Departamento de
Economía
Universidad de Concepción

Modelo de Variable Dependiente Limitada

Modelos de Sesgo de Selección

Profesor: Felipe Quezada Escalona

Asignatura: Econometría II

Medias Condicionales

Bajo el supuesto de **normalidad**, $\epsilon_i \sim N(0, \sigma^2)$, la media condicional en un modelo truncado se simplifica a:

$$E(y_i|x_i, y_i^* \geq 0) = x_i'\beta + \sigma \cdot \lambda \left(\frac{x_i'\beta}{\sigma} \right).$$

donde $\lambda = \frac{\phi(\alpha)}{1-\Phi(\alpha)}$ es la relación de Mills inversa.

- **Pregunta:** ¿Como podria entonces usar esta información para estimar con MCO con datos censurados?

Estimador de Heckman

Estimador de Heckman

- El estimador de Heckman se utiliza para corregir el sesgo de selección en modelos de regresión con datos censurados.
    - Este sesgo surge cuando la muestra observada fuera del umbral no es representativa de la población, debido a un proceso de selección no aleatorio.
- **Ejemplo:** Consideremos el mercado laboral. Una persona **entra al mercado laboral** si el **salario ofrecido** es mayor al **costo de oportunidad** de trabajar. Por ende, una muestra del mercado laboral **excluye personas** para las cuales no es rentable entrar al mercado (pero si observamos con ingresos igual a cero).
    - Estimar un modelo con solo personas en el mercado laboral genera un problema de **sesgo de selección**.
    - **OLS es inconsistente** porque el salario observado está correlacionado con la probabilidad de entrar al mercado laboral.

Departamento de Economía
Universidad de Concepción

Contexto del mercado laboral

- Observamos y_i (salario) solo para las personas que trabajan ($d_i = 1$). d_i es una **variable dummy** que indica si una persona trabaja,

$$d_i = \begin{cases} 1 & \text{si trabaja} \\ 0 & \text{en caso contrario} \end{cases}$$

- **Variables Latentes**:
  - y_i^*: Variable latente de resultado (salario, incluso para los que no trabajan).
  - d_i^*: Variable latente que describe la decisión de participar en el mercado laboral.
- **Definición**:
  - Si $d_i^* > 0$: La persona participa en el mercado laboral ($d_i = 1$).
  - Si $d_i^* \leq 0$: La persona no participa en el mercado laboral ($d_i = 0$).

Departamento de Economía
Universidad de Concepción

Contexto del mercado laboral

Para modelar el sesgo de selección, se utiliza un sistema de dos ecuaciones:

Ecuación de Selección: Modela la decisión de participar en el mercado laboral:

$$d_i^* = z_i'\gamma + \epsilon_i, \quad d_i = 1 \text{ si } d_i^* > 0.$$

donde:

- z_i: Vector de variables que influyen en la decisión de participar (e.g., educación, edad, estado civil).
- γ: Vector de coeficientes.
- ϵ_i: Término de error.

Departamento de Economía
Universidad de Concepción

Contexto del mercado laboral

Ecuación de Resultado (*outcome*): Modela la variable de interés (salario):

$$y_i = x_i'\beta + u_i, \quad \text{observada solo si } d_i = 1.$$

donde:

- x_i: Vector de variables que influyen en el salario (e.g., experiencia, educación).
- β: Vector de coeficientes.
- u_i: Término de error.

Ejemplo: d_i es Participación en el mercado laboral; y_i es Salario individual; z_i son características que afectan la participación (ej., nivel educativo); x_i son características que afectan el salario (e.g., experiencia laboral).

Estimador de Heckman

- **Selección No Aleatoria:** El modelo aborda la selección no aleatoria de la muestra, donde la decisión de participar (o ser observado) está correlacionada con la variable de resultado.
- **Correlación de Errores:** Los errores ϵ_i y u_i pueden estar correlacionados. Esta correlación es la fuente del sesgo de selección, ya que implica que las variables no observadas que afectan la participación también afectan el resultado.
- **Estimación Conjunta:** Se requiere una estimación conjunta de las dos ecuaciones para obtener resultados consistentes.
    - El estimador de Heckman (de dos pasos) es un método común para corregir el sesgo de selección en este tipo de modelos.

Método de Heckman en Dos Pasos (Heckit)

Modelo Heckit

El truncamiento incidental o sesgo de selección surge cuando la observación de la variable dependiente está condicionada a un proceso de selección. Para modelar este tipo de datos, se utiliza un modelo con dos ecuaciones: una para la selección y otra para el resultado.

Departamento de Economía
Universidad de Concepción

Ecuación de Participación:

$$y_i^{1*} = x'_{i1}\beta_1 + \epsilon_{i1}$$

Esta ecuación modela la decisión binaria de participar o no en la muestra. La variable latente y_i^{1*} representa la propensión a participar, y x_{i1} son las variables que influyen en esta decisión.

Ecuación de Outcome:

$$y_i^{2*} = x'_{i2}\beta_2 + \epsilon_{i2}$$

Esta ecuación modela la variable de resultado y_i^{2*}, que solo se observa si el individuo decide participar ($y_i^{1*} > 0$). Las variables x_{i2} son las que influyen en el resultado.

