

Problemas propuestos de Regresión Lineal Múltiple

Francisco Javier Mercader Martínez

```
knitr::opts_chunk$set(echo = TRUE)
```

Problema 1

En el fichero **cemento_RLM.xlsx**, contiene los datos correspondientes a la presencia (en %) de cuatro componentes químicos en un tipo de cemento, así como el calor emitido (en calorías por gramo de cemento) durante el proceso de endurecimiento. Se desea proponer un modelo que permita predecir el calor emitido en función de los componentes químicos presentes del cemento.

```
library(readxl)
cemento <- read_excel("../data/cemento_RLM.xlsx")
```

- 1) Realiza un análisis descriptivo previo de las variables del problema y comenta los resultados más relevantes. ¿Podemos suponer que nuestra variable respuesta es Normal?

```
colnames(cemento) <- c("A", "B", "C", "D", "Calor")
```

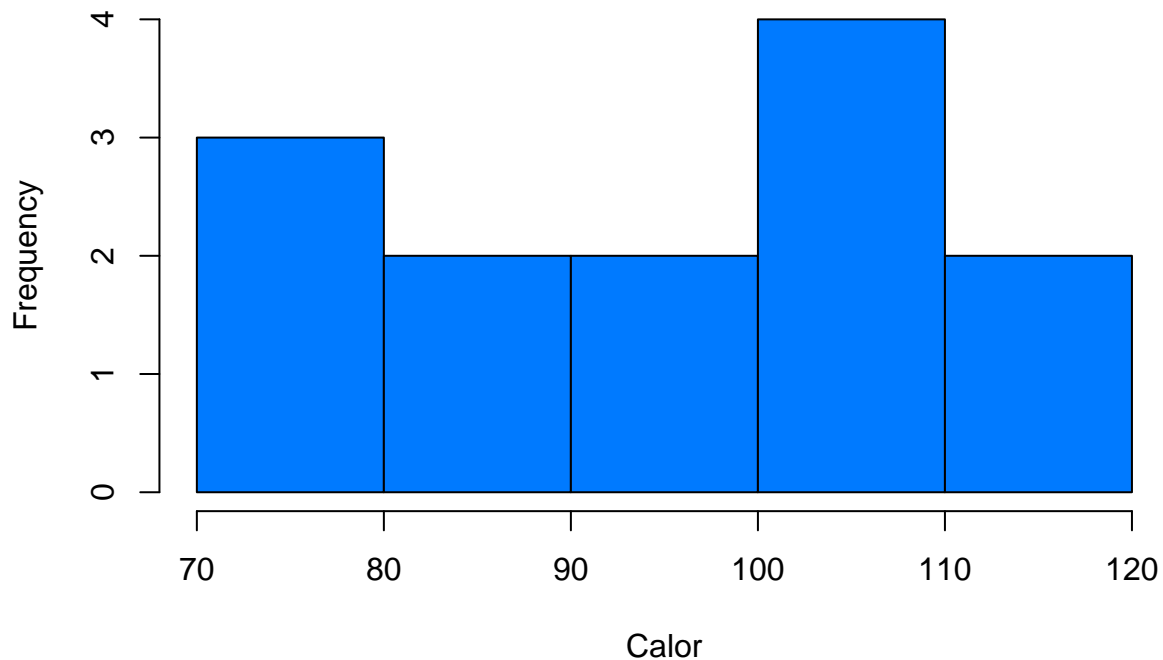
```
# Análisis descriptivo
summary(cemento)
```

##	A	B	C	D	Calor
## Min.	: 1.000	Min. :26.00	Min. : 4.00	Min. : 6	Min. : 72.50
## 1st Qu.:	2.000	1st Qu.:31.00	1st Qu.: 8.00	1st Qu.:20	1st Qu.: 83.80
## Median :	7.000	Median :52.00	Median : 9.00	Median :26	Median : 95.90
## Mean :	7.462	Mean :48.15	Mean :11.77	Mean :30	Mean : 95.42
## 3rd Qu.:	11.000	3rd Qu.:56.00	3rd Qu.:17.00	3rd Qu.:44	3rd Qu.:109.20
## Max.	:21.000	Max. :71.00	Max. :23.00	Max. :60	Max. :115.90

```
# Histograma de la variable respuesta
```

```
hist(cemento$Calor, main = "Histograma de Calor", xlab = "Calor", col = "#007AFF", border = "black")
```

Histograma de Calor



- 2) Calcula la matriz de correlaciones de las cinco variables. ¿Qué información proporciona esta matriz? ¿Qué regresores del modelo presentan una más estrecha relación lineal entre sí? ¿Cuál es la primera variable que debería entrar en el modelo?

```
cor(cemento)
```

```
##           A           B           C           D           Calor
## A      1.0000000  0.2285795 -0.8241338 -0.2454451  0.7307175
## B      0.2285795  1.0000000 -0.1392424 -0.9729550  0.8162526
## C     -0.8241338 -0.1392424  1.0000000  0.0295370 -0.5346707
## D     -0.2454451 -0.9729550  0.0295370  1.0000000 -0.8213050
## Calor  0.7307175  0.8162526 -0.5346707 -0.8213050  1.0000000
```

La matriz de correlaciones nos proporciona información sobre la relación lineal entre las variables.

- Las variables B y D tienen la correlación más fuerte entre sí (-0.9729550), lo que indica una fuerte relación lineal negativa.
 - La variable Calor (la variable de respuesta) tiene la correlación más fuerte con la variable B (0.8162526), seguida de la variable A (0.7307175).
- 3) Realiza la selección del modelo mediante regresión por pasos, hacia delante y hacia atrás. Indica el orden de entrada y salida de las variables para cada uno de los métodos. Comenta los resultados obtenidos.

```
# Regresión hacia delante
full.modelo <- lm(Calor ~ ., data = cemento)
```

- 4) Estudia si hay colinealidad entre los regresores de los modelos resultantes en el apartado anterior y en caso afirmativo explica cuál es tu decisión para solventarlo.

- 5) ¿Propondrías un único modelo o varios? ¿Cuál o cuáles y por qué?
- 6) Determina el (los) modelo(s) ajustado(s) y los intervalos de confianza al 95% para los parámetros de regresión.
- 7) Para el modelo que contempla sólo los regresores A y D, estudia si se verifican las hipótesis del modelo de regresión múltiple, comentando los procesos utilizados. Estudia si hay colinealidad entre los regresores y si aparecen observaciones influyentes, comentando los procesos utilizados. En caso de que se presente alguno de estos problemas, explica cuál es tu decisión para solventarlo.
- 8) Obtén una estimación puntual del calor emitido por el cemento sabiendo que $A=15$, $B=39$, $C=4.5$ y $D=40$. Determina también un intervalo de confianza para el calor emitido en ese caso, así como un intervalo de predicción. ¿Podemos concluir que el calor emitido por el cemento superará las 95 cal/gr? ¿Y en promedio?
- 9) Responde a la cuestión anterior sabiendo que $A=45$ y $D=40$.