URCA Hugo Harari-Kermadec

2008-2009 harari@ecogest.ens-cachan.fr

Économétrie, chapitre 4
Le modèle Tobit

I Données censurées

I.1 Introduction

Nous allons à présent envisager le cas des modèles à variable dépendante limitée : ceux sont des modèles pour lesquels la variable dépendante est continue mais n'est observable que sur un certain intervalle. Ainsi, ce sont des modèles qui se situent à mi chemin entre les modèles de régression linéaires où la variable endogène est continue et observable et les modèles qualitatifs. En effet, les modèles à variable dépendante limitée dérivent des modèles à variables qualitatives, dans le sens où l'on doit modéliser la probabilité que la variable dépendante appartienne à l'intervalle pour lequel elle est observable. Nous verrons que la structure de base des modèles à variable dépendante limitée est représentée par le modèle Tobit.
Toutefois, ces modèles sont aussi appelés modèles de régression censurées (censored regression models) ou modèle de régression tronquée (truncated regression models). Cette terminologie plus précise permet en effet d'introduire la distinction entre des échantillons tronqués et des échantillons censurés :

1. Un modèle de régression est dit tronqué lorsque toutes les observations des variables explicatives et de la variable dépendante figurant en dehors d'un certain intervalle sont totalement perdues.
2. Un modèle de régression est dit censuré lorsque l'on dispose au moins des observations des variables explicatives sur l'ensemble de l'échantillon. Nous verrons par la suite que le modèle Tobit est ainsi un modèle de régression censurée.

Dans son étude, Tobin cherche à modéliser la relation entre le revenu d'un ménage et les dépenses en biens durables. Il dispose pour cela d'un échantillon de $N = 100$ consommateurs. Tobin observe que lorsque l'on représente les couples revenus - dépenses des N consommateurs, la relation obtenue ressemble au graphique ci-dessous.

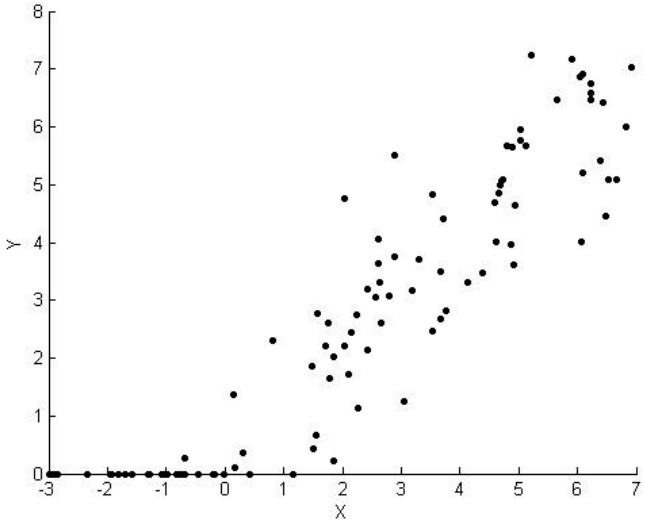

FIGURE 1 – Un échantillon censuré, avec une seule variable explicative.

Une des caractéristiques essentielles des données est que plusieurs observations pour le montant des dépenses de consommation sont nulles. En effet, ces observations sont nulles pour tous les ménages n'ayant pas acheté de biens durables sur la période. Pour ces individus, on dispose ainsi d'observations sur le revenu mais pas d'observations sur les dépenses de consommation : il s'agit d'un échantillon censuré.

