racine de l'équation $\frac{\partial \operatorname{Log} L}{\partial b} = 0$. Pour cela, on se donne une valeur initiale $\hat{b}_0$ et on cherche le plan tangent en ce point à la fonction : $b \rightarrow d = \frac{\partial \operatorname{Log} L}{\partial b}$. Ce plan, d'équation :

$$d = \frac{\partial \operatorname{Log} L}{\partial b}(\hat{b}_0) + \frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'}(\hat{b}_0)\,[b - \hat{b}_0] \quad ,$$

constitue une approximation de $d = \frac{\partial \operatorname{Log} L}{\partial b}$; d'où l'idée d'approcher la racine de : $\frac{\partial \operatorname{Log} L}{\partial b} = 0$ par la racine de :

$$\frac{\partial \operatorname{Log} L}{\partial b}(\hat{b}_0) + \frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'}(\hat{b}_0)\,[b - \hat{b}_0] = 0 \quad ,$$

c'est-à-dire par :

$$\tilde{b}_1 = \tilde{b}_0 - \left[\frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'} (\tilde{b}_0)\right]^{-1} \frac{\partial \operatorname{Log} L}{\partial b} (\tilde{b}_0) \quad .$$

On recommence alors la démarche en prenant $\tilde{b}_1$ comme valeur initiale et ainsi de suite. La formule de récurrence permettant de calculer $\tilde{b}_{h+1}$ en fonction de $\tilde{b}_h$ est :

$$\boxed{\tilde{b}_{h+1} = \tilde{b}_h - \left[\frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'} (\tilde{b}_h)\right]^{-1} \frac{\partial \operatorname{Log} L}{\partial b} (\tilde{b}_h)} \quad . \tag{1.11}$$

Si la suite des valeurs $\tilde{b}_h$ converge vers une limite $\tilde{b}$, celle-ci est forcément solution des équations de vraisemblance car :

$$\tilde{b} = \lim_{h \to \infty} \tilde{b}_{h+1} = \tilde{b} - \left[\frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'} (\tilde{b})\right]^{-1} \frac{\partial \operatorname{Log} L}{\partial b} (\tilde{b})$$

$$\Rightarrow \frac{\partial \operatorname{Log} L}{\partial b} (\tilde{b}) = 0.$$

1.5.b — La méthode du score

La méthode du score consiste à remplacer dans (1.11),

$$\frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'} (\tilde{b}_h) \quad ,$$

par son espérance conditionnelle à x. Ceci peut être justifié, lorsque le modèle possède de bonnes propriétés asymptotiques, en particulier dans le cas de l'échantillonnage. La formule de récurrence est alors :

$$\boxed{\tilde{b}_{h+1} = \tilde{b}_h + \left[E - \frac{\partial^2 \operatorname{Log} L\, (\tilde{b}_h)}{\partial b\, \partial b'}\right]^{-1} \frac{\partial \operatorname{Log} L}{\partial b} (\tilde{b}_h)} \tag{1.12}$$

Dans le cas particulier du modèle dichotomique simple, cette formule de récurrence peut être retrouvée de manière différente.

Comme $Ey_i = F(x_i b)$ et $Vy_i = F(x_i b)\,[1 - F(x_i b)]$, on peut écrire y_i sous la forme :

$$y_i = F(x_i b) + v_i \quad ,$$

$$\text{avec} \qquad Ev_i = 0 \quad , \qquad Vv_i = F(x_i b)\,[1 - F(x_i b)] \quad .$$

Supposons que nous sachions a priori que la vraie valeur de b est proche d'une valeur connue $\tilde{b}_h$, il est naturel de faire un développement limité de F $(x_i b)$ autour de $\tilde{b}_h$. On a alors :

$$y_i \;\#\; F(x_i \tilde{b}_h) + f(x_i \tilde{b}_h)\, x_i (b - \tilde{b}_h) + v_i$$

$$\Leftrightarrow y_i - F(x_i \tilde{b}_h) + f(x_i \tilde{b}_h)\, x_i \tilde{b}_h \;\#\; f(x_i \tilde{b}_h)\, x_i b + v_i \; .$$

Sous cette forme le modèle est un modèle linéaire où la i[e] observation de la variable expliquée est : $y_i - F(x_i \tilde{b}_h) + f(x_i \tilde{b}_h)\, x_i \tilde{b}_h$, la i[e] observation des variables explicatives $f(x_i \tilde{b}_h)\, x_i$. La matrice de variance-covariance des perturbations v_i dépendant de b pourra être approchée par la même matrice, où b aura été remplacé par $\tilde{b}_h$. L'estimateur des moindres carrés quasi-généralisés de b est alors donné par :

$$\left[\sum_{i=1}^{n} \frac{x_i' f^2(x_i \tilde{b}_h)\, x_i}{F(x_i \tilde{b}_h)\,[1 - F(x_i \tilde{b}_h)]}\right]^{-1}$$

$$\left[\sum_{i=1}^{n} \frac{\{y_i - F(x_i \tilde{b}_h) + f(x_i \tilde{b}_h)\, x_i \tilde{b}_h\}\, f(x_i \tilde{b}_h)\, x_i'}{F(x_i \tilde{b}_h)\,[1 - F(x_i \tilde{b}_h)]}\right]$$

$$= \left[E - [\frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'}(\tilde{b}_h)]\right]^{-1} \left[\frac{\partial \operatorname{Log} L}{\partial b}(\tilde{b}_h) + E\,[-\frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'}(\tilde{b}_h)]\, \tilde{b}_h\right]$$

$$= \tilde{b}_h + \left[E\,[-\frac{\partial^2 \operatorname{Log} L}{\partial b\, \partial b'}(\tilde{b}_h)]\right]^{-1} \frac{\partial \operatorname{Log} L}{\partial b}(\tilde{b}_h) = \tilde{b}_{h+1} \; .$$

La formule permettant après linéarisation du modèle d'obtenir l'estimateur des moindres carrés quasi-généralisés n'est donc autre que la formule de récurrence correspondant à la méthode du score.

