PRÁCTICA 3

Usaremos la siguiente notación para el modelo lineal múltiple:

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{n\times p}{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \tag{1}$$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Cada fila de la matriz X corresponde a las observaciones correspondientes a cada individuo (la fila i-ésima contiene las observaciones del individuo i-ésimo) y las columnas identifican a las variables.

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{X}\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

donde

- $\mathbf{Y}$ es un vector de respuestas
- $\boldsymbol{\beta}$ es un vector de parámetros
- X es una matriz de covariables
- $\boldsymbol{\varepsilon}$ es un vector de variables aleatorias

1. Notamos a la fila i–ésima de la matriz X (es decir, a la fila de covariables para el i–esimo individuo) por $\mathbf{x}_i^T = [1\ X_{i1}\ X_{i2}\ \cdots\ X_{i,p-1}]$, o lo que es lo mismo

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,(p-1)} \end{bmatrix}$$

Entonces

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

Asumimos que la matriz X tiene rango p, y que $n \geq p$.

a) ¿Qué dimensión tiene el producto $\mathbf{x}_i\mathbf{x}_i^T$?

b) Verificar que $X^TX = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$

c) ¿Cuánto vale X^TX cuando $p = 2$?

d) Hallar $(X^TX)^{-1}$ cuando $p = 2$. Para ello, recordar en el caso 2×2, si $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ su inversa puede obtenerse mediante $A^{-1} = \frac{1}{\det(A)}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ donde el $\det(A) = ad - bc$.[1]

[1] Puede ser útil recordar estas propiedades que probamos en la Práctica 1,

- $\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2$ y
- $\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \sum_{i=1}^n \left(X_i - \overline{X}\right) Y_i = \sum_{i=1}^n X_iY_i - n\overline{XY}$

e) Probar que la matriz $P = X(X^T X)^{-1} X^T$ es simétrica e idempotente, es decir, $P^2 = P$.

2. Probar, construyendo la matriz X y los vectores $\boldsymbol{Y}$ y $\widehat{\boldsymbol{\beta}}$ en el caso del modelo de regresión lineal simple, la expresión $\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \mathbf{Y}$ coincide con los estimadores de mínimos cuadrados previamente obtenida.

3. El conjunto de datos `prostate` del paquete `genridge` contiene observaciones de 10 variables medidas a 97 pacientes con cáncer de próstata.

    a) Ajustar un modelo lineal para explicar la variable respuesta `lpsa` (el logaritmo del antígeno prostático específico) a partir de `lcavol` (logaritmo del volumen del cáncer) y `lweight` (el logaritmo del peso de la próstata). Ajustarlo con el comando `library(lm)` con la opción `library(x = TRUE)` para que calcule la matriz X.

    b) Verificar que vale la propiedad del ejercicio ($1b$).

    c) Hallar los valores predichos, $\widehat{\mathbf{Y}}$. Calcular la correlación al cuadrado entre los valores predichos $\widehat{\mathbf{Y}}$ y los valores observados $\mathbf{Y}$ ¿Con qué numero del `summary` coincide esta cantidad?

    d) Hallar el vector de residuos. Comprobar que su promedio vale 0.

    e) Interpretar el modelo ajustado.