Laboratorio 4 – Regresión múltiple, no linealidades y teorema de Frisch–Waugh

Table of contents

El propósito de este laboratorio es seguir practicando tus habilidades de **regresión**.

Primeros pasos

Abre un nuevo script de R y carga los paquetes

```
# Instalar paquetes solo si faltan (estilo ECO/EPG)
pkgs <- c("tidyverse", "broom", "wooldridge", "modelsummary", "kableExtra")
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(modelsummary)
library(kableExtra)

# Receta ECO/EPG: forzar tablas estables (LaTeX clásico)
options(modelsummary_factory_default = "kableExtra")
```

Para este laboratorio usaremos datos de **precios de viviendas**, contenidos en el conjunto `hprice1` del paquete `wooldridge`. Cada observación es una vivienda.

```
df <- as_tibble(hprice1)
```

Revisa qué variables hay en `df` usando:

```
glimpse(df)
```

```
Rows: 88
Columns: 10
$ price    <dbl> 300.000, 370.000, 191.000, 195.000, 373.000, 466.275, 332.500~
$ assess   <dbl> 349.1, 351.5, 217.7, 231.8, 319.1, 414.5, 367.8, 300.2, 236.1~
$ bdrms    <int> 4, 3, 3, 3, 4, 5, 3, 3, 3, 3, 4, 5, 3, 3, 3, 4, 4, 3, 3, 4, 3~
$ lotsize  <dbl> 6126, 9903, 5200, 4600, 6095, 8566, 9000, 6210, 6000, 2892, 6~
$ sqrft    <int> 2438, 2076, 1374, 1448, 2514, 2754, 2067, 1731, 1767, 1890, 2~
$ colonial <int> 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1~
$ lprice   <dbl> 5.703783, 5.913503, 5.252274, 5.273000, 5.921578, 6.144775, 5~
$ lassess  <dbl> 5.855359, 5.862210, 5.383118, 5.445875, 5.765504, 6.027073, 5~
$ llotsize <dbl> 8.720297, 9.200593, 8.556414, 8.433811, 8.715224, 9.055556, 9~
$ lsqrft   <dbl> 7.798934, 7.638198, 7.225482, 7.277938, 7.829630, 7.920810, 7~
```

O, si quieres estadísticas descriptivas rápidas:

```
datasummary_skim(df, histogram = FALSE)
```

Regresión múltiple

Estimemos el siguiente modelo:

$$price = \beta_0 + \beta_1\, sqrft + \beta_2\, bdrms + u$$

donde *price* es el precio de la vivienda en **miles de dólares**.

```
est1 <- lm(price ~ sqrft + bdrms, data = df)
```

Mostramos la salida del modelo en una tabla (estable en PDF):

	Unique	Missing Pct.	Mean	SD	Min	Median	Max
price	71	0	\num{293.5}	\num{102.7}	\num{111.0}	\num{265.5}	\num{7
assess	88	0	\num{315.7}	\num{95.3}	\num{198.7}	\num{290.2}	\num{7
bdrms	6	0	\num{3.6}	\num{0.8}	\num{2.0}	\num{3.0}	\num{7
lotsize	84	0	\num{9019.9}	\num{10174.2}	\num{1000.0}	\num{6430.0}	\num{9
sqrft	85	0	\num{2013.7}	\num{577.2}	\num{1171.0}	\num{1845.0}	\num{3
colonial	2	0	\num{0.7}	\num{0.5}	\num{0.0}	\num{1.0}	\num{1
lprice	71	0	\num{5.6}	\num{0.3}	\num{4.7}	\num{5.6}	\num{6
lassess	88	0	\num{5.7}	\num{0.3}	\num{5.3}	\num{5.7}	\num{6
llotsize	84	0	\num{8.9}	\num{0.5}	\num{6.9}	\num{8.8}	\num{1
lsqrft	85	0	\num{7.6}	\num{0.3}	\num{7.1}	\num{7.5}	\num{8

```r
modelsummary(est1, output = "kableExtra") |>
  kable_styling(full_width = FALSE, position = "center", latex_options = "hold_position")
```

	(1)
(Intercept)	−19.315
	(31.047)
sqrft	0.128
	(0.014)
bdrms	15.198
	(9.484)
Num.Obs.	88
R2	0.632
R2 Adj.	0.623
AIC	984.0
BIC	993.9
Log.Lik.	−487.999
F	72.964
RMSE	61.96

Deberías obtener un coeficiente cercano a `0.128` en `sqrft` y `15.2` en `bdrms`. Interpreta estos coeficientes (puedes escribir la interpretación como comentario en tu script). ¿Te parecen razonables?

