Problemas propuestos de Regresión Lineal Múltiple

Francisco Javier Mercader Martínez

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")
print.data.frame(cemento)
```

```
##     A  B  C  D  HEAT
## 1   7 26  6 60  78.5
## 2   1 29 15 52  74.3
## 3  11 56  8 20 104.3
## 4  11 31  8 47  87.6
## 5   7 52  6 33  95.9
## 6  11 55  9 22 109.2
## 7   3 71 17  6 102.7
## 8   1 31 22 44  72.5
## 9   2 54 18 22  93.1
## 10 21 47  4 26 115.9
## 11  1 40 23 34  83.8
## 12 11 66  9 12 113.3
## 13 10 68  8 12 109.4
```

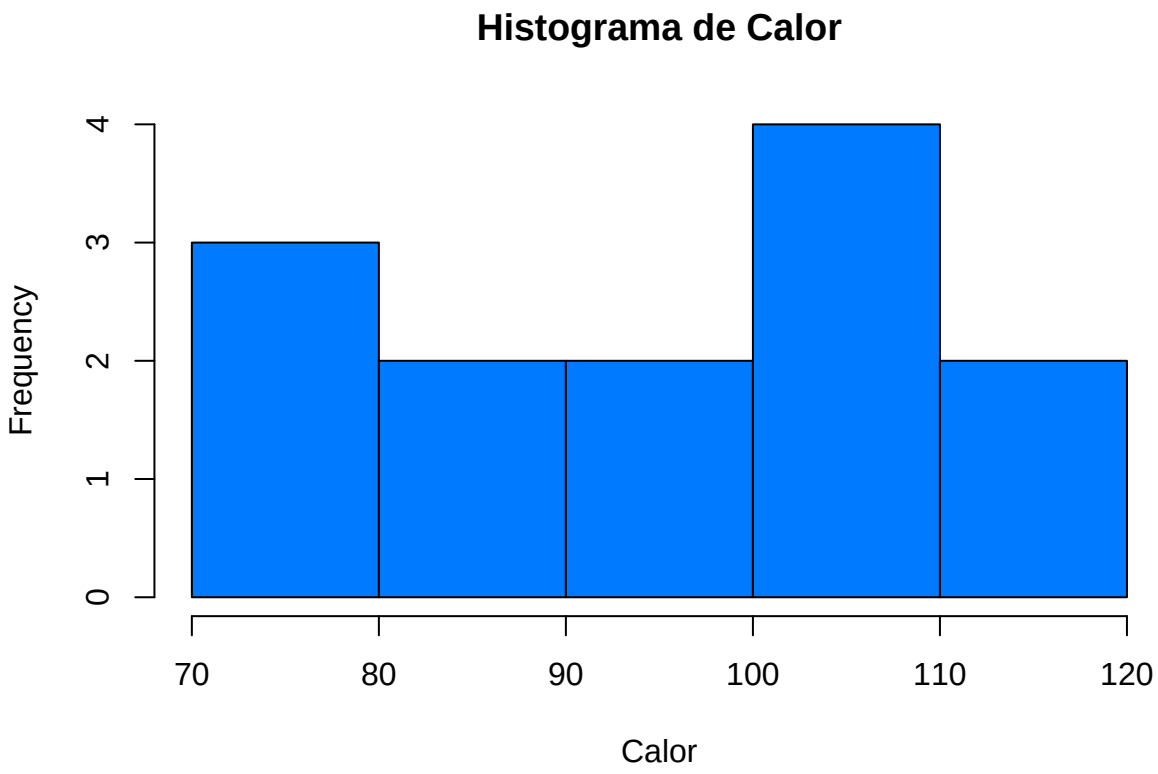
1) Realiza un análisis descriptivo previo de las variables del problema y comenta los resultados más relevantes. ¿Podemos suponer que nuestra variable respuesta es Normal?

```
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")
```



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.

