Econométrie 2

Chapitre 4 : censure et sélection.

ENSAE 2021-2022

Michael Visser

CREST-ENSAE

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.

Plan

Modèles de censure ou tobit simple

Modèles de sélection

Présentation

- On considère le modèle suivant :

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

où l'on observe seulement $Y = \max(0, Y^*) = Y^*\mathbb{1}\{Y^* > 0\}$.

- 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. On observe donc $Y = \min(s, Y^*)$ et

  $$\begin{aligned} -Y + s &= -\min(s, Y^*) + s = \max(0, -Y^* + s) \\ &= \max(0, Y^{**}) \text{ avec } Y^{**} = -Y^* + s \end{aligned}$$

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

  2) $Y =$ 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.

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 $\partial E(Y|X, Y > 0)/\partial x_k$.

- On a

$$\frac{\partial E(Y|X = x)}{\partial x_k} = \Phi\left(\frac{x'\beta_0}{\sigma_0}\right)\beta_{0k} \tag{1}$$

et

$$\frac{\partial E(Y|X = x, Y > 0)}{\partial x_k} = \beta_{0k}\left\{1 - \lambda\left(\frac{x'\beta_0}{\sigma_0}\right)\left[\frac{x'\beta_0}{\sigma_0} + \lambda\left(\frac{x'\beta_0}{\sigma_0}\right)\right]\right\} \tag{2}$$

où la fonction $\lambda(u) = \varphi(u)/\Phi(u)$ est appelée inverse du ratio de Mills.

Paramètres d'intérêt du modèle.

- **Preuve :** On a

$$
\begin{aligned}
E(Y|X=x, Y>0) &= x'\beta_0 + \sigma_0 E(\varepsilon|X=x, \varepsilon > -X'\beta_0/\sigma_0) \\
&= x'\beta_0 + \sigma_0 E(\varepsilon|\varepsilon > -x'\beta_0/\sigma_0) \\
&= x'\beta_0 + \sigma_0 \frac{\varphi\left(-x'\beta_0/\sigma_0\right)}{1-\Phi\left(-x'\beta_0/\sigma_0\right)} \\
&= x'\beta_0 + \sigma_0 \frac{\varphi\left(x'\beta_0/\sigma_0\right)}{\Phi\left(x'\beta_0/\sigma_0\right)} = x'\beta_0 + \sigma_0\lambda(x'\beta_0/\sigma_0)
\end{aligned}
\tag{3}
$$

(La 3ème égalité découle de la propriété suivante : si $Z \sim \mathcal{N}(0,1)$, alors $E(Z|Z>c) = \varphi(c)/(1-\Phi(c))$).

On peut montrer que $\lambda'(x) = -\lambda(x)\left[x+\lambda(x)\right]$, et on trouve (2).

Paramètres d'intérêt du modèle.

Par ailleurs, on a $\lambda'(x) \in \ (-1,\ 0)$. Par conséquent, l'effet marginal (2) est compris entre 0 et β_{0k}.

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

$$\begin{aligned} E(Y|X=x) &= P(Y>0|X=x)E(Y|X=x, Y>0) \\ &= P(\varepsilon > -X'\beta_0/\sigma_0|X=x)E(Y|X=x, Y>0) \\ &= \Phi\left(\frac{x'\beta_0}{\sigma_0}\right)x'\beta_0 + \sigma_0\varphi\left(\frac{x'\beta_0}{\sigma_0}\right). \end{aligned} \quad (4)$$

On a donc

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

ce qui correspond à (1). Comme précédemment, l'effet marginal est compris entre 0 et β_{0k}.

Estimation du modèle

- La régression de Y sur X par la méthode des MCO ne conduit pas à un estimateur convergent de β_0 en général. En effet, l'équation (4) montre que l'espérance de Y conditionnellement à $X = x$ est différente de $X'\beta_0$.
- Intuitivement, les MCO vont approcher l'effet marginal moyen de X sur Y, soit $P(Y > 0)\beta_0$. On s'attend alors à un biais d'atténuation vers 0 dans ce cas.
- La régression sur les données non-censurées seules ne conduit pas non plus à un estimateur convergent pour une raison analogue : l'espérance conditionnelle (3) diffère de $X'\beta_0$.
- Intuitivement, on va approcher dans ce cas l'effet marginal moyen de X sur Y sachant $Y > 0$, soit $\beta_0 \left\{1 + E\left[\lambda'\left(X'\beta_0/\sigma_0\right)\right]\right\}$. Là aussi on a un biais d'atténuation vers 0.

