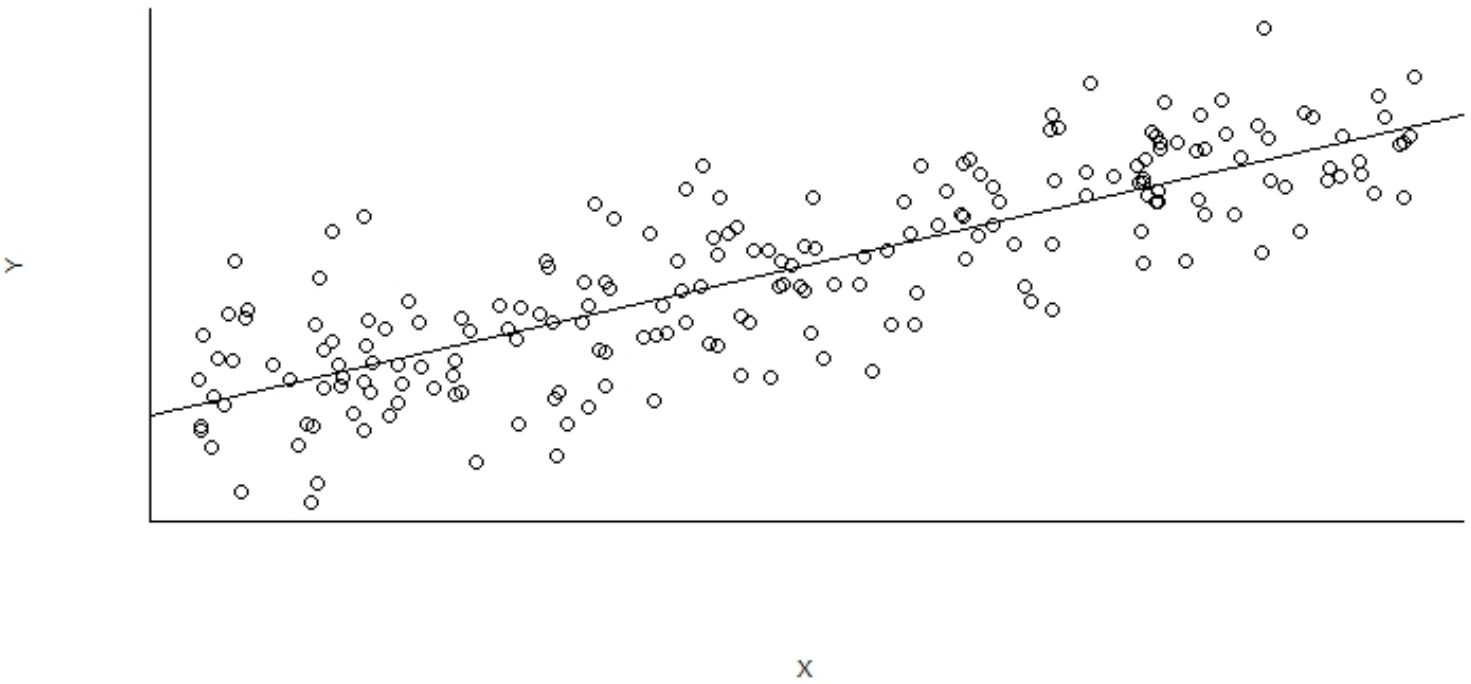
Selectividad Muestral

Walter Sosa-Escudero

Universidad de San Andres y CONICET

April 15, 2019

Motivacion



Que pasa si 'borramos' datos?. Ej: mujeres que no trabajan, compañeros que faltan a la reunion de egresados, etc. Muestra no aleatoria.

Preliminares 1: Normal truncada

$X \sim f(x)$. $X|X < a$: X *truncada* in a.

Resultado: si $X \sim N(\mu, \sigma^2)$,

$$E(X|X < a) = \mu - \sigma \frac{\phi(\alpha)}{\Phi(\alpha)},$$

con $\alpha \equiv (a - \mu)/\sigma$. $\phi(x), \Phi(x)$, densidad y fda normal estandar.

- Truncada a la derecha: esperanza a la izquierda (general).
- Normal: desplazamiento aditivo.
- Cuanto? Depende de α and σ^2
- $\lambda(z) \equiv \phi(z)/\Phi(z)$: inversa de la razon de Mills.

