Laboratorio 5 – Sesgo por variable omitida y multicolinealidad

Table of contents

El propósito de este laboratorio es **comprender mejor el sesgo por variable omitida** y la **multicolinealidad**. ## Para empezar

Abre un nuevo script de R y carga los paquetes

```
# Instalar paquetes solo si faltan (estilo ECO/EPG)
pkgs <- c("tidyverse", "broom", "wooldridge", "car")
to_install <- pkgs[!pkgs %in% rownames(installed.packages())]
if (length(to_install) > 0) {
  install.packages(to_install, repos = "http://cran.us.r-project.org")
}

library(tidyverse)
library(broom)
library(wooldridge)
library(car)        # para vif()
```

Nota: en el enunciado original se pedía instalar `car` "en la consola". Aquí lo dejamos **compatible y automático**, instalándolo solo si falta.

Cargar los datos

Usaremos un nuevo set de datos sobre salarios, llamado `wage2`.

```
df <- as_tibble(wage2)
```

Revisa qué contiene el set de datos con:

```
glimpse(df)
```

```
Rows: 935
Columns: 17
$ wage    <int> 769, 808, 825, 650, 562, 1400, 600, 1081, 1154, 1000, 930, 921~
$ hours   <int> 40, 50, 40, 40, 40, 40, 40, 40, 45, 40, 43, 38, 45, 38, 40, 50~
$ IQ      <int> 93, 119, 108, 96, 74, 116, 91, 114, 111, 95, 132, 102, 125, 11~
$ KWW     <int> 35, 41, 46, 32, 27, 43, 24, 50, 37, 44, 44, 45, 40, 24, 47, 37~
$ educ    <int> 12, 18, 14, 12, 11, 16, 10, 18, 15, 12, 18, 14, 15, 16, 16, 10~
$ exper   <int> 11, 11, 11, 13, 14, 14, 13, 8, 13, 16, 8, 9, 4, 7, 9, 17, 6, 1~
$ tenure  <int> 2, 16, 9, 7, 5, 2, 0, 14, 1, 16, 13, 11, 3, 2, 9, 2, 9, 10, 7,~
$ age     <int> 31, 37, 33, 32, 34, 35, 30, 38, 36, 36, 38, 33, 30, 28, 34, 35~
$ married <int> 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,~
$ black   <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ south   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ urban   <int> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1,~
$ sibs    <int> 1, 1, 1, 4, 10, 1, 1, 2, 2, 1, 1, 1, 2, 3, 1, 1, 3, 2, 3, 3, 0~
$ brthord <int> 2, NA, 2, 3, 6, 2, 2, 3, 3, 1, 1, 2, NA, 1, 1, 2, 3, 3, 1, 2, ~
$ meduc   <int> 8, 14, 14, 12, 6, 8, 8, 8, 14, 12, 13, 16, 12, 10, 12, 6, 12, ~
$ feduc   <int> 8, 14, 14, 12, 11, NA, 8, NA, 5, 11, 14, NA, 12, 10, 12, 8, 10~
$ lwage   <dbl> 6.645091, 6.694562, 6.715384, 6.476973, 6.331502, 7.244227, 6.~
```

Propiedades de las variables omitidas (Omitted Variable Bias)

Considera el siguiente modelo de regresión:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 IQ + u$$

donde `wage` es la **tasa salarial por hora** (en centavos, no en dólares).

Queremos verificar la propiedad mostrada en Wooldridge:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$$

donde:

- $\tilde{\beta}_1$ proviene de la regresión de $\log(wage)$ sobre `educ` (modelo "corto").
- $\tilde{\delta}_1$ proviene de la regresión de `IQ` sobre `educ`.
- $\hat{\beta}_1$ y $\hat{\beta}_2$ provienen del modelo "completo" (con `educ` e `IQ`).

