Departamento de
Economía
Universidad de Concepción

Modelo de Variable Dependiente Limitada

Truncamiento

Felipe J. Quezada-Escalona

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.

Departamento de Economía
Universidad de Concepción

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".

Departamento de Economía
Universidad de Concepción

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.

Truncamiento

Truncamiento

El truncamiento es un tipo de censura donde las observaciones fuera de un rango determinado se **excluyen completamente** de la muestra. Esto implica una **pérdida de información** tanto de la variable dependiente como de las variables independientes.

Tipos de Truncamiento:

- **Truncamiento por debajo (L):**

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

- Ejemplo: Solo se incluyen observaciones con $y^* > L$ (e.g., hogares con ingresos mayores a 10 millones).

Truncamiento

El truncamiento es un tipo de censura donde las observaciones fuera de un rango determinado se **excluyen completamente** de la muestra. Esto implica una **pérdida de información** tanto de la variable dependiente como de las variables independientes.

Tipos de Truncamiento:

- **Truncamiento por encima (U):**

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

- Ejemplo: Solo se incluyen hogares con ingresos menores a 10 millones.

Truncamiento: Función de Densidad Condicional

Para modelar datos truncados, necesitamos ajustar la función de densidad para tener en cuenta la exclusión de observaciones fuera del rango permitido. La función de densidad condicional de y dado que $y^* > L$ (truncamiento por debajo) es:

$$f(y) = \frac{f^*(y)}{P(y^* > L)},$$

- donde $f^*(y)$ es la función de densidad no censurada (o no truncada) de y^*. $P(y^* > L)$ es la probabilidad de que y^* sea mayor que L (es decir, la probabilidad de que la observación no esté truncada).
- **Probabilidad de Truncamiento** La probabilidad de truncamiento se puede expresar en términos de la función de distribución acumulada (FDA) de y^*:

$$P(y^* > L) = 1 - F^*(L),$$

- donde $F^*(L)$ es la F.D.Acumulada de y^* evaluada en L.

Truncamiento: Verosimilitud

La función de verosimilitud para datos truncados se construye considerando la densidad condicional de las observaciones que no están truncadas. En el caso de truncamiento, la log-verosimilitud incluye solo las observaciones dentro del rango permitido:

$$\ell(\theta) = \sum_{i:y_i > L} \ln \left(\frac{f^*(y_i|x_i, \theta)}{1 - F^*(L|x_i, \theta)} \right).$$

donde θ representa los parámetros del modelo.

Truncamiento: Log-Verosimilitud

Para datos truncados por debajo en L, la función de log-verosimilitud se puede escribir como:

$$\ell(\theta) = \sum_{i=1}^{n} \{\ln f^*(y_i|x_i;\theta) - \ln\left[1 - F^*(L|x_i;\theta)\right]\}$$

donde $f^*(y_i|x_i;\theta)$ es la densidad condicional de y_i^* dado x_i con parámetros θ. $F^*(L|x_i;\theta)$ es la función de distribución acumulada (FDA) de y_i^* evaluada en el umbral L, condicional en x_i.

Pregunta: ¿Que pasa si es truncamiento es por encima?

Truncamiento: Log-Verosimilitud

Para datos truncados por debajo en L, la función de log-verosimilitud se puede escribir como:

$$\ell(\theta) = \sum_{i=1}^{n} \{\ln f^*(y_i|x_i;\theta) - \ln\left[1 - F^*(L|x_i;\theta)\right]\}$$

Notar que la log-verosimilitud se compone de dos términos:

- El primer término, $\ln f^*(y_i|x_i;\theta)$, es la log-verosimilitud de la variable no truncada y_i^*.
- El segundo término, $-\ln\left[1 - F^*(L|x_i;\theta)\right]$, ajusta la verosimilitud para tener en cuenta el truncamiento. Representa la probabilidad de observar y_i^* por encima del umbral L.
- **Importancia del Ajuste:** Ignorar el truncamiento en los datos implicaría **no incluir el segundo término** en la log-verosimilitud. Esto resultaría en estimaciones sesgadas e inferencias incorrectas, ya que la muestra truncada no es representativa de la población completa.

