Selección de Heckman

Corrigiendo el Sesgo de Selección Muestral

EC3003B - Economía Laboral Aplicada

Tecnológico de Monterrey

Miercoles 18 de febrero, 2025 | 3-5pm

Contenido

Motivación: ¿Quienes Observamos?

En la ecuación de Mincer:

$$\ln(w_i) = \beta_0 + \beta_1 S_i + \beta_2 X_i + \varepsilon_i$$

Problema: Solo observamos salarios de quienes **trabajan**.

Motivación: ¿Quienes Observamos?

En la ecuación de Mincer:

$$\ln(w_i) = \beta_0 + \beta_1 S_i + \beta_2 X_i + \varepsilon_i$$

Problema: Solo observamos salarios de quienes **trabajan**.

- No observamos el salario potencial de quienes no trabajan
- La decision de trabajar no es aleatoria
- Personas con bajo salario potencial pueden elegir no trabajar

Motivación: ¿Quienes Observamos?

En la ecuación de Mincer:

$$\ln(w_i) = \beta_0 + \beta_1 S_i + \beta_2 X_i + \varepsilon_i$$

Problema: Solo observamos salarios de quienes **trabajan**.

- No observamos el salario potencial de quienes no trabajan
- La decision de trabajar no es aleatoria
- Personas con bajo salario potencial pueden elegir no trabajar

Consecuencia

Si estimamos Mincer solo con quienes trabajan, los coeficientes pueden estar **sesgados**.

Ejemplo: Participacion Laboral Femenina

Situacion:

- Mujeres con salario potencial bajo pueden no trabajar
- Solo observamos mujeres con salario "suficiente"
- La muestra de mujeres trabajadoras no es representativa

Ejemplo: Participacion Laboral Femenina

Situacion:

- Mujeres con salario potencial bajo pueden no trabajar
- Solo observamos mujeres con salario "suficiente"
- La muestra de mujeres trabajadoras no es representativa

Sesgo de selección

Si mujeres con bajo salario no trabajan, el salario promedio observado de mujeres **sobreestima** el salario promedio verdadero.
$\Rightarrow$ Subestimamos la brecha de género.

Formalizacion del Problema

Dos ecuaciones:

1. Ecuación de selección (trabajar o no):

$$D_i^* = Z_i'\gamma + u_i$$
$$D_i = 1(D_i^* > 0)$$

2. Ecuación de resultado (salario):

$$Y_i = X_i'\beta + \varepsilon_i \quad \text{observado solo si } D_i = 1$$

Formalizacion del Problema

Dos ecuaciones:

1. Ecuación de selección (trabajar o no):

$$D_i^* = Z_i'\gamma + u_i$$
$$D_i = 1(D_i^* > 0)$$

2. Ecuación de resultado (salario):

$$Y_i = X_i'\beta + \varepsilon_i \quad \text{observado solo si } D_i = 1$$

Si Corr$(u_i, \varepsilon_i) \neq 0$**:**

- Los errores están correlacionados
- OLS en la muestra seleccionada es sesgado
- Necesitamos corregir por selección

Intuicion del Método

Idea clave: El sesgo de selección es una forma de variable omitida.

$$E[Y_i|X_i, D_i = 1] = X_i'\beta + E[\varepsilon_i|D_i = 1]$$

El término $E[\varepsilon_i|D_i = 1]$ no es cero si hay selección.

Intuicion del Método

Idea clave: El sesgo de selección es una forma de variable omitida.

$$E[Y_i|X_i, D_i = 1] = X_i'\beta + E[\varepsilon_i|D_i = 1]$$

El término $E[\varepsilon_i|D_i = 1]$ no es cero si hay selección.
Solución de Heckman:

1. Modelar la probabilidad de selección
2. Calcular el "Inverse Mills Ratio" (IMR)
3. Incluir IMR como control en la ecuación de resultado

El Inverse Mills Ratio

Bajo normalidad bivariada:

$$E[\varepsilon_i | D_i = 1] = \rho\sigma_\varepsilon \cdot \lambda(Z_i'\gamma)$$

Donde $\lambda(\cdot)$ es el **Inverse Mills Ratio**:

$$\lambda(z) = \frac{\phi(z)}{\Phi(z)}$$

- $\phi(\cdot)$: función de densidad normal estándar
- $\Phi(\cdot)$: función de distribución acumulada

El Inverse Mills Ratio

Bajo normalidad bivariada:

$$E[\varepsilon_i | D_i = 1] = \rho\sigma_\varepsilon \cdot \lambda(Z_i'\gamma)$$

Donde $\lambda(\cdot)$ es el **Inverse Mills Ratio**:

$$\lambda(z) = \frac{\phi(z)}{\Phi(z)}$$

- $\phi(\cdot)$: función de densidad normal estándar
- $\Phi(\cdot)$: función de distribución acumulada

Interpretación

λ captura la "selectividad" de la muestra. Mayor λ = muestra más seleccionada.

Procedimiento en Dos Etapas

Etapa 1: Probit de selección

$$P(D_i = 1|Z_i) = \Phi(Z_i'\gamma)$$

Estimar $\hat{\gamma}$ y calcular $\hat{\lambda}_i = \lambda(Z_i'\hat{\gamma})$

Procedimiento en Dos Etapas

Etapa 1: Probit de selección

$$P(D_i = 1|Z_i) = \Phi(Z_i'\gamma)$$

Estimar $\hat{\gamma}$ y calcular $\hat{\lambda}_i = \lambda(Z_i'\hat{\gamma})$

Etapa 2: OLS con correccion

$$Y_i = X_i'\beta + \delta\hat{\lambda}_i + \eta_i$$

Estimar en la muestra con $D_i = 1$.

