Lección 2: Análisis de regresión múltiple

Módulo 4: Introducción al análisis de regresión

Magdalena Cornejo

Variables Omitidas

- Al estimar un modelo de regresión simple, hay muchas variables que pueden afectar a y que no están siendo evaluadas.
- Eventualmente, podríamos encontrarnos en el problema de **sesgo por variables omitidas** si omitimos una variable que está correlacionada con X.
- Un modelo de regresión múltiple es un caso más realista (e interesante).

Regresión múltiple

El **modelo de regresión múltiple** es una extensión del simple adicionando más regresores:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$$

donde ahora hay k variables explicativas y $k + 1$ parámetros a estimar.

El modelo es lineal en los β (coeficientes de regresión), que son los parámetros desconocidos y que estimaremos por MCO para obtener:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}$$

Interpretación de los coeficientes

- Si X_k es **continua**:

$$\beta_k = \frac{\partial E(y/X)}{\partial X_k}$$

  Por lo que nos da el cambio marginal en el valor esperado de y cuando cambia marginalmente X_k, permaneciendo constantes el resto de las variables explicativas.

- Si X_k es **binaria** (*dummy*):

$$\beta_k = E(Y/X_k = 1) - E(Y/X_k = 0)$$

Ejemplo

Imaginemos que queremos explicar el precio de los vinos de alta gama en el mercado:

- y = precio del vino (medido en \$)
- X = antigüedad del vino (medido en años)
- $D = 1$ si pasó por barrica de roble, $= 0$ en caso contrario.

Una muestra de vinos nos permitió estimar el siguiente modelo:

$$\hat{y}_i = 110 + 25X_i + 30D_i$$

Si las variables de la regresión (X y D) resultaron estadísticamente significativas, **¿cómo se interpretan dichos coeficientes?**

Variables explicativas categóricas

- Si X_k es **categórica** (**nominal** y **ordinal**). Por ejemplo:

$$(\text{NSE})\ X = \begin{cases} 1\ (\text{ABC1}) \\ 2\ (\text{C2C3}) \\ 3\ (\text{DE}) \end{cases}$$

Se puede armar un conjunto de *dummies*:

- $D_1 = 1$ si ABC1, 0 otro caso.
- $D_2 = 1$ si C2C3, 0 otro caso.

Si y es el consumo de un producto (en unidades) y estimamos por MCO:

$$\hat{y}_i = 3 + 5D_{1i} + 4D_{2i}$$

- Si pertenece al segmento ABC1: $\hat{y}_i = 3 + 5 = 8$
- Si pertenece al segmento C2C3: $\hat{y}_i = 3 + 4 = 7$
- Si pertenece al segmento DE: $\hat{y}_i = 3$

Bondad de ajuste

- El R^2 tiene la interpretación usual (si el modelo incluye constante).
- Pero tiene un **gran problema**: tiende a aumentar en la medida en que incorporamos más regresores.

¿Qué pasa si tenemos dos o más modelos alternativos para explicar a y, pero uno tiene más regresores que otro? ¿Con cuál me quedo?

Bondad de ajuste

- El R^2 tiene la interpretación usual (si el modelo incluye constante).
- Pero tiene un **gran problema**: tiende a aumentar en la medida en que incorporamos más regresores.

¿Qué pasa si tenemos dos o más modelos alternativos para explicar a y, pero uno tiene más regresores que otro? ¿Con cuál me quedo?

R^2 **ajustado**:

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\sum_i e_i^2}{\sum_i (\hat{y}_i - \overline{y})}$$

