Econometría II

Capítulo 3: Modelos elección discreta y variable dependiente limitada
3.2 Modelos truncados y censurados

Profesor: Graciela Sanroman

Facultad de Ciencias Económicas y Administración

Año 2015

Modelos truncados y censurados

Cuando analizo microdatos también tendré casos en los cuales la variable dependiente tiene un comportamiento en parte "cualitativo", en parte "cuantitativo", por ejemplo: las horas trabajadas, la cantidad invertida en maquinaria por parte de una empresa, los salarios.

Veremos en términos generales dos tipos de modelos de regresión

- truncados,
- censurados.

Nos limitaremos a ver el caso de modelos de regresión truncados y censurados en los cuales el término de error del modelo latente sigue una distribución normal.

Modelos con variable dependiente truncada

Sólo observamos el valor de la variable para un subconjunto de la muestra. Por ejemplo:

- El salario: sólo observamos el salario en aquellos casos en los que la persona está ocupada y eso se da si el salario que obtendría la persona es mayor que su salario de reserva:

  $$y_i = y_i^* \text{ si } y_i^* > \text{salario de reserva de } i$$

  ¿Cuál es sería el salario de los que no trabajan? No es cero obviamente, pero no observamos ningún valor.
- Cualquier variable en una encuesta en la cuál algunos de los encuestados declara no saber o no desear contestar. Por ejemplo: el salario de una persona que trabaja pero que no responde a la pregunta sobre el importe de su salario.

Modelos con variable dependiente censurada

La variable dependiente se observa con censura, por ejemplo:

- inversión en maquinarias (censurada inferiormente en cero)
- horas de trabajo (censurada inferiormente en cero)

$$y_i = \begin{cases} y_i^* & y_i^* > 0 \\ c & y_i^* \leq 0 \end{cases} \quad \text{o } y_i = \max(0, y_i^*)$$

- ingreso del hogar en algunas encuestas (censurada superiormente, ya que si el ingreso es mayor que un determinado valor sólo se registra eso). Por ejemplo

$$y_i = \begin{cases} y_i^* & y_i^* < c \\ c & y_i^* \geqslant c \end{cases} \quad \text{o } y_i = \min(c, y_i^*)$$

Modelos con variable dependiente censurada

- Es necesario distinguir:
  - censura dada por soluciones de esquina en el problema de decisión ecónomica del agente (inversión, horas de trabajo, etc.)
  - censura por las características de los datos (ingreso en encuestas "top coded")
  - truncamiento dado por soluciones de esquina en el problema de decisión ecónomica del agente (salario)
  - truncamiento dado por las características de los datos (no respuesta en las encuestas)

Truncamiento simple: densidad y media condicionales

Sea ε una v.a. con fda $\Pr(\varepsilon \leq c) = F(c)$ y función de densidad $f(\varepsilon)$. Supongamos que dicha distribución está sujeta a truncamiento inferior en el punto c,
si tenemos una variable aleatoria continua con una función de densidad $f(z)$, la función de densidad de la variable truncada a partir del valor c es la función de densidad condicional:

$$f(\varepsilon \mid \varepsilon > c) = \frac{f(\varepsilon)}{P[\varepsilon > c]} = \frac{f(\varepsilon)}{1 - F(c)}$$

y la media condicional de una distribución truncada es

$$\mathbb{E}(\varepsilon \quad \mid \quad \varepsilon > c) = \int_c^{+\infty} \varepsilon \frac{f(\varepsilon)}{1 - F(c)} d\varepsilon$$

mayor que $\mathbb{E}(\varepsilon)$ y mayor que c

Truncamiento simple inferior: densidad condicional en la Normal

En el caso de que $\varepsilon \sim N(\mu, \sigma^2)$ entonces, suponiendo que hay un truncamiento inferior tenemos:

$$f(\varepsilon \mid \varepsilon > c) = \frac{f\left(\frac{\varepsilon-\mu}{\sigma}\right)}{1-\Phi\left(\frac{c-\mu}{\sigma}\right)} = \frac{\left(\frac{1}{\sigma}\right)\phi\left(\frac{\varepsilon-\mu}{\sigma}\right)}{1-\Phi\left(\frac{c-\mu}{\sigma}\right)}$$

denotando $\varepsilon^* = \frac{\varepsilon-\mu}{\sigma}$ y $c^* = \frac{c-\mu}{\sigma}$

$$f(\varepsilon \mid \varepsilon > c) = \frac{1}{\sigma}\frac{\phi(\varepsilon^*)}{1-\Phi(c^*)}$$

