Ecole Nationale Supérieure de Statistique et d'Economie Appliquée-ENSEA, Abidjan, Côte d'Ivoire

Econométrie des variables qualitatives

Note pédagogique

KEHO Yaya
yayakeho@yahoo.fr
yaya.keho@ensea.ed.ci

KANGA Désiré
kangadesire@yahoo.fr
desire.kanga@ense.ed.ci

Février 2007

Chapitre 1

En guise d'Introduction

1. Position du problème

L'analyse des comportements individuels ne repose pas toujours sur des variables continues comme le revenu, la consommation, l'investissement…; elle peut aussi porter sur des phénomènes à caractère qualitatif comme, par exemple, le fait de consommer un certain bien, le fait d'adhérer à une union syndicale ou à une association, le moyen de transport utilisé, le fait de choisir une filière de formation parmi un ensemble, le fait d'exercer une certaine activité professionnelle. Dans ces cas, la variable explicative Y ne peut prendre qu'un nombre limité de modalités. Il a été démontré que dans un tel cadre, l'économétrie classique (notamment la méthode des MCO) produit des résultats biaisés et non-convergents. En lieu et place de l'application des méthodes de l'économétrie classique, plusieurs modèles ont été développés selon la nature de la variable ou du phénomène à analyser. C'est à cette fin que répond l'économétrie des variables qualitatives et des variables à domaine de définition limitée. Dans toute la suite, nous utiliserons l'appellation anglaise CLDV (Categorial and Limited Dependent Variable) pour désigner l'économétrie des variables qualitatives et des variables à domaine de définition limitée.

Il existe quatre grands groupes de modèles définis selon la nature des variables à analyser qui rentrent dans la cadre des CLDV : les modèles binaires ou dichotomiques, les modèles multinomiaux, les modèles de comptage, les modèles censurés ou tronqués.

Les modèles binaires ou dichotomiques sont élaborés pour les cas où la variable dépendante à analyser Y est susceptible de prendre deux valeurs (0 ou 1), permettant ainsi de rendre compte de l'occurrence ou non d'un événement.

Les modèles multinomiaux sont une généralisation des modèles dichotomiques aux cas où la variable dépendante à analyser Y est susceptible de prendre plus de deux valeurs. C'est le cas par exemple, du statut matrimonial, des avis donnés lors d'une enquête de satisfaction sur une échelle de plus de deux modalités (1=très satisfait, 2=satisfait, 3=pas satisfait, 4=pas du tout satisfait). Il existe une gamme de modèles multinomiaux selon que la variable est ordonnée, non ordonnée ou séquentielle.

Les modèles de comptage, quant à eux, sont élaborés pour la modélisation des variables prenant un nombre très limité de modalités positives et traduisant le plus souvent un phénomène de comptage. Par exemple, le nombre d'appel entre 12 heures et 13 heures à un poste de police, le nombre passagers à une gare de bus entre 12 heures et 14 heures.

Les modèles censurés et tronqués sont adaptés au cas de variables d'intérêt « coincés » entre deux valeurs ou présentant soit une contrainte de supériorité soit une contrainte d'infériorité.

Dans chacun des modèles ci-dessus énumérés, on veut expliquer les réalisations du phénomène observé. A cet effet, on entend croiser les réalisations de la variable Y avec celles d'un certain

nombre de variables explicatives $X_1,\ldots,X_k$ dont les réalisations peuvent être indifféremment de natures qualitative ou quantitative.

Les modèles à variables qualitatives sont de plus en plus utilisés parmi l'éventail des outils d'inférence statistique. Leurs applications se révèlent fort diverses, des études épidémiologiques aux études de marché du travail et d'allocation du temps, en passant par le marché du crédit. Dans les modèles à variable dépendante limitée, la méthode traditionnelle des Moindres Carrés Ordinaires ne semble plus adaptée car elle doit tenir compte de l'absence de continuité de la variable endogène et souvent de l'absence d'un ordre naturel entre les modalités de cette variable.

2. Pourquoi les MCO ne sont-ils pas appropriés?

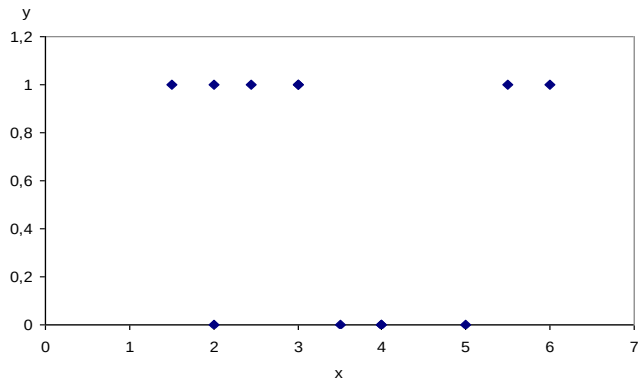
Pour illustrer le fait que les variables catégorielles « violent » les hypothèses des MCO et ne peuvent pas se prêter à cette méthodologie; nous allons utiliser le modèle binaire ou dichotomique. On considère une variable dichotomique *y* à deux modalités 0 et 1, et *x* un vecteur de *k+1* variables explicatives. On cherche à expliquer la réalisation de l'évènement *y=1* par les variables de *x*. x'_i est le vecteur (*1, k+1*) des caractéristiques de l'individu *i*. Le modèle linéaire classique s'écrit *:*

$$y_i = x'_i\, a + e_i \qquad (1)$$

Plusieurs éléments rendent inappropriée l'estimation de ce modèle par la méthode des MCO :

1) Les deux membres du modèle sont de nature différente : y_i est qualitative et $x'_i\, a + e_i$ est quantitative continue.

2) Une représentation graphique des points montre que l'approximation linéaire n'est pas adaptée.



Le nuage de points se trouve sur deux droites parallèles. On voit mal comment on peut faire passer une seule droite d'ajustement par ces points.

3) Les erreurs du modèle prennent deux valeurs :

- $e_i = 1 - x'_i\, a$ quand $y_i = 1$ avec une probabilité $\Pr(y_i = 1) = x'_i\, a$.

- $e_i = x'_i\, a$ quand $y_i = 0$ avec une probabilité $\Pr(y_i = 0) = 1 - x'_i\, a$.

Par conséquent, les erreurs ne peuvent être continues, à fortiori suivent une loi de distribution normale. Le non-respect de l'hypothèse de normalité ne permet pas d'utiliser les statistiques usuelles de test (Student, Fisher, Chi-deux).

4) La variable dépendante y_i suit une loi binomiale de paramètre $x'_i\, a$. Le terme d'erreur e_i est aussi une binomiale et $Var(e_i) = x'_i\, a(1 - x'_i\, a)$: les erreurs sont hétéroscédastiques par construction et les estimateurs par MCO ne sont pas efficaces. La variance dépend des variables explicatives. Cependant, cet inconvénient est mineur puisqu'on peut utiliser les moindres carrés pondérés pour résoudre ce problème économétrique. Après avoir estimé le modèle par MCO, on tire $\hat{\sigma}_i^2 = \hat{y}_i(1 - \hat{y}_i)$ comme estimateur de $Var(e_i) = \sigma_i^2$. Ensuite, on applique les moindres carrés pondérés, c'est-à-dire les MCO au modèle linéaire obtenu en divisant les observations par $\hat{\sigma}_i$.

5) On a par nature $y_i \in \{0;1\}$. Or, rien n'impose que les prédictions $\hat{y}_i$ appartiennent à l'intervalle [0 1]. Même si on estime sous les contraintes $0 \le x_i a \le 1$, rien ne garantit que ces contraintes soient compatibles entre elles. Le risque d'avoir des probabilités calculées négatives est présent. Il se peut ainsi que $\hat{\sigma}_i^{\,2}$ soit négative!

3. Principe d'estimation des CLDV

La méthode d'estimation de choix des CLDV est la Méthode du Maximum de Vraisemblance. Avant toute chose, il convient de bien identifier la fonction de répartition de Y ou sa fonction de densité. Cependant, compte tenu de la nature qualitative des variables, on a recours à des hypothèses sur la distribution des erreurs en utilisant une approche par les variables latentes. Soit *f* la fonction de densité retenue. Considérons un échantillon de taille *n.* On construit la fonction de vraisemblance comme suit :

$$L(\beta) = \prod_{i=1}^{n} f(y_i|X_i) \qquad (2)$$

où β est le vecteur des paramètres à estimer.

La détermination des paramètres β se fait par la maximation du logarithme de la vraisemblance, c'est-à-dire la résolution du programme :

$$\underset{\beta}{Max} \log\left(L(\beta)\right) \qquad (3)$$

Pour résoudre cette équation, on utilise les algorithmes du calcul numérique. A cet effet, plusieurs algorithmes basés sur le principe itératif sont disponibles. Le principe est le suivant :

- On part d'une valeur $\hat{\beta}_0$;
- On détermine $\hat{\beta}_1$ tel que $\hat{\beta}_1 = \hat{\beta}_0 + \varepsilon_0$, ε_0 est un incrément ;
- On vérifie si $\left\|\hat{\beta}_{MV} - \beta'\right\| < \varepsilon$ pour ε aussi petit que l'on veut ;
- Si c'est le cas, on s'arrête et l'estimateur est $\hat{\beta}_1$;
- Sinon, on continue le processus jusqu'à avoir $\left\|\hat{\beta}_{MV} - \beta'\right\| < \varepsilon$ pour ε petit ;

Toute la différence au niveau des méthodes de résolution diffèrent au niveau du choix de ε_0. Quatre algorithmes sont le plus souvent utilisés :

1. La méthode Steepest Ascent :

$$\varepsilon_m = \frac{\partial L(\hat{\beta}_m)}{\partial \beta} \ , \hat{\beta}_{m+1} = \hat{\beta}_m + \varepsilon_m \tag{4}$$

2. La méthode de Newton-Raphson :

$$\varepsilon_m = \left[\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'}\right]^{-1}_{\beta=\hat{\beta}_m} \times \frac{\partial L(\hat{\beta}_m)}{\partial \beta}, \quad \hat{\beta}_{m+1} = \hat{\beta}_m - \varepsilon_m \tag{5}$$

3. La méthode de Scoring :

$$\varepsilon_m = \left[\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'}\right]^{-1}_{\beta=\hat{\beta}_m} \times \frac{\partial L(\hat{\beta}_m)}{\partial \beta}, \ \hat{\beta}_{m+1} = \hat{\beta}_m + \varepsilon_m \tag{6}$$

4. La méthode de Bernd-Hall-Hall-Hausman (BHHH) :

$$\varepsilon_m = \left[\sum_{i=1}^{n} \left\{\frac{\partial L_i}{\partial \hat{\beta}_m}\right\} \times \left\{\frac{\partial L_i}{\partial \hat{\beta}_m}\right\}'\right]^{-1}, \ \hat{\beta}_{m+1} = \hat{\beta}_m + \varepsilon_m \tag{7}$$

4. Démarche générale d'analyse des modèles CLDV

De façon pratique, pour s'assurer que le maximum de vraisemblance fonctionne, il faut :

- 10 observations au moins par paramètre estimé sans toutefois oublier qu'il est conseillé d'atteindre au moins un échantillon de taille 100 ;

- Eviter la multicolinéarité entre les variables explicatives : on doit avoir une indépendance linéaire des colonnes de X ;

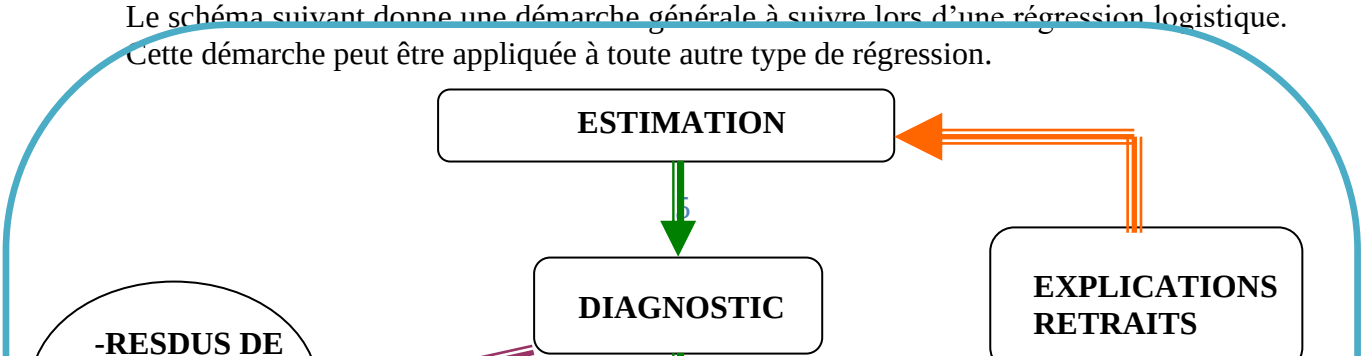
- Avoir à l'esprit le principe « GIGO » : **G**arbage **I**n, **G**arbage **O**ut

- Des modèles comme Poisson, Binomial Négatif, ZIP (Zero Inflated Poisson), ZINB (Zero Inflated Binomial) sont gourmands en observations.

L'analyse des modèles CLDV procède suivant les étapes ci-après:

1. Estimer le modèle à l'aide du MV.
2. Test d'hypothèse à l'aide des tests tels que Wald, rapport de vraisemblance, multiplicateur de Lagrange, etc.
3. Mesurer l'adéquation du modèle aux données à l'aide des critères d'information (AIC, BIC, Schwartz) et de certains tests tel que celui de Hosmer et Lemeshow.
4. Interprétation du modèle à l'aide des probabilités prédites, des odds ratios, etc.

Le schéma suivant donne une démarche générale à suivre lors d'une régression logistique. Cette démarche peut être appliquée à toute autre type de régression.



Chapitre 2

Modèles dichotomiques Probit et Logit

1. Introduction

Ce chapitre présente les modèles destinés à l'analyse des phénomènes binaires. Dans ces modèles, la variable dépendante Y prend deux valeurs (0 ou 1) indiquant l'occurrence ou non d'un événement. On parle ainsi des modèles dichotomiques ou binaires. On veut expliquer pourquoi cet événement se produit (ou, au contraire, ne se produit pas). A cet effet, on entend croiser les réalisations de la variable binaire Y avec celles d'un certain nombre de variables explicatives X_1, X_2…, X_k dont les réalisations peuvent être indifféremment de natures qualitative ou quantitative.

Le chapitre introductif a montré que pour ce type de variables, la méthode traditionnelle des Moindres Carrées Ordinaires n'est pas adaptée car elle doit tenir compte de l'absence de continuité de la variable dépendante et souvent de l'absence d'un ordre naturel dans les modalités de cette variable. Afin de pallier cette limite, en général, on a recours à deux types de modèles selon l'hypothèse faite sur la distribution des termes d'erreurs : le modèle Logit et le modèle Probit.

Ce chapitre présente les intuitions et les développements théoriques permettant de formaliser et mieux interpréter les modèles dichotomiques. Un aperçu des domaines d'application de ces modèles est également présenté.

2. Quelques domaines d'application

Il existe de nombreux domaines d'application des modèles dichotomiques. Ces modèles peuvent être utilisés à titre explicatif, pour rechercher les déterminants d'un phénomène donné, ou à titre prévisionnel, pour prédire un phénomène pour des cas nouveaux. Voici quelques domaines d'application intéressants de ces modèles.

a) Anthropométrie

On dispose de données anthropométriques relatives à un échantillon de crânes d'hommes et de crânes de femmes. On cherche à savoir quels sont les déterminants des crânes d'hommes et à déterminer le sexe (inconnu) d'un individu dont on a retrouvé le crâne lors de fouilles archéologiques. On s'intéresse donc à la probabilité que le crâne soit du sexe masculin.

b) Médecine

On dispose de mesures cliniques, biologiques... caractérisant des malades atteints de la même maladie. Après avoir observé l'évolution de ces patients sur une période, on cherche les déterminants qui expliqueraient la survie des malades. On peut ainsi prévoir le diagnostic final pour un nouveau malade atteint de la même maladie au vu de ses caractéristiques biologiques.

c) Finance et banque

Les banques sont intéressées à prévoir le comportement des demandeurs de crédits, en fonction de leurs caractéristiques qui doivent discriminer entre les « bons clients » et les « mauvais clients». A partir d'une régression logistique, on chercher à savoir quelles sont les caractéristiques des clients qui expliquent leur comportement face au crédit (bon client/mauvais client). Des méthodes en analyse discriminante permettent également de résoudre ce genre de problème (Crédit Scoring).

3. Notion de variable latente et modélisation des variables dichotomiques

Pour pallier les insuffisances de la spécification linéaire, une approche classique consiste à considérer la variable endogène *y* comme étant la manifestation d'une variable « cachée » y^* inobservable ; cette dernière étant reliée à un ensemble de variables explicatives *X*. Nous allons illustrer cette approche en considérons trois exemples.