Preliminares 2: probits y variables latentes

$$y^* = x'\beta^* + u, \qquad u \sim N(0, \sigma^2)$$

- Si se observa $(y_i^*, x_i),\ i = 1, \ldots, n$ es posible estimar β y σ^2 consistentemente.
- Que es posible estimar si se observa $(y, x_i),\ i = 1, \ldots, n$, con $y_i = 1[y_i^* > 0]$?

Si $y^* = x'\beta^* + u$:

$$\begin{aligned} P(y=1|x) &= P(y^* > 0|x) \\ &= P(u > -x'\beta^*|x) \\ &= P(u/\sigma < x'\beta^*/\sigma \mid x) \quad \text{(Simetria)} \\ &= \Phi(x'\beta) \end{aligned}$$

con, $\beta \equiv \beta^*/\sigma$ y $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$.

- $P(y=1|x) = \Phi(x'\beta)$ es el modelo probit.
- Es posible estimar β por MV en base a $(y_i, x_i), i = 1, \ldots, n$.

$$y^* = x'\beta^* + u, \qquad u \sim N(0, \sigma^2)$$

- Con (y_i^*, x_i) es posible recuperar β^* y σ^2.
- Con (y_i, x_i) solo $\beta = \beta^*/\sigma^2$.
- σ^2 y β^* no estan identificados en una muestra (y_i, x_i).
- *Ejemplo:* $\beta^* = 10$ y $\sigma^2 = 2$ producen los mismos (y_i, x_i) que $\beta^* = 5$ y $\sigma^2 = 1$.

Preliminares 3: Omision de variables

Recordar que si en

$$Y = X_1\beta_1 + X_2\beta_2 + u$$

omitimos X_2, MCO de regresar Y en X_1 es en general sesgado. Se arregla controlando por X_2.

Selectividad muestral

$$y_i^* = x_i'\beta + u_i$$

s_i, variable de *selectividad*: $s_i = 1$ observado, 0 si no.

- 'Super muestra' de tamaño N de (y_i^*, x_i, s_i), por solo observamos una 'sub muestra' (y_i^*, x_i) solo cuando $s_i = 1$.
- Ejemplo: productividad de mujeres en el mercado laboral.
- Ejemplo: el efecto 'reunion de egresados'

Resultado: MCO bajo selectividad en general conduce a sesgos.

MCO bajo selectividad

Con una muestra aleatoria, (y_i^*, x_i), consistencia/insesgadez depende de

$$E(u_i|x_i) = 0,$$

que implica $E(y_i^*|x_i) = x_i'\beta$.

Con una muestra no aleatoria (condicionada en $s_i = 1$):

$$E(y_i|x_i, s_i = 1) = x_i'\beta + E(u_i|x_i, s_i = 1)$$

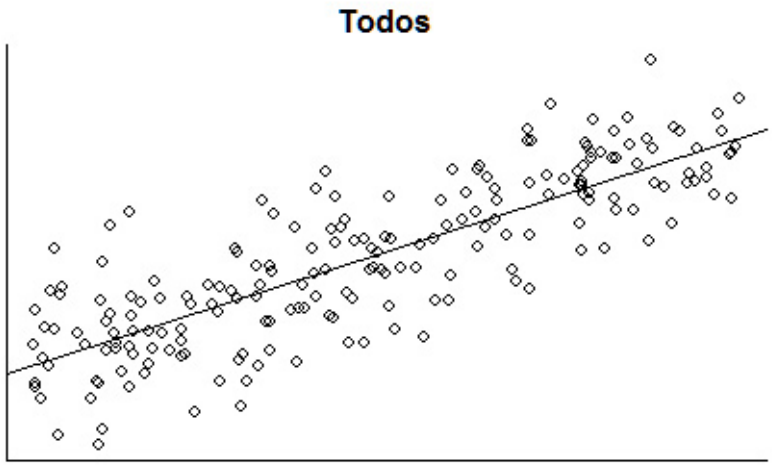
MCO bajo selectividad es inconsistente, a menos que $E(u_i|x_i, s_i = 1) = 0$.

