Modelos truncados y censurados

Gabriel V. Montes-Rojas

Modelo Tobit

- Ejemplo: Horas trabajadas. El número de horas de trabajo no puede ser negativa, $h \geq 0$. El valor $h = 0$ es verdadero.
- Ejemplo: Cantidad gastada en artículos electrónicos (TV, DVD). Puede ser que algunos meses se gaste 0.
- En ambos casos, el 0 es un valor verdadero. Estos modelos se llaman **modelos Tobit de tipo I**, a partir de un paper de Tobin (1958) sobre gastos en consumos durables.
- Consideremos el siguiente modelo de variable latente:

$$y^* = \boldsymbol{x}\boldsymbol{\beta} + u, \; u|\boldsymbol{x} \sim N(0, \sigma^2)$$

- Sin embargo no se observa y^*, pero

$$y = max\{0, y^*\}$$

  En este caso puede ser que la variale latente no se pueda interpretar, dado que y es interpretable directamente.
- En este caso y está **truncada** en 0, o sea, no puede tomar valores negativos.

Modelo Tobit

- Podríamos entonces plantear el siguiente modelo de verosimilitud

$$\ell_i(\boldsymbol{\beta}, \sigma) = \mathbf{1}[y_i = 0] \log[1 - \Phi(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma)] + \mathbf{1}[y_i > 0] \log[(1/\sigma)\phi\left((y_i - \boldsymbol{x}_i\boldsymbol{\beta})/\sigma\right)]$$

- Notar que a diferencia del modelo probit, acá sí podemos identificar σ. En general para obtener mayor estabilidad en la estimación se usa σ^2 como parámetro a estimar:

$$\ell_i(\boldsymbol{\beta}, \sigma) = \mathbf{1}[y_i = 0] \log[1 - \Phi(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma)] - \mathbf{1}[y_i > 0] \left[\log(\sigma^2)/2 + (y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2/2\sigma^2\right]$$

- Hallar las funciones scores (y probar que la esperanza es cero):

$$\partial\ell_i(\boldsymbol{\beta}, \sigma)/\partial\boldsymbol{\beta} = -\mathbf{1}[y_i = 0]\phi(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma)(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma)/[1 - \Phi(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma)] + \mathbf{1}[y_i > 0](y_i - x_i\boldsymbol{\beta})x_i/\sigma^2$$

$$\partial\ell_i(\boldsymbol{\beta}, \sigma)/\partial\sigma^2 = \mathbf{1}[y_i = 0]\phi(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma)(\boldsymbol{x}_i\boldsymbol{\beta})/\{2\sigma^2[1 - \Phi(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma)]\}$$
$$+\mathbf{1}[y_i > 0]\{(y_i - x_i\boldsymbol{\beta})^2/(2\sigma^4) - 1/(2\sigma^2)\}$$

Álgebra de normales truncadas

Para probar los resultados de la slide anterior necesitamos estas igualdades.

- Tomemos $z \sim N(0,1)$, $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$. Entonces, $\phi'(z) = -z\phi(z)$, $\phi''(z) = -\phi(z) + z^2\phi(z) = \phi(z)(z^2 - 1)$.
- Calculemos entonces

$$\int_A^B z\phi(z)dz = -\int_A^B \phi'(z)dz = \phi(A) - \phi(B)$$

.

- Usando estos resultados con $E[z] = E[1(z > c)z]P[z > c] + E[1(z \leq c)z]P[z \leq c]$ y $lim_{c\to\pm\infty}\phi(c) = 0$, tenemos

$$E[1(z > c)z] = E[1(z > c)z|z > c]P[z > c] = E[z|z > c]P[z > c] = \phi(c)$$

tal que

$$E[z|z > c] = \phi(c)/(1 - \Phi(c))$$

- Finalmente,

$$\int_A^B (z^2 - 1)\phi(z)dz = \int_A^B \phi''(z)dz = \phi'(B) - \phi'(A) = A\phi(A) - B\phi(B)$$

(también $lim_{c\to\pm\infty} c\phi(c) = 0$)