1.5.c — La méthode de Berndt-Hall-Hall-Hausman

Pour passer de l'algorithme de Newton-Raphson à la méthode du score, il suffit de remplacer :

$$\frac{-\,\partial^2 \operatorname{Log} L}{\partial b\, \partial b'} \text{ par } E\left[\frac{-\,\partial^2 \operatorname{Log} L}{\partial b\, \partial b'}\right] \; .$$

Mais cette dernière matrice peut également s'écrire :

$$E\left[\sum_{i=1}^{n} - \frac{\partial^2 \operatorname{Log} L_i}{\partial b\, \partial b'}\right] = \sum_{i=1}^{n} E\left[\frac{\partial \operatorname{Log} L_i}{\partial b}\, \frac{\partial \operatorname{Log} L_i}{\partial b'}\right] \; ,$$

où L_i désigne la vraisemblance associée à la i[e] observation.

Enlevant le signe espérance, on obtient un nouvel algorithme, qui présente l'avantage de ne faire intervenir que les dérivées premières. La formule de récurrence est donnée par :

$$\text{(1.13)} \quad \boxed{\hat{b}_{h+1} = \hat{b}_h + \left[\sum_{i=1}^{n} \frac{\partial \operatorname{Log} L_i}{\partial b}(\hat{b}_h) \frac{\partial \operatorname{Log} L_i}{\partial b'}(\hat{b}_h)\right]^{-1} \frac{\partial \operatorname{Log} L}{\partial b}(\hat{b}_h)} \; .$$

Pour assurer la convergence de ces divers algorithmes et réduire le délai de convergence, il est recommandé d'utiliser une valeur initiale b_0 proche de la vraie valeur du paramètre. Il faut donc disposer d'une méthode d'estimation permettant de trouver un tel b_0 et qui soit plus simple que la méthode du maximum vraisemblance. Ce sera le but de la section suivante.

Signalons également que lorsqu'on utilise directement les formules d'itération précédentes, la convergence des algorithmes peut être assez lente ; il est possible d'améliorer celle-ci en choisissant à chaque étape une bonne direction. Ainsi prenons le cas du dernier algorithme ; on introduit à la h^e itération un coefficient λ_h $(0 \leq \lambda_h < 1)$ et on définit :

$$\hat{b}_{h+1} = \hat{b}_h + \lambda_h \left[\sum_{i=1}^{n} \frac{\partial \operatorname{Log} L_i}{\partial b}(\hat{b}_h) \frac{\partial \operatorname{Log} L_i}{\partial b'}(\hat{b}_h)\right]^{-1} \frac{\partial \operatorname{Log} L}{\partial b}(\hat{b}_h) \; .$$

Le coefficient λ_h est choisi par balayage de façon à rendre la plus grande possible la log-vraisemblance calculée en $\hat{b}_{h+1}$.

I.6. OBSERVATIONS RÉPÉTÉES

1.6.a — Formalisation

Lorsque les modèles qualitatifs servent à décrire une expérience contrôlée, on peut souvent admettre qu'à chaque valeur des caractéristiques exogènes correspondent plusieurs observations du caractère qualitatif. Ceci traduit la possibilité de répéter plusieurs fois l'expérience sous les mêmes conditions. Ce type d'approche peut aussi être utilisé pour décrire le comportement qualitatif d'agents (par exemple des ménages), pouvant être regroupés en classes homogènes (classement des ménages par taille, lieu de domicile, tranche d'âge, revenu du chef de ménage...).

Appelons x_j, $j = 1 \ldots J$, les J valeurs possibles des K variables explicatives et supposons que l'on effectue n_j expériences sous les conditions x_j, indépendamment les unes des autres ; les valeurs observées de la variable dichotomique sont notées y_{ij}, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$.

Dans ce cas la vraisemblance (I.6) a la forme suivante :

$$L(y\,;b) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} \{F(x_j b)^{y_{ij}} [1 - F(x_j b)]^{1-y_{ij}}\}$$

$$= \prod_{j=1}^{J} \{F(x_j b)^{\sum_{i=1}^{n_j} y_{ij}} [1 - F(x_j b)]^{n_j - \sum_{i=1}^{n_j} y_{ij}}\} \quad .$$

Appelons $\hat{p}_j$ la proportion d'observations pour lesquelles la variable dichotomique prend la valeur 1 dans une expérience de type j ; on a :

$$\hat{p}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \,,$$

et remplaçant dans l'expression de la vraisemblance, on obtient :

$$L(y\,;b) = \prod_{j=1}^{J} \{F(x_j b)^{n_j \hat{p}_j} [1 - F(x_j b)]^{n_j(1-\hat{p}_j)}\} \quad .$$

On remarque que l'ensemble des fréquences observées $\hat{p}_j$, $j = 1, ..., J$, constitue une statistique exhaustive. Il suffit donc de raisonner directement sur la loi de cette statistique.

Les variables $\hat{p}_j$, $j = 1, ..., J$ sont indépendantes et sont telles que $n_j\,\hat{p}_j$ suive une loi binomiale de paramètre n_j, $p_j = F(x_j b)$.

$$L(\hat{p}\,;b) = \prod_{j=1}^{J} \{C_{n_j}^{n_j \hat{p}_j} F(x_j b)^{n_j \hat{p}_j} [1 - F(x_j b)]^{n_j(1-\hat{p}_j)}\} \quad .$$

1.6.b — Méthode de Berkson

Nous allons approcher le modèle qualitatif par un modèle linéaire en supposant que chaque expérience est répétée un grand nombre de fois, c'est-à-dire en supposant que les n_j sont suffisamment grands et ceci quel que soit $j = 1 ... J$ (le nombre J d'expériences est lui considéré comme fixe). Il s'agit donc d'une approche asymptotique différente de celle étudiée en 1.4.d., où l'on supposait que seul le nombre total d'observations :

$$n = \sum_{j=1}^{J} n_j \text{ était grand.}$$

$\hat{p}_j = \dfrac{1}{n_j} \sum\limits_{i=1}^{n_j} y_{ij}$ est une moyenne empirique de variables aléatoires réelles indépendantes de même loi. Lorsque n_j est grand, $\hat{p}_j$ est proche de $E\,y_{ij} = F(x_j b)$ (loi des grands nombres) et la variable

$\sqrt{n_j}\,(\hat{p}_j - p_j)$ suit approximativement une loi normale de moyenne nulle et de variance :

$V(y_{ij}) = F(x_j b)\,[1 - F(x_j b)] = p_j(1 - p_j)$ (Théorème central limite).

