Multicolinealidad y Variables Cualitativas

Fundamentos de Econometría

Juan Palomino[1]

[1]Magister en Economía Aplicada con Mención Estudios Regionales
juan.palominoh@pucp.pe

Departamento de Economía

PUCP

Índice

Índice

Multicolinealidad perfecta

- **Definición:** una de las variables explicativas es una combinación lineal exacta de las otras explicativas.
- Por ejemplo, dado el modelo poblacional:

$$PBI_i = \beta_0 + \beta_1 export + \beta_2 import_i + \beta_3 balance + \varepsilon_i$$

- Ocurrira multicolinealidad perfecta si podemos expresar $balance = export - import$.
- En este caso β_1 no podrá medir el efecto de incrementar $balance$ manteniendo constante $export$ e $import$ dada su relación lineal.

Multicolinealidad perfecta

- No podemos encontrar de forma única $\hat{\beta}_{OLS} = (X^{'}X)^{-1}X^{'}Y$

$$rango(X) < K \Rightarrow \left|X^{'}X\right| = 0 \Rightarrow \nexists (X^{'}X)^{-1}$$

- ¿Cómo detectarlo?
    - Los programas se quejaran de que no podemos invertir la matriz $(X^{'}X)$
- ¿Cómo corregirlo?
    - Se deben a errores del investigador al introducir las explicativas
    - Al aparecer mensaje de error, corregiremos las explicativas.

Multicolinealidad perfecta

- Corrección: en el ejemplo de balanza comercial sabemos: $balance = export - import$

$$PBI_i = \beta_0 + \beta_1 export + \beta_2 import_i + \beta_3 balance + \varepsilon_i$$

- Excluimos la variable *balance*:

$$PBI_i = \beta_0 + \beta_1 export + \beta_2 import_i + \varepsilon_i$$

Índice

Multicolinealidad Imperfecta

- **Definición:** la correlación entre las variables explicativas es alta, pero no perfecta.
- Por ejemplo:

$$Consumo_i = \beta_1 + \beta_2 Ingreso_i + \beta_3 Riqueza_i + \varepsilon_i$$

- No hay multicolinealidad perfecta, pero es probable que riqueza e ingresos esten altamente correlacionadas.

Multicolinealidad Imperfecta

- Otros ejemplos son:

$$Rendimiento_i = \beta_1 + \beta_2 NivEducPadre_i + \beta_3 NivEducMadre_i + \varepsilon_i$$

$$\begin{aligned} GastoTelefonia_i = \beta_1 + \beta_2 Miembros_i + \beta_3 GastoVivienda_i + \\ \beta_4 Ingresos_i + \beta_5 DisposInternet + \varepsilon_i \end{aligned}$$

Multicolinealidad Imperfecta

- Podemos encontrar de forma única $\hat{\beta}_{OLS} = (X^{'}X)^{-1}X^{'}Y$

$$rango(X) = K \Rightarrow \left|X^{'}X\right| \neq 0 \Rightarrow \exists(X^{'}X)^{-1}$$

- Estimadores cumplen las propiedades de MCO.
- ¿Qué problemas genera en la estimación?
  - Para entenderlo, consideremos un modelo de regresión lineal con k variables:

$$Y_i = \beta_1 + \beta_2 W_{2i} + \ldots + \beta_{k-1} W_{k-1i} + \beta_z Z_i + \varepsilon_i$$

  - Y definimos:

$$R_z^2 = 1 - \frac{SCR}{STC} = 1 - \frac{\varepsilon_z^{'}\varepsilon_z}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}$$

Multicolinealidad Imperfecta

- **¿Qué problemas genera en la estimación?**
  - Podemos demostrar por teorema de Frisch-Waugh que:

$$\hat{\beta}_Z = (z^{'}M_W z)^{-1} z^{'} M_W y$$

  - Entonces:

$$\begin{aligned} Var(\hat{\beta}_Z) &= \sigma^2 (z^{'} M_W z)^{-1} \\ &= \frac{\sigma^2}{z^{'} M_W^{'} M_W z} \end{aligned}$$

$$Var(\hat{\beta}_Z) = \frac{\sigma^2}{\hat{\varepsilon}_z^{'} \hat{\varepsilon}_z}$$

Multicolinealidad Imperfecta

- Sabiendo que $\hat{\varepsilon}_Z$ son los residuos de la regresión de Z_i contra las variables W. Sabiendo $\varepsilon_z^{'}\varepsilon_z$ tenemos que:

$$Var(\hat{\beta}_Z) = \frac{\sigma^2}{(1 - R_z^2)\sum_{i=1}^{n}(Z_i - \bar{Z})^2}$$

- Implicaciones:
  - Si $R_z^2 \to 1 \Rightarrow Var(\hat{\beta}_Z) \to \infty$
  - Si Z no se relacionara con las demás variables, entonces $R_z^2 = 0$.
  - Estimación imprecisa e intervalos de confianza muy grandes.