Estimation du modèle

- On utilise le maximum de vraisemblance pour estimer les paramètres β_0 et σ_0.
- Y a une distribution continue sur $]0, +\infty[$ mais une masse en 0.
- Notons $g(\cdot|x)$ la densité de Y conditionnellement à $X = x$. On a

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

- Par ailleurs, $P(Y \leq y|X = x) = P(Y^* \leq y|X = x)$ pour tout $y > 0$. Donc $g(y|x) = g^*(y|x)$ pour tout $y > 0$, où $g^*(\cdot|x)$ est la densité de Y^* conditionnellement à $X = x$.
- Comme $Y^*|X = x \sim \mathcal{N}(x'\beta_0, \sigma_0^2)$, on a $g^*(y|x) = \frac{1}{\sigma_0}\varphi\left(\frac{y - x'\beta_0}{\sigma_0}\right)$.
- La fonction de densité s'écrit alors

$$\begin{aligned} g(y|x) &= \mathbb{1}\{y = 0\}g(0|x) + \mathbb{1}\{y > 0\}g(y|x) \\ &= \mathbb{1}\{y = 0\}\Phi\left(\frac{-x'\beta_0}{\sigma_0}\right) + \mathbb{1}\{y > 0\}\frac{1}{\sigma_0}\varphi\left(\frac{y - x'\beta_0}{\sigma_0}\right). \end{aligned}$$

Estimation du modèle

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

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

  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 :

$$\widetilde{\ell_n}(b,s) = \sum_{i|Y_i=0} \ln\Phi(-X_i'b) - \frac{1}{2}\sum_{i|Y_i>0}(sY_i - X_i'b)^2 + N_+\ln s - N_+\ln\sqrt{2\pi}.$$

- Comme d'habitude, l'estimateur du maximum de vraisemblance $(\widehat{\beta},\widehat{\sigma})$ est asymptotiquement normal.

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) :

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

- Pour estimer γ_0, il suffit d'effectuer un probit sur $W = \mathbb{1}\{Y > 0\}$.
- On estime β_0 et σ_0 en régressant les $\ln Y_i$ sur les X_i sur l'échantillon des $Y_i > 0$.

Application

- Déterminants de la consommation de tabac aux Etats-Unis.
- Code Stata :

```
keep if age<100 & educa2_r != 99

gen cig_jours = smokperday
replace cig_jours= . if inlist(smokperday,777,888,999)
replace cig_jours= 0 if inlist(smokperday,666,.)

gen femme = (gender==2) if inlist(gender,1,2)

gen black = (raceethnic_r==2)
gen hisp = (raceethnic_r==5)
gen other_race = inlist(raceethnic_r,3,4,9)

char educa2_r[omit] 2

xi: tobit cig_jours age femme black hispanic other_race ///
        i.educa2_r Tax, 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()`).

Application : résultats

```
Tobit regression                                Number of obs     =    115,165
                                                LR chi2(11)       =    5837.55
                                                Prob > chi2       =     0.0000
Log likelihood = -84086.208                     Pseudo R2         =     0.0335
```

cig_jours	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.3435799	.0096744	-35.51	0.000	-.3625416	-.3246182
femme	-3.684466	.3131422	-11.77	0.000	-4.29822	-3.070712
black	-2.926718	.5809287	-5.04	0.000	-4.065329	-1.788107
hispanic	.8919963	.3195028	2.79	0.005	.2657757	1.518217
other_race	2.950636	.6149396	4.80	0.000	1.745364	4.155908
_Ieduca2_r_1	7.952429	.5520324	14.41	0.000	6.870454	9.034404
_Ieduca2_r_3	-3.21447	.4560256	-7.05	0.000	-4.108273	-2.320667
_Ieduca2_r_4	-4.797549	.4777661	-10.04	0.000	-5.733964	-3.861135
_Ieduca2_r_5	-18.95434	.4966429	-38.16	0.000	-19.92776	-17.98093
_Ieduca2_r_6	-25.26274	.6112839	-41.33	0.000	-26.46085	-24.06464
Tax	-1.248026	.1874222	-6.66	0.000	-1.615371	-.8806815
_cons	-10.8541	.8667528	-12.52	0.000	-12.55292	-9.15528
/sigma	31.09331	.2426334			30.61775	31.56886