Truncamiento simple inferior: media condicional en la Normal

En el caso de que $\varepsilon \sim N(\mu, \sigma^2)$ defino ε^* tal que $\varepsilon = \mu + \sigma\varepsilon^*$ y $\varepsilon^* \sim N(0, 1)$:

$$\begin{aligned}
\mathbb{E}(\varepsilon \quad | \quad \varepsilon > c) &= \int_c^{+\infty} \varepsilon \; f(\varepsilon \mid \varepsilon > c) d\varepsilon \\
&= \int_{c^*}^{+\infty} (\mu + \sigma\varepsilon^*) \frac{1}{\sigma} \frac{\phi(\varepsilon^*)}{1 - \Phi(c^*)} \sigma d\varepsilon^* \\
&= \mu + \sigma \int_{c^*}^{+\infty} \varepsilon^* \frac{\phi(\varepsilon^*)}{1 - \Phi(c^*)} d\varepsilon^*
\end{aligned}$$

se puede demostrar que

$$\mathbb{E}(\varepsilon \mid \varepsilon > c) = \mu + \sigma \frac{\phi(c^*)}{1 - \Phi(c^*)} = \mu + \sigma\lambda(c^*)$$

$\lambda(c^*) = \frac{\phi(c^*)}{1 - \Phi(c^*)}$ se conoce como la "inversa del ratio de Mills"

Truncamiento simple superior: densidad y media condicional en la Normal

La densidad con truncamiento superior si la distribución es normal es

$$f(\varepsilon \mid \varepsilon < c) = \frac{1}{\sigma}\frac{\phi(\varepsilon^*)}{\Phi(c^*)}$$

mientras que la media

$$\begin{aligned}\mathbb{E}(\varepsilon \quad &| \quad \varepsilon < c) = \mu - \sigma\frac{\phi(c^*)}{\Phi(c^*)} \\ &= \quad \mu - \sigma\lambda^*(c^*)\end{aligned}$$

$\lambda^*(c^*) = \frac{\phi(c^*)}{\Phi(c^*)}$ se conoce como el complementario de la "inversa del ratio de Mills"

Modelos de Regresión Truncada

Los modelos de regresión truncada se concentran en explicar el valor esperado de una variable endógena y truncada superior o inferiormente condicional a los valores de las variables explicativas $\mathbf{x}$.
El modelo de regresión truncada refleja un modelo poblacional que cumple con los supuestos del modeo lineal clásico:

$$y_i^* = x_i^{'}\beta + \varepsilon_i \qquad \varepsilon_i\,|x_i \sim N(0, \sigma^2)$$

si observaramos y_i^* el procedimiento MCO produce los estimadores lineales e insesgados de mínima varianza.
El problema es que sólo se observarán valores de y en algunas situaciones:

- en el caso de soluciones de esquina cuando sobrepasa un cierto umbral mínimo c, por lo que para estimar el vector β y σ necesitamos conocer la distribución de y dado que $y_i^* > c$, en definitiva una distribución de probabilidad truncada $y_i = y_i^*$ si $y_i^* > c$
- en el caso de no respuesta no hay un umbral particular