Le premier exemple et le plus célèbre est tiré de la biologie, celui de l'insecticide : on diffuse dans un espace clos un insecticide et l'on cherche à déterminer la dose minimale permettant de tuer les insectes. Pour cela, on observe au terme d'une période fixé les insectes morts pour lesquels on adopte le code $y_i = 0$ et ceux encore vivants codés $y_i = 1$. On suppose alors que chaque insecte dispose d'une capacité de résistance propre qui se traduit par un seuil inobservable de produit, noté y_i^*, telle que si la dose de produit γ est supérieure à ce seuil l'insecte meurt ($y_i = 0$), il reste vivant (mais malade peut être) pour une dose γ inférieure ($y_i = 1$). On cherche alors à modéliser la probabilité de survie de l'insecte i en fonction de la dose d'insecticide et des observations faites sur y_i. Le problème peut s'écrire de la façon suivante :

$$y_i = \begin{cases} 1 & si \quad y_i^* > \gamma \\ 0 & si \quad y_i^* \le \gamma \end{cases} \tag{2.1}$$

La tolérance y_i^* peut s'écrire comme la somme d'une combinaison linéaire de caractéristiques propres à chaque insecte et d'un terme d'erreur.

$$y_i^* = x'_i\, a + e_i \tag{2.2}$$

Un autre exemple, toujours tiré de la biologie, concerne la probabilité pour un mineur i de ne pas contracter une maladie des poumons (événement codé $y_i = 1$). Le mineur contracte la maladie ($y_i = 0$) lorsque sa tolérance inobservable, notée y_i^*, aux conditions de travail et en particulier aux poussières de charbon, est inférieure à certain seuil γ inconnu. On suppose que la tolérance est liée à l'âge du mineur noté x_i. De la même façon que précédemment, ce problème peut s'écrire sous la forme :

$$y_i = \begin{cases} 1 & si \quad y_i^* = \beta_0 + \beta_1 x_i + e_i \succ \gamma \\ 0 & \sin on \end{cases} \tag{2.3}$$

Le troisième exemple s'intéresse à la consommation d'un certain bien C (par exemple un poste radio avec lecteur CD). On définit la variable $y_i = 1$ si l'individu *i* a acheté le bien C, $y_i = 0$ sinon. La variable « *y*= acheter le bien C» est celle qui est observée ; mais en réalité elle est le résultat d'un arbitrage en terme d'utilité. Si $U(x_i,1)$ représente l'utilité que procure l'achat du bien C à l'individu *i* de caractéristiques x_i et $U(x_i,0)$ l'utilité liée à la non consommation du bien, on peut poser que $y_i = 1 \Leftrightarrow U(x_i,1) > U(x_i,0) + c$. La variable inobservée $y_i^* = U(x_i,1) - U(x_i,0)$ est la variable latente qui sous-tend l'achat du bien C. On a :

$$y_i = \begin{cases} 1 & si \quad y_i^* = x'_i\, a + e_i > c \\ 0 & si \quad y_i^* \le c \end{cases} \tag{2.4}$$

Dans cet exemple, nous avons assimilé la variable latente y_i^* à une différence d'utilité. Mais cette variable inobservable peut représenter n'importe quelle grandeur économique susceptible d'affecter le comportement d'achat du bien C.

Tout modèle dichotomique peut s'écrire sous la forme suivante :

$$y_i = \begin{cases} 1 & si \quad y_i^* > c \\ 0 & si \quad y_i^* \le c \end{cases} \tag{2.5}$$

où la variable latente y_i^* inobservable est liée à un ensemble de caractéristiques observables x_i et à une perturbation e_i :

$$y_i^* = x'_i\, a + e_i \tag{2.6}$$

Les erreurs sont supposées $i.i.d\left(0, \sigma^2\right)$.

Si les y_i^* étaient observables, on pouvait estimer directement le modèle (2.6) à l'aide des MCO, mais malheureusement ce n'est pas le cas. On peut cependant estimer la probabilité de réalisation de l'événement $(y_i = 1)$:

$$\Pr(y_i = 1) = \Pr(y_i^* > c) = \Pr(e_i > c - x'_i\, a) = 1 - \Pr(e_i \le c - x'_i\, a) = 1 - F(c - x'_i\, a) \tag{2.7}$$

où F est la fonction de répartition des erreurs. On fait l'hypothèse que la distribution des erreurs est symétrique autour de sa moyenne : $f(x) = f(-x)$ et $F(x) = 1 - F(-x)$. Dans ces conditions, on a :

$$\Pr(y_i = 1 / x_i) = F(x'_i\, a - c) \tag{2.8}$$

Remarque : La variable latente y^* n'a pas toujours une interprétation économique claire, elle n'est qu'un artefact destiné à modéliser la réalisation de la variable observée.

4. Estimation du modèle

L'estimation du modèle dichotomique se fait généralement par la méthode du maximum de vraisemblance. Pour cela, on écrit la vraisemblance de l'échantillon. Lorsque les observations individuelles y_i , $i=1,...,n$, sont indépendantes, cette vraisemblance s'écrit comme le produit des probabilités. La méthode du maximum de vraisemblance consiste alors à trouver les valeurs des paramètres qui rendent l'observation des données la plus vraisemblable, c'est-à-dire à maximiser la fonction de vraisemblance. Il s'agit en fait de chercher à faire dire au modèle la même chose que la nature.

Si les observations sont indépendantes et identiquement distribuées, la probabilité jointe est le produit des probabilités associées à chaque observation :

$$L(y,x,a)=\prod_{i=1}^{n}\left(F(x'_i\,a)\right)^{y_i}\left(1-F(x'_i\,a)\right)^{1-y_i} \tag{2.9}$$

Cette fonction de vraisemblance peut aussi s'écrire :

$$L(y,x,a)=\prod_{i=1}^{n}\left(F(\delta_i x_i a)\right) \tag{2.10}$$

avec

$\delta_i = 1$ si $y_i = 1$

$\delta_i = -1$ si $y_i = 0$

Les conditions de premier ordre de la maximisation de la fonction de log-vraisemblance s'écrivent:

$$\begin{aligned} S(a)=\frac{\partial l}{\partial a}&=\sum_{i=1}^{n}\left[\frac{y_i}{F(x'_i\,a)}-(1-y_i)\frac{1}{1-F(x'_i\,a)}\right]f(x'_i\,a)x'_i= \\ \sum_{i=1}^{n}&\left[\frac{y_i-F(x'_i\,a)}{F(x'_i\,a)(1-F(x'_i\,a))}\right]f(x'_i\,a)x'_i=0 \end{aligned} \tag{2.11}$$

Pour résoudre cette équation, il faut expliciter la forme fonctionnelle de F. En pratique, deux lois de distribution sont utilisées: la loi logistique et la loi normale.

4.1 Le modèle Logit

On pose ici l'hypothèse que les erreurs ont une distribution logistique. La fonction de répartition s'écrit :

$$F(x)=\frac{e^x}{1+e^x}=\frac{1}{1+e^{-x}}=\Lambda(x) \tag{2.12}$$

Le modèle Logit impose la variance des erreurs égale à $\pi^2/3$. La fonction de log-vraisemblance prend la forme $l=a'\sum_{i=1}^{n}x'_i y_i-\sum_{i=1}^{n}\ln\left(1+e^{x'_i a}\right)$ et les conditions de premier ordre donnent:

$$S(a) = \sum_{i=1}^{n} \left(y_i - \frac{1}{1+e^{-x'_i a}} \right) x'_i = 0 \qquad (2.13)$$

4.2 Le modèle Probit

Dans le cas du modèle Probit, la fonction de répartition F est celle de la loi normale centrée réduite :

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2} t^2} dt = \Phi(x) \qquad (2.14)$$

Les conditions de premier ordre s'écrivent:

$$S(a) = \frac{\partial l}{\partial a} = \sum_{i=1}^{n} \left[\frac{y_i}{\Phi_i} - (1-y_i) \frac{1}{1-\Phi_i} \right] \phi_i x'_i = \sum_{i=1}^{n} \left[\frac{y_i - \Phi_i}{\Phi_i (1-\Phi_i)} \right] \phi_i x'_i = 0 \qquad (2.15)$$

Le système d'équations défini par les conditions du premier ordre est non-linéaire. On est contraint de rechercher une solution numérique (et non pas analytique) pour ce problème. Pour cela, on devra utiliser un algorithme d'optimisation numérique de la fonction de vraisemblance (voir Gouriéroux,(1989) page 20 pour les algorithmes de résolution).

Remarques 1 : Problèmes d'identification

L'estimation des modèles Logit / Probit pose deux problèmes :

- le seuil c qui détermine la modalité 0 ou 1 ne peut être estimé indépendamment de la constante ;
- la variance σ de l'erreur e ne peut être estimée indépendamment des coefficients.

Pb n° 1 : Le seuil ne peut être identifié

$$\Pr(y_i = 1) = \Pr(y_i^* > c) = \Pr(e_i > c - x'_i a) = \Pr(e_i > (c - a_0) - \sum_{j=1}^{k} a_j x_{ij}) \qquad (2.16)$$

La constante a_0 et le seuil c ne peuvent être dissociés $\rightarrow$ par la suite on fera « comme si » c = 0.

Pb n° 2 : La variance de l'erreur ne peut être identifiée

$$\Pr(y_i = 1 / x_i) = \Pr(e_i > -x'_i a) = \Pr\left(\frac{e_i}{\sigma} > -x'_i \frac{a}{\sigma} \right) = \Phi\left(x'_i \frac{a}{\sigma} \right) \qquad (2.17)$$

$\rightarrow$ Il est impossible de dissocier σ de a. Les estimateurs des coefficients ne sont donc identifiés qu'à une constante multiplicative près $(1/\sigma)$. On peut faire «comme si» $\sigma = 1$.

Remarque 2: Lorsqu'on est en présence de mesures répétées ou que les données présentent une « structure hiérarchique », comme c'est le cas lorsqu'on échantillonne des ménages et que l'on s'intéresse aux caractéristiques des membres de ces ménages, l'hypothèse d'indépendance des données n'est pas plausible. Dans ce cas, il faut utiliser d'autres méthodes qui prennent en compte la corrélation des données (ex : modèle marginal, modèle logistique conditionnel, modèle mixte). Les situations de mesures répétées sont plus fréquentes en biologie (expériences répétées) qu'en économie.

5. Propriétés des estimateurs Logit et Probit

Les estimateurs Logit et Probit obtenus par la méthode du maximum de vraisemblance possèdent les propriétés asymptotiques suivantes.

a. L'estimateur $\hat{a}$ converge en probabilité vers la vraie valeur a. Cela signifie que plus la taille de l'échantillon est grande, plus l'estimation tend vers la vraie valeur.

$$p\lim \hat{a} = a \tag{2.18}$$

b. Il est asymptotiquement normalement distribué:

$$\hat{a} \to N(a, I(a)^{-1}) \tag{2.19}$$

où $I(a) = -\left[E\left(\frac{\partial^2 \log L}{\partial a \partial a'} \right) \right]$ est la matrice d'information de Fisher. On montre que :

$$I(a) = \sum_{i=1}^{n} \left[\frac{f^2(x'_i a)}{F(x'_i a)(1 - F(x'_i a))} \right] x'_i x_i \tag{2.20}$$

c. L'estimateur $\hat{a}$ est asymptotiquement efficace: en grands échantillons, l'estimateur du maximum de vraisemblance utilise de façon optimale l'information contenue dans les données.

6. Qualité d'ajustement du modèle

Dans les modèles qualitatifs, plusieurs statistiques peuvent être utilisées pour juger de la qualité de l'ajustement. Les plus courantes sont le test du rapport de vraisemblance et le pseudo R^2 de Mc-Fadden.

6.1 Le test du rapport de vraisemblance (LR-test)

Le test du rapport de vraisemblance consiste à comparer deux modèles, à savoir le modèle estimé avec la constante seule et le modèle estimé avec toutes les variables explicatives (qu'on appelle modèle saturé). C'est donc l'analogue du test de Fisher dans le cas des modèles estimés par maximum de vraisemblance. L'hypothèse nulle de ce test s'écrit :

$$H_0 : a_1 = a_2 = ... = a_k = 0 \tag{2.21}$$

Puisque le modèle sous H_0 est emboîté dans le modèle saturé, la vraisemblance est augmente au fur et à mesure qu'on ajoute de nouvelles variables: un modèle explique mieux la réalité avec davantage de variables explicatives. La vraisemblance du modèle saturé est donc supérieure à celle du modèle contraint. Suivant la vraisemblance, on aura tendance à choisir le modèle contraint. Mais s'il se trouve que l'écart entre les deux vraisemblances est non significatif, alors on choisira le modèle contraint, car il explique aussi bien la réalité que le modèle saturé avec moins de variables. On le retient si on préfère les modèles parcimonieux.

Le test du rapport de vraisemblance est donc basé sur l'écart entre les log-vraisemblances des deux modèles. La statistique du test est défini par :

$$LR = -2(l_o - l) \tag{2.22}$$

où l_o est la log-vraisemblance du modèle estimé avec la constante seule comme variable explicative, c'est-à-dire sous l'hypothèse nulle ; l est la log-vraisemblance du modèle saturé.

Sous H_0, on a :

$$LR \xrightarrow{\infty} \chi^2(k) \tag{2.23}$$

On rappelle que k est le nombre de variables explicatives véritables sans la constante. Pour un niveau de confiance donné, on lit la valeur critique associée à la loi du khi-deux à k degrés de liberté χ_α^{2*}. Si $LR < \chi_\alpha^{2*}$ alors on accepte l'hypothèse H_0, c'est-à-dire les variables explicatives du modèle n'apportent pas grande chose dans l'explication du phénomène. Dans le cas contraire, on conclut que les variables sont globalement significatives, c'est-à-dire qu'il existe au moins une qui apporte une information significative dans l'interprétation du modèle.

6.2 Les pseudo-R^2

Plusieurs auteurs ont proposé des pseudo-R^2 pour les modèles qualitatifs pour juger la qualité de l'ajustement du modèle aux données, avec l'idée d'en faire des équivalents du coefficient de détermination R^2 du modèle linéaire classique. On les appelle des pseudo-R^2, car ils ne s'interprètent pas en termes de rapport de variances, comme dans le cas du modèle linéaire. Néanmoins, ils permettent d'évaluer le pouvoir prédictif du modèle. Une valeur proche de 1 indique que le pouvoir prédictif du modèle est acceptable.

a) Le R^2 de Mc-Fadden

On le définit par :

$$R^2_{Mc} = 1 - \frac{LogL}{LogL_o} \tag{2.24}$$

Comme $L \geq L_o$ (la vraisemblance d'un modèle libre est toujours supérieure à celle du modèle contraint), alors $R^2_{Mc} \in [0,1]$.

Remarque : Sous EViews, le R^2 de Mc-Fadden n'est pas calculé lorsque le modèle est spécifié sans constante.

b) Le R^2 de Cragg et Uhler

On le définit par :

$$R^2_{CU} = \frac{L^{1/n} - L_0^{2/n}}{(1 - L_0^{2/n})L^{2/n}} \tag{2.25}$$

c) Le R^2 d'Efron

Il est définit par :

$$R^2_E = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.26}$$

avec $\hat{y}_i = F(x'_i \hat{a})$.

d) Le R^2 de Count

Ce coefficient est définit par :

$$R^2_{Count} = \frac{1}{n}\sum_j n_{jj} \tag{2.27}$$

où n_{jj} le nombre d'individus biens classés.

e) Le R^2 de Count ajusté

On le définit par :

$$R^2_{AC} = \frac{\sum_j n_{jj} - \max(n_{1+}, n_{2+})}{n - \max(n_{1+}, n_{2+})} \tag{2.28}$$

où n_{j+} est le nombre d'individus possédant la modalité *j* dans la base de départ.

6.2 Test d 'adéquation de Hosmer Lemeshow

Le test de Hosmer et Lemeshow teste l'adéquation des probabilités calculées aux probabilités théoriques (inobservables) de l'événement $y = 1$. Il est basé sur un regroupement des probabilités prédites par le modèle en *J* groupes, déciles par exemple. On calcule, ensuite, pour chacun des groupes le nombre observé de réponses positives *y* = 1 et négatives *y* = 0, que l'on compare au nombre espéré prédit par le modèle. On calcule alors une distance entre les effectifs observés et les effectifs « espérés » au moyen d'une statistique du Chi-deux. Lorsque cette distance est petite on considère que le modèle est bien calibré

Soit n_j^1 le nombre d'individus qui présentent effectivement la valeur $y = 1$ dans la classe j.
Pour chaque classe j on calcule la probabilité moyenne de $y = 1$, par :

$$\bar{p}_j = \frac{1}{n_j} \sum_{i \in j} \hat{p}_{ij} \tag{2.29}$$

Si les probabilités sont correctement évaluées, la statistique de Hosmer et Lemeshow est définie par:

$$HL = \sum_{j=1}^{J} \frac{\left(n_j - n_j \bar{p}_j\right)^2}{n_j \bar{p}_j (1 - \bar{p}_j)} \sim \chi^2(J-2) \tag{2.30}$$

On doit noter que $n_j \bar{p}_j$ est l'espérance calculée du nombre d'individus présentant la modalité $y = 1$ dans la classe j et que $n_j \bar{p}_j (1 - \bar{p}_j)$ est la variance calculée de n_j^1.

