Programa de Especialización en Econometría Aplicada
Centro de Formación Continua -UNI
Modelos Variable Dependiente Limitada
Clase 2

Edinson Tolentino
MSc Economics
email: edinson.tolentino@gmail.com

Twitter: @edutoleraymondi

Universidad Nacional de Ingeneria

24 de noviembre de 2024

Contenido

Introducción

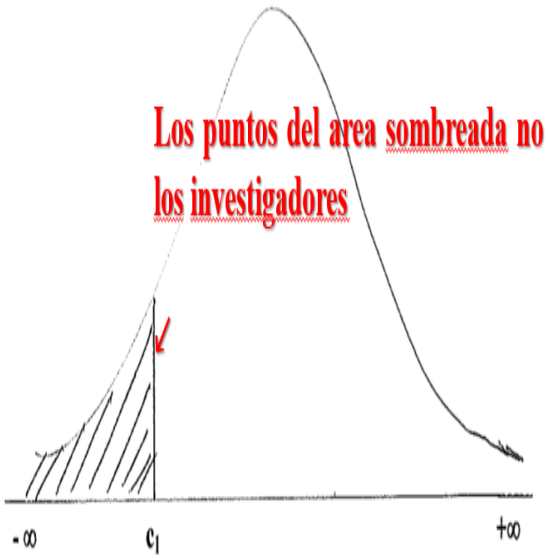
Los econometristas hacen una distinción entre modelos de regresión **censurados** y **truncados**

- Modelos de regresión censurado : se observa todos los casos de x, sin embargo no para los valores de y cuando el evento ocurre $(i.e\mathbf{y=0})$.
- Modelos de regresión truncada : no se observa todos los casos de x y tampoco para los valores de y cuando el evento ocurre $(i.e\mathbf{y=0})$.

Entonces, podemos seguir la siguiente expresión para el promedio de la población (μ):

$$E(X) = \int_{-\infty}^{+\infty} xf(x)\partial x$$

Función de distribución



Caso de Truncación por debajo:

- Truncación esta por debajo de la distribución del umbral o corte c_1
- Solo observaciones encima de dicho corte son observados

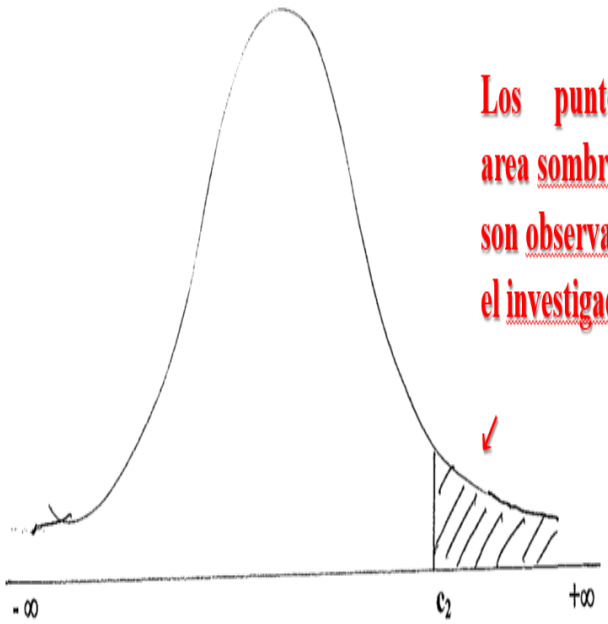
$$1 - F(c_1) = \int_{c_1}^{+\infty} f(x)\partial x$$

- Dividido entre $1 - F(c_1)$:

$$1 = \int_{c_1}^{+\infty} \frac{f(x)}{1 - F(c_1)}\partial x$$

$$f(x; c_1) = f(x \mid x \geq c_1) = \frac{f(x)}{1 - F(c_1)}$$

Función de distribución



El investigador solo observa los puntos desde abajo de c_2 y no las del area

Caso de Truncación por encima:

- Truncado esta por encima de la distribución a través del umbral o corte c_2
- Solo observaciones debajo de dicho corte son observados

$$F(c_2) = \int_{-\infty}^{c_2} f(x)\partial x$$

- Dividido entre $F(c_2)$:

$$1 = \int_{-\infty}^{c_2} \frac{f(x)}{F(c_2)}\partial x$$

$$f(x; c_2) = f(x \mid x \leq c_2) = \frac{f(x)}{F(c_1)}$$

Función de distribución

Promedio de la distribución normal truncada:

