

Cours d'économétrie 2

Première partie : variables dépendantes limitées

# Chapitre 3

## Censure et sélection

Xavier d'Haultfœuille

## 1 Introduction

On s'intéresse maintenant à une variable continue mais imparfaitement observée :

- On observe  $Y = \max(0, Y^*)$  où  $Y^*$  suit un modèle linéaire : modèle de censure ou tobit simple.
- On observe  $Y$  seulement lorsque  $D = 1$  : modèle de sélection.

Dans ces deux cas, l'estimateur des MCO n'est pas convergent en général.

## 2 Modèles de censure (ou tobit simple)

### 2.1 Présentation

On considère le modèle suivant :

$$\begin{cases} Y^* = X'\beta_0 + \sigma_0\varepsilon, & \varepsilon|X \sim \mathcal{N}(0, 1) \\ Y = \max(0, Y^*) = Y^*\mathbf{1}\{Y^* > 0\} \end{cases}$$

Ce type de situation peut survenir principalement pour deux raisons :

- 1) Problème d'observation des données : on observe  $Y^*$  si  $Y^*$  est inférieur à un seuil  $s$ , et le seuil sinon. En d'autres termes, on observe  $Y = \min(s, Y^*)$  et donc

$$-Y + s = -\min(s, Y^*) + s = \max(0, -Y^* + s) \equiv \max(0, Y^{**}) \text{ avec } Y^{**} = -Y^* + s.$$

Exemple : revenus, score à un test, demandes de réservation pour un train ou un avion etc.

- 2)  $Y$  est la solution d'un programme de maximisation sur  $[0; +\infty[$  qui peut admettre une solution en coin. Exemple : consommation d'un bien.

**N.B.** : on parle encore de tobit de type I pour ces modèles.

## 2.2 Paramètres d'intérêt du modèle.

- Dans le cas de données censurées, nous sommes intéressés par l'effet marginal de  $X$  sur la “vraie” variable  $Y^*$  :

$$\frac{\partial E(Y^* | X_k = x_k, X_{-k} = x_{-k})}{\partial x_k} = \beta_{0k}.$$

- Dans le cas de solutions en coin, la variable d'intérêt est  $Y$  et non  $Y^*$  et les paramètres d'intérêt sont plutôt  $\partial E(Y|X)/\partial x_k$  et  $\partial E(Y|X, Y > 0)/\partial x_k$ .

Pour les calculer, notons que pour  $U \sim \mathcal{N}(0, 1)$ , on a (cf. chap. 2, 2.4) :

$$E(U|U > c) = \frac{\varphi(c)}{1 - \Phi(c)}$$

Ici, on a

$$\begin{aligned} E(Y|X = x, Y > 0) &= x'\beta_0 + \sigma_0 E(\varepsilon|X = x, \varepsilon > -X'\beta_0/\sigma_0) \\ &= x'\beta_0 + \sigma_0 E(\varepsilon|\varepsilon > -x'\beta_0/\sigma_0) \\ &= x'\beta_0 + \sigma_0 \frac{\varphi(-x'\beta_0/\sigma_0)}{1 - \Phi(-x'\beta_0/\sigma_0)} \\ &= x'\beta_0 + \sigma_0 \frac{\varphi(x'\beta_0/\sigma_0)}{\Phi(x'\beta_0/\sigma_0)} \end{aligned}$$

## 2.2 Paramètres d'intérêt du modèle.

La fonction  $\lambda(x) = \varphi(x)/\Phi(x)$  est appelée inverse du ratio de Mills. On a alors :

$$E(Y|X = x, Y > 0) = x'\beta_0 + \sigma_0\lambda\left(\frac{x'\beta_0}{\sigma_0}\right).$$

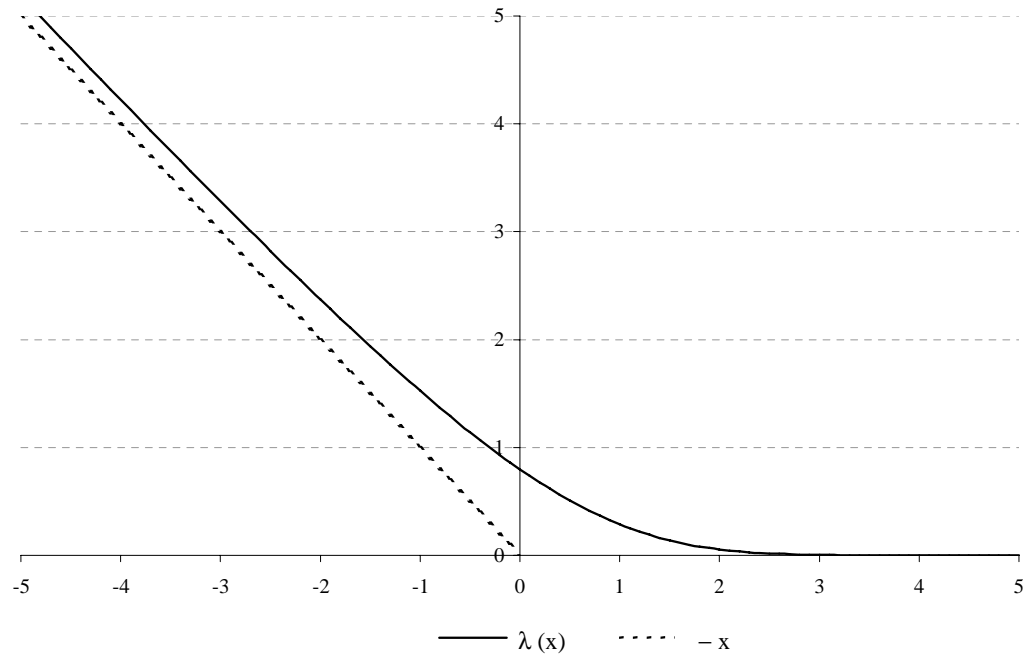


FIG. 1 – Inverse du ratio de Mills.

