

El Modelo de Selección de Heckman en gretl

Microeconomía Cuantitativa

R. Mora

Departamento de Economía
Universidad Carlos III de Madrid

Esquema

- 1 Introducción: El modelo de Heckman
- 2 Heckit y gretl

Introducción

El modelo de selección de Heckman

observamos w_i si $s_i = 1$

- ecuación output: $w = \beta_0 + \beta x + \varepsilon$
- ecuación de participación: $s = 1(\gamma'z + v)$
- $\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho \\ \rho & 1 \end{bmatrix}\right)$

Estimación

- MCO es inconsistente.
- Estimación MV es consistente: la expresión de la probabilidad es más complicada que la del modelo probit y tobit, ya que requiere la obtención de la distribución conjunta de w y s
 - En general, la función de probabilidad no es globalmente cóncava, y puede tener máximos locales
- El procedimiento de dos etapas de Heckman basado en la expectativa condicional da estimaciones consistentes y es fácil de implementar.
 - Se puede utilizar para obtener condiciones iniciales para Estimación MV.
 - Los errores estándar habituales obtenidos en la segunda regresión no son válidos.

Heckman y gretl

Comandos y funciones para Estimación Heckit

- `heckit`: computa el modelo de selección de Heckman
- `restrict`: constrasta hipótesis para los parámetros de las dos ecuaciones

heckit *output* *x__vars* ; *selection* *z__vars* —two-step

- *output* representa la variable dependiente en la ecuación *output*
- *x__vars* representa la lista de controles en la ecuación *output*
- *selection* representa la variable dependiente en la ecuación de participación
 - *selection* debe ser una variable binaria $\{0,1\}$
- *z__vars* representa la lista de controles en la ecuación de participación
- —two-step: ejecuta el procedimiento de dos etapas de Heckman, mostrando estimaciones consistentes de los errores estandar (MV se ejecuta por defecto)

Un ejemplo simple

Participación

- $U_m - U_h = -0.5 + 0.03 * educ - 1.5 * kids + v$

Ecuación de salarios

- $wage = 5 + 0.07educ + u$

- $cov(educ, u) = 0$

- $\begin{bmatrix} u \\ v \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} \right)$

Un ejemplo sencillo: estimación MCO ($\beta_{educ} = .07, \rho = .9$)

```
ols wage const educ
```

Model 1: OLS, using observations 1-5000 (n = 1112)
 Missing or incomplete observations dropped: 3888
 Dependent variable: wage
 Heteroskedasticity-robust standard errors, variant HC1

	coefficient	std. error	t-ratio	p-value	
const	6.12441	0.0979021	62.56	0.0000	***
educ	0.0561435	0.00689433	8.143	1.03e-15	***
Mean dependent var	6.904610	S.D. dependent var	0.826190		
Sum squared resid	713.5680	S.E. of regression	0.801782		
R-squared	0.059060	Adjusted R-squared	0.058212		
F(1, 1110)	66.31550	P-value(F)	1.03e-15		
Log-likelihood	-1331.197	Akaike criterion	2666.394		
Schwarz criterion	2676.422	Hannan-Quinn	2670.186		

El sesgo en MCO

- En el ejemplo, tenemos lo siguiente:
 - Los verdaderos rendimientos a la educación son aproximadamente 7% ($\beta_{educ} = .07$).
 - La puntuación para la participación también depende de la educación ($\gamma_{educ} = .03$).
 - Es importante destacar que los factores no observables (por el econometra) tanto de los salarios como de la decisión de participación están correlacionados positivamente ($\rho = 0.9$).
- Esta correlación positiva implica que los participantes en el mercado de trabajo con menores niveles de educación tienden a tener errores positivos en la ecuación salarial.
- OLS subestima los retornos a la educación: El intervalo de confianza al 95%: (4.26, 6.97)

Poniendo $wage = 0$ para los missing ($\beta_{educ} = .07, \rho = .9$)

Model 5: OLS, using observations 1–5000

Dependent variable: wage2

Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	Std. Error	t-ratio	p-value
const	0.819650	0.157393	5.2077	0.0000
educ	0.0567067	0.0117025	4.8457	0.0000

Mean dependent var	1.581584	S.D. dependent var	2.925883
Sum squared resid	42586.28	S.E. of regression	2.919018
R^2	0.004886	Adjusted R^2	0.004687
$F(1,4998)$	23.48076	P-value(F)	1.30e-06
Log-likelihood	-12449.93	Akaike criterion	24903.86
Schwarz criterion	24916.89	Hannan-Quinn	24908.42

El sesgo de MCO usando toda la muestra

- No se corrige el sesgo si ponemos $wage = 0$ en las observaciones de individuos que no participan.
- Cuando reemplazamos el verdadero salario por 0 cuando $s = 0$, entonces el modelo es:
 - $w = \begin{cases} \beta_0 + \beta x + \varepsilon & \text{if } s = 1 \\ 0 & \text{if } s = 0 \end{cases}$
 - $s = 1(\gamma'z + v)$
 - $\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho \\ \rho & 1 \end{bmatrix}\right)$
- Es fácil demostrar que $E(w|x) \neq \beta_0 + \beta_1 x + \varepsilon$.