```
# Ajustar el modelo de regresión lineal completo
modelo_completo <- lm(Calor ~ ., data = cemento)

# Selección de modelo hacia adelante
```

```r
modelo_forward <- step(lm(Calor ~ 1, data = cemento), scope = list(lower = ~1, upper =
 ↪  modelo_completo), direction = "forward")
```

```
## Start:  AIC=71.44
## Calor ~ 1
##
##        Df Sum of Sq     RSS    AIC
## + D     1   1831.90  883.87 58.852
## + B     1   1809.43  906.34 59.178
## + A     1   1450.08 1265.69 63.519
## + C     1    776.36 1939.40 69.067
## <none>               2715.76 71.444
##
## Step:  AIC=58.85
## Calor ~ D
##
##        Df Sum of Sq    RSS    AIC
## + A     1    809.10  74.76 28.742
## + C     1    708.13 175.74 39.853
## <none>               883.87 58.852
## + B     1     14.99 868.88 60.629
##
## Step:  AIC=28.74
## Calor ~ D + A
##
##        Df Sum of Sq    RSS    AIC
## + B     1    26.789 47.973 24.974
## + C     1    23.926 50.836 25.728
## <none>               74.762 28.742
##
## Step:  AIC=24.97
## Calor ~ D + A + B
##
##        Df Sum of Sq    RSS    AIC
## <none>               47.973 24.974
## + C     1   0.10909 47.864 26.944
```

```r
# Selección de modelo hacia atrás
modelo_backward <- step(modelo_completo, direction = "backward")
```

```
## Start:  AIC=26.94
## Calor ~ A + B + C + D
##
##        Df Sum of Sq    RSS    AIC
## - C     1    0.1091 47.973 24.974
## - D     1    0.2470 48.111 25.011
## - B     1    2.9725 50.836 25.728
## <none>               47.864 26.944
## - A     1   25.9509 73.815 30.576
##
## Step:  AIC=24.97
## Calor ~ A + B + D
##
##        Df Sum of Sq    RSS    AIC
```

```
## <none>               47.97 24.974
## - D     1      9.93  57.90 25.420
## - B     1     26.79  74.76 28.742
## - A     1    820.91 868.88 60.629
```

```
# Imprimir los modelos
summary(modelo_forward)
```

```
##
## Call:
## lm(formula = Calor ~ D + A + B, data = cemento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0919 -1.8016  0.2562  1.2818  3.8982
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.6483    14.1424   5.066 0.000675 ***
## D            -0.2365     0.1733  -1.365 0.205395
## A             1.4519     0.1170  12.410 5.78e-07 ***
## B             0.4161     0.1856   2.242 0.051687 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.309 on 9 degrees of freedom
## Multiple R-squared:  0.9823, Adjusted R-squared:  0.9764
## F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

```
summary(modelo_forward)
```

```
##
## Call:
## lm(formula = Calor ~ D + A + B, data = cemento)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0919 -1.8016  0.2562  1.2818  3.8982
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.6483    14.1424   5.066 0.000675 ***
## D            -0.2365     0.1733  -1.365 0.205395
## A             1.4519     0.1170  12.410 5.78e-07 ***
## B             0.4161     0.1856   2.242 0.051687 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.309 on 9 degrees of freedom
## Multiple R-squared:  0.9823, Adjusted R-squared:  0.9764
## F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

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.