## 2.2 Paramètres d'intérêt du modèle.

On a par ailleurs :

$$\lambda'(x) = -\lambda(x) [x + \lambda(x)] \text{ et } \lambda'(x) \in ] - 1, 0[.$$

Par conséquent,

$$\begin{aligned} \frac{\partial E(Y|X_k = x_k, X_{-k} = x_{-k}, Y > 0)}{\partial x_k} &= \beta_{0k} \left\{ 1 + \lambda' \left( \frac{x' \beta_0}{\sigma_0} \right) \right\} \\ &= \beta_{0k} \left\{ 1 - \lambda \left( \frac{x' \beta_0}{\sigma_0} \right) \left[ \frac{x' \beta_0}{\sigma_0} + \lambda \left( \frac{x' \beta_0}{\sigma_0} \right) \right] \right\} \end{aligned}$$

Cet effet est donc compris entre 0 et  $\beta_{0k}$ . Comme pour les modèles binaires, on peut s'intéresser (par exemple) à l'effet marginal moyen :

$$E \left[ \frac{\partial E(Y|X, Y > 0)}{\partial x_k} \right] = \beta_{0k} \left\{ 1 + E \left[ \lambda' \left( \frac{X' \beta_0}{\sigma_0} \right) \right] \right\}.$$

## 2.2 Paramètres d'intérêt du modèle.

Pour calculer  $E(Y|X)$ , on remarque que

$$\begin{aligned}
 E(Y|X = x) &= P(Y > 0|X = x)E(Y|X = x, Y > 0) \\
 &= P(\varepsilon > -X'\beta_0/\sigma_0|X = x)E(Y|X = x, Y > 0) \\
 &= \Phi\left(\frac{x'\beta_0}{\sigma_0}\right)x'\beta_0 + \sigma_0\varphi\left(\frac{x'\beta_0}{\sigma_0}\right).
 \end{aligned}$$

Par conséquent,

$$\begin{aligned}
 \frac{\partial E(Y|X_k = x_k, X_{-k} = x_{-k})}{\partial x_k} &= \varphi\left(\frac{x'\beta_0}{\sigma_0}\right)\frac{\beta_{0k}}{\sigma_0}x'\beta_0 + \Phi\left(\frac{x'\beta_0}{\sigma_0}\right)\beta_{0k} - \sigma_0\frac{x'\beta_0}{\sigma_0}\varphi\left(\frac{x'\beta_0}{\sigma_0}\right) \times \frac{\beta_{0k}}{\sigma_0} \\
 &= \Phi\left(\frac{x'\beta_0}{\sigma_0}\right)\beta_{0k} \\
 &\left( = P(Y > 0|X = x)\beta_{0k} \right)
 \end{aligned}$$

Comme précédemment, l'effet marginal est compris entre 0 et  $\beta_{0k}$ . Notons que l'effet marginal moyen a une forme très simple :

$$E\left[\frac{\partial E(Y|X)}{\partial x_k}\right] = E[P(Y > 0|X)]\beta_{0k} = P(Y > 0)\beta_{0k}.$$

### 2.3 Estimation du modèle

L'estimateur des MCO n'est pas convergent dans ce modèle. En effet,

$$\begin{aligned}
 E(Y|X) &= P(Y^* \geq 0|X) \times E(Y|X, Y^* \geq 0) + P(Y^* < 0|X) \times E(Y|X, Y^* < 0) \\
 &= P(Y^* \geq 0|X) \times E(Y^*|X, Y^* \geq 0) + P(Y^* < 0|X) \times 0 \\
 &> P(Y^* \geq 0|X) \times E(Y^*|X, Y^* \geq 0) + P(Y^* < 0|X) \times E(Y^*|X, Y^* < 0) \\
 &= E(Y^*|X) = X'\beta_0
 \end{aligned}$$

Donc, en général,  $E(XY) = E(XE(Y|X)) \neq E(XX')\beta_0$  et ainsi

$$\text{plim}\hat{\beta}_{MCO} = E(XX')^{-1}E(XY) \neq \beta_0.$$

De même, la régression sur les données censurées seules ne conduit pas à un estimateur convergent.

En effet :

$$E(Y|X, Y > 0) = X'\beta_0 + \sigma_0\lambda \left( \frac{X'\beta_0}{\sigma_0} \right) > X'\beta_0 = E(Y^*|X). \quad (1)$$

## 2.3 Estimation du modèle

On utilise le maximum de vraisemblance.

Problème :  $Y$  a une distribution continue sur  $]0, +\infty[$  mais une masse en 0 :

$$P(Y = 0|X = x) = 1 - \Phi\left(\frac{x'\beta_0}{\sigma_0}\right) = \Phi\left(\frac{-x'\beta_0}{\sigma_0}\right) > 0.$$

Donc  $P^{Y|X=x}$  n'est pas absolument continue par rapport à la mesure de Lebesgue  $\lambda$ .

En revanche,  $P^{Y|X=x}$  est dominée par la mesure  $\mu = \lambda + \delta_0$  où  $\lambda$  est la mesure de Lebesgue et  $\delta_0$  est la mesure de Dirac en 0 (la mesure de Dirac en  $a$  est telle que pour tout ensemble mesurable  $A$ ,  $\delta_0(A) = 1$  si  $a \in A$ , 0 sinon).