Estimación MV ($\beta_{educ} = .07, \rho = .9$)

heckit wage const educ ; work const educ kids

Model 3: ML Heckit, using observations 1-5000
 Dependent variable: wage
 Selection variable: work

	coefficient	std. error	t-ratio	p-value	
const	4.97764	0.110538	45.03	0.0000	***
educ	0.0700091	0.00705706	9.920	3.39e-23	***
lambda	0.933032	0.0374084	24.94	2.62e-137	***

Selection equation

const	-0.473329	0.0877324	-5.395	6.85e-08	***
educ	0.0257985	0.00617236	4.180	2.92e-05	***
kids	-1.46115	0.0438994	-33.28	6.57e-243	***

Mean dependent var	6.923428	S.D. dependent var	0.815325
sigma	1.036613	rho	0.900076
Log-likelihood	-3142.936	Akaike criterion	6291.873
Schwarz criterion	6306.835	Hannan-Quinn	6297.538

Total observations: 5000
 Censored observations: 3917 (78.3%)

Estimación en dos etapas ($\beta_{educ} = .07, \rho = .9$)

```
heckit wage const educ ; work const educ kids --two-step
```

Model 2: Two-step Heckit, using observations 1-5000
 Dependent variable: wage
 Selection variable: work

	coefficient	std. error	t-ratio	p-value	
const	5.05706	0.121918	41.48	0.0000	***
educ	0.0690590	0.00713587	9.678	3.75e-22	***
lambda	0.874918	0.0541856	16.15	1.20e-58	***

Selection equation

const	-0.473118	0.0894129	-5.291	1.21e-07	***
educ	0.0261450	0.00629308	4.155	3.26e-05	***
kids	-1.48345	0.0470196	-31.55	1.82e-218	***

Mean dependent var 6.923428 S.D. dependent var 0.815325
 sigma 1.009690 rho 0.866521

Total observations: 5000
 Censored observations: 3917 (78.3%)

- Tanto el estimador MV y el procedimiento en dos etapas dan estimaciones consistentes.
- El estimador MV es más preciso. Esto es cierto para muestras grandes.
- Entre los resultados, también obtenemos estimaciones de la correlación de los errores.
 - Si $\rho = 0$, entonces no hay sesgo de selección de la muestra.
- La inferencia para la significatividad de los parámetros en las ecuaciones de salida y de participación puede llevarse a cabo como de costumbre.
- La predicción no se puede implementar utilizando `fcast`

restrict

- El comando `restrict` permite contrastar simultáneamente varias restricciones.
 - El contraste se realiza sobre las estimaciones del último modelo estimado antes de invocar el comando.
 - Después de `heckit`, la numeración de los parámetros sigue el orden de la visualización de la salida.
 - La opción `--quiet` hace que el resultado de la estimación del modelo restringido no aparezca.
- Para ponerlo en práctica, creamos un bloque:
`restrict`
aquí insertamos tantas líneas como restricciones a contrastar
`end restrict`

Ejemplo de contraste

```
? restrict --quiet
? b[lambda]=0
? end restrict
Restriction:
  b[lambda] = 0

Test statistic: chi^2(1) = 732.782, with p-value = 2.22441e-161

? restrict --quiet
? b[5]=0.03
? b[6]=-1.5
? end restrict
Restriction set
  1: b[educ] = 0.03
  2: b[kids] = -1.5

Test statistic: chi^2(2) = 0.635692, with p-value = 0.727715
```

Constraste de la significatividad de λ

- Podemos contrastar la significatividad del parámetro asociado a λ en la expectativa condicional de la ecuación output.
- Este test es un contraste de selección de la muestra aleatoria.
- Si el parámetro no es significativo, entonces no rechazamos la hipótesis nula de selección aleatoria (y MCO será consistente).
- En el ejemplo anterior, la hipótesis nula es fuertemente rechazada: encontramos pruebas de selección no aleatoria.

Efectos marginales

```
# marginal effects of another year of education
genr coeff=$coeff
genr beta=coeff[1:2]
genr gamma=coeff[4:6]
series educ0=educ
matrix x0={const,educ0}
series educ1=educ+1
matrix x1={const,educ1}
series x1b = x1*beta
series x0b = x0*beta
genr Mg_educ = mean(x1b-x0b)
```

Generated scalar Mg_educ = 0.0700091

Efectos de los salarios observados

- El efecto marginal anterior está en la expectativa incondicional de los salarios.
- Esta es la noción relevante si lo que queremos es el efecto en las ofertas salariales.
- Si queremos calcular el efecto sobre los salarios observados promedio, tenemos que restringir la muestra a las ofertas salariales observadas (los salarios de los individuos que trabajan).
- La mejor manera de hacer esto es mediante el uso de la expresión analítica de la esperanza condicional:

$$E[w | x, z, s = 1] = x\beta + \rho\lambda(z\gamma)$$

Resumen

- gret1 permite la estimación del modelo de selección de Heckman.
- Tanto con el procedimiento en dos etapas como con la estimación MCO.
- El contraste y la predicción se computa como de costumbre.