Anatomia de MCO

Walter Sosa-Escudero

Universisad de San Andres, UNLP y CONICET

Anatomia de MCO

- $\hat{\beta} = (X'X)^{-1}X'Y$
- Mejor: Gauss/Markov
- Bueno: sesgo varianza

Que hace a MCO: a) sesgado b) impreciso?

Fuente de sesgos

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

- X_1, X_2, dos variables o dos grupos de variables.
- Interes en β_1. Regresar Y en X_1 y X_2.
- Supongamos que regresamos Y en solo X_1 omitiendo X_2:

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

Omision de variables relevantes

Resultado (Sesgo por variable omitida):

$$E(\hat{\beta}_1^*|X) = \beta_1 + \underbrace{\delta_{21}\ \beta_2}_{\text{sesgo}},$$

$\delta_{21} = (X_1' X_1)^{-1} X_1' X_2$

- δ_{21}?
- Sesgo: δ_{21} y $\beta_2 \neq 0$
- Sesgo por omision de variables: omitir variables relevantes ($\beta_2 \neq 0$) y correlacionadas con las incluidas ($\delta_{21} \neq 0$) sesga al estimador de MCO.

Tatuense esto un brazo.

Prueba:

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

Idea clave: lo que dejemos fuera del modelo pertenece al termino de error.

Fuentes de imprecision

Imprecision: varianza alta. Resultado: si $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$,

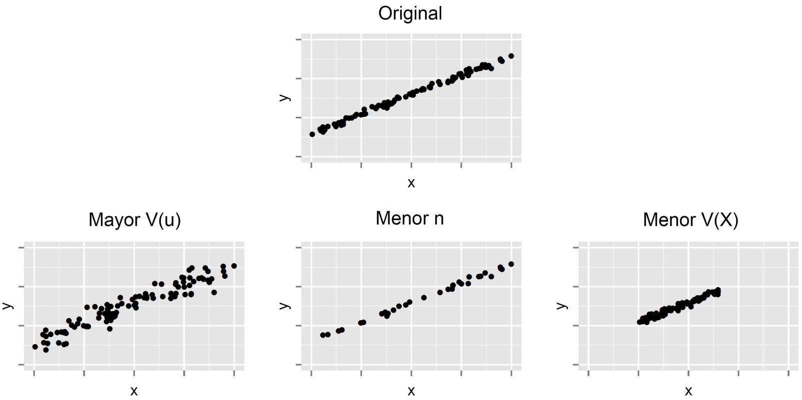
$$V(\hat{\beta}_j) = \frac{\sigma^2}{n\,(1 - R_j^2)\,V(X_j)},$$

Para un modelo dado, cuatro factores contribuyen a la varianza:

1. σ^2: ignorancia.
2. $V(X_j)$: variabilidad de X_j.
3. n: tamaño de muestra
4. R_j^2: multicolinealidad.

Tatuense esto en un brazo. Dejen lugar para una mas... Dejaremos la demostracion para un curso de posgrado.

Fuentes de impresicion: Dos variables



Multicolinealidad

$$V(\hat{\beta}_j) = \frac{\sigma^2}{n\,(1 - R_j^2)\,V(X_j)}$$

Con K variables, fuente adicional de imprecision: multicolinealidad.

- R_j^2: R^2 de regresar la $j-$esima variable en todas las otras variables explicativas.
- Multicolinealidad exacta: $R_j^2 = 1$. Viola los supuestos clasicos ($\rho(X) \neq K$). $\hat{\beta}$ no existe.
- Multicolinealidad alta: R_j^2 'alto'.
- Cuestion de grado: mayor multicolinealidad, mayor varianza

$$V(\hat{\beta}_j) = \frac{\sigma^2}{n\,(1 - R_j^2)V(X_j)}$$

- Multicolinealidad alta: modelo poco claro. Muchas variables
- Mayor n *compensa* multicolinealidad, pero no la resuelve.
- Sacar variables? Peligroso, como veremos a continuacion.

Inclusion de variables irrelevantes

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

Interes en β_1. Que pasa si sabemos que $\beta_2 = 0$ y asi y todo incluimos X_2?: inclusion de variables irrelevantes.

- $\hat{\beta}$, MCO de regresar Y en X_1 y X_2 (incluye variables irrelevantes).
- $\hat{\beta}$ es insesgado. Bajo supuestos clasicos, $E(\hat{\beta}) = \beta)$ *para todo* β, no importa que uno de sus componentes valga cero.
- $\hat{\beta}$ es ineficiente. Por el Teorema de Gauss Markov, el insesgado y eficiente es el que regresa Y en X_1 solamente (basado en el verdadero modelo, ahora con $\beta_2 = 0$)

- Cuidado: la inclusion de variables irrelevantes no aparece en nuestra formula de factores que afectan a la varianza. Porque esta refiere a un modelo *dado*.
- La ineficiencia de incluir variables irrelevantes se sigue del Teorema de Gauss/Markov para el modelo correcto.

En sintesis

Omitir variables relevantes:

- Sesga al estimador de MCO, si las variables omitidas son relevantes y correlacionadas con las incluidas.

Incluir variables irrelevantes:

- MCO que incluye X_1 y X_2, pero es ineficiente (mayor varianza)

Modelos grandes vs chicos

$Y = X_1\beta_1 + X_2\beta_2 + u$, Interes en β_1. Que hacer con X_2?

- $\beta_2 = 0$. Omitir. Por que?
- $\beta_2 \neq 0$
  - $\delta_{21} = 0$, se puede omitir X_2.
  - $\delta_{21} \neq 0$, incluir X_2 para evitar sesgo.

En la practica, incluir X_2 o no?

Trade-off sesgo/varianza: incluir variables reduce el sesgo pero aumenta la varianza, y viceversa. Modelos chicos, tienden a ser eficientes pero sesgados.

Sesgo por variables omitidas: un ejemplo

Datos ficticios, compatibles con Appleton, French and Vanderpump ("Ignoring a Covariate: an Example of Simpon's Paradox", The American Statistician, 50, 4, 1996)

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

```
. reg y smoke

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  194.34
       Model |  7613.25147     1  7613.25147           Prob > F      =  0.0000
    Residual |  3839.18734    98  39.1753811           R-squared     =  0.6648
-------------+------------------------------           Adj R-squared =  0.6614
       Total |  11452.4388    99    115.6812           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
-------------+------------------------------           F(  2,    97) = 5424.58
       Model |  11350.9524     2  5675.47622           Prob > F      =  0.0000
    Residual |  101.486373    97  1.04625126           R-squared     =  0.9911
-------------+------------------------------           Adj R-squared =  0.9910
       Total |  11452.4388    99    115.6812           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
```

Toda la econometria basica en 3 formulas

1. MCO: $\hat{\beta} = (X'X)^{-1}X'Y$
2. Varianza: $V(\hat{\beta}_j) = \frac{\sigma^2}{n\ (1-R_j^2)V(X_j)}$
3. Sesgo: $E(\hat{\beta}_1^*) = \beta_1 + \delta_{21}\ \beta_2$