6.3 Indicateurs de « prédictions » correctes, spécificité et sensibilité

On peut également juger de la qualité du modèle en évaluant son aptitude à reproduire les valeurs effectivement observées de Y sur l'échantillon qui a servi à l'estimation des coefficients. Pour cela, on doit convenir d'un seuil au-delà duquel la valeur calculée de y_i^* se concrétiserait par une valeur prédite de y_i égale à 1. On peut, par exemple, convenir d'un seuil égal à 50 % (quoique ce seuil soit totalement arbitraire) et retenir la règle suivante :

$$\hat{y}_i = 1 \text{ si } F(x'_i \hat{a}) > 0{,}5 \tag{2.31}$$

Le choix d'un seuil égal à 0,5 a ses désavantages. En particulier, il attribuera le même résultat à deux individus ayant l'un une probabilité estimée de 0,45 et l'autre une probabilité de 0,001. Dans certains cas, on peut être amené à réviser ce seuil en fonction de l'évènement étudié (cas d'échantillons déséquilibrés ou des phénomènes rares par exemple).

On peut construire une matrice de confusion indiquant les réalisations et les prédictions.

Réalisation (y)	**Prédiction ($\hat{y}$)**		**Total**
	1	**0**	
1	n_{11}	n_{10}	$n_{1.}$
0	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	n

On définit ainsi le taux de prédictions correctes global (TPC) et le taux de prédictions correctes de chacun des événements $(TPC(1), TPC(0))$.

$$TPC = \frac{n_{11} + n_{00}}{n} \times 100 \tag{2.32}$$

$$TPC(1) = \frac{n_{11}}{n_{1.}} \times 100 \tag{2.33}$$

$$TPC(0) = \frac{n_{00}}{n_{0.}} \times 100 \tag{2.34}$$

En général, le taux global de prédictions correctes offre une mesure optimiste de la qualité prédictive du modèle. La sensibilité est définie comme la probabilité de bien classer un individu de la catégorie $y = 1$, c'est-à-dire la probabilité de classer l'individu dans la catégorie $y = 1$ étant donné qu'il est effectivement observé dans celle-ci :

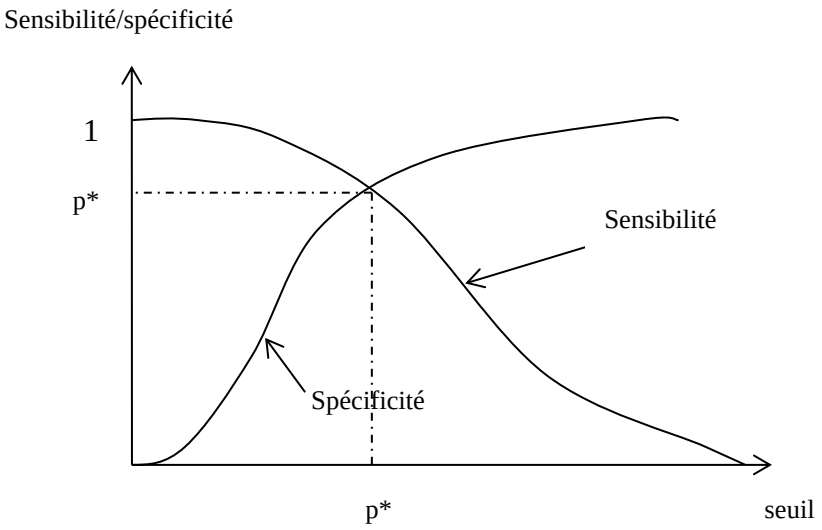
$$Sensibilité = \Pr(\hat{y} = 1 / y = 1) \tag{2.35}$$

La spécificité est définie comme la probabilité de bien classer un individu de la catégorie $y = 0$, c'est-à-dire la probabilité de classer l'individu dans la catégorie $y = 0$ étant donné qu'il est effectivement observé dans celle-ci :

$$Spécificité = \Pr(\hat{y} = 0 / y = 0) \tag{2.36}$$

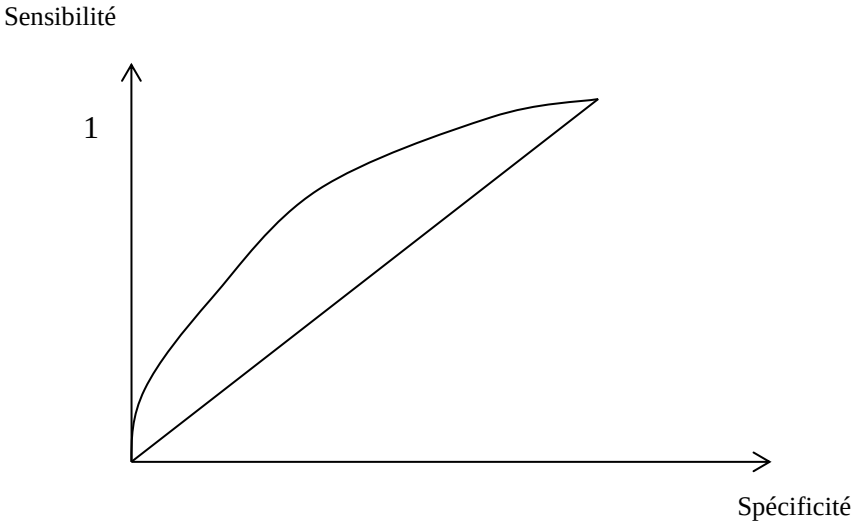
Les taux $TPC(1)$ et $TPC(0)$ fournissent respectivement une estimation de la **sensibilité** et de la **spécificité**.

Lorsque le seuil varie, le taux global de cas bien classés, la sensibilité et la spécificité changent, puisque le classement est modifié. Afin de représenter les valeurs pour toutes les possibilités de seuil, on dessine sur un graphique des courbes de sensibilité et de spécificité.



En fixant un seuil p^*, on obtient un classement avec une sensibilité et une spécificité égales à p^*.

Comme indicateur de la capacité du modèle à discriminer, on utilise la courbe ROC (Receiving Operating Curve) qui indique la sensibilité en fonction de la spécificité. La courbe ROC se présente comme ci-dessous :



La surface sous cette courbe permet d'évaluer la capacité du modèle à discriminer entre *y=1* et *y=0*.

ROC=0,5	→	pas de discrimination
0,7≤ROC<0,8	→	discrimination acceptable
0,8≤ROC<0,9	→	très bonne discrimination
0,9≤ROC	→	Discrimination exceptionnelle

7. Détection des outliers et des observations influentes

Une observation est dite outlier lorsqu'elle ne suit pas le mouvement général des autres observations de la série. La détection des outliers se fait via le résidu de Pearson défini par :

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \tag{2.37}$$

avec $\hat{\pi}_i = P(y_i | x_i, \hat{a})$.

On peut aussi utiliser celui de Pearson standardisé qui est défini par :

$$r_i^{std} = \frac{r_i}{1-h_{ii}} \tag{2.38}$$

avec $h_{ii} = \hat{\pi}_i(1-\hat{\pi}_i)x_i \operatorname{var}(\hat{a})x_i^{'}$

Une observation sera déclarée outlier lorsque la valeur absolue de ces résidus est plus grande que 2.

La quantité h_{ii} est appelé « puissance » (leverage) et permet d'identifier les observations avec une valeur extrême sur une variable explicative. Ces observations sont appelées points avec une puissance élevée (high leverage)[1].

[1] La puissance est une mesure de la distance avec laquelle une variable indépendante dévie de son point moyen. Ces points de puissance peuvent avoir un effet sur l'estimation des coefficients de régression.

Au-delà des outliers, il existe parfois dans la base de données des observations qui ont des influences significatives sur les coefficients de la régression, c'est-à-dire qui peuvent changer aussi bien le signe que les coefficients lorsque celles-ci sont retirées. La détection de ces observations peut être faite en utilisant la distance de Cook et les Dbeta.
La distance de Cook est donnée par l'expression :

$$C_i = \frac{r_i^2 h_{ii}}{(1-h_{ii})^2} \tag{2.39}$$

Au regard de ce critère, une observation sera suspecte si $C_i > \frac{4}{n}$.

Le tableau synthétique ci-après fournit pour chaque indicateur le seuil tolérable :

Indicateurs	Seuils
abs(r_i)	2
abs(r_i^{std})	2
h_{ii}	2*k/n
C	4/n

Dans la pratique, ces indicateurs doivent être combinés afin de produire des résultats intéressants au risque d'exclure toutes les observations de la base de données. En fonction du nuage des résidus, il est possible de modifier un temps soit peu les seuils prédéfinis pour les mêmes raisons que précédemment.

8. Tests de significativité des coefficients

Pour tester la significativité des coefficients, trois statistiques de test sont généralement utilisées: le test de Wald, le test du rapport de vraisemblance (LR test) et le test du multiplicateur de Lagrange (LM test). Ces trois statistiques de test sont utilisées pour tester plus généralement des restrictions sur les coefficients des modèles. Elles sont asymptotiquement équivalentes, mais elles ont des comportements différents en petits échantillons. Nous allons développer les deux premières statistiques de test.

8.1 Test de Wald

Dans le cas où l'on veut tester la significativité d'un seul coefficient, la statistique de Wald est définie à partir de la statistique :

$$z = \frac{\hat{a}_k}{\hat{\sigma}_k} \rightarrow N(0,1) \tag{2.40}$$

On a donc que :

$$W = z^2 \rightarrow \chi^2(1) \tag{2.41}$$

Si w_α^* désigne la valeur critique au seuil α d'un Khi-deux à 1 degré de liberté, alors la stratégie de test est la suivante:

- Si $W < w_\alpha^*$: on accepte l'hypothèse que le coefficient a_k n'est pas significativement différent de zéro. En d'autres termes, la variable correspondante x_k n'est pas significative dans l'explication du phénomène étudié.

- Si $W \geq w_\alpha^*$: on accepte que la variable x_k est significative pour le modèle spécifié.

8.2 Test du rapport de vraisemblance

Pour tester la significativité du coefficient a_j à l'aide du test du rapport de vraisemblance, on compare le modèle saturé au modèle estimé en enlevant la variable x_j de la liste des variables explicatives. La statistique de ce test est:

$$\lambda_j = -2(l_0 - l) \rightarrow \chi^2(1) \qquad (2.42)$$

où l_o est la log-vraisemblance du modèle estimé sans la variable explicative x_j, c'est-à-dire sous l'hypothèse $H_0 : a_j = 0$, et l est la log-vraisemblance du modèle sous l'hypothèse alternative.

Le test du rapport de vraisemblance est plus performant que le test de Wald. Dans certains cas, le dernier peut accepter l'hypothèse nulle alors que le coefficient en question est bien significatif.

On peut également utiliser un test du rapport de vraisemblance pour tester la significativité de plusieurs coefficients du modèle. Le principe du test reste toujours le même : on estime le modèle sous les deux hypothèses et on calcule la statistique du rapport de vraisemblance.

9. Interprétation des coefficients et calcul des effets marginaux

Nous avons vu que dans les modèles Probit et Logit, les paramètres du modèle ne sont identifiés qu'à une constante multiplicative près. Ils ne peuvent être identifiés sans imposer des restrictions sur la moyenne et la variance du terme d'erreur. Toutefois, les conditions d'identification n'affectent pas la probabilité de l'événement. Il ne faut pas également perde de vue que les coefficients estimés reflètent la relation entre les variables explicatives et la variable latente. Par conséquent, l'ordre de grandeur des coefficients n'a, en lui-même que peu d'importance. Les seules informations vraiment directement interprétables sont les signes et les valeurs relatives des coefficients. Le signe d'un coefficient indiquera si la variable explicative associée influence la probabilité de l'événement à la hausse ou à la baisse. Un coefficient a_j positif signifie qu'un accroissement de x_j joue dans le sens d'une plus grande probabilité d'observer l'évènement $y = 1$.

9.1 Calcul des effets marginaux

En pratique, on se sert des effets marginaux pour étudier l'effet d'une variable explicative sur la probabilité de l'événement étudié. L'effet marginal d'une variable x_j est la dérivée de la probabilité estimée par rapport à cette variable :

$$Em(x_j) = \frac{\partial F(x'\hat{a})}{\partial x_j} = f(x'\hat{a}) \times \hat{a}_j \tag{2.43}$$

Pour un Logit :

$$Em(x_j)\frac{\partial F(x'\hat{a})}{\partial x_j} = \frac{e^{x'a}}{\left(1+e^{x'a}\right)^2} \times \hat{a}_j \tag{2.44}$$

Pour un Probit :

$$Em(x_j) = \frac{\partial \Phi(x'\hat{a})}{\partial x_j} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x'a)^2} \times \hat{a}_j \tag{2.45}$$

Dans le modèle linéaire l'effet marginal des variables explicatives sur la probabilité de réalisation de l'événement $y = 1$ est constant. Au contraire, ici cet effet marginal varie en fonction du point à partir duquel il est apprécié. Néanmoins, le signe de l'effet marginal est celui du coefficient.

Cependant on peut évaluer un effet marginal synthétique dans l'échantillon, qui renseignerait sur l'impact moyen d'une variation unitaire de la variable explicative. Deux solutions sont alors envisageables. On peut calculer l'effet marginal moyen en remplaçant les valeurs individuelles x_i par leurs moyennes empiriques calculées sur toutes les observations. Ce qui donne :

$$\overline{Em}(x_j) = \frac{\partial F(\bar{x}'\hat{a})}{\partial x_j} = f(\bar{x}'\hat{a}) \times \hat{a}_j \tag{2.46}$$

On peut également considérer la moyenne des effets marginaux individuels.

$$\overline{Em}(x_j) = \frac{1}{n}\sum_i \frac{\partial F(x'\hat{a})}{\partial x_j} \tag{2.47}$$

Pour une variable qualitative binaire *s*, l'effet marginal s'obtient en faisant la différence des probabilités:

$$Em(s) = P(y = 1/x^*, s = 1) - P(y = 1/x^*, s = 0) \tag{2.48}$$

On parle dans ce cas d'effet discret.

On peut également calculer une élasticité qui présente l'avantage par rapport à l'effet marginal d'être indépendante de l'unité de mesure de la variable explicative:

$$\varepsilon_j = \frac{\partial F(x'\hat{a})}{\partial x_j} \cdot \frac{x_j}{F(x'\hat{a})} = \hat{a}_j x_j \frac{f(x'\hat{a})}{F(x'\hat{a})} \tag{2.48}$$

Cette expression prend une forme simplifiée dans le cas du modèle Logit:

$$\varepsilon_j = \hat{a}_j x_j \frac{1}{1+e^{x'\hat{a}}} \tag{2.49}$$

9.2 Comparaison des estimations des modèles Logit et Probit

Les restrictions sur les moyennes et les variances des erreurs permettent certes d'identifier les modèles; en revanche, ces restrictions d'identification rendent les valeurs numériques des paramètres arbitraires. En effet, la différence de variance (1 pour le Probit et $\pi^2/3$ pour le Logit) implique une différence dans les valeurs numériques des coefficients estimés Logit et Probit. Les coefficients Logit et Probit ne peuvent être comparés directement qu'à la condition de prendre la précaution de pré-multiplier les coefficients Probit par $\pi/\sqrt{3}$ (ou de diviser les coefficients Logit par $\pi/\sqrt{3}$). Autrement dit, si on prend en compte la différence de variance, on a l'approximation :

$$\hat{a}_{\text{logit}} = \frac{\pi}{\sqrt{3}} \hat{a}_{\text{Probit}} \cong 1{,}8 \times \hat{a}_{\text{Probit}} \tag{2.50}$$

Les résultats des modèles Probit et Logit sont généralement similaires si l'on tient compte des problèmes de normalisation. Toutefois, il convient d'être prudent dans l'utilisation des approximations pour comparer ces deux modèles. Il est toujours préférable de raisonner en termes de probabilité et non en termes d'estimation des coefficients pour comparer ces résultats (Amemiya, 1981).

9.3 Odds-Ratio

Les coefficients du modèle Logit ont une interprétation intéressante qui justifie son utilisation intensive en épidémiologie. On a en effet :

$$\ln\left[\frac{\Pr ob(y_i = 1/x_i)}{\Pr ob(y_i = 0/x_i)}\right] = \ln\left(\frac{p_i}{1-p_i}\right) = a_0 + a_1 x_{1i} + \ldots + a_k x_{ki} \tag{2.51}$$

Les coefficients a_j sont les effets marginaux des variables explicatives sur le logarithme du rapport des côtes $p_1/1-p_1$. L'équation (2.51) est appelée transformation Logit et est notée $\log it(P(y=1/x)$. Si on pose $c_i = p_i/1-p_i$, on interprète ce rapport en disant qu'il y a c_i fois plus de chance que l'événement $y_i = 1$ se réalise qu'il ne se réalise pas.