De plus, on peut montrer (cf. exercice 2 du TD1 de statistique 1) que

$$\begin{aligned}\frac{dP^{Y|X=x}}{d\mu}(y) &= \mathbf{1}\{y = 0\}P(Y = 0|X = x) + \mathbf{1}\{y > 0\}f_{Y^*|X=x}(y) \\ &= \mathbf{1}\{y = 0\}\Phi\left(\frac{-x'\beta_0}{\sigma_0}\right) + \mathbf{1}\{y > 0\}\frac{1}{\sigma_0}\varphi\left(\frac{y - x'\beta_0}{\sigma_0}\right)\end{aligned}$$

## 2.3 Estimation du modèle

La log-vraisemblance d'un échantillon i.i.d s'écrit donc :

$$\begin{aligned} l_n(\beta, \sigma) &= \sum_{i/Y_i=0} \ln \Phi \left( \frac{-X'_i \beta}{\sigma} \right) + \sum_{i/Y_i>0} \ln \varphi \left( \frac{Y_i - X'_i \beta}{\sigma} \right) - N_+ \ln \sigma \\ &= \sum_{i/Y_i=0} \ln \Phi \left( \frac{-X'_i \beta}{\sigma} \right) - \frac{1}{2} \sum_{i/Y_i>0} \left( \frac{Y_i - X'_i \beta}{\sigma} \right)^2 - N_+ \ln \sigma - N_+ \ln \sqrt{2\pi} \end{aligned}$$

où  $N_+$  est le nombre d'observations non censurées. Pour rendre le programme de maximisation concave on effectue le changement de variables  $b = \beta/\sigma$  et  $s = 1/\sigma$ . Il s'agit alors de maximiser :

$$\tilde{l}_n(b, s) = \sum_{i/Y_i=0} \ln \Phi(-X'_i b) - \frac{1}{2} \sum_{i/Y_i>0} (sY_i - X'_i b)^2 + N_+ \ln s - N_+ \ln \sqrt{2\pi}$$

Comme d'habitude, on peut montrer que l'estimateur du maximum de vraisemblance  $(\hat{\beta}, \hat{\sigma})$  est asymptotiquement normal (exercice : calculer sa variance asymptotique).

## 2.4 Extensions

Deux directions peuvent être prises :

- Modélisations alternatives des solutions en coin ;
- relâcher l'hypothèse  $\varepsilon|X \sim \mathcal{N}(0, 1)$  (*hors programme*)

## Modélisations alternatives

Une limite du modèle tobit : un mécanisme unique détermine  $Y > 0$  vs  $Y = 0$  et la quantité  $Y$  sachant  $Y > 0$ . Dans certains cas il est plus judicieux de supposer qu'il existe deux mécanismes ("two-tiered" model) :

$$\begin{aligned} P(Y = 0|X) &= 1 - \Phi(X'\gamma_0) \\ \ln Y|X, Y > 0 &\sim \mathcal{N}(X'\beta_0, \sigma_0^2) \end{aligned}$$

Pour estimer  $\gamma_0$ , il suffit d'effectuer un probit sur  $W = \mathbf{1}\{Y > 0\}$ . On estime  $\beta_0$  et  $\sigma_0$  en régressant les  $\ln Y_i$  sur les  $X_i$  sur l'échantillon des  $Y_i > 0$ .

Autre modèle possible : Cragg (1971) :

$$\begin{aligned} P(Y = 0|X = x) &= 1 - \Phi(x'\gamma_0) \\ f_{Y|X, Y > 0}(y, x) &= \frac{\varphi\left(\frac{y-x'\beta_0}{\sigma_0}\right)}{\sigma_0 \Phi\left(\frac{x'\beta_0}{\sigma_0}\right)} \mathbf{1}\{y > 0\} \end{aligned}$$

Le modèle tobit est un cas particulier de ce modèle, lorsque  $\gamma_0 = \beta_0/\sigma_0$ . On peut donc tester la validité du modèle tobit contre une alternative plus général du modèle de Cragg (par un test du score par exemple).

## Approche semiparamétrique (*hors programme*)

L'estimateur du maximum de vraisemblance n'est pas convergent en général si les résidus sont non normaux et/ou hétéroscédastiques.

Approche semiparamétrique : restriction sur la médiane (Powell, 1984). On suppose que  $\varepsilon$  a une loi continue et

$$\text{Med}(\varepsilon|X) = 0.$$

$\Rightarrow$  Pas d'hypothèse de loi ni d'indépendance entre  $\varepsilon$  et  $X$ .

Pour toute v.a.r.  $U$ , on définit la médiane de  $U$  par :

$$\text{Med}(U) = \inf\{x/P(U \leq x) \geq 1/2\}.$$

## Approche semiparamétrique (*hors programme*)

1. si  $x'\beta_0 \leq 0$ ,

$$\begin{aligned} P(Y \leq 0|X = x) &= P(Y = 0|X = x) \\ &= P(\varepsilon \leq -x'\beta_0|X = x) \\ &\geq P(\varepsilon \leq 0|X = x) = \frac{1}{2} \end{aligned}$$

De plus, pour tout  $y < 0$ ,

$$P(Y \leq y|X = x) = 0.$$

Donc  $\text{Med}(Y|X = x) = 0$ .

2. si  $x'\beta_0 > 0$ ,

$$P(Y \leq x'\beta_0|X = x) = P(\varepsilon \leq 0|X = x) = \frac{1}{2}$$

Et pour tout  $y < x'\beta_0$ ,

$$\begin{aligned} P(Y \leq y|X = x) &\leq P(\varepsilon \leq y - x'\beta_0|X = x) \\ &< P(\varepsilon \leq 0|X = x) = \frac{1}{2} \end{aligned}$$

Donc  $\text{Med}(Y|X = x) = x'\beta_0$

Finalement,

$$\text{Med}(Y|X) = \max(0, X'\beta_0).$$

## Approche semiparamétrique (*hors programme*)

Pour estimer  $\beta_0$ , on utilise ensuite le lemme suivant :