I.2 Modèle Tobit

Plus formellement, considérons N couples de variables (X_i, Y_i^*) où la variable Y_i^* est telle que $\mathbb{E}[Y_i^*|X_i] = X_i\theta$, où θ est un vecteur de paramètres. jusqu'ici, le modèle linéaire est parfaitement adaptée à la situation. Malheureusement, et comme pour les variables qualitatives, la variable Y_i^* est une variable latente qui n'est pas toujours observable : on ne l'observe que si sa valeur est supérieure à un certain seuil c. On peut ainsi construire une variable Y_i observable, qui est égale à Y_i^* lorsque celle-ci est observable et qui vaut c par convention lorsque Y_i^* n'est pas observable. Le modèle Tobit est un modèle **censuré** : contrairement à Y_i^*, on observe X_i pour **tout l'échantillon**.
Cette propriété remet en cause l'hypothèse de linéarité et montre que les moindres carrés ordinaires ne sont pas la méthode pertinente pour estimer une telle relation. De façon plus générale, on ne peut pas utiliser ici une densité continue pour expliquer la distribution conditionnelle des dépenses par rapport au revenu : en effet, une distribution continue est incompatible avec le fait que plusieurs observations des dépenses soient nulles. C'est donc dans ce contexte que Tobin propose son modèle à variable dépendante limitée.
L'analyse économique de cette situation est la suivante : l'agent choisit la consommation en optimisant son utilité, sous une contrainte de budget. Si la consommation optimale est positive, il consomme la quantité optimale. Si la consommation optimale est négative, il ne consomme pas. L'agent est donc limité par une contrainte de positivité.
Le modèle s'écrit alors :

$$\begin{aligned} Y_i^* &= X_i\theta + \varepsilon_i \\ Y_i &= \begin{cases} Y_i^* & \text{si } Y_i^* > 0 \\ 0 & \text{si } Y_i^* \leq 0 \end{cases} \end{aligned}$$

θ et les X_i sont dans $\mathbb{R}^p$ et les résidus ε_i sont gaussiens homoskédastiques : $\varepsilon_i \sim \mathcal{N}(0, \sigma)$. On peut bien sûr spécifier une distribution non gaussienne, comme la loi de Weibull ou la loi Logistique.

I.3 Application des MCO à l'ensemble des observations

On va démontrer que l'estimateur des moindres carrés, pertinent dans le modèle linéaire, est biaisé dans notre cas. L'estimateur des MCO appliqué à l'ensemble des N couples d'observations (Y_i, X_i) est défini par la relation suivante :

$$\hat{\theta}_{MCO} = (X'X)^{-1} X'Y$$

On a alors, dans le cas où les X_i sont déterministes, un estimateur biaisé :

$$\mathbb{E}[\hat{\theta}_{MCO}] = (X'X)^{-1} X'\mathbb{E}[Y].$$

On remarque en effet que

$$\begin{array}{ll} \mathbb{E}[Y_i] & = \mathbb{E}[Y_i \mathbb{1}_{Y_i^*>0}] + \mathbb{E}[Y_i \mathbb{1}_{Y_i^*\leq 0}] \\ & = \mathbb{E}[(X_i\theta + \varepsilon_i)\mathbb{1}_{\varepsilon_i > -X_i\theta}] \\ & = X_i\theta(1-\Phi(-X_i\theta/\sigma)) + \sigma\int_{-X_i\theta/\sigma}^{+\infty} -\phi'(t)dt \end{array} \quad \begin{array}{l} = \mathbb{E}[Y_i^* \mathbb{1}_{Y_i^*>0}] + 0\ \mathbb{P}(Y_i^* \leq 0) \\ = X_i\theta\ \mathbb{P}(\varepsilon_i > -X_i\theta) + \sigma\int_{-X_i\theta}^{+\infty} t\phi(t)dt \\ = X_i\theta\ \Phi(X_i\theta/\sigma) + \sigma\phi(X_i\theta/\sigma), \end{array}$$

car $\varphi'(t) = -t\varphi(t)$.
On a donc

$$\mathbb{E}[Y_i] = X_i\theta\ \Phi(X_i\theta/\sigma) + \sigma\phi(X_i\theta/\sigma) \neq X_i\theta \text{ et donc } \mathbb{E}[Y] \neq X\theta.$$

L'espérance de notre estimateur, $\mathbb{E}[\hat{\theta}_{MCO}]$, n'est donc pas θ : l'estimateur des moindres carrés est **biaisé**.