On peut donc écrire de manière approchée :

$$\hat{p}_j = F(x_j b) + v_j \quad , \quad j = 1 \ldots J \quad ,$$

où les perturbations v_j sont indépendantes et suivent des lois normales de moyenne nulle, de variances :

$$\frac{p_j(1 - p_j)}{n_j} \; .$$

Le modèle précédent n'est cependant pas linéaire dans le paramètre b et il est nécessaire de le transformer. Pour cela examinons les propriétés de la variable $F^{-1}(\hat{p}_j)$, lorsque le nombre d'observations n_j est grand.

D'après le théorème de Slutsky (voir Monfort [1981] : théorie des probabilités, p. 166, théorème 5), $F^{-1}(\hat{p}_j)$ est proche de :

$$F^{-1}(p_j) = F^{-1}F(x_j b) = x_j b \quad ,$$

et la variable :

$$\sqrt{n_j}\,[F^{-1}(\hat{p}_j) - F^{-1}(p_j)] = \sqrt{n_j}\,[F^{-1}(\hat{p}_j) - x_j b] \quad ,$$

suit approximativement une loi normale de moyenne 0, de variance :

$$p_j(1 - p_j)\left(\left[\frac{d\,F^{-1}(t)}{dt}\right]_{t = p_j}\right)^2 = \frac{p_j(1 - p_j)}{f\,[F^{-1}(p_j)]^2} = \frac{p_j(1 - p_j)}{f\,(x_j b)^2} \; .$$

Finalement nous pouvons écrire le modèle conditionnel approché :

(1.14)

$$F^{-1}(\hat{p}_j) = x_j b + w_j \; , \quad j = 1, \ldots, J \quad ,$$

où les perturbations w_j sont indépendantes, normales, de moyennes nulles et ont pour variances :

$$V\,w_j = \frac{p_j(1 - p_j)}{n_j\,f^2(x_j b)} = \frac{F(x_j b)\,[1 - F(x_j b)]}{n_j\,f^2(x_j b)} \; .$$

L'approximation obtenue apparaît comme un modèle dont la partie déterministe est linéaire dans le paramètre b, dont les perturbations sont non corrélées et hétéroscédastiques. Remarquons cependant que les variances des perturbations dépendent du paramètre b intervenant dans la partie déterministe. On peut alors estimer b en employant une méthode de type moindres carrés quasi-généralisés, par exemple en remplaçant $V\,w_j$ par :

$$\frac{\hat{p}_j\,(1-\hat{p}_j)}{n_j\,f^2\,[F^{-1}(\hat{p}_j)]}\,.$$

Dans le cas d'un modèle logit, la variance de la perturbation devient :

$$V\,w_j = \frac{F\,(x_j b)\,[1 - F\,(x_j b)]}{n_j\,f^2\,(x_j b)} = \frac{1}{n_j\,F\,(x_j b)\,[1 - F\,(x_j b)]}\,.$$

1.6.c — La méthode du χ^2 minimum

La méthode consiste à chercher les valeurs des paramètres rendant minimale une mesure de l'écart entre les fréquences observées $\hat{p}_j$ et les fréquences théoriques $F\,(x_j b)$. Généralement la mesure retenue est du type distance du χ^2 ; pour les observations de type j, cette distance est donnée par :

$$\varphi_{1j}\,(b) = n_j \left\{\frac{[\hat{p}_j - F\,(x_j b)]^2}{F\,(x_j b)} + \frac{[1 - \hat{p}_j - (1 - F\,(x_j b))]^2}{1 - F\,(x_j b)}\right\}$$

$$= n_j\,\frac{[\hat{p}_j - F\,(x_j b)]^2}{F\,(x_j b)\,[1 - F\,(x_j b)]}\,,$$

et pour l'ensemble des observations par :

$$\varphi_1\,(b) = \sum_{j=1}^{J} \varphi_{1j}\,(b) = \sum_{j=1}^{J} n_j\,\frac{[\hat{p}_j - F\,(x_j b)]^2}{F\,(x_j b)\,[1 - F\,(x_j b)]}\,.$$

Remarquons que ce critère pourrait être remplacé par :

$$\varphi_2\,(b) = \sum_{j=1}^{J} n_j\,\frac{[\hat{p}_j - F\,(x_j b)]^2}{\hat{p}_j\,(1 - \hat{p}_j)}\,.$$

L'estimateur obtenu en minimisant φ_2 n'est autre qu'un estimateur des moindres carrés non linéaires quasi-généralisés appliqué au modèle : $\hat{p}_j = F\,(x_j b) + v_j$. Les minimisations de l'un ou l'autre de ces critères conduisent intuitivement à des résultats assez proches dès que les nombres d'observations n_j sont assez grands.

L'écriture des conditions du premier ordre correspondant à ces minimisations n'entraîne pas de simplifications notables, même si on spécifie la forme de F (voir III.4), et il est nécessaire d'utiliser des procédures numériques pour obtenir les estimations correspondantes des paramètres.

I.7. UN EXEMPLE : LA RÉUSSITE DES ÉTUDIANTS EN PREMIÈRE ANNÉE DE MÉDECINE

Cette étude est décrite dans Lassibile [1979]. Il s'agit d'un modèle logit univarié dichotomique. La variable observée pour chaque étudiant est la réussite ($y_i = 1$) ou l'échec ($y_i = 0$) à l'examen de fin d'année.

La probabilité de réussite a été choisie sous forme logistique :

$$P(y_i = 1) = \frac{1}{1 + \exp(-x_i b)} \cdot$$

Les variables explicatives retenues et les valeurs des coefficients estimés par le maximum de vraisemblance sont données ci-dessous.

Variable	Estimation du coefficient	Significativité du coefficient : + 10 % ++ 5 % +++ 1 %
Taille de la commune de résidence des parents	− 0.31	+++
Ressource des parents	1.16	+++
Age	− 3.65	+++
Note obtenue à un test logique	0.18	+++
Moyenne d'écrit au baccalauréat	5.27	+++
Etudes précédentes	1.85	++
Origine du secondaire	0.37	+
Baccalauréat scientifique	1.39	+++
Baccalauréat non scientifique	− 15.13	+
Constante	− 0.53	+++

Les quatre dernières variables (la constante mise à part) sont des variables dichotomiques :

— études précédentes : 1, si l'étudiant était déjà dans l'enseignement supérieur l'année précédente, 0 sinon,

— origine du secondaire : 1, si les études secondaires ont été faites dans un établissement public, 0 sinon,

— baccalauréat scientifique : 1, si l'étudiant possède le bac série C, 0 sinon,

— baccalauréat non scientifique : 1, si l'étudiant possède un bac série A, B, F ou G, 0 sinon.

