Laboratorio 5 - Regresión lineal

Métodos Cuantitativos

Carlos Eduardo Molina Berumen

El Colegio de México

14 de noviembre de 2025

Regresión líneal

Concepto básico:

$$\mathrm{y} = \beta_1 + \beta_2 \cdot \mathrm{x} + \mathrm{error}.$$

Objetivo de hoy:

Entender una regresión lineal. Por ejemplo, Queremos ver si los pingüinos con aletas más largas tienden a pesar más.

Carguemos los datos y un par de librerías

```
#install.packages("palmerpenguins"), lo mismo con el resto

library(palmerpenguins)
library(ggplot2)
library(stargazer)
```

Exploremos los datos

```
head(penguins)
```

```
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_lengt
  <fct>   <fct>              <dbl>         <dbl>           <
1 Adelie  Torgersen           39.1          18.7
2 Adelie  Torgersen           39.5          17.4
3 Adelie  Torgersen           40.3          18
4 Adelie  Torgersen           NA            NA
5 Adelie  Torgersen           36.7          19.3
6 Adelie  Torgersen           39.3          20.6
# i 2 more variables: sex <fct>, year <int>
```

Hay una forma sencilla de sacar la estadística descriptiva.

```
summary(penguins)
```

```
      species          island    bill_length_mm  bill_depth_mm
 Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
 Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
 Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
                                 Mean   :43.92   Mean   :17.15
                                 3rd Qu.:48.50   3rd Qu.:18.70
                                 Max.   :59.60   Max.   :21.50
                                 NA's   :2       NA's   :2
 flipper_length_mm  body_mass_g       sex           year
 Min.   :172.0     Min.   :2700   female:165   Min.   :2007
 1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
 Median :197.0     Median :4050   NA's  : 11   Median :2008
 Mean   :200.9     Mean   :4202                Mean   :2008
 3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
 Max.   :231.0     Max.   :6300                Max.   :2009
```

Veamos cómo se mueven nuestras variables juntas

```
# Quitamos los NA para evitar errores
penguins_clean <- na.omit(penguins)

# Covarianza entre masa corporal y longitud de aleta
cov(penguins_clean$body_mass_g,
    penguins_clean$flipper_length_mm)
```

```
[1] 9852.192
```

Veamos si nuestras variables siguen una distribución normal:

```
shapiro.test(penguins_clean$body_mass_g)
```

```
    Shapiro-Wilk normality test

data:  penguins_clean$body_mass_g
W = 0.95801, p-value = 3.568e-08
```

```
shapiro.test(penguins_clean$flipper_length_mm)
```

```
    Shapiro-Wilk normality test

data:  penguins_clean$flipper_length_mm
W = 0.95171, p-value = 5.393e-09
```

Los datos no están normalmente distribuidos.

Hagamos un modelo de regresión líneal:

```
m1 <- lm(body_mass_g ~ flipper_length_mm, data = penguins)
summary(m1)
```

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max
-1058.80  -259.27   -26.88   247.33  1288.69

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5780.831    305.815  -18.90   <2e-16 ***
flipper_length_mm    49.686      1.518   32.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podemos hacerlo mejor

```
stargazer(m1, type = "text")
```

```
================================================
                         Dependent variable:
                     ---------------------------
                              body_mass_g
------------------------------------------------
flipper_length_mm              49.686***
                               (1.518)

Constant                     -5,780.831***
                              (305.815)

------------------------------------------------
Observations                      342
R2                                0.759
```

¿Cómo se vería esto en una fórmula?

