

PRÁCTICA 1

Correlación

1. El conjunto de datos `bdims` del paquete `openintro` que se habilita en el workspace del R con los comandos

```
library(openintro)
```

```
data(bdims, package = "openintro")
```

El archivo consiste en medidas del diámetro y circunferencia de distintas partes del cuerpo (21 variables), así como edad, peso, altura y género de 507 personas físicamente activas. Para más detalle, tipear

```
help(bdims, package = "openintro")
```

- Calcular las correlaciones muestrales entre las 21 variables que miden el diámetro o circunferencia de las distintas partes del cuerpo. ¿Cuántas correlaciones debe calcular? ¿Cuál sería la mejor manera de exhibir esta información? ¿Están positiva o negativamente correlacionadas estas variables?
 - Encontrar las dos variables con mayor correlación entre sí. Realizar un scatter plot. ¿Le parece que este número resume adecuadamente el vínculo entre ambas variables?
 - Repetir con las de menor correlación.
 - Hacer un scatter plot de peso en el eje y y altura en el eje x y calcular la correlación muestral o de Pearson. ¿Le parece que este número resume adecuadamente el vínculo entre ambas variables?
 - Hacer scatter plots de la variable `bia_di`, que es la distancia biacromial (informalmente, la distancia entre los hombros) con las siguientes cuatro variables y calcular las correlaciones de a pares para ambas. Observar cómo se comportan los scatterplots para distintos valores de la correlación.
 - `age`, la edad
 - `bii_di`, el ancho de la pelvis
 - `che_de`, la profundidad del pecho
 - `wri_di`, la circunferencia de la muñeca
2. Sean $(X_i, Y_i)_{1 \leq i \leq n}$ observaciones bivariadas, la covarianza muestral entre X e Y basada en las observaciones se define por

$$\widehat{\text{cov}}((X_1, Y_1), \dots, (X_n, Y_n)) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Por simplicidad, en vez de escribir $\widehat{\text{cov}}((X_1, Y_1), \dots, (X_n, Y_n))$ a veces escribiremos $\widehat{\text{cov}}(X_i, Y_i)$

- Sean $a, b \in \mathbb{R}$ constantes.
 - Definimos $X_i^* = X_i + a$, $i = 1, \dots, n$. Probar que $\widehat{\text{cov}}(X_i^*, Y_i) = \widehat{\text{cov}}(X_i, Y_i)$.
 - Definimos $X_i^* = bX_i + a$, $i = 1, \dots, n$. Probar que $\widehat{\text{cov}}(X_i^*, Y_i) = b\widehat{\text{cov}}(X_i, Y_i)$.
- Sean $X_i^* = X_i - \bar{X}$ y $Y_i^* = Y_i - \bar{Y}$ $i = 1, \dots, n$. Probar que $\widehat{\text{cov}}(X_i^*, Y_i^*) = \widehat{\text{cov}}(X_i^*, Y_i) = \widehat{\text{cov}}(X_i, Y_i)$.
- Probar que vale lo siguiente:

$$\widehat{\text{cov}}((X_1, Y_1), \dots, (X_n, Y_n)) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}) Y_i$$

- Probar que la covarianza muestral puede escribirse de la siguiente forma

$$\widehat{\text{cov}}(X_i, Y_i) = \frac{1}{n-1} \left[\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y} \right]$$

- Probar que

$$\widehat{\text{cov}}(X_i, X_i) = S_X^2$$

donde $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ es la varianza muestral de las X 's.

3. Sean $(X_i, Y_i)_{1 \leq i \leq n}$ observaciones bivariadas, el coeficiente de correlación muestral o coeficiente de correlación de Pearson entre X e Y basado en las observaciones se define por

$$\hat{\rho}((X_1, Y_1), \dots, (X_n, Y_n)) = \frac{\widehat{\text{cov}}(X_i, Y_i)}{S_X S_Y} = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

y el denominador es el producto de los desvíos muestrales de cada muestra.

- a) Sean $a, b \in \mathbb{R}$ constantes.

i. Definimos $X_i^* = X_i + a$, $i = 1, \dots, n$. Probar que $\hat{\rho}(X_i^*, Y_i) = \hat{\rho}(X_i, Y_i)$.

ii. Definimos $X_i^* = bX_i + a$, $i = 1, \dots, n$. Probar que $\hat{\rho}(X_i^*, Y_i) = \hat{\rho}(X_i, Y_i)$ si $b > 0$ y $\hat{\rho}(X_i^*, Y_i) = -\hat{\rho}(X_i, Y_i)$ si $b < 0$.

