Université Abderrahmane Mira-Bejaia
Faculté des Sciences Économiques, Commerciales et des Sciences de Gestion

Département des sciences économiques

Polycopié pédagogique

Dossier numéro : .....

Titre

Econométrie des variables qualitatives

Cours destiné aux étudiants de

Master II spécialité Economie Quantitative

Année : 2020/2021

Intitulé du cours	**Econométrie des variables qualitatives**
Niveau/Semestre	Master 2/ S1
Enseignant	Dr.Rafika ZIDAT
Volume horaire cours	50
Volume horaire TD	
Modalité d'évaluation: Examen à la fin du cours	

Coordonnées de l'équipe pédagogique:

E-mail: rafika.zidat@univ-bejaia.dz

Introduction

La formation *Master Economie Quantitative* a permet aux étudiants de ladite spécialité l'acquisition et la maîtrise des modèles à variables dépendantes quantitatives. Toutefois, l'afflux considérable des données de nature microéconomique, à partir des années 60 et 70, très souvent de nature qualitative, exemple: les catégories socioprofessionnelles, le type d'étude suivie, travailler ou être en chômage, acheter ou ne pas acheter...etc., a donné naissance à de nouveaux modèles économétriques vu que les méthodes d'inférence traditionnelles ne peuvent traiter et estimer des modèles économétriques à variables dépendantes qualitatives.

Notre cours traite, ainsi, les modèles d'estimations spécifiques, les plus courantes en économétrie des variables qualitatives. Ces modèles ont été très répandues dans les domaines de biologie, sociologie et de la psychologie. Or, son utilisation, sur des données économiques, est initiée par les travaux de McFadden, D.L (1974) et de Heckman, J. (1976).

Dans ce cours, nous allons, donc, traiter, dans un premier temps, les modèles dits dichotomiques si $K=2$. Par la suite, nous évoquerons les modèles polytomiques $(K \in N)$. Enfin, nous concerterons le modèle *Tobit* qui est entre les modèles à variables qualitatives et les modèle linéaire usuels.

- ✓ **Objectifs du cours :**

Ce cours couvre les modèles économétriques dont les variables dépendantes sont discrètes. Ils incluent, donc, les modèles à réponse dichotomiques, polytomiques, les modèles pour données censurées et tronquées. L'objectif étant d'expliquer la variable dépendante de nature qualitative à travers les modèles : Logit, Probit et les modèles Tobit.

- Identifier le modèle économétrique le plus adapté à un problème empirique ;
- Interprétation des résultats des simulations.

- ✓ **Contenu :**

Introduction

Chapitre 01 : Introduction aux variables qualitatives.

Chapitre 02 : Les modèles dichotomiques simples.

Chapitre 03 : Les modèles à choix multiples.

Chapitre 04 : Les modèles tronqués et censurés.

Conclusion

- ✓ **Modalités d'évaluation :**

Tout d'abord, nous allons procéder par une évaluation diagnostique. Cette dernière se situe avant la séquence d'apprentissage, et permet de faire le point sur les connaissances de l'étudiant. Ensuite, un contrôle continu sera comptabilisé sur la base de différents critères d'évaluation à savoir : L'assiduité qui sera notée sur 5 (5/5), une interrogation (10/10) ainsi que la participation notée sur 5 (5/5). Enfin, un examen sommatif à la fin du Semestre 1.

Chapitre 01 : Introduction aux variables qualitatives

Dans ce chapitre, nous allons évoquer, dans un premier temps la classification des variables qualitatives. Par la suite nous arborerons les problèmes liés à l'estimation par la méthode des moindres carrés ordinaires des modèles à variables dépendantes qualitatives.

1.1. Classification des variables qualitatives

Les variables qualitatives sont assimilées à des variables discrètes qui prennent leurs valeurs dans un ensemble d'entiers naturels. Nous considérons, ainsi, deux grandes catégories de variables discrètes : les variables binaires et les variables polytomiques**.**

1.1.1. Les variables binaires:

Une variable qualitative binaire possède au maximum deux modalités. Elles sont aussi appelées variables ***dichotomiques*** qui sont codées par *0* et *1*. Considérons l'exemple de la variable qualitative Y = «emploi» pouvant prendre 2 modalités : "Être au chômage ", "avoir un emploi", pour coder cette variable qualitative, il faudrait associer à y une variable quantitative x qui peut prendre deux valeurs réelles distinctes; 0 et 1.Ainsi, on définit, par exemple, la variable x de la façon suivante :

$$x = \begin{cases} 1 & si\ Y = \text{ avoir un emploi} \\ 0 & si\ Y = \text{Être } au \text{ chômage} \end{cases}$$

Le 1 indique l'existence de la caractéristique, 0 son absence. L'espérance d'une variable binaire qui prend les valeurs 0 et 1 est la fréquence empirique observée sur un échantillon de cas où la variable vaut 1. Donc l'espérance d'une variable binaire *x* est:

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

Où: N: la taille de l'échantillon;

p(*x*): la fréquence empirique d'individus ayant la caractéristique *x=1* sur l'échantillon;

Si la probabilité théorique que x =1 peut être estimée par la fréquence empirique, l'espérance d'une variable binaire est synonyme de probabilité d'occurrence de la caractéristique à laquelle *x* fait référence [1]

[1]Thomas, A. (2000). « Économétrie des variables qualitatives», éd DUNOD, paris. P2.

1.1.2. Les variables polytomiques:

Elles prennent plus de deux modalités possibles. Prenant l'exemple de la variable qualitative y = «niveau d'étude» pouvant prendre 3 modalités : «licence», «master», «doctorat». Afin de coder la variable qualitative y, il s'agit tout simplement d'associer à y une variable quantitative x pouvant prendre trois valeurs réelles distinctes (a, b, c) ∈ R^3.

$$x = \begin{cases} 1 & \text{si y = "Licence"} \\ 2 & \text{si y = "Master"} \\ 3 & \text{si y = "Doctorat"} \end{cases}$$

Considérant d'autres exemples où nous distinguons plusieurs catégories de variables multinomiales:

- La variable polytomique ordonnée: «Revenu»

$$\text{Revenu} = \begin{cases} 1 & \text{si l'individu gagne moins de 24000 da/mois} \\ 2 & \text{si l'individu gagne entre 24000 et 38 000 da/mois} \\ 3 & \text{si l'individu gagne entre 38000 et 46000 da/mois} \\ 4 & \text{si l'individu gagne entre 38000 et 46000 da/mois} \\ 5 & \text{si l'individu gagne entre46000 et 69000da /mois} \\ 6 & \text{si l'individu gagne plus de 69000da/mois} \end{cases}$$

- la variable polytomique non ordonnée: «Emploi»

$$\text{Emploi} = \begin{cases} 1 & \text{si ouvrier} \\ 2 & \text{si employé} \\ 3 & \text{si cadre} \\ 4 & \text{si sans profession} \end{cases}$$

- la variable polytomique séquentielle «Éducation»: chaque niveau est conditionné par l'obtention du diplôme du niveau précédent.

$$Education = \begin{cases} 1 & si\ l'individu\ a\ le\ baccalauréat\ mais\ pas\ la\ Licence \\ 2 & si\ l'individu\ a\ \ la\ licence\ mais\ pas\ le\ master \\ 3 & si\ l'individu\ a\ le\ master\ mais\ pas\ le\ doctorat \\ 4 & si\ l'individu\ a\ \ le\ doctorat \end{cases}$$

1.2. Les problèmes de la spécification binaire:

Supposons qu'on observe N observations, Yi; ∀i = 1;….. N; d'une variable endogène dichotomique, et parallèlement on observe K variables exogènes Xi = (X_{i1};… X_{iK});et soit $\beta = \beta_1 \dots \dots \beta_K$ le vecteur des paramètres à estimer ; ∀i = 1; ……N.

Dans ce cas le modèle linéaire simple s'écrit :

$$Y_i = \beta X_i + \varepsilon_i \forall \mathrm{i} = 1, \dots \mathrm{N}.$$

Pour mieux comprendre, appliquons une modélisation linéaire simple au cas d'une variable endogène dichotomique :

On pose :

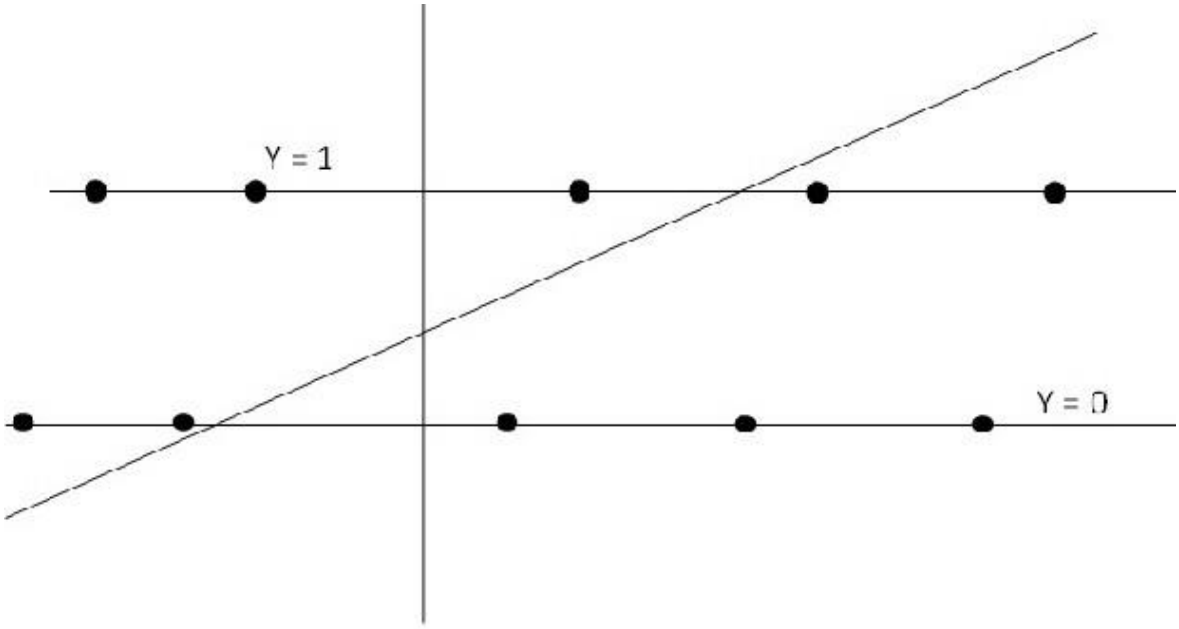
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \forall \mathrm{i} = 1, \dots \mathrm{N}.$$

$\beta_0 \, et \, \beta_1 \, paramètres \, du \, modèle$;

$X_i : variables \, explicatives$

Premièrement, une étude graphique montre que l'approximation linéaire n'est pas adaptée au problème posé, il suffit de se placer dans un repère (X_i;Y_i) et de reproduire les différents couples (X_i;Y_i) $\forall \mathrm{i} = 1, \dots \mathrm{N}$. Le nuage de points ainsi obtenu, se situe soit sur la droite Y=0, soit sur la parallèle Y = 1. Ainsi la figure ci-dessous illustre les deux droites parallèles:

Figure01 : Ajustement linéaire



Il est impossible d'ajuster de façon satisfaisante, pour une seule droite, le nuage de points associé à une variable dichotomique qui, par nature, est réparti sur deux droites parallèles. Deuxièmement, la spécification linéaire standard ne convient pas, aux variables dichotomiques, et plus généralement aux variables qualitatives, car elle pose un certain nombre de problèmes mathématiques. Quand la variable à expliquer, notée Y_i, est dichotomique c.-à-d. : Y_i prend la valeur *0* ou *1* en fonction de variables explicatives X_i, le modèle à estimer prend généralement la structuration d'une spécification de la probabilité à observer $Y_i = 1$ sachant les variables exogènes X_i.Autrement dit : $P_i = P\ (Y_i=1/\ x_i)$. Cette probabilité définit carrément la loi conditionnelle de Y_i sachant X_i. Par ailleurs, la probabilité d'observer $Y_i = 0$ correspond à la probabilité $(1- P_i)$.

Le modèle dichotomique, dans ce cas, correspond à une fonction linéaire ; $P_i=F(X_i\ \beta)$. Ces modèles à probabilité linéaire sont pratiquement assimilés aux modèles linéaires. Ainsi, la fonction F(.) définit la distribution de la probabilité de Y_i et son espérance conditionnelle.

$$E(Y_i/x_i) = 1 \times prob(Y_i = 1) + 0 \times prob(Y_i = 0) = P_i$$

Soit :

$$P_i = \beta_1 X_i\ , \forall\ i.$$

De ce principe, le modèle est transformé en une équation de régression telle que :

$$Y_i = F(x_i\ \beta) + \varepsilon_i$$

.

Si la fonction F(.) est supposée linéaire, le modèle sera définit, donc, par la relation :

$$E(Y_i/x_i) = P(Y_i = 1/x_i) = x_i\beta.$$

Le modèle pourrait, donc, être estimé par les méthodes d'estimation linéaires classiques, à savoir : les méthodes d'estimation MCO et MCG. [2]

Cependant, la modélisation linéaire avec les méthodes d'ajustement linéaires n'exige pas le respect des valeurs à prendre par Y_i dans l'intervalle [0, 1]. Autrement dit, l'estimation du modèle par les méthodes linéaire, généralement les moindres carrés

[2]Crépon, B. Jacquemet, N. (2010), « Econométrie : Méthode et applications », Edition De Boeck, Bruxelles, P331.

ordinaires, n'oblige pas $P(Y_i = 1/x_i) = x_i\beta$ à prendre ses valeurs dans l'intervalle [0, 1] : $0 \leq \beta X_i \leq 1$

Ainsi, les valeurs prédites en termes de probabilités peuvent être supérieurs à 1 ou négatives.

Par ailleurs, la spécification linéaire implique, aussi, que le terme aléatoire ε_i ne peut prendre, lui aussi, que deux valeurs, conditionnellement au vecteurX_i:

- $\varepsilon_i = 1 - Xi\ \beta$ avec une probabilité de pi $=$ P (Yi $=$ 1).
- $\varepsilon_i = -Xi\ \beta$ avec une probabilité de$(1 - \text{pi}) =$ P (Yi $=$ 0).