Multicolinealidad Imperfecta

- Pruebas estadísticas informales:
  - Análisis de la matriz de correlación
  - Analizar el R^2, su test F asociado y las pruebas t individuales.
    - Si el R^2 es alto y el test F asociado es estadísticamente significativo pero las pruebas t individuales son bajas, muy probablemente exista multicolinealidad.

Multicolinealidad Imperfecta

- Pruebas estadísticas formales:
  - Analizar el VIF (variance inflation factor): mide la redundancia que una variable inserta en el modelo en términos de la información que comparte con las otras variables:

$$
\begin{aligned}
Var(\hat{\beta}_Z) &= \frac{\sigma^2}{(1-R_z^2)\sum_{i=1}^{n}(Z_i-\bar{Z})^2} \\
&= \frac{1}{(1-R_z^2)}\frac{\sigma^2}{\sum_{i=1}^{n}(Z_i-\bar{Z})^2} \\
&= VIF\frac{\sigma^2}{\sum_{i=1}^{n}(Z_i-\bar{Z})^2}
\end{aligned}
$$

Multicolinealidad Imperfecta

- Existen dos reglas para interpretar el VIF:
  - Si el VIF de una variable es mayor a 10, podemos afirmar que esta inserta al modelo una considerable inflación de la varianza.
  - Si el promedio de los VIF es considerablemente mayor a 1, podemos afirmar que el modelo sufre un problema de multicolinealidad.

Índice

Variables Cualitativas

- Tenemos las siguientes variables cualitativas:

$$Sexo = \left\{ \begin{array}{c} Hombre \\ Mujer \end{array} \right\}$$

- Para la variable sexo podemos tener:

$$S_{1i} = \left\{ \begin{array}{c} 1 \ si \ es \ hombre \\ 0 \ si \ es \ mujer \end{array} \right\} \text{ o } S_{2i} = \left\{ \begin{array}{c} 1 \ si \ es \ mujer \\ 0 \ si \ es \ hombre \end{array} \right\}$$

Variables Cualitativas

- Supongamos que el modelo es:

$$ln(ingreso)_i = \beta_0 + \beta_1 educ_i + \varepsilon_i$$

- Incorporar la variable sexo:

$$ln(ingreso)_i = \beta_0 + \beta_1 educ_i + \beta_2 S_{1i} + \beta_3 S_{2i} + \varepsilon_i$$

- Este modelo no puede ser estimado. La manera correcta es:

$$ln(ingreso)_i = \beta_0 + \beta_1 educ_i + \beta_2 S_{1i} + \varepsilon_i$$

- En la función de regresión poblacional

$$E[ingreso_i \mid S_{1i} = 1, educ] = (\beta_0 + \beta_2) + \beta_1 educ_i$$
$$E[ingreso_i \mid S_{1i} = 0, educ] = \beta_0 + \beta_1 educ_i$$

Variables Cualitativas

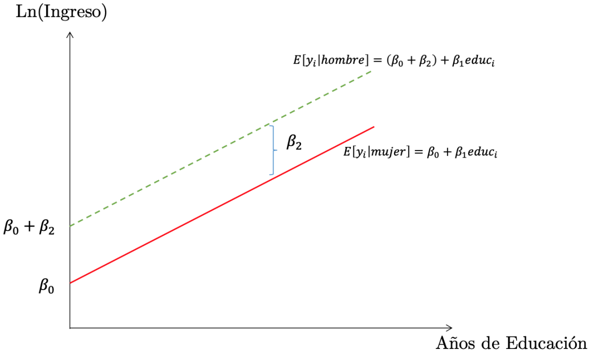

Figura: FRP de los hombres y mujeres con cambios en intercepto

Índice

Variables Cualitativas: Dummy Interactivas

- Supongamos que el modelo es:

$$ln(ingreso)_i = \beta_0 + \beta_1 educ_i + \beta_2 S_{1i} + \beta_3 (S_1 \times educ)_i + \varepsilon_i$$

- El valor esperado condicional es:

$$E[ingreso_i \mid S_{1i} = 1, educ] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) educ_i$$
$$E[ingreso_i \mid S_{1i} = 0, educ] = \beta_0 + \beta_1 educ_i$$

Variables Cualitativas: Dummy Interactivas

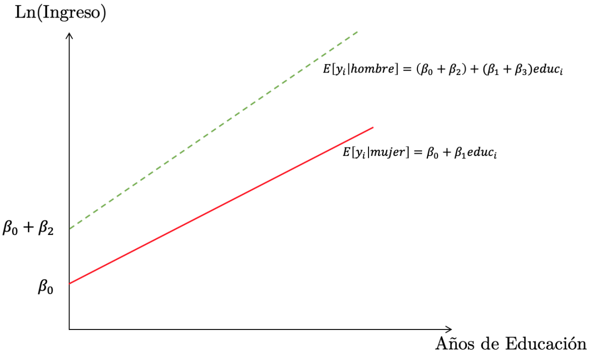

Figura: FRP de hombres y mujeres con cambios en intercepto y pendiente

Índice

Variables Cualitativas

- Tenemos las siguientes variables cualitativas:

$$\textit{Región} = \left\{ \begin{array}{c} Costa \\ Sierra \\ Selva \end{array} \right\}$$

