Laboratorio 2 – Regresión Lineal Simple

Table of contents

El propósito de este laboratorio es practicar el uso de **R** para realizar **pruebas de hipótesis** y ejecutar una **regresión MCO (OLS)** básica.

Primeros pasos

Abre un nuevo script de R y carga los paquetes

```
# Instalar paquetes solo si faltan (igual que el estilo de tus labs)
pkgs <- c("tidyverse", "modelsummary", "wooldridge", "broom", "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(modelsummary)
library(broom)
library(wooldridge)
library(kableExtra)
```

```
# --- CLAVE: evitar que modelsummary use backend "tinytable" (que puede generar tblr + \num)
# Con esto, tus tablas en PDF salen en LaTeX clásico vía kableExtra.
options(modelsummary_factory_default = "kableExtra")
```

Carga el conjunto de datos `audit` del paquete `wooldridge`:

```
df <- as_tibble(audit)
```

Prueba unilateral

El conjunto de datos `audit` contiene tres variables: `w`, `b` y `y`.

- `b` indica si el currículum de la persona **negra** de un par fue seleccionado para una oferta de trabajo.
- `w` indica lo mismo, pero para la persona **blanca** del par.
- `y` corresponde a la diferencia entre ambas variables, es decir: `y = b - w`.

Queremos probar la siguiente hipótesis:

$$H_0 : \mu = 0 \quad \text{vs} \quad H_a : \mu < 0$$

donde $\mu = \theta_B - \theta_W$, es decir, la diferencia en las tasas de oferta laboral entre personas negras y blancas.

La función `t.test()` en R

Para realizar una prueba t en R, basta con entregar la información apropiada a la función `t.test()`.

Ahora realizamos la prueba de hipótesis descrita anteriormente:

```
t.test(df$y, alternative = "less")
```

```
    One Sample t-test

data:  df$y
t = -4.2768, df = 240, p-value = 1.369e-05
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
```

```
        -Inf -0.08151529
sample estimates:
 mean of x
-0.1327801
```

R calcula automáticamente el estadístico t usando la fórmula:

$$\frac{\overline{y} - \mu}{SE_{\bar{y}}}$$

Interpretación del resultado de `t.test()`

Al ejecutar el script, deberías obtener una salida similar a la siguiente:

```
> t.test(df$y, alternative="less")

    One Sample t-test

data:  df$y
t = -4.2768, df = 240, p-value = 1.369e-05
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
        -Inf -0.08151529
sample estimates:
 mean of x
-0.1327801
```

R reporta el valor del estadístico t, los grados de libertad y el p-value asociado. En este caso, el p-value es aproximadamente 0.00001369, mucho menor que 0.05 (nuestro nivel de significancia). Por lo tanto, rechazamos H_0.

Prueba bilateral

Ahora supongamos que queremos probar si las tasas de oferta laboral para personas negras son **distintas** de las de personas blancas.

Queremos probar:

$$H_0 : \theta_b = \theta_w \quad \text{vs} \quad H_a : \theta_b \neq \theta_w$$

El código para realizar esta prueba es:

```r
t.test(df$b, df$w, alternative = "two.sided", paired = TRUE)
```

```
    Paired t-test

data:  df$b and df$w
t = -4.2768, df = 240, p-value = 2.739e-05
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.1939385 -0.0716217
sample estimates:
mean difference
     -0.1327801
```

Tu primera regresión (de este curso)

```r
df <- as_tibble(countymurders)
```

```r
glimpse(df)
```

```
Rows: 37,349
Columns: 20
$ arrests     <int> 2, 3, 2, 7, 3, 1, 1, 2, 0, 5, 0, 1, 5, 3, 4, 5, 8, 4, 9, 8~
$ countyid    <int> 1001, 1001, 1001, 1001, 1001, 1001, 1001, 1001, 1001, 1001~
$ density     <dbl> 54.05000, 53.66000, 53.75000, 53.78000, 53.91000, 54.11000~
$ popul       <int> 32216, 31984, 32036, 32056, 32128, 32248, 32888, 33264, 33~
$ perc1019    <dbl> 20.63000, 20.19000, 19.66000, 19.10000, 18.54000, 18.06000~
$ perc2029    <dbl> 15.28000, 15.55000, 15.73000, 15.88000, 15.92000, 15.87000~
$ percblack   <dbl> 22.33000, 22.07000, 21.80000, 21.53000, 21.26000, 20.96000~
$ percmale    <dbl> 40.25000, 40.36000, 40.42000, 40.47000, 40.51000, 40.45000~
$ rpcincmaint <dbl> 167.670, 167.990, 166.630, 176.530, 166.250, 153.120, 151.~
$ rpcpersinc  <dbl> 8780.80, 8232.80, 8327.61, 8545.55, 8965.16, 9254.02, 9885~
$ rpcunemins  <dbl> 29.160, 43.920, 71.410, 72.220, 40.360, 44.540, 38.350, 35~
$ year        <int> 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989~
$ murders     <int> 2, 1, 3, 7, 2, 2, 4, 1, 0, 3, 1, 1, 1, 1, 1, 5, 7, 4, 6, 7~
$ murdrate    <dbl> 0.6208096, 0.3126563, 0.9364465, 2.1836790, 0.6225100, 0.6~
$ arrestrate  <dbl> 0.6208095, 0.9379690, 0.6242977, 2.1836790, 0.9337650, 0.3~
$ statefips   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ countyfips  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3~
```

```
$ execs       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ lpopul      <dbl> 10.38022, 10.37299, 10.37462, 10.37524, 10.37748, 10.38121~
$ execrate    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Regresar tasa de asesinatos sobre tasa de ejecuciones

```
est <- lm(murders ~ execs, data = df)
```

```
tidy(est)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)     6.84     0.242      28.3 3.97e-174
2 execs          65.5      2.15       30.5 7.44e-202
```

```
glance(est)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic   p.value    df   logLik    AIC    BIC
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>    <dbl>  <dbl>  <dbl>
1    0.0243        0.0243  46.6      930. 7.44e-202     1 -196508. 3.93e5 3.93e5
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Tabla final del modelo (PDF estable)

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

	(1)
(Intercept)	6.838
	(0.242)
execs	65.465
	(2.146)
Num.Obs.	37 349
R2	0.024
R2 Adj.	0.024
AIC	393 021.2
BIC	393 046.8
Log.Lik.	−196 507.599
F	930.365
RMSE	46.64