Ainsi le terme aléatoire ε_i du modèle admet nécessairement une loi discrète, ce qui exclut en particulier l'hypothèse de normalité des résidus.

Enfin, l'estimation des variables qualitatives dépendantes par la méthode des moindres carrés, demeurer une problématique. Ceci est lié à la présence d'hétéroscédasticité. En effet, nous constatons, immédiatement, que dans le modèle $Y_i = \beta X_i + \varepsilon_i$ la matrice de variance covariance des résidus varie entre les individus en fonction de leurs caractéristiques associées aux exogènes puisque:

$$V(\varepsilon_i) = Xi\ \beta(1 - Xi\ \beta)\forall i = 1, \ldots \text{N}.$$

Pour démontrer ce résultat il suffit de considérer des résidus et de calculer la variance:

$$V(\varepsilon_i^2) = (1 - Xi\ \beta)^2 P(Y_i = 1) + (-Xi\ \beta)^2 P(Y_i = 0)$$
$$= (1 - Xi\ \beta)^2 p_i + (-Xi\ \beta)^2 (1 - p_i)$$

Sachant que d'après la relation précédente, on a :

$$p_i = Xi\ \beta$$

On en déduit que:

$$V(\varepsilon_i) = (1 - Xi\ \beta)^2 Xi\ \beta + (-Xi\ \beta)^2 (1 - Xi\ \beta) = (1 - Xi\ \beta) Xi\ \beta[(1 - Xi\ \beta) + Xi\ \beta]$$
$$= (1 - Xi\ \beta) Xi\ \beta$$

De plus le problème d'hétéroscédasticité ne peut pas être résolu par la méthode d'estimation des Moindres carrés Généralisés. Puisque la matrice de variance covariance des erreurs dépend du vecteur des paramètres à estimer dans la spécification linéaire, qui par nature sont supposés inconnus.

Pour ces raisons la spécification linéaire des variables dichotomique, n'est jamais utilisée.

- **A retenir**

a) L'estimation de l'équation$Y_i = X_i\beta + \varepsilon_i$, avec Y_i= 0 ou Y_i=1, est totalement arbitraire. A vrai dire, si nous estimons l'équation avec la méthode des moindres carrés ordinaires en prenant une codification différente de [0, 1] ; exemple [0, 5], les coefficients β_k seraient certainement différents.

b) L'hypothèse de normalité des erreurs n'est pas versifiée. Puisque la variable dépendante y_i prend deux valeurs, l'erreur ne peut prendre que deux valeurs :

$\varepsilon_i = 1 - (X_i\beta)\ pour\ \ P_i = P(Y_i{=}1)$
$\varepsilon_i = -(X_i\beta)\ \ pour\ \ (1 - P_i) = P(Y_i{=}0)$

c) L'existence d'hétéroscédasticité d'après l'équation suivante:

$$E(\varepsilon_i^2) = V(\varepsilon_i) = P_i(\text{la valeur de } \varepsilon_i \text{ si } Y_i = 1)^2 + (1 - P_i)(\text{la valeur de } \varepsilon_i \text{ si } Y_i = 0)^2$$

D'où:

$$V(\varepsilon_i) = E(\varepsilon_i^2) = P_i(1 - \beta_0 - \beta_1 x_i)^2 + (1 - P_i) \times (-\beta_0 + \beta_1 x_i)^2$$

Or :

$$P_i = \beta_0 + \beta_1 x_i$$

Nous aurons donc :

$$V(\varepsilon_i) = E(\varepsilon_i^2) = P_i(1 - P_i)^2 + (1 - P_i) \times (P_i)^2 = P_i(1 - P_i)$$

d) La contrainte $0 \leq X_i\beta \leq 1$ n'est pas imposée dans l'équation suivante:

$$V(\varepsilon_i) = E(\varepsilon_i^2) = P_i(1 - P_i)$$

Chapitre02 : Les modèles dichotomiques simples:

Afin de remédier aux lacunes de la méthode des moindres carrés ordinaires sur des modèles économétriques à variables dépendantes qualitatives à savoir:

- La présence d'hétéroscédasticité;
- La nature discrète du résidu;
- Les valeurs estimées peuvent être en dehors de l'intervalle [0, 1] : $0 \leq X_i\beta \leq 1$

La solution est que la réalisation de la variable dépendante $Y_i = 0\ ou$ à 1 provient d'une règle de décision. Cette dernière est un système reliant les variables explicatives x_i à la réalisation de l'événement $\{Y_i = 0\}$ ou$\{Y_i = 1\}$. À ce stade, nous supposons que la réalisation de l'évènement $\{Y_i = 1\}$ est attribuée à des valeurs élevées des $x_i$et la réalisation de l'évènement $\{Y_i = 0\}$ est associée à des valeurs faibles desx_i.

Il devrait, donc, exister un certain seuil de valeur qui dépend de l'équation linéaire; $\boldsymbol{x_i\beta}$, au-delà de laquelle la proportion des $\{Y_i = 1\}$ l'emporte sur celle des $\{Y_i = 0\}$.

Notons cette valeur seuil « c ». Nous supposons également que cette règle de décision n'est pas déterministe. Autrement dit, certaines observations Y_i sont égales à 0 tandis que les valeurs des variables explicatives x_i sont élevées. Par conséquent, nous devons intégrer à notre équation linéaire un terme d'erreur notéε_i. La règle de décision est donc:

- ***Règle déterministe***: $\{Y_i = 1\}$ proportion élevée pour $x_i\beta > c$
  $\{Y_i = 0\}$Proportion faible pour $x_i\beta \leq c$

- ***Règle probabiliste*** (aléatoire): $\text{Prob}(y_i = 1) = \text{Prob}(x_i\beta + \varepsilon_i > c\ ) \cdots\cdots\cdots\cdots$ A
  $\text{Prob}(Y_i = 0) = \text{Prob}(x_i\beta + \varepsilon_i \leq c\ )$ $\cdots\cdots\cdots\cdots$ B

2.1. Le modèle latent:

La variable latente Y_i^* est une variable inobservable contrairement à la variable binaireY_i, qui est donnée (observée). De ce fait, les variables latentes s'agissent de variables continues, non observables et représentatives du phénomène étudié (Y_i) et reflète donc les mécanismes économiques à l'œuvre. Par contre, la variable Y_i est le résultat d'un processus sous-jacent.

La variable latente n'est qu'une solution au problème de mis en application des méthodes d'estimation des moindres carrés ordinaires(MCO) sur les variables dépendantes discrètes.

En considérant, donc, la variable latenteY_i^* :

$$Y_i = \begin{cases} 1 & si\ Y_i^* > c \\ 0\ si & Y_i^* \leq c \end{cases}$$

Où Y_i^* est une variable aléatoire en présence du terme de l'erreur :

$$Y_i^* = x_i\beta + \varepsilon_i$$

La règle de décision probabiliste devient alors :

$$\begin{cases} Prob(Y_i = 1) = Prob(x_i\beta + \varepsilon_i > c) = 1 - Prob(\varepsilon_i < c - x_i\beta) \\ Prob(Y_i = 0) = Prob(x_i\beta + \varepsilon_i \leq c) = \quad Prob(\varepsilon_i \leq c - x_i\beta) \end{cases}$$

Nous notons que la valeur seuil « c » est identique pour toutes les observations. Nous pouvons fixer arbitrairement la valeur seuil à 0 si « c » est un scalaire et si le vecteur des variables explicatives contient un terme constant. Ainsi, nous supposons que la proportion des $(Y_i = 1)$ est élevée pour :

$$Y_i = \begin{cases} 1 & si\ Y_i^* > 0 & si\ \varepsilon_i > -x_i\beta \\ 0 & sinon & si\ \varepsilon_i \leq -x_i\beta \end{cases}$$

Soit P la probabilité que $Y_i^* > 0$

$$P_i = prob(Y_i = 1) = Prob(Y_i^* > 0) = Prob(x_i\beta + \varepsilon_i > 0) = Prob(\varepsilon_i > -(x_i\beta))$$

L'écriture probabiliste de notre règle de décision dépend, donc, de la distribution statistique de la variable aléatoire de l'équation ε_i.

Par conséquent les probabilités de la réalisation des évènements $\{Y_i = 1\}$ et $\{Y_i = 0\}$ seront calculées en faisant référence à la loi statistique du terme d'erreur. Nous supposons que la distribution de ε_i est centrée par rapport à la moyenne ; de moyenne nulle nous obtenons :

$$Prob(\varepsilon_i > -(X_i\beta) = Prob(\varepsilon_i < (X_i\beta))$$
$$P_i = Prob(Y_i = 1) = Prob(\varepsilon_i < X_i\beta)$$

P_i dépend, donc, de la distribution de $\varepsilon_i$du modèle de décision. Nous distinguons deux lois statistiques les plus utilisées en pratique :

- ε_i suit une loi logistique afin d'estimer un modèle ***Logit ;***
- $\varepsilon_i$suit une loi de Plass Gauss (loi normale) pour estimer un modèle ***Probit.***

✓ ***Exemple de variable latente : (****tiré du polycopié de cours : HURLIN, C. (2003))*[3]

L'exemple le plus répandu, afin d'expliquer une variable latente, trouve ses origines en bio-économétrie, précisément celui de l'insecticide :

Dans un espace fermé, un insecticide est diffusé dans l'air afin de déterminer la dose minimale permettant de tuer les insectes. De ce fait, à l'issu d'une période fixée, on observe les insectes i morts codifié par Y_i= 0 et ceux encore vivants dont la valeur est Y = 1.

On suppose que Y_I^* traduit la capacité de résistance de chaque insecte, de manière à ce que la dose de produit est supérieure à ce seuil l'insecte est mort (Yi = 0), et qu'il reste vivant pour une dose inférieure (Y_i= 1). Il s'agit alors de modéliser la probabilité de survie de l'insecte i en fonction de la dose d'insecticide et des observations faites sur Y_i. On suppose pour cela qu'un certain dosage « c » est diffusé sur l'ensemble des insectes. On voit immédiatement que ce problème peut s'écrire de la façon suivante :

$$Y_i = \begin{cases} 1 si\ Y_i^* > c \\ 0 \quad sinon \end{cases}$$

Où la variable latente Y_i^* peut s'écrire comme la somme d'une combinaison linéaire de caractéristiques propres à chaque insecte et d'un terme aléatoire.

$$Y_i^* = X_i\beta + \varepsilon_i$$

[3] HURLIN, C. (2003), « Econométrie des Variables Qualitatives : chapitre I : Modèles dichotomiques univariés », Polycopié de cours, Université d'Orléans, P21.

Si le terme aléatoire ε_i est distribué selon une loi normale, nous allons donc estimer un modèle Probit, si ce terme est distribué selon une loi logistique nous estimons le modèle logit.

2.2. Modèles binaires Logit-Probit :

Il a été souligné, précédemment, que la variable latente est un élément crucial dans la modélisation de la survenu d'un évènement $\{Y_i = 0\}$ ou $\{Y_i = 1\}$, en utilisant sa probabilité. Pour que cette dernière notée « P_i » représente de manière optimale la variable d'intérêt « Y_i », il faut que la fonction de probabilité P_i soit croissante et bornée ; supérieurement par 1 et inférieurement par 0.

Il est vrai que, la fonction de répartition de n'importe quelle loi de probabilité remplie ces propriétés. A cet effet, pratiquement, deux lois de probabilités sont le plus utilisées afin de définir un modèle à variable dépendante qualitative.

2.2.1 Le modèle Probit :

Ce type de modèle repose sur la fonction de répartition de la variable aléatoire $\varepsilon_i$telle que :

$$\boldsymbol{P_i} = \int_{-\infty}^{\boldsymbol{\beta_0+\beta_1 x_i}} \boldsymbol{f(t)dt} = \boldsymbol{F}(.)$$

$$\text{Où : } \boldsymbol{f(t)} = \frac{\mathbf{1}}{\sqrt{\mathbf{2\pi}}}\boldsymbol{e^{-t^2/2}}$$

Le terme aléatoireε_i suit une normale centrée réduite$N(0,1)$. La probabilité associée à la variable expliquée dans le modèle *Probit* est donc :

$$P(Y_i = 1) = F(\beta X_i)$$

2.2.2 Le modèle Logit :

Ce modèle est donné par la fonction logistique **F :**

$$\boldsymbol{P_i} = \mathbf{F}(\boldsymbol{\beta X_i}) = \frac{\mathbf{exp}(\boldsymbol{X_i\beta})}{\mathbf{1} + \mathbf{exp}(\boldsymbol{X_i\beta})} = \frac{\mathbf{1}}{\mathbf{1} + \mathbf{exp}(-(\boldsymbol{X_i\beta}))}$$ [4]

La fonction Logit pourra se transformer, si nous notons P_i la probabilité que$P(Y_i = 1)$, nous aurons, alors, la représentation suivante :

[4] Greene, W. (2002). «Econometric analysis», Ed.Prentice Hall, (seventh edition), New Jersey.P667.

$$Log\left(\frac{P_i}{1-P_i}\right) = X_i\beta \Rightarrow Log\left(\frac{P_i}{1-P_i}\right) = Y_i^* = X_i\beta + \varepsilon_i,$$

$$\frac{P_i}{1-P_i} est\ La\ probabilité\ relative\ au\ choix\ Y_i = 1$$

2.3. Estimation des modèles dichotomiques :

Le maximum de vraisemblance est la méthode d'estimation la plus convenable aux modèles à variables dépendantes discrètes.

2.3.1. La définition de la vraisemblance :

L'estimation du modèle dichotomique par la méthode de maximum de vraisemblance consiste à choisir le vecteur de paramètres β de façon à maximiser la vraisemblance de Y_i

La vraisemblance est, donc, la probabilité d'observer un échantillon, étant donné les paramètres du processus ayant engendré les données. La vraisemblance est donnée par [5]:

$$\mathcal{L}(x,\beta) = \prod_{i=1}^{N} f(x_i,\beta)$$

Où : N: taille de l'échantillon constitué de paires $\{y_i, x_i\}$;

$i = 1,2 \cdots \cdots N$;

$y_i : variable\ dépendante$;

x_i: $variables\ explicatives$; $vecteur$ à $1 \times k\ dimension$.