**Lemme 1** *Pour toute variable aléatoire  $U$ , on*

$$\text{Med}(U) \in \arg \min_a E[|U - a|]$$

**Preuve** (pour des v.a. continues). On a :

$$\begin{aligned} E[|U - a|] &= \int_{-\infty}^{+\infty} |u - a| f_U(u) du \\ &= \int_{-\infty}^a (a - u) f_U(u) du + \int_a^{\infty} (u - a) f_U(u) du \end{aligned}$$

Donc l'application  $g(a) = E[|U - a|]$  est dérivable et

$$\begin{aligned} g'(a) &= (a - a) f_U(a) + \int_{-\infty}^a f_U(u) du - (a - a) f_U(a) + \int_a^{\infty} f_U(u) du \\ &= 2F_U(a) - 1 \end{aligned}$$

Ainsi, la fonction  $g$  est convexe et son minimum vérifie  $1 - 2F_U(a_0) = 0$ , soit encore  $a_0 = F^{-1}(1/2)$   $\square$

## Approche semiparamétrique (*hors programme*)

Ici, on a donc

$$\beta_0 \in \arg \min_{\beta} E \left[ |Y - \max(0, X'\beta)| \mid X \right]$$

Et sous des conditions de régularité sur  $X$ ,

$$\beta_0 = \arg \min_{\beta} E [|Y - \max(0, X'\beta)|]$$

On estime alors  $\beta_0$  par :

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n |Y_i - \max(0, X_i'\beta)|.$$

On appelle cet estimateur LAD (pour Least Absolute Deviation). On peut montrer sous certaines conditions (cf. Powell, 1984) que  $\hat{\beta}$  est convergent et asymptotiquement normal. Il peut cependant être difficile à implémenter (la fonction objectif est non différentiable). Commande `clad` sous Stata.

## 2.5 Application

Nombres d'heures travaillées, à partir de l'enquête PCV (Insee).

Code Stata :

```
tobit nb_heures dipl sexe adfe adfe2 exp exp2, ll(0)
```

On peut également spécifier une borne supérieure, par `ul(valeur)`. Par ailleurs, on peut faire varier la censure d'un individu à l'autre avec la procédure `cnreg` (l'indicatrice de censure étant précisée via la commande `censored()`).

Code SAS :

```
proc qlim data=donnees;  
  class dipl sexe;  
  model nb_heures = dipl sexe adfe adfe2 exp exp2;  
  endogenous nb_heures ~ censored(lb=0);  
run;  
quit;
```

Le Système SAS

The QLIM Procedure

Statistiques descriptives des réponses continues

Variable	Moyenne	Erreur std	Type	Borne inférieure	Borne supérieure	Borne inférieure N obs.	Borne supérieure N obs.
nb_heures	19.11582	19.942458	Censored	0		11799	

Informations sur le niveau de classe

Classe	Niveaux	Valeurs
dipl	6	>+2 B+2 BAC BDC BEP CEP
SEXE	2	1 2

Model Fit Summary

Number of Endogenous Variables	1
Endogenous Variable	nb_heures
Number of Observations	24123
Log Likelihood	-69862
Maximum Absolute Gradient	7.41215
Number of Iterations	158
AIC	139748
Schwarz Criterion	139845

Algorithm converged.

Résultats estimés des paramètres

Paramètre		Estimation	Erreur standard	Valeur du test t	Proba. Approx. >  t
<b>Intercept</b>		-21.716457	2.346328	-9.26	<.0001
<b>dipl</b>	<b>&gt;+2</b>	17.121113	1.109880	15.43	<.0001
<b>dipl</b>	<b>B+2</b>	14.092050	1.007891	13.98	<.0001
<b>dipl</b>	<b>BAC</b>	13.763340	0.949818	14.49	<.0001
<b>dipl</b>	<b>BDC</b>	6.856848	1.048811	6.54	<.0001
<b>dipl</b>	<b>BEP</b>	10.761496	0.733852	14.66	<.0001
<b>dipl</b>	<b>CEP</b>	0	.	.	.
<b>SEXE</b>	<b>1</b>	10.691830	0.487308	21.94	<.0001
<b>SEXE</b>	<b>2</b>	0	.	.	.
<b>adfe</b>		0.395544	0.160116	2.47	0.0135
<b>adfe2</b>		0.001153	0.002315	0.50	0.6184
<b>exp</b>		0.457800	0.071656	6.39	<.0001
<b>exp2</b>		-0.006314	0.001569	-4.02	<.0001
<b>_Sigma</b>		34.208645	0.250024	136.82	<.0001

### 3 Modèles de sélection

#### 3.1 Introduction

On s'intéresse à l'influence de  $X$  sur  $Y$  mais on n'observe  $Y$  que si  $D = 1$ . Ceci peut correspondre à plusieurs cas de figure :

- Non-réponse à une enquête. L'échantillon utilisé est composé des seuls répondants ( $D = 1$ ) à l'enquête ;
- Du fait du mode d'échantillonnage, on n'observe  $Y$  que lorsque  $Y > 0$  (troncature).
- Autosélection : on observe le salaire d'un individu que lorsque ce dernier a choisi d'être actif.

**Remarque** : parfois,  $X$  est toujours observé, parfois (non-réponse totale, troncature) il ne l'est que lorsque  $D = 1$ .

### 3.2 Sélection exogène

Ce cas correspond à la situation où  $Y \perp\!\!\!\perp D|X$  : on a par exemple  $D = \mathbf{1}\{X'\gamma + \eta \geq 0\}$  avec  $\eta \perp\!\!\!\perp (X, \varepsilon)$ .

Dans ce cas, on peut ignorer le problème de sélection car la loi de  $Y|X, D = 1$  est identique à celle de  $Y|X$ .

**Exemple (important)** : si  $Y = X'\beta_0 + \varepsilon$  avec  $E(\varepsilon|X) = 0$ , alors :

$$E(Y|X, D = 1) = E(Y|X) = X'\beta_0.$$

Donc l'estimateur des MCO sur les individus tels que  $D = 1$  converge vers  $\beta_0$ .