- **Caso de No truncada**

$$E(X) = \int_{-\infty}^{+\infty} xf(x)\partial x$$

- **Caso de truncada por debajo**

$$E(X|X \geq c_1) = \int_{c_1}^{+\infty} \frac{xf(x)}{1 - F(c_1)}\partial x$$

- **Caso de truncada por encima**

$$E(X|X \leq c_2) = \int_{+\infty}^{c_2} \frac{xf(x)}{F(c_2)}\partial x$$

Si se asume una distribucion normal $\mu \sim N(\mu, \sigma^2)$ y evaluamos estas dos integrales

- **Caso de truncada por debajo**

$$E(X|X \geq c) = \mu + \sigma \frac{\phi(c^*)}{(1 - \Phi(c^*))}$$

- **Caso de truncada por encima**

$$E(X|X < c) = \mu - \sigma \frac{\phi(c^*)}{(1 - \Phi(c^*))}$$

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

Introducción

- El modelo de regresión censurada da lugar cuando los valores de la variable dependiente son restringidos. Es decir, la variable dependiente es igual a cierto valor siempre que la variable este por debajo (o sobre) el umbral de censura.
- El valor especifico de la variable dependiente no es observado si cae por debajo o sobre cierto umbral. Para las observaciones censuradas, sabemos que los valores de la variable depndiente cae en un rango de censura, pero no sabemos el valor especifico. Mientras las observaciones no censuradas los valores son exactamente observado.
- Cuando los datos están censurdos tenemos acceso al total de observaciones de la muestra, solo que para cierto grupo de observaciones no conocemos los valores exactos de la variable dependiente, siendo no el caso de las variables explicativas.

Ejemplo:

- Datos de ingresos son codificados (censurada a la derecha).
- Gasto en bienes durables pueden tomar valores de cero y positivos.

Modelo (Tobit)

- El modelo Tobit opera bajo la idea de que existe un único punto de censura el cual es fijo y generalmente normalizado a cero.
- Consideremos un variable latente y_i^* que depende linealmente de x_i:

$$y_i^* = x_i'\beta + \varepsilon_i \quad \text{donde} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\Longrightarrow y_i^* \sim \mathcal{N}(x_i'\beta, \sigma^2)$$

- Ahora, los valores observados están censurados debajo de 0.

$$y_i = \begin{cases} y_i^* & \text{si } y_i^* > 0 \\ \\ 0 & \text{si } y_i^* \leq 0 \end{cases}$$

- La variable observada es una variable aleatorio mixta con valores de 0 y valores continuos para observaciones positivas de y_i^*.

Modelo (Tobit)

Estamos interesados:

$$E(y_i|x_i) = (1 - \Phi(\alpha))\,[y_i|y_i > 0] + \Phi(\alpha)\,[y_i|y_i = 0]$$

$$E(y_i|x_i) = \left[x_i'\beta + \sigma \underbrace{\frac{\phi(\alpha)}{[1 - \Phi(\alpha)]}}_{\lambda(\alpha)} \right] . [1 - \Phi(\alpha)]$$

Dado que la censura es en 0, se tiene que:

$$\alpha = \frac{0 - x_i'\beta}{\sigma} = \frac{-x_i'\beta}{\sigma}$$

$$\lambda(\alpha) = \frac{\phi(\alpha)}{1 - \Phi(\alpha)} = \frac{\phi(-x_i'\beta/\sigma)}{1 - \Phi(-x_i'\beta/\sigma)} = \frac{\phi(x_i'\beta)}{1 - (1 - \Phi(-x_i'\beta/\sigma))} = \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)}$$

Modelo (Tobit)

Como resultados, tenemos:

$$E(y_i|x_i) = \left[x_i'\beta + \Sigma\lambda\right]\Phi(x_i'\beta) = x_i'\beta\Phi(\frac{x_i'\beta}{\sigma}) + \sigma\phi(\frac{x_i'\beta}{\sigma})$$

La expresión sugiere que la media de la variable censurada es una función no lineal de x, β y σ.