Modelo Tobit

- Notemos que,

$$E(y|\boldsymbol{x}) = P[y > 0|\boldsymbol{x}] \times E[y|y > 0, \boldsymbol{x}] + P[y = 0|\boldsymbol{x}] \times 0$$

$$= P[y > 0|\boldsymbol{x}] \times E[y|y > 0, \boldsymbol{x}]$$

- Notar que $P[y > 0|\boldsymbol{x}] = P[u > -\boldsymbol{x}\beta|\boldsymbol{x}] = P[u/\sigma > -\boldsymbol{x}\beta/\sigma|\boldsymbol{x}] = \Phi(\boldsymbol{x}\beta/\sigma)$ puede ser estimado con un modelo probit.
- Necesitamos ciertas herramientas matemáticas: $E(y|y > 0, \boldsymbol{x}) = \boldsymbol{x}\beta + E(u|u > -\boldsymbol{x}\boldsymbol{\beta}) = \boldsymbol{x}\boldsymbol{\beta} + \sigma\phi(\boldsymbol{x}\boldsymbol{\beta})/\Phi(\boldsymbol{x}\boldsymbol{\beta})$. (usando propiedades de las variables aleatorias normales, $\phi(-c) = \phi(c)$ y $1 - \Phi(-c) = \Phi(c)$)
- Así,

$$E(y|y > 0, \boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} + \sigma\lambda(\boldsymbol{x}\boldsymbol{\beta}/\sigma)$$

  donde $\lambda(.)$ es la **inversa de Mills** (inverse Mills ratio) el ratio de pdf y cdf de una normal.
- Además,

$$E(y|\boldsymbol{x}) = \Phi(\boldsymbol{x}\beta/\sigma)E(y|y > 0, \boldsymbol{x}) = \Phi(\boldsymbol{x}\beta/\sigma)[\boldsymbol{x}\beta + \sigma\lambda(\boldsymbol{x}\beta/\sigma)]$$

Modelo Tobit

Efectos marginales

- Hay dos tipos de efectos marginales de interés para x_j contínua:
- $$\partial E(y|y>0,\boldsymbol{x})/\partial x_j = \beta_j + \beta_j \frac{d\,\lambda}{d\,c}(\boldsymbol{x}\beta/\sigma)$$

  donde $\frac{d\,\lambda}{d\,c} = -\lambda(c)[c+\lambda(c)]$, entonces

  $$= \beta_j \{1 - \lambda(\boldsymbol{x}\beta/\sigma)\,[\boldsymbol{x}\beta + \sigma\lambda(\boldsymbol{x}\beta/\sigma)]\}$$

- $$\partial E(y|\boldsymbol{x})/\partial x_j = \beta_j \Phi(\boldsymbol{x}\beta/\sigma)$$

  (luego de simplificar bastante)
  Notar que si $\Phi()$ está cercano a 1, entonces es improbable observar $y=0$, y por lo tanto el factor de ajuste se vuelve insignificante.
- Para variables x que son dummies lo correcto es calcular la diferencia para $x_j = 1$ y $x_j = 0$.

Modelo Tobit

Inconsistencia de MCO

- Supongamos que usamos los datos para los cuales $y_i > 0$, y que estamos interesados en estimar $E(y|y > 0, \boldsymbol{x})$. De acuerdo a los resultados anteriores tenemos $E(y|y > 0, \boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} + \sigma\lambda(\boldsymbol{x}\boldsymbol{\beta}/\sigma)$ e implícitamente el modelo

$$y_i = \boldsymbol{x}_i\beta + \sigma\lambda(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma) + e_i$$

  donde $E(e_i|y_i > 0, \boldsymbol{x}_i) = 0$.
- Esto implica $E(e_i|y_i > 0, \lambda_i, \boldsymbol{x}_i) = 0$ donde $\lambda_i = \lambda(\boldsymbol{x}_i\boldsymbol{\beta}/\sigma)$.
- Entonces si usamos la regresión de y en x, estaríamos omitiendo la variable λ, con lo cual tendríamos un estimador inconsistente de β por un problema de variables omitidas.