```r
# Comprobar la colinealidad

vif(modelo_completo)
```

```
##         A         B         C         D
##  38.49621 254.42317  46.86839 282.51286
```

```r
vif(modelo_forward)
```

```
##        D        A        B
## 18.94008  1.06633 18.78031
```

```r
vif(modelo_backward)
```

```
##        A        B        D
##  1.06633 18.78031 18.94008
```

En el `modelo_completo` y el `modelo_forward`, todos los regresores tienen un VIF muy alto, lo que indica una fuerte colinealidad. Para solucionar esto, podrías considerar eliminar uno o más de los regresores, o combinarlos de alguna manera si tiene sentido en el contexto de tus datos.

En el `modelo_backward`, los regresores A tienen un VIF bajo, lo que indica que no hay colinealidad. Sin embargo, B y D tienen un VIF mayor a 5, lo que sugiere alguna colinealidad.

5) ¿Propondrías un único modelo o varios? ¿Cuál o cuáles y por qué?

El `modelo_backward` muestra una colinealidad moderada entre las variables B y D, pero la variable A no muestra colinealidad. Por lo tanto, este modelo puede ser más adecuado para la predicción.

6) Determina el (los) modelo(s) ajustado(s) y los intervalos de confianza al 95% para los parámetros de regresión.

```r
# Modelo ajustado
modelo_backward
```

```
##
## Call:
## lm(formula = Calor ~ A + B + D, data = cemento)
##
## Coefficients:
## (Intercept)            A            B            D
##     71.6483       1.4519       0.4161      -0.2365
```

```r
confint(modelo_backward, level=0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept) 39.655990254 103.6406237
## A            1.187271016   1.7166049
## B           -0.003770331   0.8359899
## D           -0.628544442   0.1554640
```

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.

```r
# Comprobar la colinealidad
modelo_ajustado <- lm(Calor ~ A + D, data = cemento)
vif(modelo_ajustado)
```

```
##        A        D
## 1.064105 1.064105
```

Los valores de `vif` para los regresores A y D son ambos 1.064105. En este caso, los valores de `vif` son muy bajos, lo que indica que no hay colinealidad entre los regresores A y D. Por lo tanto, no es necesario tomar ninguna medida para tratar la colinealidad en este modelo.

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?

```
predict(modelo_ajustado, newdata = data.frame(A = 15, D = 40), interval = "confidence",
 ↪  level = 0.95)
```

```
##        fit      lwr      upr
## 1 100.1386 96.87177 103.4055
```

```
predict(modelo_ajustado, newdata = data.frame(A = 15, D = 40), interval = "prediction",
 ↪  level = 0.95)
```

```
##        fit      lwr      upr
## 1 100.1386 93.22567 107.0515
```

El intervalo de confianza al 95% para el calor emitido por el cemento es (96.87177, 103.4055) cal/gr. El intervalo de predicción al 95% para el calor emitido por el cemento es (93.22567, 107.0515) cal/gr.

Dado que la estimación puntual del calor emitido por el cemento es de 100.1386 cal/gr, que está por encima de 95 cal/gr, podemos concluir que, en promedio, es probable que el calor emitido por el cemento supere las 95 cal/gr.

9) Responde a la cuestión anterior sabiendo que A=45 y D=40.

```
predict(modelo_ajustado, newdata = data.frame(A = 45, D = 40), interval = "confidence",
 ↪  level = 0.95)
```

```
##        fit      lwr      upr
## 1 143.3374 131.3281 155.3467
```

```
predict(modelo_ajustado, newdata = data.frame(A = 45, D = 40), interval = "prediction",
 ↪  level = 0.95)
```

```
##        fit      lwr      upr
## 1 143.3374 129.8711 156.8036
```

Dado que la estimación puntual del calor emitido por el cemento es de 143.3374 cal/gr, que está muy por encima de 95 cal/gr, podemos concluir que, en promedio, es más que probable que el calor emitido por el cemento supere las 95 cal/gr.

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.

```r
motor <- read.table("../../data/motor.dat", header = TRUE)
```

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 logarítmica sería adecuada y realizarla.

