Seleccion Muestral

Walter Sosa Escudero

wsosa@udesa.edu.ar

Universidad de San Andrés

Cuestiones preliminares: Variables latentes y modelos probit

Recordemos el modelo probit:

$$Pr(y = 1|x) = \Phi(x'\beta)$$

β puede ser consistentemente estimado en base a una muestra aleatoria (y_i, x_i), $i = 1, \ldots, n$, por el metodo de maxima verosimilitud.

Consideremos el siguiente modelo de regresion, al cual llamaremos modelo de *variables latentes*:

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

Supongamos que y^* no es directamente observable, sino que se observa $y = 1[y^* > 0]$. La pregunta es qué es posible estimar (y que no) en base a esta informacion.

A tal efecto, notar que:

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

lo cual corresponde al modelo *probit* con $\beta \equiv \beta^*/\sigma$.

Consecuentemente, en base a una muestra (Y_i, x_i), el estimador MV permite estimar consistentemente β, aun cuando no es posible estimar consistentemente β^* y σ^2 por separado.

- En el modelo de variables latentes hay un *problema de identificacion*, es decir, existen varias configuraciones de los parametros que son compatibles con el mismo modelo estimable. Por ejemplo, en el caso de una sola variable explicativa, un modelo con $\beta^* = 10$ y $\sigma^2 = 2$ es observacionalmente identico (desde el punto de vista del modelo estimable en base a la muestra disponible) que uno en donde $\beta^* = 5$ y $\sigma^2 = 1$.
- En sintesis, si bien los parametros del modelo de variables latentes no estan identificados, es posible estimar consistentemente $\beta = \beta^*/\sigma$.

Introduccion

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

s, variable de seleccion, $s = 1$ si observado, 0 si no.

• Podemos pensar que existe una 'super muestra' de tamaño N de y_i^*, x_i, s_i y que observamos la 'sub muestra' y_i^*, x_i solo cuando $s = 1$.

• Ejemplo: productividad de mujeres. Solo observamos salarios para mujeres que trabajan, para el resto no observamos nada.

Consistencia de MCO bajo seleccion

Si tuviesemos una muestra aleatoria (y_i^*, x_i), consistencia descansa en que:

$$E(u|x) = 0$$

El problema es que ahora *no tenemos una muestra aleatoria*, sino que está condicionada en que, $s = 1$. Tomando esperanza condicional:

$$E(y|x, s = 1) = x'\beta + E(u|x, s = 1)$$

entonces, MCO usando la muestra seleccionada sera inconsistente, a menos que $E(u|x, s = 1) = 0$.

- No cualquier mecanismo de seleccion hace que MCO sea inconsistente. Si u es independiente de s, entonces MCO sigue siendo consistente.
- Otro ejemplo: el mecanismo de seleccion depende de x, por ejemplo, seleccionar en base a x no induce inconsistencia. Entonces, el caso relevante a estudiar es el caso en donde $E(u|x,s) \neq 0$.
- *Ejemplos:* Salarios y educacion. Muestra para hombres con DNI impar? Muestra para hombres con alguna educacion formal?

Un modelo simple de selectividad

Consideremos el siguiente sistema de ecuaciones:

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

para una muestra *inicial* o *completa* de tamaño $i = 1, 2, \ldots, N$. Definamos la variable binaria $y_{2i} = 1[y^*_{2i} > 0]$. La primer ecuacion recibe el nombre de *ecuacion de regresion* y la segunda de *ecuacion de seleccion.*

Ejemplo: y_{1i} = salario, la ecuacion de regresion es una ecuacion de determinacion de salarios en base a caracteristicas de la persona (x_{1i}). y_{2i} = utilidad neta de trabajar, x_{2i} son determinantes de la utilidad.

Supuestos:

1. (y_{2i}, x_{2i}) se observa para *todos* los elementos de la muestra completa.
2. (y_{1i}, x_{1i}) se observa solo si $y_{2i} = 1$. A esta muestra la llamaremos *muestra seleccionada*.
3. (u_{1i}, u_{2i}) son independientes de x_{2i} y tienen media cero.
4. $u_{2i} \sim N(0, \sigma_2^2)$
5. $E(u_{1i}|u_{2i}) = \gamma\, u_2$. Esto permite que los no-observables de ambas ecuaciones esten relacionados.

Sesgo por selectividad

Notar que en este caso, $s_i \equiv y_{2i}$: ecuacion de seleccion probit.