También deberías obtener $R^2 \approx 0.632$. A partir de ese número, ¿crees que es un buen modelo para explicar precios de vivienda?

Verifica que el promedio de los residuos es (aproximadamente) cero:

```
mean(est1$residuals)
```

```
[1] -2.863674e-16
```

Agregando no linealidades

El modelo anterior estimó un intercepto cercano a `-19.3`, lo que implicaría que una casa sin dormitorios y sin superficie tendría un precio esperado de **-$19,300**.

Para asegurar que el modelo siempre prediga un precio positivo, usemos $log(price)$ como variable dependiente. Además, agreguemos términos cuadráticos para `sqrft` y `bdrms`, permitiendo **rendimientos marginales decrecientes**.

Primero, usemos `mutate()` para crear las nuevas variables:

```
df <- df %>%
  mutate(
    logprice = log(price),
    sqrftSq  = sqrft^2,
    bdrmSq   = bdrms^2
  )
```

Ahora estimamos el nuevo modelo:

```
est2 <- lm(logprice ~ sqrft + sqrftSq + bdrms + bdrmSq, data = df)
```

Tabla del modelo (estable para PDF):

```
modelsummary(est2, output = "kableExtra") |>
  kable_styling(full_width = FALSE, position = "center", latex_options = "hold_position")
```

Si quieres ver más decimales:

```
modelsummary(est2, output = "kableExtra", fmt = 10) |>
  kable_styling(full_width = FALSE, position = "center", latex_options = "hold_position")
```

Los coeficientes nuevos tienen magnitudes mucho más pequeñas. Explica por qué podría ocurrir eso.

El nuevo $R^2 \approx 0.595$, menor que el 0.632 del modelo anterior. ¿Eso significa necesariamente que este modelo es peor? ¿Por qué?

	(1)
(Intercept)	5.074
	(0.325)
sqrft	0.000
	(0.000)
sqrftSq	0.000
	(0.000)
bdrms	−0.130
	(0.145)
bdrmSq	0.020
	(0.018)
Num.Obs.	88
R2	0.595
R2 Adj.	0.576
AIC	−28.7
BIC	−13.8
Log.Lik.	20.347
F	30.523
RMSE	0.19

Teorema de Frisch–Waugh: obtener efectos parciales

Probemos el teorema de **Frisch–Waugh**, que afirma:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \hat{r}_{i1}\, y_i}{\sum_{i=1}^{N} \hat{r}_{i1}^2}$$

donde $\hat{r}_{i1}$ es el residuo de una regresión de x_1 sobre $x_2, \dots, x_k$.

Apliquémoslo al modelo recién estimado. Primero, regresamos `sqrft` sobre el resto de los X y guardamos los residuos en `df`:

```
aux <- lm(sqrft ~ sqrftSq + bdrms + bdrmSq, data = df)
df <- df %>% mutate(sqrft.resid = aux$residuals)
```

Ahora, si estimamos una regresión simple de `logprice` sobre `sqrft.resid`, deberíamos obtener el mismo coeficiente que el de `sqrft` en la regresión original (aprox. `3.74e-4`).

```
fw_est <- lm(logprice ~ sqrft.resid, data = df)
```

	(1)
(Intercept)	5.073 869 982 0
	(0.325 409 544 1)
sqrft	0.000 374 152 6
	(0.000 247 441 3)
sqrftSq	0.000 000 000 7
	(0.000 000 050 8)
bdrms	−0.130 182 503 5
	(0.144 982 999 9)
bdrmSq	0.019 899 343 4
	(0.017 800 957 2)
Num.Obs.	88
R2	0.595
R2 Adj.	0.576
AIC	−28.7
BIC	−13.8
Log.Lik.	20.347
F	30.523
RMSE	0.19

Salida en tabla (estable para PDF):

```
modelsummary(fw_est, output = "kableExtra") |>
  kable_styling(full_width = FALSE, position = "center", latex_options = "hold_position")
```

	(1)
(Intercept)	5.633
	(0.032)
sqrft.resid	0.000
	(0.000)
Num.Obs.	88
R2	0.011
R2 Adj.	0.000
AIC	43.9
BIC	51.4
Log.Lik.	−18.963
F	0.970
RMSE	0.30

Frisch–Waugh “a mano”

También podemos calcular la fórmula de Frisch–Waugh directamente:

```r
beta1 <- sum(df$sqrft.resid * df$logprice) / sum(df$sqrft.resid^2)
print(beta1)
```

```
[1] 0.0003741526
```

Debería coincidir (aproximadamente) con el coeficiente estimado para `sqrft` en el modelo con todas las covariables.