- b) Si tomamos $X_i^* = \frac{X_i - \bar{X}}{S_X}$, $i = 1, \dots, n$, probar que $\hat{\rho}(X_i^*, Y_i) = \hat{\rho}(X_i, Y_i)$.

4. El conjunto de datos `datasaurus_dozen` del paquete `datasauRus` tiene 13 conjuntos de observaciones bivariadas (X_i, Y_i) distintos. El data set al que pertenecen está codificado en la variable categórica "`dataset`". El conjunto de datos se carga al R con las instrucciones

```
library(datasauRus)
```

```
data(package="datasauRus")
```

- a) Realizar el scatter plot de las `x` e `y` cuyo `dataset` es "`dino`". Calcular las medias muestrales y los desvíos estándares muestrales de dichas `x` e `y`, y la correlación entre ambas. Las siguientes instrucciones pueden ser útiles.

```
aa <- "dino"
plot(datasaurus_dozen$x[datasaurus_dozen$dataset == aa],
     datasaurus_dozen$y[datasaurus_dozen$dataset == aa],
     main = c("datasaurus_dozen ", paste(aa)), xlab = "x", ylab = "y", pch = 20)
```

- b) Repetir (a) para las `x` e `y` cuyo `dataset` es "`star`".

- c) Repetir (a) para las `x` e `y` cuyo `dataset` es "`circle`".

- d) Repetir (a) para las `x` e `y` cuyo `dataset` es "`slant_up`".

- e) Repetir (a) para las `x` e `y` cuyo `dataset` es "`x_shape`".

- f) Consultar el help de `datasaurus_dozen` para ver instrucciones de cómo plotearlos todos. Concluir. La moraleja de este ejercicio es que siempre que podamos, debemos hacer scatterplots de nuestros conjuntos de datos, previamente a resumirlos con medidas de resumen numéricas.

5. El conjunto de datos `anscombe` que está en el R base, corresponde a cuatro conjuntos de datos generados y publicados en 1973 por Francis Anscombe. Se los conoce con el nombre del cuarteto de Anscombe. Corresponde a 4 conjuntos de 11 datos bivariados cada uno.

- a) Cargar los datos al R. Hacer cuatro scatterplots, graficando la `y` versus la `x` de cada uno.

- b) Calcular las medias muestrales y los desvíos estándares muestrales de dichas `x` e `y` separadas por grupo, y la correlación entre ambas.

Regresión lineal simple

6. Sean $(X_i, Y_i)_{1 \leq i \leq n}$ observaciones que siguen el modelo lineal

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (1)$$

donde $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, es el vector que tiene los errores. Asumimos que $E(\varepsilon) = \mathbf{0}$.

- a) Transformamos las X_i en $X_i^* = X_i - \bar{X}$, $i = 1, \dots, n$, es decir, las centramos. Indicar cómo cambia esto los parámetros del nuevo modelo lineal para (X_i^*, Y_i) ,

$$Y_i = \beta_0^* + \beta_1^* X_i^* + \varepsilon_i^*, \quad i = 1, 2, \dots, n, \quad (2)$$

Es decir, escribir a β_0^* y a β_1^* en términos de β_0 y a β_1 .

Sugerencia: A partir de (1) sumar y restar términos adecuados hasta obtener (2). ¿Cambian ambos parámetros, o alguno queda igual?

- b) ¿Cómo cambian sus estimadores? ¿Cambian ambos estimadores, o alguno queda igual?
- c) Ahora centremos también a las Y_i . Sean $Y_i^* = Y_i - \bar{Y}$, $i = 1, \dots, n$. Responder (a) y (b) para las (X_i^*, Y_i^*) ,

7. Problema de simulación 1.

- a) Generar $n = 40$ datos de la siguiente manera. Tomar $X_i \sim \mathcal{E}(1)$ y definir

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \dots, n$$

donde $\beta_0 = 5$, $\beta_1 = -2$ y $\varepsilon_i \sim N(0, \sigma^2 = 3)$. Graficar los datos. Repetir varias veces y observar la forma de los scatterplots.

- b) Repetir (a) pero ahora tomar los errores con distribución $U(-3, 3)$.
- c) Repetir (a) pero ahora tomar los errores con distribución: $\varepsilon_i - 30 \sim \Gamma(\alpha = 0.3, \lambda = 0.01)$.
- d) Repetir (a) pero ahora tomar los errores con distribución $N(0, \sigma^2 = 25)$.