> **Relación con el Modelo Tobit:** El modelo Tobit es un caso particular de este modelo donde `$y_i^{1*} = y_i^{2*}$`, es decir, la variable latente que determina la selección es la misma que la variable de resultado.

Supuesto de Normalidad

Se asume que los errores de ambas ecuaciones, ϵ_{i1} y ϵ_{i2}, siguen una distribución normal bivariada:

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim \mathcal{N}\left(0, \quad \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right).$$

- **Implicaciones:** ρ: El parámetro ρ representa la correlación entre los errores de la ecuación de participación y la ecuación de resultado.
- **Consistencia:** Si $\rho \neq 0$, existe una correlación entre las variables no observadas que afectan la participación y las que afectan el resultado. Ignorar esta correlación al estimar la ecuación de resultado resultará en estimaciones sesgadas e inconsistentes.

Idea general de los pasos:

Relación entre los Errores:

Bajo el supuesto de normalidad bivariada, la relación entre los errores de las dos ecuaciones se puede escribir como:

$$\epsilon_{i2} = \rho \epsilon_{i1} + \xi_i,$$

donde ξ_i es independiente de ϵ_{i1} y $\xi_i \sim \mathcal{N}(0, \sigma_\xi^2)$.

Media Condicional:

La media condicional del outcome (y_i^{2*}) dado que el individuo participa ($y_i^{1*} > 0$) es:

$$\mathbb{E}(y_i^{2*} | y_i^{1*} > 0) = x_{i2}'\beta_2 + \rho\lambda(x_{i1}'\beta_1),$$

donde $\lambda(\cdot)$ es el **Inverse Mills Ratio** $\lambda(x) = \frac{\phi(x)}{\Phi(x)}$.

Modelo Heckit

Corrección del Sesgo: El término $\rho\lambda(x'_{i1}\beta_1)$ corrige el sesgo de selección. La relación de Mills inversa ($\lambda(x)$) captura la información sobre la selección no aleatoria de la muestra.

- El método de Heckman, también conocido como Heckit, es un procedimiento en dos pasos que se utiliza para corregir el sesgo de selección en modelos de regresión. Este método es una **alternativa** a la estimación por máxima verosimilitud completa y es especialmente útil cuando se asume una distribución normal bivariada para los errores.

Departamento de Economía
Universidad de Concepción

Paso 1: Probit para la Participación

- Se estima un modelo probit para la variable indicadora de participación d_i, donde $d_i = 1$ si el individuo participa ($y_i^{1*} > 0$) y $d_i = 0$ en caso contrario.
- La probabilidad de participación se modela como:

$$P(d_i = 1|x_{i1}) = \Phi(x_{i1}'\beta_1),$$

donde $\Phi(\cdot)$ es la función de distribución acumulada de la normal estándar.

- Se obtienen las estimaciones de los coeficientes $\hat{\beta}_1$ del modelo probit.
- Se calcula la relación de Mills inversa (λ) para cada individuo utilizando las estimaciones del probit:

$$\hat{\lambda}_i = \lambda(x_{i1}'\hat{\beta}_1) = \frac{\phi(x_{i1}'\hat{\beta}_1)}{\Phi(x_{i1}'\hat{\beta}_1)},$$

Departamento de Economía
Universidad de Concepción

Paso 2: Regresión OLS Aumentada

- Se estima la ecuación de resultado (y_i^{2*}) mediante una regresión OLS que incluye la relación de Mills inversa ($\hat{\lambda}_i$) como una variable explicativa adicional:

$$y_i^{2*} = x_{i2}'\beta_2 + \rho\sigma_2\hat{\lambda}_i + \nu_i,$$

donde ρ es la correlación entre los errores de las dos ecuaciones y σ_2 es la desviación estándar del error en la ecuación de resultado.

- La inclusión de $\hat{\lambda}_i$ corrige el sesgo de selección.

Ventajas:

- **Simplicidad:** Es fácil de implementar, ya que solo requiere la estimación de un modelo probit y una regresión OLS.
- **Amplitud:** Es aplicable a una amplia variedad de modelos de selección, como el análisis del mercado laboral, la participación en programas sociales y las decisiones de inversión.
- **Supuestos:** Requiere menos supuestos que la máxima verosimilitud completa, pero asume la normalidad conjunta de los errores ϵ_{i1} y ϵ_{i2}.

Departamento de Economía
Universidad de Concepción

Prueba de Hipótesis:

- **Matriz de Varianza:** Es importante tener en cuenta que la matriz de varianza y los errores estándar deben ajustarse para considerar la estimación en dos pasos. Se utilizan métodos como el método de Murphy-Topel o el bootstrap para obtener errores estándar consistentes.
- Se puede realizar una prueba de hipótesis para determinar si la correlación entre los errores (ρ) es significativamente diferente de cero: $H_0 : \rho = 0$.
- Si se rechaza la hipótesis nula (H_0), se recomienda utilizar la máxima verosimilitud completa (MLE) en lugar de OLS, ya que OLS no captura adecuadamente el sesgo de selección cuando $\rho \neq 0$.

¡Muchas gracias!

¿Preguntas?

Felipe J. Quezada-Escalona

Departamento de Economía
Universidad de Concepción

felipequezada.com