Ainsi, si les paires $\{y_i, x_i\}$ sont indépendamment distribuées, la probabilité de l'échantillon est le produit des probabilités associées à chaque paire.

Soit $f(y_i, x_i, \beta)$ la fonction de densité conjointe associée au couple $\{y_i, x_i\}$ où $f(.)$ est une fonction croissante à valeurs positives et β un vecteur $k \times 1$ de paramètres.[6]

[5] Thomas, A. (2000). « Économétrie des variables qualitatives», éd DUNOD, p28.

[6] Wooldridge, F. (2002). «Econometric analysis of cross section and panel data», The MIT Press, Cambridge. P387.

2.3.2. Propriétés de l'estimateur :

L'estimateur du maximum de vraisemblance est définit par la solution du problème qui suit :

$$\boldsymbol{Max\ \mathcal{L}(x,\beta) = Max_{\beta} Log\ \mathcal{L}(x,\beta) \Leftrightarrow \frac{\partial Log(x,\beta)}{\partial \beta} = 0}$$

L'estimateur $\hat{\beta}$ est caractérisé par les propriétés suivantes :

- Convergence : plim$\beta = \beta$;
- Invariance à une transformation paramétrique : si $\beta^* = t(\beta)$ alors $\hat{\beta}^* = t(\hat{\beta})$;
- Efficacité asymptotique :$\hat{\beta}$ atteint la borne inférieure de Cramér-Rao ;
- Normalité asymptotique : $\hat{\beta} \sim N(\beta, [-E\partial^2 Log\mathcal{L}/\partial\beta'\partial\beta']^{-1})$

Ainsi, à partir d'une fonction de répartition connue, P de densité p, la probabilité conditionnelle $P(y_i = 1/x_i) = F(x_i, \beta)$ détermine la probabilité d'observer la variable expliquée y_i :

$$P(y_i = 1/x_i) = P(y_i = 1/x_i)^{y_i}[1 - P(y_i = 1/x_i)]^{1-y_i} = F(x_i\beta)^{y_i}[1 - F(x_i\beta)]^{1-y_i}$$

La vraisemblance d'un échantillon d'observations indépendantes s'écrit donc :

$$\mathcal{L}(y_i/x_i) = \prod_{i=1}^{N} P\left(\frac{y_i}{x_i}\right) = \prod_{i=1}^{N} F(x_i\,\beta_i)^{y_i}[1 - F(x_i\beta_i)]^{1-y_i}$$

Et la Log-vraisemblance :

$$Log\mathcal{L}_N = \sum_{i=1}^{N} [y_i\ log\ F(x_i\beta_i) + (1 - y_i)\log(1 - F(x_i\beta_i))]$$

La valeur des paramètres qui maximise la vraisemblance de l'échantillon est celle qui satisfait la condition :

$$\frac{\partial Log\ \mathcal{L}_N(\hat{\beta}_{Mv})}{\partial\beta}(\hat{\beta}_{Mv}) = \sum_{i=1}^{N} [Y_i \frac{f(x_i\hat{\beta}_{Mv})}{F(x_i\hat{\beta}_{Mv})} + (1 - Y_i)\frac{-f(x_i\hat{\beta}_{Mv})}{1 - F(x_i\hat{\beta}_{Mv})}]x_i'$$

$$= \sum_{i=1}^{N} [Y_i - F(x_i\hat{\beta}_{Mv})]\frac{f(x_i\hat{\beta}_{Mv})}{F(x_i\hat{\beta}_{Mv})[1 - F(x_i\hat{\beta}_{Mv})]}x_i' = 0$$

Notons que si $\boldsymbol{x_i}$contient un terme constant, les conditions de premier ordre impliquent que la moyenne des probabilités prédite doit être égale à la proportion de celle de l'échantillon. Cette implication présente une certaine similitude avec les équations linéaires estimées par la méthode des moindres carrés ordinaires, si nous supposons le terme $\left(Y_i - F(x_i\hat{\beta}_{Mv})]\frac{f(x_i\widehat{\beta}_{Mv})}{F(x_i\widehat{\beta}_{Mv})[1-F(x_i\widehat{\beta}_{Mv})]}\right)$ comme un résidu. [7]

Pour un échantillon de taille N, nous supposons l'exemple d'une vraisemblance déduite d'une distribution normale. La fonction de vraisemblance devient, ainsi, comme suit :

$$\varepsilon_i = Y_i - x_i\beta \sim N(0, \delta^2)$$

$$Log\ \mathcal{L} = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\delta^2) - \frac{1}{2\delta^2}\sum_{i=1}^{N}\varepsilon_i^2$$

Et sous la condition :

$$\partial log\mathcal{L}/\partial\delta^2 = 0$$

Nous avons :

$$\hat{\delta}^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - x_i\beta)^2$$

Et si nous la reportons dans l'expression de vraisemblance :

$$Log\mathcal{L}^* = -\frac{N}{2} - \frac{Nlog(2\pi)}{2} - \frac{N}{2}log[\frac{1}{N}\sum_{i=1}^{N}(Y_i - x_i\beta)^2]$$

Afin d'estimer les paramètres β_i de dimension k+1, nous devons, donc, résoudre la condition au premier ordre qui est la nullité du gradient de la log-vraisemblance. Cette dernière est non linéaire à cause des expressions $f(X_i\beta)$ et $F(X_i\beta)$ il n'est pas possible de donner une expression analytique simple de ces estimateurs, et leur calcul se fait généralement par la mise en œuvre d'un algorithme d'optimisation qui nécessite des prérequis solides pour les étudiants. Pour cela, nous avons jugé plus utile, de faire référence au logiciel Eviews afin de résoudre cette étape, en pratique.

[7] Greene, W. (2002), Op cit, P671.

2.3.3. Tests statistiques :

Afin de tester la pertinence d'une hypothèse sur les paramètres estimés par la méthode du maximum de vraisemblance, particulièrement la nullité des coefficients, nous construisons une statistique de test. Cette dernière n'est qu'une fonction aléatoire des données et des paramètres estimés. Par la suite, nous comparons la valeur de la statistique avec la valeur théorique tabulée, en choisissant le niveau de risque α (généralement pris à 5%).

Les tests statistiques les plus utilisés dans le cadre du maximum de vraisemblance sont le rapport de vraisemblance (***LRT*** : Likelihood Ratio Test), le test de ***Wald*** et le test du Multiplicateur de Lagrange (***LM*** : Lagrange Multiplier). Ces trois tests suivent asymptotiquement une distribution de Khi-deux (χ^2) avec k degrés de liberté (nombre de variables explicatives).

Toutefois, le test le plus naturel est le rapport de vraisemblance :

$$L_R = -2[log\mathcal{L}(\hat{\beta}) - log\mathcal{L}(\hat{\beta}^c)]$$

Où : $\hat{\beta}^c$: $est\ l'estimateur du\ maximum\ de\ vraisemblance\ sous\ la\ contraite.$

Sur le plan pratique, afin de tester l'hypothèse H_0: $\beta_1 = \beta_2 = \cdots\cdots = \beta_k = 0$, nous faisons référence au rapport de Log-vraisemblance$L_R = -2\big(Ln(LR)\big) - Ln(Lu))$. Comme il a été précisé ci-dessus L_R suit une loi de χ^2 sous l'hypothèse H_0 à k degrés de liberté. Si la statistique L_R est supérieure au χ^2théorique tabulée pour un risque α, l'hypothèse H_0sera rejetée. Ainsi, le modèle estimé comporte au moins une variable explicative significative.

Quant à la significativité des coefficients, elle est évaluée à l'aide des ratios « $\mathcal{Z}$-statistique », puisque le rapport de l'estimateur sur son écart type suit une loi normale. La statistique « $\mathcal{Z}$ » s'interprète, donc, à partir des probabilités critiques.

Par ailleurs, les coefficients estimés ne sont pas directement interprétables, en termes de valeur. Contrairement à ces modèles, les modèles à variables qualitatives (Logit comme Probit) détiennent l'inconvénient que la variation marginale d'un coefficient a un effet différent d'un individu à un autre. Autrement dit, dans les modèles Logit ou/et Probit, seuls les signes des estimateurs peuvent nous renseigner sur les signes des estimateurs peuvent nous informer de l'effet des variables explicatives sur la variable $y_i = 1$ de référence. Vu que l'effet marginal d'une variable explicative est :

$$\frac{\partial E(y_i/x_i)}{\partial x_i^k} = P'(x_i\beta)\beta_k$$

L'élasticité quant à elle, est de :

$$\frac{\partial \log \mathrm{E}(y_i/x_i)}{\partial x_i^k} = \frac{P'(x_i\beta)}{P(x_i\beta)}\beta_k$$

En raison de la présence de la probabilité dans l'équation, l'effet marginal diffère donc, d'un groupe à un autre ($Y_i = 1\ ou\ Y_i = 0$) . Pour cela, nous nous contenterons de l'interprétation du signe associé au régresseur.

Afin de tester l'ajustement global du modèle; nous ferons appel à la statistique nommée Pseudo $-\ R^2$ mesurée par :

$$R^2 = 1 - \frac{Log\ (L_U)}{Log\ (L_R)}\ (R^2\ de\ McFadden)$$

2.3.4. Comparaison entre le modèle Logit et Probit :

La modélisation de variables dichotomique tend vers l'utilisation de deux types de modèles similaires en termes d'ajustement statistique. Leurs deux distributions, normale et logistique, sont de la famille lois exponentielles. En effet, les différences en termes d'estimation sont apparentes lorsqu'il s'agit d'un échantillon très grand, vu que les deux distributions divergent aux extrémités (faibles et fortes valeurs de fonction de répartition).

Ainsi, la différence apparente est celle des valeurs estimées qui ne sont pas directement comparables. Ceci dit, afin de confronter le Probit au Logit, nous devons multiplier les coefficients du modèle Probit par$\frac{\pi}{\sqrt{3}} \approx 1.81$. Inversement, il faut multiplier les estimateurs du Logit par $\frac{\sqrt{3}}{\pi}$ pour les comparer aux coefficients issus d'une modélisation Probit.

2.4. Exercices d'application sur les modèles dichotomiques *: estimation d'un Logit Binaire :*

2.4.1. Exercice 01 :

La question du choix de régime de change reste l'une des problématiques les plus controversées en finance internationale. A cet effet, nous allons tenter de répondre à la question du choix de régime de change pour l'Algérie sur la période allant de 1970 à 2010. L'étude est menée à l'aide d'un modèle logit binaire.

Nous avons considéré, à cette fin, un large éventail de déterminants du choix de régime de change. De ce fait, les déterminants faisant l'objet de l'analyse sont : le degré d'ouverture de l'économie (**DEGOUV**), le niveau du développement économique (**PIBHAB**), le taux de croissance de la masse monétaire (**TCM2R**), le taux de croissance du PIB réel (**TCPIBR**), le taux d'inflation (**TXINF**) et la performance des finances publiques (**PERFINPUB**).

Pour estimer le choix d'un régime de change optimal, nous utilisons une variable discrète Y_t qui prend une valeur égale à zéro (**0**) si le régime de « **change fixe »** est choisi durant la période t et égale à un (**1**) si le régime de change « **flottant géré »** est choisi.

Les résultats de la manipulation économétrique du modèle **LOGIT** est reportée ci-dessous.

On vous demande de :

1- Procéder à la validation du modèle à travers l'interprétation statistique des résultats (tests statistiques et lecture de la table de prédiction des résultats).

2- Faire l'interprétation économique des résultats de l'estimation du modèle logit binaire.

Tableau 01

Estimation des paramètres par un logit binaire

Variable	Coefficient	Std. Error	z-Statistic	Prob
C	115.174	77.735	1.481	0.1384
DEGOUV	70.461	28.106	2.506	0.0122
PIBHAB	-12.156	8.406	-1.446	0.1481
TCM2R	-0.229	0.13	-1.766	0.0774
TCPIBR	0.25	0.141	1.766	0.0773
TXINF	-4.303	2.262	-1.902	0.0572
PERFINPUB	-5317946	3176706	-1.674	0.0941

McFadden R-squared	0.711635	Mean dependent var	0.390244
S.D. dependent var	0.493865	S.E. of regression	0.266096
Akaike info criterion	0.727214	Sum squared resid	2.407443
Schwarz criterion	1.019775	Log Likelihood	-7.90788
Hannan-Quinn criter	0.833748	Deviance	15.81576
Restr. Deviance	54.84628	Restr. log Likelihood	-27.42314
LR statistic	39.03052	Avg. log Likelihood	-0.192875
Prob(LR statistic)	0.000001		
Obs with Dep=0	25	Obs with Dep=1	16

Source : établi sous eviews 7.

Tableau 02

Table de prédictions et des résultats attendus

	Equation estimée			Probabilité constante		
	Dep =0	Dep =1	Total	Dep=0	Dep=1	Total
P(Dep=1)≤C	24	2	26	25	16	41
P (Dep=1)>C	1	14	15	0	0	0
Total	25	16	41	25	16	41
Correct	24	14	38	25	0	25
%Correct	96.0	87.50	92.6	100.0	0.00	60.9
% Incorrect	4.00	12.50	7.32	0.00	100.0	39.0
Gain total*	-4.00	87.50	31.71			
Pourcentage du gain**	NA	87.50	81.25			

Source : établi sous eviews 7.

- ***Solution de l'exercice 01:***

Compte tenu de la nature discrète du choix de régime de change, nous supposons que l'Algérie choisira un régime de change fixe si la variable latente Y_i^*est inférieure à un certain seuil $Y_i^* \leq c_1$. De même que le régime de change flottant administré sera choisi si seulement si$Y_t^* > c_2/c_2 > c_1$.

$$Y_i = \begin{cases} 1 \; si \; Y_i^* > c_2 \\ 0 \; sinon \end{cases} i = 1,2, \cdots \cdots N.$$

1- Interprétation statistique

Nous avons fait recours à la technique de maximum de vraisemblance qui s'est étendue sur un échantillon de 41 observations dont 25 ayant une valeur 0 et 16 ayant une valeur 1.