Plus généralement, les procédures (maximum de vraisemblance, moindres carrés non linéaires...) appliquées à la sélection seule sont convergentes. On peut donc ignorer le problème de sélection (on parle de non-réponse ignorable).

### 3.3 Modèles de troncature (*hors programme*)

Dans ce cas, on observe  $(Y, X)$  seulement si  $Y > 0$  ( $D = \mathbb{1}\{Y > 0\}$ ). Autrement dit, on observe  $(DY, DX)$ .

**Exemple :** on s'intéresse à la consommation d'un bien. Pour cela, on demande à quelques acheteurs de remplir un questionnaire. Problème : on n'échantillonne pas de non-acheteurs !

**Remarque :** ce modèle est proche du modèle de censure mais ici,  $X$  n'est observé que lorsque  $D = 1$ .

Approche paramétrique : supposons que la densité de  $Y$  sachant  $X$  soit paramétrée par  $\theta$  (on la note  $f_{Y|X;\theta}$ ). On estime le modèle à partir de la vraisemblance conditionnelle :

$$L_{1c}(y | x; \theta) = f_{Y|X,D=1;\theta}(y, x) = \frac{f_{Y|X;\theta}(y, x) \mathbb{1}\{y > 0\}}{P(Y > 0 | X = x; \theta)} = \frac{f_{Y|X;\theta}(y, x) \mathbb{1}\{y > 0\}}{\int_0^{+\infty} f_{Y|X;\theta}(y, x) dy}$$

On considère alors la log-vraisemblance conditionnelle :

$$l_{nc}(\theta) = \sum_{i=1}^n D_i \ln L_{1c}(Y_i | X_i; \theta)$$

### 3.3 Modèles de troncature (*hors programme*)

Notons que  $\theta \mapsto l_{nc}(\theta)$  n'est pas une log-vraisemblance même si elle en partage les propriétés. La validité de la procédure est basée sur le fait que

$$\frac{1}{n}l_{nc}(\theta) \xrightarrow{P} l_{\infty}(\theta) = E [D \ln L_{1c}(Y | X; \theta)]$$

et sur le lemme suivant :

**Lemme 2** *Supposons que pour tout  $x \in A$  (avec  $P(A) > 0$ ),*

$$L_{1c}(y | x; \theta) = L_{1c}(y | x; \theta_0) \forall y \implies \theta = \theta_0. \quad (2)$$

*Alors,*

$$\theta_0 = \arg \max_{\theta} E [D \ln L_{1c}(Y | X; \theta)]$$

**Preuve :** on a, pour tout  $\theta \neq \theta_0$ ,

$$E [\ln L_{1c}(Y | X; \theta) | D = 1, X = x] - E [\ln L_{1c}(Y | X; \theta_0) | D = 1, X = x] = \int \ln \frac{L_{1c}(y | x; \theta)}{L_{1c}(y | x; \theta_0)} f_{Y|X, D=1; \theta_0}(y, x) dy$$

Donc d'après l'inégalité de Jensen appliquée à  $u \mapsto \ln(u)$ ,

$$\begin{aligned} E [\ln L_{1c}(Y | X; \theta) | D = 1, X = x] - E [\ln L_{1c}(Y | X; \theta_0) | D = 1, X = x] &\leq \ln \left( \int \frac{L_{1c}(y | x; \theta)}{L_{1c}(y | x; \theta_0)} L_{1c}(y | x; \theta_0) dy \right) \\ &\leq \ln \left( \int L_{1c}(y | x; \theta) dy \right) = 0. \end{aligned}$$

Pour tout  $x \in A$ , et  $\theta \neq \theta_0$ ,  $y \mapsto L_{1c}(y | x; \theta)$  et  $y \mapsto L_{1c}(y | x; \theta_0)$  ne sont pas égales, elles sont donc non colinéaires. L'inégalité précédente est donc stricte. Le lemme est alors une conséquence de l'égalité

$$E [D \ln L_{1c}(Y | X; \theta)] = P(D = 1)E \{E [\ln L_{1c}(Y | X; \theta) | D = 1, X] | D = 1\} \quad \square$$

Supposons maintenant que

$$Y = X'\beta_0 + \sigma_0\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1).$$

On a dans ce cas :

$$\begin{aligned} L_{1c}(y | x; \beta, \sigma) &= \frac{f_{Y|X;\beta}(y, x)}{P(Y > 0 | X = x; \beta)} \mathbf{1}\{y > 0\} \\ &= \frac{\varphi\left(\frac{y-x'\beta}{\sigma}\right)}{\sigma\Phi\left(\frac{x'\beta}{\sigma}\right)} \mathbf{1}\{y > 0\} \end{aligned}$$

(Exercice : montrer que dans ce cas-là, la condition d'identification (2) est satisfaite.)

On maximise alors :

$$l_{nc}(\beta, \sigma) = \sum_{i=1}^n D_i \left[ \ln \varphi\left(\frac{Y_i - X_i'\beta}{\sigma}\right) - \ln \Phi\left(\frac{X_i'\beta}{\sigma}\right) \right] - \left( \sum_{i=1}^n D_i \right) \ln \sigma.$$

On peut estimer ce modèle sous SAS avec la `proc qlim`.

### 3.3 Modèles de troncature (*hors programme*)

Solution semiparamétrique (Powell, 1986) : basée sur l'hypothèse de symétrie de la distribution de  $\varepsilon$  sachant  $X$ .

Problème : la distribution de  $\varepsilon$  sachant  $Y > 0$  n'est pas symétrique. L'idée de Powell est alors de "resymétriser" la distribution des résidus, en coupant les  $\varepsilon$  au-delà de  $X'\beta_0$ .