I.4 Application des MCO aux observations strictement positives

On pourrait pensé que le biais provient des observations censurées, et chercher à utiliser l'estimateur MCO sur la partie linéaire du modèle. L'estimateur des MCO appliqué aux couples d'observations non censurées est défini par la relation suivante :

$$\hat{\theta}_{MCO2} = \left(\sum_{i|Y_i>0} X_i'X_i\right)^{-1} \sum_{i|Y_i>0} X_i'Y_i$$

On a à nouveau, dans le cas où les X_i sont déterministes, un estimateur biaisé :

$$\mathbb{E}[\hat{\theta}_{MCO2}] = \left(\sum_{i|Y_i>0} X_i'X_i\right)^{-1} \sum_{i|Y_i>0} X_i'\mathbb{E}[Y_i|Y_i>0].$$

On remarque en effet que

$$\begin{array}{ll} \mathbb{E}[Y_i|Y_i>0] & = \mathbb{E}[Y_i\mathbb{1}_{Y_i>0}]/\mathbb{P}(Y_i>0) \\ & = X_i\theta + \sigma\frac{\phi(X_i\theta/\sigma)}{\Phi(X_i\theta/\sigma)} \end{array} \quad \begin{array}{l} = \mathbb{E}[Y_i^*\mathbb{1}_{Y_i^*>0}]/\mathbb{P}(Y_i^*>0) \\ \neq X_i\theta. \end{array}$$

L'espérance de notre estimateur, $\mathbb{E}[\hat{\theta}_{MCO2}]$, n'est donc pas θ : l'estimateur des moindres carrés sur les données positives est à nouveau **biaisé**.

II Analyse statistique

II.1 Estimation par maximum de vraisemblance

La procédure d'estimation la plus utilisée aujourd'hui est celle du maximum de vraisemblance. Commençons par définir la log-vraisemblance associée au modèle Tobit :

$$Y_i = \begin{cases} Y_i^* & \text{si } Y_i^* = X_i\theta + \varepsilon_i \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

La vraisemblance de ce modèle est définie par :

$$L(\theta, \sigma^2) = \prod_{i:Y_i=0} \left[1 - \Phi\left(\frac{X_i\theta}{\sigma}\right)\right] \prod_{i:Y_i>0} \frac{1}{\sigma}\phi\left(\frac{Y_i - X_i\theta}{\sigma}\right).$$

Le premier produit est similaire à celui obtenu pour le modèle probit, puisque les deux modélisations sont identiques pour l'événement $Y_i = 0$. Le second produit correspond à celui obtenu dans le modèle linéaire, puisqu'il s'agit simplement de la vraisemblance d'un échantillon gaussien. Ceci est cohérent avec notre modèle tobit, intermédiaire entre modèle linéaire et modèle dichotomique.
La log-vraisemblance s'écrit

$$\log L(\theta, \sigma^2) = \sum_{i:Y_i=0} \log\left[1 - \Phi\left(\frac{X_i\theta}{\sigma}\right)\right] - N_1 \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i:Y_i>0}(Y_i - X_i\theta)^2.$$

où N_1 est le nombres d'observations positives. On en tire les scores :

$$\frac{\partial \log L(\theta,\sigma^2)}{\partial \theta'} = -\frac{1}{\sigma}\sum_{i:Y_i=0}\frac{\phi\left(\frac{X_i\theta}{\sigma}\right)X_i'}{1-\Phi\left(\frac{X_i\theta}{\sigma}\right)} + \frac{1}{\sigma^2}\sum_{i:Y_i>0}(Y_i - X_i\theta)X_i',$$

et

$$\frac{\partial \log L(\theta,\sigma^2)}{\partial \sigma^2} = \frac{1}{2\sigma^3}\sum_{i:Y_i=0}\frac{\phi\left(\frac{X_i\theta}{\sigma}\right)X_i\theta}{1-\Phi\left(\frac{X_i\theta}{\sigma}\right)} - \frac{N_1}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i:Y_i>0}(Y_i - X_i\theta)^2.$$

L'estimateur du maximum de vraisemblance, $(\hat{\theta}, \widehat{\sigma^2})$ est sans biais, asymptotiquement gaussien de matrice de variance-covariance égale à l'inverse de la matrice d'information de Fisher.