Ces deux dernières variables dichotomiques permettent en fait de décrire un caractère à trois modalités « avoir un bac C » (section Mathématiques-Physique), « avoir un bac A, B, F ou G » (sections littéraires), « avoir un bac D » (section Sciences Naturelles).

Les valeurs numériques des coefficients n'ont pas d'interprétation directe ; en revanche leur signe et le fait qu'ils soient ou non significatifs sont interprétables. Le signe permet de savoir si la probabilité de réussite est une fonction croissante ou décroissante de la variable explicative correspondante (toutes choses égales par ailleurs). Ainsi la réussite est une fonction décroissante de l'âge.

La non-significativité de certains coefficients (origine du secondaire) permet de repérer des variables expliquant peu la réussite ou la non-réussite.

La « qualité » du modèle a été évaluée par le rapport du maximum de vraisemblance, la valeur trouvée 86 correspondant à une loi du χ^2 à 9 degrés de liberté était très significative.

Le modèle a ensuite été utilisé par Lassibile pour la prévision de la probabilité de réussite d'un étudiant ayant des caractéristiques données.

D'autres applications sont envisageables, telle « comment effectuer le recrutement des étudiants pour avoir un taux de réussite donné ? ».

I.8. ERREURS DE SPÉCIFICATION

Les méthodes d'estimation que nous avons introduites : maximum de vraisemblance, méthodes de Berkson, χ^2 — minimum possèdent de bonnes propriétés, lorsque le modèle (1.4) sur lequel elles sont fondées est correct, c'est-à-dire contient la « vraie loi » des observations. Il est évidemment important de regarder comment se comportent ces méthodes en présence d'erreurs de spécification. Ces erreurs peuvent porter par exemple sur la forme retenue pour la fonction F, sur la liste des variables exogènes, sur l'hypothèse de non corrélation des perturbations, sur le mode de tirage des observations (exercices 6 et 7).

1.8.a — Erreur sur la forme de F

i) Choix entre un modèle logit et un modèle probit

Les modèles logit ont initialement été introduits comme approximations de modèles probit, permettant des calculs plus simples. Quelle est la conséquence de l'emploi d'une fonction F logistique, si la véritable spécification est normale ? Comme l'a vérifié Morimune (1980) par des études de Monte-Carlo, les estimations des paramètres et de leur précision obtenues par ces deux modèles sont généralement peu différentes. Ceci s'explique par la proximité des familles de lois logistiques et normales.

Les deux familles étant définies à un changement d'origine et un paramètre d'échelle près, il suffit de vérifier, par exemple, que la loi normale centrée réduite peut être bien approchée par une distribution de type logistique.

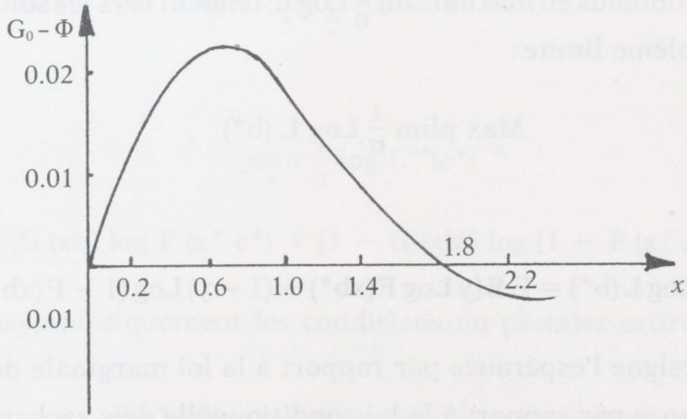
La loi logistique habituelle, de fonction de répartition :

$$\frac{1}{1 + \exp(-x)},$$

a pour moyenne 0, pour variance $\Pi^2/3$; il est donc naturel de comparer la loi N (0,1) avec la loi logistique de fonction de répartition :

$$G_0(x) = \frac{1}{1 + \exp\left(-\frac{\Pi x}{\sqrt{3}}\right)} .$$

La figure ci-dessous donne en fonction de x, la différence $G_0(x) - \Phi(x)$ des deux fonctions de répartition.



ii) Cas où les variables explicatives sont normales

Nous nous proposons dans cette section d'examiner la convergence des estimateurs du maximum de vraisemblance obtenus en maximisant :

$$(1.15)\quad \mathrm{Log}\, L = \sum_{i=1}^{n} \{y_i \,\mathrm{Log}\, F(x_i b^*) + (1 - y_i)\,\mathrm{Log}\,[1 - F(x_i b^*)]\} \;,$$

alors que la véritable fonction de répartition de $\frac{u}{\sigma}$ est G.

Propriété 1.16 :

Si le modèle comporte un terme constant :

$$x_i b = b_0 + x_i^1 b_1 + \ldots + x_i^K b_K \quad ,$$

si les observations $[x_i^1, \ldots, x_i^K]'$ sont indépendantes, extraites d'une même loi normale, les estimateurs b_k obtenus en maximisant la log-vraisemblance incorrecte 1.15 sont tels que :

$$\forall\, k, \ell : \frac{\hat{b}_k}{\hat{b}_\ell} \text{ converge vers } \frac{b_k}{b_\ell} \quad .$$

Les coefficients des vraies variables explicatives sont donc estimés de manière convergente à un coefficient de proportionalité près. Nous donnons ci-dessous une idée de la démonstration et renvoyons à l'article de Ruud (1983) pour une preuve plus rigoureuse.

Preuve

Asymptotiquement les estimateurs du maximum de vraisemblance, obtenus en maximisant $\frac{1}{n}\mathrm{Log}\, L$, tendent vers les solutions du problème limite :

$$\underset{b^*}{\mathrm{Max}}\ \underset{n}{\mathrm{plim}}\ \frac{1}{n}\,\mathrm{Log}\, L\,(b^*) \quad ,$$

avec :

$$\underset{n}{\mathrm{plim}}\frac{1}{n}\mathrm{Log}\, L\,(b^*) = \underset{x}{E}\, E\,[y\, \mathrm{Log}\, F(xb^*) + (1-y)\,\mathrm{Log}\,(1 - F(xb^*))] \quad ,$$

où $\underset{x}{E}$ désigne l'espérance par rapport à la loi marginale de x, E l'espérance par rapport à la loi conditionnelle de y sachant x et plim la limite en probabilité.

