# Regresión Lineal IV

Probabilidad Aplicada

3602

# Análisis de regresión lineal múltiple

En este capítulo veremos cómo gestionar un análisis de regresión lineal considerando una variable respuesta y múltiples variables predictoras.

Trabajaremos con la bese **insurance** y **datos_personas**

# Regresión lineal múltiple: variables predictoras cualitativas dicotómicas.

Hasta ahora hemos supuesto que las variables $x$ son cuantitativas continuas. Sin embargo los modelos de regresión lineal permiten combinar variables independientes cuantitativas con cualitativas. Para ello se utilizan las variables conocidas como *dummy*.

Si se quiere incluir una variable categórica con sólo 2 categorías (dicotómica), será necesario introducir únicamente una variable *dummy* que tomará el valor $0$ para una de las categorías, que será la de *referencia* y el $1$ en la otra categoría.

A modo de ejemplo, cargamos la base **datos_personas**

```
datos_personas <- read_csv("E:/INSPT-UTN/2024/02-Informática/Proba/Unidad 7/datos_personas.csv")
```

```
## Rows: 100 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Sexo
## dbl (3): Peso, Altura, Ancho
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Hacemos un *head()* para ver las primeras 6 filas-

```
head(datos_personas)
```

```
## # A tibble: 6 x 4
##    Peso Altura Sexo   Ancho
##   <dbl>  <dbl> <chr>  <dbl>
## 1    88    158 Hombre    37
## 2    78    173 Hombre    41
## 3    64    150 Mujer     50
## 4    92    193 Mujer     50
## 5    57    157 Mujer     54
## 6    70    173 Mujer     51
```

En la base se registra:

Tabla de variables

| Variable | Representa | Tipo de variable |
|---|---|---|
| **Peso** | Peso de una persona en kg. | Cuantitativa continua |
| **Altura** | Altura de una persona en cm. | Cuantitativa continua |
| **Sexo** | Sexo de una persona (hombre-mujer) | Cualitativa dicotómica |
| **Ancho** | Ancho de la espalda en cm | Cuantitativa continua |

A modo de ejemplo, la siguiente regresión lineal modeliza el peso de una persona en función de su altura y sexo (Por ahora dejamos la variable Ancho de espalda).

Mientras que $Altura Altura$ es una variable continua, $Sexo Sexo$ está planteada como una variable cualitativa con 2 categorías, tomando los valores **hombre** o **mujer**, entonces para nuestro modelo es necesario incorporar la variable $Sexo Sexo$ como una variable *dummy* tomando el valor $0 0$ para mujer y el $1$ para el hombre.

Hacemos esta transformación con la función *mutate()* y *case_when()* del paquete **dyplr**

```
datos_personas <- datos_personas %>%
  mutate(Sexo_dummy = case_when( #la nueva columna o variable se llamará Sexo_dummy
    Sexo == "Hombre" ~ 1, #cuando Sexo tome el valor hombre se reemplazará por 1
    Sexo == "Mujer" ~ 0 #cuando Sexo tome el valor mujer se reemplazará por 0
  ))
```

Esta designación es arbitraria y se podría hacer al contrario.

El modelo que buscamos tiene la forma:

$$Peso = \beta_0 + \beta_1 Altura + \beta_2 SexoDummy Peso = \beta_0 + \beta_1 Altura + \beta_2 SexoDummy$$

Siendo:

$Peso$ :$Peso$ : Variable respuesta. Peso de un sujeto en kg.

$Altura$ :$Altura$ : Variable predictora. Altura en cm.

$SexoDummy$ :$SexoDummy$ : Variable predictora, con valor $0$ para mujer y $1$ para hombre.

$\beta_0$ :$\beta_0$ : Intercepto. Peso de una persona que está en la categoría de referencia, es decir una mujer ($Sexo = 0 Sexo = 0$) que tuviera $Altura = 0 Altura = 0$. Tiene sentido matemático pero carece de sentido real.

$\beta_1$ :$\beta_1$ : Aumento medio de Peso por cada unidad de aumento en Altura.

$\beta_2$ :$\beta_2$ : Aumento medio de Peso de los Hombres con respecto a las mujeres que son el grupo de *referencia*.

Armamos el modelo usando *lm()* (R base) y *tidy()* paquete **broom**

```
modelo <- lm(Peso ~ Altura + Sexo_dummy, data = datos_personas)
tidy(modelo)
```

```
## # A tibble: 3 x 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   57.4      18.1       3.17  0.00204
## 2 Altura         0.100     0.104     0.970 0.335
## 3 Sexo_dummy    -1.87      2.91     -0.644 0.521
```

Entonces el modelo obtenido es:

$$Peso = 57.4 + 0.100 Altura - 1.87 SexoDummy Peso = 57.4 + 0.100 Altura - 1.87 SexoDummy$$

Este modelo nos dice que el peso aumenta en 0.100 kg por cada centímetro de aumento en la altura y disminuye 1.87 kg si la persona es hombre frente a que sea una mujer (variable de referencia).

ATENCIÓN: ¿Qué observas en nuestro modelo? ¿Qué sucede si se incorpora la variable Ancho?

# Regresión lineal múltiple: interacción entre variables.

Además de introducir variables cuantitativas y categóricas a un modelo también se pueden incluir combinaciones de ellas. Estas combinaciones son llamadas **interacciones** y se incorporan en el caso de que además del efecto lineal que tiene una variable independiente en la dependiente, una variable independiente pueda modular el efecto de otra independiente debido a que interaccionan entre sí.

Siguiendo el ejemplo anterior vamos a formular un modelo de regresión lineal múltiple donde se estima el peso de una persona a partir de su altura el ancho de su espalda y la interacción de ambas variables.

El esquema del modelo sería:

$$Peso = \beta_0 + \beta_1 Altura + \beta_2 Ancho + \beta_3 Altura Ancho Peso = \beta_0 + \beta_1 Altura + \beta_2 Ancho + \beta_3 Altura Ancho$$

Siendo:

$Peso$ :$Peso$ : Variable respuesta. Peso de un sujeto en kg.

$Altura$ :$Altura$ : Variable predictora. Altura en cm.

$Ancho$ :$Ancho$ : Variable predictora. Ancho de espalda en cm.

$\beta_0$ :$\beta_0$ : Intercepto. Peso de una persona cuando $Altura = 0 Altura = 0$ y $Ancho = 0 Ancho = 0$. Tiene sentido matemático pero carece de sentido real.

$\beta_1$ :$\beta_1$ : Aumento medio de Peso por cada unidad de aumento en Altura.

$\beta_2$ :$\beta_2$ : Aumento medio de Peso por cada unidad de aumento en Ancho de espalda.

$\beta_3$ :$\beta_3$ : Interacción de los efectos lineales de $Altura Altura$ y $Ancho Ancho$.

En R, la interacción de variables se analiza con el signo * en la fórmula de **lm()**.

```
modelo <- lm(Peso ~ Altura + Ancho + Altura * Ancho, data = datos_personas)
tidy(modelo)
```

```
## # A tibble: 4 x 5
##   term          estimate std.error statistic p.value
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  162.      130.          1.24    0.218
## 2 Altura        -0.439     0.752      -0.584   0.560
## 3 Ancho         -2.40      2.89       -0.828   0.409
## 4 Altura:Ancho   0.0123    0.0166      0.739   0.462
```

ATENCIÓN: ¿Qué observas en nuestro modelo? ¿Qué sucede si se incorpora la variable Sexo_dummy?