Université d'Orléans - Maitrise Econométrie
Econométrie des Variables Qualitatives

Examen Novembre 2002

Exercice 1 (12 points) : On considère le problème traditionnel du dosage d'un insecticide (Gurland, Lee et Dahm 1960), dont la structure peut naturellement être reprise dans de nombreux problèmes économiques (crises de change, défaillance dans les relations bancaires, décision d'achat etc..). Le problème est le suivant : on diffuse dans un espace clos un insecticide et l'on cherche à déterminer la dose minimale permettant de tuer les insectes ainsi que les principales caractéristiques permettant de modéliser leur tolérance à cette insecticide.

A l'issue de l'expérience, on observe les N insectes de l'échantillon, indicés $i = 1.., N$. On adopte le code $y_i = 1$ lorsque le $i^{ème}$ insecte est mort à l'issue de l'expérience et $y_i = 0$ lorsque l'insecte est vivant. On admet que dans tous les insectes reçoivent la même dose d'insecticide, notée $\widetilde{\gamma}$. Or, chaque insecte dispose d'une capacité de résistance propre représentée par une tolérance à l'insecticide, notée y_i^*. Si la tolérance à l'insecticide de l'individu i est inférieure à la dose de produit reçu, l'insecte décède ($y_i = 1$), et il reste vivant ($y_i = 0$) lorsque la tolérance est au contraire supérieure à la dose d'insecticide reçue. La tolérance y_i^* est inobservable, toutefois on suppose qu'elle peut être modélisée sous la forme d'une combinaison linéaire d'un ensemble de K caractéristiques propres à chaque insecte (poids, âge, sexe etc..), représentées par un vecteur x_i (sans constante) et d'un résidu noté ε_i

$$y_i^* = x_i \widetilde{\beta} + \varepsilon_i \tag{1}$$

avec $x_i = \left(x_i^1..x_i^K\right)$, $\forall i = 1,..,N$, $\widetilde{\beta} = \left(\widetilde{\beta}_1...\widetilde{\beta}_K\right)' \in \mathbb{R}^K$, ε_i *i.i.d.* $\left(0, \sigma_\varepsilon^2\right)$ et où **la variable réduite** $\varepsilon_i/\sigma_\varepsilon$ suit une loi de fonction de répartition $F(.)$.

1. **Question 1** *(1 point)* : Compte tenu des éléments précédents exprimez la probabilité que le $i^{ème}$ insecte soit mort à l'issue de l'expérience en fonction des paramètres $\widetilde{\beta}$, $\widetilde{\gamma}$ et σ_ε, des caractéristiques x_i et de la fonction $F(.)$.

2. **Question 2** *(1 point)* : Discutez brièvement les avantages et les inconvénients respectifs d'une modélisation logit ou probit de ce problème.

3. **Question 3** *(2.5 points)* : On considère une modélisation logit où $F(.)$ désigne la fonction de répartition d'une loi logistique :

$$F(x) = \Lambda(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}} \quad \forall x \in \mathbb{R} \tag{2}$$

Sachant que la variance d'une variable distribuée selon une loi logistique est égale à $\pi^2/3$, montrez que le choix d'un modèle logit revient en fait à normaliser les coefficients $\widetilde{\beta}$ et $\widetilde{\gamma}$ en imposant :

$$Prob\left(y_i = 1\right) = \Lambda\left(\gamma - x_i \beta\right) \tag{3}$$

Exprimez les paramètres normalisés γ et β en fonction des paramètres $\widetilde{\beta}$ et $\widetilde{\gamma}$ du modèle initial. *Qu'en concluez vous quant à l'interprétation de ces paramètres normalisés ?* Ecrivez les résidus normalisés du modèle transformé, notés v_i, en fonction des ε_i. Vérifiez que ces résidus normalisés ont une variance égale à $\pi^2/3$.