```
  102,639  left-censored observations at cig_jours <= 0
   12,526     uncensored observations
        0 right-censored observations
```

Application : résultats

Variable	tobit	MCO
age	-0.037	-0.034
femme	-0.400	-0.656
black	-0.318	-0.77
hispanic	0.097	0.14
other_race	0.321	0.218
educa2_r_1	0.864	1.557
educa2_r_3	-0.349	-0.502
educa2_r_4	-0.521	-0.700
educa2_r_5	-2.06	-2.006
educa2_r_6	-2.745	-2.249
Tax	-0.136	-0.165

Table 1 – Comparaison des effets marginaux estimés par tobit et MCO

Questions :

- Pourquoi ne pas inclure le prix des cigarettes dans le modèle ?
  Réponse : Le prix des cigarettes est le prix d'équilibre (où la demande est égale à l'offre), et dépend donc des déterminants observés et inobservés de la demande et de l'offre. Le terme d'erreur ϵ capte ces déterminants inobservés. Le prix est alors potentiellement une variable endogène.
- A quelle condition peut-on interpréter le coefficient de Tax comme l'effet causal des taxes sur la consommation de cigarettes ?
  Réponse : Si le modèle défini à la page 4 est bien spécifié et si le terme d'erreur suit bien une loi $\mathcal{N}(0, 1)$.

Plan

Modèles de censure ou tobit simple

Modèles de sélection

Introduction

- On s'intéresse ici à des situations où l'on n'observe Y que lorsque $D = 1$ ($D = 0$ sinon).
- On considère le modèle suivant :

$$Y^* = X'\beta_0 + \varepsilon,$$

(contrairement au modèle de censure, la distribution de ε n'est pas spécifiée) et on observe $Y = DY^*$.

- Exemples :
  - non-réponse partielle dans une enquête ;
  - auto-sélection : on n'observe le salaire des individus que s'ils ont décidé de se porter sur le marché du travail.
- N.B. : il existe également des situations où on n'observe (Y, X) que lorsque $D = 1$: non-réponse totale, modèle de troncature. Les méthodes d'estimation, différentes, ne sont pas abordées ici.

Sélection exogène

- Ce cas correspond à la situation où $Y^* \perp\!\!\!\perp D|X$.
- En d'autres termes, D et ε sont indépendants conditionnellement à X. Exemple :

$$D = \mathbb{1}\{X'\gamma_0 + \eta \geq 0\}, \quad \text{avec } \eta \perp\!\!\!\perp (X, Y^*).$$

- Dans ce cas, on peut ignorer le problème de sélection car la loi de $Y^*|X, D = 1$ est identique à celle de $Y^*|X$.
- Si par ailleurs $E(\varepsilon|X) = 0$, alors :

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

- Donc l'estimateur des MCO sur le sous-échantillon $\{i \mid D_i = 1\}$ converge vers β_0.

Modèle de sélection généralisée : présentation

- On ne suppose plus maintenant que $Y^* \perp\!\!\!\perp D|X$. Mais on dispose d'un instrument corrélé à D et qui n'affecte pas directement Y^*.
- Noter l'analogie avec l'approche instrumentale dans le modèle linéaire avec variables endogènes.
- Cas du modèle linéaire (Tobit généralisé ou Tobit II) :

$$\begin{aligned} Y^* &= X'\beta_0 + \varepsilon \\ D &= \mathbb{1}\{Z'\gamma_0 + \eta \geq 0\} \end{aligned}$$

où Z contient au moins une composante qui est exclue de X.

- N.B. : ε et η sont a priori corrélés (sélection endogène)

$\Rightarrow$ l'estimateur des MCO de Y sur X sur $\{i \mid D_i = 1\}$ n'est pas convergent en général.

Modèle de sélection généralisée : exemple

- Exemple canonique : modèle d'offre de travail (Gronau, 1974). On s'intéresse à l'effet de caractéristiques X sur le salaire horaire offert W.
