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**. El laboratorio debe completarse en tu grupo. Para obtener crédito, sube tu script **.R** al lugar correspondiente en Canvas.

Para empezar

Abre un nuevo script de R (llámalo `ICL5_XYZ.R`, donde `XYZ` son tus iniciales) y agrega el "preámbulo" habitual al inicio:

```
# 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, 1,~
$ black   <int> 0, 0, 0, 0, 0, 1, 0, 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, 0,~
$ urban   <int> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 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 $\tilde{\delta}_1$)

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

```
Call:
lm(formula = IQ ~ educ, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-50.228  -7.262   0.907   8.772  37.373

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.6872     2.6229   20.47   <2e-16 ***
educ          3.5338     0.1922   18.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.9 on 933 degrees of freedom
Multiple R-squared:  0.2659,    Adjusted R-squared:  0.2652
F-statistic:   338 on 1 and 933 DF,  p-value: < 2.2e-16
```

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)
summary(est2)
```

```
Call:
lm(formula = logwage ~ educ, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.94620 -0.24832  0.03507  0.27440  1.28106

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.973062   0.081374   73.40   <2e-16 ***
educ        0.059839   0.005963   10.04   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4003 on 933 degrees of freedom
Multiple R-squared:  0.09742,   Adjusted R-squared:  0.09645
F-statistic: 100.7 on 1 and 933 DF,  p-value: < 2.2e-16
```

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)
summary(est3)
```

```
Call:
lm(formula = logwage ~ educ + IQ, data = df)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.01601 -0.24367  0.03359  0.27960  1.23783

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.6582876  0.0962408  58.793  < 2e-16 ***
educ        0.0391199  0.0068382   5.721 1.43e-08 ***
IQ          0.0058631  0.0009979   5.875 5.87e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3933 on 932 degrees of freedom
Multiple R-squared:  0.1297,    Adjusted R-squared:  0.1278
F-statistic: 69.42 on 2 and 932 DF,  p-value: < 2.2e-16
```

Verifica que se cumple la identidad:

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

# Verificar la identidad
cat("  (modelo corto):", beta1_tilde, "\n")
```

```
  (modelo corto): 0.05983921
```

```r
cat("  (modelo completo):", beta1_hat, "\n")
```

```
  (modelo completo): 0.0391199
```

```r
cat("  (modelo completo):", beta2_hat, "\n")
```

```
  (modelo completo): 0.005863132
```

```r
cat("  (IQ sobre educ):", delta1_tilde, "\n")
```

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

```r
cat("\nVerificación:  _tilde =  _hat +  _hat ×  _tilde\n")
```

```
Verificación:  _tilde =  _hat +  _hat ×  _tilde
```

```r
cat("Lado izquierdo:", beta1_tilde, "\n")
```

```
Lado izquierdo: 0.05983921
```

```r
cat("Lado derecho:", beta1_hat + beta2_hat * delta1_tilde, "\n")
```

```
Lado derecho: 0.05983921
```

```r
cat("¿Son iguales?",
    abs(beta1_tilde - (beta1_hat + beta2_hat * delta1_tilde)) < 0.0001, "\n")
```

```
¿Son iguales? TRUE
```

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

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)
summary(est_a)
```

```
Call:
lm(formula = math4 ~ pctsgle, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-47.791  -8.310   1.600   8.092  50.317

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 96.77043    1.59680   60.60   <2e-16 ***
pctsgle     -0.83288    0.07068  -11.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.48 on 227 degrees of freedom
Multiple R-squared:  0.3795,    Adjusted R-squared:  0.3768
F-statistic: 138.9 on 1 and 227 DF,  p-value: < 2.2e-16
```

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:

```
correlacion <- cor(df$lmedinc, df$free)
cat("Correlación entre lmedinc y free:", round(correlacion, 4), "\n")
```

```
Correlación entre lmedinc y free: -0.747
```

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)
summary(est_b)
```

```
Call:
lm(formula = math4 ~ pctsgle + lmedinc + free, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-34.919  -7.195   0.931   7.313  50.152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.72322   58.47814   0.884    0.377
pctsgle     -0.19965    0.15872  -1.258    0.210
lmedinc      3.56013    5.04170   0.706    0.481
free        -0.39642    0.07035  -5.635  5.2e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.7 on 225 degrees of freedom
Multiple R-squared:  0.4598,    Adjusted R-squared:  0.4526
F-statistic: 63.85 on 3 and 225 DF,  p-value: < 2.2e-16
```

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)
print(vif_values)
```

```
 pctsgle  lmedinc     free
5.740981 4.118812 3.188079
```

VIFs de 10 o más suelen considerarse problemáticos, porque:

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

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

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,