Les valeurs des coefficients de PIBHAB, TCM2R, TXINF et PERFINPUB sont négatives. Cela s'explique par le fait que ces variables diminuent la probabilité du régime de change flottant géré. Alors que la constante (C), DEGOUV et TCPIBR ont des coefficients positifs. Ainsi, la hausse de ces coefficients accroit la probabilité de la réponse.

Cependant la constante est insignifiante suite à sa probabilité qui est de **0.13> 0.05.**

Le log-vraisemblance (LL) est de **-7.9078**, ce qui fait que la moins double log-vraisemblance est de **15.8157** (-2LL). La statistique L_R est calculée à l'aide de $-2\big(Ln(L_R)\big) - Ln(L_u)\big)$ et est un analogue de F-statistique dans les modèles de régression linéaire et teste la significativité globale du modèle. La statistique du log vraisemblance

est égale à $\boldsymbol{LR = 39.0305}$que nous comparons à un Khi-deux lu dans la table à un seuil de 0.95 et à 6 degré de liberté, qui est égale à$1.63 < 39.0305$. Ainsi l'hypothèse $H_0: \beta_1 = \beta_2 = \cdots \cdots = \beta_k = 0$ est rejetée et le modèle comporte au moins une variable explicative significative. Le pseudo R^2 de McFadden est un analogue de R^2 et est donné par :

$$R^2 = 1 - \frac{\log(L_U)}{\log(L_R)} = 1 - \frac{-7.9078}{-27.4231} = 0.71163$$

$$R^2 = 0.71$$

Ainsi le modèle est validé statistiquement.

Dans l'objectif d'appréhender les qualités prévisionnelles du modèle sur l'échantillon, nous avons fait recours à la table de prédiction, en comparant la probabilité estimée pour un individu i d'être $Y_i = 1(p(Dep = 1))$ au seuil arbitraire de 50% à la valeur observée des $Y_i = 0\ ou\ 1$.

Pour cette application, les 25 des observations pour les quelles$Y_i = 0$, le modèle indique que 24 observations ont une probabilité estimée que le régime de change soit flexible géré inférieur à 50%. Dans 96% des cas le régime de change fixe est correctement prévu.

Pour 16 des observations pour lesquelles$Y_i = 1$, le modèle indique 14 observations qui ont une probabilité estimée supérieure à 50%. Dans 87.50% le régime de change flexible administré est correctement prévu. Ainsi, globalement, le modèle estimé prédit 60.98% des observations. Le taux d'erreur est donc faible.

2- Interprétation économique des résultats

Les coefficients positifs (négatifs) signifient que l'augmentation d'une variable a pour conséquence une hausse (baisse) de l'utilité du régime de change flexible géré. Ainsi, nous noterons l'augmentation de sa probabilité d'être adopter. Les variables estimées jouent un rôle important pour la détermination d'un régime de change car la majorité des coefficients sont significatifs.

Contrairement à ce que prédit la théorie des zones monétaires optimales, en matière de choix de régime de change, nos résultats indiquent que l'augmentation du degré d'ouverture de l'économie algérienne (**DEGOUV**) conduit à un choix d'un régime de change flexible géré. En effet, Barry Eichengreen et Masson (1998) considèrent que la grande ouverture de l'économie est associée à des taux de change variables afin d'amortir

les chocs en provenance de l'extérieur. Michael Mussa et al (2000) affirment, lui aussi, que les taux de change variables sont plus favorables aux économies plus ouvertes car ils constituent un ajustement meilleur des chocs externes. A contrario, Von Hagen et al (2002) affirment la théorie des zones monétaires optimales en termes d'ouverture économique. Autrement dit, l'augmentation de l'ouverture économique est associée à des régimes de change fixes. De même, Eduardo Levy-Yeyati.E. et al (2010) soutiennent empiriquement la théorie des zones monétaires optimales (y compris le critère d'ouverture de l'économie). Quant à la variable **PIBHAB**, elle est moins importante que les autres variables puisque son coefficient est insignifiant. Par contre le coefficient de la variable **TCM2R** est significatif et de signe négatif. Ainsi, toute augmentation de la masse monétaire réelle engendre une utilité d'adoption d'un régime de change fixe.

En effet, toute augmentation de la masse monétaire sans contrepartie de production engendre une baisse de la valeur de la monnaie nationale ce qui implique une augmentation des prix à l'importation et ceci entrainera une inflation. Donc l'adoption d'un régime fixe éloignera ce scénario. Le coefficient de la variable **TCPIBR** est significativement positif. L'augmentation du **PIB** réel engendre, ainsi, une préférence pour le choix d'un régime de change flottant administré. Cela s'explique par la réaction instantanée des régimes flexibles quant aux chocs réels alors qu'en régime fixe la réaction des prix est lente.

La variable **TXINF** est significative et de signe négatif. L'utilité du régime de change fixe est donc plus grande. Cela est expliqué entre autre par la hausse des prix à l'exportation suite à une hausse des taux d'inflation interne. Cet état de fait explique l'intérêt à adopter un régime de change fixe afin d'éliminer les effets négatifs des taux d'inflation sur les prix à l'exportation et à l'importation.

La variable **PERFINPUB** est significative et de signe négatif. De ce fait, l'utilité du régime de change fixe est plus grande. A long terme, une hausse permanente de l'offre de monnaie et/ou une baisse des dépenses publiques entrainent une dépréciation du taux de change. Dans le modèle de Mundell- Fleming, une hausse des dépenses publiques entraine une appréciation du taux de change.

2.4.2. Exercice 02 : *Estimation des modèles Probit et Logit binaires explicatifs des facteurs de la réussite en Licence (Extrait du livre : Bourbonnais, R. (2018)).*[8]

Nous avons relevé sur un échantillon de 60 étudiants inscrits en dernière année de Licence d'Économie, les variables suivantes susceptibles d'expliquer la réussite ou l'échec à l'examen de Licence (variable **REUSSITE** = **0** si échec, **1** sinon) :

NENFANTS = variable discrète représentant le nombre de frères et sœurs de l'étudiant,
NECONO = la note d'économétrie sur 20 obtenue en Licence,
NMICRO = la note de micro-économie sur 20 obtenue en Licence,
GENRE = variable muette, (1 = masculin, 0 = féminin).
Un extrait des données est présenté dans le tableau 1.

Tableau 03 – Extrait de données

OBS	REUSSITE	NENFANTS	NECONO	NMICRO	GENRE
1	0	2	3,6	0	1
2	0	5	3,8	0	1
…	…	…	…	…	…
59	1	0	16,2	12	0
60	1	2	17	4	0

On demande :
1) d'estimer un modèle de type Logit permettant de prévoir la probabilité de réussite d'un étudiant en Licence,
2) de comparer les résultats avec un modèle de type Probit.
3) de donner la probabilité de réussite, à l'aide du modèle Logit estimé, pour un étudiant dont les caractéristiques sont les suivantes : NENFANTS = 1 ; NECONO = 12 ; NMICRO = 13,5 ; GENRE = masculin.

[8] Bourbonnais, R. (2018), « Econométrie », édition DUNOD, p352.

Tableau 04 : Une estimation d'un modèle Logit

Dependent Variable : REUSSITE
Method: ML – Binary Logit
Included observations: 60

Variable	Coefficient	Std. Error	z-Statistic	Prob.
NENFANTS	– 0.682523	0.378870	– 1.801470	0.0716
NECONO	0.632062	0.239564	2.638382	0.0083
GENRE	– 3.761718	1.437068	– 2.617633	0.0089
NMICRO	0.155322	0.188916	0.822173	0.4110
C	– 3.265634	2.020060	–1.616602	0.1060

Tableau 05 : ***Une estimation d'un nouvel modèle Logit***

Dependent Variable : REUSSITE
Method: ML – Binary Logit
Included observations: 60

Variable	Coefficient	Std. Error	z-Statistic	Prob.
NENFANTS	– 0.746742	0.378942	– 1.970596	0.0488
NECONO	0.695857	0.231789	3.002112	0.0027
GENRE	– 3.634605	1.410945	– 2.576008	0.0100
C	– 2.859277	1.910377	– 1.496708	0.1345

Mean dependent var	0.516667	S.D. dependent var	0.503939
S.E. of regression	0.287086	Akaike info criterion	0.645890
Sum squared resid	4.615432	Schwarz criterion	0.785512
Log likelihood	– 15.37669	Hannan-Quinn criter.	0.700504
Restr. log likelihood	– 41.55549	Avg. log likelihood	– 0.25627
LR statistic (3 df)	52.35761	McFadden R-squared	0.629972
Probability(LR stat)	2.51E – 11		
Obs with Dep=0	29	Total obs	60
Obs with Dep=1	31		

Avec:

$$\boldsymbol{L_u = Log\ likelihood};\ \boldsymbol{L_R = Restr.\log likelihood};\ \boldsymbol{LR = LR\ statistic};\ \boldsymbol{L_u/n = Avg.\log likelihood}$$

Le critère d'information de Hannan-Quinn permet des comparaisons entre les modèles (comme les critères d'Akaike ou Schwarz) en termes d'arbitrage: apport d'information lié

à l'ajout de variables explicatives et perte de degrés de liberté. En cas de modèle concurrent, celui ayant le plus faible critère d'information sera retenu.

Tableau 06 : Table de prédiction des résultats

Dependent Variable : REUSSITE
Method: ML – Binary Logit
Included observations: 60
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	26	4	30	0	0	0
P(Dep =1) > C	3	27	30	29	31	60
Total	29	31	60	29	31	60
Correct	26	27	53	0	31	31
% Correct	89.66	87.10	88.33	0.00	100.00	51.67
% Incorrect	10.34	12.90	11.67	100.00	0.00	48.33
Total Gain*	89.66	– 12.90	36.67			
Percent Gain**	89.66	NA	75.86			

Tableau 07 : Estimation d'un modèle Probit :

Included observations: 60

Variable	Coefficient	Std. Error	z-Statistic	Prob.
NENFANTS	– 0.428197	0.219223	– 1.953247	0.0508
NECONO	0.363148	0.110230	3.294454	0.0010
GENRE	– 1.824203	0.650426	– 2.804629	0.0050
C	– 1.491466	1.108767	– 1.345157	0.1786

- ***Solution de l'exercice 02* :**

1- Interprétation des résultats :

À la lecture des résultats du ***tableau 04***, nous constatons que :

– la variable NMICRO à une probabilité critique de 0,41, elle n'est donc pas significative,

– la variable NENFANTS à une probabilité critique de 0,07, elle est donc faiblement significative.

Nous procédons à une nouvelle estimation en retirant la variable NMICRO dont le coefficient n'est pas significativement différent de 0. A cet effet, Les résultats complets fournis par Eviews sont reportés dans le ***tableau 05***. Les coefficients sont tous significativement différents de 0, hormis le terme constant. La statistique de la Log vraisemblance est égale à *LR* = 52,35 que l'on compare à un χ^2 lu dans la table à un seuil de 0,95 % et à 3 degrés de liberté,

$$\chi_3^{2\,0,95} = 9{,}28 < 52{,}35 \rightarrow rejet\ de\ H_0$$

Le pseudo-R^2 est donné par :

$$1 - \frac{Log(L_u)}{Log(L_R)} = 1 - \frac{-15{,}38}{-41{,}56} = 1 - 0{,}37 = 0{,}63$$

Le modèle est validé sur le plan statistique.

Le logiciel Eviews propose une table permettant d'appréhender les qualités prévisionnelles du modèle sur l'échantillon en comparant la probabilité estimée pour un individu *i* d'être y_i = 1*(P(Dep* = 1*))* au seuil arbitraire de 50 % à la valeur observée des y_i= 0 ou 1.

D'après le ***tableau 06***, les individus (29) pour lesquels yi = 0, le modèle indique que 26 individus ont une probabilité estimée de réussite inférieure à 50 %. Dans 89,66 % des cas, les échecs sont donc correctement prévus. Pour les individus (31) pour lesquels yi = 1, le modèle indique que 27 individus ont une probabilité estimée supérieure à 50 %. Dans 87,10 % des cas, les réussites sont correctement prévues. Le taux d'erreur est donc faible.

- Le modèle s'écrit :

$$Ln\left(\frac{P_i}{1-P_i}\right) = -0.75NENFANTS_{(1,97)} + 0{,}70NECONO_{(3,00)} - 3{,}63GENRE_{(2,57)} - 2{,}86 + e_i$$

(.) = $\mathcal{Z}$-Statistique

e_i= Résidu d'estimation

– Le nombre de frères et sœurs du foyer agit négativement, les étudiants issus de familles nombreuses ont un taux de réussite plus faible.

– La note d'économétrie est un facteur positif de réussite.

– Enfin, les étudiants de genre masculin réussissent en général moins bien (signe négatif) que les étudiants de genre féminin.

2- D'après ***le tableau 07***, les valeurs des coefficients sont de même signe mais différentes par rapport au modèle Logit car la spécification n'est pas la même. Cependant,

nous pouvons retrouver, approximativement, les valeurs estimées du modèle Logit en multipliant chacun des coefficients des variables explicatives par la constante $\frac{\pi}{\sqrt{3}} \approx 1.81$.

3- Soit les caractéristiques de l'étudiant : NENFANTS = 1 ; NECONO = 12 ; NMICRO = 13,5 ; GENRE = masculin. Le modèle Logit estimé (la note de micro-économie ne figurant pas dans le modèle final, elle n'est pas intégrée dans le calcul, *cf.* question 1) est le suivant :

$$Ln\left(\frac{P_i}{1-P_i}\right) = -0.75NENFANTS_{(1,97)} + 0,70NECONO_{(3,00)} - 3,63GENRE_{(2,57)} - 2,86 + e_i$$

$$Ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = -0,75 \times 1 + 0,70 \times 12 - 3,63 \times 1 - 2,86 - 2,86 = 1,109$$

$$Ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = e^{1,109} = 3,033 \rightarrow \hat{P}_i = \frac{3,033}{1+3,033} = 0,75$$

La probabilité de réussite de cet étudiant de licence est donc de 75 %.

Chapitre 03 : Les modèles à choix multiple

Les modèles à variables discrètes multiples, appelés aussi les modèles polytomiques, sont une généralisation des modèles à variable dépendante dichotomique.

