Universidad de Concepción

Modelo de Variable Dependiente Limitada

Censura

Felipe J. Quezada-Escalona
Departamento de Economía

Introducción

- En estos modelos, la variable de interés `Y` es continua (nuevamente).
- Sin embargo, por alguna razón, la variable `Y` está observada en forma incompleta o limitada:
  - Truncada
  - Censurada.
- Por ende, OLS no es válido ya que la muestra **no es representativa** de la población.

Introducción

Truncamiento: Las observaciones están **sistemáticamente excluidas** (var. dep. y explicativas eliminadas o perdidas). Es decir, no hay datos completos.

- Ejemplo: Encuesta de hogares donde no se incluye a hogares con ingresos muy altos. Por alguna razón, la muestra de ingreso de hogares no incluye los hogares que ganan 10 millones de pesos al mes.

Censura: Todas las observaciones son incluidas. Sin embargo, la variable dependiente y **se observa dentro de un rango**; por encima o por debajo de cierto **umbral** son tratados como si estuvieran en el umbral.

- Ejemplo: Encuesta de hogares donde se reemplaza ingreso "mayor a un millon" por un valor.
- Una persona que tiene un ingreso de "mil millones", se registra en la base como "un millón o mas" y en la base solo se observa "1.000.000".

Introducción

Truncamiento incidental a sesgo de selección:

- Hay un truncamiento donde la posibilidad de obtener la muestra en particular se relaciona de forma con la variable de interés. **Ejemplo:** Encuesta de innovación de Chile que solo incluye a las empresas que innovan.
- Hay un sesgo de selección. Por ejemplo, en un estudio de salarios femeninos que se base solo en usar datos de mujeres que trabajan las estimaciones enfrentan un sesgo ya que la decisión de trabajar no es aleatoria y depende del nivel de la variable de interés.

> **Ejemplo: Heckman (1979).** El ejemplo más famoso es el de *Heckman (1979)*, quien estudia el sesgo de selección en la estimación de salarios. El problema surge porque los salarios solo se observan para las personas que trabajan, es decir, la muestra no incluye a quienes deciden no participar en el mercado laboral. Si se estima una regresión de salarios usando solo a los trabajadores observados, el error estará correlacionado con la decisión de participar, generando estimadores sesgados.

Censura

Censura

Nuestro gran objetivo es plantear la función de verosimilitud, así que partamos por entender el contexto del modelo.

Mecanismo

Sea y el valor observado, la parte incompleta de y^*. En censura observamos toda la información de X_i, pero la censura en y^* puede ser:

Censura por debajo

$$y = \begin{cases} y^* & \text{si } y^* > L \\ L & \text{si } y^* \leq L \end{cases}$$

- Es decir, si la variable latente es menor a un umbral ($y^* \leq L$), entonces la variable y toma el valor del umbral L. Y si $y^* > L$, entonces $y = y^*$.
- Por ejemplo, $L = 0$ en una encuesta de gasto en bienes durables. Otro ejemplo, es lo que intentaba Tobit: **modelar una solución de esquina.**

Censura

Nuestro gran objetivo es plantear la función de verosimilitud, así que partamos por entender el contexto del modelo.

Mecanismo

Sea y el valor observado, la parte incompleta de y^*. En censura observamos toda la información de X_i, pero la censura en y^* puede ser:

Censura por encima

$$y = \begin{cases} y^* & \text{si } y^* \leq U \\ U & \text{si } y^* > L \end{cases}$$

- Por ejemplo, en una encuesta de hogares con ingreso mayor a $U = 10^6$.

Observaciones

- Supongamos que en los datos solo vemos la variable dependiente hasta un umbral. Por ejemplo, ingresos hasta un millón: todos los ingresos reportan $< U$ y los ingresos mayores se consideran censurados.
- Censura superior o inferior. Ejemplo: censura por debajo en los ingresos cuando $L = 0$. En este caso: para deuda negativa sustituimos 0.