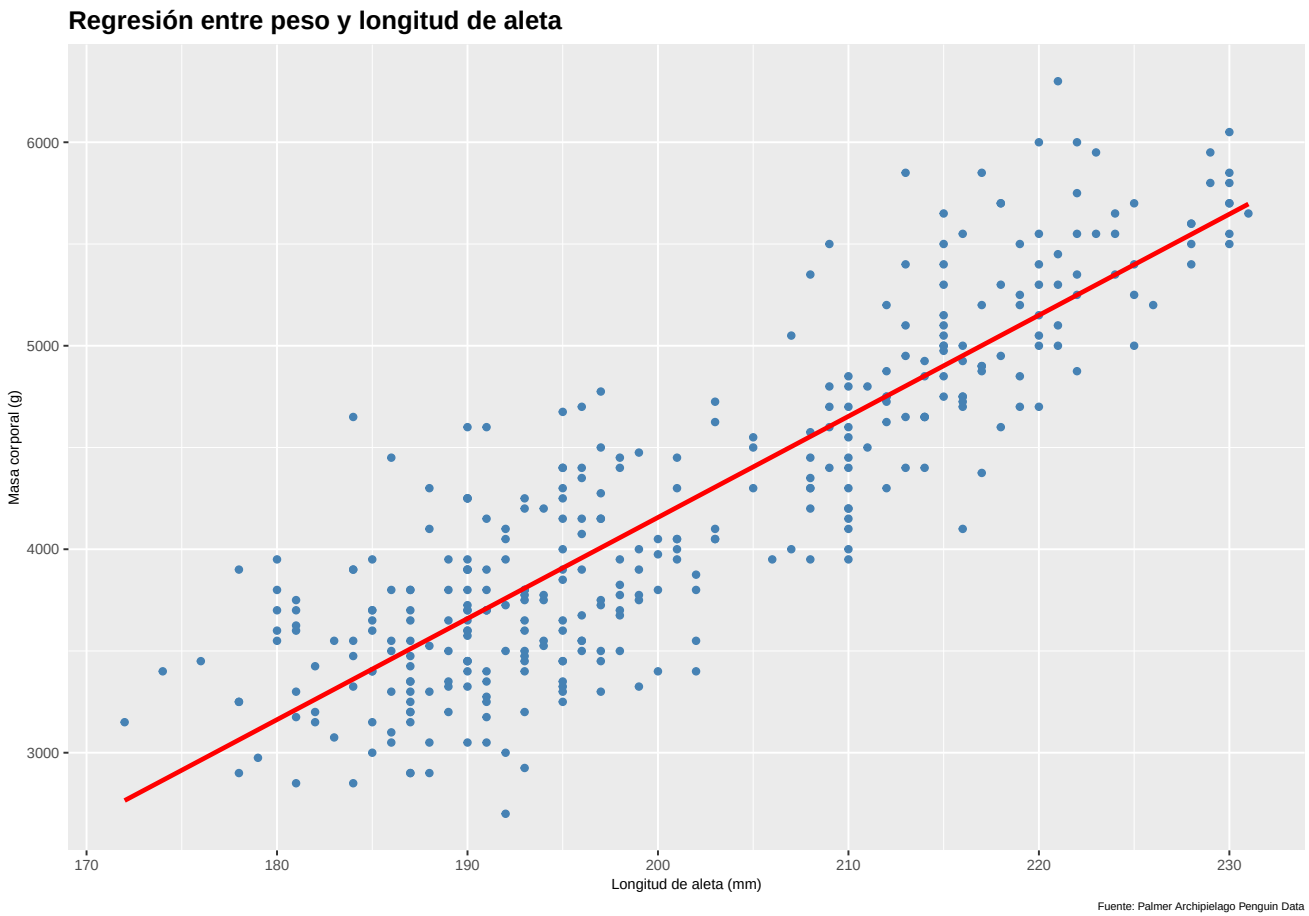
$$\text{Body Mass}_i = -5780.81 + 49.686\,(\text{Flipper Length}_i) + \varepsilon_i$$

¿Cómo interpretamos eso?

Grafiquemos nuestros datos

```
p1 <- ggplot(penguins, aes(x = flipper_length_mm,
                           y = body_mass_g)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE,
              color = "red", lwd = 1.2) +
  labs(title = "Regresión entre peso y longitud de aleta",
       x = "Longitud de aleta (mm)",
       y = "Masa corporal (g)",
        caption = "Fuente: Palmer Archipielago Penguin Data")
  theme(
    plot.title=element_text(size=14, face="bold"),
    plot.subtitle=element_text(size=8),
    plot.caption=element_text(size=6),
    axis.title=element_text(size=8),
    axis.text=element_text(size=8)
)
```

```
print(p1)
```


Regresión entre peso y longitud de aleta
Masa corporal (g)
6000
5000
4000
3000
170
180
190
200
210
220
230
Longitud de aleta (mm)
Fuente: Palmer Archipielago Penguin Data