4. **Question 4** *(3 points)* : On considère à présent un **modèle logit** où la tolérance s'écrit :

$$y_i^* = x_i \beta + v_i \tag{4}$$

où $x_i = \left(x_i^1..x_i^K\right)$, $\forall i = 1,..,N$, $\widetilde{\beta} = \left(\widetilde{\beta}_1...\widetilde{\beta}_K\right)' \in \mathbb{R}^K$, v_i *i.i.d.* $\left(0, \pi^2/3\right)$. On note γ le seuil mortel d'insecticide tel que :

$$y_i = \begin{cases} 1 & \text{si } y_i^* < \gamma \\ 0 & \text{sinon} \end{cases} \quad \forall i = 1.., N \tag{5}$$

Ecrivez la log-vraisemblance $\log L(y, \beta, \gamma)$ associée à l'échantillon des N insectes et dérivez le gradient de cette log-vraisemblance par rapport au vecteur de paramètres β. *Commentez la forme du gradient de ce problème par rapport à celle vue en cours et indiquez brièvement comment obtenir un estimateur du maximum de vraisemblance des paramètres du modèle.*

5. **Question 5** *(2.5 points)* : On suppose pour simplifier que $\gamma = 0$ et que $K = 1$. On note $\widehat{\beta}$ l'estimateur du maximum de vraisemblance (MV) du paramètre β associé à l'unique caractéristique x_i. En supposant que les conditions nécessaire à la normalité asymptotique de l'estimateur du MV sont vérifiées, construisez un intervalle de confiance pour un risque de première espèce de $\alpha\%$, sur le paramètre β.

6. **Question 6** *(2 points)* : En conservant les hypothèses $\gamma = 0$ et $K = 1$, calculez l'élasticité de la probabilité associée à la mort du $i^{ème}$ insecte par rapport à l'unique variable explicative x_i. *Commentez.* Déterminez la cote associée à cet évenement.

Exercice 2 (10 points) : On cherche ici à expliquer par un modèle à réponse binaire la probabilité de vote pour le candidat démocrate Jimmy Carter aux éléctions présidentielles américaines de 1976 en fonction de caractéristiques propres aux Etats. La variable expliquée $y_i = 1$ prend la valeur 1 si dans l'état indicé i, les votes ont été majoritaires pour le candidat démocrate Jimmy Carter, et 0 si au contraire le candidat républicain Gerry Ford a obtenu la majorité des voix dans cet état. Les variables explicatives sont les suivantes : *Inc* désigne le revenu médian de l'état en 1975, *School* désigne la médiane du nombre d'années de scolarité suivies par les habitants de l'état âgés de plus de 18 ans, $Urban$ désigne le pourcentage de la population vivant en zone urbaine et la variable *Region* est une variable dummy prenant la valeur 1 pour la région Nord Est, 2 pour le Sud Est, 3 pour le Sud et le Middle West et 4 pour l'Ouest. Les données couvrent les 51 états américains et sont tirées de Greene[1] (1997). Sur la figure (1) sont reportés les résultats d'estimation d'un modèle probit sous Eviews.

Figure 1: Estimation d'un Modèle Probit

Dependent Variable: Y
Method: ML - Binary Probit
Date: 11/03/02 Time: 21:27
Sample: 1 51
Included observations: 51
Convergence achieved after 7 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
INC	-0.000406	0.000156	-2.598739	0.0094
SCHOOL	0.269580	0.139611	1.930938	0.0535
URBAN	0.035902	0.012489	2.874685	0.0040

Mean dependent var	0.470588	S.D. dependent var	0.504101
S.E. of regression	0.455714	Akaike info criterion	1.241275
Sum squared resid	9.968416	Schwarz criterion	1.354912
Log likelihood	-28.65252	Hannan-Quinn criter.	1.284699
Avg. log likelihood	-0.561814		
Obs with Dep=0	27	Total obs	51
Obs with Dep=1	24		

[1] Greene W. (1997) "*Econometric Analysis*".