```r
summary(motor) # Los valores de la columna EMPUJE son caracteres, por lo que necesitamos
 ↪  convertirlos a numéricos
```

```
##       VRP             VRS           PRESION         TEMP_ESC       TEMP_AMB
##  Min.   :1403   Min.   :17008   Min.   :130.0   Min.   :1500   Min.   : 85.00
##  1st Qu.:1586   1st Qu.:17822   1st Qu.:154.0   1st Qu.:1558   1st Qu.: 89.00
##  Median :1802   Median :19141   Median :173.5   Median :1630   Median : 94.00
##  Mean   :1835   Mean   :19028   Mean   :175.3   Mean   :1622   Mean   : 93.86
##  3rd Qu.:2094   3rd Qu.:20107   3rd Qu.:197.0   3rd Qu.:1678   3rd Qu.: 99.00
##  Max.   :2300   Max.   :20950   Max.   :220.0   Max.   :1749   Max.   :102.00
##     LN_RFC             EMPUJE
##  Length:200         Length:200
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```r
motor$EMPUJE <- gsub(",", ".", motor$EMPUJE) # Reemplazar las comas por puntos para que
 ↪  el programa los reconozca
options(digits = 10) # Aumentar el número de dígitos para ver los valores completos
motor$EMPUJE <- as.numeric(motor$EMPUJE) # Convertir la columna EMPUJE a numérica
summary(motor) # Verificar que la columna EMPUJE ahora es numérica
```

```
##       VRP              VRS             PRESION          TEMP_ESC
##  Min.   :1403.000   Min.   :17008.00   Min.   :130.00   Min.   :1500.00
##  1st Qu.:1585.750   1st Qu.:17822.25   1st Qu.:154.00   1st Qu.:1558.25
##  Median :1802.000   Median :19140.50   Median :173.50   Median :1630.00
##  Mean   :1835.045   Mean   :19027.88   Mean   :175.29   Mean   :1622.08
##  3rd Qu.:2093.750   3rd Qu.:20106.75   3rd Qu.:197.00   3rd Qu.:1678.00
##  Max.   :2300.000   Max.   :20950.00   Max.   :220.00   Max.   :1749.00
##     TEMP_AMB          LN_RFC             EMPUJE
##  Min.   : 85.000   Length:200         Min.   :2808.546
##  1st Qu.: 89.000   Class :character   1st Qu.:3848.581
##  Median : 94.000   Mode  :character   Median :4315.945
##  Mean   : 93.865                      Mean   :4359.991
##  3rd Qu.: 99.000                      3rd Qu.:4838.706
##  Max.   :102.000                      Max.   :6378.240
```

```r
shapiro.test(motor$EMPUJE)
```

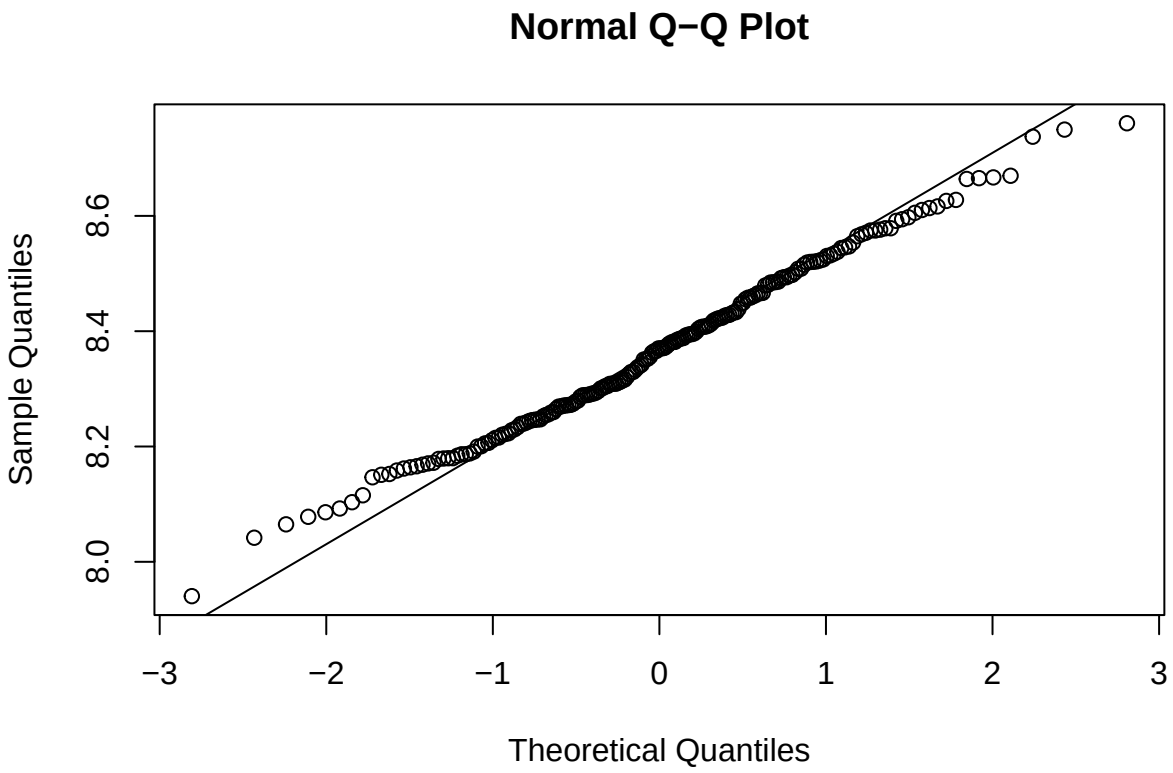
```
##
##  Shapiro-Wilk normality test
##
## data:  motor$EMPUJE
## W = 0.97831071, p-value = 0.003448391
```

El `p-value` es 0.003448391, es menor que 0.05. Por lo tanto, rechazaríamos la hipótesis nula y concluiríamos que los datos no están normalmente distribuidos.