Función de densidad con censura

Para y^*, la función de densidad $f^*(y^*|x, \theta)$ se define de la forma usual, donde θ representa los parámetros del modelo. Sin embargo, para la variable observada y, la densidad se ajusta para considerar la censura.

Censura por debajo en L

- Para los valores de $y > L$, la densidad de y es la misma que la densidad de y^*:

$$f(y|x) = f^*(y|x, \theta) \text{ si } y > L$$

- Para $y \leq L$, la densidad de y se concentra en el punto L:

$$f(y|x) = P(y^* \leq L|x, \theta) = F^*(L|x, \theta) \text{ si } y = L$$

donde $F^*(L|x, \theta)$ es la función de distribución acumulada de y^* evaluada en L.

Función de Verosimilitud

Para una observación i, la contribución a la función de verosimilitud se define como:

$$L_i(\theta) = \begin{cases} f^*(y_i|x_i,\theta) & \text{si } y_i > L \\ F^*(L|x_i,\theta) & \text{si } y_i = L \end{cases}$$

Podemos escribir esto de forma compacta usando una variable indicadora $d_i = \mathbb{I}(y_i > L)$:

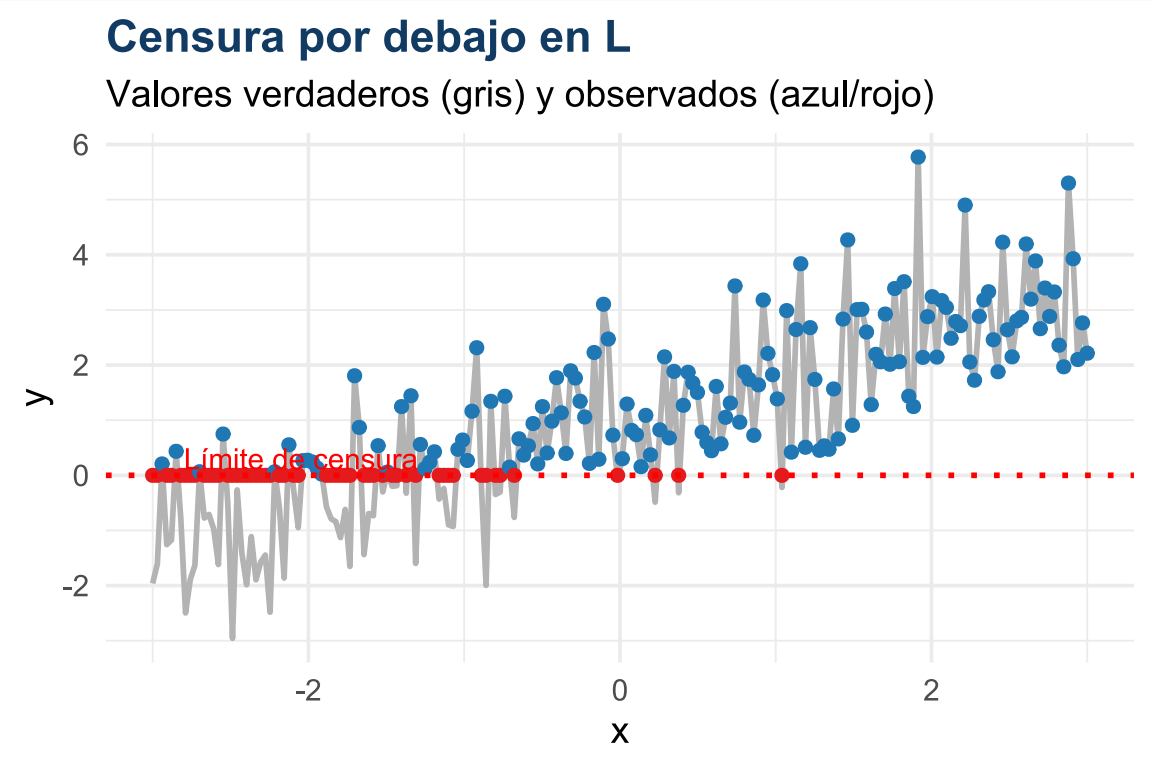
$$L_i(\theta) = [f^*(y_i|x_i,\theta)]^{d_i} \cdot [F^*(L|x_i,\theta)]^{1-d_i}$$