On définit le Odds Ratio (OR) associé à une variable x_j par :

$$OR_{x_j} = \frac{\dfrac{P(y_i = 1/x_{ij} = 1)}{1-P(y_i = 1/x_{ij} = 1)}}{\dfrac{P(y_i = 1/x_{ij} = 0)}{1-P(y_i = 1/x_{ij} = 0)}} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \tag{2.52}$$

où p_1 représente la probabilité que $y = 1$ pour un individu pour lequel $x_j = 1$ et p_0 celle pour un individu pour lequel $x_j = 0$.

Un Odds Ratio de 1 signifie que la probabilité de $y=1$ est la même chez les individus $x_j=1$ que chez ceux $x_j=0$. Autrement dit, la réalisation de $y=1$ n'est pas associée à x_j. En revanche, un Odds Ratio différent de 1 signifie qu'il y a une association entre l'événement $y=1$ et la variable x_j. Si Odds Ratio est >1, cela signifie que le numérateur est plus grand que le dénominateur et, par conséquent, que les individus $x_j=1$ ont une plus grande occurrence de l'événement $y=1$ que ceux $x_j=0$. C'est le contraire s'il est <1.

En utilisant l'expression (2.52) dans le cadre d'un modèle Logit, on a :

$$OR_{x_j}=\frac{e^{a_0+\sum_{l\neq j}^{k}a_l x_l+a_j}}{e^{a_0+\sum_{l\neq j}^{k}a_l x_l}}=e^{a_j} \tag{2.53}$$

L'exponentiel du coefficient d'une variable explicative dichotomique s'interprète comme son Odds Ratio (l'Odds Ratio (OR) associé au passage de la catégorie de référence $x_j=0$ à la catégorie $x_j=1$).

Lorsque la variable explicative est continue, on calcule un Odds Ratio associé à un accroissement unitaire :

$$OR_{x_j}=\frac{e^{a_0+\sum_{l\neq j}^{k}a_l x_l+a_j(x_j+1)}}{e^{a_0+\sum_{l=1}^{k}a_l x_l}}=e^{a_j} \tag{2.54}$$

On notera que l'Odds Ratio dépend de l'unité de mesure de la variable.

Exemple : On observe un échantillon de 170 candidats à un concours d'entrée dans une grande école. On s'intéresse à l'association entre l'option et l'admission.

	Echec		
Option	Oui (1)	Non (0)	Total
Eco	17	73	90
Maths	46	115	161
Total	103	148	251

Risque d'échec chez les Eco=17/90=0,63.
Risque d'échec chez les Math=46/161=0,28.

Risque relatif RR=0,63/0,28=2,21 : le risque d'échec est 2,21 fois plus élevé chez les Eco que chez les Math.

Odds chez les Eco=0,63/(1-0,63)=1,73 : les Eco ont 1,74 fois plus de risque d'échouer que de réussir.

Odds chez les Math=0,28/(1-0,28)=0,4 : les Math ont 2,5 fois plus de chances de réussir que d'échouer au concours.

Odds-ratio OR=0,63/(1-0,63)/ 0,28/(1-0,28)= 1,73 /0,4=4,31.

10. Critères de comparaison de plusieurs modèles

Il n'existe pas de critère absolu permettant de comparer plusieurs modèles alternatifs estimés à partir d'un même échantillon. Selon le critère retenu, un modèle peut sembler plus performant qu'un autre et moins performant pour un autre critère. Néanmoins, il existe un certain nombre de critères statistiques permettant de juger de la performance des modèles, le critère le plus discriminant dépend de l'objectif assigné au modèle. Nous présentons ci-après les deux groupes de critères les plus utilisés.

10.1 Taux de bonnes prédictions

On peut comparer les performances de deux modèles en comparant leur pouvoir prédictif, c'est-à-dire leur capacité à classer correctement les observations. Pour cela, il faut définir une stratégie de prédiction ou d'affectation sous la forme :

On décide que $y_i = 1$ quand $\hat{p}_i \geq \overline{p}$ et $y_i = 0$ sinon.

A partir de cette stratégie, on construit la table de vérité croisant les prédictions et les observations réelles de la variable *y*. De cette façon, on calcule le pourcentage d'observations bien prédites, qui fournit un critère de performance du modèle. Toutefois, ce critère est trop optimiste pour trancher de façon pertinente entre deux modèles concurrents.

10.2 Critères d'information

Plusieurs critères statistiques sont couramment utilisés dans les logiciels d'économétrie. Ces critères fournissent une mesure de la quantité d'information donnée par le modèle. Il s'agit notamment du :

- Critère d'Akaike :

$$AIC = -2\log(L)/n + 2k/n \qquad (2.55)$$

- Critère de Schwarz :

$$SC = -2\log(L)/n + k\log(n)/n \qquad (2.56)$$

- Critère d'Hannan-Quinn:

$$HQ = -2Log(L)/n + 2k\log(Log(n))/n \qquad (2.57)$$

- Critère d'Information Bayésien :

$$BIC = (2\log(L) - 2\log(L_0)) - ddl * \ln(n) \qquad (2.58)$$

- Critère d'Information Bayésien Modifié :

$$BIC^{'} = (2\log(L) - 2\log(L_0)) - k * \ln(n) \tag{2.59}$$

Le ''meilleur modèle'' est celui qui fournit un critère minimal.

11. Test d'hétéroscédasticité résiduelle

Le test d'hétéroscédatsicité est très important dans les modèles de choix dichotomique. En effet, l'hétéroscédasticité biaise l'estimation de la matrice de variance-covariance, ainsi que les tests statitiques, car en présénce d'hétéroscédatsicité les estimateurs ne sont pas asymptotiquement efficaces.

Pour tester l'hypothèse d' hétéroscédatsicité résiduelle, on considère la formulation générale due à Harvey (1976):

$$\sigma_i^2 = e^{2z'_i\eta} \tag{2.60}$$

où z'_i est un vecteur de variables de dimension $(1 \times g)$. Cette spécification est compatible avec le modèle Probit seulement. L'hypothèse nulle d'homoscédatsicité est $H_0 : \eta = 0$. Le vecteur z'_i ne contient pas de terme contant. Pour tester cette hypothèse, on peut utiliser le test du rapport de vraisemblance ou celui du multiplicateur de Lagrange.

11.1 Test du rapport de vraisemblance

La log-vraisemblance du modèle hétéroscédastique est :

$$\ln L = \sum_{i=1}^{n} y_i \ln \Phi\left[\frac{x'_i a}{\exp(z'_i \eta)}\right] + (1 - y_i)\ln\left[1 - \Phi\left(\frac{x'_i a}{\exp(z'_i \eta)}\right)\right] \tag{2.61}$$

Les conditions de premier ordre s'écrivent:

$$\frac{\partial \ln L}{\partial a} = \sum_{i=1}^{n}\left[\frac{\phi_i(y_i - \Phi_i)}{\Phi_i(1-\Phi_i)}\right] e^{-z'_i\eta} x'_i = 0 \tag{2.62}$$

$$\frac{\partial \ln L}{\partial \eta} = \sum_{i=1}^{n}\left[\frac{\phi_i(y_i - \Phi_i)}{\Phi_i(1-\Phi_i)}\right] e^{-z'_i\eta} z'_i \left(-x'_i a\right) = 0 \tag{2.63}$$

On évalue les log-vraisemblances du modèle libre et du modèle contraint. La statistique du rapport de vraisemblance est définie par :

$$LR = -2\left(\ln L_0 - \ln L\right) \rightarrow \chi^2(g) \tag{2.64}$$

11.2 Test du multiplicateur de Lagrange

Le test du multiplicateur de Lagrange est basé sur les conditions de premier ordre du programme de maximisation de la fonction de vraisemblance sous l'hypothèse alternative. On vérifie si ces conditions sont violées lorsqu'on se situe sous l'hypothèse nulle. Autrement dit, dans l'hypothèse nulle $H_0 : \eta = 0$, on doit vérifier la condition suivante :

$$\frac{\partial \ln L}{\partial \eta} = \sum_{i=1}^{n} \left[\frac{\phi_i (y_i - \Phi_i)}{\Phi_i (1 - \Phi_i)} \right] z'_i \left(- x'_i a \right) = 0 \tag{2.65}$$

Cette condition implique l'orthogonalité de $z_i^{'}(-x_i^{'}a)$ avec le résidu normalisé du modèle.

Pratiquement, ce test s'effectue simplement à partir de la régression du modèle artificiel suivant:

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i (1 - \hat{p}_i)}} = \frac{\phi(x'_i \hat{a})}{\sqrt{\hat{p}_i (1 - \hat{p}_i)}} x'_i \beta + \frac{\phi(x'_i \hat{a}) x'_i \hat{a}}{\sqrt{\hat{p}_i (1 - \hat{p}_i)}} z'_i \gamma + e_i \tag{2.66}$$

La variable expliquée de l'équation de test est un résidu normalisé. $\hat{a}$ et $\hat{p}_i = \Phi(x'_i \hat{a})$ sont estimées sous l'hypothèse nulle. La statistique de test est égale à la somme des carrés des $\hat{n}$:

$$LM = \sum_{i=1}^{n} \hat{r}_i^2 = n \overset{l}{R}^2 \rightarrow \chi^2(g) \tag{2.67}$$

11.3 Calcul des effets marginaux en présence d'hétéroscédasticité

En présence d'hétéroscédasticité, l'effet marginal d'une variable w_k pouvant être dans *x* ou *z* est :

$$EM(w_k) = \frac{\partial \Pr ob(y = 1)}{\partial w_k} = \phi \left[\frac{x' \hat{a}}{\exp(z' \hat{\eta})} \right] \frac{\hat{a}_k - (x' \hat{a}) \hat{\eta}_k}{\exp(z' \hat{\eta})} \tag{2.68}$$

Si w_k apparaît seulement dans *x* alors l'effet marginal se réduit à:

$$EM(w_k) = \phi \left[\frac{x' \hat{a}}{\exp(z' \hat{\eta})} \right] \frac{\hat{a}_k}{\exp(z' \hat{\eta})} \tag{2.69}$$

Si w_k apparaît seulement dans *z* alors l'effet marginal se réduit à:

$$EM(w_k) = -\phi \left[\frac{x' \hat{a}}{\exp(z' \hat{\eta})} \right] \frac{(x' \hat{a}) \hat{\eta}_k}{\exp(z' \hat{\eta})} \tag{2.70}$$

12. Variables explicatives polytomiques

Dans les applications, il est souvent fréquent que des variables qualitatives figurent parmi les variables explicatives dans les modèles de regression. Etant donné que les codes associés à ces variables sont arbitraires, le codage ne servant qu'à reperer les catégories et n'a pas de sens numérique, il est conseillé d'introduire une variable indicatrice ou binaire par modalité. Par exemple, pour une variable comme la catégorie socio-professionnelle (CSP) ayant 3 modalités codées 1, 2 et 3, on définit 3 variables indicatrices de la façon suivante: CSP1=1 si CSP=1, 0 sinon; CSP2=1 si CSP=2, 0 sinon; CSP3=1 si CSP=3, 0 sinon.

Cependant, on ne gardera pas ces trois variables dans le modèle, car elles ne sont pas linéairement indépendantes. En effet, chaque individu a une et une seule CSP, donc CSP1+CSP2+CSP3=1. Il y a un problème de multicolinéarité si le modèle contient une constante.

Il est suggéré de supprimer une des 3 variables indicatrices. La modalité répresentant la situation la plus courante sert de modalité de référence et on supprime la variable correspondante. Cela revient à dire que son coefficient est nul. Dans le choix de la variable à inclure on peut comparer les modèles obtenus avec les différents codages et retenir celui qui a la plus grande vraisemblance. Dans tous les cas, l'interprétation des coefficients se fait par rapport à la modalité de référence. Dans un modèle de regression sur variables binaires, l'ensemble des situations de référence est representé par la constante.

On peut utiliser un test du rapport de vraisemblance pour tester l'effet d'une variable polytomique sur la probabilité de réalisation de l'événement. Soi M le nombre de modalités de la variable. Si cette variable est remplacée dans le modèle par M - 1 variables binaires, alors tester l'effet de la variable polytomique revient à tester la nullité simultanée des M - 1 coefficients associés aux différentes modalités. La statistique de test du rapport de vraisemblance suit un Khi-deux à M - 1 degrés de liberté.

Lorsque la variable explicative est polytomique à modalités ordonnées, le choix de la modalité de référence est moins difficile. Dans ce cas, en effet, on a très souvent intérêt à prendre comme référence la modalité la plus basse. On peut alors commenter les coefficients comme s'il s'agissait de variables continues.

Chapitre 3

Modèles multinomiaux

1. Introduction

Les modèles multinomiaux sont des modèles où la variable dépendante est une variable qualitative à plusieurs modalités. Il existe trois grandes catégories de modèles multinomiaux qui se distinguent de la façon de modéliser le processus aléatoire ayant engendré les réalisations de la variable dépendante et/ou par le choix du codage de la variable.

- Les modèles ordonnés
- Les modèles non ordonnés
- Les modèles séquentiels

Dans la pratique, les modèles polytomiques non ordonnés sont les plus fréquents. Dans cette catégorie, on trouve notamment le modèle Logit multinomial et le modèle Logit conditionnel de McFadden, qui sont les modèles les plus utilisés et qui constituent une extension du Logit binaire.

Si ces modèles sont simples, ils posent toutefois un problème de cohérence en raison d'une propriété peu réaliste d'Indépendance des Alternatives non Pertinentes. C'est pourquoi des modèles alternatifs ont été développés comme le modèle Logit hiérarchisé ou le Probit multinomial. Ces derniers requièrent toutefois des techniques d'estimation relativement complexes.

2. Modèles polytomiques ordonnés

Dans les modèles polytomiques ordonnés, la variable dépendante est une variable qualitative ordonnée, c'est-à-dire dont les modalités peuvent être hiérarchisées comme dans les exemples suivants :

$$y = \begin{cases} 1\, aucun \quad niveau \\ 2\ primaire \\ 3\, sécondaire \\ 4\, \text{sup}\, érieur \end{cases} \qquad y = \begin{cases} 1\, pas \quad du\, tout \quad d'accord \\ 2\ pas \quad d'accord \\ 3\, d'accord \\ 4\ parfaitement \quad d'accord \end{cases}$$

2.1 Modélisation d'une variable polytomique ordonnée

On considère une variable dépendante ordonnée y prenant J modalités. Pour modéliser cette variable, on peut adopter une approche en termes de variable latente en posant que :

$$y_i^* = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + e_i = x'_i \beta + e_i \quad (3.1)$$

où les $x_1 \ \ldots \ x_k$ sont les variables susceptibles d'expliquer y^*. Comme dans le cas binomial, la modalité de Y dépendrait directement de la position de y^* par rapport à différents seuils:

$$y = \begin{cases} 1 \, si \quad y_i^* < c_1 \\ 2 \, si \quad c_1 \le y_i^* < c_2 \\ . \\ . \\ J \, si \quad y_i^* \succ c_{J-1} \end{cases} \quad (3.2)$$

On peut écrire de façon plus compacte :

$$y_i = j \ \text{ si } \ c_{j-1} \le y_i^* \le c_j \ \text{ avec } \ c_0 = -\infty \, \text{et} \ c_J = +\infty \quad (3.3)$$

Si on désigne par *F* la fonction de répartition du terme d'erreur, on a :

$$\Pr(y_i = 1) = \Pr(x'_i \beta + e_i \ \text{p} \ c_1) = F(c_1 - x'_i \beta) \quad (3.4)$$

$$\Pr(y_i = j) = \Pr(c_{j-1} \le x'_i \beta + e_i \le c_j) = F(c_j - x'_i \beta) - F(c_{j-1} - x'_i \beta), \ 2 \le j \le J-1 \quad (3.5)$$

$$\Pr ob(y_i = J) = \Pr ob(x'_i \beta + e_i \ \text{f} \ c_j) = 1 - F(c_{J-1} - x'_i \beta) \quad (3.6)$$

Selon que *F* est la fonction de répartition de la loi normale ou logistique, on a un modèle Probit ordonné ou un Logit ordonné.

2.2 Fonction de vraisemblance et estimation

La fonction de vraisemblance du modèle s'écrit :

$$L(y, \beta, c_1, ..., c_J) = \prod_{i=1}^{n} \prod_{j}^{J} \Pr ob(y_i = j)^{y_{ij}} = \prod_{i=1}^{n} \prod_{j}^{J} \left[F(c_j - x'_i \beta) - F(c_{j-1} - x'_i \beta) \right]^{y_{ij}} \quad (3.7)$$

où la variable y_{ij} est définit par $y_{ij} = 1 \, \text{si} \, y_i = j$.

Les coefficients du modèle sont estimés par la méthode du maximum de vraisemblance :

$$\hat{\beta} = Arg \underset{\beta}{Max} \log L(y, \beta, c_1, ..., c_J) \quad (3.8)$$

$$\hat{c}_j = Arg \underset{c_j}{Max} \log L(y, \beta, c_1, ..., c_J) \quad (3.9)$$

On rencontre ici les mêmes difficultés que celles qui ont déjà été évoquées dans le cas binaire.