II.2 Interprétation

Pour connaitre l'effet d'une variable explicative, on étudie les variations de l'espérance de Y. On a déjà calculé

$$\mathbb{E}[Y_i] = \mathbb{E}[Y_i^* \mathbb{1}_{Y_i^*>0}] = X_i\theta\Phi\left(\frac{X_i\theta}{\sigma}\right) + \sigma\varphi\left(\frac{X_i\theta}{\sigma}\right),$$

On peut calculer les effets marginaux :

$$\begin{aligned}\frac{\partial \mathbb{E}[Y_i]}{\partial X_{i1}} &= \theta_1\Phi(X_i\theta/\sigma) + X_i\theta\frac{\theta_1}{\sigma}\varphi(X_i\theta/\sigma) + \theta_1\varphi'(X_i\theta/\sigma)\\ &= \theta_1\Phi(X_i\theta/\sigma) + X_i\theta\frac{\theta_1}{\sigma}\varphi(X_i\theta/\sigma) - \theta_1\frac{X_i\theta}{\sigma}\varphi(X_i\theta/\sigma) = \theta_1\Phi(X_i\theta/\sigma),\end{aligned}$$

car $\varphi'(t) = -t\varphi(t)$.
L'effet marginal de X_{i1} est donc du même signe que θ_1. On peut remarquer que l'effet est modulé par $\Phi(X_i\theta/\sigma)$:
- Si $X_i\theta/\sigma$ est grand et négatif, $\Phi(X_i\theta/\sigma)$ est proche de 0 et l'on a beaucoup de chances d'être censuré, l'effet de X_1 est donc diminué d'un facteur $\Phi(X_i\theta/\sigma)$.
- Si $X_i\theta/\sigma$ est grand et positif, $\Phi(X_i\theta/\sigma)$ est proche de 1 et l'on a peu de chances d'être censuré, l'effet de X_1 est pratiquement plein, c'est-à-dire égal à θ_1.

La modulation joue au même niveau pour toutes les variables explicatives.

III Procedure Lifereg

La version 8 de SAS n'offre malheureusement pas de commande directe pour estimer le modèle Tobit. Un usage astucieux de la procédure LIFEREG, plus spécifiquement destinée à l'estimation des modèles de durée, permet néanmoins de pallier cette lacune. On décrit simplement la manière d'opérer.

III.1 Entrée

Soit Y la variable à expliquer, censurée à gauche et égale à 0 lorsque la variable y^* sous-jacente n'est pas observable. Il faut construire une variable **BINF** qui indique les données censurées. BINF doit être **manquante** quand Y est nulle, c'est-à-dire BINF="." . Soient X_1, X_2, X_3 les explicatives envisagées, on utilise la forme suivante de la procédure LIFEREG :

PROC LIFEREG DATA=donnees[1] ;
  BY sexe[2]
  MODEL (BINF, Y)[3] = X_1 X_2 X_3 / D = NORMAL[4] ALPHA= 0.05[5] ;
RUN ;

1. Nom de la table à utiliser.
2. BY sexe : lance la procédure sur les sous-populations définies par les valeurs de la variable sexe. Il faut avoir ordonner la table avant, avec :
   PROC SORT DATA=données ; BY sexe ; run ; Attention : si la table est ouverte, il est impossible de la manipuler et donc de l'ordonner.
3. Spécification de la censure.
4. D= WEIBUL (par défaut) ou NORMAL ou LOGISTIC : la loi des résidus.
5. ALPHA= 0.05 (par défaut) ou 0.1 ou 0.01 : c'est le niveau de confiance des IC.

La version 9 de SAS remédie opportunément à la lacune précédente et la nouvelle procédure QLIM, du module ETS, permet d'estimer de manière simple et naturelle les modèles de Tobin et connexes.