Les variables polytomiques sont très répandues dans les enquêtes afin de faciliter le traitement des informations à travers des codes. De ce fait, en pratique, il existe trois types de modèles multinomiaux ; à savoir : les modèles ordonnés, séquentiels et non ordonnés.

De manière générale, le modèle multinomial est définit comme suit :

Pour un individu $i/i = 1 \cdots \cdots \cdots N$,La variable qualitative y_i peut prendre $m+1$ modalités : 0, 1, 2, ……m, ainsi :

$$\forall i: \sum_{i=0}^{m+1} Prob(y_i = j) = 1$$

La probabilité associée à chaque modalité est :

$$Prob(Y_i = j) = \mathrm{F}_{ij}(x\beta)$$

$$i = 1 \cdots \cdots \cdots N, j = 0, 1, 2, \cdots \cdots \cdots m$$

La vraisemblance du modèle peut s'écrire comme le produit des probabilités associées aux différentes modalités :

$$\mathcal{L} = \prod_{i=1}^{N} \prod_{j=0}^{m_i} \Phi(x\beta)^{Y_{ij}}$$

3.1. Les modèles ordonnés :

Dans ce type de modèles, les valeurs prises par la variable dépendante y_{ij} sont hiérarchisées ; elles correspondent à une classe ou à une catégorie. Nous distinguons, ainsi, deux types de modèle à choix multiples ordonnés :

- « discrétisation » d'une variable continue sous forme de tranche ; exemple : tranche de revenu ;

- L'appartenance à une catégorie ; exemple : mention baccalauréat : assez bien, bien, très bien et excellent.

L'introduction d'une variable latente dans ce type de modèle, comme dans les modèles binaires, est indispensable pour leur estimation. De ce fait :

$$Y_i^* = \beta_k x_i + \varepsilon_i$$

Les variables prises par la variable discrète $Y_i$correspondent à des intervalles dans lesquels se trouve la seule variable latente (inobservable) continue Y_i^* :

$$\begin{cases} Y_i = 0 & si\ Y_i^* \leq c_1 \\ Y_i = 1 & si\ \ c_1 < Y_i^* \leq c_2 \\ Y_i = 2 & si\ c_2 < Y_i^* \leq c_3 \\ & \vdots \\ & \vdots \\ & \vdots \\ Y_i = m & si \quad c_m < Y_i^* \end{cases}$$

Où : $c_{1,} \cdots\cdots\cdots c_m$sont des bornes à estimer, définissant les extrémités des intervalles. La variable latente Y_i^*est, donc, une combinaison linéaire des variables explicatives x_i :

$$Y_i^* = x_i\beta + \varepsilon_i$$

Où ε_i admet une fonction de répartition F où les probabilités associées aux réalisations sont :

$P_{i0} = Prob(Y_i = 0) = F\big(c_1 - (\beta_0 + \beta_1 x_i)\big)$

$P_{i1} = Prob(Y_i = 1) = F\big(c_2 - (\beta_0 + \beta_1 x_i)) - \Phi(c_1 - (\beta_0 + \beta_1 x_i)\big)$

.

.

.

$P_{im} = Prob(Y_i = m) = 1 - \mathrm{F}(c_m - (\beta_0 + \beta_1 x_i))$

Si la probabilité *P* est associée à une densité symétrique et une fonction de répartition *F*, nous aurons donc :

$$Prob\left(Y = \frac{j}{x}, \beta, c\right) = F\big(c_{j+1} - x\beta\big) - F\big(c_j - x\beta\big)$$

Où :$j = 0,1,2, \cdots\cdots\cdots mc_0 = -\infty$, $\quad c_{m+1} = +\infty$, $\quad c_j \leq c_{j+1}$.

Les bornes c_0 et c_{m+1} garantissent la condition de la somme des probabilités, sur tous les intervalles, qui est égale à 1. Nous aurons la fonction de répartition F de la loi de probabilité normale ou logistique.

$$\mathrm{F(t)} = \frac{e^t}{1+e^t} \;\; et \sum_{i=0}^{m} P_i = 1$$

.

3.1.1. Exemples :

Nous prenons ici l'exemple des individus qui veulent acquérir un bien immobilier, les biens sont regroupés en trois catégories en fonction de leur prix. Par contre, ces derniers sont inobservables, et seule son appartenance à l'une des trois catégories est observée :

$$Y_i = \begin{cases} 0 & \text{si le prix du bien i acquis est inférieur à 48,999€} \\ 1 & \text{si le prix du bien i acquis est compris entre 49,000€ et 74,999€} \\ 2 & \text{si le prix du bien i acquis est supérieur à 75,000€} \end{cases}$$

Nous modélisant la variable polytomique $Y_i = 0, 1, 2$ selon l'appartenance de la variable latente y_i^*à trois catégories différentes :

$$Y_i = \begin{cases} 0 & \text{si} \quad Y_i^* < c_1 \\ 1 & \text{si } c_1 \leq Y_i^* < c_2 \\ 2 & \text{si } Y_i^* > c_2 \end{cases} \qquad \forall i = 1 \ldots \ldots . N.$$

La variable latente$Y_i^* \sim N(x_i\beta, \sigma^2)$, où le vecteur $x_i$inclut l'ensemble des particularités du bien citées précédemment. Le vecteur x_i ne comporte pas de constante pour une raison de colinéarité. Il faudrait, donc, passer à l'estimation des paramètres structurels du modèle, à savoir :c_1,c_2,σ et les $k$coefficients du vecteurβ. Subséquemment, nous estimons $k + 3$ paramètres structurels.

Les fonctions de répartition normale et logistiques définissent respectivement les modèles Logit et Probit. Et comme les modèles dichotomiques, l'estimation des coefficients β_k et les valeurs des seuils $C_j$des modèles multinomiaux ordonnés se fait à travers des algorithmes de maximisation d'une fonction log-vraisemblance définie parP_{ij}.

Le signe des paramètres estimés nous renseigne sur l'impact positif ou négatif que pourrait exercer les variables explicatives sur la variableY_{ij}. La significativité des paramètres estimés est apprécié par le test $\mathcal{Z}$-statistique et la significativité globale du modèle par l'intermédiaire de la statistique $LR = -2(L_n(L_R) - L_n(L_u))$. Le Pseudo-$R^2$évalue quant à lui l'ajustement global du modèle :

$$R^2 = 1 - \frac{\log(L_u)}{\log(L_R)}$$

3.1.2. Application :(*Extrait de : Hurlin, C, (2003))*[9]

Nous considérons une application qui porte sur l'évaluation des performances des institutions de dépôt (banques et autres institutions financières) aux Etats Unis. Cette appreciation est faite par plusieurs instances de contrôle, à savoir: *Office of the Comptroller of the Currency* (OCC), *Board of Governors of the Federal Reserve System* (FRB), *Office of Thrift Supervision* (OTS), and *Federal Deposit Insurance Corporation* (FDIC).

A cet effet, une notation (ou rating) est alors accordé selon quatre modalité:

$$Y_i = \begin{cases} 1 \ si\ la\ performance\ est\ remarquable \\ 2\ si\ la\ performance\ est\ satisfaisante \\ 3 \quad si\ la\ performance\ est\ \text{à}\ améliorer \\ 4 \quad si\ la\ performance\ est\ déplorable \end{cases}$$

Cette analyse est tenue sur 350 observations en fonction de plusieurs variables explicatives désignées respectivement :*loa, prl, equ, roa, sec, ass, metro* et *growth*.

La variable *loa* désigne le ratio prêt sur actif total de la banque ;

La variable *prl* désigne le ratio actif douteux sur actif total ;

La variable *equ* désigne le ratio capital propre sur actif ;

La variable *roa* désigne le ratio dividende sur actifs ;

La variable *sec* désigne le ratio investissements de valeurs sur actifs ;

La variable *ass* le logarithme de l'actif de la banque ;

La variable metro prend une valeur 1 si la banque à son siège dans une MSA et 0 sinon ;

La variable growth désigne le taux de croissance du PIB de l'Etat dans lequel la banque opère.

Dans le tableau ci-dessous sont reproduit les valeurs ces différentes variables pour les 10 premiers individus de l'échantillon.

[9] Hurlin, C, (2003), « Econométrie des variables qualitatives : chapitre 02 : Modèles multinomiaux », polycopié de cours, université d'Orléans, P10.

Tableau 08 : extrait des données

CRA	EQU	GROWTH	LOA	METRO	PRL	ROA	SEC
CRA	EQU	CROWTH	LOA	METRO	PRL	ROA	SEC
1.000000	0.068826	0.055908	0.398216	0.000000	0.012070	0.003437	0.395789
1.000000	0.109697	0.037609	0.622300	0.000000	0.005999	0.008949	0.329522
1.000000	0.062175	0.059176	0.520229	1.000000	0.002058	0.013376	0.316544
1.000000	0.071209	0.061640	0.559965	0.000000	0.009662	0.013883	0.323823
1.000000	0.111563	0.037609	0.653102	0.000000	0.003479	0.011397	0.269915
1.000000	0.107558	0.077307	0.654570	1.000000	0.000995	0.013553	0.166809
1.000000	0.108573	0.031546	0.581331	0.000000	0.009516	0.014816	0.361508
1.000000	0.091364	0.060035	0.544357	0.000000	0.000905	0.014636	0.171662
1.000000	0.075470	0.101828	0.633069	1.000000	0.001361	0.010469	0.247411
1.000000	0.121436	0.018210	0.250721	0.000000	0.039951	-0.015693	0.495006

Cette problématique sera traitée à travers un modèle Probit multinomial ordonné. En effet, dans ce cas précis, les valeurs prises par la variable multinomiale peuvent correspondre à des intervalles dans lesquels va se trouver une variable latente inobservable. Nous avons un ordre naturel sur les modalités allant de la satisfaction la plus complète à la performance déplorable. Pour modéliser ce Probit ordonné sous Eviews, nous choisissons *Estimate Equation* dans le menu *Quick*, et nous retenons la méthode *ORDERED – Ordered Choice* avec une *Error Distribution* de type *Normal*. Nous indiquons par la suite la variable polytomique ainsi que les variables explicatives. Les coefficients des variables *roa* et *sec* sont alors non significativement différents de zéro. Nous les retirons donc et les résultats obtenus pour le Probit ordonnés sont alors les suivants :

Tableau09 : Estimation d'un probit ordonné

Dependent Variable: CRA
Method: ML - Ordered Probit
Date: 10/15/02 Time: 11:07
Sample: 1 350
Included observations: 350
Number of ordered indicator values: 4
Convergence achieved after 5 iterations
Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
ASS	-0.250326	0.053268	-4.699375	0.0000
EOU	5.249539	1.720978	3.050324	0.0023
GROWTH	-6.107240	2.790589	-2.188513	0.0286
LOA	-1.724511	0.338800	-5.090059	0.0000
METRO	0.768531	0.133197	5.769886	0.0000
PRL	10.74869	2.697211	3.985113	0.0001
Limit Points				
LIMIT_2:C(7)	-3.645509	0.609564	-5.206606	0.0000
LIMIT_3:C(8)	-2.725115	0.678875	-4.014162	0.0001
LIMIT_4:C(9)	-1.614912	0.677212	-2.384648	0.0171

Akaike info criterion	2.397249	Schwarz criterion	2.496453
Log likelihood	-410.5185	Hannan-Quinn criter.	2.436735
Restr. log likelihood	-473.1244	Avg. log likelihood	-1.172910
LR statistic (6 df)	125.2117	LR index (Pseudo-R2)	0.132324
Probability(LR stat)	0.000000		

- **Interprétation des résultats :**

Les coefficients ainsi que les trois seuils sont tous significativement différents de 0 (leurs probabilités critiques sont inférieures à 0,05).

La statistique de la Log vraisemblance est égale à *LR* = 125,21 que l'on compare à un χ^2 lu dans la table à un seuil de 0,95 % et à 3 degrés de liberté, $\chi_6^{2\ 0.95}$<125,21 → rejet de H0.

Le modèle est donc validé sur le plan statistique.

Les seuils c_1, c_2 et c_3 sont respectivement de :(-3,64), (-2.72) et (-1.61).

D'après les résultats du **tableau 08**, nous constatons que les variables ASS, GROWTH et LOA agissent négativement sur la performance des banques. En revanche, les variables explicatives ; EOU, METRO et PRL agissent positivement sur la performance des institutions financières.

3.2. Les modèles non ordonnés :

Le logit multinomial est le plus pratique vu qu'il est une extension directe du Logit binaire, de par leur facilité d'utilisation. A cet effet, nous nous contenterons, dans cette partie, d'étudier le Logit multinomial non ordonné. D'après *Daniel McFadden (1974)* ces derniers sont modélisés en associant à chaque modalités un niveau d'utilité (gain) tel que :

$$U_{ij} = \beta_0 + \beta_1 x_i + \frac{\varepsilon_i}{j} = 1 \dots \dots m$$

L'équation représente donc le modèle latent qui est relié à la variable observée en sachant que la modalité choisie par l'individu est celle qui lui procure le gain maximal.

$$Y_i = j \; si \; U_{ij} = Max(U_{i0}, U_{i1}, \dots \dots .., U_{im})$$

La généralisation du Logit binaire au cas multinomial est faisable à partir de l'équation de Prob$(Y_i = j)$. Soit m+1 modalités et $(P_0, P_1, P_2 \dots \dots \dots . P_m)$les probabilités associées. La probabilité 0 est prise comme référence, nous posons donc :

$$\frac{P_1}{P_1 + P_0} = H(x_i\beta_1), \frac{P_2}{P_2 + P_0} = H(x_i\beta_2) \dots \dots \dots \frac{P_m}{P_m + P_0} = \; H(x_i\beta_m)$$

H(.) est une fonction continue et croissante :

$$\frac{P_j}{P_j + P_0} = \; H^{-1}(x_i\beta_j)$$

Le rapport entre deux probabilités P_j et P_k $(\, j, k = 1,2, \dots \dots . m)$ vaut :

$$exp[x_i(\beta_j - \beta_k)]$$

Et le vecteur des paramètres β_0 est normalisé à 0.