```r
motor$EMPUJE <- log(motor$EMPUJE)
shapiro.test(motor$EMPUJE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  motor$EMPUJE
## W = 0.99289164, p-value = 0.4462751
```

```r
qqnorm(motor$EMPUJE)
qqline(motor$EMPUJE)
```



Utilizando la tranformación logarítmica los valores de la variable respuesta se han normalo y hemos obtenido `p-value` con un valor mayor que 0.05, concluyendo que ahora los datos están normalmente distribuidos.

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).

```r
# Para poder hacer la matriz de correlaciones primero habría que convertir en valores
# numéricos la columma LN_RFC ya que nos encontramos el mismo problema que con la
# columma EMPUJE
motor$LN_RFC <- gsub(',', '.', motor$LN_RFC)
motor$LN_RFC <- as.numeric(motor$LN_RFC)
cor(motor)
```

```
##                    VRP             VRS        PRESION       TEMP_ESC
## VRP       1.00000000000 -0.017426715437  0.05409888834  0.02254576302
```

```
## VRS      -0.01742671544  1.000000000000  0.04551434622  0.02927089880
## PRESION   0.05409888834  0.045514346221  1.00000000000 -0.09176940580
## TEMP_ESC  0.02254576302  0.029270898800 -0.09176940580  1.00000000000
## TEMP_AMB -0.01405090116 -0.119992371325 -0.15821952684 -0.07087704285
## LN_RFC   -0.01277237617 -0.004834900563 -0.10585768483 -0.06196525548
## EMPUJE    0.02237367287  0.064205503639  0.84392883530  0.14832414747
##               TEMP_AMB          LN_RFC         EMPUJE
## VRP      -0.01405090116 -0.012772376166  0.02237367287
## VRS      -0.11999237133 -0.004834900563  0.06420550364
## PRESION  -0.15821952684 -0.105857684826  0.84392883530
## TEMP_ESC -0.07087704285 -0.061965255478  0.14832414747
## TEMP_AMB  1.00000000000  0.216247834698 -0.15570056528
## LN_RFC    0.21624783470  1.000000000000 -0.24113467616
## EMPUJE   -0.15570056528 -0.241134676156  1.00000000000
```

Mirando tu matriz de correlaciones, la correlación más alta es entre las variables `PRESION` y `EMPUJE`, con un coeficiente de correlación de 0.84392883530. Esto indica una fuerte correlación positiva entre estas dos variables, lo que significa que cuando `PRESION` aumenta, `EMPUJE` también tiende a aumentar, y viceversa.

Por lo tanto, `PRESION` sería la primera variable que debería entrar en el modelo, ya que es la que tiene la correlación más alta con la variable de respuesta.

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.