Modelos de Regresión Truncada

En adelante nos concentraremos en el caso de regresión truncada inferiormente debido a soluciones de esquina. Consideraremos además el caso particular de $c = 0$ entonces

$$\mathbb{E}(y_i \,|x_i, y_i^* > 0\,) = x_i^{'}\beta + \sigma\lambda_i$$

con

$$\lambda_i = \lambda_i\left(-x_i^{'}\frac{\beta}{\sigma}\right) = \frac{\phi\left(x_i^{'}\frac{\beta}{\sigma}\right)}{\Phi\left(x_i^{'}\frac{\beta}{\sigma}\right)}$$

Modelos de Regresión Truncada

Consideremos el caso $c = 0$. Sea

$$d_i = 1(y_i^* > 0)$$

Cuando $d_i = 1$ la contribución del individuo i a verosimilitud (bajo el supuesto de normalidad de las perturbaciones de la ecuación estructural) estará dada por la función de densidad de una normal truncada:

$$f\left(y_i \mid x_i, y_i > 0\right) = \frac{\left(\frac{1}{\sigma}\right) \phi\left[\frac{y_i - x_i^{'}\beta}{\sigma}\right]}{1 - \Phi\left[\frac{-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right]}$$

Cuando $d_i = 0$ la información del individuo i no realizará ninguna contribución a la verosimilitud

Notar: sólo la información de los individuos con $y_i^* > 0$ es utilizada en el modelo de regresión truncada

Modelos de Regresión Truncada

Por lo que la función de verosimilitud de un modelo de regresión truncada, será:

$$\begin{aligned}\mathcal{L}(\beta,\sigma) &= \prod_{i=1}^{N}\left[\frac{\left(\frac{1}{\sigma}\right)\phi\left[\frac{y_i-\mathbf{x}\prime_i\boldsymbol{\beta}}{\sigma}\right]}{1-\Phi\left[\frac{-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right]}\right]^{d_i} \\ &= \prod_{i=1}^{N}\left[\frac{\left(\frac{1}{\sigma}\right)\phi\left[\frac{y_i-\mathbf{x}\prime_i\boldsymbol{\beta}}{\sigma}\right]}{\Phi\left[\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right]}\right]^{d_i}\end{aligned}$$

y la de log-verosimilitud será (eliminado los términos constantes):

$$L(\beta,\sigma) = -\sum_{i=1}^{N} d_i\left\{\frac{1}{2}\ln(\sigma^2)+\frac{1}{2\sigma^2}(y_i-x_i^{/}\beta)^2+\ln\left[\Phi\left(\frac{x_i^{/}\beta}{\sigma}\right)\right]\right\}$$

a partir de esta función se obtienen los estimadores máximo verosímiles para β y σ; el estimador máximo verosímil de la matriz de covarianzas puede obtenerse de la inversa de la matriz hessiana.

Notar que hay identificación separada de β y σ.

Efectos parciales en el modelo de regresión truncada

Tenemos

$$\frac{\partial \mathbb{E}\left(y_i^* \,|x_i\right)}{\partial x_j} = \beta_j$$

pero aquí estimo un modelo de esperanza condicional $\mathbb{E}\left(y_i^* \,|x_i, y_i^* > 0\right)$, Tenemos

$$\mathbb{E}(y_i \,|x_i, y_i^* > 0) = x_i^{'}\beta + \sigma\lambda_i$$

con $\lambda_i = \lambda_i\left(-x_i^{'}\frac{\beta}{\sigma}\right) = \frac{\phi\left(x_i^{'}\frac{\beta}{\sigma}\right)}{\Phi\left(x_i^{'}\frac{\beta}{\sigma}\right)}$, entonces

$$\begin{aligned}\frac{\partial \mathbb{E}(y_i \,|x_i, y_i^* > 0)}{\partial x_j} &= \beta_j + \sigma\frac{\partial \lambda_i}{\partial x_j} \\ &= \beta_j\left[1 - \left(x_i^{'}\frac{\beta}{\sigma} + \lambda_i\right)\lambda_i\right]\end{aligned}$$

Distribuciones censuradas (censura inferior)

Existen casos en los cuales se dispone de observaciones de las variables aleatorias en el punto límite pero no a la izquierda de éste, lo que se define como censura.