- Mais on n'observe W que si l'individu a décidé d'être actif. 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 travaillés et A correspond aux revenus non salariaux.
- Si l'on note $s(h) = u(Wh + A, h)$, on a

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

Modèle de sélection généralisée : exemple

- 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 n'observe W que si $W \geq W^r$. Si l'on suppose que

$$\begin{array}{rcl} \ln W & = & X'\beta_0 + \varepsilon \\ \ln W^r & = & \widetilde{X}'\beta_1 + \nu \end{array}$$

  où $\widetilde{X}$ a au moins une composante exclue de X, alors :

$$\begin{array}{rcl} \ln W & = & X'\beta_0 + \varepsilon \\ D & = & \mathbb{1}\{\ln W - \ln W^r \geq 0\} \equiv \mathbb{1}\{Z'\gamma_0 + \eta \geq 0\} \end{array}$$

  avec D l'indicatrice d'activité, Z la réunion de X et $\widetilde{X}$ et $\eta = \varepsilon - \nu$.
- En général, ε et η seront corrélés.

Modèle de sélection généralisée : identification

- Supposons que :

$$\begin{cases} Y^* & = & X'\beta_0 + \varepsilon \\ D & = & \mathbb{1}\{Z'\gamma_0 + \eta \geq 0\} \end{cases} \quad (5)$$

  avec :

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

- Les hypothèses 2 et 3 sont satisfaites lorsque (ε, η) est gaussien mais sont plus faibles en général.
- On a alors :

$$\begin{aligned} E(Y^*|X, Z, \eta) &= X'\beta_0 + E(\varepsilon|X, Z, \eta) \\ &= X'\beta_0 + E(\varepsilon|\eta) \\ &= X'\beta_0 + \delta_0\ \eta. \end{aligned}$$

- Par conséquent (toujours avec $Y = DY^*$),

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

- Par ailleurs, γ_0 est identifié puisque la deuxième équation de (5) est un probit.
- Donc β_0 et δ_0 sont identifiés par la régression de Y sur $(X, \lambda(Z'\gamma_0))$ (conditionnellement à $D = 1$).
- N.B. : stricto sensu, on peut identifier β_0 et δ_0 même si $X = Z$. Dans ce cas l'identification repose sur la non-linéarité de la fonction $\lambda(X'\gamma_0)$ en X.

Modèle de sélection généralisée : estimation

- La méthode d'estimation suit la même démarche (méthode d'Heckman en deux étapes ou "Heckit", en référence à Heckman, 1976) :
  1. Estimer le probit de D_i sur Z_i $\Rightarrow \widehat{\gamma}$.
  2. Régresser Y_i sur X_i et $\lambda(Z_i'\widehat{\gamma})$ sur les $\{i|D_i = 1\} \Rightarrow \widehat{\beta}$ et $\widehat{\delta}$.
- Cette procédure conduit à des estimateurs convergents et asymptotiquement normaux.
- N.B. : l'erreur commise sur γ_0 en première étape a un impact sur la variance asymptotique de $\widehat{\beta}$ et $\widehat{\delta}$ (sauf lorsque $\delta_0 = 0$).
- Il est également possible d'estimer le modèle par maximum de vraisemblance en supposant (ε, η) gaussien.
- Mais cet estimateur a l'inconvénient de reposer sur des hypothèses plus fortes que le Heckit.

Modèle de sélection généralisée : application

- Exemple : équation de salaire horaire des femmes en couple (ici à partir de l'enquête emploi 2012).
- Relation d'exclusion : le nombre d'enfants de moins de 6 ans est supposé joué sur la probabilité d'être active mais pas sur le salaire horaire.
- On peut estimer le modèle en utilisant l'option `heckman`.

Modèle de sélection généralisée : application

Code Stata :