Log-verosimilitud

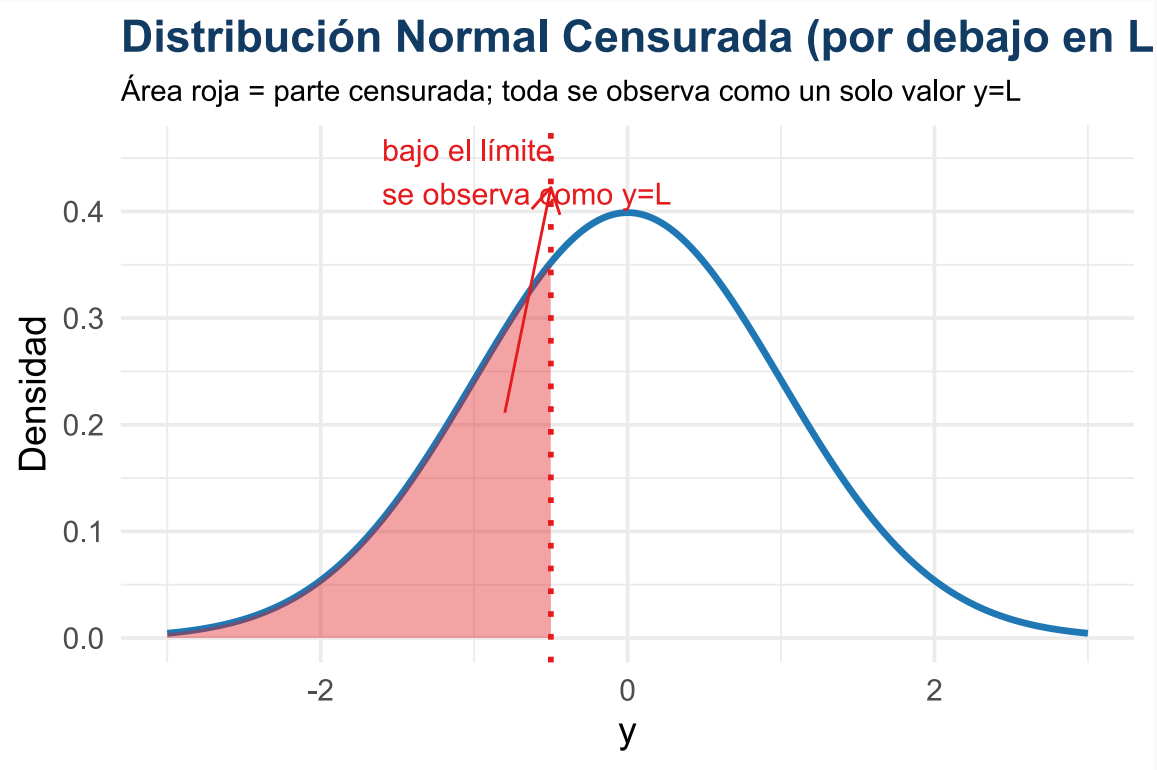
La log-verosimilitud para la muestra completa se obtiene sumando las contribuciones individuales:

$$\ell(\theta) = \sum_{i=1}^{n} [d_i \cdot \ln f^*(y_i|x_i, \theta) + (1 - d_i) \cdot \ln F^*(L|x_i, \theta)]$$

Censura: Grafico



Censura: Grafico



Incorporación del supuesto de Normalidad

- En muchos casos, se asume que la variable latente y^* sigue una distribución normal.
- Bajo el supuesto de normalidad, podemos derivar expresiones explícitas para la función de verosimilitud, lo que facilita la estimación e inferencia.