La variable censurada inferiormente se define como:

$$y_i = \begin{cases} y_i^* & y_i^* > c \\ c & y_i^* \leq c \end{cases} \quad \text{o } y_i = \max(c, y_i^*)$$

La distribución de una variable censurada es una mezcla de una distribución discreta (con punto de acumulación en c) y una distribución de densidad continua.

Recordar que la censura puede estar asociada a

- soluciones de esquina
- características de los datos.

Distribuciones censuradas (censura inferior)

Consideremos el caso de una v.a. ε^* con fda $\Pr(\varepsilon^* \leq c) = F(c)$ y función de densidad $f(\varepsilon^*)$ con $-\infty < \varepsilon^* < \infty$, se cumple

$$\varepsilon = \begin{cases} \varepsilon^* & \varepsilon^* > c \\ c & \varepsilon^* \leq c \end{cases} \qquad \varepsilon = \max(c, \varepsilon^*)$$

$$\begin{array}{ll} \varepsilon^* > c & f(\varepsilon^*) \;\; \text{con } c < \varepsilon^* < \infty \\ \varepsilon^* \leq c & \qquad \Pr(\varepsilon = c) = F(c) \end{array}$$

la media de ε

$$\mathbb{E}(\varepsilon) = cF(c) + \mathbb{E}(\varepsilon \mid \varepsilon > c)\left[1 - F(c)\right]$$

Si $\varepsilon^* \sim N\left(\mu, \sigma^2\right)$

$$\mathbb{E}(\varepsilon) = \mu + \sigma\left\{c^*\Phi(c^*) + \phi(c^*)\right\}$$

con $c^* = \frac{c-\mu}{\sigma}$

Distribuciones censuradas: modelo Tobit

Cierto tipo de modelos censurados son denominados modelos Tobit, en honor al economista James Tobin que estudió la demanda de bienes durables en un artículo de 1959. El modelo Tobit puede ser más sencillamente presentado como un modelo de variable latente:

$$y_i^* = x_i^{'}\beta + u_i, \;\; u_i \sim N(0, \sigma^2)$$

$$y_i = \max(0, y_i^*) = \max(0, x_i^{'}\beta + u_i)$$

La variable latente cumple con las suposiciones del modelo lineal clásico. En el caso de censura a la izquierda la variable observada y es y^*cuando $y^* \geqslant 0$, y y es 0 cuando $y^* < 0$.

Distribuciones censuradas: modelo Tobit

La distribución de la variable y es mixta, ya que es discreta en 0 (cuando $y^* < 0$):

$$\begin{aligned} P(y_i &= 0 \mid x_i^{'}\beta) = P(y_i^* \leqslant 0 \mid x_i^{'}\beta) = P(u_i \leqslant -x_i^{'}\beta) \\ &= P(\frac{u_i}{\sigma} \leqslant \frac{-x_i^{'}\beta}{\sigma}) = \Phi(\frac{-x_i^{'}\beta}{\sigma}) \end{aligned}$$

y continua en los demás valores. Supondremos en adelante que $c = 0$.

Distribuciones censuradas: modelo Tobit

Sea

$$d_i = 1(y_i^* > 0)$$

Cuando $d_i = 1$ la contribución del individuo a verosimilitud estará dada por la función de densidad de una normal:

$$f(y_i \mid x_i) = \left(\frac{1}{\sigma}\right) \phi \left[\frac{y - x_i^{'}\beta}{\sigma}\right]$$

Cuando $d_i = 0$ la contribución del individuo a la verosimilitud estará dada por

$$\begin{aligned} \Pr(d_i &= 0 \mid x_i) = \Pr(y_i^* \leq 0 \mid x_i) \\ &= \Pr(u_i \leqslant -x_i^{'}\beta \mid x_i) \\ &= 1 - \Phi(\frac{x_i^{'}\beta}{\sigma}) \end{aligned}$$

