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'')
```

   a) 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?

   b) 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?

   c) Repetir con las de menor correlación.

   d) 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?

   e) Hacer scatter plots de la variable `bia_di`, que es la distancia biacromial (informalmente, la distancia emtre 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

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

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

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

      i. Definimos $X_i^* = X_i + a$, $i = 1,\ldots,n$. Probar que $\widehat{\text{cov}}\left(X_i^*,Y_i\right) = \widehat{\text{cov}}\left(X_i,Y_i\right)$.

      ii. Definimos $X_i^* = bX_i + a$, $i = 1,\ldots,n$. Probar que $\widehat{\text{cov}}\left(X_i^*,Y_i\right) = b\,\widehat{\text{cov}}\left(X_i,Y_i\right)$.

   b) Sean $X_i^* = X_i - \overline{X}$ y $Y_i^* = Y_i - \overline{Y}$ $i = 1,\ldots,n$. Probar que $\widehat{\text{cov}}\left(X_i^*,Y_i^*\right) = \widehat{\text{cov}}\left(X_i^*,Y_i\right) = \widehat{\text{cov}}\left(X_i,Y_i\right)$.

   c) Probar que vale lo siguiente:

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

   d) Probar que la covarianza muestral puede escribirse de la siguiente forma

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

   e) Probar que

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

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

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

$$\widehat{\rho}\left((X_1,Y_1),\ldots,(X_n,Y_n)\right) = \frac{\widehat{\mathrm{cov}}\left(X_i,Y_i\right)}{S_X\, S_Y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}$$

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

   a) Sean $a, b \in \mathbb{R}$ constantes.
      i. Definimos $X_i^* = X_i + a$, $i = 1,\ldots,n$. Probar que $\widehat{\rho}\left(X_i^*, Y_i\right) = \widehat{\rho}\left(X_i, Y_i\right)$.
      ii. Definimos $X_i^* = bX_i + a$, $i = 1,\ldots,n$. Probar que $\widehat{\rho}\left(X_i^*, Y_i\right) = \widehat{\rho}\left(X_i, Y_i\right)$ si $b > 0$ y $\widehat{\rho}\left(X_i^*, Y_i\right) = \widehat{\rho}\left(X_i, Y_i\right)$ si $b < 0$.

   b) Si tomamos $X_i^* = \frac{X_i - \overline{X}}{S_X}$, $i = 1,\ldots,n$, probar que $\widehat{\rho}\left(X_i^*, Y_i\right) = \widehat{\rho}\left(X_i, Y_i\right)$.

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 pertencen está codficado 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

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

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

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

$$Y_i = \beta_0^* + \beta_1^* X_i^* + \varepsilon_i^*, \quad i = 1, 2, \ldots, n, \tag{2}$$

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 $Y_i^* = Y_i - \overline{Y}$, $i = 1, \ldots, n$. 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

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

donde $\beta_0 = 5$, $\beta_1 = -2$ y $\varepsilon_i \sim N(0, \sigma^2 = 3)$. ¿Cuánto vale la media y la varianza de los errores? Graficar los datos. Repetir varias veces y observar la forma de los scatterplots. ¿Es razonable ajustar estos datos con un modelo lineal?

b) Repetir (a) pero ahora tomar los errores con distribución $U(-3, 3)$. ¿Cuánto vale la media y la varianza de los errores?

c) Repetir (a) pero ahora tomar los errores con distribución: $\varepsilon_i + 3 \sim \Gamma(\alpha = 3, \lambda = 1)$. ¿Cuánto vale la media y la varianza de los errores? Hacer un gráfico de la densidad (la función `curve` de `R` puede ser útil). Observar que para generar una gamma en `R` con un valor λ_0 prefijado, debemos setear el argumento `rate` $= \lambda_0$ y el valor de α se setea con el argumento `shape` de la función `rgamma`. ¿Es razonable pensar que un modelo lineal simple ajustará bien a estos datos?

d) Repetir (c) pero con $\beta_1 = -400$. ¿Qué cambia en el gráfico? ¿Es razonable pensar que un modelo lineal simple ajustará bien a estos datos?

e) Repetir (a) pero ahora tomar los errores con distribución $N(0, \sigma^2 = 25)$. ¿Es razonable pensar que un modelo lineal simple ajustará bien a estos datos?

8. **Problema de simulación 2.**

Generar $n = 20$ datos de la siguiente manera. Tomar $X_i \sim U(0, 1)$ y dejarlos fijos y definir

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

donde $\beta_0 = 5$, $\beta_1 = 3$ y $\varepsilon_i \sim N(0, 1)$. Graficar los datos. Ajustar un modelo lineal y agregar al gráfico la recta de regresión estimada, los valores ajustados y el punto $(\overline{x}, \overline{y})$.

a) Repetir el experimento $N = 1000$ veces y guardar los valores de $\widehat{\beta}_1$ que se obtienen cada vez, $\widehat{\beta}_1^{(1)}, \widehat{\beta}_1^{(2)}, \ldots, \widehat{\beta}_1^{(N)}$. ¿Cuál es la media (muestral) de estos valores? ¿Cuál su varianza muestral? Realizar un histograma de $\widehat{\beta}_1^{(1)}, \widehat{\beta}_1^{(2)}, \ldots, \widehat{\beta}_1^{(N)}$. Grafique un estimador de la densidad a los $\{\widehat{\beta}_1^{(j)}\}_{1 \leq j \leq 1000}$. ¿Qué distribución parecen tener?

b) Repetir (a) con $n = 5$.

c) Repetir (a) con $n = 20$ y los errores con distribución $\epsilon_i - 1 \sim \mathcal{E}(\lambda = 1)$.

d) Repetir (a) con $n = 5$ y los errores con distribución $\epsilon_i - 1 \sim \mathcal{E}(\lambda = 1)$.