```
use "W:\Cours\Econométrie 2\Données\eec_indiv.dta", clear

destring age fordat, replace

* On garde les femmes en couple entre 18 et 64 ans
keep if age>=18 & age<=64 & sexe=="2" & inlist(TYPMEN5,"3","4","5")

* log(salaire horaire)
gen logsal_hor = log(salred*52/(12*nbhp))

* Indicatrice d'observation du salaire(=salarié et répondant)
gen indic_obs = (logsal_hor != .)

* Expérience potentielle et exp. pot. au carré
gen exp = 2012 - fordat
gen exp2 = exp^2

char ddipl[omit] 7

xi: heckman logsal_hor i.ddipl exp exp2, select(indic_obs = NBENF6 exp ///
            exp2 i.ddipl) twostep
```

Modèle de sélection généralisée : application

```
. xi: heckman logsal_hor i.ddipl exp exp2, select(indic_obs = NBENF6 exp ///
>                           exp2 i.ddipl) twostep
i.ddipl           _Iddipl_1-6          (_Iddipl_6 for ddipl==7 omitted)

Heckman selection model -- two-step estimates   Number of obs      =     106878
(regression model with sample selection)        Censored obs       =      89296
                                                   Uncensored obs     =      17582

                                                   Wald chi2( 7)       =    5490.36
                                                   Prob > chi2        =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
logsal_hor   |
   _Iddipl_1 |   .6924857   .0116691    59.34   0.000     .6696146    .7153568
   _Iddipl_2 |   .5026082   .0139363    36.06   0.000     .4752936    .5299229
   _Iddipl_3 |    .305763   .0122888    24.88   0.000     .2816773    .3298487
   _Iddipl_4 |    .149917   .0117749    12.73   0.000     .1268387    .1729953
   _Iddipl_5 |   .1580453    .013888    11.38   0.000     .1308254    .1852653
         exp |   .0276042   .0014949    18.47   0.000     .0246742    .0305342
        exp2 |  -.0004318   .0000382   -11.31   0.000    -.0005067    -.000357
       _cons |   4.276439   .0857463    49.87   0.000     4.108379    4.444499
-------------+----------------------------------------------------------------
indic_obs    |
      NBENF6 |  -.1443794   .0091964   -15.70   0.000     -.162404   -.1263548
         exp |   .0281243    .001419    19.82   0.000     .0253431    .0309055
        exp2 |  -.0008314   .0000293   -28.37   0.000    -.0008888    -.000774
   _Iddipl_1 |   .0607705   .0175224     3.47   0.001     .0264273    .0951138
   _Iddipl_2 |   .2154777   .0171016    12.60   0.000     .1819592    .2489962
   _Iddipl_3 |   .1455301   .0163727     8.89   0.000     .1134402    .1776201
   _Iddipl_4 |   .1406145   .0155556     9.04   0.000     .1101261    .1711029
   _Iddipl_5 |   .1193359   .0203689     5.86   0.000     .0794137    .1592581
       _cons |  -1.111414   .0205948   -53.97   0.000    -1.151779   -1.071049
-------------+----------------------------------------------------------------
mills        |
      lambda |   .2169139   .0503266     4.31   0.000     .1182756    .3155523
-------------+----------------------------------------------------------------
         rho |    0.53462
       sigma |  .40573343
------------------------------------------------------------------------------
```

Modèle de sélection généralisée : application

Questions :

- Quel est l'effet marginal de l'expérience potentielle sur le salaire potentiel ? Réponse : La dérivée du log(salaire horaire) par rapport à l'expérience vaut 0.027-0.0004 * 2 *exp=0.027-0.0008 * exp. Lorsque exp=10, alors une année d'expérience supplémentaire augmente le salaire horaire de 1.9% (0.027-0.008).
- A quoi correspondent `lambda`, `rho` et `sigma` ? Réponse : `lambda`=$\delta_0 \equiv \frac{Cov(\epsilon,\eta)}{V(\eta)} = Cov(\epsilon,\eta)$; `sigma`=$\sqrt{V(\epsilon)}$; `rho`=$\frac{Cov(\epsilon,\eta)}{\sqrt{V(\epsilon)V(\eta)}} = \frac{\delta_0}{\sqrt{V(\epsilon)}}$.
- La sélection est-elle significativement endogène ici ? Dans quel sens joue-t-elle ? Réponse : Oui car on rejette l'hypothèse nulle que $\delta_0 = 0$. Comme $\widehat{\delta} = 0.217$ (signe positif), le salaire horaire *potentiel* des femmes non-actives est inférieur à celui des femmes actives.
- L'instrument a-t-il un effet significatif ? Réponse : Oui, la variable `NBENF6` est statistiquement significative.

L'essentiel

- Modèle tobit I / tobit simple :
  - Cas d'applications ;
  - Paramètres d'intérêt ;
  - Vraisemblance.
- Modèle de sélection généralisée (tobit II) :
  - Cas d'applications ;
  - Hypothèses et relation d'exclusion ;
  - Méthode d'estimation en deux étapes.