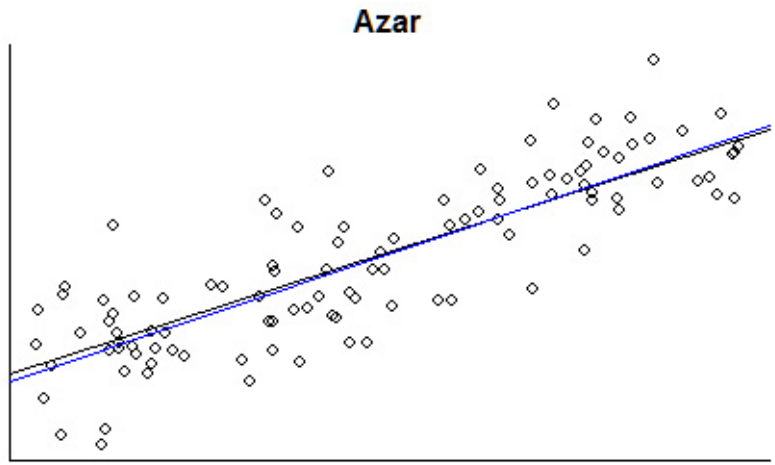
Discusion

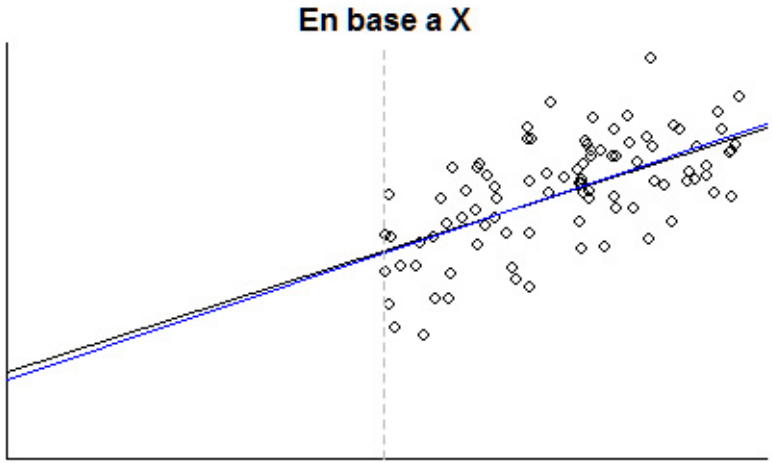
$$E(y_i|x_i, s_i = 1) = x_i'\beta + \underbrace{E(u_i|x_i, s_i = 1)}_{=0?}$$

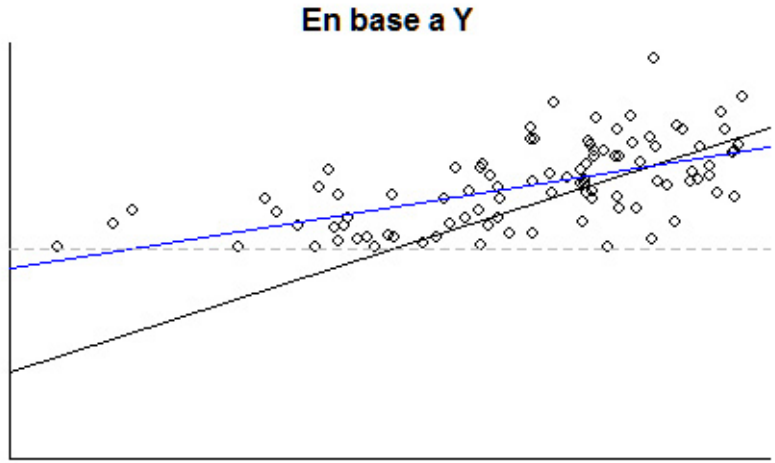
- No todo mecanismo de selectividad sesga a MCO.
- Si u independiente de s, MCO insesgado (por que?).
- Si $s = g(x)$, MCO insesgado.
- Cuatro ejemplos: salarios y educacion (hombres)
  - DNI par
  - Terminaron la primaria
  - Pasaron un test de inteligencia
  - Salarios, educacion, inteligencia pero para personas que pasaron un test de inteligencia

El sesgo depende de la conformacion del modelo


Todos


Azar


En base a X


En base a Y

Un modelo estimable bajo selectividad

$$\begin{cases} y_{1i} &= x'_{1i}\,\beta_1 + u_{1i} & \text{(regresion)} \\ y^*_{2i} &= x'_{2i}\,\beta_2 + u_{2i} & \text{(selectividad)} \end{cases}$$

$y_{2i} = 1[y^*_{2i} > 0]$.

Ejemplo: y_{1i} = salarios. Regresion: salarios en base a productividad. Selectividad: decision de trabajar (y^*_{2i} = utilidad neta del trabajo). x_1, determinantes de la productividad, x_{2i}, determinantes de la decision de trabajar.