III.2 Sortie

On veut expliquer la consommation Y en produits bio par l'age, le sexe et le revenu. Une fois les variables pré-traitées et la variable de censure BINF construite, on tape donc :

```
PROC LIFEREG DATA=bio2;
    MODEL (BINF, Y) = age revenu sexe/ D = normal;
RUN;
```

On obtient :

```
                              Le Système SAS        10:27 Monday, January 29, 2007   9
                           The LIFEREG Procedure
                            Model Information
                     Data Set                        WORK.BIO2
                     Dependent Variable                  lower
                     Dependent Variable                      y
                     Number of Observations                405
                     Noncensored Values                    294
                     Right Censored Values                   0
                     Left Censored Values                  111
                     Interval Censored Values                0
                     Name of Distribution               Normal
                     Log Likelihood               -1578.060367

                      Number of Observations Read          405
                      Number of Observations Used          405
Algorithm converged.

                        Analyse des effets de Type III
                                      Khi 2
                   Effet        DF       de Wald    Pr > Khi 2
                   AGE           1       15.4386        <.0001
                   REVENU        1        2.5547        0.1100
                   SEXE          1        0.0045        0.9463

                  Analyse des Résultats estimés du paramètre
                                Erreur    95Limites de
Paramètre DF Estimation standard    confiance %      Khi 2 Pr > Khi 2
Intercept  1    -2.5731   9.3834 -20.9642  15.8180    0.08     0.7839
AGE        1     5.6333   1.4337   2.8233   8.4433   15.44     <.0001
REVENU     1    -2.3255   1.4550  -5.1772   0.5261    2.55     0.1100
SEXE       1     0.2897   4.3033  -8.1446   8.7240    0.00     0.9463
Scale      1    39.0366   1.6605  35.9141  42.4306
```

On retrouve certaines sorties de la procédure LOGISTIC.

Statistiques générales

SAS commence par donner quelques informations descriptives sur les données et le modèle. On retrouve en particuliers la log-vraisemblance. le critère AIC, c'est-à-dire la vraisemblance corrigée du nombre p de variables explicatives n'est plus donnée. On la retrouve en retirant $2p$ à la vraisemblance.

Analyse des effets de type III

Donne des informations sur les effets des variables explicatives, utiles pour un modèle non-linéaire. Dans le cas du modèle Tobit, on retrouve exactement les informations ci-dessous.

Analyse des Résultats estimés du paramètre

Donne les valeurs estimées pour les $\hat{\theta}_j$ et des indications de leur significativité. La première ligne donne l'estimation de la constante $\hat{\theta}_0$. Les lignes suivantes donnent les coefficients de variables explicatives. La dernière ligne donne l'estimation de l'écart-type σ des résidus.

DF
: Rappelle la dimension du paramètre.

Estimation
: Donne $\hat{\theta}_j$.

Erreur std
: Donne l'estimation de l'écart-type : $\sqrt{I_{jj}(\hat{\theta})}$.

Khi 2 de Wald
: Donne la valeur du test de Wald pour la nullité du coefficient θ_j.

Pr > Khi 2
: Donne la p-value de ce test.

IV TP SAS

On va s'attaquer au jeu de donnée pbio, disponible à l'URL

http ://www.info.univ-angers.fr/pub/gh/Datasets/pbio.htm

De nombreuses variables de ce jeu de données sont des variables socio-démographiques (ex : Etat-civil) qui ne peuvent être utilisées directement. Il faut les traduire, sans les effacer, avant de les utiliser (ex : en construisant une indicatrice de "marié"). Pensez à fermer la table avant d'entamer une étape DATA.

Cette fois, le rapport doit être rendu en fin de session et servira d'examen.

1. Mise en jambe
   (a) Importer les données dans la librairie WORK, nommer la table BIO1.
   (b) Faire une copie BI02.
   (c) Lire le questionnaire Pbio.dsc joint à la base de donnée. Les données ne sont pas directement exploitables. Construire une table BIO3 avec les variables utiles, et en particulier une variable Y traduisant en valeurs numériques les réponses contenues dans CONSOM et la variable BINF donnant la censure.
2. Analyse statistique globale
   (a) Chercher le meilleur modèle possible pour expliquer la Y. Vous pouvez améliorer le modèle en choisissant au mieux les variables explicatives et la loi des résidus (Weibull, normal ou logistique). Attention à l'usage des variables socio-démographiques !
   (b)
3. Analyse statistique stratifiée
   (a) Reprendre l'étude séparément pour les femmes et les hommes.