Econometría

Diplomado Banco Central de Honduras

Instituto de Economía

Pontificia Universidad Católica de Chile

Juan Ignacio Urquiza – Junio 2022

Modelo de Regresión Lineal Clásico

- <u>Supuestos</u>:
  - Linealidad en parámetros (RLM.1).
  - Muestreo aleatorio (RLM.2).
  - Colinealidad imperfecta (RLM.3).
  - Media condicional cero (RLM.4).
  - Homocedasticidad (RLM.5).
- En la clase 2 demostramos que:
  - Bajo los supuestos RLM.1 a RLM.4, los estimadores de MCO son insesgados.
  - Bajo los supuestos RLM.1 a RLM.5, los estimadores de MCO son MELI – Teorema de Gauss-Markov.

Errores de especificación

- En la clase anterior vimos qué pasaba cuando se levantaba el supuesto (RLM.2) de muestreo aleatorio y el supuesto (RLM.5) de homocedasticidad.
- Ahora veremos qué ocurre cuando cometemos algún error de especificación.
- Consideraremos 3 tipos de errores de especificación:
  - Omisión de variables relevantes.
  - Inclusión de variables irrelevantes.
  - Errores de medición.
- Veremos que en algunos casos esto implica un desvío respecto del supuesto (RLM.4) de media condicional nula tal que los estimadores de MCO dejan de ser insesgados.

Omisión de variables relevantes

- Considere el siguiente MRL múltiple:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Sin embargo, imagine que por alguna razón (falta de datos, ignorancia, inobservabilidad) trabajamos con un modelo de regresión que no incluye a la variable explicativa X_2:

$$y = \beta_0 + \beta_1 X_1 + u$$

- Decimos que se ha omitido una variable relevante (en este caso X_2) cuando $\beta_2 \neq 0$.
- ¿Qué consecuencias tiene dicha omisión?
  - Al omitir X_2, su efecto pasa a formar parte del término del error u.