Comme Ey = G (xb), où b est la vraie valeur du paramètre, on a :

$$\operatorname*{plim}_{x} \frac{1}{n} \operatorname{Log} L(b^*)$$

$$= \operatorname*{E}_{x} \{G(xb) \operatorname{Log} F(xb^*) + [1 - G(xb)] \operatorname{Log} [1 - F(xb^*)]\} .$$

De manière à faire apparaître les rapports des coefficients des vraies variables explicatives, nous considérons le changement de paramètres suivant :

$$c_1^* = \frac{b_1^*}{b_1} \quad , \quad c_k^* = b_k^* - c_1^* \, b_k \qquad k \geq 2 \quad , \qquad c_0^* = b_0^* - c_1^* \, b_0 \; ,$$

et posons :

$$\tilde{c}^* = (c_2^*, \ldots, c_K^*)' \quad , \quad \tilde{x}^* = (x^2, \ldots, x^K) \quad , \quad x^{1*} = xb$$

On a :

$$\begin{aligned} xb^* &= b_0^* + x^1 b_1^* + x^2 b_2^* + \ldots + x^K b_K^* \\ &= c_0^* + c_1^* b_0 + x^1 b_1 c_1^* + x^2 [b_2 c_1^* + c_2^*] \\ &\quad + \ldots + x^K [b_K c_1^* + c_K^*] \\ &= c_0^* + x^{1*} c_1^* + \tilde{x}^* \tilde{c}^* = x^* c^* . \end{aligned}$$

Dans les nouveaux paramètres le problème de maximisation limite est :

$$\operatorname*{Max}_{c^*} \operatorname*{plim}_{n} \frac{1}{n} \log L^* (c^*) \quad ,$$

avec :

$$\operatorname*{plim}_{n} \frac{1}{n} \log L^* (c^*)$$

$$= \operatorname*{E}_{x} \{G(xb) \log F(x^* c^*) + [1 - G(xb)] \log [1 - F(x^* c^*)]\} \quad .$$

Asymptotiquement les conditions du premier ordre s'écrivent :

$$\operatorname*{E}_{x} \left\{ \frac{f(x^* c^*)}{F(x^* c^*) [1 - F(x^* c^*)]} [G(xb) - F(x^* c^*)] \, x^{*\prime} \right\} = 0 \quad .$$

Il nous faut regarder si ces équations admettent une solution c^* pour laquelle $\tilde{c}^* = 0$.

De manière équivalente ceci revient à regarder si le système de $K + 1$ équations :

$$\underset{x}{E} \frac{f(c_0^* + x^{1*} c_1^*)[G(x_1^*) - F(c_0^* + x^{1*} c_1^*)]}{F(c_0^* + x^{1*} c_1^*)[1 - F(c_0^* + x^{1*} c_1^*)]} \begin{bmatrix} 1 \\ x^{1*} \\ \tilde{x}^{*\prime} \end{bmatrix} = 0 \quad ,$$

admet une solution unique en c_0^* et c_1^*. Cette condition est effectivement satisfaite car les $K - 1$ dernières équations s'écrivent encore :

$$\underset{x}{E} \left\{ \frac{f(c_0^* + x^{1*} c_1^*)[G(x_1^*) - F(c_0^* + x^{1*} c_1^*)]}{F(c_0^* + x^{1*} c_1^*)[1 - (c_0^* + x^{1*} c_1^*)]} \underset{x}{E} [\tilde{x}^{*\prime}/1, x^{1*}] \right\} = 0 \quad ,$$

et sont conséquences des deux premières puisque, du fait de l'hypothèse de normalité $\underset{x}{E} [\tilde{x}^{*\prime}/1, x^{1*}]$ est une fonction linéaire de 1 et x^{1*}.

Q.E.D.

1.8.b — Méthodes non paramétriques

Une méthode du type maximum de vraisemblance fondée sur une spécification erronée de la fonction F ne permet pas comme nous venons de le voir d'estimer de manière convergente tous les paramètres.

Il peut alors être intéressant d'introduire des méthodes d'estimation non paramétriques, ne faisant aucune hypothèse sur la forme de F. De telles méthodes sont plus robustes que celles décrites précédemment, mais en contre-partie sont moins précises, lorsque la forme de F avait été approximativement bien choisie.

i) La méthode du score maximum

Nous décrivons ci-dessous une telle approche proposée par Manski [1975]; cette méthode est habituellement dite du score maximum (on se gardera de la confondre avec la méthode du score 1.5.b). Elle conduit à des estimateurs convergents, lorsque F est la fonction de répartition d'une loi de médiane 0.

La démarche est la suite :

— pour des valeurs données de b, évaluer x_i b,
— pour chaque j compter le nombre de fois où l'alternative choisie est 1 et où $x_i b > 0$ et le nombre de fois où l'alternative choisie est 0 et où $x_i b \leq 0$,
— on somme alors sur $i = 1, ..., n$ et on retient pour b la valeur du paramètre rendant maximum cette somme.

Propriété 1.17 :

La méthode du score maximum fournit des estimateurs convergents.

Preuve :

Souvent les méthodes d'estimation employées consistent à maximiser (ou à minimiser) par rapport au paramètre b une fonction critère du type $\sum_{i=1}^{n} g(y_i, x_i, b)$, où g est une fonction connue.

Une première approche pour voir si l'estimateur ainsi obtenu existe et est fortement convergent est la suivante [voir Jennrich (1969) pour le détail des hypothèses].

a) On regarde si la quantité $\frac{1}{n} \sum_{i=1}^{n} g(y_i, x_i, b)$ converge presque sûrement vers une limite, lorsque n tend vers l'infini. Sous des hypothèses du type : « conditionnellement aux x_i, les variables y_i sont indépendantes », « la distribution empirique des x_i converge en loi vers la distribution de densité $\nu(x)$ »..., on montre que :

$$\frac{1}{n} \sum_{i=1}^{n} g(y_i, x_i, b) \quad ,$$

converge presque sûrement vers $\underset{x}{E}\, E\, g(y, x, b) \quad ,$

où E désigne l'espérance par rapport à la vraie loi de y sachant x et $\underset{x}{E}$ l'espérance par rapport à la loi de densité $\nu(x)$.

b) Il suffit alors de regarder le problème de maximisation limite :

$$\underset{b}{\mathrm{Max}}\, E\, \underset{x}{E}\, g(y, x, b) \quad .$$

Si ce problème admet pour solution unique la vraie valeur b_0 du paramètre, l'estimateur obtenu par maximisation du critère existe presque sûrement et est fortement convergent.