- Il est impossible de dissocier l'estimation de la constante β_0 de celle des seuils c_1, … c_{J-1} .
- Dans le cas du modèle Probit ordonné, il est impossible de dissocier l'estimation des différents coefficients de celle de la variance de l'erreur (qu'on pose par convention égale à l'unité) : les coefficients estimés ne nous renseignent donc sur les valeurs théoriques de ceux-ci qu'à un facteur multiplicatif près. Seuls comptent les signes et les valeurs relatives de ces coefficients.

Toutefois, si les seuils c_j sont connus (discrétisation d'une variable continue par exemple), les paramètres β et σ sont identifiés dès lors que $J > 2$ car la variabilité dans la variable de seuil permet d'identifier $1/\sigma$ (dans un modèle dichotomique $J = 2$, il n'y a pas de variabilité des seuils).

2.3 Test de régression parallèle

Avant de discuter de l'interprétation d'un modèle ordonné, il est indispensable de comprendre et tester une hypothèse implicite de ce modèle connu sous les noms d'hypothèse de régression parallèle, pour le modèle Logit ordonné, et d'hypothèse de proportion des odds. Les équations (3.4) à (3.6) peuvent être utilisées pour la dérivation des probabilités cumulées qui s'écrivent sous la forme simplifiée par :

$$\Pr ob(y_i \le j) = F(c_j - x'_i \beta), \; 1 \le j \le J-1 \qquad (3.10)$$

Ces dernières équations montrent que le modèle de régression ordonné est équivalent à *J-1* régressions binaires sous l'hypothèse fondamentale que les coefficients estimés par rapport aux variables explicatives sont identiques dans chacune des équations. Par exemple, avec *J=4* et une seule variable explicative *x*, nous avons en contraignant l'ordonnée à l'origine à 0 :

$$\Pr ob(y_i \le 1) = F(c_1 - \beta x_i) \qquad (3.11)$$

$$\Pr ob(y_i \le 2) = F(c_2 - \beta x_i) \qquad (3.12)$$

$$\Pr ob(y_i \le 3) = F(c_3 - \beta x_i) \qquad (3.13)$$

Si nous représentons l'argument de F sur un graphique de dimension deux, nous avons que la pente β est identique (droite parallèle) au niveau de chacune des équations.

Cette hypothèse de constance des coefficients β est implicite mais doit être testée pour assurer la validité du modèle de régression ordinale. Le test est effectué en comparant les coefficients issus des *J-1* équations binaires obtenues ci-dessus. Ces équations seront réécrites en modifiant β de sorte à ce qu'ils varient d'une équation à une autre comme suit :

$$\Pr ob(y_i \le j) = F(c_j - x'_i \beta_j), \; 1 \le j \le J-1 \qquad (3.14)$$

L'hypothèse de régression parallèle implique l'égalité $\beta_1 = \beta_2 = \mathsf{K} = \beta_{J-1}$. L'hypothèse sera vérifiée à condition que les $\hat{\beta}_1 = \hat{\beta}_2 = \ldots = \hat{\beta}_{J-1}$ soient très proches. Le test commun utilisé pour

s'assurer que cette hypothèse est vérifiée est celui développé par Brant (1990). Toutefois, il existe un autre test basé sur le rapport de vraisemblance développé par Wolfe et Gould (1998).

Une violation de cette hypothèse entraîne la non-validité du modèle ordonnée sous la forme présentée ci-dessus. Il faut recourir soit aux modèles ordonnés généralisées soit aux modèles multinomiaux non ordonnés.

2.4 Interprétation des coefficients

Les coefficients ne sont pas directement interprétables. On doit calculer l'influence marginale des variables sur les probabilités en dérivant les probabilités conditionnelles. Les effets marginaux par rapport à une variable explicative x_k quelconque sont donnés par les formules suivantes:

$$\frac{\partial \Pr ob(y_i = 1/x_i)}{\partial x_{ik}} = -\beta_k f(c_1 - x'_i \beta) \tag{3.15}$$

$$\frac{\partial \Pr ob(y_i = j/x_i)}{\partial x_{ik}} = -\beta_k \left[f(c_j - x'_i \beta) - f(c_{j-1} - x'_i \beta) \right] \tag{3.16}$$

$$\frac{\partial \Pr ob(y_i = J/x_i)}{\partial x_{ik}} = \beta_k f(c_{J-1} - x'_i \beta) \tag{3.17}$$

A partir des effets marginaux individuels, on peut calculer des effets marginaux globaux sur l'échantillon, qui renseignent sur l'impact moyen des variables explicatives sur la probabilité des différents événements. Deux méthodes peuvent être utilisées. La première évalue l'effet moyen en prenant la moyenne simple des effets marginaux individuels :

$$\overline{em}_k = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \Pr ob(y_i = j/x_i)}{\partial x_{ik}} = \frac{-\beta_k}{n} \sum_{i=1}^{n} \left[f(c_j - x'_i \beta) - f(c_{j-1} - x'_i \beta) \right] \tag{3.18}$$

La deuxième méthode calcule l'effet marginal global au point moyen :

$$\overline{em}_k = \left. \frac{\partial \Pr ob(y = j/x)}{\partial x_k} \right|_{x=\bar{x}} = -\beta_k \left[f(c_j - \bar{x}'\beta) - f(c_{j-1} - \bar{x}'\beta) \right] \tag{3.19}$$

L'effet marginal de la variable x_k sur la probabilité d'avoir $y = J$ est de même signe que le coefficient β_k tandis qu'il est de signe opposé sur la probabilité d'avoir $y = 1$. Pour les modalités intermédiaires ($j = 2...J-1$), le signe de l'effet marginal n'est pas forcément celui du coefficient, la quantité entre crochets étant de signe indéterminé. On interprétera donc un coefficient β_k positif en disant que tout accroissement de x_k contribue à rendre plus probable la modalité la plus élevée de y. Un coefficient négatif signifie *a contrario* que tout accroissement de x_k contribue à tirer y vers ses modalités les plus faibles.

L'interprétation des coefficients d'un modèle ordonné est donc délicate, surtout pour les modalités intermédiaires. On doit pour ce type de modèle toujours calculer les effets marginaux et ne pas se contenter de présenter les coefficients estimés.

2.5 Pouvoir prédictif du modèle

On peut appliquer aux modèles ordonnés les mêmes calculs d'indicateurs de performance que ceux mis en œuvre pour les modèles binomiaux. Le R^2 de McFadden se calcule de la même manière. Quant à la table des bonnes prédictions, on peut considérer que la valeur de y_i prédite par le modèle est celle qui correspond à la probabilité la plus forte.

3. Modèles multinomiaux non ordonnés

3.1 Logit multinomial

3.1.1 Modélisation

Ce modèle généralise le modèle Logit binaire. On modélise plusieurs choix non ordonnés. Par exemple, le choix du mode de transport : le bus, transport public, voiture, autre (vélo, marche à pieds, etc.). L'ordre dans lequel sont rangées les différentes occurrences de Y est sans importance et ne doit pas affecter le calcul des probabilités de ces occurrences.

Soit y la variable dépendante prenant les modalités $1,2,...,J$. La probabilité d'occurrence d'une modalité j s'écrit :

$$\Pr(y_i = j / x_i) = \frac{e^{x'_i \beta_j}}{\sum_{j=1}^{J} e^{x'_i \beta_j}} \tag{3.20}$$

Les coefficients β dépendent de la catégorie à laquelle appartient l'individu. Pour chaque variable explicative, on estime autant de coefficients que de modalités de y, chacun mesurant l'effet de la variable sur l'appartenance à l'une des J modalités de y. On est cependant confronté à un problème d'identification : en remplaçant β_j par $\beta_j + \delta$, la probabilité ne change pas. Une infinité de valeurs de β_j sont donc possibles, qui conduisent à une même valeur de la probabilité. Pour résoudre ce problème, on doit imposer aux coefficients une condition d'identification. Celle qui est souvent retenue est d'imposer la nullité de tous les paramètres relatifs à une catégorie donnée, appelée *modalité de référence*. Le choix de cette catégorie de référence est arbitraire. Par exemple, si on décide que la modalité de référence correspond à $j = 1$ alors la condition d'identification $\beta_1 = 0$ implique que :

$$\beta_{01} = \beta_{11} = ... = \beta_{k1} = 0 \tag{3.21}$$

Avec cette condition identifiante, l'équation de probabilité devient :

$$\Pr(y_i = j / x_i) = \frac{e^{x'_i \beta_j}}{1 + \sum_{j=2}^{J} e^{x'_i \beta_j}}, \quad j \geq 2 \tag{3.22}$$

La probabilité que $y = 1$ ne sera pas modélisée car elle est connue à partir des autres probabilités. En conséquence, les coefficients ne peuvent être estimés que pour les J-1 modalités, sans la modalité de référence. La conséquence importante de cette contrainte d'identification est que le modèle mesure l'effet d'une variable explicative non sur la probabilité d'appartenir à une catégorie donnée, mais sur la probabilité d'appartenir à la catégorie plutôt qu'à la catégorie de référence, ou, plus précisément, sur le rapport entre la probabilité d'appartenir à la catégorie et la probabilité d'appartenir à la catégorie de référence. En effet, il est facile de montrer que :

$$\ln\left[\Pr(y_i = j) / \Pr ob(y_i = 1)\right] = x'_i (\beta_j - \beta_1) = \beta'_{0j} + \beta'_{1j} x_{1i} + ... + \beta'_{kj} x_{ki} \tag{3.23}$$

Ainsi l'interprétation des coefficients d'un modèle Logit multinomial se fait en termes d'écart au référentiel. Par exemple si $\beta_k > 0$, tout accroissement de x_k contribue à rendre plus probable le choix de la modalité j par rapport à celui de la modalité de référence.

Les coefficients β_{kj} sont obtenus par maximisation de la log-vraisemblance de l'échantillon d'estimation :

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{J} \delta_{ij} \left[x'_i \beta_j - \ln\left(1 + \sum_{j=2}^{J} e^{x'_i \beta_j}\right)\right] = \sum_{i=1}^{n} \sum_{j=2}^{J} \delta_{ij} x'_i \beta_j - \sum_{i=1}^{n} \ln\left(1 + \sum_{j=2}^{J} e^{x'_i \beta_j}\right) \tag{3.24}$$

où la variable δ_{ij} est définit par $\delta_{ij} = 1 \text{si } y_i = j$.

3.1.2 Interprétation des coefficients d'un Logit multinomial

L'interprétation des coefficients d'un modèle multinomial est délicate. En effet, si on calcule l'effet marginal d'une variation de x_k sur la probabilité que l'individu choisisse j (plutôt que 1), on obtient:

$$\frac{\partial \Pr(y = j / x)}{\partial x_k} = \Pr(y = j / x)\left[\hat{\beta}_{kj} - \sum_{h=1}^{J} \hat{\beta}_{kh} \Pr(y = h / x)\right] \tag{3.25}$$

où $\hat{\beta}_{kj}$ est la $k^{ième}$ composante de $\hat{\beta}_j$ associée à la variable explicative x_k.

Pour chaque variable x_k, on doit calculer J effets marginaux associés aux probabilités $p_{ij} = \Pr(y_i = j)$, $j = 1,..., J$.

On constate que l'effet marginal n'est pas de même signe que celui du coefficient. Il dépend des valeurs de tous les coefficients et non seulement de $\hat{\beta}_{kj}$. De plus, on note que la valeur de

l'effet marginal dépend du point à partir duquel on le mesure. Pour cette raison, on le calcule le plus souvent au point moyen.

Le problème d'interprétation des coefficients se complique lorsque la variable explicative est une variable qualitative polytomique, puisqu'il faut imposer, à la variable, une modalité de référence à laquelle toutes les autres modalités de la variable doivent être comparées. Dans ce cas, la lecture des résultats doit « gérer » deux références : la catégorie de référence de la variable dépendante et la modalité de référence de la variable explicative.

On peut calculer le rapport des probabilités comme suit:

$$\frac{\Pr(y_i = j)}{\Pr(y_i = l)} = e^{x'_i(\beta_j - \beta_l)} \tag{3.26}$$

Ce rapport est indépendant des autres modalités : le rapport des probabilités associées au choix entre deux modalités ne dépend pas des autres modalités. Ajouter ou supprimer une tierce modalité, ou bien modifier les caractéristiques d'une modalité déjà incluse, ne change pas le rapport entre ces probabilités. C'est ce qu'on qualifie de propriété d'indépendance des alternatives non pertinentes (IIA : Independance of Irrelevant Alternatives).

Remarque : Le modèle Logit multinomial est formellement équivalent à une analyse discriminante linéaire si toutes les variables explicatives sont continues et distribuées selon une loi normale multidimensionnelle de telle manière que les *J* lois conditionnelles à l'appartenance de l'individu à l'une des *J* classes ont la même variance (Amemiya, 1981 ; Maddala, 1983 ; Sautory et Vong, 1992; Bardos, 2001). On peut donc utiliser ce modèle pour répondre aux deux objectifs de l'analyse discriminante : trouver la fonction linéaire des variables individuelles qui sépare au mieux les classes (les catégories) ; affecter à une classe un nouvel individu dont on connaît seulement les caractéristiques.

3.1.3 Tests d'hypothèses sur les coefficients

Les principaux tests d'hypothèse examinés ici portent sur la nullité d'un ou plusieurs paramètres du modèle. A cet effet, on peut utiliser la statistique de Student ou celle du rapport de vraisemblance.

A) Significativité d'un coefficient

On veut tester la nullité du paramètre β associé à une variable x_j caractéristique de choix dans un Logit conditionnel, ou du paramètre β_j d'une variable individuelle, associé à la catégorie *j*, dans un Logit multinomial. Pour ce faire, on utilise la statistique de Student définit par le rapport de la valeur estimée du paramètre à son écart-type estimé.

B) Significativité de plusieurs coefficients

Si on veut tester la nullité simultanée de plusieurs paramètres, on utilise le test du rapport de vraisemblance. Ce test consiste à comparer la vraisemblance L_0 d'un modèle contraint à celle L_1 d'un modèle non contraint. La statistique de test $LR = -2(\ln L_0 - \ln L_1)$ suit asymptotiquement une loi du Khi-deux dont le nombre de degrés de liberté est égal à la

différence entre le nombre de paramètres du modèle non contraint et le nombre de paramètres du modèle contraint.

Le test du rapport de vraisemblance peut être utilisé après l'estimation d'un modèle Logit multinomial pour tester l'effet d'une variable explicative x_k sur l'appartenance à une quelconque des *J* catégories, c'est-à-dire si au moins un des paramètres $\beta_2, \beta_2, ..., \beta_J$ de la variable est non nul. Cela revient à tester la nullité des *J-1* coefficients: $\beta_2 = \beta_2 = ... = \beta_J = 0$.

Le principe du teste consiste à calculer la vraisemblance du modèle complet (L_1) et celle du modèle contraint (L_0) obtenu en supprimant la variable explicative x_k. La statistique du test $LR = -2(\ln L_0 - \ln L_1)$ suit asymptotiquement une loi du Khi-deux à $J-1$ degrés de liberté. Le rejet du modèle contraint signifie qu'un des paramètres au moins n'est pas nul : la variable x_k a bien un effet.

C) Test de l'hypothèse IIA

Certains auteurs ont montré que dans certaines occasions, l'hypothèse d'indépendance des alternatives non pertinentes est trop restrictive pour modéliser correctement les comportements des individus (voir l'exemple du « bus bleu/bus rouge » de McFadden (1973) repris dans Horowitz et Savin, 2001 ; et celui du métro dans Thomas, 2000).

La propriété d'IIA peut être testée. L'hypothèse nulle est celle d'IIA. L'idée du test proposé par Hausman est de comparer deux estimateurs des coefficients sous les hypothèses nulle et alternative. Les étapes de ce test sont les suivantes :

C.1 Test de Hausman

- On estime le modèle complet avec toutes les *J* alternatives :

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \cdot \\ \cdot \\ \hat{\beta}_J \end{pmatrix} \tag{3.27}$$

- On estime le modèle contraint en élevant la ou les alternatives concernées et en excluant les individus qui ont choisi ces modalités:

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_1 \\ \cdot \\ \cdot \\ \tilde{\beta}_L \end{pmatrix} \tag{3.28}$$

- On compare les valeurs des coefficients obtenues dans ces deux estimations. Si la propriété IIA est valide, elles doivent être proches. L'écart entre les deux ensembles de paramètres s'estime en calculant la statistique de test :

$$H = (\hat{\beta} - \tilde{\beta})'\left[\text{var}(\hat{\beta}) - \text{var}(\tilde{\beta})\right]^{-1}(\hat{\beta} - \tilde{\beta}) \sim \chi^2(k) \qquad (3.29)$$

où k est la dimension du vecteur de paramètres.