Truncamiento: Normalidad

A menudo se asume que la variable y_i^* sigue una distribución normal. En este caso, es importante comprender cómo el truncamiento afecta los momentos (media y varianza) de la distribución:

- Supongamos que z sigue una distribución normal $N(\mu, \sigma^2)$ y que está truncada por debajo en L. La función de densidad de la distribución normal truncada es:

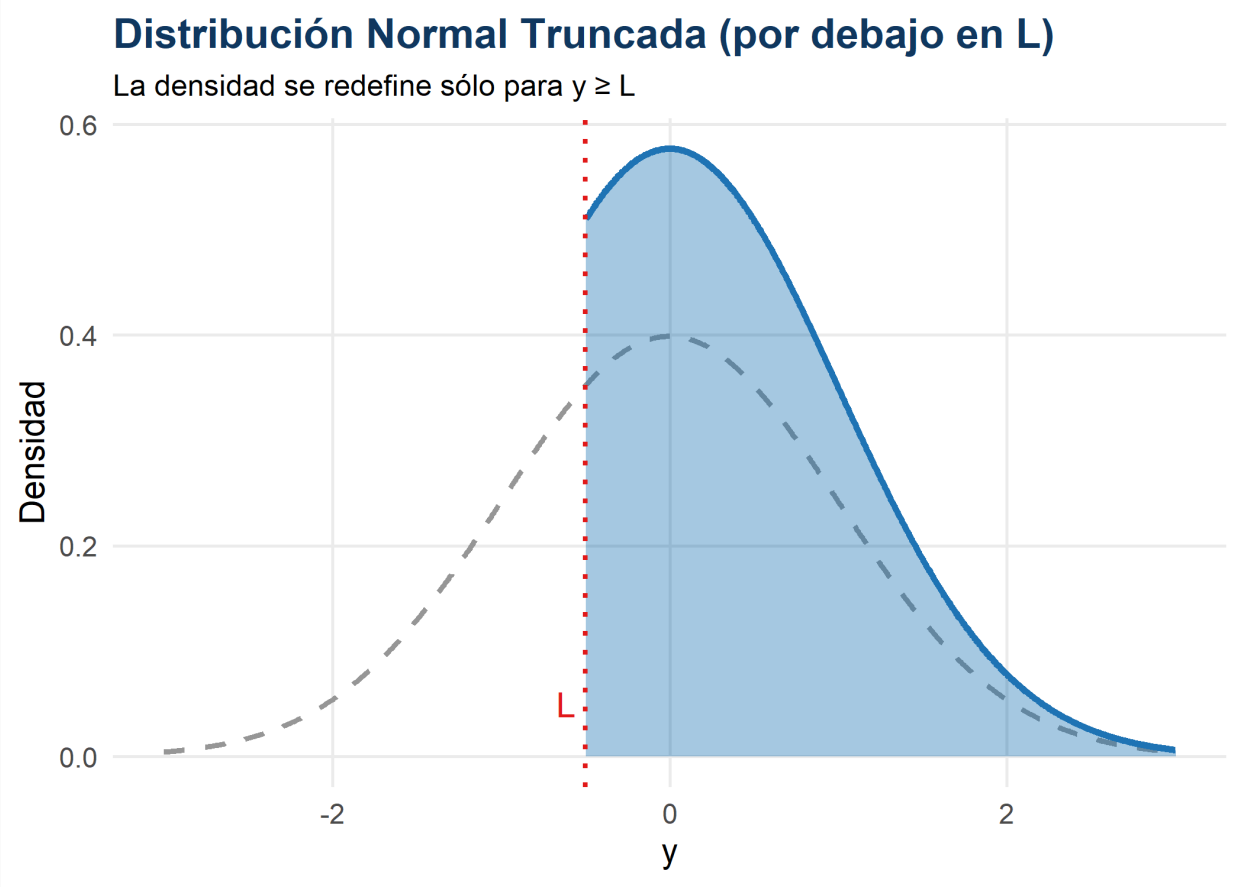
$$f(z|z \geq L) = \frac{f(z)}{1 - \Phi(\alpha)},$$

donde $\alpha = \frac{L-\mu}{\sigma}$ (valor estandarizado del umbral de truncamiento) y $f(z)$ es la densidad normal. $\Phi(\alpha)$ es la función de distribución acumulada normal estándar.

Truncamiento: Gráficamente

Departamento de Economía
Universidad de Concepción

La densidad normal truncada tiene forma similar a la densidad normal estándar pero restringida al intervalo $[L, \infty)$. La densidad truncada redefine la probabilidad sobre el rango permitido ($y \geq L$), y el área excluida corresponde a $\Phi(\alpha_i)$.