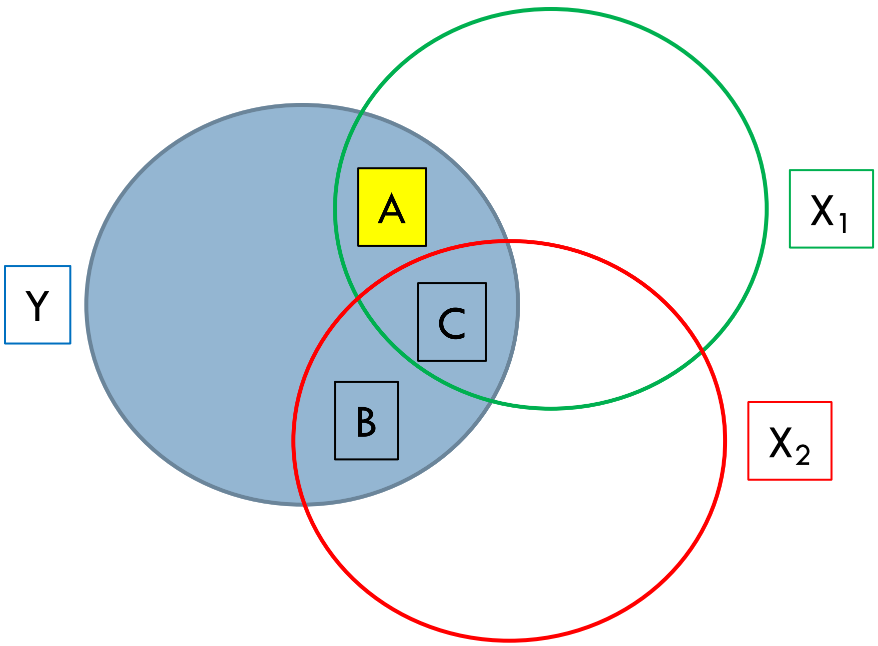
Omisión de variables relevantes

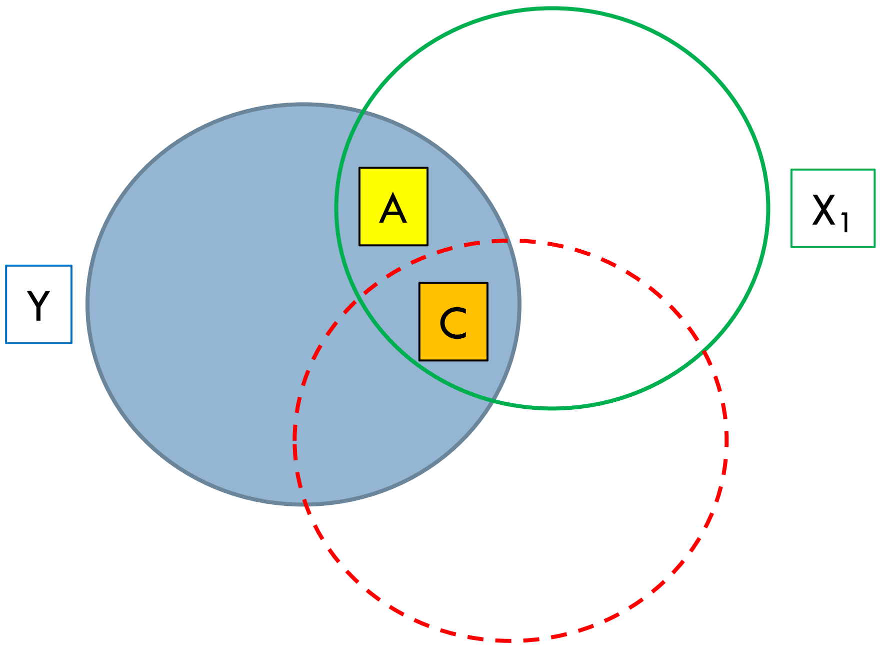
- Entonces:

$$\begin{aligned} E(u|X_1) &= E(\varepsilon + \beta_2 X_2 | X_1) \\ &= E(\varepsilon|X_1) + E(\beta_2 X_2|X_1) \\ &= E\big[E(\varepsilon|X_1, X_2)\big|X_1\big] + \beta_2 \times E(X_2|X_1) \\ &= \beta_2 \times E(X_2|X_1) \neq 0 \end{aligned}$$

- Por lo tanto, vemos que la omisión de variables relevantes conduce a sesgos en la estimación, a menos que las variables omitidas no estén correlacionadas con las variables incluidas.
- El coeficiente de X_1 no recoge el efecto *ceteris paribus* de un cambio en X_1 sobre Y, sino que recoge el efecto de un cambio en X_1 más un efecto indirecto de X_1 sobre X_2.

Gráficamente





Omisión de variables relevantes

- En el caso simple, podemos resumir el sesgo en la estimación de β_1 cuando se omite X_2 mediante la siguiente tabla:

	$C(X_1, X_2) > 0$	$C(X_1, X_2) < 0$
$\beta_2 > 0$	$+$	$-$
$\beta_2 < 0$	$-$	$+$

- En particular, se puede demostrar que:

$$\widetilde{\beta_1} = \widehat{\beta_1} + \widehat{\beta_2} \times \widehat{\delta_1}$$