Distribuciones censuradas: modelo Tobit

De esta forma la función de verosimilitud será:

$$\mathcal{L}(\beta, \sigma) = \prod_{i=1}^{N} \left[1 - \Phi\left(\frac{x_i' \beta}{\sigma} \right) \right]^{1-d_i} \left[\left(\frac{1}{\sigma} \right) \phi\left(\frac{y - x_i' \beta}{\sigma} \right) \right]^{d_i}$$

y la log-verosimilitud

$$L(\beta, \sigma) = \sum_{i=1}^{N} \left\{ (1 - d_i) \log \left[1 - \Phi\left(\frac{x_i' \beta}{\sigma} \right) \right] + d_i \left[\frac{1}{\sigma} \phi\left(\frac{y_i - x_i' \beta}{\sigma} \right) \right] \right\}$$

A partir de la maximización de la ecuación anterior se obtienen los estimadores β y σ; el estimador máximo verosímil de la matriz de covarianzas puede obtenerse de la inversa de la matriz de información.

Distribuciones censuradas: modelo Tobit

Si multiplicamos la ecuación de la verosimilitud antes definida por la siguiente expresión:

$$\prod_{i=1}^{N}\left[\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]^{d_i}\prod_{i=1}^{N}\frac{1}{\left[\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]^{d_i}}$$

reacomodando se llega a que:

$$\mathcal{L}(\beta,\sigma)=\prod_{i=1}^{N}\left\{\frac{\left(\frac{1}{\sigma}\right)\phi\left[\frac{y-x_i'\beta}{\sigma}\right]}{\Phi\left[\frac{x_i'\beta}{\sigma}\right]}\right\}^{d_i}\prod_{i=1}^{N}\left[1-\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]^{1-d_i}\left[\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]^{d_i}$$

donde la primer productoria es un modelo truncado y la segunda corresponde a un probit que modela si la observaciòn es censurada o no. Esto sugiere que un modelo tobit es una combinación de un modelo probit, que determina las observaciones que son censuradas y las que no, y un modelo truncado para las observaciones no censuradas.

Modelo Tobit

En el modelo Tobit

$$y_i^* = x_i^{'}\beta + u_i, \;\; u_i \sim N(0, \sigma^2)$$

$$y_i = \max(0, y_i^*) = \max(0, x_i^{'}\beta + u_i)$$

Una expresión de importancia es

$$E(y_i \mid y_i^* > 0, x_i) = x_i^{'}\beta + \sigma \frac{\phi(\frac{x_i^{'}\beta}{\sigma})}{\Phi(\frac{x_i^{'}\beta}{\sigma})} = x_i^{'}\beta + \sigma\lambda_i$$

donde nuevamente aparece la razón $\lambda_i = \lambda(\frac{x_i^{'}\beta}{\sigma})$ la inversa del ratio de Mills.

Efectos parciales en el modelo Tobit

Efectos parciales en el modelo Tobit

Pero otros parámetros de importancia podrían ser:

- El efecto parcial sobre la variable latente, variable que en ocasiones tiene un sentido económico y en otros no:

$$\frac{\partial E(y_i^* \mid x_i)}{\partial x_j} = \beta_j$$

- Efecto parcial en la esperanza "condicional" (condicional en que la observación no está censurada)

$$E(y_i \mid y_i^* > 0, x_i) = x_i^{'}\beta + \sigma\lambda_i$$

respecto a $\mathbf{x}_j$, .se obtiene que

$$\frac{E(y_i \mid y_i^* > 0, x_i)}{\partial \mathbf{x}_j} = \beta_j \left\{ 1 - \lambda_i \times \left(\frac{x_i^{'}\beta}{\sigma} + \lambda_i \right) \right\}$$