C.2 Test de Small-Hsiao

Comme Hausman, Small et Hsiao ont développé un test qui permet également de vérifier l'hypohèse IIA. Ce test procède comme suit :

- Diviser l'échantillon en deux (S_1 et S_2) de tailles à peu près égales.
- Estimer dans chacun des échantillons le modèle complet. On obtient $\hat{\beta}^{S_1}$ et $\hat{\beta}^{S_2}$ à partir desquels on calcule $\hat{\beta}^{S_1S_2} = \frac{1}{\sqrt{2}}\hat{\beta}^{S_1} + (1 - \frac{1}{\sqrt{2}})\hat{\beta}^{S_2}$.
- Estimer le modèle contraint dans l'échantillon 2 et obtenir $\hat{\beta}_c^{S_2}$.
- Calculer la statistique $SH_c = -2\left[L(\hat{\beta}^{S_1S_2}) - L(\hat{\beta}_c^{S_2})\right]$ qui suit une loi de Khi-deux admettant comme paramètres le nombre de paramètres dans le modèle contraint.

Remarque : Dans la pratique, les tests d'Hausman et Small-Hsiao peuvent donner des résultats contradictoires. Cheng et Long (2005) ont montré que :

- La puissance du test d'Hausman est faible même si la taille de l'échantillon atteint 1 000.
- Pour certains types de données, le test Small-Hsiao a une bonne puissance pour les échantillons de taille supérieure ou égale à 500. Pour d'autres échantillons, ce test a une faible puissance indépendamment de la taille de l'échantillon.

3.2 Modèle Logit multinomial conditionnel

3.2.1 Modélisation

Le Logit multinominal admet que les valeurs prises par les variables explicatives ne sont pas influencées par la nature du choix et que les probabilités attachées aux différentes modalités ne diffèrent les unes des autres que par le fait qu'à chaque modalité est attaché un jeu spécifique de coefficients. Il existe cependant une autre possibilité : considérer un vecteur de coefficients constants quel que soit l'individu et la modalité et autoriser les variables explicatives à dépendre des modalités. Cette possibilité est à la base du modèle Logit conditionnel de McFadden (1973).

En reprenant la démarche utilisée pour le logit multinomial et en remplaçant formellement $x'_i \beta_j$ par $x'_{ij} \beta$, le modèle conditionnel s'écrit :

$$\Pr(y_i = j) = \frac{e^{x'_{ij}\beta}}{\sum_{j=1}^{J} e^{x'_{ij}\beta}} = \frac{e^{x^{*'}_{ij}\beta}}{1 + \sum_{j=2}^{J} e^{x^{*'}_{ij}\beta}} = \frac{e^{\beta_0 + \beta_1 x^*_{1ij} + \ldots + \beta_k x^*_{kij}}}{1 + \sum_{j=2}^{J} e^{\beta_0 + \beta_1 x^*_{1ij} + \ldots + \beta_k x^*_{kij}}} \qquad (3.30)$$

avec $x^{*'}_{ij} = x'_{ij} - x'_{i1}$.

Parce que toutes les variables explicatives dépendent de chaque choix *j*, le problème d'identification rencontré avec le Logit multinomial ne se pose pas : il n'y a pas à imposer des contraintes sur β. Les coefficients β s'interprètent comme associés aux différences des variables de chaque modalité par rapport aux variables du cas de référence (modalité 1).

Prenons l'exemple du choix des modes de transport. On considère les modes suivants : bus (modalité 2), la voiture (modalité 3) et les autres modes de transport (modalité 1). Les variables explicatives sont exprimées en différences par rapport à leurs valeurs prises dans la modalité 0. Il s'agit par exemple du temps de transport moyen du domicile au lieu de travail pour le mode j, noté $t_{ij} = x_{1ij}$ et le coût au kilomètre de ce mode, noté $c_{ij} = x_{2ij}$. Si on suppose que ce sont les deux seules variables explicative, on a $\beta = (\beta_0, \beta_1, \beta_2)'$. La probabilité qu'un individu i caractérisé par des temps relatifs $(t_{i2,} t_{i3})$ et des coûts relatifs $(c_{i2,} c_{i3})$ choisisse le mode de transport $j = 2,3$ s'écrit :

$$\Pr(y_i = j) = \frac{e^{\beta_0 + \beta_1 t_{ij} + \beta_2 c_{ij}}}{1 + \sum_{j=2}^{3} e^{\beta_0 + \beta_1 t_{ij} + \beta_2 c_{ij}}} = \frac{e^{x_{ij}^{*'}\beta}}{1 + \sum_{j=2}^{3} e^{x_{ij}^{*'}\beta}} \tag{3.31}$$

avec $x_{ij}^{*'} = (1, t_{ij}, c_{ij})$.

La log-vraisemblance associée à un modèle Logit conditionnel s'écrit :

$$\log L = \sum_{i=1}^{n}\sum_{j=1}^{J} \delta_{ij} x'_i \beta_j - \sum_{i=1}^{n} \ln\left(\sum_{j=1}^{J} e^{x'_i \beta_j}\right) = \sum_{i=1}^{n}\sum_{j=2}^{J} \delta_{ij} x_{ij}^{*'} \beta_j - \sum_{i=1}^{n} \ln\left(1 + \sum_{j=2}^{J} e^{x_{ij}^{*'}\beta_j}\right) \tag{3.32}$$

On détermine le vecteur de coefficients β en maximisant cette fonction.

3.2.2 Interprétation des coefficients d'un Logit conditionnel

Effets marginaux

Les effets marginaux mesurent les variations de la probabilité de choisir la modalité *j* quand la variable explicative x_k « augmente » d'une unité. On peut calculer deux types d'effets marginaux.

Effet direct :

$$efmar_j = \frac{\partial \Pr(y = j)}{\partial x_{kj}} = P(y = j) \times (1 - P(y = j)) \times \hat{\beta}_k \tag{3.33}$$

Effet croisé :

$$efmar_{hj} = \frac{\partial \Pr(y = h)}{\partial x_{kj}} = -P(y = j) \times (P(y = h) \times \hat{\beta}_k, \; h \neq j \tag{3.34}$$

Les effets marginaux dépendent donc des valeurs des variables explicatives.

L'effet marginal direct (dérivée par rapport à x_j de $\Pr(y = j)$) est toujours du signe de β, tandis que l'effet marginal croisé est toujours du signe opposé à celui de β. Cette propriété est une conséquence directe de la forme fonctionnelle des probabilités. On vérifie de plus que :

$$\sum_{h=1}^{J} \frac{\partial \Pr(y = h)}{\partial x_j} = 0 \tag{3.35}$$

La variation d'une des probabilités à l'augmentation d'une caractéristique est compensée par les variations concomitantes des autres probabilités.

On peut calculer de la même manière les élasticités directes et croisées.

Elasticité directe :

$$e_j = \frac{\partial \Pr(y = j)}{\partial x_j} \times \frac{x_j}{\Pr(y = j)} = (1 - P(y = j)) \times x_j \hat{\beta} \tag{3.36}$$

Elasticité croisée :

$$e_{hj} = \frac{\partial \Pr(y = h)}{\partial x_j} \times \frac{x_j}{\Pr(y = h)} = -P(y = j) \times x_j \hat{\beta}, \; h \neq j \tag{3.37}$$

Les élasticités directes mesurent l'effet, sur la probabilité de choisir *j*, d'une augmentation de la caractéristique *x* de *j*. Les élasticités croisées mesurent les effets sur les probabilités des autres choix, d'une augmentation de la caractéristique *x* de *j*, elles décrivent les substitutions possibles entre *j* et *h* du fait de l'augmentation de x_j . On remarquera que les élasticités croisées ne dépendent pas de *h*. Elles sont les mêmes pour tous les choix autres que *j*.

Probabilité d'un événement virtuel

Le modèle Logit conditionnel permet d'estimer la probabilité associée à une modalité virtuelle de la façon suivante :

$$\Pr(y_i = j) = \frac{e^{x_{iJ+1}^{*'} \hat{\beta}}}{1 + \sum_{j=2}^{J} e^{x_{ij}^{*'} \hat{\beta}} + e^{x_{iJ+1}^{*'} \hat{\beta}}} \tag{3.38}$$

où $\hat{\beta}$ désigne un estimateur convergent de β obtenu sur la base des modalités existantes ; $x_{iJ+1}^{*'} = x_{iJ+1}^{'} - x_{i1}^{'}$ représente les caractéristiques exogènes de l'individu associées à la $J+1^{ième}$ modalité virtuelle.

Considérons l'exemple précédent sur les modes de transport et supposons maintenant qu'on cherche à évaluer la probabilité que la population adopte un nouveau mode de transport public (le métro par exemple), en plus de ceux déjà existants. Soient $\hat{t}_{i4}$ et $\hat{c}_{i4}$ les évaluations du temps de trajet et du coût au kilomètre du nouveau mode de transport. La probabilité qu'un individu i utilise le nouveau mode de transport (modalité 4) lorsque celui-ci sera effectivement mis en en service est:

$$\Pr(y_i = 4) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 t_{i4} + \hat{\beta}_2 c_{i4}}}{1 + \sum_{j=2}^{3} e^{\hat{\beta}_0 + \hat{\beta}_1 t_{ij} + \hat{\beta}_2 c_{ij}} + e^{\hat{\beta}_0 + \hat{\beta}_1 t_{i4} + \hat{\beta}_2 c_{i4}}} \tag{3.39}$$

On obtient une estimation de la probabilité que l'individu i choisisse le nouveau mode de transport plutôt que les autres modes de transport.

La propriété IIA

Le logit conditionnel partage avec le logit multinomial la propriété IIA que le rapport de deux probabilités de choix j et h est indépendant des autres modalités (de leur nombre, leurs configurations, etc) :

$$\frac{\Pr(y_i = j)}{\Pr(y_i = h)} = e^{(x'_{ij} - x'_{ih})\beta} = e^{\hat{\beta}_1 (x_{1ij} - x_{1ih}) + \ldots + \hat{\beta}_k (x_{kij} - x_{kih})} \tag{3.40}$$

On peut donc interpréter un coefficient comme une semi-élasticité : un coefficient positif signifie que tout accroissement du différentiel dans les variables explicatives contribue à accroître la probabilité de choisir la modalité j par rapport à la modalité h.

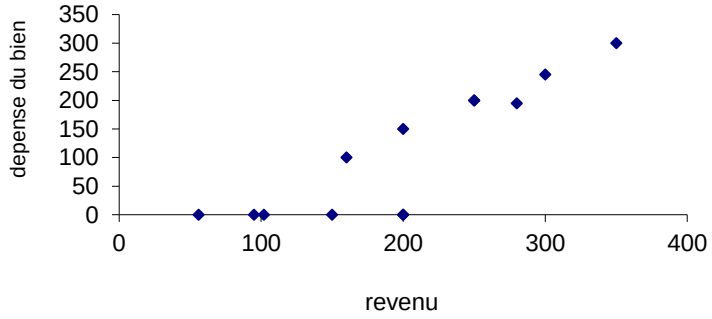
Cette propriété peut être testée en suivant la même procédure que dans le modèle Logit multinomial.

Chapitre 4

Modèles Tobit

1. Introduction

Les modèles de régression classiques supposent à travers l'hypothèse de normalité sur la distribution des termes d'erreur que la variable dépendante est une variable aléatoire continue. Par conséquent, elle ne peut prendre une ou plusieurs valeurs données avec une probabilité non nulle. Cependant, pour certains phénomènes économiques, cette hypothèse semble irréaliste dans la mesure où la variable dépendante est continue mais peut prendre des valeurs isolées avec des probabilités finies. Il s'agit en effet des modèles à variables dépendantes limitées. Dans ces modèles, la variable dépendante n'est observable que sur un certain intervalle. Par exemple, sur un échantillon aléatoire de ménages, on cherche à expliquer les dépenses d'un bien (par exemple le logement) en tenant compte du fait que, pour une partie de l'échantillon la dépense est nulle. Pour cet échantillon la valeur nulle est observée avec une probabilité différente de 0.



De tels échantillons sont appelés des échantillons censurés car la variable dépendante n'est observable que pour certains ménages (ménages locataires notamment). L'échantillon est dit tronqué lorsque pour une partie de l'échantillon les observations relatives à la variable dépendante et aux caractéristiques individuelles ne sont pas disponibles. Dans ce cas, l'échantillon tronqué n'est plus aléatoire et l'estimation utilisant cet échantillon pourrait donner des résultats biaisés. Le modèle censuré le plus simple est le modèle Tobit (Tobin's Probit) considéré comme une extension du modèle Probit permettant de traiter un certain nombre de situations.

2. Spécification du modèle Tobit simple

Reprenons l'exemple des dépenses d'un bien. Soit d_i la dépense du ménage i consacrée à ce bien. Pour un certain nombre de ménages on n'observe pas la dépense, par contre, pour d'autres on a $d_i > 0$. Soit x_i le vecteur des caractéristiques du ménage i. d_i est fonction de x_i à travers un modèle linéaire :

$$d_i = x'_i\, a + u_i \qquad (4.1)$$

La variable y_i observée est définie de la façon suivante:

$$y_i = d_i \text{ si } d_i > 0,\ y_i = 0 \text{ sinon} \tag{4.2}$$

Autrement dit, on reporte une dépense nulle pour les ménages n'ayant pas révélé un montant de dépense. On obtient de cette façon un modèle dit **censuré à gauche**. Le nuage de points sera mal décrit par la relation linéaire précédente puisqu'il contient deux parties nettement différentes (voir graphique ci-dessus).

Comment expliquer les variations des dépenses entre les différents ménages de l'échantillon alors que cette variable n'est positive que pour certains ménages? Les autres ménages ont une valeur nulle pour cette variable mais leurs caractéristiques sont néanmoins observées.

D'une façon générale, le modèle Tobit simple est spécifié sous la forme :

$$\begin{cases} y_i^* = x_i' a + u_i \qquad u_i \approx Niid(0, \sigma^2) \\ \\ y_i = \begin{cases} y_i^* \quad si \quad y_i^* > 0 \\ \\ 0 \quad si \quad y_i^* \le 0 \end{cases} \end{cases} \tag{4.3}$$

On suppose que la borne de censure $c = 0$, ce qui implique que si le vecteur x_i contient un terme constant, celui-ci se confond au seuil.

3. Vraisemblance du modèle Tobit et méthodes d'estimation

Pour estimer le modèle Tobit, la méthode utilisée est celle du maximum de vraisemblance. Pour écrire la fonction de vraisemblance du modèle, il faut remarquer que la distribution de la variable dépendante y_i est un mélange d'une variable discrète et d'une variable continue (normale). Si on désigne par ϕ et Φ respectivement la densité et la fonction de répartition de la loi normale standard, on a :

$$\Pr(y_i = 0) = \Pr(y_i^* \le 0) = \Pr\left(\frac{u_i}{\sigma} \le -x_i \frac{a}{\sigma}\right) = \Phi\left(-x_i \frac{a}{\sigma}\right) = 1 - \Phi\left(x_i \frac{a}{\sigma}\right) \tag{4.4}$$

Lorsque $y_i > 0$, sa densité s'écrit :

$$f(y_i / y_i > 0) = \frac{1}{\sigma} \frac{\phi((y_i - x_i a)/\sigma)}{\Pr(y_i > 0)} \tag{4.5}$$

La fonction de log-vraisemblance du modèle censuré s'écrit donc :

$$l = \sum_{y_i=0} Log(\Pr(y_i = 0)) + \sum_{y_i>0} Log(\Pr(y_i > 0) f(y_i / y_i > 0) = \sum_{y_i=0} Log\Phi\left(-\frac{x_i a}{\sigma}\right) + \sum_{y_i>0} Log\left(\frac{1}{\sigma}\phi(y_i - x_i a)/\sigma\right)$$

$$l = \sum_{y_i>0} \frac{-1}{2}\left[\log 2\pi + \log\sigma^2 + \left(\frac{y_i - x_i a}{\sigma}\right)\right]^2 + \sum_{y_i=0} \log\left[\Phi\left(-\frac{x_i a}{\sigma}\right)\right] \qquad (4.6)$$

L'estimation par la méthode du maximum de vraisemblance consiste à maximiser la fonction *l*. On sait que cette technique fournit des estimateurs convergents et asymptotiquement efficaces. On note que, contrairement au modèle Probit, on peut identifier ici séparément les paramètres a et σ.

3.1 Pourquoi les MCO ne sont pas appropriés ?

Que se passe-t-il si on ignore le problème de censure ou de troncature et qu'on estime par MCO le modèle $y_i = x_i a + u_i$ sur l'échantillon des individus tels que $y_i > 0$?