donde $\widetilde{\beta_1}$ es la estimación de β_1 cuando se omite X_2, $\widehat{\beta_1}$ es la estimación de β_1 cuando se incluye X_2, y $\widehat{\delta_1}$ es la estimación de la pendiente en la regresión de X_2 sobre X_1.

```
. reg testscr str

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     22.58
       Model |  7794.11004         1  7794.11004   Prob > F        =    0.0000
    Residual |  144315.484       418  345.252353   R-squared       =    0.0512
-------------+----------------------------------   Adj R-squared   =    0.0490
       Total |  152109.594       419  363.030056   Root MSE        =    18.581

------------------------------------------------------------------------------
     testscr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .4798256    -4.75   0.000     -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000     680.3231    717.5428
------------------------------------------------------------------------------
```

```
. reg testscr str el_pct

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(2, 417)       =    155.01
       Model |  64864.3011         2  32432.1506   Prob > F        =    0.0000
    Residual |  87245.2925       417  209.221325   R-squared       =    0.4264
-------------+----------------------------------   Adj R-squared   =    0.4237
       Total |  152109.594       419  363.030056   Root MSE        =    14.464

------------------------------------------------------------------------------
     testscr | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .3802783    -2.90   0.004    -1.848797   -.3537945
      el_pct |  -.6497768   .0393425   -16.52   0.000    -.7271112   -.5724423
       _cons |   686.0322   7.411312    92.57   0.000     671.4641    700.6004
------------------------------------------------------------------------------
```

```
. reg el_pct str

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =     15.25
       Model |  4932.98526         1  4932.98526   Prob > F        =    0.0001
    Residual |    135170.2       418  323.373684   R-squared       =    0.0352
-------------+----------------------------------   Adj R-squared   =    0.0329
       Total |  140103.185       419  334.375143   Root MSE        =    17.983

------------------------------------------------------------------------------
      el_pct | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         str |   1.813719   .4643735     3.91   0.000     .9009206    2.726517
       _cons |  -19.85405   9.162604    -2.17   0.031    -37.86458   -1.843531
------------------------------------------------------------------------------
```

$$\rightarrow \widetilde{\beta_1} = \widehat{\beta_1} + \widehat{\beta_2} \times \widehat{\delta_1}$$

$$-2.2798085 = -1.101296 - 0.6497768 \times 1.813719$$

Inclusión de variables irrelevantes

- Se dice que un modelo está sobre-especificado cuando incluye variables que no forman parte del modelo poblacional.
- Considere el siguiente MRL múltiple:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + u$$

  que satisface todos los supuestos de MCO, pero que en la población se cumple que $\beta_{k+1} = 0$.
- Sabemos que MCO será insesgado: $E\left(\widehat{\beta_{k+1}} \middle| \boldsymbol{X}\right) = \beta_{k+1} = 0$.
- Por lo tanto, la inclusión de una o más variables irrelevantes no influye sobre el insesgamiento de MCO, pero posiblemente afecte sus varianzas a causa de la multicolinealidad.
- Cuánto más correlacionada esté la v. irrelevante con las v. relevantes incluidas, mayor será la pérdida de precisión.

Errores de medición

- Hasta ahora hemos considerado casos en que las variables X e Y siempre estaban disponibles.
- Sin embargo, en muchas aplicaciones, no es posible contar con datos sobre la variable económica que realmente nos interesa.
- Por ejemplo:
  - Ingreso permanente vs. ingreso disponible.
  - Ingreso verdadero vs. ingreso declarado.
  - Calorías consumidas vs. calorías compradas.
- Hablaremos de errores de medición en aquellos casos en que se utilice una medida imprecisa de la variable de interés.
- ¿Qué consecuencias tiene esto sobre los estimadores de MCO?

Variable dependiente

- Considere el siguiente MRL simple:

$$y^* = \beta_0 + \beta_1 x + u$$

- La variable que nos interesa es y^* pero sólo contamos con una medida imprecisa (y) tal que:

$$y = y^* + e_0$$

$$\rightarrow e_0 = y - y^*$$

donde e_0 es el error de medición.

- ¿Qué pasa si estimamos la regresión con y en lugar de y^*?
- La propiedades de MCO dependerán de la relación entre e_0 y las variables explicativas.

Variable dependiente

- Para obtener el modelo estimable se reemplaza para y^*:

$$y^* = y - e_0 = \beta_0 + \beta_1 x + u$$

$$\rightarrow y = \beta_0 + \beta_1 x + v$$

donde $v = (u + e_0)$.

- Sabemos que:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\left(\frac{1}{n}\right) \times \sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\left(\frac{1}{n}\right) \times \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\rightarrow plim\left(\widehat{\beta_1}\right) = \frac{plim\left[\left(\frac{1}{n}\right) \times \sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})\right]}{plim\left[\left(\frac{1}{n}\right) \times \sum_{i=1}^{n}(x_i - \bar{x})^2\right]}$$

Variable dependiente

- Para obtener el modelo estimable se reemplaza para y^*:

$$y^* = y - e_0 = \beta_0 + \beta_1 x + u$$
$$\rightarrow y = \beta_0 + \beta_1 x + v$$

donde $v = (u + e_0)$.