```
modelo_completo <- lm(EMPUJE ~ ., data = motor)
modelo_forward <- step(modelo_completo, scope = list(lower = ~ 1, upper =
 ↪  modelo_completo), direction = "forward")
```

```
## Start:  AIC=-1047.47
## EMPUJE ~ VRP + VRS + PRESION + TEMP_ESC + TEMP_AMB + LN_RFC
```

```
modelo_backward <- step(modelo_completo, direction = "backward")
```

```
## Start:  AIC=-1047.47
## EMPUJE ~ VRP + VRS + PRESION + TEMP_ESC + TEMP_AMB + LN_RFC
##
##            Df Sum of Sq       RSS         AIC
## - VRS       1 0.0020035 0.9930036 -1049.06766
## - TEMP_AMB  1 0.0033364 0.9943366 -1048.79938
## - VRP       1 0.0040770 0.9950772 -1048.65047
## <none>                  0.9910002 -1047.47158
## - LN_RFC    1 0.0892856 1.0802858 -1032.21835
## - TEMP_ESC  1 0.2175019 1.2085021 -1009.78714
## - PRESION   1 3.1969955 4.1879956  -761.21902
##
## Step:  AIC=-1049.07
## EMPUJE ~ VRP + PRESION + TEMP_ESC + TEMP_AMB + LN_RFC
##
##            Df Sum of Sq       RSS         AIC
## - TEMP_AMB  1 0.0028064 0.9958101 -1050.50322
## - VRP       1 0.0042014 0.9972050 -1050.22325
## <none>                  0.9930036 -1049.06766
## - LN_RFC    1 0.0886698 1.0816735 -1033.96160
## - TEMP_ESC  1 0.2187311 1.2117347 -1011.25287
## - PRESION   1 3.2055758 4.1985794  -762.71423
```

```
##
## Step:  AIC=-1050.5
## EMPUJE ~ VRP + PRESION + TEMP_ESC + LN_RFC
##
##            Df Sum of Sq       RSS         AIC
## - VRP       1 0.0042166 1.0000266 -1051.65815
## <none>                  0.9958101 -1050.50322
## - LN_RFC    1 0.0859013 1.0817113 -1035.95461
## - TEMP_ESC  1 0.2162675 1.2120776 -1013.19629
## - PRESION   1 3.2470543 4.2428643  -762.61576
##
## Step:  AIC=-1051.66
## EMPUJE ~ PRESION + TEMP_ESC + LN_RFC
##
##            Df Sum of Sq       RSS         AIC
## <none>                  1.0000266 -1051.65815
## - LN_RFC    1 0.0857092 1.0857358 -1037.21189
## - TEMP_ESC  1 0.2147855 1.2148122 -1014.74558
## - PRESION   1 3.2440710 4.2440976  -764.55763
```

```
# Comprobar la colinealidad
vif(modelo_completo)
```

```
##         VRP         VRS     PRESION    TEMP_ESC    TEMP_AMB      LN_RFC
## 1.004189049 1.017043151 1.047891635 1.020796633 1.090071202 1.058908827
```

```
vif(modelo_forward)
```

```
##         VRP         VRS     PRESION    TEMP_ESC    TEMP_AMB      LN_RFC
## 1.004189049 1.017043151 1.047891635 1.020796633 1.090071202 1.058908827
```

```
vif(modelo_backward)
```

```
##     PRESION    TEMP_ESC      LN_RFC
## 1.021358329 1.013805844 1.016660497
```

Todos los valores de `VIF` son muy bajos por lo que suponemos que no hay colinealidad.

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?