C'est cette approche que nous allons suivre pour étudier les propriétés asymptotiques de l'estimateur du score maximum.

Le « vrai modèle » est défini par :

$$y_i = \begin{cases} 1 \quad , & \text{si } u_i \geq -x_i b_0 \quad , \\ 0 \quad , & \text{sinon} \quad , \end{cases}$$

avec une perturbation u_i suivant une loi F_0 de médiane 0. b_0 et F_0 sont évidemment inconnus et le problème est d'estimer b_0.

La méthode du score consiste à choisir comme estimateur de b_0 la solution du problème de maximisation de :

$$\sum_{i=1}^{n} (\mathbf{1}_{x_i b > 0} \quad \mathbf{1}_{y_i = 1} + \mathbf{1}_{x_i b \leqslant 0} \quad \mathbf{1}_{y_i = 0}) \quad .$$

Le problème limite associé consiste à maximiser :

$$\underset{x}{E}\; E\,[\mathbf{1}_{xb > 0} \quad \mathbf{1}_{y = 1} + \mathbf{1}_{xb \leqslant 0} \quad \mathbf{1}_{y = 0}]$$

$$= \underset{x}{E}\, \{\mathbf{1}_{xb > 0}\, F_0\,(xb_0) + \mathbf{1}_{xb \leqslant 0}\,[1 - F_0\,(xb_0)]\}$$

$$= \int G\,(x, b)\, \nu\,(x)\, dx \quad ,$$

avec :

$$G\,(x, b) = \mathbf{1}_{xb > 0}\, F_0\,(xb_0) + \mathbf{1}_{xb \leqslant 0}\,[1 - F_0\,(xb_0)] \quad .$$

Or :

$$G\,(x, b) - G\,(x, b_0)$$

$$= (\mathbf{1}_{xb > 0} - \mathbf{1}_{xb_0 > 0})\, F_0\,(xb_0) + (\mathbf{1}_{xb \leqslant 0} - \mathbf{1}_{xb_0 \leqslant 0})\,[1 - F_0\,(xb_0)]$$

$$= (\mathbf{1}_{xb > 0}\, \mathbf{1}_{xb_0 \leqslant 0} - \mathbf{1}_{xb_0 > 0}\, \mathbf{1}_{xb \leqslant 0})\, F_0\,(xb_0)$$

$$+ (\mathbf{1}_{xb \leqslant 0}\, \mathbf{1}_{xb_0 > 0} - \mathbf{1}_{xb > 0}\, \mathbf{1}_{xb_0 \leqslant 0})\,[1 - F_0\,(xb_0)]$$

$$= \mathbf{1}_{xb > 0}\, \mathbf{1}_{xb_0 \leqslant 0}\,[2\, F_0\,(xb_0) - 1] + \mathbf{1}_{xb \leqslant 0}\, \mathbf{1}_{xb_0 > 0}\,[1 - 2\, F_0\,(xb_0)] \quad .$$

Comme F_0 est de médiane 0, on a $F_0\,(xb_0) < \frac{1}{2}$, si $xb_0 < 0$

et

$$F_0\,(xb_0) \geqslant \frac{1}{2} \quad , \text{si } xb_0 > 0 \quad .$$

La quantité $G\,(x, b) - G\,(x, b_0)$ est donc toujours négative, et si ν est, par exemple, une loi chargeant $\mathbb{R}$, $\underset{x}{E}\,[G\,(x, b) - G\,(x, b_0)]$ est strictement négatif dès que b est différent de b_0.

Le problème d'optimisation limite admet donc comme solution unique la vraie valeur b_0 ; on en déduit l'existence et la convergence presque sûre de l'estimateur du score maximum.

Q.E.D.

La démarche précédente ne s'intéresse pas à l'estimation de la fonction F_0. Cependant cette fonction de répartition pourrait aussi être estimée par des méthodes non paramétriques (Cosslett (1981), Manski (1982), (1985)).

ii) Analyse discriminante semi-paramétrique

Une démarche différente de celle proposée par Manski et Cosslett a été récemment introduite par [Klein-Spady (1986)]. Elle s'appuie sur les deux remarques suivantes.

Il existe diverses méthodes non paramétriques pour estimer des densités conditionnelles, en particulier des approches de type noyau. Elles pourraient donc théoriquement être utilisées pour estimer la forme de la probabilité conditionnelle :

$$P\,(y = 1/x) \quad .$$

Malheureusement de telles méthodes nécessitent un très grand nombre d'observations dans le cas de plusieurs variables exogènes. Ainsi pour un même degré de précision sur l'estimation, il faut approximation 17 fois plus d'observations dans le cas de 3 variables explicatives que dans le cas d'une seule, 210.000 fois plus dans le cas de 10 variables exogènes.

Il est possible de se ramener facilement à une seule variable conditionnante. Si on suppose les variables (x_i, y_i) $i = 1, ..., n$ indépendantes, de même loi, on peut utiliser la formule de Bayes pour écrire :

$$P\,(y = 1/x) = P\,(y = 1)\,\frac{g\,(x/y = 1)}{g\,(x)} \;,$$

en notant $g\,(x)$ la densité de la loi de x et $g\,(x/y = 1)$ celle de la loi de x sachant $y = 1$ (on considère le cas où x est continue).

b) L'aspect multidimensionnel est alors dans la variable conditionnée. Cet aspect disparaît si le paramètre b est connu, puisque la loi de y sachant x dépend de x par l'intermédiaire de la variable scalaire xb :

$$P\,[y = 1/x] = P\,[y = 1]\,\frac{\tilde{\tilde{g}}\,(xb/y = 1)}{\tilde{\tilde{g}}\,(xb)} \;.$$

Si b était connu, on pourrait approcher la probabilité conditionnelle :

$$P\,[y = 1/x] = p\,(x\,;\,b)$$

par :

$$\hat{p}\,(xb) = \bar{y}\,\frac{\hat{g}\,(xb/y = 1)}{\hat{g}\,(xb)}\,, \tag{1.18}$$

où $\hat{g}$ (xb) et $\hat{g}$ (xb/y = 1) sont des estimateurs non paramétriques des densités g (xb) et g (xb/y = 1) .