- Por LGN, se puede demostrar que:

$$plim\left(\widehat{\beta_1}\right) = \frac{plim\left[\left(\frac{1}{n}\right) \times \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]}{plim\left[\left(\frac{1}{n}\right) \times \sum_{i=1}^{n}(x_i - \bar{x})^2\right]} = \frac{C(x, y)}{V(x)}$$

$$= \frac{C(x, y^* + e_0)}{V(x)} = \frac{C(x, y^*)}{V(x)} + \frac{C(x, e_0)}{V(x)}$$

Variable dependiente

- Por lo tanto, se puede demostrar que:

$$\rightarrow plim\left(\widehat{\beta_1}\right) = \frac{C(x, y^*)}{V(x)} + \frac{C(x, e_0)}{V(x)} = \beta_1 + \frac{C(x, e_0)}{V(x)}$$

- En resumen, el error de medición en la v. dependiente puede causar sesgo e inconsistencia en MCO cuando esté relacionado sistemáticamente con una o más v. explicativas.
- Si el error de medición es sólo un error aleatorio, asociado al reporte de los datos, que es independiente de las variables explicativas, entonces MCO es perfectamente apropiado.
- Sin embargo, se traduce en una mayor varianza del error, lo que resulta en varianzas mayores de los estimadores de MCO.

Variable independiente

- Vuelva a considerar el MRL simple:

$$y = \beta_0 + \beta_1 x^* + u$$

- Ahora la variable que nos interesa es x^* pero sólo contamos con una medida imprecisa (x) tal que:

$$x = x^* + e_1$$

$$\rightarrow e_1 = x - x^*$$

donde e_1 es el error de medición.

- Hablaremos de errores clásicos en variables (ECV) cuando:

$$E(e_1) = 0 \quad , \quad C(e_1, u) = 0 \quad , \quad C(e_1, x^*) = 0$$

- ¿Qué pasa si estimamos la regresión con x en lugar de x^*?

Variable independiente

- Para obtener el modelo estimable se reemplaza para x^*:

$$y = \beta_0 + \beta_1(x - e_1) + u$$

$$\rightarrow y = \beta_0 + \beta_1 x + w$$

donde $w = (u - \beta_1 e_1)$.

- Al igual que antes, se puede demostrar que:

$$plim\left(\widehat{\beta_1}\right) = \frac{C(x,y)}{V(x)} = \frac{C(x^* + e_1, y)}{V(x^* + e_1)}$$

$$= \frac{C(x^*, y) + C(e_1, y)}{V(x^*) + V(e_1)} = \frac{C(x^*, y)}{V(x^*) + V(e_1)}$$

Variable independiente

- Por lo tanto, tenemos que:

$$plim\left(\widehat{\beta_1}\right) = \frac{C(x^*, y)}{V(x^*) + V(e_1)} = \frac{\beta_1 \times V(x^*)}{V(x^*) + V(e_1)}$$

$$\rightarrow plim\left(\widehat{\beta_1}\right) = \beta_1 \times \lambda \neq \beta_1$$

donde $\lambda = \frac{V(x^*)}{V(x^*) + V(e_1)} < 1$.

- En resumen, el error de medición en una variable explicativa nos lleva a subestimar (en valor absoluto) el efecto de dicha variable – sesgo de atenuación.
- La magnitud del sesgo dependerá de cuán grande sea $V(x^*)$ en comparación con $V(e_1)$.

Conclusiones

- La omisión de v. relevantes conduce a sesgos en la estimación, salvo en el caso en que no estén correlacionadas con las v. incluidas.
- La inclusión de v. irrelevantes no genera sesgos pero implica una pérdida de precisión.
- Bajo el supuesto de errores clásicos, el error de medición en la variable dependiente no genera sesgos pero sí aumentos en la varianza de los estimadores.
- Por el contrario, el error de medición en una v. independiente genera sesgo de atenuación tal que el estimador de MCO del coeficiente sobre dicha variable estará sesgado hacia cero.