Figure 2: Nombres de Prédictions Fausses

Dependent Variable: Y
Method: ML - Binary Probit
Date: 11/03/02 Time: 21:27
Sample: 1 51
Included observations: 51
Prediction Evaluation (success cutoff C = 0.5)

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	20	6	26	27	24	51
P(Dep=1)>C	7	18	25	0	0	0
Total	27	24	51	27	24	51
Correct	20	18	38	27	0	27
% Correct	74.07	75.00	74.51	100.00	0.00	52.94
% Incorrect	25.93	25.00	25.49	0.00	100.00	47.06
Total Gain*	-25.93	75.00	21.57			
Percent Gain*	NA	75.00	45.83			

7. **Question 1** *(2 points)* : Dans quelle mesure le revenu médian des états a-t-il eu une influence sur le vote démocrate aux éléctions de 1976 ? Même question pour le niveau de scolarité médian et l'importance de la population urbaine.

8. **Question 2** *(2 points)* : A partir des estimations du modèle probit, donnez la probabilité qu'un Etat comme le Texas, ayant un revenu médian de 12672$, un niveau médian d'étude de 12.4, et un taux d'urbanisation de 71.4 ait voté démocrate aux éléctions de 1976. *Calculez la cote du vote démocrate dans cet état.* Quelle utilisation pourrait on envisager pour ce modèle probit dans le cadre de nouvelles éléctions présidentielles ? Précisez les limites d'un tel exercice.

9. **Question 3** *(1.5 points)* : Compte tenu des différents éléments à votre disposition (figures 1 et 2), établissez un diagnostic quant à la qualité de la modélisation probit de la probabilité du vote démocrate.

10. **Question 4** *(1.5 points)* : Le choix d'un modèle probit vous semble–t-il justifié dans ce cas ? Proposez une valeur approximée de la réalisation des estimateurs des différents paramètres des variables explicatives dans le cas d'un modèle logit et d'un modèle linéaire simple pour cet échantillon.

11. **Question 5** *(2 points)* : Calculez l'élasticité pour l'état du Texas de la probabilité du vote démocrate par rapport au revenu médian. Vous utiliserez pour cela les élements et résultats de la question 2. Quel serait l'impact d'une augmentation de 20% du revenu médian dans cet état en matière de vote aux présidentielles ?

12. **Question 6** *(1 point)* : L'estimation d'un probit avec pour seule variable explicative la variable $Urban$ donne comme valeur pour la log vraisemblance $Loglikelihood - 35.28002$. Testez la nullité des coefficients des variables de revenu Inc et de scolarité $School$.

Correction Examen Novembre 2002

Christophe Hurlin

January 21, 2003

1 Exercice 1

Q1 : Soit y_i^* la tolérance. On a un modèle du type :

$$y_i = \begin{cases} \text{si } y_i^* = x_i\widetilde{\beta} + \varepsilon_i < \widetilde{\gamma} \\ \text{sinon} \end{cases}$$

Dès lors sachant que ε_i iid $(0, \sigma_\varepsilon)$. On a

$$\begin{aligned} p_i &= \text{Prob}\left(y_i = 1\right) \\ &= \text{Prob}\left(\frac{\varepsilon_i}{\sigma_\varepsilon} < \frac{\widetilde{\gamma}}{\sigma_\varepsilon} - x_i\frac{\widetilde{\beta}}{\sigma_\varepsilon}\right) \end{aligned}$$

Si $F(.)$ est la répartition associée à la loi de $\varepsilon_i/\sigma_\varepsilon$, on a donc :

$$p_i = F\left(\frac{\widetilde{\gamma}}{\sigma_\varepsilon} - x_i\frac{\widetilde{\beta}}{\sigma_\varepsilon}\right) \qquad ((1 \text{ pt}))$$

Q2 : cf. cours. (1 point)