Estimer ce modèle par MCO revient en fait à supposer que $E(y_i / x_i, y_i > 0) = x_i a$. Si cette condition n'est pas vérifiée, l'estimateur des MCO sera biaisé. En effet, on a :

$$E(u_i / y_i > 0) = \frac{\sigma\phi(x_i a/\sigma)}{\Phi(x_i a/\sigma)} \neq 0 \qquad (4.7)$$

$$E(y_i / x_i, y_i > 0) = x_i a + \sigma\frac{\phi\left(\frac{x_i a}{\sigma}\right)}{\Phi\left(\frac{x_i a}{\sigma}\right)} \Rightarrow E(\hat{a}_{mco}) \neq 0 \qquad (4.8)$$

(On utilise le résultat : Si $z \approx N(0,1)$ alors $E(z / z \geq z^*) = \frac{\phi(z^*)}{\Phi(-z^*)}$ et $E(z / z \leq z^*) = \frac{-\phi(z^*)}{\Phi(z^*)}$)

Dans un modèle Tobit censuré, on a :

$$E(y_i / x_i, y_i > 0) = x_i a + \sigma\frac{\phi\left(\frac{x_i a}{\sigma}\right)}{\Phi\left(\frac{x_i a}{\sigma}\right)} \qquad (4.9)$$

Il y a donc oubli de variable explicative lorsqu'on estime le modèle par MCO ; la variable omise est définie par :

$$\lambda_i = \frac{\phi\left(\frac{x_i a}{\sigma}\right)}{\Phi\left(\frac{x_i a}{\sigma}\right)} \qquad (4.10)$$

Cette variable est appelée **« ratio inverse de Mills ».**

On peut montrer que l'estimateur des MCO est biaisé en écrivant simplement les conditions de premier ordre du maximum de vraisemblance.

$$\frac{\partial l}{\partial a} = 0 \Rightarrow \frac{1}{\sigma}\sum_{y_i>0}(y_i - x'_i\, a)x_i = \sum_{y_i=0} x_i \lambda_{0i} \tag{4.11}$$

Cette condition s'écrit matriciellement :

$$\frac{1}{\sigma} X'_1 \left(Y_1 - X_1 a\right) = X'_0\, \lambda_0 \tag{4.12}$$

On en déduit que :

$$\hat{a}_{MV} = (X'_1 X_1)^{-1} X'_1 Y_1 - \sigma (X'_1 X_1)^{-1} X'_0 \lambda_0 = \hat{a}_{mco} - \sigma (X'_1 X_1)^{-1} X'_0 \lambda_0 \tag{4.13}$$

L'estimateur du MV est en général supérieur en valeur absolue à celui des MCO.

Ainsi, un modèle Tobit censuré estimé par MCO fournit des estimateurs biaisés non convergents du fait de l'oubli d'une certaine variable explicative, appelée le ratio inverse de Mills.

Que se passe-t-il si on estime par MCO le modèle $y_i = x_i a + u_i$ sur l'échantillon tout entier? Cela suppose que $E(y_i / x_i) = x_i a$. Or, ici encore on a :

$$\begin{aligned} E(y_i / x_i') &= \Pr(y_i > 0).E(y_i / y_i > 0) + \Pr(y_i = 0).E(y_i / y_i = 0) \\ &= \Phi(x_i a / \sigma)\left(x'_i\, a + \sigma\lambda_i\right) + (1 - \Phi_i).0 = \Phi(x'_i\, a / \sigma)\left(x'_i\, a + \sigma\lambda_i\right) \end{aligned} \tag{4.14}$$

Les estimateurs obtenus seront biaisés et non-convergents.

3.2 Procédure d'estimation en deux étapes d'Heckman

Le développement précédent ne signifie pas qu'on ne peut pas utiliser les MCO pour estimer un modèle Tobit censuré. En effet, il suffit d'ajouter au modèle "ce qui manque" et d'estimer le modèle par MCO pour avoir des estimateurs convergents. Le modèle augmenté à estimer s'écrit sous la forme :

$$y_i = x_i a + \lambda_i \sigma + u_i \tag{4.15}$$

Tout le problème revient à trouver un moyen de construire la variable λ qui, manifestement, dépend des paramètres inconnus a et σ.

Heckman (1976) propose une procédure d'estimation utilisant successivement les parties qualitative et quantitative du modèle. Les étapes de cette procédure sont décrites comme suit:

1- Estimer la probabilité de censure à l'aide d'un Probit:

$$\Pr(z_i = 1) = \Pr(y_i > 0) = \Pr obit(x_i a / \sigma) \tag{4.16}$$

Cela fournit un estimateur convergent $\hat{\alpha}'$ de $\alpha' = \frac{a}{\sigma}$.

2- Utiliser cet estimateur pour construire un estimateur convergent du ratio inverse de Mills :

$$\hat{\lambda}_i = \frac{\phi(x_i\hat{\alpha}')}{\Phi(x_i\hat{\alpha}')} \tag{4.17}$$

3- Estimer par MCO le modèle linéaire augmenté $y_i = x_i a + \hat{\lambda}_i\sigma + u_i$ sur le sous-échantillon des observations pour lesquelles $y_i > 0$.

Par construction, les résidus u_i de cette régression sont hétéroscédastiques. En effet, sur ce sous-échantillon, la variance des u_i n'est pas égale à σ^2, elle vaut :

$$\text{var}(u_i / y_i > 0) = \text{var}(u_i / u_i > -x'_i\, a) = \sigma^2\left(1 - \lambda_i^2 + \lambda_i x_i'\alpha\right) \tag{4.18}$$

On utilise le résultat de la théorie des probabilités selon lequel : si $z \to N(0,\sigma)$ alors $\text{var}(z / z < a) = \sigma^2\left(1 - \lambda^2 - \lambda\frac{a}{\sigma}\right)$ avec $\lambda = \frac{\phi(a/\sigma)}{\Phi(a/\sigma)}$.

On peut donc appliquer les MCO pondérés pour corriger l'hétéroscédasticité des erreurs : après avoir estimé α', on estime ensuite λ, on calcule la variance des erreurs et on pondère les observations par $\sqrt{\hat{V}(u_i)}$.

4. Calcul des effets marginaux

Les effets marginaux dans un modèle de régression correspondent à des prévisions sur une variable continue lorsqu'une variable explicative donnée est modifiée. Dans un modèle Tobit, il y a trois effets marginaux possibles selon la distribution de la variable considérée. En effet, on a :

- $$\frac{\partial E(y_i^* / x_i)}{\partial x_k} = a_k \tag{4.19}$$

- $$\frac{\partial E(y_i / x_i)}{\partial x_k} = a_k.\Phi(x_i\frac{a}{\sigma}) \tag{4.20}$$

  En effet, on montre que $E(y_i / x_i) = \Phi(\frac{x_i a}{\sigma})(x_i a + \sigma\lambda_i)$ (voir Greene (1997) page 910).

- $$\frac{\partial E(y_i / x_i, y^*_i > 0)}{\partial x_k} = a_k\left[1 - x_i\alpha\lambda_i - \lambda_i^2\right] \tag{4.21}$$

  En effet: $E(y_i / x_i, y_i > 0) = x_i a + \sigma\lambda_i$.

5. Test d'hétéroscédasticité

La présence d'hétéroscédasticité implique la non-convergence de l'estimateur du maximum de vraisemblance (Gourieroux, 1989, p.210). Il faut donc prendre en compte l'hétéroscédasticité quand elle est présente, lors de l'estimation. Une façon de tester l'hétéroscedasticité est de spécifier une relation scédastique sous la forme :

$$\sigma_i^{\,2} = \sigma_0^2 \exp(z_i'\gamma) \tag{4.22}$$

où z_i est un vecteur de g variables responsables de l'hétéroscedasticité. En pratique, on prend souvent certaines variables explicatives. L'hypothèse nulle d'homoscédasticité est équivalente à :

$$\mathrm{H}_0 : \gamma = 0 \tag{4.23}$$

Cette hypothèse peut être testée à l'aide de la statistique du ratio de vraisemblance:

$$LR = -2(l_0 - l) \approx \chi^2(g) \tag{4.24}$$

où l_0 est la log-vraisemblance sous l'hypothèse d'homoscédasticité et l la log-vraisemblance sous l'hypothèse d'hétéroscédasticité.

On peut également utiliser le test du multiplicateur de Lagrange. Si l'on évalue la matrice $G(a, b)$ de dimension $(n, k+g)$ contenant les dérivées de la fonction de log-vraisemblance pour chaque observation, le terme général de cette matrice est donné par :

$$G_{ij} = \frac{\partial l(x_i, y_i, \theta)}{\partial \theta_j} \quad , j = 1 \ldots k+g \tag{4.25}$$

avec $\theta = (a,b)$ un vecteur de $k+g$ éléments.

La statistique de test est définie par :

$$LM = e'_n \, G(\hat{a}) \left[G(\hat{a})' G(\hat{a}) \right]^{-1} G(\hat{a})' e_n \tag{4.26}$$

où $e_n = (1,1,...,1)'$.

On utilise un résultat beaucoup plus simple:

$$LM = nR^2 \rightarrow \chi^2(g) \tag{4.27}$$

où R^2 est le coefficient de determination de la régression du vecteur $e_n = (1,1,...,1)'$ sur la matrice G, évaluée sous l'hypothèse d'homoscédasticité.

Sous l'hypothèse d'une hétéroscédasticité multiplicative de la forme $\sigma_i^{\,2} = \sigma_0^2 \exp(z_i'\gamma)$, le gradient de la fonction log-vraisemblance donne les équations ci-après:

$$\frac{\partial l}{\partial a} = \sum_{i=1}^{n} x_i \left[k_i u_i / \sigma_0^2 - (1 - k_i)\lambda_i / \sigma_0^2 \right] \tag{4.28}$$

$$\frac{\partial l}{\partial \sigma_0^2} = \sum_{i=1}^{n} k_i \left[u_i^2 / \sigma_0^2 - 1 \right] / 2\sigma_0^2 + (1 - k_i)\lambda_i x_i' a / 2\sigma_0^3 \tag{4.29}$$

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{n} \sigma_0^2 k_i \left[k_i' (u_i^2 / \sigma_0^2 - 1) / 2\sigma_0^2 + (1 - k_i)\lambda_i x_i' a / 2\sigma_0^3 \right] \tag{4.30}$$

6. Modèle Tobit généralisé (Tobit avec sélection)

Dans le modèle Tobit simple (Tobit I), les deux parties du modèle (qualitative et quantitative) sont modélisées simultanément. Dans l'exemple de l'achat du bien, l'individu décide simultanément du fait qu'il va ou non consommer le bien et de la dépense qu'il va affecter à l'achat du bien. En fait, ce modèle ne modélise pas explicitement la partie qualitative, c'est-à-dire la décision d'acheter ou non le bien. Autrement dit, le modèle Tobit simple suppose que les déterminants de la décision d'acheter le bien et la somme dépensée sont les mêmes. Un modèle alternatif, plus approprié à l'étude de nombreux phénomènes (offre de travail, dépenses de transferts…), consiste à supposer un comportement séquentiel séparant les deux parties du modèle. Dans un premier temps, l'individu décide s'il va consommer ou non le bien. Cette première décision peut être modélisée par un modèle binaire basé sur une variable latente s_i^* : si $s_i^* > 0$ alors l'individu achète le bien, sinon il ne l'achète pas.

Dans une seconde étape, il fixe la somme y_i^* qu'il va consacrer à l'achat du bien. La variable observée y_i est alors définie par :

$$y_i = \begin{cases} y_i^* & si \quad s_i^* > 0 \\ 0 & si \quad s_i^* \leq 0 \end{cases} \tag{4.31}$$

Cette spécification généralise le modèle Tobit simple qui correspond au cas particulier où $y_i^* = s_i^*$.

De façon générale, le modèle Tobit généralisé (Tobit de type II) est un modèle dans lequel le phénomène de censure est basé sur la valeur d'une variable *s* différente de la variable dépendante. La structure du modèle Tobit II est la suivante:

$$\begin{cases} y_i^* = x_i' a + u_i \\ s_i^* = z'_i \alpha + v_i \\ y_i = \begin{cases} y_i^* & si \quad s_i^* > 0 \\ 0 & si \quad s_i^* \leq 0 \end{cases} \end{cases} \tag{4.32}$$

La deuxième équation $s_i^* = z'_i\,\alpha + v_i$ définit l'équation de sélection.

On suppose que :

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \approx Niid\left(0, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix}\right) \tag{4.33}$$

où ρ représente le coefficient de corrélation entre u_i et v_i. La formulation Tobit II permet donc de faire apparaître la plus ou moins grande corrélation existant entre les deux décisions. La restriction que la variance de v_i est égale à 1 est imposée parce que seul le signe de s_i^* sera observé. De fait, les variables réellement observées sont y_i et s_i. *x* et *z* sont des vecteurs de variables explicatives de dimensions k_1 et k_2 respectivement. Certaines variables explicatives peuvent être communes à *x* et *z*, mais *a priori* rien n'impose que ces variables soient les mêmes. En effet, une variable peut très bien expliquer les dépenses de consommation d'un bien sans pour autant être déterminante dans la décision d'achat du bien.

On montre que:

$$E(y_i / s_i^* > 0) = x'_i\,a + \rho\sigma\frac{\phi(z_i b)}{\Phi(z_i b)} \tag{4.34}$$

On peut appliquer la procédure d'estimation en deux étapes d'Heckman. Dans une première étape, on modélise par un Probit ordinaire la partie qualitative du modèle:

$$s_i = \begin{cases} 1 \quad si \quad s_i^* > 0 \\ \\ 0 \quad si \quad s_i^* \le 0 \end{cases} \tag{4.35}$$

L'estimation Probit de ce modèle permet d'obtenir un estimateur convergent $\hat{\alpha}$ des paramètres de l'équation de sélection $\Pr(s_i = 1) = \Pr obit(z_i\alpha)$.

Dans une seconde étape, le regresseur de sélection $\lambda_i = \dfrac{\phi(z_i\alpha)}{\Phi(z_i\alpha)}$ est évalué en $\hat{\alpha}$ et le modèle augmenté $y_i = x'_i\,a + \theta\hat{\lambda}_i + \varepsilon_i$ est estimé par *MCO* à l'aide des observations pour lesquelles $s_i = 1$.

On obtient des estimateurs asymptotiquement sans biais de *a* et $\rho\sigma$. Mais ces estimateurs ne sont pas asymptotiquement efficaces, car les résidus de la régression sont hétéroscédastiques par construction. En effet, on a:

$$V(\varepsilon_i) = \sigma^2 - (\rho\sigma)^2\left[z'_i\,\alpha.\lambda_i + \lambda_i^2\right] \tag{4.36}$$

Ainsi, pour obtenir une inférence paramétrique correcte, il est nécessaire de corriger cette hétéroscédasticité par la méthode des moindres carrés pondérés. Pour obtenir une estimation de σ, on considère les résidus de la dernière régression:

$$\hat{\varepsilon}_i = y_i - x'_i \hat{a} - \hat{\theta}\hat{\lambda}_i \tag{4.37}$$

Puisque :

$$\sigma^2 = V(\varepsilon_i) + (\rho\sigma)^2 \left[z'_i \, \alpha . \lambda_i + \lambda_i^2 \right] \tag{4.38}$$

On obtient un estimateur convergent de σ^2 donné par :

$$\hat{\sigma}^2 = \frac{1}{n_1} \sum_{s_i=1} \hat{\varepsilon}_i^2 + \frac{\hat{\theta}^2}{n_1} \sum_{s_i=1} (z'_i \, \hat{\alpha}\hat{\lambda}_i + \hat{\lambda}_i^2) \tag{4.39}$$

où n_1 représente l'effectif des observations pour lesquelles $s_i = 1$.

Cette méthode d'estimation en deux étapes fournit aussi bien un test pour la sélection d'échantillon qu'une technique d'estimation. Le coefficient du regresseur de sélection est égal à $\rho\sigma$. Puisque $\sigma \neq 0$ alors on peut utiliser un test de Wald pour tester l'hypothèse que $\rho = 0$. Sous cette hypothèse, la statistique de test suit asymptotiquement une loi du Khi-deux à un degré de liberté. Si l'hypothèse n'est pas rejetée, on peut conclure que la sélection n'introduit pas de biais dans l'estimation par la méthode des MCO.

7. Modèle à régime

On suppose ici que l'échantillon peut être scindé en deux sous-echantillons suivant un critère donné. Par exemple, marié/non mariés, utilise ou n'utilise pas une méthode contraceptive, etc. On considère la variable de régime $R_i = 1$ si l'individu *i* satisfait le critère (régime 1), 0 sinon (régime 2). La variable latente associée à R_i est notée R_i^*.

On cherche à modéliser une variable dépendante *y* sachant que celle-ci est expliquée de façon différente selon le régime :

$$\begin{cases} y_{1i} = x_{1i} a_1 + u_{1i} & si \quad R_i = 1 \\ y_{2i} = x_{2i} a_2 + u_{2i} & si \quad R_i = 0 \end{cases} \qquad \begin{cases} R_i^* = Z_i \alpha + v_i \\ R_i = 1_{\{Z_i \alpha + v_i > 0\}} \end{cases} \tag{4.40}$$

Les caractéristiques explicatives x_1 et x_2 ne sont pas nécessairement les mêmes, elles peuvent être communes ou bien différentes selon le régime. Par exemple, si on considère le régime relatif à l'utilisation d'une méthode contraceptive, on peut prendre en compte les variables telles que la source d'information sur cette méthode, la fréquence d'utilisation, pour les individus utilisant la méthode (*régime 1*).