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

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

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

- La inconsistencia tiene que ver con omitir el termino z_i y la posible correlacion entre z_i y x_{1i}. Heckman (1979): sesgo de seleccion como un 'error de especificacion'.
- La inconsistencia se debe a la correlacion existente entre u_{1i} y u_{2i}, o sea, a que $\gamma \neq 0$.

Un estimador consistente en dos etapas

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

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

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

- Si x_{1i} y z_i fuesen observables cuando $y_{2i} = 1$, MCO de regresar y_{1i} en x_{1i} y z_i usando la muestra seleccionada daria estimaciones consistentes de β_1 y γ.
- Problema: $z_i \equiv \lambda(x_{2i}'\beta_2/\sigma_2)$ no es observable ya que depende de β_2 y σ_2.

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

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

$P(y_{2i} = 1)$ corresponde a un modelo *probit* con coeficiente δ.

x_{2i} y y_{2i} son observados para la muestra *completa*: δ puede ser consistentemente estimado en base al modelo probit.

Importante: no es posible identificar β_{2i} y σ_{2i} por separado, sino $\delta = \beta_{2i}/\sigma_{2i}$. Entonces, β_{2i} puede ser consistentemente estimado utilizando el siguiente procedimiento en dos etapas:

- *1a Etapa:* Obtener estimaciones $\hat{\delta}$ en base al modelo probit $P(y_{2i} = 1) = \Phi(x'_{2i}\delta)$ utilizando la muestra *completa*. Estimar z_i utilizando $\hat{z}_i = \lambda(x'_{2i}\hat{\delta})$.

- *2a Etapa:* Regresar y_{2i} en x_{1i} y $\hat{z}_i$ utilizando la muestra *seleccionada*, lo cual produce estimaciones consistentes de β_{1i} y γ.

- Este metodo es conocido con el nombre de *metodo de Heckman en dos etapas*, o *Heckit*. La consistencia del estimador de la segunda etapa se sigue porque estamos reemplazando z_i por el estimador consistente $\hat{z}_i$, de modo que por continuidad del estimador MCO con respecto a z_i, la consistencia de este estimador no se altera si reemplazamos z_i por un estimador consistente.
- Importante que para poder estimar la primer etapa necesitamos la muestra *censurada* y para la segunda, la *truncada*.

Cuestiones practicas y extensiones

- El estimador en dos etapas es asintoticamente normal. Los metodos estandar (tests 't', etc.) funcionan correctamente.
- El problema es la estimacion de la matriz de varianzas asintotica. En primer lugar, es facil mostrar que el modelo de la segunda etapa es heterocedastico. Esto es facil de corregir si z_i fuese directamente observable. El segundo problema es que z_i no es observable, lo cual implica tener que hacer ajustes al estimador de la varianza de la segunda etapa. Ver Greene (Cap. 20). La mayoria de los paquetes econometricos hacen este ajuste. No es un ajuste simple tipo White (lo seria si z_i fuese observable).

- Un test de $H_0 : \gamma = 0$ provee una test simple de 'sesgo por seleccion'. Bajo H_0 el modelo de regresion con la muestra seleccionada es homocedastico, por lo que el test puede realizarse sin hacer correcciones por heterocedasticidad.
- Problema clasico con Heckman: alta correlacion entre x_1 y $\lambda(.)$. Esta es monotona creciente, con poca variabilidad se asemeja a una funcion lineal. Si $x_1 \simeq x_2$, la correlacion entre x_{1i} y z_i puede ser muy alta. De hecho, la posibilidad de identificar γ tiene que ver estrictamente con que λ no es lineal. Si x_1 y x_2 tienen muchos elementos en comun, la estimacion de la segunda etapa estara sujeta a un problema de multicolinealidad alta. Esto se manifiesta en valores poco significativos para γ en la segunda etapa. Resulta crucial especificar que va en la ecuacion de regresion y que la de seleccion.

- Estimacion maximo-verosimil: Bajo el supuesto (mas restrictivo) de que (u_1, u_2) tienen distribucion normal bivariada es posible construir un estimador maximo-verosimil para los parametros del modelo. Esta estrategia es usualmente descartada porque la funcion de verosimilitud es muy complicada y no-concava, conduciendo a resultados muy inestables (rara vez se usa en la practica).
- Sin embargo, ver Nawata y Nagase.