Q3 : On pose $F(.) = \Lambda(.)$. On sait que si X suit une loi logistique, alors

$$E(X) = 0 \qquad E(X^2) = \frac{\pi^2}{3}$$

Dans notre problème :

$$\begin{aligned} p_i &= \text{Prob}\left(\frac{\varepsilon_i}{\sigma_\varepsilon} < \frac{\widetilde{\gamma}}{\sigma_\varepsilon} - x_i\frac{\widetilde{\beta}}{\sigma_\varepsilon}\right) \\ &= \text{Prob}\left(\frac{\pi}{\sqrt{3}}\frac{\varepsilon_i}{\sigma_\varepsilon} < \frac{\pi}{\sqrt{3}}\frac{\widetilde{\gamma}}{\sigma_\varepsilon} - x_i\frac{\pi}{\sqrt{3}}\frac{\widetilde{\beta}}{\sigma_\varepsilon}\right) \end{aligned}$$

On pose $v_i = \dfrac{\pi}{\sqrt{3}} \dfrac{\varepsilon_i}{\sigma_\varepsilon}$ et l'on pose

$$\gamma = \frac{\pi}{\sqrt{3}} \frac{\widetilde{\gamma}}{\sigma_\varepsilon} \qquad \beta = \frac{\pi}{\sqrt{3}} \frac{\widetilde{\beta}}{\sigma_\varepsilon}$$

Dès lors, la probailité devient :

$$p_i = \text{ Prob}(v_i < \gamma - x_i\beta)$$

où les résidus normalisés v_i ont une variance égale à $\pi^2/3$ par construction

$$E(v_i) = \frac{\pi}{\sigma_\varepsilon\sqrt{3}} E\left(\varepsilon_i\right) = 0 \qquad ((0.5 \text{ pt}))$$

$$E(v_i^2) = \left(\frac{\pi}{\sigma_\varepsilon\sqrt{3}}\right)^2 .E(\varepsilon_i^2) = \frac{\pi^2}{\sqrt{3}}$$

Donc si l'on suppose que v_i est distribué selon une loi logistique, on obtient un modèle logit :

$$p_i = \text{Prob}(y_i = 1) = \Lambda\left(\gamma - x_i\beta\right) \qquad (1)$$

$$\text{avec } \gamma = \frac{\pi}{\sqrt{3}} \frac{\widetilde{\gamma}}{\sigma_\varepsilon} \qquad \beta = \frac{\pi}{\sqrt{3}} \frac{\widetilde{\beta}}{\sigma_\varepsilon} \qquad ((1.5 \text{ point}))$$

Conséquences :

- interprétation des paramètres impossibles :

(0.5 pt)

- seuls les signes sont interprétables

Q4 : Soit

$$y_i = \begin{cases} 1 & \text{si } y_i^* < \gamma \\ 0 & \text{sinon} \end{cases}$$

On a donc :

$$L(y, \beta) = p_i y_i (1 - p_i)^{(1-y_i)}$$

(ou)

$$\log\ L(y,\beta) = \sum_{i=1}^{N} y_i\ \log\ \Lambda\left(\gamma - x_i\beta\right) + (1 - y_i)\ \ \log\left[1 - \Lambda\left(\gamma - x_i\beta\right)\right]$$

Connaissant $\Lambda(.)$ on a donc

$$\log\ L(y,\beta) = \sum_{i=1}^{N} y_i\ \log\ \left(\frac{e^{\gamma - x_i\beta}}{1 + e^{\gamma - x_i\beta}}\right) + (1 - y_i)\ \ \log\left[1 - \frac{e^{\gamma - x_i\beta}}{1 + e^{\gamma - x_i\beta}}\right]$$

(1 pt)