1) Regresión de IQ sobre $educ$ (para obtener $\tilde{\delta}_1$)

```
est1 <- lm(IQ ~ educ, data = df)
```

** Modelo 1: IQ ~ educ **

Coeficientes estimados:

Estadísticas del modelo:

Estadístico	Valor
R^2	0.2659
R^2 ajustado	0.2652
Error estándar residual	12.9036
Estadístico F	338.02
N	935

2) Regresión de $\log(wage)$ sobre $educ$ (para obtener $\tilde{\beta}_1$)

Primero crea la variable $\log(wage)$ (si no recuerdas `mutate()`, revisa los labs anteriores):

```
df <- df %>% mutate(logwage = log(wage))
```

Ahora estima el modelo "corto":

```
est2 <- lm(logwage ~ educ, data = df)
```

** Modelo 2: log(wage) ~ educ **

Coeficientes estimados:

Estadísticas del modelo:

Estadístico	Valor
R^2	0.0974
R^2 ajustado	0.0964
Error estándar residual	0.4003
Estadístico F	100.70
N	935

3) Regresión completa de $\log(wage)$ **sobre** $educ$ **y** IQ **(para obtener** $\hat{\beta}_1$ **y** $\hat{\beta}_2$**)**

```
est3 <- lm(logwage ~ educ + IQ, data = df)
```

** Modelo 3: log(wage) ~ educ + IQ **

Coeficientes estimados:

Estadísticas del modelo:

Estadístico	Valor
R^2	0.1297
R^2 ajustado	0.1278
Error estándar residual	0.3933
Estadístico F	69.42
N	935

Verifica que se cumple la identidad:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$$

```
# Extraer coeficientes
beta1_tilde <- coef(est2)["educ"]
beta1_hat <- coef(est3)["educ"]
beta2_hat <- coef(est3)["IQ"]
delta1_tilde <- coef(est1)["educ"]
```

```r
# Mostrar los valores
cat("Coeficientes extraídos:\n")
```

```
Coeficientes extraídos:
```

```r
cat("----------------------------------------\n")
```

```
----------------------------------------
```

```r
cat(sprintf("~  (modelo corto):     %.6f\n", beta1_tilde))
```

```
~  (modelo corto):     0.059839
```

```r
cat(sprintf("^  (modelo completo):  %.6f\n", beta1_hat))
```

```
^  (modelo completo):  0.039120
```

```r
cat(sprintf("^  (modelo completo):  %.6f\n", beta2_hat))
```

```
^  (modelo completo):  0.005863
```

```r
cat(sprintf("~  (IQ sobre educ):    %.6f\n", delta1_tilde))
```

```
~  (IQ sobre educ):    3.533829
```

```r
cat("\n")
```

```r
# Verificar la identidad
lado_izq <- beta1_tilde
lado_der <- beta1_hat + beta2_hat * delta1_tilde

cat("Verificación de la identidad:\n")
```

```
Verificación de la identidad:
```

```
cat("----------------------------------------\n")
```

```
----------------------------------------
```

```
cat(sprintf("Lado izquierdo (β̃₁):          %.6f\n", lado_izq))
```

```
Lado izquierdo (β̃₁):          0.059839
```

```
cat(sprintf("Lado derecho (β̂₁ + δ̂₁·γ̃₁):    %.6f\n", lado_der))
```

```
Lado derecho (β̂₁ + δ̂₁·γ̃₁):    0.059839
```

```
cat(sprintf("Diferencia:                    %.10f\n", abs(lado_izq - lado_der)))
```

```
Diferencia:                    0.0000000000
```

```
cat(sprintf("¿Son iguales? %s\n", ifelse(abs(lado_izq - lado_der) < 1e-10, "SÍ", "NO")))
```

```
¿Son iguales? SÍ
```

Pregunta: ¿$\tilde{\beta}_1$ es mayor o menor que $\hat{\beta}_1$? ¿Qué significa esto en términos de **sesgo por variable omitida**?

Tabla comparativa de los tres modelos:

Modelo	Especificación	R^2	Coef. educ	Coef. IQ	N
Modelo 1	IQ ~ educ	0.2659	3.533829	—	935
Modelo 2	log(wage) ~ educ	0.0974	0.059839	—	935
Modelo 3	log(wage) ~ educ + IQ	0.1297	0.039120	0.005863	935

Multicolinealidad

Ahora veamos cómo calcular diagnósticos de multicolinealidad. Recuerda de Wooldridge que la multicolinealidad puede interpretarse mejor como "un problema de tamaño muestral pequeño".

Usaremos el set de datos `meapsingle` del paquete `wooldridge`. Nos interesa la variable `pctsgle`, que entrega el porcentaje de familias monoparentales que residen en el mismo ZIP code que la escuela. La variable resultado es `math4`, que corresponde al porcentaje de estudiantes que aprobaron el test estatal de matemáticas de 4° grado.