Estimación

- Al igual que en el caso de truncamiento, la estimación de β por MCO resulta ser sesgada.
- Bajo el supuesto de normalidad, es posible obtener estimadores de Máxima Verosimilitud. Notese que la función de distribución de y_i es mixta, es decir, es discreta en cero y continua en el resto de valores.

$$
\begin{aligned}
Pr(y_i = 0) &= Pr(y_i^* \leq 0) \\
&= Pr(x_i'\beta + \varepsilon_i \leq 0) \\
&= Pr(\varepsilon_i \leq -x_i'\beta) \\
&= Pr(\varepsilon_i/\sigma \leq -x_i'\beta/\sigma) \\
&= \Phi(-x_i'\beta) = 1 - \Phi(x_i'\beta)
\end{aligned}
$$

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

Estimación

Con esta informaión, podemos construir la función de Verosimilitud:

$$L(\beta, \sigma, y, x) = \prod_{i=1}^{n} f(y_i = 0)^{I(y_i=0)} f(y_i, y_i > 0)^{I(y_i>0)}$$

$$= \prod_{i=1}^{n} \left[1 - \Phi(x_i'\beta/\sigma)\right]^{I(y_i=0)} \left[\frac{1}{\sigma}\phi\left(\frac{(y_i - x_i'\beta)}{\sigma}\right)\right]^{I(y_i>0)}$$

Y el logaritmo de la función de verosimilitud:

$$\ln L(\cdot) = \sum_{y_i=0} \ln\left[1 - \Phi(x_i'\beta/\sigma)\right] + \sum_{y_i>0} \ln\left[\sigma^{-1}\phi\left(\frac{(y_i - x_i'\beta)}{\sigma}\right)\right]$$

La cual puede ser maximizada con respecto a los parámetros de interés utilizando métodos numéricos. Esta función de verosimilitud es un tanto particular puesto que la primera parte suma probabilidades y la segunda densidades.

Efectos marginales

La interpretación de los parámetros depende de la pregunta de investigación.

- Si estamos interesados en la relación lineal subyacente de toda la población, el coeficiente β puede ser directamente interpretado.

$$\frac{\partial E(y_i^*|x_i)}{\partial x_k} = \beta_k$$

- Sin embargo, si estamos interesados en el efecto sobre el valor esperado observado censurado, el efecto marginal es:

$$\frac{\partial E(y_i^*|x_i)}{\partial x_k} = \frac{\partial \left[x_i'\beta\Phi\left(\frac{x_i'\beta}{\sigma}\right) + \sigma\phi\left(\frac{x_i'\beta}{\sigma}\right)\right]}{\partial x_k} = \beta_k\Phi\left(\frac{x_i'\beta}{\sigma}\right)$$

- Los efectos marginales dependen de las características individuales (x_i) por lo que tendremos que elegir entre EMP, EMM y EMEVR.

Efectos marginales

La interpretación de los parámetros depende de la pregunta de investigación.

- Dado la siguiente variable dependiente latente:

$$y_i^* = x_i'\beta + \gamma D_i + \varepsilon_i$$

- Donde D_i es una variable dummy (toma el valor de 1 o 0), donde:

$$E(y_i^*|x_i, D_i = 1) = \triangle_1 = \Phi\left(\frac{x_i\beta+\gamma}{\sigma}\right)\left[x_i'\beta + \gamma\sigma\frac{\phi\left(\frac{x_i\beta+\gamma}{\sigma}\right)}{\Phi\left(\frac{x_i\beta+\gamma}{\sigma}\right)}\right]$$

$$E(y_i^*|x_i, D_i = 0) = \triangle_0 = \Phi\left(\frac{x_i\beta}{\sigma}\right)\left[x_i'\beta + \sigma\frac{\phi\left(\frac{x_i\beta}{\sigma}\right)}{\Phi\left(\frac{x_i\beta}{\sigma}\right)}\right]$$

- Entonces el **efecto impacto** en este caso es dado por :

$$\triangle = \triangle_1 - \triangle_0$$

Implementación en Stata

Stata estima modelos Tobit estándar (o censurados en general) mediante el comando `tobit`, cuya sintaxis es:

Syntax

tobit `depvar [indepvars] [if] [in] [weight] [,options]`

Las opciones más importantes son:

- `ll(varname|#)`: indica el límite inferior para un censura a la izquiera.
- `ul(varname|#)`: indica el límite superior para un censura a la derecha.
- `vce(type)`: corrige los errores estandar: `robust`.

Es posible utilizar las opciones `ll(#)` y `ul(#)` al mismo tiempo. Asimismo, los comandos `post estimation` más comunes son `margins`, `marginsplot` y `predict`.