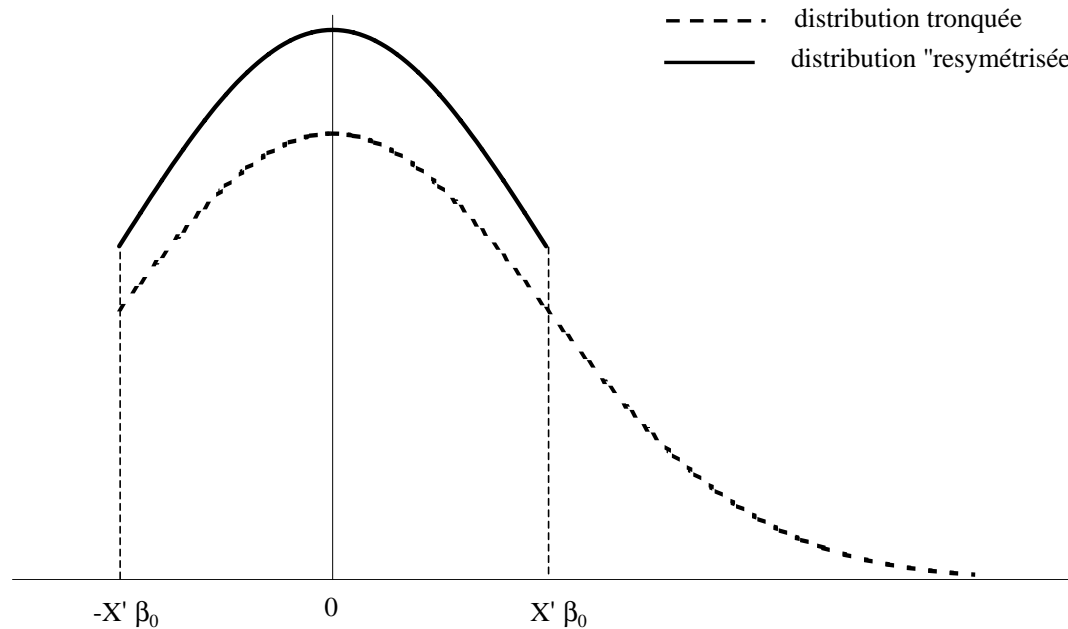


FIG. 2 – Distribution des résidus tronquées et "resymétrisée".

### 3.3 Modèles de troncature (*hors programme*)

On considère donc seulement les données telles que  $\varepsilon < X'\beta_0 \Leftrightarrow Y < 2X'\beta_0$ . On a

1) si  $x'\beta_0 < 0$ ,

$$E[\mathbf{1}\{Y < 2X'\beta_0\}(Y - X'\beta_0)|X = x, Y > 0] = E[0 \times (Y - X'\beta_0)|X = x, Y > 0] = 0$$

2) si  $x'\beta_0 \geq 0$ ,

$$\begin{aligned} E[\mathbf{1}\{Y < 2X'\beta_0\}(Y - X'\beta_0)|X = x, Y > 0] &= P(Y > 0|X = x)E[\mathbf{1}\{0 < Y < 2X'\beta_0\}(Y - X'\beta_0)|X = x] \\ &= P(Y > 0|X = x)E[\mathbf{1}\{-X'\beta_0 < \varepsilon < X'\beta_0\} \varepsilon | X = x] \\ &= P(Y > 0|X = x) \int_{-x'\beta_0}^{x'\beta_0} u f_{\varepsilon|X=x}(u) du \\ &= 0 \end{aligned}$$

Donc

$$E[\mathbf{1}\{Y < 2X'\beta_0\}(Y - X'\beta_0)|X, Y > 0] = 0$$

Et ainsi

$$E[DX\mathbf{1}\{Y < 2X'\beta_0\}(Y - X'\beta_0)] = 0 \tag{3}$$

### 3.3 Modèles de troncature (*hors programme*)

On peut montrer (cf. Powell, 1986) que sous certaines conditions, l'équation (3) permet d'identifier  $\beta_0$ . Elle s'écrit de plus comme la CPO du programme :

$$\min_{\beta} E \left\{ D \left[ Y - \max \left( \frac{Y}{2}, X' \beta \right) \right]^2 \right\}.$$

L'estimateur proposé par Powell est l'estimateur des moindres carrés suivant :

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n D_i \left[ Y_i - \max \left( \frac{Y_i}{2}, X_i' \beta \right) \right]^2.$$

Il est convergent et asymptotiquement normal.

**Remarque 1 :** cette méthode peut également être utilisée (moyennant de légères adaptations) pour estimer des modèles avec censure. L'hypothèse initiale est cependant légèrement plus forte que l'approche par la médiane conditionnelle.

**Remarque 2 :** il n'existe pas de procédures "officielles" sous Stata ou SAS pour estimer ce modèle.

### 3.4 Modèles de sélection généralisée

#### a) *Tobit II*

On considère maintenant le modèle de sélection suivant :

$$\begin{cases} Y = X'_1\beta_0 + \varepsilon \\ D = \mathbf{1}\{X'\gamma_0 + \eta \geq 0\} \end{cases} \quad (4)$$

On suppose que l'on observe  $(D, X)$  mais  $Y$  seulement si  $D = 1$ .  $\varepsilon$  et  $\eta$  sont a priori corrélés (sélection endogène). Un tel modèle est dit modèle tobit généralisé (ou tobit de type II).