- Para la variable región podemos tener:

$$R_{1i} = \left\{ \begin{array}{c} 1 \ si \ vive \ en \ la \ Costa \\ 0 \ en \ otro \ caso \end{array} \right\}$$

$$R_{2i} = \left\{ \begin{array}{c} 1 \ si \ vive \ en \ la \ Sierra \\ 0 \ en \ otro \ caso \end{array} \right\}$$

$$R_{3i} = \left\{ \begin{array}{c} 1 \ si \ vive \ en \ la \ Selva \\ 0 \ en \ otro \ caso \end{array} \right\}$$

Variables Cualitativas

- Supongamos que el modelo es:

$$ln(ingreso)_i = \beta_0 + \beta_1 educ_i + \beta_2 S_{1i} + \beta_3 R_{2i} + \beta_4 R_{3i} + \varepsilon_i$$

Variables Cualitativas

- Mujeres de la Costa:

$$E[ingreso_i \mid S_{1i}=0, R_{2i}=0, R_{3i}=0, educ] = \beta_0 + \beta_1 educ_i$$

- Hombres de la Costa:

$$E[ingreso_i \mid S_{1i}=1, R_{2i}=0, R_{3i}=0, educ] = (\beta_0 + \beta_2) + \beta_1 educ_i$$

- Mujeres de la Sierra:

$$E[ingreso_i \mid S_{1i}=0, R_{2i}=1, R_{3i}=0, educ] = (\beta_0 + \beta_3) + \beta_1 educ_i$$

- Hombres de la Sierra:

$$E[ingreso_i \mid S_{1i}=1, R_{2i}=1, R_{3i}=0, educ] = (\beta_0 + \beta_2 + \beta_3) + \beta_1 educ_i$$

- Mujeres de la Selva:

$$E[ingreso_i \mid S_{1i}=0, R_{2i}=0, R_{3i}=1, educ] = (\beta_0 + \beta_4) + \beta_1 educ_i$$

- Hombres de la Selva:

$$E[ingreso_i \mid S_{1i}=1, R_{2i}=0, R_{3i}=1, educ] = (\beta_0 + \beta_2 + \beta_4) + \beta_1 educ_i$$

Variables Cualitativas

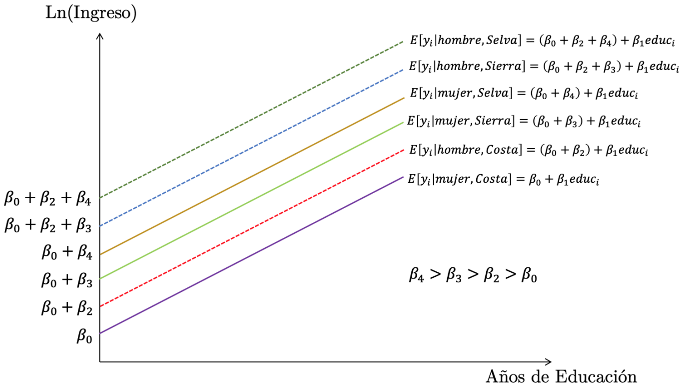

Figura: FRP de los hombres y mujeres en Regiones

Variables Cualitativas

- Supongamos que el modelo es:

$$ln(ingreso)_i = \beta_0 + \beta_1 educ_i + \beta_2 S_{1i} + \beta_3 R_{2i} + \beta_4 R_{3i} + \\ \beta_5 (S_1 \times R_2)_i + \beta_6 (S_1 \times R_3)_i + \varepsilon_i$$

Variables Cualitativas

- Diferencial Sierra-Costa (hombres)

$$
\begin{aligned}
E[y_i \mid S_{1i}=1, R_{2i}=1, R_{3i}=0, educ] - E[y_i \mid S_{1i}=1, R_{2i}=0, R_{3i}=0, educ]\\
(\beta_0+\beta_1 educ_i+\beta_2+\beta_3+\beta_5)-(\beta_0+\beta_1 educ_i+\beta_2)\\
=\beta_3+\beta_5
\end{aligned}
$$

- Diferencial Sierra-Costa (mujeres)

$$
\begin{aligned}
E[y_i \mid S_{1i}=0, R_{2i}=1, R_{3i}=0, educ] - E[y_i \mid S_{1i}=0, R_{2i}=0, R_{3i}=0, educ]\\
(\beta_0+\beta_1 educ_i+\beta_3)-(\beta_0+\beta_1 educ_i)\\
=\beta_3
\end{aligned}
$$

Referencias

- **Capítulo 9 y 10.** Gujarati, D., & Porter, D. (2010). Econometría (Quinta edición ed.). & P. Carril Villareal, Trad.) México: Mc Graw Hill educación.
- **Capítulo 7.** Wooldridge, J. M. (2006). Introducción a la Econometría. 4 edición. Un enfoque moderno. México: Editorial Thomson-Learning. 849 p.