À la suite des probabilités définies comme des rapports d'exponentiels, diverses propriétés s'affichent. Premièrement, le rapport entre deux probabilités est invariant à une transposition dans les vecteurs des paramètres :

$$Prob(Y_i = j)/Prob(Y_i = k) = exp(x_i\beta_j)/\exp(x_i\beta_k)$$
$$= \exp[x_i(\beta_j - \beta_0)]/exp[x_i(\beta_k - \beta_0)]$$

Deuxièmement, l'interprétation des paramètres estimés est considérée par rapport aux écarts au référentiel (modalité 0).

L'estimation du modèle Logit multinomial non ordonné se fait en maximisant la log-vraisemblance par rapport aux vecteurs de paramètres $(\beta_1, \beta_2, \dots\dots\dots \beta_m)$:

$$Log\mathcal{L} = \sum_{i=1}^{N}\sum_{k=1}^{m} Y_{ik}x_i\beta_k - (m+1)\sum_{i=1}^{N} Log(1 + \sum_{k=1}^{m}\exp(x_i\beta_k))$$

Les valeurs des coefficients ne sont pas directement interprétables en termes d'élasticité β_k , mais plutôt les signes qui importent le plus en indiquant si la variable agit positivement ou négativement sur la probabilité relative de choisir ***j*** plutôt que 0. Quant à l'interprétation statistique des résultats, elle se fait de la même manière que le modèle précédent (**3.1**.).

3.3. Les modèles séquentiels :

Ce type de modèles sont utilisés pour rendre compte des choix effectués ou des évènements, selon une séquence bien précise. Cette dernière est généralement conditionnée par le temps et les faits successifs conditionnent l'ensemble des modalités futures. A chaque étape, ces modèles détiennent l'avantage de construire la séquence des évènements comme étant le produit des probabilités primordiales à la réalisation d'un seul choix.

3.3.1. Exemple de modèle séquentiel :

Considérons l'exemple le plus répandu dans ce type de modèle ; la réussite au master. Nous cherchons à estimer la probabilité qu'un étudiant obtienne son Master. On note $y_i = 1$ si l'étudiant « i » a obtenu le baccalauréat mais pas la licence, $y_i = 2$ si l'étudiant a obtenu la licence mais pas le master et $y_i = 3$si l'étudiant a obtenu le master.

Chapitre 04 : Les modèles à variables dépendantes limitées

Nous analysons dans cette partie la modélisation d'une variable dépendante continue qui est observable sur un certain intervalle. Nous tenterons, donc, d'expliquer la valeur de la variable continue à travers la modélisation en regression linéaire. Quoique, cette variable n'est observable que si elle appartient à un certain intervalle. Nous parlons donc la notion de probabilité et son estimation est semblable à celle des modèles à variables dépendantes discrètes. Exemple d'une étude sur le salaire des individus. Leur revenu n'est connu que si l'individu possède un emploi.

De ce fait les modèles *Tobit* sont adaptés à cette catégorie d'analyse. Nous distinguons, pratiquement, deux extensions du modèles *Tobit simple* : les modèles *Tronqués* et les modèles *Censurés*. A cet effet, dans cette partie, nous présenterons les modèles *Tronqués*, *Censurés*, le *Tobit simple* et le *Tobit généralisé*.

4.1. Le modèle de regression censuré et tronqué :

Soit la variable latente Y_i^*qui est non observable et la variable dépendante observée Y_i :

$$Y_i^* = x_i\beta_i + \varepsilon_i$$

Où :

$x_i est\ la\ matrice\ des\ valeurs\ de\ variables\ explicatives$;

$\beta : les\ paramètres\ du\ modèle$;

ε_i: $le\ terme\ d'erreur\ avec\ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

Ce qui fait :

$$Y_i^* \sim N\ (x_i\beta,\qquad \sigma_\varepsilon^2)$$

D'où :

$$\begin{cases} Y_i = Y_i^*\ si\ Y_i^* > 0 \\ Y_i\ = 0\ si\ Y_i^* \leq 0 \end{cases}$$

D'une manière générale :

$$\begin{cases} Y_i = Y_i^*\ si\ Y_i^* > c_i \\ Y_i\ = c_i\ si\ Y_i^* \leq c_i \end{cases}$$

L'exemple donné préalablement s'applique si Y_i^*est le salaire des interrogés et Y_i est la réponse des individus interrogés ; s'ils possèdent un emploi ou pas. $c_i = 0 \forall i$ etx_i est la matrice des caractéristiques des individus interrogés. Si l'individu ne possède pas de travail $Y_i = 0$ et $Y_i = Y_i^*$dans le cas contraire. Par conséquent, la décision de l'individu de répondre 0 dépend de Y_i^*qui est certainement≤ 0. Dans ce cas l'échantillon est dit *Censuré* vu que la variable dépendante n'est observée que sur l'intervalle $]0, +\infty[$.

Par ailleurs, si l'enquêteur décide d'omettre les observations telles que $Y_i^* \leq 0$ de l'échantillon, dans ce cas l'échantillon est dit *Tronqué*.

Dans ce cours, nous nous intéressons qu'aux modèles *Censurés.* L'estimation de ces derniers se fait à travers l'adoption de la distribution de la loi normale pour la variable latente Y_i^*de moyenne $x_i\beta$ et de variance σ_ε^2. Dans un échantillon censuré, la probabilité d'avoir $Y_i = c_i$ est de :

$$Prob(Y_i = c_i) = Prob(Y_i^* \leq c_i) = Prob\left[\left(\frac{Y_i^* - x_i\beta}{\sigma}\right) \leq \left(\frac{c_i - x_i\beta}{\sigma}\right)\right] = F(\frac{c_i - x_i\beta}{\sigma})$$

Comme Y_i^*est non observable, nous modélisons la probabilité de l'évènement qu'elle appartienne à un intervalle donné à travers la fonction de répartition de la loi normale *F(.).* D'autre part, si $Y_i > c_i$ la variable latente Y_i^*admet, donc, une distribution continue. Par conséquent, la probabilité associée à $Y_i^*/Y_i > c_i$ est modélisée par l'intermédiaire de la fonction de densité de la loi normale *f(.)* :

$$Prob(Y_i > c_i) = Prob(Y_i = Y_i^*) = \frac{1}{\sigma\sqrt{2\pi\pi}} \exp\left[-\frac{(Y_i - x_i\beta)^2}{2\sigma^2}\right]$$

$$Prob(Y_i = Y_i^*) = \frac{1}{\sigma} f(\frac{Y_i - x_i\beta}{\sigma})$$

Par conséquent, l'espérance d'une variable censurée :$Y_i^* \sim N(\mu, \sigma^2)$et $Y_i = c_i$ si $(Y_i^* \leq c_i), (y_i = Y_i^*)$:

$$E(Y) = F(\frac{c - \mu}{\sigma}) \times c[1 - F(\frac{c - \mu}{\sigma})] \times [\mu + \sigma \frac{f[(c - \mu)/\sigma]}{1 - \mathrm{F}[c - \mu)/\sigma]}$$

Sachant que :

$$\mu = x_i\beta$$

$$E(Y/Y > c) = E(Y^*/Y^* > c) = \mu + \sigma\frac{f[(c-\mu)/\sigma]}{1 - \mathrm{F}[c-\mu)/\sigma]}$$

En revanche, quand l'échantillon est tronqué, nous disposons que des observations où$Y_i^* > c_i$. L'expression de la densité conditionnelle est :

$$Prob\ (Y_i^* > c_i) = \frac{1}{\sigma} f(\frac{Y_i^* - x_i\beta}{\sigma})/[1 - F(\frac{c_i - x_i\beta}{\sigma})]$$

$(1 - F)$ est la constante de normalisation assurant que :

$$\int_{c_i}^{+\infty} Prob(Y_i^* > c_i) = 1$$

L'espérance conditionnelle d'une variable normale tronquée $Y \sim N(\mu, \sigma^2)$:

- Y est observée si $Y > c$:

$$E(Y/Y > c) = \mu + \sigma\frac{f[(c-\mu)/\sigma]}{1 - \mathrm{F}[c-\mu)/\sigma]}$$

- Y est observée si $Y \leq c$:

$$E\left(\frac{Y}{Y} \leq c\right) = \mu - \sigma\frac{f[(c-\mu)/\sigma]}{\mathrm{F}[c-\mu)/\sigma]}$$

Le terme $f/_F$ est appelé « le rapport inverse de Mills » (Inverse Mills Ratio).

4.2. Modèle Tobit simple

Le modèle Tobit simple est similaire au modèle à variable dépendante qualitative binaire en termes d'hypothèse pour la variable latenteY_i^*. Par contre, le terme d'erreur $\varepsilon_i$suit une distribution normale dans les modèles Tobit simple ; $\varepsilon_i \sim N(0, \sigma^2)$. Le modèle s'écrit donc :

$$\begin{cases} Y_i^* = & x_i\beta + \varepsilon_i \\ Y_i = & Y_i^*\ si\ Y_i^* > 0 \\ Y_i = & 0\ si\ Y_i^* \leq 0 \end{cases}$$

Pour les observations où $Y_i > 0$:

$$P(Y_i > 0) \times P(Y_i/Y_i > 0) = F(\frac{x_i\beta}{\sigma})\frac{P(\varepsilon_i)}{F(x_i\beta/\sigma)} = P(\varepsilon_i) = \frac{1}{\sigma} f(\frac{Y_i - x_i\beta}{\sigma})$$

Pour les observations $Y_i = 0$:

$$P(Y_i = 0) = P(\varepsilon_i < -x_i\beta) = 1 - F(\frac{x_i\beta}{\sigma})$$

De ce fait, la fonction de log-vraisemblance du modèle Tobit, s'écrit en combinant les deux probabilités :

$$Log\mathcal{L} = \sum_0 log[1 - F(\frac{x_i\beta}{\sigma})] + \sum_1 \log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \sum_1 \frac{(Y_i - x_i\beta)^2}{2\sigma^2}$$

Où :

$\sum_0 sommation\ des\ observations\ des\ Y_i = 0$

$\sum_1 sommation\ des\ observations\ des\ Y_i > 0$

L'espérance conditionnelle de la variable Y_i est de :

$$E(Y_i/x_i, Y_i > 0) = \frac{E(Y_i/x_i)}{Prob(Y_i > 0)} = x_i\beta + \sigma\frac{f(-x_i\beta/\sigma)}{1 - F(-x_i\beta/\sigma)}$$

4.2.1. Estimation par le maximum de vraisemblance :

La condition de 1er ordre pour maximiser une fonction log-vraisemblance est :

$$\frac{\partial log\mathcal{L}}{\partial\beta} = \sum_0 \frac{(1/\sigma)f(Y_i - x_i\beta)x_i'}{1 - F_i} + \frac{1}{\sigma^2}\sum_1 (Y_i - x_i\beta)\, x_i' = 0$$

$$\frac{\partial log\mathcal{L}}{\partial\sigma^2} = \frac{1}{\sigma^2}\sum_0 \frac{x_i\beta(1/\sigma)f(Y_i - x_i\beta)}{1 - F_i} - \frac{N_1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_1 (Y_i - x_i\beta)^2 = 0$$

De ces équations ressort l'expression de l'estimateur par maximum de vraisemblance β (*Olsen 1980*) :

$$\beta = (X_1'X_1)^{-1}X_1'Y_1 - \sigma(X_1'X_1)^{-1}X_0'Y_0$$

4.2.2. Estimation en deux étapes :

Bien que 'il existe des progrès en terme d'estimation des modèles Tobit, la méthode du maximum de vraisemblance reste la plus adaptée. Par ailleurs, dans cette partie, nous

approcherons tout de même cette notion d'estimation en deux étapes. Cette dernière est issue de la procédure d'*Heckman (1976).*

Dans une première étape, l'estimation d'une fonction linéaire par la méthode des moindres carrés ordinaires fournira les résultats des paramètres β et σ à travers cette fonction :

$$Log\ \mathcal{L} = \sum_{i=1}^{1} \left\{ I_1 \left[\mathrm{F}\left(\frac{x_i\beta}{\sigma}\right)\right] + (1 - I_i) log \left[1 - f\left(\frac{x_i\beta}{\sigma}\right)\right]. \right.$$

En imposant $\sigma = 1$ les estimateurs du Maximum de vraisemblance deviennent convergent. Ainsi, la première étape d'estimation fournira les résultats β/σ et f/F.

Dans une deuxième étape les observations seront pondérée par $[\widehat{Var}(U_{ii})]^{1/2}$

4.3. Les modèles Tobit généralisés :

Cinq modèles appropriés à différentes situations peuvent être distingués, à savoir : le modèle Tobit simple dit du ***Type I***, le modèle Tobit du Type II, le modèle Tobit de Type III, IV et du type V. Dans cette partie nous nous intéressons au modèle Tobit du type II.

Dans les modèles Tobit du Type II, le modèle de censure dans lequel la variable d'intérêt est observée selon un processus distinct ; composé de deux variables latentes :

$$\begin{cases} Y_{1i}^* = x_{1i}\beta_1 + \varepsilon_{1i} \\ Y_{2i}^* = x_{2i}\beta_2 + \varepsilon_{2i} \\ Y_{2i} = Y_{2i}^*\ si\ Y_{1i}^* > 0 \\ Y_{2i} = 0\ si\ \ Y_{1i}^* \leq 0 \end{cases}$$

Où : x_{1i} et x_{2i} : Matrice de variables explicatives ;

$\beta_1$et β_2 : Deux vecteurs de paramètres ;

ε_{1i}etε_{2i} : Termes d'erreur qui suivent une distribution normale ;

Y_{2i}^*est observée seulement si $Y_{1i}^* > 0$.

Le modèle Tobit de type II peut être censuré si x_{1i} et x_{2i} sont observable même si $Y_{1i}^* \leq 0$ et il peut être tronqué si x_{2i} n'est pas observable lorsque $Y_{1i}^* \leq 0$.