Ainsi si $K(z) = \frac{1}{\sqrt{2\Pi}} \exp - \frac{z^2}{2}$ est un noyau gaussien, on peut par exemple retenir :

$$\left\{ \begin{array}{l} \hat{g}\,(xb/y = 1) = \dfrac{\sum\limits_{j=1}^{n} y_j \dfrac{1}{h} K\left[\dfrac{xb - x_j\, b}{h}\right]}{\sum\limits_{j=1}^{n} y_j}\,, \\[3ex] \hat{g}\,(xb) = \dfrac{\sum\limits_{j=1}^{n} \dfrac{1}{h} K\left[\dfrac{xb - x_j\, b}{h}\right]}{n}\,, \end{array} \right. \tag{1.20}$$

où h est choisi assez petit.

Comme le paramètre b est inconnu, $\hat{p}$ (xb) est pour l'instant une approximation non utilisable de la vraie probabilité conditionnelle P (y = 1/x).

La démarche est alors complétée de la façon suivante :

a)dans une première étape, on recherche un estimateur du « maximum de vraisemblance approché » en prenant une solution $\hat{b}$ du problème :

$$\underset{b}{\text{Max}} \sum_{i=1}^{n} \{y_i \log \hat{p}\,(x_i\, b) + (1 - y_i) \log [1 - \hat{p}\,(x_i\, b)]\}\,. \tag{1.21}$$

b) On en déduit alors une estimation de la fonction de réponse P [y = 1/x] en prenant $\hat{p}\,(x\hat{b})$. L'estimateur $\hat{b}$ ainsi défini est convergent [voir Klein-Spady (1986)]. Il est vraisemblablement asymptotiquement normal bien que le résultat ne soit, pour le moment, pas montré.

L'intérêt d'une telle démarche est de donner en plus de l'estimation de b, la forme estimée de la fonction de réponse et ceci sans avoir fait d'hypothèse a priori sur la forme de cette fonction.

1.8.c — Oubli de variables

i) Erreurs sur la liste des variables explicatives

Il est bien connu que dans le cas du modèle linéaire, les estimateur des moindres carrés des paramètres peuvent être utilisés, lorsque les variables oubliées sont orthogonales aux variables figurant dans le modèle. Cette propriété n'est pas valable en général pour les modèles qualitatifs estimés par la méthode du maximum de vraisemblance.

Nous le vérifierons à partir d'un modèle logit comportant deux variables explicatives x^1 et x^2:

$$P(y_i = 1) = F(x_i^1 b_1 + x_i^2 b_2) \quad ,$$

$$\text{avec } F(x) = \frac{1}{1 - \exp(-x)} \quad .$$

Supposons que pour estimer b_1 nous ayons appliqué la méthode du maximum de vraisemblance au modèle erroné: $P^* [y_i = 1] = F(x_i^1 b_1^*)$.

L'estimateur est solution de

$$\sum_{i=1}^{n} [y_i - F(x_i^1 b_1^*)] x_i^1 = 0 \quad ,$$

et converge vers la solution du système limite:

$$\underset{x}{E} \, E[y - F(x^1 b_1^*)] x^1 = 0$$

$$\Leftrightarrow \underset{x}{E} [F(x^1 b_1 + x^2 b_2) - F(x^1 b_1^*)] x^1 = 0 \quad .$$

L'estimateur est convergent de la vraie valeur b_1, si $b_1^* = b_1$ est solution de cette équation, c'est-à-dire si x^1 est orthogonale à

$$F(x^1 b_1 + x^2 b_2) - F(x^1 b_1) \quad .$$

Cette condition de convergence dépend des paramètres inconnus et est donc non utilisable. Elle diffère évidemment de la condition d'orthogonalité entre x^1 et x^2: $E(x^1 x^2) = 0$.

ii) Prise en compte de l'erreur d'oubli de variables

Lorsque chaque expérience j est répétée n_j fois, nous avons vu en I.6 que:

$$\hat{p}_j = F(x_j b) + v_j \quad , \quad j = 1, \ldots, J \quad ,$$

où v_j est une perturbation de moyenne nulle et de variance égale à:

$$\frac{F(x_j b)[1 - F(x_j b)]}{n_j} \quad .$$

Il n'est pas rare en pratique, qu'on dispose d'un nombre très élevé d'observations par strate ; dans ce cas v_j est approximativement nulle et le modèle : $\hat{p}_j = F(x_j\, b)$, ne comportant plus d'aléa, ne peut être ajusté de manière correcte aux données.

Cette difficulté provient du fait que seul a été pris en compte l'aléa de comportement. Or il existe d'autres sources d'erreurs : erreurs de mesure sur les variables observées, erreurs dues à l'omission volontaire ou non de certaines variables, ... Nous allons étudier quelle modification apporte la prise en compte par exemple de cette dernière cause d'erreurs.

Admettons que la variable latente satisfasse un modèle du type :

$$y_{ij}^* = x_{ij}^*\, b^* + u_{ij}, \qquad j = 1, ..., J \quad ; \quad i = 1, ..., n_j \quad ,$$

mais que seules soient utilisées les variables exogènes x_j. On peut alors écrire :

$$y_{ij}^* = x_j\, b + \varepsilon_{ij} + u_{ij} \quad ,$$

où ε_{ij} traduit l'ensemble des variables oubliées. Cette nouvelle erreur est indépendante de l'erreur de comportement u_{ij}. D'autre part, il est vraisemblable que parmi les variables omises, certaines sont spécifiques de la strate, d'autres de l'individu. Il est alors naturel de décomposer ε_{ij} en deux termes :

$$\varepsilon_{ij} = \alpha_j + \beta_{ij} \quad ,$$

où α_j et β_{ij} sont deux perturbations de moyennes nulles que nous supposerons indépendantes entre elles et indépendantes des x_j.

On a finalement :

$$y_{ij}^* = x_j\, b + \alpha_j + u_{ij}^* \quad , \text{avec} \qquad u_{ij}^* = \beta_{ij} + \varepsilon_{ij} \quad .$$

La différence essentielle entre ce modèle et celui utilisé jusqu'à maintenant provient de la corrélation existant entre les observations d'une même strate.