1) Cargar el set de datos y estimar *math4* sobre *pctsgle*

```
df <- as_tibble(meapsingle)
est_a <- lm(math4 ~ pctsgle, data = df)
```

** Modelo A: math4 ~ pctsgle **

Coeficientes estimados:

Estadísticas del modelo:

Estadístico	Valor
R^2	0.3795
R^2 ajustado	0.3768
Error estándar residual	12.4798
Estadístico F	138.85
N	229

Pregunta: Interpreta el coeficiente de pendiente de esta regresión. ¿El efecto parece grande?

2) Correlación entre *lmedinc* y *free*

Ahora considera el mismo modelo, pero agregando *lmedinc* y *free* como regresores adicionales.

- *lmedinc* es el log del ingreso mediano del hogar del ZIP code.
- *free* es el porcentaje de estudiantes que califican para almuerzo gratis o con descuento.

¿Crees que podría haber una correlación fuerte entre *lmedinc* y *free*? Calcula la correlación:

```r
correlacion <- cor(df$lmedinc, df$free)
cat(sprintf("Correlación entre lmedinc y free: %.4f\n", correlacion))
```

```
Correlación entre lmedinc y free: -0.7470
```

Preguntas:

- ¿El signo de la correlación es el esperado?
- ¿Está lo suficientemente cerca de 1 en valor absoluto como para violar la suposición de "no colinealidad perfecta"?

3) Modelo con *pctsgle*, *lmedinc* y *free*

```r
est_b <- lm(math4 ~ pctsgle + lmedinc + free, data = df)
```

** Modelo B: math4 ~ pctsgle + lmedinc + free **

Coeficientes estimados:

Estadísticas del modelo:

Estadístico	Valor
R^2	0.4598
R^2 ajustado	0.4526
Error estándar residual	11.6958
Estadístico F	63.85
N	229

Comparación de coeficiente de pctsgle:

Modelo	Coef. pctsgle	Error estándar	Estadístico t	R^2
Modelo A (simple)	-0.832881	0.070682	-11.784	0.3795
Modelo B (múltiple)	-0.199645	0.158716	-1.258	0.4598

Pregunta: Comenta el valor del coeficiente de *pctsgle* comparado con la primera regresión. ¿Qué puedes decir de *lmedinc* y *free* como variables de confusión (confounders)?

4) Cálculo de VIF (Variance Inflation Factors)

Un diagnóstico común para multicolinealidad es el **VIF**. Podemos usar la función `vif()` del paquete `car` para esto. Calcula los VIF del modelo anterior:

```
vif_values <- vif(est_b)
```

Factores de Inflación de Varianza (VIF):

	Variable	VIF	R^2 implícito	Interpretación
pctsgle	pctsgle	5.741	0.8258	Multicolinealidad moderada
lmedinc	lmedinc	4.119	0.7572	Multicolinealidad baja
free	free	3.188	0.6863	Multicolinealidad baja

VIFs de 10 o más suelen considerarse problemáticos, porque:

$$VIF_j = \frac{1}{1 - R_j^2}$$

lo que implica $R_j^2 > 0.9$. Ver p. 86 de Wooldridge.

Interpretación: Si $VIF_j = 10$, entonces:

$$R_j^2 = 1 - \frac{1}{10} = 0.9$$

Esto significa que el 90% de la variación en x_j puede ser explicada linealmente por las demás variables explicativas, indicando una colinealidad alta.

5) ¿Es la multicolinealidad un problema?

La multicolinealidad suele ser un problema principalmente en sets de datos con **muestra pequeña**. A medida que el tamaño muestral aumenta, R_j^2 podría disminuir. Además, la variación total en x_j (i.e., SST_j) aumenta con el tamaño muestral. Por esto, la multicolinealidad generalmente no es un problema que preocupe demasiado en muestras grandes.

La fórmula de la varianza del estimador MCO es:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j \cdot (1 - R_j^2)} = \frac{\sigma^2}{SST_j} \cdot VIF_j$$

donde:

- σ^2 es la varianza del error
- $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ es la suma de cuadrados total de x_j
- R_j^2 es el R^2 de la regresión auxiliar de x_j sobre todas las demás variables explicativas

Observa que incluso con un VIF_j alto, si SST_j es suficientemente grande (muestra grande), la varianza del estimador puede seguir siendo razonablemente pequeña.