Distribución Normal Censurada

Consideremos una variable aleatoria z^* que sigue una distribución normal con media μ y varianza σ^2.

Supongamos que z^* está censurada por debajo en un umbral L. Esto significa que solo observamos $z = z^*$ si $z^* > L$, y observamos $z = L$ si $z^* \leq L$.

$$z^* \sim \mathcal{N}(\mu, \sigma^2) \quad ; \quad z = \begin{cases} z^*, & \text{si } z^* > L \\ L, & \text{si } z^* \leq L \end{cases}$$

Modelo Tobit

- El modelo Tobit es un modelo de **regresión censurada** donde la variable dependiente está censurada, típicamente en **cero**.
- Originalmente fue propuesto para soluciones de esquina (ej. adquirir un seguro agrícola).
- Es una extensión del modelo de regresión lineal que permite manejar la censura en la variable dependiente.
- Se asume que existe una variable latente y_i^* que sigue un modelo lineal con errores normales:

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

- La variable observada y_i es una versión censurada de y_i^*:

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

Modelo Tobit

Probabilidad de Censura

La probabilidad de que la variable latente sea menor o igual a cero (y por lo tanto censurada) es:

$$F^*(0) = P(y^* \leq 0) = P(\varepsilon \leq -X'\beta) = \Phi(-X'\beta/\sigma).$$

¿Por qué aparece la división por σ?

La división por σ ocurre porque la **CDF normal estándar** $\Phi(\cdot)$ asume una varianza igual a 1. Por tanto, debemos "escalar" el error para expresarlo en unidades de desviación estándar:

$$z = \frac{\varepsilon}{\sigma} \sim N(0,1)$$

Modelo Tobit

Función de log-verosimilitud

Para estimar los parámetros del modelo Tobit (β y σ), se utiliza el método de máxima verosimilitud. La función de log-verosimilitud se construye considerando la densidad de la variable observada y_i, que se ajusta para tener en cuenta la censura:

$$\ell(\beta, \sigma) = \sum_{i=1}^{n} \left[d_i \left(-\ln(\sqrt{2\pi}\sigma) - \frac{(y_i - X_i'\beta)^2}{2\sigma^2} \right) + (1 - d_i) \ln[1 - \Phi(X_i'\beta/\sigma)] \right]$$

donde $d_i = 1$ si $y_i > 0$, y $d_i = 0$ si $y_i \leq 0$. Es decir, el primer término dentro de la suma corresponde a las observaciones no censuradas, y el segundo término a las observaciones censuradas.

Modelo Tobit

Condiciones de Primer Orden (CPO) del Modelo Tobit

Derivadas de la función de log-verosimilitud respecto a los parámetros:

$$\frac{\partial \ell}{\partial \beta} = \sum_i \left[d_i \cdot \frac{(y_i - X_i'\beta)X_i}{\sigma^2} - (1-d_i) \cdot \frac{\phi(X_i'\beta/\sigma)}{1-\Phi(X_i'\beta/\sigma)} \cdot \frac{X_i}{\sigma} \right] = 0$$

$$\frac{\partial \ell}{\partial \sigma} = \sum_i \left[d_i \left(-\frac{1}{\sigma} + \frac{(y_i - X_i'\beta)^2}{\sigma^3} \right) + (1-d_i) \cdot \frac{\phi(X_i'\beta/\sigma)}{1-\Phi(X_i'\beta/\sigma)} \cdot \frac{X_i'\beta}{\sigma^2} \right] = 0$$

Estas ecuaciones no tienen una solución analítica cerrada y se utilizan métodos numéricos, como el algoritmo de Newton-Raphson, para encontrar las estimaciones de máxima verosimilitud de β y σ.

¡Muchas gracias!

Felipe J. Quezada-Escalona

Department of Economics

¿Preguntas?