Supuestos:

1. (y_{2i}, x_{2i}) se observa para todos.
2. (y_{1i}, x_{1i}) se observa solo si $y_{2i} = 1$ (muestra bajo selectividad).
3. (u_{1i}, u_{2i}) indepdendientes de x_{2i}, con esperanza nula.
4. $u_{2i} \sim N(0, \sigma_2^2)$.
5. $E(u_{1i}|u_{2i}) = \gamma\, u_2$. Los no observables *pueden* estar relacionados.

Sesgo por selectividad

$$y_{1i} = x'_{1i}\,\beta_1 + u_{1i}$$

Notar que aqui $s_i \equiv y_{2i}$.

$$\begin{aligned}
E(y_{1i}|x_{1i}, y_{2i}=1) &= x'_{1i}\beta_1 + E(u_{1i} \mid x_{1i}, y_{2i}=1) \\
&= x'_{1i}\beta_1 + E\big[E(u_{1i}|u_{2i}) \mid x_{1i}, y_{2i}=1\big] \\
&= x'_{1i}\beta_1 + E\big[\gamma u_{2i} \mid x_{1i}, y_{2i}=1\big] \\
&= x'_{1i}\beta_1 + \gamma E\big[u_{2i} \mid x_{1i}, y^*_{2i} > 0\big] \\
&= x'_{1i}\beta_1 + \gamma E\big[u_{2i} \mid x_{1i}\ ,\ u_{2i} < x'_{2i}\beta_2\big] \\
&= x'_{1i}\beta_1 - \gamma\sigma_2\ \lambda(x'_{2i}\beta_2/\sigma_2) \\
&= x'_{1i}\beta_1 - \gamma\sigma_2\ z_i \neq x'_{1i}\beta_1
\end{aligned}$$

con $z_i \equiv \lambda(x'_{2i}\beta_2/\sigma_2)$. MCO con la muestra seleccionada es inconsistente.

$$E(y_{1i}|x_{1i}, y_{2i} = 1) = x'_{1i}\beta_1 - \gamma\sigma_2 \; z_i \neq x'_{1i}\beta_1$$

- Inconsistencia: omision de z_i. Heckman (1979): sesgo de selectividad como especificacion incorrecta.
- Fuente de inconsistencia: correlation between u_{1i} y u_{2i} ($\gamma \neq 0$).

Estimador de Heckman en dos etapas

Definamos $u_{1i}^* \equiv y_{1i} - x_{1i}'\beta_1 - \gamma^* z_i$, con $\gamma^* \equiv -\gamma\sigma_2$. Despejando:

$$y_{1i} = x_{1i}'\beta + \underbrace{\gamma^* z_i + u_{1i}^*}_{u_{1i}}$$

Por construccion $E(u_{1i}^* | x_{1i}, y_{2i} = 1) = 0$.

- Si x_{1i}, z_i fuesen observables cuando $y_{2i} = 1$: MCO de y_{1i} en x_{1i} y z_i usando la muestra seleccionada estima consistemente a β_1 y γ^*.
- Problema: $z_i \equiv \lambda(x_{2i}'\ \beta_2/\sigma_2)$ no observable, depende de β_2 and σ_2.

Notar que $u_{2i} \sim N(0, \sigma_2^2)$, entonces:

$$P(y_{2i} = 1) = P(y_{2i}^* > 0) = P(u_{2i}/\sigma_2 < x_{2i}' \, \beta_2/\sigma_2) = \Phi(x_{2i}'\delta)$$

- $P(y_{2i} = 1)$ es un modelo *probit* concoeficiente desconocido δ.
- x_{2i} y y_{2i} se observan para todos: δ se puede estimar por MV via probit.
- Importante: no podemos indentificar β_{2i} y σ_2 por separado, pero no hace falta (tan solo $\delta = \beta_{2i}/\sigma_{2i}$).

Metodo en dos estpas (Heckman):

- *Etapa 1:* Estimar $\hat{\delta}$ via probit $P(y_{2i} = 1) = \Phi(x'_{2i}\delta)$ usando toda la muestra. Obtener $\hat{z}_i = \lambda(x'_{2i}\hat{\delta})$.
- *Etapa 2:* Regresar y_{2i} en x_{1i} y $\hat{z}_i$ usando la muestra seleccionada. Esto estima consistentemente a β_1 y γ^*.

En la practica

- Consistente y asintoticamente normal (metodo de momentos)
- Cuidado con la varianza asintotica. La segunda etapa requiere una correcion (usar software especifico).
- Test de $H_0 : \gamma = 0$ como test de selectividad.
- Muy baja potencia cuando x_1 es muy similar a x_2.
- MV? Requiere normalidad bivariada, verosimilitud muy inestable.