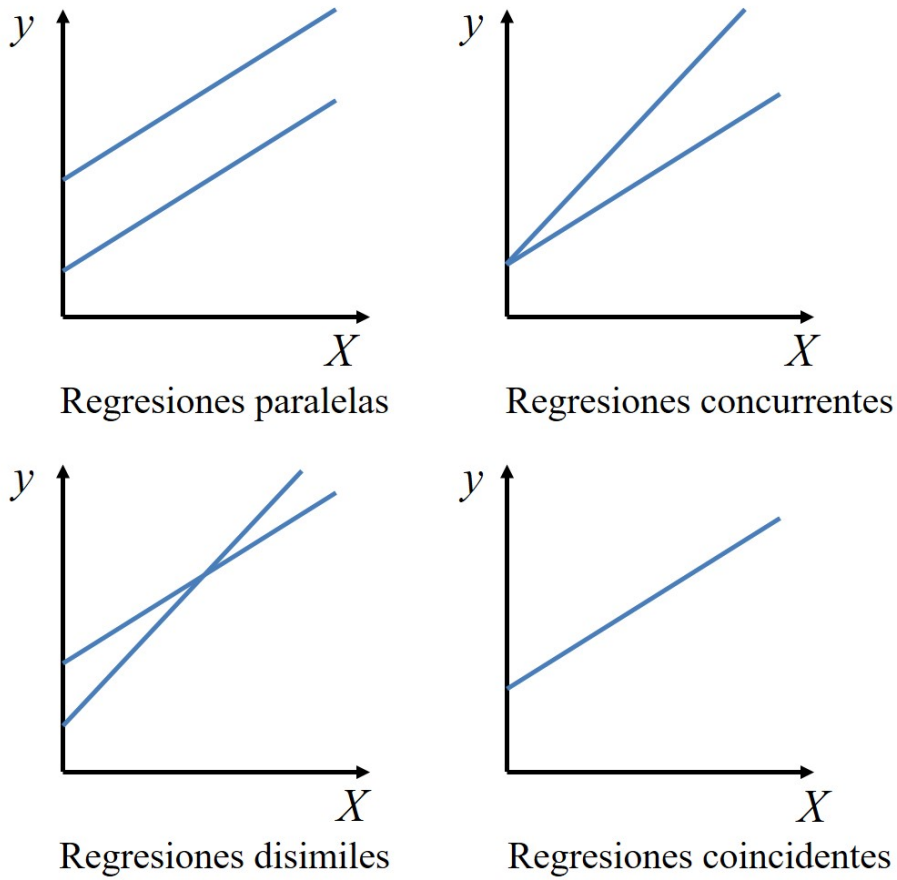
Observar que:

- El ratio $(n-1)/(n-k-1)$ es siempre mayor a 1 (en regresión múltiple), entonces el R^2 ajustado es menor que el R^2.
- Agregar un regresor tiene dos efectos contrapuestos:
  - Cae la $\sum e_i^2$, aumentando el R^2 ajustado.
  - Aumenta el factor $(n-1)/(n-k-1)$, disminuyendo el R^2 ajustado.

Interacción entre una variable continua y otra binaria

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + \varepsilon_i$$

El parámetro β_2 produce cambios en la ordenada al origen, mientras que el parámetro β_3 produce cambios en la pendiente.



Aplicación

¿Cuántas horas de fútbol ve a la semana?

Se ha estimado el siguiente modelo (suponga que todas las variables on significativas):

$$\hat{y}_i = 1 - 0.5X_i + 2D_i + 0.8D_iX_i$$

- y es el número de horas de fútbol que un individuo ve en televisión a la semana.
- X es el número de canales de televisión al que se tiene acceso.
- D es una variable binaria (=1 si es hombre y 0 si es mujer).

Entonces, si:

- $D = 0$ (mujer) $\hat{y}_i = 1 - 0.5X_i$
- $D = 1$ (hombre) $\hat{y}_i = 3 + 0.3X_i$

Modelos en logs

- Permite una interpretación económica clara de los coeficientes.
- Pero no es gratis, ya que se asume cierta forma funcional.
- NO aplica si X es binaria o toma valores negativos.

Modelo	Variable Dependiente	Variable Explicativa	Interpretación de β_1
lin-lin	y	x	$\Delta y = \beta_1 \Delta x$
lin-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
log-lin	$\log(y)$	x	$\%\Delta y = (100\beta_1)\,\Delta x$
log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

$$\widehat{log(salario)} = 4.822 + 0.257 log(ventas)$$

El coeficiente del logaritmo de las ventas es la elasticidad estimada del salario con respecto a las ventas.

Términos cuadráticos

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

- Nos indica la concavidad o convexidad del efecto.
- Existe un **punto de inflexión**.

Advertencia: cambia la interpretación ya que ahora β_1 no es el efecto parcial de X_k en y. Se requiere imputar algún X para conocer el efecto (ahora el efecto depende de X, antes no):

$$\frac{\partial E(y/X)}{\partial X_i} = \beta_1 + 2\beta_2 X_{1i}$$

Términos cuadráticos