**Exemple canonique :** modèle d'offre de travail (Gronau, 1974). On s'intéresse à l'effet de caractéristiques  $X$  sur le salaire horaire offert  $W$ . Le problème est qu'on observe  $W$  que si l'individu a décidé d'être actif. Plus précisément, si l'on considère un choix d'activité hebdomadaire, l'individu résout :

$$\max_h u(Wh + A, h) \quad \text{s. c. } 0 \leq h \leq 168$$

où  $u$  est l'utilité de l'individu (de dérivées partielles  $u_1 > 0$  et  $u_2 < 0$ ),  $h$  le nombre d'heures travaillées et  $A$  correspond aux revenus non salariaux. Si l'on note  $s(h) = u(Wh + A, h)$ , on a

$$s'(h) = Wu_1(Wh + A, h) + u_2(Wh + A, h).$$

### 3.4 Modèles de sélection généralisée

Si  $s'(0) \leq 0$ , alors l'individu choisit 0 heures d'activité. Il travaillera donc si et seulement si

$$W \geq -\frac{u_2(A, 0)}{u_1(A, 0)} = W^r.$$

$W^r$  est appelé le *salaire de réserve*. On observe  $W$  que si  $W \geq W^r$ . Si l'on suppose que

$$\ln W = X_1' \beta_0 + \varepsilon$$

$$\ln W^r = X_2' \beta_1 + \nu$$

où  $(X_1, X_2)$  ont éventuellement des éléments en commun. Alors on a le système :

$$\ln W = X_1' \beta_0 + \varepsilon$$

$$D = \mathbb{1}\{\ln W - \ln W^r \geq 0\} \equiv \mathbb{1}\{X' \gamma_0 + \eta \geq 0\}$$

avec  $D$  l'indicatrice d'activité,  $X$  la réunion de  $X_1$  et  $X_2$  et  $\eta = \varepsilon - \nu$ . En général,  $\varepsilon$  et  $\eta$  seront corrélés.

### 3.4 Modèles de sélection généralisée

Identification et estimation du modèle (4).

On suppose que :

1.  $(\varepsilon, \eta)$  sont indépendants de  $(X_1, X)$  ;
2.  $\eta \sim \mathcal{N}(0, 1)$ .
3.  $E(\varepsilon|\eta) = \delta_0 \eta$ .

Les hypothèses 2 et 3 sont satisfaites lorsque  $(\varepsilon, \eta)$  est gaussien mais sont plus faibles en général.

On a alors :

$$E(Y|X_1, X, \eta) = X_1' \beta_0 + E(\varepsilon|X_1, X, \eta) = X_1' \beta_0 + E(\varepsilon|\eta) = X_1' \beta_0 + \delta_0 \eta.$$

Maintenant,

$$\begin{aligned} E(Y|X_1, X, D = 1) &= E [E(Y|X_1, X, D = 1, \eta)|X_1, X, D = 1] \\ &= E [E(Y|X_1, X, \eta)|X_1, X, \eta \geq -X' \gamma_0] \\ &= E [X_1' \beta_0 + \delta_0 \eta|X_1, X, \eta \geq -X' \gamma_0] \\ &= X_1' \beta_0 + \delta_0 \lambda(X' \gamma_0) \end{aligned}$$

### 3.4 Modèles de sélection généralisée

Par ailleurs,  $\gamma_0$  est identifié puisque la deuxième équation de (4) est un probit. Donc  $\delta_0$  et  $\beta_0$  sont identifiés par la régression de  $Y_i$  sur  $(X_{1i}, \lambda(X_i' \gamma_0))$  (conditionnellement à  $D = 1$ ).

La méthode d'estimation suit la même démarche :

**Procédure (“Heckit”, en référence à Heckman, 1976) :**

1. Estimer le probit de  $D_i$  sur  $X_i \Rightarrow \hat{\gamma}$ .
2. Régresser  $Y_i$  sur  $X_{1i}$  et  $\lambda(X_i' \hat{\gamma}) \Rightarrow \hat{\beta}$  et  $\hat{\delta}$ .

**Remarque 1 :** cette procédure conduit à des estimateurs convergents et asymptotiquement normaux. Cependant, l'erreur commise sur  $\gamma_0$  en première étape a un impact sur la variance asymptotique de  $\hat{\beta}$  et  $\hat{\delta}$  (sauf lorsque  $\delta_0 = 0$ ).

**Remarque 2 :** il n'est pas formellement nécessaire de supposer  $X_1 \neq X$  pour identifier les paramètres du modèle. Cependant, si  $X_1 = X$ , l'identification de  $(\beta_0, \delta_0)$  repose uniquement sur la non-linéarité de la fonction  $\lambda(\cdot)$ . En pratique la procédure sera très peu robuste dans ce cas.

### 3.4 Modèles de sélection généralisée

Alternative : estimation par maximum de vraisemblance. On maintient l'hypothèse 1 et on remplace 2 et 3 par la condition plus forte :

$$\begin{pmatrix} \varepsilon \\ \eta \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho_0 \sigma_0 \\ \rho_0 \sigma_0 & 1 \end{pmatrix} \right).$$

Rappelons qu'on a alors :

$$\eta | \varepsilon \sim \mathcal{N} \left( \frac{\rho_0}{\sigma_0} \varepsilon, (1 - \rho_0^2) \right).$$

Posons  $Y_1 = DY$ . La densité de  $(Y_1, D)$  (par rapport à  $\mu \otimes \nu$  où  $\nu$  est la mesure de comptage) s'écrit :

$$\begin{aligned} f_{Y_1, D | X_1 = x_1, X = x}(0, 0) &= P(D = 0 | X_1 = x_1, X = x) = 1 - \Phi(x' \gamma_0) \\ f_{Y_1, D | X_1 = x_1, X = x}(y, 1) &= f_{Y, D | X_1 = x_1, X = x}(y, 1) \\ &= P(D = 1 | X_1 = x_1, X = x, Y = y) f_{Y | X_1 = x_1, X = x}(y) \\ &= P(-\eta \leq X' \gamma_0 | X_1 = x_1, X = x, \varepsilon = y - x'_1 \beta_0) \frac{1}{\sigma_0} \varphi \left( \frac{y - x'_1 \beta_0}{\sigma_0} \right) \\ &= \Phi \left[ \frac{1}{\sqrt{1 - \rho_0^2}} \left( x' \gamma_0 + \rho_0 \frac{y - x'_1 \beta_0}{\sigma_0} \right) \right] \frac{1}{\sigma_0} \varphi \left( \frac{y - x'_1 \beta_0}{\sigma_0} \right) \end{aligned}$$