La vraisemblance du modèle d'écrit comme suit :

$$\frac{\sigma_{12}}{\sigma_2^2(Y_{2i} - x_{2i}\,\beta_2)}$$

$$\mathcal{L} = \prod_0 [1 - \mathrm{F}\,(\frac{x_{1i}\beta_1}{\sigma_1})] \prod_1 \mathrm{F}\,(\frac{x_{1i}\beta_1 + (\sigma_{12}/\sigma_2^2)(Y_{2i} - x_{2i}\,\beta_2)}{\sqrt{\sigma_1^2 - (\sigma_{12}^2/\sigma_2^2)}} \times \frac{1}{\sigma^2} f(\frac{Y_{2i} - x_{2i}\,\beta_2}{\sigma^2})$$

4.4. Estimation et interprétation des résultats

L'estimation des modèles Tobit se fait à travers l'application du maximum de vraisemblance du fait de l'apparition fréquente des valeurs seuils « *c* » dans les équations de la variable dépendante. A cet effet, la méthode des moindres carrés ordinaires n'est guère applicable dans la résolution de ce type de modèle.

Contrairement aux modèles binaires et multinomiaux, l'interprétation des valeurs des paramètres du modèle estimé sont directement interprétables en termes de propension marginale surY_i^*, du fait de la continuité dans l'intervalle des seuils de la variable à expliquer. En revanche, La significativité des estimations du modèle est testée à l'aide des ratios appelés *Z-Statistique*.

4.5. Application :

Exercice extrait du livre : Bourbonnais, R. (2018)[10]

Prévision de la demande d'électricité pour un fournisseur à capacité limitée

Dans le cadre de l'ouverture du marché de l'électricité à destination des industriels, un fournisseur d'électricité, qui n'est pas l'opérateur historique, propose de l'énergie électrique à bas prix dans la limite de ses capacités fixées à 3 000 mégawatts : la demande supérieure à ce seuil ne peut donc pas être servie. Au-delà de ses capacités les clients sont délestés et sont donc dans l'obligation de basculer vers une autre source. La demande *(Y_t)* exprimée en mégawatts à la période *t* est fonction de trois facteurs explicatifs :

x_{1t}= indicateur d'écart de prix par rapport à la concurrence en *t*, la valeur indique le % de réduction, pour le jour considéré, accordé par l'opérateur historique (si 0 pas de réduction de prix),

x_{2t}= nombre de clients industriels alimentés en *t*,

x_{3t}= variable indicatrice signalant les jours particuliers à forte consommation tels que le lendemain de jour férié, …

Soit les données quotidiennes sur 60 jours dont cet opérateur dispose :

[10] *Bourbonnais, R. (2018), op cit, P366.*

Tableau 10 : Extrait de donnés

Jour	y	x_1	x_2	x_3
1	2717	0	61	0
2	2126	0	32	0
…	…	…	…	…
59	2683	0,1	61	0
60	3000	0	79	0
61		1	98	1
62		0	60	0

On demande :

1) d'estimer un modèle Logit permettant de prévoir la demande quotidienne à partir des facteurs explicatifs proposés et de commenter les résultats ;

2) de comparer les résultats avec l'estimation par la méthode, non adaptée, des moindres carrés ordinaires ;

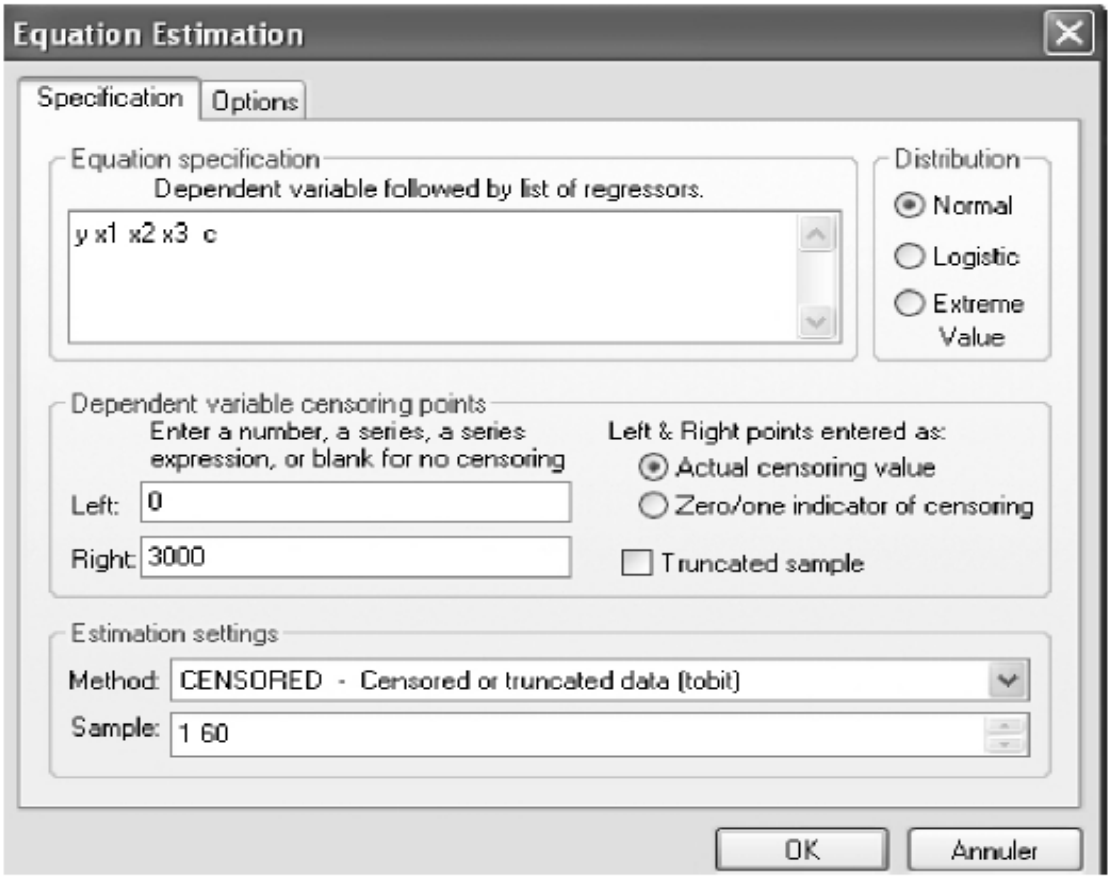
3) d'effectuer une prévision pour les jours 61 et 62 sachant que :

$x_{1,61} = 1; x_{2,61} = 98 x_{3,61} = 1$ Et $x_{1,62} = 0 x_{2,62} = 60 x_{3,62} = 0$.

- ***Solution :***

1) La consommation est censurée car les valeurs de la variable à expliquer (la demande) ne sont pas connues lorsqu'elles sortent de l'intervalle [0 ; 3 000] puisque au-delà de3 000, la demande ne peut être satisfaite.

Figure 02 : fenêtre de paramétrage pour un modèle Tobit (logiciel Eviews)



Pour estimer le modèle Tobit, le logiciel Eviews permet de paramétrer les seuils de censure à gauche *($c_1 = 0$)* et à droite *(c_2= 3 000)* et de choisir la distribution de l'erreur (ici normale). Nous ne sélectionnons donc pas l'option distribution tronquée puisque la distribution est censurée.

Les résultats d'estimation sont les suivants :

Tableau 11 : Estimation du modèle Tobit

Dependent Variable : Y
Method: ML – Censored Normal (TOBIT) (Quadratic hill climbing)
Included observations: 60 after adjustments
Left censoring (value) series: 0
Right censoring (value) series: 3000

	Coefficient	Std. Error	*z*-Statistic	Prob.
X1	– 27.69316	7.582649	– 3.652175	0.0003
X2	20.50361	0.480937	42.63261	0.0000
X3	186.2357	34.88046	5.339256	0.0000
C	1473.642	23.60984	62.41642	0.0000
Error Distribution				
SCALE: SIG	47.41983	4.617038	10.27062	0.0000
R-squared	0.981952	Mean dependant var		2542.667
S.E. of regression	46.09561	Akaike info criterion		9.355239
Sum squared resid	116864.3	Schwarz criterion		9.529768
Log likelihood	– 275.6572	Hannan-Quinn criter		9.423507
Avg. log likelihood	– 4.594286			
Left censored obs	0	Right censored obs		8
Uncensored obs	52	Total obs		60

a) Interprétation statistique

Les coefficients sont tous significativement différents de 0 (les probabilités critiques des coefficients sont toutes inférieures à 0,05), le modèle est validé sur le plan statistique. Eviews indique, sur l'avant dernière ligne, le nombre de données censurées : 0 à gauche et 8 à droite.

b) Interprétation économique

– L'indicateur d'écart de prix agit négativement sur la demande : en cas de réduction tarifaire de la concurrence, la demande diminue.

– Le nombre de clients connectés au réseau et la variable muette « type de jour » ont un effet positif sur la demande.

– La variable d'échelle (estimateur de σ) est égale à 47,41.

Les coefficients ont bien le signe attendu, le modèle est validé sur le plan économique.

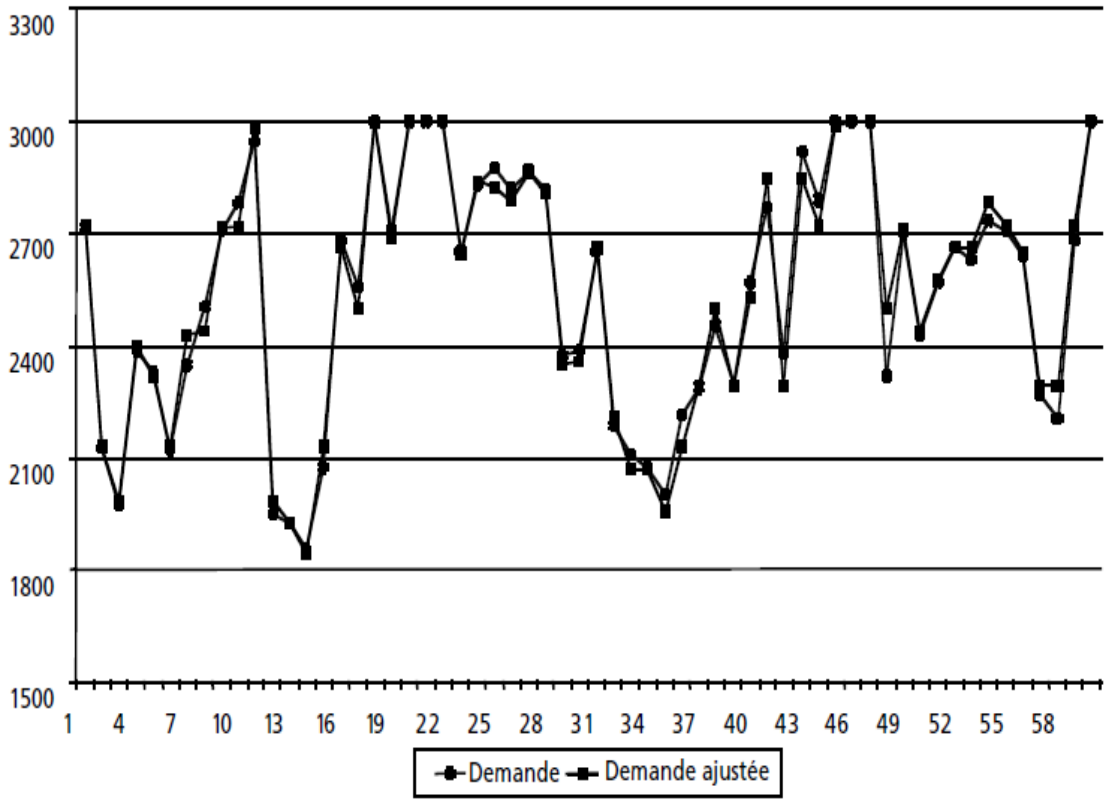
Le modèle Tobit s'écrit :

$$\widehat{Y_t^*} = -27{,}69 \times x_{1t} + 20{,}50 \times x_{2t} + 186{,}23 \times x_{3t} + 1473{,}64$$

$$Y_t = 0 \times F((0 - \widehat{Y_t^*})/47{,}41) + 3000 \times (1 - F((3000 - \widehat{Y_t^*})/47.41)) + (F((3000 - \widehat{Y_t^*})/47{,}41) - F((0 - \widehat{Y_t^*})/47{,}41) > 0) \times (\widehat{Y_t^*} \times F((3000 - \widehat{Y_t^*})/47.41) - F((0 - \widehat{Y_t^*})/47{,}41)) + 47{,}41 \times (-f((3000 - \widehat{Y_t^*})/47{,}41) + f((0 - \widehat{Y_t^*})/47{,}41))).$$

Avec Fla fonction de répartition de la loi normale et f la fonction de densité associée.Les graphiques de la série brute et de la série ajustée présentent la qualité de l'ajustement : les valeurs sont limitées par le seuil maximum, soit 3 000 mégawatts.