On pose que u_{1i}, u_{2i} et v_i suivent une loi normale trivariée avec :

- $E(u_{1i}^2) = \sigma_1^2$ (4.41)
- $E(u_{2i}^2) = \sigma_2^2$ (4.42)
- $E(v^2{}_i) = \sigma_v^2 = 1$ (4.43)
- $E(u_{1i}u_{2i}) = \sigma_{12}$ (4.44)
- $corr(u_{1i}, v_i) = \rho_{1v}$ (4.45)
- $corr(u_{2i}, v_i) = \rho_{2v}$ (4.46)

Sous ces hypothèses, on établit que :

$$E(y_{1i} / R_i = 1) = x_{1i}a_1 + \rho_{1v}\sigma_1 \frac{\phi(z_i\alpha)}{\Phi(z_i\alpha)} \quad (4.47)$$

$$y_{2i} = x_{2i}a_2 - \rho_{2v}\sigma_2 . \frac{\phi(z_i\alpha)}{(1-\Phi(z_i\alpha)} + \varepsilon_{2i} \quad (4.48)$$

On peut utiliser une procédure d'estimation en deux étapes à la Heckman :

- On estime à l'aide d'un Probit l'équation de régime $\Pr(R_i = 1) = \Pr obit(z_i\alpha)$. On obtient un estimateur convergent de α, à partir duquel on estime les regresseurs de sélection $\frac{\phi(z_i\alpha)}{\Phi(z_i\alpha)}$ et $\frac{\phi(z_i\alpha)}{(1-\Phi(z_i\alpha))}$.
- On estime séparément, par MCO, les modèles $y_{1i} = x_{1i}a_1 + \rho_{1v}\sigma_1 \frac{\phi(z_i\hat{\alpha})}{\Phi(z_i\hat{\alpha})} + \varepsilon_{1i}$ et $y_{2i} = x_{2i}a_2 - \rho_{2v}\sigma_2 . \frac{\phi(z_i\hat{\alpha})}{(1-\Phi(z_i\hat{\alpha}))} + \varepsilon_{2i}$, sous les hypothèses habituelles du modèle linéaire multiple.

Chapitre 5

Modèles de Comptage

1. Introduction

Le but de ce chapitre est de trouver le ou les modèles appropriés pour analyser une variable de comptage. Une variable Y est dite de comptage si elle désigne le nombre de fois qu'un événement survient.

Ces variables sont la résultante de certains phénomènes prenant des petits nombres de valeurs discrètes positives, mais non catégorielles, comme par exemple le nombre d'accidents, le nombre d'enfants, le nombre d'années d'étude, le nombre d'arrivée journalière à une gare, le nombre de fois où un individu change d'emploi.

Pour expliquer comment les réalisations de telles variables dépendent d'autres variables quantitatives ou qualitatives, le modèle linéaire classique se révèle inadéquat pour les mêmes raisons que dans le modèle dichotomique :

1) le nuage des observations n'a pas une forme adaptée à un ajustement linéaire ;
2) l'hypothèse de normalité ne peut être plausible puisque la variable endogène prend des valeurs discrètes avec des probabilités non nulles ;
3) les prévisions de la variable dépendante donnent des valeurs que ne peut prendre *y*.

La formulation la plus courante consiste à supposer que les réalisations de la variable sont issues d'une loi de Poisson, dont le paramètre dépend des valeurs prises par des variables exogènes.

2. Distribution de Poisson

Soit y une variable aléatoire indiquant le nombre de fois un événement s'est produit durant un intervalle de temps. On dit que y a une distribution de Poisson de paramètre $\lambda > 0$ si :

$$\Pr ob(y = k) = e^{-\lambda} \frac{\lambda^k}{k!} \tag{5.1}$$

Le paramètre λ est appelé taux d'incidence et on a :

$$E(y) = \lambda = Var(y) \tag{5.2}$$

L'égalité de la moyenne et la variance est qualité d'équidispersion. En pratique, les variables de comptage ont souvent une variance plus grande que la moyenne, ce qui est qualifié de surdispersion. Le développement des modèles de comptage essaie de prendre en compte cette surdispersion.

Une autre hypothèse du modèle de Poisson est que les événements sont indépendants. Cela signifie que quand un événement se produit, il n'affecte pas la probabilité de réalisation de l'événement dans le futur. Par exemple, considérons le nombre de visites chez un médecin. L'hypothèse d'indépendance implique que quand un individu visite un médecin, son taux de visite ne change pas. Les visites passées n'affectent pas les visites futures. De même, si on considère le nombre d'enfants nés durant une période, l'hypothèse d'indépendance implique que le fait d'avoir des enfants n'affecte pas le taux de naissance.

L'une des explications de l'échec de la distribution de Poisson à ajuster correctement les données empiriques est que le paramètre λ diffère selon les individus. On qualifie cette situation d'hétérogénéité entre les individus. L'hétérogénéité dans les caractéristiques des individus est la cause de la surdispersion dans la distribution marginale de la variable.

3. Le modèle de régression de Poisson

3.1 Présentation du modèle

Soit y_i une variable de comptage à valeur dans *N*. la probabilité que $y_i = k$, avec $k \in \{0,1,...\}$, est donnée par:

$$\Pr(y_i = k) = e^{-\lambda_i} \frac{\lambda_i^k}{k!} \tag{5.3}$$

où λ_i est le paramètre de distribution tel que $E(y_i) = V(y_i) = \lambda_i$. Pour introduire des variables explicatives $x_i' = (x_{1i},...,x_{ki})$, on pose la relation suivante :

$$\lambda_i = \exp(x'_i\ \beta) = \exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right) \text{ ou } \log(\lambda_i) = x'_i\ \beta \tag{5.4}$$

Le choix de la forme fonctionnelle liant le paramètre aux exogènes s'explique essentiellement par la nécessité d'avoir des λ_i positifs. En effet, une spécification $\lambda_i = x'_i\ \beta$ conduirait à une des λ_i négatif. De plus, lorsque les variables explicatives sont prises en log, les coefficients s'interprètent comme des élasticités :

$$\beta_j = \frac{\partial \log E(y_i)}{\partial \log x_{ik}} \tag{5.5}$$

La probabilité conditionnelle de y_i s'écrit :

$$\Pr(y = y_i) = e^{-\lambda_i(x_i)} \frac{\lambda_i(x_i)^{y_i}}{y_i!} = \exp\left[-\lambda_i(x_i) + y_i \log(\lambda_i(x_i)) - \log(y_i!)\right] \tag{5.6}$$

La log-vraisemblance du modèle de Poisson est :

$$\log L = \sum_{i=1}^{n} \left(-\exp(x'_i \beta) + y_i x'_i \beta - \log(y_i!) \right) \tag{5.7}$$

Le vecteur de paramètres β s'obtient par maximisation de cette fonction :

$$\hat{\beta} = \underset{\beta}{ArgMax} \log L \tag{5.8}$$

3.2 Interprétation

Les outils d'interprétation d'un modèle de comptage diffèrent selon que l'on désire connaître la valeur espérée ou la distribution de la variable.

3.3 Effet marginal sur la moyenne conditionnelle

La valeur espérée de y conditionnellement aux variables explicatives est :

$$E(y) = \exp(x'\beta) = \exp\left(\sum_{j=1}^{k} x_j \beta_j \right) \tag{5.9}$$

L'effet marginal d'une variable explicative x_j sur la valeur espérée de y est :

$$\frac{\partial E(y/x)}{\partial x_j} = \beta_j \exp(x'\beta) = \beta_j E(y/x) \tag{5.10}$$

L'effet marginal dépend de x_j mais aussi de toutes les autres variables.

On peut calculer l'effet sur la moyenne en termes relatifs ou de pourcentage. Pour une variation absolue de x_j de δ (x_j passe de x_j^* à $x_j^* + \delta$), on a :

$$\frac{E(y/x, x_j = x_j^* + \delta)}{E(y/x, x_j = x_j^*)} = \exp(\beta_j \delta) \tag{5.11}$$

Si toutes les variables autres que x_j sont maintenues constantes, alors toute variation d'une unité de x_j ($\delta = 1$) entraîne une variation de la valeur espérée de $\exp(\beta_j)$.

Alternativement, on peut exprimer cette variation en pourcentage :

$$\frac{E(y/x, x_j = x_j^* + \delta) - E(y/x, x_j = x_j^*)}{E(y/x, x_j = x_j^*)} \times 100 = 100 \times \left[\exp(\beta_j \delta) - 1 \right] \tag{5.12}$$

Enfin, on peut calculer le changement suite à un changement discret dans la variable x_j, par exemple x_j passe d'une valeur x_E à une valeur x_S :

$$\Delta_j = \frac{\Delta E(y/x)}{\Delta x_j} = E(y/x, x_j = x_E) - E(y/x, x_j = x_S) \qquad (5.13)$$

Ainsi, pour un changement de x_j de x_E à x_S, la valeur espérée change de Δ_j, toutes choses étant égales par ailleurs. Les cas les plus courants où un changement discret a lieu sont obtenus en faisant varier x_j de :

- sa valeur minimum à sa valeur maximum ;
- 0 à 1 (pour une variable binaire).

3.4 Probabilités prédites

Connaissant x'_i on peut calculer la probabilité que y_i prenne n'importe quelle valeur de son ensemble de définition :

$$\Pr(y_i = k / x_i) = \frac{\exp(x'_i \hat{\beta})^k \exp(-\exp(x'_i \hat{\beta}))}{k!} \qquad (5.14)$$

Cette probabilité est calculée pour chaque observation et pour chaque valeur de *k*. La probabilité prédite moyenne permet, pour chaque valeur de *k*, de résumer le pourvoir prédictif du modèle. Elle est donnée par :

$$\overline{\Pr(y = k)} = \frac{1}{n} \sum_{i=1}^{n} \Pr(y_i = k / x_i) \qquad (5.15)$$

3.5 Prise en compte du temps d'exposition

Dans ce qui précède, nous n'avons pris en compte le temps d'exposition des individus à l'événement d'intérêt. Rappelons que les y_i sont indépendantes et que $\lambda_i = E(y_i / x'_i)$ dans une unité d'intervalle de temps.

Désignons par t_i la durée de temps d'exposition de l'individu à l'événement. Au bout du temps t_i, le nombre d'événements espéré est :

$$\mu_i = t_i \times \lambda_i = \exp(\ln(t_i) + x'_i \beta) = \exp(z'_i \tilde{\beta}) \qquad (5.16)$$

On peut donc intégrer le temps t_i dans la régression à l'aide de la variable $\ln(t_i)$ dont le coefficient est forcé égal à 1.

4. Modèle binomial négatif

4.1 Spécification

Le modèle de Poisson impose que l'espérance conditionnelle est égale à la variance conditionnelle. Cette hypothèse est parfois peu réaliste. Le problème souvent rencontré est celui

de la surdispersion: $V(y_i / x_i) > E(y_i / x_i)$. Ce problème provient de l'hétérogénéité non observable.

Dans le modèle de Poisson, la moyenne conditionnelle de y sachant x est connue : $\lambda = \exp(x'\beta)$. Dans le modèle binomial négatif (*Modèle NegBin : Negative Binomial Model*), la moyenne conditionnelle est une variable aléatoire :

$$\tilde{\lambda}_i = \exp(x'_i \beta + e_i) \tag{5.17}$$

où e_i est un terme d'erreur aléatoire supposé non-corrélé avec x_i. Dans le modèle de Poisson, les variations de λ_i résultent de l'hétérogénéité observée entre les individus. A différentes valeurs de x_i sont associées différentes valeurs de λ et tous les individus ayant les mêmes caractéristiques observables x ont la même valeur de λ. Dans le modèle binomial négatif, les variations de $\tilde{\lambda}_i$ sont dues à la fois aux variations de x_i et à l'hétérogénéité non observable captée par la variable e_i. Pour des valeurs données de x_i, il existe une distribution de valeurs de $\tilde{\lambda}_i$ plutôt qu'une seule valeur.

La relation entre les moyennes conditionnelles du modèle de Poisson et du modèle binomial négatif est donnée par la relation suivante:

$$\tilde{\lambda}_i = \exp(x'_i \beta)\exp(e_i) = \lambda_i \exp(e_i) = \lambda_i \delta_i \tag{5.18}$$

Pour permettre que le modèle ait la même moyenne conditionnelle que le modèle de Poisson, on pose que :

$$E(\delta_i) = 1 \tag{5.19}$$

La distribution conditionnelle des observations est toujours une loi de Poisson :

$$\Pr(y = y_i / x_i) = e^{-\tilde{\lambda}_i} \frac{\tilde{\lambda}_i^{y_i}}{y_i!} = \frac{\exp(-\lambda_i \delta_i)(\lambda_i \delta_i)^{y_i}}{y_i!} \tag{5.20}$$

Cependant, étant donné que le paramètre δ_i est inconnu, on ne peut pas calculer cette distribution de probabilité. On impose une distribution de probabilité pour le paramètre δ_i. L'hypothèse la plus souvent faite est de supposer que δ_i suit une distribution Gamma de paramètre ν_i :

$$g(\delta_i) = \frac{\nu_i^{\nu_i}}{\Gamma(\nu_i)} \delta_i^{\nu_i - 1} \exp(-\delta_i \nu_i),\ \nu_i > 0 \tag{5.21}$$

On montre que $E(\delta_i) = 1$ et $Var(\delta_i) = 1/\nu_i$.

Sous ces hypothèses, la variance conditionnelle de y_i est définie par :

$$Var(y_i / x_i) = \lambda_i \left(1 + \frac{\lambda_i}{\nu_i} \right) \tag{5.22}$$

Si l'expression de la moyenne conditionnelle permet d'identifier le paramètre λ_i, il se pose un problème d'identification pour la valeur de ν_i. Si ce paramètre varie suivant l'individu, alors il y aura plus de paramètres que d'observations. L'hypothèse souvent faite est de supposer que le paramètre ν_i est constant :

$$\nu_i = \frac{1}{\alpha} \tag{5.23}$$

Cette hypothèse implique que la variance de δ_i est constante. Le coefficient α est appelé paramètre de dispersion car la variance conditionnelle de y croit avec α.

4.2 Estimation

Le modèle binomial négatif peut être estimé par la méthode du maximum de vraisemblance. La fonction de vraisemblance du modèle est la suivante :

$$L(\beta / y, x) = \prod_{i=1}^{n} \Pr(y = y_i / x_i) = \prod_{i=1}^{n} \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1}) y!} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i} \right)^{\alpha^{-1}} \left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i} \right)^{y_i} \tag{5.24}$$

où $\lambda_i = \exp(x'_i \beta)$.

4.3 Test de l'hypothèse de surdispersion

Il est important de tester l'hypothèse de surdispersion lorsqu'on utilise le modèle de Poisson afin de vérifier si l'hypothèse sous-jacente au modèle est vérifiée. La spécification de Poisson peut être facilement testée à travers l'hypothèse $H_0 : \quad \alpha = 0$. Sous l'hypothèse nulle, le modèle binomial négatif se réduit au modèle de Poisson. On peut utiliser un *z-test* unilatéral pour tester la significativité de α. On peut aussi utiliser la statistique du rapport de vraisemblance définie par :

$$LR = 2(\log L_{MBN} - \log L_{MP}) \rightarrow \chi^2(1) \tag{5.25}$$

où L_{MBN} est la vraisemblance du modèle binomial négatif et L_{MP} la vraisemblance du modèle de Poisson.

Références Bibliographiques

[1.] Alban T. (2000), *Econométrie des variables qualitatives*, Dunod, Paris.

[2.] Amemiya T. (1981), ”Qualitative Response Models : A Survey”, *Journal of Economic Litterature,* 19(4), 481-536

[3.] Colin A. C. et Pravin K. T. (1998), *Regression Analysis of Count Data*, Econometric Society Monographs.

[4.] Gouriéroux C. (1989), *Econométrie des Variables Qualitatives*, Paris, Economica.

[5.] Gouriéroux C., Monfort A.(1981), "Asymptotic properties of the Maximum Likelihood Estimator in Dichotomous Logit Models", *Journal of Econometrics*, n°17, pp.83-97.

[6.] Greene W.H.(1997), *Econometric Analysis*, Londres, Prentice Hall, 3ième edition.

[7.] Gujarati, D. (1995), *Basic Econometrics*, Third Edition, New York, McGraw-Hill.

[8.] Maddala G.S. (1983), *Limited-dependent and Qualitative Variables in Econometrics*, Econometric Society Monographs, N°3, Cambridge, Cambridge University Press.

[9.] Pindyck, R.S., Rubinfeld, D.L. (1981), *Econometric Models and Economic Forecasts*, Second Edition, McGraw-Hill.

[10.] Scott L. et Freese J. (2000), *Regression Models for Categorical Dependent Variables Using Stata*, 2ième édition, A stata Press Publication, Texas.