Ici on a :

$$\begin{aligned}
G(p) &= \frac{\partial \log \ L(y,\beta)}{\partial \beta} \\
&= \sum_{i=1}^{N} y_i \frac{\lambda\left(\gamma - x_i\beta\right)}{\Lambda\left(\gamma - x_i\beta\right)}(-x_i') + (1-y_i)\,\frac{-\lambda\left(\gamma - x_i\beta\right)}{1-\Lambda\left(\gamma - x_i\beta\right)}(-x_i')
\end{aligned}$$

où $\lambda\left(x\right)$ désigne la densité de loi logistique. Sachant que $\lambda(x) = \Lambda(x)[1 - \Lambda(x)]$. On a donc :

$$G\left(\beta\right) = \sum_{i=1}^{N} -y_i\left[1-\Lambda(\gamma - x_i\beta)\right]x_i' + (1-y_i)\Lambda(\gamma - x_i\beta)x_i'$$

$$\Rightarrow G(\beta) = \sum_{i=1}^{N} -y_i x_i' + \Lambda(\gamma - x_i\beta)x_i'$$

D'où :

$$G(\beta) = \sum_{i=1}^{N}\left[\Lambda\left(\gamma - x_i\beta\right) - y_i\right]x_i' \tag{2}$$

Commentaire (**0.5 pt**)

→ Forme non linéaire sur β, γ
→ Pas d'expression analytique par estimateur du MV $\widetilde{\beta}$ et $\widetilde{\gamma}$

→ Recours à des algorithmes d'optimisation numérique.

Q5 : On suppose $\gamma = 0$, $K = 1$. Soit $\widehat{\beta}$ l'estimateur du MV de β. Sous les conditions mentionnées, on sait que :

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow[N\to\infty]{L} N(0, I(\beta)^{-1})$$

Soit $\widehat{I}(\widehat{\beta})$ un estimateur convergent de la matrice d'information de Fischer. Dans un modèle Logit, on sait que :

$$\begin{aligned}
I(\beta) &= -E\left[\frac{\partial^2 \text{log} L(y,\beta)}{\partial \beta^2}\right](0.5pt) \\
\Rightarrow \quad I(\beta) &= \sum_{i=1}^{N} \frac{\exp(x_i\beta)}{[1+\exp(x_i\beta)]^2}x_i^2
\end{aligned}$$

Dès lors, on sait que :

$$\widehat{I}(\widehat{\beta}) = \sum_{i=1}^{N} \frac{\exp(x_i\widehat{\beta})}{[1+\exp(x_i\widehat{\beta})]^2}x_i^2 \tag{1.5}$$

Donc, si on définit C^α le quantile de la loi normale centrée réduite à $\alpha\%$. On a :

$$\text{prob }[\widehat{\beta} - C^\alpha \widehat{I}(\widehat{\beta})^{-1/2}\ ,\ \widehat{\beta} + C^\alpha \widehat{I}(\widehat{\beta})^{-1/2}] = 1 - \alpha \qquad \text{(1 pt)}$$

Q6 : On pose $\gamma = 0$, $F = 1$, dès lors il vient :

$$\frac{\partial p_i}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{\exp\ (x_i\beta)}{1+\exp(x_i\beta)}\right)$$

$$\frac{\partial p_i}{\partial x_i} = \frac{e^{x_i\beta}}{(1+e^{x_i\beta})^2}\beta$$

Dès lors :

$$\begin{aligned} \varepsilon_{p_i/x_i} &= \frac{\partial p_i}{\partial x_i}.\frac{x_i}{p_i} \\ \Rightarrow\ \varepsilon_{p_i/x_i} &= \frac{e^{x_i\beta}}{(1+e^{x_i\beta})^2}\frac{x_i\beta}{e^{x_i\beta}}(1+e^{x_i\beta}) \end{aligned}$$

$$\varepsilon_{p_i/x_i} = \frac{x_i\beta}{1+e^{x_i\beta}} \qquad \text{((1.5 pt))}$$

Si x_i augmente de 1%, alors la probabilité de l'évènement $y_i = 1$ augmente de $\varepsilon_{p_i/x_i}\%$.

$$\text{Cote} = c_i = \frac{p_i}{1-p_i} = e^{x_i\beta} \qquad \text{((0.5 pt))}$$

Il y a $e^{x_i\beta}$ fois plus de chances que l'évenement $y_i = 1$ se réalise qu'il ne se réalise pas.