Efectos parciales en el modelo Tobit

- El efecto parcial en la esperanza "no condicional" (toda la población):

$$\begin{aligned}
E(y_i \mid x_i) &= 0 \times \Pr(y_i = 0 \mid x_i) + E(y_i \mid y_i^* > 0, x_i) \times \Pr(y_i > 0 \mid x_i) \\
&= E(y_i \mid y_i^* > 0, x_i) \times \Pr(y_i > 0 \mid x_i) \\
&= E(y_i \mid x_i) = \Phi(\frac{x_i^{'}\beta}{\sigma})\left(x_i^{'}\beta + \sigma\lambda_i\right) = m(x_i^{'}\beta, \sigma^2)
\end{aligned}$$

derivando y desarrollando la expresión anterior con respecto a $x_j$se llega a:

$$\frac{\partial E(y_i \mid x_i)}{\partial x_j} = \beta_j \Phi\left(\frac{x_i^{'}\beta}{\sigma}\right)$$

Elasticidades en el modelo Tobit

Si estamos interesados en elasticidades. Consideremos la esperanza "no condicional" en el modelo Tobit (punto de censura es 0)

$$\begin{aligned} \mathbb{E}\left(y_i \,|x_i\right) &= \mathbb{E}\left(y_i \,|x_i, y_i^* > 0\right) \Pr\left(y_i^* > 0\right) \\ \ln \mathbb{E}\left(y_i \,|x_i\right) &= \ln \mathbb{E}\left(y_i \,|x_i, y_i^* > 0\right) + \ln \Pr\left(y_i^* > 0\right) \end{aligned}$$

la elasticidad puede escribirse como

$$x_{ji} \frac{\partial \ln \mathbb{E}\left(y_i \,|x_i\right)}{\partial x_j} = x_{ji} \frac{\partial \ln \mathbb{E}\left(y_i \,|x_i, y_i^* > 0\right)}{\partial x_j} + x_{ji} \frac{\partial \ln \Pr\left(y_i^* > 0\right)}{\partial x_j}$$

a estos términos se le suele denominar "margen intensivo" y "margen extensivo", para indicar que el efecto de un cambio en x_j afecta:

- la esperanza condicional
- la probabilidad de que la observación pertenezca al intervalo donde es observable

Nota: $\Pr(y_i^* > 0 \,|x_i) = \Phi\left(x_i' \frac{\beta}{\sigma}\right)$ y $\frac{\partial \Pr(y_i^* > 0 | x_i)}{\partial x_j} = \frac{\beta_j}{\sigma} \phi\left(x_i' \frac{\beta}{\sigma}\right)$ equivale a un modelo Probit

Modelos de Regresión Censurada

Los modelos de regresión censurada son extensiones del modelo de Tobit, las generalizaciones pueden ser diferentes:

- censura superior
- censura inferior y superior
- censuras que dependen de características del individuo
- modelos censurados en donde la censura o el truncamiento se da respecto de otra variable

(ver Amemiya (1985), Wooldridge (2002))

Truncamiento respecto a otra variable
La densidad con truncamiento respecto a otra variable se puede escribir

$$
\begin{aligned}
f(\varepsilon_1, \varepsilon_2 \quad | \quad \varepsilon_2 > c_2) &= \frac{f(\varepsilon_1, \varepsilon_2)}{\Pr(\varepsilon_2 > c_2)} \qquad \begin{array}{c} -\infty < \varepsilon_1 < +\infty \\ c_2 \leq \varepsilon_2 < +\infty \end{array} \\
\Pr(\varepsilon_2 > c_2) &= \int_{c_2}^{\infty} \int_{-\infty}^{\infty} f(\varepsilon_1, \varepsilon_2) d\varepsilon_1 d\varepsilon_2
\end{aligned}
$$

entonces

$$
f(\varepsilon_1 \mid \varepsilon_2 > c_2) = f(\varepsilon_1) \frac{\Pr(\varepsilon_2 > c_2 \,|\varepsilon_1)}{\Pr(\varepsilon_2 > c_2)}
$$

Truncamiento respecto a otra variable: la normal bivariada

Supongamos

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ & \sigma_2^2 \end{bmatrix} \right)$$