Graphique 01 : Série brute et ajustée de la variable y_t : Demande (modèle Tobit)



2) L'application de la méthode des moindres carrés ordinaires conduit aux résultats suivants:

Tableau 12 : Estimation par la méthode des moindres carrés ordinaires

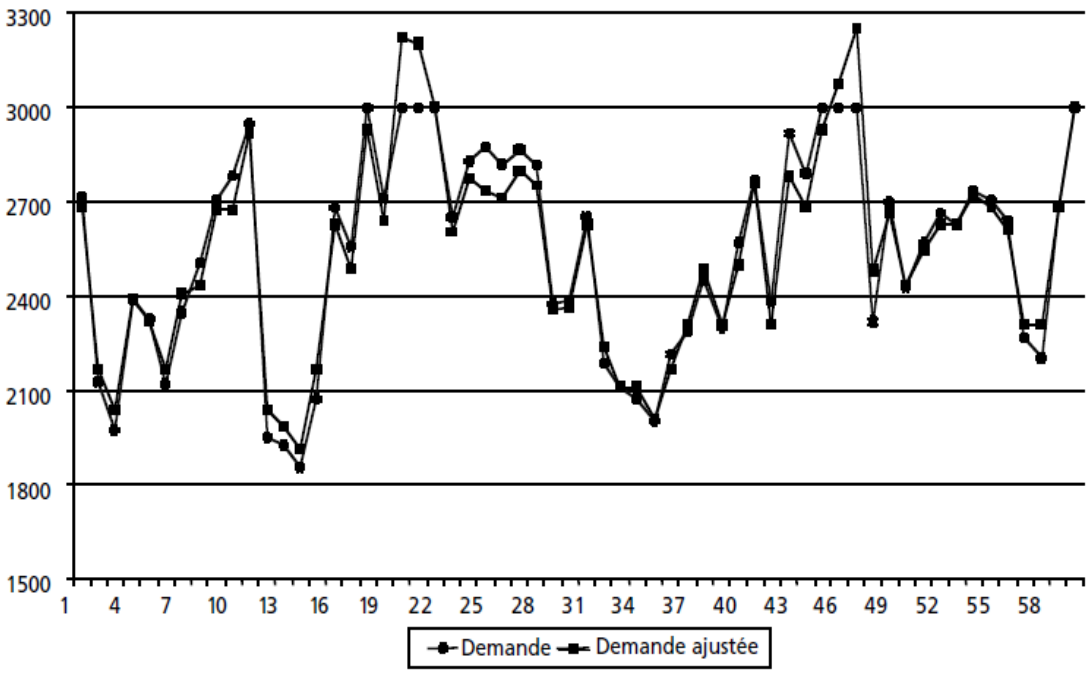
Dependent Variable : Y
Method: Least Squares
Included observations: 60 after adjustments

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
X1	– 31.95356	11.44578	– 2.791732	0.0072
X2	17.83772	0.659615	27.04260	0.0000
X3	156.0019	60.45055	2.580653	0.0125
C	1594.516	34.80522	45.81254	0.0000
R-squared	0.941213	F-statistic		298.8649
Sum squared resid	380662.1	Prob(F-statistic)		0.000000

Les coefficients du modèle, tous significativement différents de 0, sont biaisés. Nous constatons que les valeurs estimées des coefficients sont légèrement différentes, bien que de même signe, de celles des valeurs estimées à l'aide du modèle Tobit.

Les graphiques de la série brute et de la série ajustée indiquent que l'utilisation de la méthode des MCO peut fournir des valeurs estimées parfois supérieures à 3 000 mégawatts ce qui est techniquement impossible.

Graphique 02 : Série brute et ajustée de la variable y_t : Demande (Méthode MCO)



3) Le calcul de la prévision pour les jours 61 et 62 est directement effectué par application du modèle Tobit estimé. Sachant que :

$x_{1,61} = 1; x_{2,61} = 98 x_{3,61} = 1$ Et $x_{1,62} = 0 x_{2,62} = 60 x_{3,62} = 0$.

La prévision pour le jour 61 est donnée par :

$$Y_{61}^{*} = -27{,}69 \times 1 + 20{,}50 \times 98 + 186{,}23 \times 1 + 1473{,}64 = 3641{,}53$$

$$Y_{61}^{*} > c_2 \rightarrow Y_{61} = c_2 = 3000$$

La prévision pour le jour 62 est donnée par :

$$Y_{62}^{*} = -27{,}69 \times 0 + 20{,}50 \times 60 + 186{,}23 \times 0 + 1473{,}64 = 2703{,}85$$

$$c_1 < Y_{62}^{*} < c_2 \rightarrow Y_{62} = Y_{62}^{*} = 2703{,}85$$

Le calcul de la prévision à l'aide du modèle estimé par la méthode des MCO conduit à des résultats légèrement différents comme l'illustre le **tableau 13** :

***Tableau 13** : Comparaison des prévisions par le modèle Tobit et par la méthode MCO*

Jour	***Modèle Tobit***	***MCO***
61	*3000*	*3544,63*
62	*2703,86*	*2687,26*

Conclusion :

Les modèles économétriques étudiés, par les étudiants de Master spécialité Economie Quantitative à l'université de Bejaia, préalablement à ce cours, considèrent les modèles à variables dépendantes quantitatives. En revanche, ce cours traite des extensions directes des modèles linéaires conduisant in fine à des estimations non linéaires.

Ce cours met en exergue trois extensions de modèles à variable dépendante qualitatives. Les modèles dichotomiques qui permettent d'étudier les variables endogènes à deux valeurs, telle que $Y_i \in \{0, 1\}$. Ils définissent, en conséquence, la probabilité d'observer $Y = 1$ comme fonction $P(x_i\beta)$. Les coefficients du modèle sont estimés par la méthode du maximum de vraisemblance.

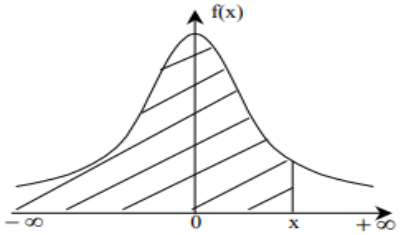
Quant aux modèles polytomiques présentés, ils relient la décision d'observation à la maximisation de l'utilité puisque Y prend ses valeurs dans un ensemble fini$\{0, 1, \dots \dots m\}$. Nous avons distingués le modèle multinomial ordonné, séquentiel et non ordonné.

Les modèles à variables limitées ont fait l'objet de la dernière partie traitée dans ce cours. Ces modèles sont caractérisés par une variable dépendante continue mais qui n'est observable que dans un intervalle donné. Nous avons, ainsi, distingué les variables tronquées et censurées qui caractérisent le modèle dit Tobit simple de type I. Nous avons, ensuite évoqué dans cette même dernière partie les modèles Tobit de type II.

Annexes :

Loi Normale centrée réduite

Probabilité de trouver une valeur inférieure à x.

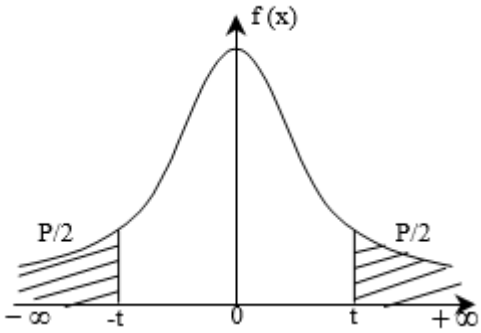


$$F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du$$

X	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Table de la loi de Student

Valeurs de t ayant la probabilité P d'être dépassées en valeur absolue.

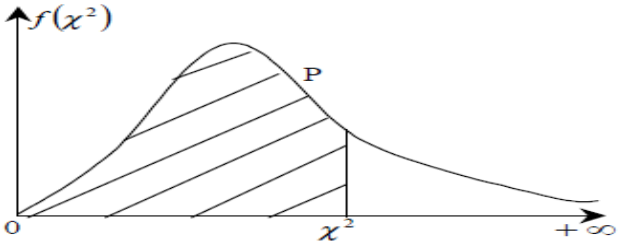


ϖ\ P	90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	1%
1	0,1584	0,3249	0,5095	0,7265	1,0000	1,3764	1,9626	3,0777	6,3137	12,7062	63,6559
2	0,1421	0,2887	0,4447	0,6172	0,8165	1,0607	1,3862	1,8856	2,9200	4,3027	9,9250
3	0,1366	0,2767	0,4242	0,5844	0,7649	0,9785	1,2498	1,6377	2,3534	3,1824	5,8408
4	0,1338	0,2707	0,4142	0,5686	0,7407	0,9410	1,1896	1,5332	2,1318	2,7765	4,6041
5	0,1322	0,2672	0,4082	0,5594	0,7267	0,9195	1,1558	1,4759	2,0150	2,5706	4,0321
6	0,1311	0,2648	0,4043	0,5534	0,7176	0,9057	1,1342	1,4398	1,9432	2,4469	3,7074
7	0,1303	0,2632	0,4015	0,5491	0,7111	0,8960	1,1192	1,4149	1,8946	2,3646	3,4995
8	0,1297	0,2619	0,3995	0,5459	0,7064	0,8889	1,1081	1,3968	1,8595	2,3060	3,3554
9	0,1293	0,2610	0,3979	0,5435	0,7027	0,8834	1,0997	1,3830	1,8331	2,2622	3,2498
10	0,1289	0,2602	0,3966	0,5415	0,6998	0,8791	1,0931	1,3722	1,8125	2,2281	3,1693
11	0,1286	0,2596	0,3956	0,5399	0,6974	0,8755	1,0877	1,3634	1,7959	2,2010	3,1058
12	0,1283	0,2590	0,3947	0,5386	0,6955	0,8726	1,0832	1,3562	1,7823	2,1788	3,0545
13	0,1281	0,2586	0,3940	0,5375	0,6938	0,8702	1,0795	1,3502	1,7709	2,1604	3,0123
14	0,1280	0,2582	0,3933	0,5366	0,6924	0,8681	1,0763	1,3450	1,7613	2,1448	2,9768
15	0,1278	0,2579	0,3928	0,5357	0,6912	0,8662	1,0735	1,3406	1,7531	2,1315	2,9467
16	0,1277	0,2576	0,3923	0,5350	0,6901	0,8647	1,0711	1,3368	1,7459	2,1199	2,9208
17	0,1276	0,2573	0,3919	0,5344	0,6892	0,8633	1,0690	1,3334	1,7396	2,1098	2,8982
18	0,1274	0,2571	0,3915	0,5338	0,6884	0,8620	1,0672	1,3304	1,7341	2,1009	2,8784
19	0,1274	0,2569	0,3912	0,5333	0,6876	0,8610	1,0655	1,3277	1,7291	2,0930	2,8609
20	0,1273	0,2567	0,3909	0,5329	0,6870	0,8600	1,0640	1,3253	1,7247	2,0860	2,8453
21	0,1272	0,2566	0,3906	0,5325	0,6864	0,8591	1,0627	1,3232	1,7207	2,0796	2,8314
22	0,1271	0,2564	0,3904	0,5321	0,6858	0,8583	1,0614	1,3212	1,7171	2,0739	2,8188
23	0,1271	0,2563	0,3902	0,5317	0,6853	0,8575	1,0603	1,3195	1,7139	2,0687	2,8073
24	0,1270	0,2562	0,3900	0,5314	0,6848	0,8569	1,0593	1,3178	1,7109	2,0639	2,7970
25	0,1269	0,2561	0,3898	0,5312	0,6844	0,8562	1,0584	1,3163	1,7081	2,0595	2,7874
26	0,1269	0,2560	0,3896	0,5309	0,6840	0,8557	1,0575	1,3150	1,7056	2,0555	2,7787
27	0,1268	0,2559	0,3894	0,5306	0,6837	0,8551	1,0567	1,3137	1,7033	2,0518	2,7707
28	0,1268	0,2558	0,3893	0,5304	0,6834	0,8546	1,0560	1,3125	1,7011	2,0484	2,7633
29	0,1268	0,2557	0,3892	0,5302	0,6830	0,8542	1,0553	1,3114	1,6991	2,0452	2,7564
30	0,1267	0,2556	0,3890	0,5300	0,6828	0,8538	1,0547	1,3104	1,6973	2,0423	2,7500
40	0,1265	0,2550	0,3881	0,5286	0,6807	0,8507	1,0500	1,3031	1,6839	2,0211	2,7045
50	0,1263	0,2547	0,3875	0,5278	0,6794	0,8489	1,0473	1,2987	1,6759	2,0086	2,6778
60	0,1262	0,2545	0,3872	0,5272	0,6786	0,8477	1,0455	1,2958	1,6706	2,0003	2,6603

80	0,1261	0,2542	0,3867	0,5265	0,6776	0,8461	1,0432	1,2922	1,6641	1,9901	2,6387
100	0,1260	0,2540	0,3864	0,5261	0,6770	0,8452	1,0418	1,2901	1,6602	1,9840	2,6259
120	0,1259	0,2539	0,3862	0,5258	0,6765	0,8446	1,0409	1,2886	1,6576	1,9799	2,6174
200	0,1258	0,2537	0,3859	0,5252	0,6757	0,8434	1,0391	1,2858	1,6525	1,9719	2,6006
∞	0,1257	0,2533	0,3853	0,5244	0,6745	0,8416	1,0364	1,2816	1,6449	1,9600	2,5758

Table de la loi de Khi-deux

Valeur de χ^2 *ayant la probabilité P d'être dépassée*



ddl/P	0,5%	1,0%	2,5%	5,0%	10,0%	50,0%	90,0%	95,0%	97,5%	99,0%	99,5%
1	0,000	0,000	0,001	0,004	0,016	0,455	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	1,386	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	2,366	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	3,357	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	4,351	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	5,348	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	6,346	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	7,344	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	8,343	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	9,342	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	10,341	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	11,340	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,041	12,340	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	13,339	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	14,339	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	15,338	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	16,338	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	17,338	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	18,338	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	19,337	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	20,337	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	21,337	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	22,337	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	23,337	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	24,337	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	25,336	35,563	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	18,114	26,336	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	27,336	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	28,336	39,087	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	20,599	29,336	40,256	43,773	46,979	50,892	53,672
31	14,458	15,655	17,539	19,281	21,434	30,336	41,422	44,985	48,232	52,191	55,002
32	15,134	16,362	18,291	20,072	22,271	31,336	42,585	46,194	49,480	53,486	56,328
33	15,815	17,073	19,047	20,867	23,110	32,336	43,745	47,400	50,725	54,775	57,648
34	16,501	17,789	19,806	21,664	23,952	33,336	44,903	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	24,797	34,336	46,059	49,802	53,203	57,342	60,275

Lorsque $\nu > 30$ on peut admettre que la quantité $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ suit une loi normale centrée réduite.

Bibliographie :

- Bourbonnais, R. (2018), « Econométrie », 10ème Edition Dunod.
- Crépon, B. Jacquemet, N. (2010), « Econométrie : Méthode et applications », Edition De Boeck, Bruxelles.
- Gourieroux, C. (1984), « Econométrie des variables qualitatives », Ed. Economica.
- Greene, W. (2002). «Econometric analysis», Ed.Prentice Hall, (seventh edition) Chapters 17 - 18 – 19, New Jersey.
- Hurlin, C. (2003), « Econométrie des variables qualitatives », chapitres 1-2-3, polycopié de cours, université d'Orléans.
- Maddala, G.S. (1986), « Limited-dependent and qualitative variables in econometrics», Cambridge University Press.
- Thomas, A. (2000). « Économétrie des variables qualitatives», éd DUNOD.
- Wooldridge, F. (2002). «Econometric analysis of cross section and panel data» Chapters 15 - 16 - 17 – 19, The MIT Press, Cambridge.

Table des matières :