Procedimiento en Dos Etapas

Etapa 1: Probit de selección

$$P(D_i = 1|Z_i) = \Phi(Z_i'\gamma)$$

Estimar $\hat{\gamma}$ y calcular $\hat{\lambda}_i = \lambda(Z_i'\hat{\gamma})$

Etapa 2: OLS con correccion

$$Y_i = X_i'\beta + \delta\hat{\lambda}_i + \eta_i$$

Estimar en la muestra con $D_i = 1$.

Importante

$\delta = \rho\sigma_\varepsilon$. Si $\delta \neq 0$, hay evidencia de sesgo de selección.

Identificacion: La Exclusion Restriction

Problema: Si $X = Z$, el modelo está debilmente identificado.
Solución: Necesitamos al menos una variable que:

1. Afecte la **selección** (trabajar o no)
2. NO afecte el **resultado** (salario)

Identificacion: La Exclusion Restriction

Problema: Si $X = Z$, el modelo está debilmente identificado.
Solución: Necesitamos al menos una variable que:

1. Afecte la **selección** (trabajar o no)
2. NO afecte el **resultado** (salario)

Ejemplos clásicos:

- Número de hijos pequeños (afecta si trabaja, no cuanto gana)
- Ingreso del conyuge
- Tasa de desempleo local

Sin exclusion restriction

El modelo se identifica solo por la no linealidad de λ, lo cual es frágil.

Comando heckman

```
* Método de maxima verosimilitud (preferido)
heckman ln_salario escolaridad experiencia experiencia2, ///
    select(trabaja = escolaridad experiencia hijos_peq ingreso_conyuge)

* Método de dos etapas
heckman ln_salario escolaridad experiencia experiencia2, ///
    select(trabaja = escolaridad experiencia hijos_peq ingreso_conyuge) ///
    twostep
```

Notas:

- `select()` específica la ecuación de selección
- Variables de exclusion: `hijos_peq, ingreso_conyuge`
- MLE es más eficiente pero requiere normalidad

Interpretación de Resultados

```
. heckman ln_salario escolaridad experiencia, ///
    select(trabaja = escolaridad experiencia hijos_peq)

                        |      Coef.   Std. Err.
------------------------+------------------------
ln_salario              |
           escolaridad  |     0.095      0.003
           experiencia  |     0.042      0.002
------------------------+------------------------
select                  |
           escolaridad  |     0.085      0.005
           experiencia  |     0.025      0.003
             hijos_peq  |    -0.350      0.020
------------------------+------------------------
              /athrho   |     0.280      0.050
             /lnsigma   |    -0.450      0.025
------------------------+------------------------
                  rho   |     0.272
                sigma   |     0.638
```

Interpretación de Parametros

Parametro	Interpretación
ρ (rho)	Correlación entre errores de selección y resultado
σ (sigma)	Desviacion estándar del error en ec. de salario
λ (lambda)	$= \rho \times \sigma$, coef. del IMR

Test de selección:

- $H_0 : \rho = 0$ (no hay sesgo de selección)
- Si rechazamos H_0, la correccion de Heckman es necesaria
- Stata reporta test de Wald automaticamente

Aplicación: Brecha de Género Corregida

Sin correccion:

- Brecha de género = 15 %
- Pero solo observamos mujeres que decidieron trabajar

Aplicación: Brecha de Género Corregida

Sin correccion:

- Brecha de género = 15 %
- Pero solo observamos mujeres que decidieron trabajar

Con correccion de Heckman:

- Estimamos salario potencial de todas las mujeres
- Brecha corregida puede ser mayor (20-25 %)
- Mas realista del mercado laboral completo

Limitaciones del Método

1. **Supuesto de normalidad**
   - Errores deben ser normales bivariados
   - Violación puede sesgar resultados

Limitaciones del Método

1. **Supuesto de normalidad**
   - Errores deben ser normales bivariados
   - Violación puede sesgar resultados
2. **Exclusion restriction**
   - Difícil encontrar variables validas
   - Sin ella, identificacion es frágil

Limitaciones del Método

1. **Supuesto de normalidad**
   - Errores deben ser normales bivariados
   - Violación puede sesgar resultados
2. **Exclusion restriction**
   - Difícil encontrar variables validas
   - Sin ella, identificacion es frágil
3. **Forma funcional**
   - Linealidad en ambas ecuaciones
   - Parametrico vs semi/no parametrico

En la práctica

Comparar resultados con y sin correccion. Si son similares, el sesgo puede no ser grave.

Resumen

El problema:

- Solo observamos a quienes trabajan
- Selección no aleatoria
- OLS puede estar sesgado

La solución:

- Modelo de dos ecuaciones
- Inverse Mills Ratio
- Correccion de Heckman

Comando Stata

```
heckman y x1 x2, select(d = z1 z2 z3)
```

¿Preguntas?

Próxima Sesión:

M08: Panel - Efectos Fijos

Jueves 19 de febrero, 3-5pm

Entrega E2 (Fichas): HOY 11:59pm