La media

$$\mathbb{E}(\varepsilon_1 \mid \varepsilon_2 > c_2) = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2} \lambda(\frac{c_2 - \mu_2}{\sigma_2})$$

El modelo generalizado de selección (Heckman, 1979)

El modelo Tobit no es apropiado cuando el proceso que genera que algunos individuos estén en el punto de censura no es aleatorio, por ejemplo cuando responde a situaciones en las cuales el individuo está restringido respecto a las decisiones a tomar (por ejemplo desempleo involuntario). Heckman propone descomponer el modelo censurado en dos procesos, de manera que tendremos un modelo bivariante con dos ecuaciones:

$$\begin{aligned} y_{1i}^* &= x_i^{'}\beta + u_{1i} \\ y_{2i}^* &= z_i^{'}\gamma + u_{2i} \end{aligned}$$

se observa $\{y_i, d_i, x_i, z_i\}$:

$$\begin{aligned} d_i &= \mathbf{1}\left(y_{2i}^* > c_2\right) \\ y_i &= y_{1i}^* \text{ si } d_i = 1 \end{aligned}$$

Si u_{1i} y u_{2i} están correlacionados estaremos ante un caso de "selección endógena" de la muestra, debido a la selección basada en y_{2i}^*.

El modelo generalizado de selección (Heckman, 1979)
Podemos escribir:

- la versión censurada del modelo de selección muestral

$$y_i = d_i y_{1i}^* + (1 - d_i)c_1 = \begin{cases} c_1 \text{ si } d_i = 0 \\ y_{1i}^* \text{ si } d_i = 1 \end{cases}$$

- la versión truncada del modelo de selección muestral:

$$y_i = y_{1i}^* \text{ si } d_i = 1$$

El modelo generalizado de selección (Heckman, 1979)

Consideremos el caso del modelo censurado con $c_2 = 0$,

$$
\begin{aligned}
y_{1i}^* &= x_i^{'}\beta + u_{1i} \\
y_{2i}^* &= z_i^{'}\gamma + u_{2i}
\end{aligned}
$$

se observa $\{y_i, d_i, x_i, z_i\}$:

$$
\begin{aligned}
d_i &= \mathbf{1}\left(y_{2i}^* > 0\right) \\
y_i &= y_{1i}^* \text{ si } d_i = 1
\end{aligned}
$$

se supone

$$
\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Omega = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ & 1 \end{bmatrix} \right)
$$

El modelo generalizado de selección (Heckman, 1979)

La verosimilitud del modelo la podemos escribir:

$$\mathcal{L}(\beta,\gamma,\Omega)=\prod_{i=1}^{N}\left[\Pr\left(d_i=0\mid z_i\right)\right]^{1-d_i}\left[\Pr\left(d_i=1\mid z_i\right)f\left(y_{1i}^{*}\mid x_i,d_i=1\right)\right]^{d_i}$$

tomando logaritmos y desarrollando obtenemos:

$$\begin{aligned}L(\beta,\gamma,\Omega) &= \sum_{i=1}^{N}\left\{(1-d_i)\ln\left[1-\Phi\left(z_i'\gamma\right)\right]\right.\\ &\quad +d_i\left[-\frac{1}{2}\ln\sigma_1^2+\ln\phi\left(\frac{y_i-x_i'\beta}{\sigma_1}\right)\right]\\ &\quad +d_i\ln\left[\Phi\left(\frac{z_i'\gamma+\sigma_{12}\left(y_i-x_i'\beta\right)}{\sqrt{1-\sigma_{12}}}\right)\right]\end{aligned}$$

y el modelo se estima por máxima verosimilitud.