Si F désigne la fonction répartition de $-\,u_{ij}^*$,

$$\text{si} \qquad y_{ij} = \begin{cases} 0 & , \text{ si } \ y_{ij}^* \geqslant 0 \quad , \\ 1 & , \text{ si } \ y_{ij}^* < 0 \quad , \end{cases} \qquad \text{désignent}$$

les variables dichotomiques observées et

XI. MODÈLES DE DURÉE

BARTHOLOMEW, D.J. (1963), « The Sampling Distribution of an Estimate Arising in Life Testing », *Technometrics* 5, 361-374.

CHESHER, A. (1984), « Testing for Neglected Heterogeneity », *Econometrica* 52, 865-872.

CHESHER, A. et LANCASTER, T. (1983), « The Estimation of Models of Labour Market Behaviour », *R.E.S.*, 609-624.

COX, D.R. (1972), « Regression Models and Life Tables », JRSS, B. 34, 187-220.

COX, D.R. (1975), « Partial Likelihood », *Biometrika* 62, 269-276.

COX, D.R. et OOAKES, D. (1984), « Analysis of Survival Data », Chapman-Hall.

ELBERS, C. et RIDDER, G. (1982), « True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model », *Review of Economic Studies* 49, 403-410.

FLINN, C.J. et HECKMAN, J.J. (1982), « New Methods for Analyzing Structural Models of Labor Force Dynamics », *Journal of Econometrics* 18, 115-168.

FLINN, C.J. et HECKMANN, J.J. (1982), « Models for the Analysis of Labor Force Dynamics », in *Advances in Econometrics,* vol. 1, Ed. by Basmann et Rhodes, J.A.I. Press 35-95.

FLINN, C.J. et HECKMANN, J.J. (1983) « The Likelihood Function for the Multistate-Multiepisode Model in Models for the Analysis of Labor Force Dynamics », dans *Advances in Econometrics,* vol. 2, J.A.I. Press, 225-231.

FOUGERE, D. (1986), « La Recherche d'Emploi: Analyse Théorique et Étude Économétrique », Thèse, Univ. de Toulouse.

HECKMAN, J.J. (1981), « Heterogeneity and State Dependence », dans *Studies in labor markets,* Ed. S. Rosen, Univ. of Chicago Press.

HECKMAN, J.J. et BORJAS, G.J. (1980), « Does Unemployment Cause Future Unemployment ? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence », *Economica* 47, 247-283.

HECKMAN, J.J. et SINGER, B. (1982), « The Identification Problem in Econometric Models for Duration Data », dans *Advances in Econometrics,* Ed. Hildenbrand, Cambridge University press, 39-77.

HECKMAN, J.J. et SINGER, B. (1974), « Econometric Duration Analysis », *Journal of Econometrics,* 24, 63-132.

HECKMAN, J.J. et SINGER, B. (1974), « A Method for Minimizing the Impact of Distributional Assumption in Econometric Models for Duration Data », *Econometrica,* 52, 271-320.

HECKMAN, J.J. et SINGER, B. (1984), « The Identifiability of the Proportional Hazard Model », *Rev. of Econ. Studies,* 231-241.

HECKMAN, J.J. et SINGER, B. (1985), « Longitudinal Analysis of Labor Market Data », Cambridge University Press.

KALBFLEISCH, J et PRENTICE, R. (1980), « The Statistical Analysis of Failure Time Data », Wiley.

KIEFER, N. (1984), « A Simple Test for Heterogeneity in Exponential Models of Duration », *Journal of Labor Economics,* 2, 539-549.

KIEFER, N. (1985), « Specification Diagnostics Based on Laguerre Alternatives for Econometric Models of Duration », *Journal of Econometrics* 28, 135-154.

KIEFER, N., LUNDBERG, S. et NEUMANN, G. (1984), « How Long is a Spell of Unemployment ? Illusions and Biases in the Use of CPS data », *NBER D.P.* 1467.

KIEFER, N. et NEUMAN, G. (1979), « An Empirical Search Model with a Test of the Constant Reservation Wage Hypothesis », *Journal of Political Economy* 87, 89-107.

KIEFER, N. et NEUMANN, G. (1981), « Individual Effects in a Nonlinear Model : Explicit Treatment of Heterogeneity in the Empirical Job-Search Model », *Econometrica* 49, 965-978.

LANCASTER, T. (1979), « Econometric Methods for the Duration of Unemployment », *Econometrica* 47, 939-956.

LANCASTER, T. (1983), « Generalised Residuals and Heterogeneous Duration Models : the Exponential Case », *Bulletin of Economic Research* 35, 71-85.

LANCASTER, T. (1985), « Generalised Residuals and Heterogenous Duration Models : with Applications to the Weibull Model », *Journal of Econometrics* 28, 155-169.

LANCASTER, T. (1986), « Some Remarks on Wage and Duration Econometrics », dans Blundell-Walker, *Unemployment, search and labour supply*, Cambridge University Press.

LANCASTER, T. et CHESHER, A. (1983), « An Econometric Analysis of Reservation Wages », *Econometrica* 51, 1661-1676.

LANCASTER, T. et CHESHER, A. (1985), « Residual Analysis for Censored Duration Data », *Economic Letters*, 18, 35-38.

LANCASTER, T. et CHESHER, A. (1985), « Residuals, Tests and Plots with a Job Matching Illustration », *Annales de l'INSEE* 7-70.

LANCASTER, T. et NICKEL, S. (1980), « The Analysis of Re-employment Probabilities of the Unemployed », *JRSS*, A, 141-165.

LEE, L.F. (1984), « Maximum Likelihood Estimation and a Specification Test for Normal Distributional Assumption for Accelerated Failure Time Models », *Journal of Econometrics* 24, 159-179.

LIPPMAN, S. et MCCALL, J. (1976), « The Economics of Job Search : a Survey », Part 1, *Economic Inquiry*, 14, 155-189.

LIPPMAN, S. et MCCALL, J. (1976), « Job Search in a Dynamic Economy », *J.E.T.* 12, 365-390.

LYNCH, L. (198?), « State Dependency in South Unemployment : a Lost Generation ? » *Journ. of Econ.*, à paraître.

NARENDRANATHAN, W. et NICKEL, S.J. (198?), « Modelling the Process of Job Search », *Journ. of Econ.*, à paraître.

NICKELL, S. (1979), « Estimating the Probability of Leaving Unemployement », *Econometrica* 47, 1249-1266.

NICKELL, S. (1979), « The Effect of Unemployement and Related Benefits on the Duration of Unemployment », *Economic Journal*, 34-49.