

Errores de Especificacion: Precision y Sesgo

Walter Sosa-Escudero

Econ 507. Econometric Analysis. Spring 2009

May 10, 2009

Modelo linear:

- 1 Linealidad: $Y = X\beta + u$.
- 2 Exogeneidad: $E(u) = 0$
- 3 No Multicolinealidad: $\rho(X) = K$
- 4 No heteroscedasticidad/correlacion serial: $V(u|X) = \sigma^2 I_n$.

Gauss/Markov: $\hat{\beta} = (X'X)^{-1}X'Y$ es el mejor estimador lineal insesgado.

Result: $\hat{\beta}$ es *insesgado*, esto es $E(\hat{\beta}) = \beta$.

Prueba:

$$\begin{aligned}
 \hat{\beta} &= (X'X)^{-1}X'Y \\
 &= (X'X)^{-1}X'(X\beta + u) \\
 &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \\
 &= \beta + (X'X)^{-1}X'u \\
 E(\hat{\beta}) &= \beta + E((X'X)^{-1}X'u) \\
 &= \beta + (X'X)^{-1}X'E(u) \\
 &= \beta
 \end{aligned}$$

Cual es el supuesto 'mas importante' para esta prueba?

Errores de especificacion, sesgo e imprecision

Hasta ahora hemos considerado al modelo lineal $Y = X\beta + u$ como 'correcto'

Consideremos el siguiente caso

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

Valen todos los supuestos clasicos. K_1 y K_2 son las columnas de X_1 and X_2 . Trivialmente, el modelo original se corresponde con $X = [X_1 \ X_2]$, with $K = K_1 + K_2$.

Consideremos los siguientes escenarios posibles en relacion a β_2 y las correspondientes estrategias de estimacion:

- **Omision de variables relevantes:** $\beta_2 \neq 0$, pero procedemos, erroneamente, como si $\beta_2 = 0$, esto es, regresamos Y en X_1 solamente.
- **Inclusion de variables irrelevantes:** $\beta_2 = 0$, pero procedemos erroneamente como si β_2 pudiese ser $\neq 0$, esto es, regresamos Y en X_1 y X_2 , cuando podriamos haber ignorado X_2 .

Sesgos

Comparemos los resultados de la estimacion de β_1 en ambos escenarios.

I) Omision de variables relevantes

Primero, notar que en este caso

$$Y = X_1\beta_1 + u^*$$

con $u^* = X_2\beta_2 + u$.

Sea $\hat{\beta}_1^* = (X_1'X_1)^{-1}X_1'Y$.

Este es el estimador de MCO cuando regresamos Y on X_1 *solamente* (omitimos X_2 .)

Resultado: $\hat{\beta}_1^*$ sera *sesgado* a menos que $X_2'X_1 = 0$.

$X_2'X_1 = 0$ implica que X_2 y X_1 no estan correlacionados.

En palabras: la omision de variables relevantes ($\beta_2 \neq 0$) conduce a sesgos en la estimacion, a menos que las variables omitidas no esten correlacionadas con las incluidas.

Prueba:

$$\begin{aligned}\hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'Y \\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'u \\ E(\hat{\beta}_1) &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'E(u) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2\end{aligned}$$

Esta ultima expresion se reduce a β_1 si $X_1'X_2 = 0$.

II) Inclusion de Variables Irrelevantes

En este caso estimariamos β_1 en conjunto con β_2 , regresando Y en X_1 y X_2 , esto es, $\hat{\beta}_1$ es un subvector de

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'X)^{-1}X'Y$$

Es importante ver que bajos los supuestos clasicos $\hat{\beta}$ y en consecuencia $\hat{\beta}_1$ seran **insesgados**. *Porque?*

Varianzas

Resultado: el estimador que omite X_2 tiene menor varianza.

Trade-off sesgo y varianza

En sintesis:

- En la practica no sabemos cual es el verdadero modelo (el chico o el grande).
- Trade-off: estimar un modelo chico (omitir variables) implica una ganancia en precision a riesgo de un sesgo. Un modelo grande es menor probable que sea sesgado pero sera mas ineficiente.
- La omision de variables no necesariamente induce a sesgos.

Sesgo por omision de variables: un ejemplo

Datos generados por computadora, pero en base a Appleton, French and Vanderpump ("Ignoring a Covariate: an Example of Simpson's Paradox", The American Statistician, 50, 4, 1996)

- Y = riesgo de muerte.
- $SMOKE$ = consumo de cigarrillos.

```
. reg y smoke
```

Source	SS	df	MS
Model	7613.25147	1	7613.25147
Residual	3839.18734	98	39.1753811
Total	11452.4388	99	115.6812

Number of obs = 100
 F(1, 98) = 194.34
 Prob > F = 0.0000
 R-squared = 0.6648
 Adj R-squared = 0.6614
 Root MSE = 6.259

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
smoke	-1.819348	.1305081	-13.94	0.000	-2.078337	-1.560359
_cons	158.5975	4.774249	33.22	0.000	149.1231	168.0718

. reg y smoke age

Source	SS	df	MS	Number of obs = 100		
Model	11350.9524	2	5675.47622	F(2, 97) = 5424.58		
Residual	101.486373	97	1.04625126	Prob > F = 0.0000		
Total	11452.4388	99	115.6812	R-squared = 0.9911		
				Adj R-squared = 0.9910		
				Root MSE = 1.0229		

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
smoke	.9431267	.050902	18.53	0.000	.8421004	1.044153
age	.9804631	.0164039	59.77	0.000	.9479059	1.01302
_cons	12.84084	2.560392	5.02	0.000	7.759169	17.92251

. cor y smoke age
(obs=100)

	y	smoke	age
y	1.0000		
smoke	-0.8153	1.0000	
age	0.9797	-0.9080	1.0000