### 3.4 Modèles de sélection généralisée

La vraisemblance d'un échantillon s'écrit donc (à une constante près et en notant  $\theta$  l'ensemble des paramètres) :

$$l_n(\theta) = \sum_{i/D_i=0} \ln [1 - \Phi(X'_i\gamma)] + \sum_{i/D_i=1} \left\{ \ln \Phi \left[ \frac{1}{\sqrt{1 - \rho^2}} \left( X'_i\gamma + \rho \frac{Y_i - X'_{1i}\beta}{\sigma} \right) \right] - \frac{1}{2} \left( \frac{Y_i - X'_{1i}\beta}{\sigma} \right)^2 \right\} - N_+ \ln \sigma$$

où  $N_+ = \sum_{i=1}^n D_i$ . En effectuant le changement de variables  $s = 1/\sigma$ ,  $b = \beta/\sigma$  (et  $\tilde{\theta}$  les nouveaux paramètres), on obtient :

$$\tilde{l}_n(\tilde{\theta}) = \sum_{i/D_i=0} \ln [1 - \Phi(X'_i\gamma)] + \sum_{i/D_i=1} \left\{ \ln \Phi \left[ \frac{1}{\sqrt{1 - \rho^2}} (X'_i\gamma + \rho(sY_i - X'_{1i}b)) \right] - \frac{1}{2} (sY_i - X'_{1i}b)^2 \right\} + N_+ \ln s$$

On peut montrer qu'à  $\rho$  fixé, cette vraisemblance est concave en  $(\gamma, s, b)$ . On peut donc facilement maximiser cette vraisemblance à  $\rho$  fixé et calculer la valeur  $M(\rho)$  de  $\tilde{l}_n$  en ce point. Pour obtenir le maximum global, on maximise ensuite  $M(\rho)$  sur  $] - 1, 1[$ .

### 3.4 Modèles de sélection généralisée

Exemple : équation de salaire des femmes en couple.

- Code stata :

```
xi: heckman logsal i.ddipl exp exp2, select(active = enfants exp exp2 i.ddipl) twostep
```

Oter l'option **twostep** pour obtenir une estimation par maximum de vraisemblance.

- Code SAS (EMV) :

```
proc qlim data=emploi;  
  model active = enfants exp exp2 dipl_bac2 dipl_bac dipl_bep dipl_brevet /discrete;  
  model logsal = exp exp2 dipl_bac2 dipl_bac dipl_bep dipl_brevet / select(active=1);  
run;
```

N.B. : il n'y a pas de procédure officielle pour faire une estimation en deux étapes mais il existe des macros en ligne (de David Jaeger notamment).

Probit de première étape :

Variable	Coefficient	Ecart-type
Constante	-0,10	0,08
Nombre d'enfants	-0,26	0,02
Expérience	0,11	0,01
Expérience <sup>2</sup>	-0,003	0,0001
> Bac + 2	0,29	0,07
Bac + 2	0,44	0,07
Bac	0,25	0,06
BEP ou CAP	0,22	0,05
Brevet	0,10	0,08

L'instrument joue significativement sur la probabilité d'être active.

Variable	Tobit II		MCO	
	Coefficient	Ecart-type	Coefficient	Ecart-type
Constante	6,56	0,081	6,20	0,043
Expérience	0,01	0,005	0,03	0,003
Expérience <sup>2</sup>	-0,00003	0,0001	-0,0005	0,0001
> Bac + 2	0,95	0,044	1,02	0,04
Bac + 2	0,67	0,044	0,77	0,037
Bac	0,48	0,039	0,54	0,035
BEP ou CAP	0,23	0,035	0,29	0,032
Brevet	0,3	0,048	0,34	0,045
$\sigma_\varepsilon$	0,62		0,55	
Lambda	-0,39	0,07		

Il y a bien une corrélation entre les deux résidus...Mais les différences avec les MCO sont faibles!

Exercice : calculer  $\text{corr}(\varepsilon, \eta)$ .

### 3.4 Modèles de sélection généralisée

b) *Tobit III (hors programme)*

Cas où l'on observe davantage que  $D$  :

$$\begin{cases} Y_1 = X_1' \beta_0 + \varepsilon \\ Y_2 = \max(0, X' \gamma_0 + \eta) \\ D = \mathbb{1}\{Y_2 > 0\} \end{cases}$$

On observe toujours  $Y_2$  mais  $Y_1$  uniquement si  $D = 1$ .

**Exemple :** modèle de participation au marché du travail avec  $Y_1$  le log du salaire horaire et  $Y_2$  le nombre d'heures travaillées.

Sous les hypothèses 1, 2 et 3 précédentes, on a :

$$E(Y_1 | X_1, X, D = 1, \eta) = E(Y_1 | X_1, X, \eta) = X_1' \beta_0 + \delta_0 \eta$$

Contrairement à précédemment, il est possible d'estimer de façon convergente  $\eta$  lorsque  $D = 1$ . On adopte alors la procédure suivante :

**Procédure :**

1. Estimer le tobit des  $Y_{2i}$  sur  $X_i \Rightarrow \hat{\eta}_i = Y_{2i} - X_i \hat{\gamma}$  (lorsque  $Y_{2i} > 0$ );
2. Régresser les  $Y_{1i}$  sur les  $X_{1i}$  et  $\hat{\eta}_i$  (sur les individus tels que  $D_i = 1$ )  $\Rightarrow \hat{\beta}$  et  $\hat{\delta}$ .

**Remarque :** il n'est pas nécessaire ici de supposer  $X_1 \neq X$  pour obtenir une estimation robuste.