El modelo generalizado de selección (Heckman, 1979)

Retomemos las ecuaciones:

$$
\begin{aligned}
y_{1i}^{*} &= x_i^{'}\beta + u_{1i} \\
y_{2i}^{*} &= z_i^{'}\gamma + u_{2i} \\
y_i &= x_i^{'}\beta + u_{1i} \text{ si } y_{2i}^{*} > 0
\end{aligned}
$$

el modelo de esperanza condicional para las observaciones no censuradas queda:

$$E(y_{1i}^{*} \mid y_{2i}^{*} > 0) = x_i^{'}\beta + \sigma_{12}\lambda\left(z_i^{'}\gamma\right)$$

donde $\lambda_i = \lambda\left(z_i^{'}\gamma\right)$ es la inversa del ratio de Mills, también denominada en el contexto de este modelo "lambda de Heckman".

El modelo generalizado de selección (Heckman, 1979)
Podemos re-escribir el modelo en forma de ecuación de error como

$$y_i = x_i^{'}\beta + \sigma_{12}\lambda_i + v_i$$

Notar que:

- el segundo término desaparece cuando $\sigma_{12} = 0$
- el término $\lambda_i = \sigma_{12}\lambda\left(z_i^{'}\gamma\right)$ se conoce en la literatura como "Sesgo de selección",
- se podría estimar en forma consistente β y σ_{12} si λ_i fuera observable.

Procedimiento en dos etapas de Heckman

Heckman también propone estimar el modelo en un procedimiento bietápico

- ETAPA 1: Estimar γ en el modelo

$$\begin{aligned} y_{2i}^{*} &= z_i^{'}\gamma + u_{2i} \\ d_i &= \mathbf{1}\left(y_{2i}^{*} > 0\right) \end{aligned}$$

  mediante un PROBIT de d_i sobre z_i y calcular

$$\widehat{\lambda}_i = \lambda\left(z_i^{'}\widehat{\gamma}\right) = \frac{\phi\left(z_i^{'}\widehat{\gamma}\right)}{\Phi\left(z_i^{'}\widehat{\gamma}\right)}$$

- ETAPA 2: Estimar β y σ en

$$y_i = x_i^{'}\beta + \sigma_{12}\widehat{\lambda}_i + v_i$$

  a través de una regresión MCO de y_i sobre x_i y $\widehat{\lambda}_i$ usando UNICAMENTE las observaciones para las cuales $d_i = 1$.

Procedimiento en dos etapas de Heckman
Comentarios:

- Se puede utilizar en situaciones en las que la variable dependiente es truncada o censurada
- IDENTIFICACION: Restricción de exclusión z y x pueden compartir variables, pero z debe contener al menos una variable que sea determinante del proceso de selección $\left(y_2^*\right)$ pero no de y_1^*.
- Si no se satisface la restricción de exclusión el modelo está identificado por "forma funcional"
- Contraste de existencia de sesgo de selección $Ho : \sigma_{12} = 0$
- Comparar coeficientes directamente con una MCO de y_i sobre x_i y $\widehat{\lambda}_i$ usando UNICAMENTE las observaciones para las cuales $d_i = 1$.
- TERMINO DE ERROR $v_i = u_{1i} + \sigma_{12}(\lambda_i - \widehat{\lambda}_i)$: cálculo de los errores estándar caso heterocedástico pero no se corrige con White.

Referencias

*** Wooldridge, J. M. (2001) Introducción a la Econometría: un enfoque moderno, Thomson Learning, México. (2a. Edición en español, 2006). Capítulo 17.

** Wooldridge, J. (2002) Econometric Analysis of Cross Section and Panel Data, MIT Press.

1a. Edición. Capítulo 15 (15.1 a 15.6 15.9 y 15.10), Capítulo 16 (16.1 a 16.4), Capítulo 17 (17.1 a 17.3 y 17.4.1).

2a. Edición: Capítulo 15 (15.1 a 15.6), Capítulo 16, Capítulo 17 (17.1 a 17.4), Capítulo 19 (19.1 a 19.5 y 19.6.1).

** Mroz, T. (1987) "The Sensitivity of an Empirical Models of Married Women's of Work to Economic and Statistical Assumptions", Econometrica, 55(4), 765-799.

* Cameron A. C. y P.K. Trivedi (2009) Microeconometrics Using Stata, Stata Press. Capítulos 14, 15 y 16.