Aplicación

- Nos interesa estudiar los determinantes del gasto en bebidas alcohólicas y tabaco por parte de los hogares peruanos.

$$y_i = \beta_1 \text{edad} + \beta_2 \text{Educación} + \beta_3 \text{sexo} + \beta_4 \text{Enfermedad} + \beta_5 \ln .\text{ingresos}$$

- Variable dependientes (y_i): **Gasto mensual del hogar en bebidas alcohólicas y tabaco**.
- Regresores (x_i):
  1. Edad.
  2. Nivel educativo (primaria, secundaria y superior).
  3. Sexo.
  4. Enfermedad crónica.
  5. Ln. del ingreso mensual del hogar.
- La fuente de la información es la Encuesta Nacional de Hogares.

Introducción

- La inclusión de una unidad económica en la muestra depende de una decisión previa que no es exogena, por lo que resulta ser una muestra no aleatorio.
- De otra manera, el sesgo se genera cuando el componente no observable de la decisión de pertenecer a la muestra está correlacionado con el componente no observable del fenómeno bajo análisis.

Ejemplo:

1. Análisis de la oferta laboral.
2. Rendimiento estudiantil para el caso de escuelas privadas.

Modelo

- Analicemos la **oferta laboral**. Si la persona está laborando en el momento de la encuesta, entonces se observa su **salario**. Caso contrario, no es posible observar su salario. Por tanto, el truncamiento de la oferta salarial es incidental debido a que depende de otra variable: **la participación en la fuerza laboral**.
- La decisión de participar en el mercado laboral puede enmarcarse en un modelo de variable latente, donde z_i^* representa la utilidad de participar en el mercado laboral.

$$z_i^* = w_i'\delta + \varepsilon_i \;\; \rightarrow \textbf{Ecuación de selección}$$

- Esta variable no es directamente observable. Lo que si se observa es sí la persona i participa o no en el mercado laboral, resultado que se refleja cuando la utilidad supera cierto umbral (a).

$$\text{Binaria observable} \longrightarrow z_i = \begin{cases} 1 & \text{si } z_i^* > a \longrightarrow \text{Participa} \\ 0 & \text{si } y_i^* \leq a \longrightarrow \text{No participa} \end{cases}$$

Modelo

- La ecuación de interés es la oferta laboral, modelada tambien por una variable latente $(y_i^* \rightarrow$ salario por hora de los trabajadores$)$.

$$y_i^* = x_i'\beta + u_i \;\; \rightarrow \textbf{Ecuación de interés}$$

- La muestra no tiene observaciones de la distribución completa de y_i^* sino solo de aquellas observaciones provenientes de participar en el mercado laboral. La variable dependiente observada (y_i) viene dada por:

$$y_i = \begin{cases} y_i^* & \text{si } z_i^* > a \\ N.A & \text{si } z_i^* \leq a \end{cases}$$

Modelo

- Tomando esperanza de la ecuación de log(wage) :

$$E(y_i^*|y_i^* > 0) = x_i'\beta + E\left[\varepsilon_i|y_i^* > 0\right]$$

$$E(y_i^*|y_i^* > 0) = x_i'\beta + E\left[\varepsilon_i|\mu_i > -\frac{w_i\gamma}{\sigma_u}\right]$$

- Denotado $z_i = \frac{w_i\gamma}{\sigma_\mu}$
- Rellamamos

$$E\left[\mu_i|\mu_i > -z_i\right] = \sigma_\mu \frac{\phi(-z_i)}{(1-\Phi(-z_i))}$$

$$E(y_i^*|y_i^* > 0) = x_i'\beta + \frac{\sigma_{\mu\varepsilon}}{\sigma_\mu^2}\sigma_\mu\left[\frac{\phi(-z_i)}{(1-\Phi(-z_i))}\right]$$

Modelo

- El coeficiente de correlación entre dos variables aleatorias es:
- $\rho = \frac{\sigma_{\mu\varepsilon}}{\sigma_\varepsilon \sigma_\mu}$
- $\rho\sigma_\mu\sigma_\varepsilon = \sigma_{\mu\varepsilon}$

$$E(y_i^*|y_i^* > 0) = x_i'\beta + \frac{\rho\sigma_\mu^2}{\sigma_\mu^2}\sigma_\varepsilon \left[\frac{\phi(-z_i)}{(1-\Phi(-z_i))}\right]$$

$$E(y_i^*|y_i^* > 0) = x_i'\beta + \rho\sigma_\varepsilon \left[\frac{\phi(z_i)}{\Phi(z_i)}\right]$$

$$E(y_i^*|y_i^* > 0) = x_i'\beta + \rho\sigma_\varepsilon \left[\frac{\phi(\frac{w_i\gamma}{\sigma_\mu})}{\Phi(\frac{w_i\gamma}{\sigma_\mu})}\right]$$