Departamento de Economía
Universidad de Concepción

Media

La media de la distribución normal truncada por debajo en L es:

$$E(Z \mid Z \geq L) = \mu + \sigma\lambda,$$

donde $\lambda = \dfrac{\phi(\alpha)}{1 - \Phi(\alpha)}$ es la **razón de Mills inversa**, y $\alpha = \dfrac{L - \mu}{\sigma}$ es la posición estandarizada del punto de truncamiento.

Intuición:

- Al truncar por debajo, eliminamos los valores bajos de la distribución.
- Por tanto, la media se **desplaza hacia la derecha** (aumenta) respecto a μ.
- Este desplazamiento está determinado por λ, que mide cuánto peso "perdimos" en la cola izquierda.

En el límite, si $L \to -\infty$, entonces $\Phi(\alpha) \to 0$, $\lambda \to 0$, y recuperamos $E(Z) = \mu$, como debería ser.

Departamento de Economía
Universidad de Concepción

Media

¿Por qué aparece σ en la fórmula?

$$E(Z \mid Z \geq L) = E(\mu + \sigma X \mid X \geq \alpha) = \mu + \sigma E(X \mid X \geq \alpha) = \mu + \sigma\lambda.$$

- La **media de la normal truncada no estándar** se obtiene **multiplicando por σ** porque la escala (la dispersión) de la variable original Z es σ veces la del estándar X

Departamento de Economía
Universidad de Concepción

Varianza

La varianza de la distribución normal truncada por debajo en L es:

$$V(Z \mid Z \geq L) = \sigma^2(1 - \delta),$$

donde $\delta = \lambda(\lambda - \alpha)$ es un **factor de ajuste**.

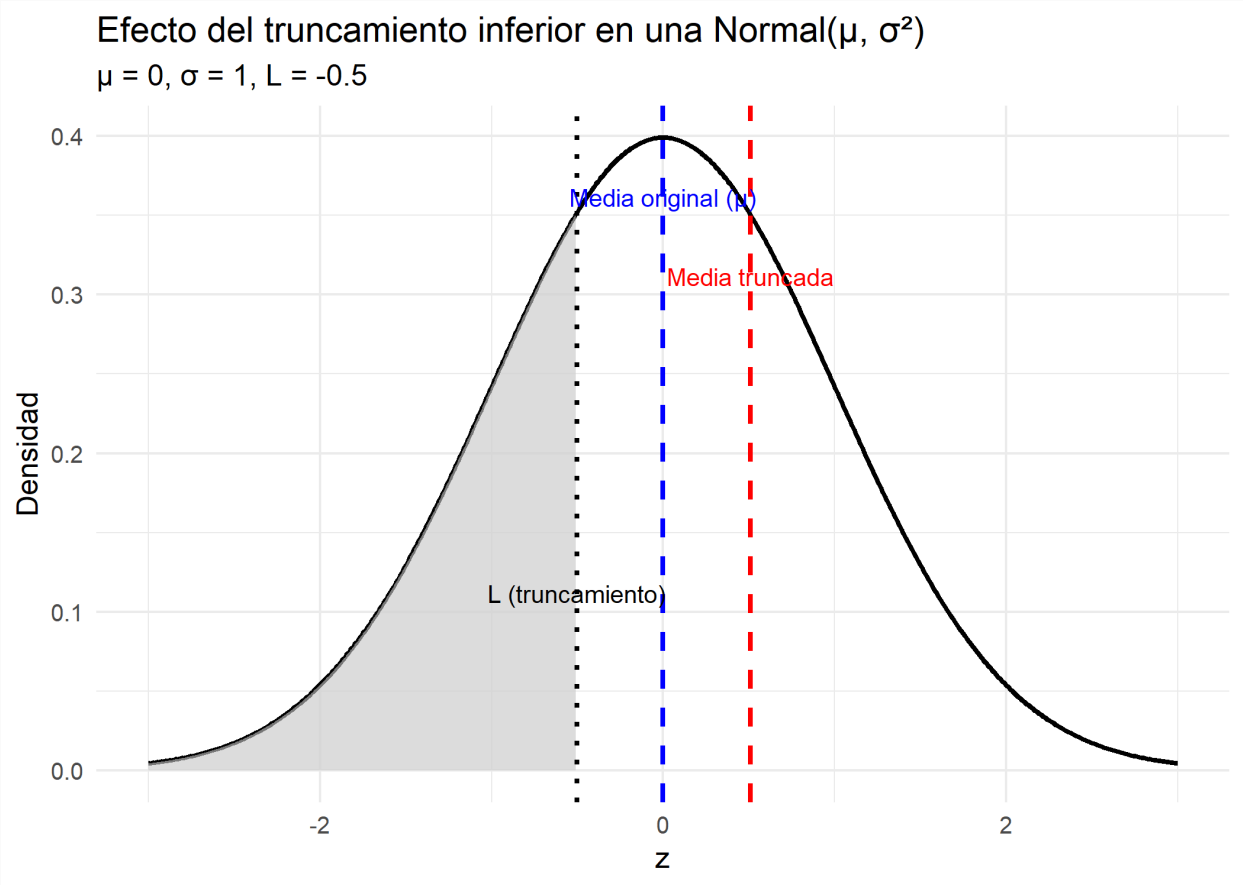
Intuición:

- La truncación elimina parte de la dispersión en la cola inferior.
- Por ello, la varianza **disminuye** respecto a la varianza original σ^2.
- El término δ mide exactamente cuánto se reduce la variabilidad debido al truncamiento.

De nuevo, si no truncamos ($L \to -\infty$), $\lambda \to 0$, $\delta \to 0$, y la varianza vuelve a ser σ^2.

Truncamiento: Momentos normal truncada

Departamento de Economía
Universidad de Concepción



Truncamiento: Modelo de Regresión Truncada

El modelo de regresión truncada se utiliza cuando la variable dependiente y^* está **truncada**, es decir, solo observamos valores de y^* que caen dentro de un rango determinado.

Estructura del Modelo

Se asume que la variable latente y_i^* sigue un modelo lineal:

$$y_i^* = x_i'\beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Departamento de Economía
Universidad de Concepción

El modelo de regresión truncada se utiliza cuando la variable dependiente y^* está **truncada**, es decir, solo observamos valores de y^* que caen dentro de un rango determinado.

Distribución de y_i^* (condicional al truncamiento)

Condicional en x_i, la variable y_i^* sigue una distribución normal con media $x_i'\beta$ y varianza σ^2:

$$y_i^* \mid x_i \sim N(x_i'\beta, \sigma^2).$$

Su **función de densidad** es:

$$f(y_i^* \mid x_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i^* - x_i'\beta)^2}{2\sigma^2}\right]$$

Truncamiento: Modelo de Regresión Truncada

El modelo de regresión truncada se utiliza cuando la variable dependiente y^* está **truncada**, es decir, solo observamos valores de y^* que caen dentro de un rango determinado.

Probabilidad acumulada hasta el umbral L

$$F^*(L) = \Phi\left(\frac{L - x_i'\beta}{\sigma}\right)$$

Densidad truncada completa

$$f^*(y_i \mid x_i; \beta, \sigma) = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right]}{1 - \Phi\left(\frac{L - x_i'\beta}{\sigma}\right)}.$$

Departamento de Economía
Universidad de Concepción

Log-verosimilitud

Al tomar logaritmos y sumar sobre $i = 1, \ldots, n$, la función de log-verosimilitud para el modelo de regresión truncada es:

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

Truncamiento: Modelo de Regresión Truncada

El modelo de regresión truncada se utiliza cuando la variable dependiente y^* está **truncada**, es decir, solo observamos valores de y^* que caen dentro de un rango determinado.

Observaciones:

- **OLS con y y x censurados o truncados:**
  - **Inconsistencia:** Si aplicamos OLS directamente a datos censurados o truncados, las estimaciones de los coeficientes serán inconsistentes. Esto se debe a que la muestra censurada o truncada no es representativa de la población, lo que introduce un sesgo en las estimaciones.

Truncamiento: Modelo de Regresión Truncada

El modelo de regresión truncada se utiliza cuando la variable dependiente y^* está **truncada**, es decir, solo observamos valores de y^* que caen dentro de un rango determinado.

Observaciones:

- **Aproximaciones:**
  - **Mínimos Cuadrados Ponderados:** Para corregir el sesgo, se pueden utilizar métodos de mínimos cuadrados ponderados, donde las ponderaciones se ajustan para tener en cuenta la censura o el truncamiento. Estos métodos son similares al **procedimiento de Heckman** para la corrección del sesgo de selección.
- **Supuesto de Normalidad:**
  - En muchos casos, se asume que los errores ϵ_i siguen una distribución normal. Este supuesto facilita la derivación de las expresiones para la función de verosimilitud y la estimación de los parámetros.

¡Muchas gracias!

¿Preguntas?

Felipe J. Quezada-Escalona

Departamento de
Economía
Universidad de Concepción

felipequezada.com