PROBLEMAS PROPUESTOS DE REGRESIÓN LINEAL MÚLTIPLE
ANÁLISIS ESTADÍSTICO MULTIVARIANTE
GRADO EN CIENCIA E INGENIERÍA DE DATOS

PROBLEMA 1

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 en el cemento.

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?
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?
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.
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.

PROBLEMA 2

En el fichero **motor.dat** se encuentran los datos correspondientes a 200 ensayos, donde se midieron las siguientes variables: VRP (velocidad de rotación primaria), VRS (velocidad de rotación secundaria), Presion (presión), Temp_Esc (temperatura de escape), Temp_Amb (temperatura ambiente a la hora de efectuar la prueba), LN_RFC (logaritmo neperiano de la rapidez de flujo de combustible) y Empuje (empuje del motor). Se desea proponer un modelo que permita predecir el "Empuje del motor" en función del resto de variables, analizando si serían necesarias todas o no.

1) Indica la variable respuesta y los regresores del problema. Las variables del problema, ¿presentan datos atípicos? NO eliminies ningún dato. ¿Podemos suponer que nuestra variable respuesta es Normal? En caso negativo, justificar si la transformación logarítimica sería adecuada y realizarla.

2) Calcula la matriz de correlaciones de las variables del problema. ¿Existen regresores altamente correlados dos a dos?. ¿Cuál es la primera variable que debería entrar en el modelo? (indica el coeficiente de correlación en cada caso e interprétalo).

3) Realiza la selección del modelo mediante regresión por pasos, hacia delante y hacia atrás. Para cada uno de los tres métodos, indica el modelo teórico resultante y estudia si existe multicolinealidad.

4) ¿Qué modelo(s) de regresión propondrías y por qué? Indica el modelo ajustado que explica el "empuje del motor" y comenta la bondad del ajuste.

5) Para el modelo propuesto, estudia si se verifican las hipótesis del modelo de regresión múltiple y si existen observaciones influyentes. Comenta los procesos utilizados.

6) Porporciona una estimación puntual del "empuje del motor" para un ensayo de las siguientes características:

VRP= 2000, VRS=19000, LN_RFC= 10.3089, Presion = 180, Temp_Esc = 1700 y Temp_Amb= 95.

Determinar también un intervalo predicción individual para el "empuje" en ese caso, así como un intervalo de confianza para el "empuje" promedio. ¿Podemos concluir que el "empuje del motor" será superior a 4000? ¿Y en promedio para los ensayos de esas características?

PROBLEMA 3

Con los datos del Problema 1 (fichero **cemento_RLM.xlsx)**, responder a las siguientes cuestiones:

1) Obtener la ecuación del modelo ajustado por mínimos cuadrados usando todos los predictores. Realizar el ajuste de tres formas diferentes:
    a. Primero con la función *lm()* de R.
    b. Después usando la inversa de (t(M)*M), siendo M la matriz de diseño y t(M) su traspuesta.
    c. Por último, usando el método GD (Gradiente Decendente). En este caso, debes probar con diferentes valores del número de iteraciones, learning rate y valores iniciales.

    ¿Se obtienen los mismos resultados? ¿A qué se debe?

2) Repetir el apartado anterior, pero usando sólo los predictores A y D.