- Donde $\sigma_\mu = 1$ en el threshold o modelo seleccion

$$E(y_i^*|y_i^* > 0) = x_i'\beta + \rho\sigma_\varepsilon \left[\frac{\phi(w_i\gamma)}{\Phi(w_i\gamma)}\right]$$

Estimación

- El modelo de selección es estimado por un **modelo probit** y el retorno del vector de γ
- El problema econométrico de truncación es tratado como una redvariable omitida en el marco de Heckman
- La variable $\frac{\phi(w_i\gamma)}{\Phi(w_i\gamma)}$ son los pseudo-residuos del modelo probit para el caso donde el evento ocurre.
  - Si el efecto de selección se encuentra presente y $\frac{\phi(w_i\gamma)}{\Phi(w_i\gamma)}$ es excluido de la ecuación: entonces la ecuación OLS presenta estimadores β que se encuentra **sesgado** e **inconsistente**
  - Si $\Phi(w_i\gamma) \longrightarrow 1$ sugiere que la probabilidad de selección es extremadamente alta y la variable de selección (psudo-residuo) es extremadamente baja dado $\Phi(w_i\gamma) \longrightarrow 0$

Estimación

La estimación del modelo puede realizarse a través de dos alternativas: MCO y Máxima Verosimilitud.

Estimación en 2 Etapas (Heckman, 1979)

En la primera etapa se estima la ecuación de selección, que caracteriza la forma en que las observaciones son incluidas en la ecuación principal. En la segunda etapa, se estima el modelo principal en la muestra no truncada incidentalmente.

- **Primera etapa**
  Se estima la **forma reducida** del modelo de selección usando un probit , dado el método de ML recuperar el estimador γ
  El modelo de la **forma reducida** deberia contener todas las **variables exogenas** relevantes y deberias tambien contener las variables que no se incluyen en el vector **x** (identificación de instrumentos)
  Use el estimado probit para construir la inversa **Ratio de Mills** (pseudo-residuo) , definido como

$$\lambda_i = \frac{\phi(w_i\gamma)}{\Phi(w_i\gamma)}$$

Estimación

- **Segunda etapa**
  Se aplica un OLS dada la siguiente ecuación:

$$y_i = x_i^{'}\beta + \theta\lambda_i + \zeta_i$$

  Donde $\theta = \rho\sigma_\varepsilon$
  - El procedimiento de Heckman provee un test parala muestra de selectividad (no aleatorio) sesgo
  - Un t-test de $H_o : \theta = 0$ versus $H_a : \theta \neq 0$ provee la base para el test de muestra de sesgo de selectividad
  - Si $\rho = 0$ y no existe correlación entre el error en la ecuación de seleccion y la ecuacion de regressión entre $\theta = 0$

Implementación en Stata

Stata estima modelos de regresión con sesgo de selección mediante el comando `heckman`, cuya sintaxis básica es:

Syntax

heckman `depvar [indepvars]`, **select**`(depvar_s = varlist_s) [twostep]`

Donde `depvar` $= y$, `[indepvars]` $= x$, `depvar` $= z$ y `varlist_s` $= w$.
Stata estima por default el método de máxima verosimilitud. Sin embargo, para especificar el método de dos etapas se debe utilicar la opción `twostep`. Las principales opciones son:

- `mle`: estimador de máxima verosimilitud (default).
- `twostep`: estimador de 2 etapas (Heckman).
- `select()`: especificación de la ecuación de selección.

Aplicación

- Estudiar los determinantes del salario de las mujeres casadas del Perú.

$$Pr(trabajar) = \beta_1 \text{Educación} + \beta_2 \text{Experiencia} + \beta_3 Experiencia^2 + \beta_4 \text{Niños menores de 6años} + \beta_5 \text{Niños entre 6 y 12 años} + \beta_6 \text{Pareja}$$

$$Salario = \beta_1 \text{Educación} + \beta_2 \text{Edad} + \beta_3 Edad^2$$

- Variables dependientes $(z_i \ ; \ y_i)$: **Dummy que toma el valor de 1 si se encuentra laborando** y **Salario mensual**, respectivamente.
- Regresores de la ecuación de selección (w_i):
  1. Años der educación.
  2. Años de experiencia.
  3. Niños menores de 6 años en el hogar.
  4. Niños entre 6 y 12 en el hogar.
  5. Si la persona tiene pareja
- Regresores de la ecuación de interes (x_i):
  1. Años der educación.
  2. Años de Edad
- La fuente de la información es la Encuesta Nacional de Hogares.