Modelo Tobit

STATA

- `tobit y x1 x2` (estimación de modelo tobit)
- `margins, dydx(*) predict(ystar(0,.))` ($\partial E(y|y>0, \boldsymbol{x}=\bar{x})/\partial x_j$)
- `margins, dydx(*) predict(e(0,.))` ($\partial E(y|\boldsymbol{x}=\bar{x})/\partial x_j$)
- `http://www.stata.com/manuals13/rtobit.pdf`

Modelo Tobit

Ejemplo: Oferta de trabajo de las mujeres casadas

- `http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge17.html`

Modelo de selección muestral

- Supongamos un modelo donde las observaciones están censuradas, y usemos una variable s_i que $=1$ si la observación se observa, $=0$ si no se observa. Entonces nuestra muestra aleatoria es $\{s_i, y_i, \boldsymbol{x}_i\}_{i=1}^{N}$.
- La condición para que la censura no nos afecte los resultados de la regresión es $E(\boldsymbol{x}'u|s=1)=0$.
- Tenemos que evaluar cómo es el mecanismo de selección. Dos modelos: (i) *missing completely at random* (MCAR) cuando s es independiente de $(y, \boldsymbol{x})$, (ii) *missing at random* (MAR) cuando s es independiente de u pero no de $\boldsymbol{x}$ (también llamada selección en observables).

Modelo de selección muestral

Heckman, J. J. (1979) "Sample selection bias as a specification error," Econometrica 47(1), 153-161.

Consideremos un modelo que tiene dos ecuaciones.

1. La **ecuación de resultados** es

$$y_i^* = \boldsymbol{x}_i\beta + u_i,\ E(u_i|\boldsymbol{x}_i) = 0$$

Sin embargo, solo observamos la variable dependiente si ocurre un mecanismo de "selección". Es decir, $y_i = y_i^*$ si $s_i = 1$, pero no observamos y_i^* cuando $s_i = 0$. La variable s_i determina un mecanismo de censura de la variable dependiente.

2. La **ecuación de selección** es un modelo de variable latente binaria para s_i, tal que

$$s_i = 1\ si\ \boldsymbol{z}_i\gamma + v_i > 0,$$

$$s_i = 0\ si\ \boldsymbol{z}_i\gamma + v_i \leq 0.$$

Entonces,

$$y_i = y_i^* \times \mathbf{1}[\boldsymbol{z}_i\gamma + v_i > 0]$$

Modelo de selección muestral

Heckman, J. J. (1979) "Sample selection bias as a specification error," Econometrica 47(1), 153-161.

Bajo ciertas condiciones el estimador de MCO es sesgado. ¿Cuáles?
Asumamos que u y v están correlacionados, i.e. $corr(u, v) = \rho$. También $(u, v) \sim Normal\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u \\ \rho\sigma_u & 1 \end{pmatrix}\right)$.
Así,

$$E(y|y > 0, \boldsymbol{x}) = E(y|y > 0, \boldsymbol{x}, \boldsymbol{z}\gamma + v > 0) = \boldsymbol{x}\beta + E(u|\boldsymbol{z}\gamma + v > 0)$$
$$= \boldsymbol{x}\beta + \rho\sigma_u\lambda(\boldsymbol{z}\gamma),$$

donde $\lambda()$ es la inversa de Mills definida anteriormente.

Modelo de selección muestral

Heckman, J. J. (1979) "Sample selection bias as a specification error," Econometrica 47(1), 153-161.

De esta manera tenemos el siguiente cálculo para los efectos marginales:

$$\partial E(y|y > 0, \boldsymbol{x})/\partial x_j = \beta_j + \rho\sigma_u \partial\lambda(\boldsymbol{z}\gamma)/\partial x_j$$

Esto nos dice que el sesgo de selección se puede ver como un problema de **variables omitidas**, donde $\lambda(\boldsymbol{z}\gamma)$ es la variable omitida.
Esto también sugiere una forma el estimar el modelo en **dos etapas**, el estimador **Heckit**:

1. Modelo probit para estimar γ: $P(y_i > 0|z_i) = \Phi(z_i\gamma)$, usando TODAS las observaciones $i = 1, 2, ..., N$. Construir $\hat{\lambda}_i = \lambda(z_i\hat{\gamma})$.
2. Correr un modelo MCO para las observaciones con y_i observado usando

$$y_i = \boldsymbol{x}_i\beta + \delta\hat{\lambda}_i + error_i$$

Sample selection models

STATA

Dos maneras de estimar el modelo (ver `http://www.stata.com/manuals13/rheckman.pdf`):

1. Máxima verosimilitud: `heckman y x1 x2, select(c= z1 z2)`
2. Heckman's two-step estimator:
   `heckman y x1 x2, select(c= z1 z2) twostep`
3. Heckman's two-step (a mano):

```
probit c z1 z2
predict xb, xb
gen pdf=normalden(xb)
gen cdf=normalprob(xb)
gen invmills=pdf/cdf
reg y x1 x2 if c==1
```