2 Exercice 2

Q1 (**2 points**) : Une hausse du revenu médian diminue la probabilité de vote démocrate.
(*) Une hausse du taux de scolarisation augmente la probabilité du vote démocrate
(*) Une hausse de l'urbanisation augmente la probabilité du vote démocrate
+ discussion sur la significativité des coefficients (z-statistique).

Q2 (**2 points**):

$$\begin{aligned} x_i\beta = \ & -12672 \times 0,000406 \\ & +12,4 \times 0,269580 \\ & +714 \times 0,035902\ = 0,76 \end{aligned}$$

Donc

$$\text{prob (Texas vote démocrate)} = \Phi\,(0.76) \simeq 77.64\%$$

(*) cote $= \dfrac{\widehat{p}}{1-\widehat{p}} = \dfrac{0.7764}{1-0.7764} = 3.4723$

Il y a environ 3.5 fois plus de chance que le vote démocrate se réalise qu'il ne se réalise pas pour l'Etat du Texas .

(*) prévision de votes ultérieurs pour des éléctions présidentielles futures par projection des variables explicatives dans le futur. Limites : stabilité du modèle et hypothèse IIA sans doute pas vérifiée en matière d'élections. L'introduction d'un nouveau candidat ou parti va sans doute déformer les rapports de probabilité des partis en place.

Q3 (**1.5 points**) :

1. coefficient statistiquement $\neq$ de 0 à 5% (z statistique)

2. Nombre de prévisions fausses : Pour les 27 oservations de vote républicains, le modèle prévoit 20 vote pour les républicains et 7 pour les démocrates, soit un taux de réussite de 75%. Pour les 24 états ayant voté démocrates, le modèle prévoit 6 votes républicains et 18 démocrates, soit 75% de prévisions correctes

$\Rightarrow$relativement bon modèle

Q4 (**1.5 points**) choix du logit/probit : cf cours

soit $\widehat{\beta}_{LM}$ l'estimateur de β dans modèle linéaire et $\widehat{\beta}_L$ l'estimateur du modèle logit, $\widehat{\beta}_p$ du modèle probit. On a

$$\widehat{\beta}_L \simeq 1.6\widehat{\beta}_p$$

$$\widehat{\beta}_{LM} = 0,4\widehat{\beta}_p$$

	$\widehat{\beta}_L$	$\widehat{\beta}_{LM}$
INC	-0,0006496	-0,0001624
SCHOOL	0,431328	0,107832
URBAN	0,0574432	0,0143608

Q5 (**2 points**): On sait que pour un probit

$$\frac{\partial p_i}{\partial x_i^{(J)}} = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_i\beta)^2\right].\beta_J$$

Donc pour le Texas :

$$\begin{aligned} x_i\beta &= 0,76 \\ \beta_J &= -0,000406 \end{aligned}$$

Donc

$$\frac{\partial p_i}{\partial x_i^J} = -6,84 \cdot 10^{-5}$$

D'où :

$$\begin{aligned} \varepsilon_{p_{i/x_i^J}} &= \frac{\partial p_i}{\partial x_i^J} \cdot \frac{x_i^J}{p_i} \\ \Rightarrow \quad & \varepsilon_{p_{i/x_i}} = -6,84.10^{-5} \cdot \frac{12672}{0,7764} \\ & \Rightarrow \varepsilon_{p_{i/x_i}} = -1,11 \end{aligned}$$

Si INC baisse de 20%, la probabilité de vote démocrate baisse de 22. 2%..

Q6 : Stat

$$LRT = -2\left[\log L\left(y, \widehat{\beta}\right) - \log L\left(y, \widehat{\beta}^c\right)\right] \underset{N\to\infty}{\xrightarrow{L}} \chi^2(2) \tag{3}$$

Ici,

$$LRT = 6.6275 \tag{4}$$

Seuil à 5% d'un $\chi^2(2) = 5.99$.. On rejette l'hypothèse de nullité jointe des coefficients.