Estudio de Regresión Lineal Múltiple en los Resultados de los Aspirantes a Ingeniería Biomédica en la Universidad Autónoma de Chihuahua

Gerardo Pérez

4/12/2020

Para el presente trabajo fueron utilizadas las siguientes librerías:

```
library(tidyverse)
library(readr)
library(leaps)
```

Los datos utilizados para realizar el presente estudio fueron obtenidos de uno de los apartados de la página oficial de la Universidad Autónoma de Chihuahua: https://listas.uach.mx/. Dichos datos fueron importados, excluyendo las columnas que representan un problema de privacidad hacia los aspirantes, así como la primera columna que muestra la posición de cada aspirante ya que dichos datos se encuentran ordenados de forma descendente.

```
datos <- read_csv("datos.csv") %>% select(-N)
```

En el primer modelo realizado fueron utilizadas todas las variables presentadas, tomando como regresando la calificación global.

```
modelo1 <- lm(GLOBAL ~ ., datos)
summary(modelo1)
```

```
##
## Call:
## lm(formula = GLOBAL ~ ., data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4387  -0.3408  -0.0539   0.1988  14.8572
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.353903   1.723553  -1.366    0.174
## PMA          0.173348   0.002038  85.066   <2e-16 ***
## PAN          0.175362   0.001781  98.449   <2e-16 ***
## ELE          0.172059   0.001881  91.493   <2e-16 ***
## CLE          0.177687   0.001587 111.960   <2e-16 ***
## MAT          0.077967   0.001617  48.228   <2e-16 ***
```

```
## BIO          0.074682   0.001499  49.829   <2e-16 ***
## LES          0.074435   0.001289  57.756   <2e-16 ***
## ING          0.076661   0.001035  74.079   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.575 on 186 degrees of freedom
## Multiple R-squared:  0.9996, Adjusted R-squared:  0.9996
## F-statistic: 5.969e+04 on 8 and 186 DF,  p-value: < 2.2e-16
```

De este análisis podemos apreciar la distribución de los residuales, en donde notamos que el 50% de ellos se encuentra entre -0.3408 y 0.1988, así como notar que el residual mínimo es de -12.4387 y el máximo de 14.8572. Dado el comportamiento del 50% se prevé que estos valores son atípicos.

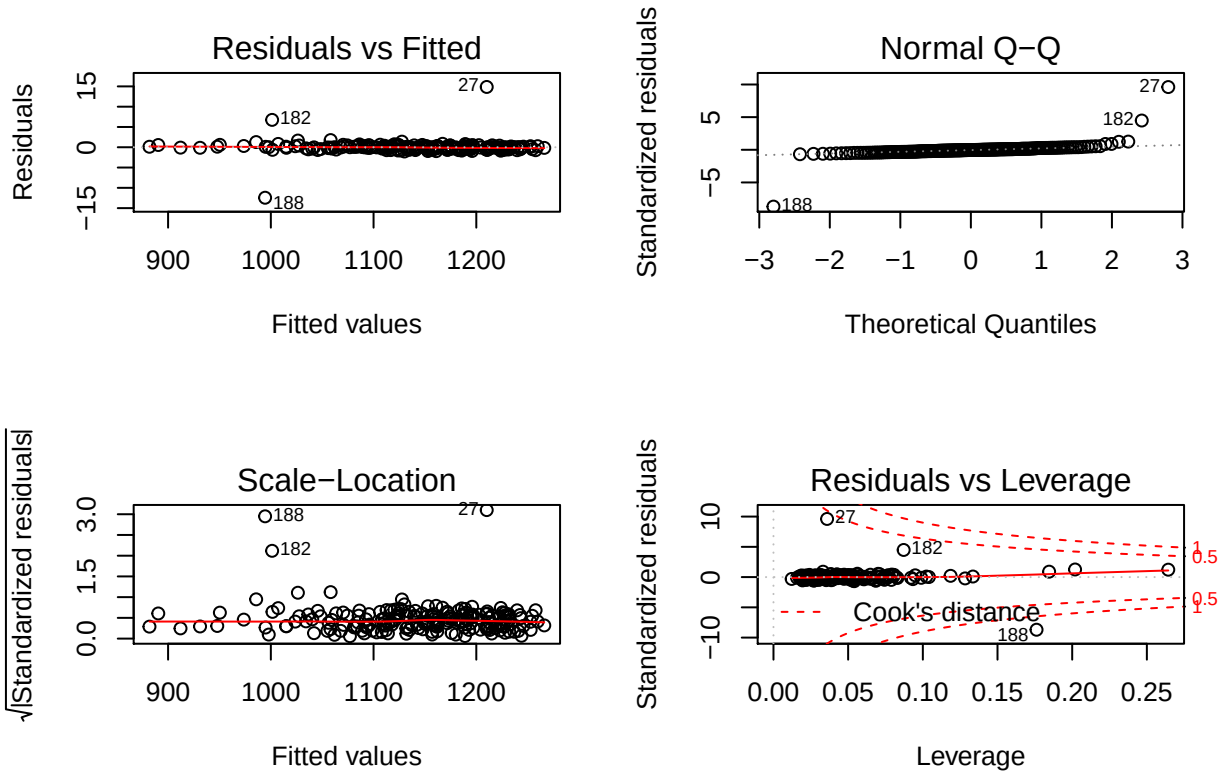
Por otro lado, podemos observar que las variables regresoras tienen un valor de P mucho menor a 0.05 lo que indica que estas variables entran dentro del nivel de significancia del 5% aceptado en este estudio. Además, dichas variables muestran un coeficiente positivo, lo que implica que el aumento en alguna de ellas representa un aumento general, relación lógica dado el estudio.

La ecuación, de esta forma, viene dada por

$$GLOBAL = -2.353903 + 0.173348PMA + 0.175362PAN + 0.172059ELE + 0.177687CLE + 0.077967MAT + 0.074682BIO + 0.074435LES + 0.076661ING$$

la cual se ajusta un 99.96% a los datos.

Los residuales se muestran en el siguiente gráfico:



Como se puede observar, los errores se distribuyen de formar adecuada según los supuestos del modelo y como se supuso inicialmente, existen valores atípicos en la posición 27 y 188. Estos valores se comportan

de esta manera por la variación entre los puntajes obtenidos mayormente en inglés, que contrastan con los puntajes previos.

Además se realizó una prueba de los mejores para determinar, bajo diversas circunstancias, qué modelo tiene un mejor comportamiento. Para ayudar en esta decisión, se creó una matriz que muestra el coeficiente de Mallows, la R^2 y la R^2 ajustada.

```
bs <- regsubsets(GLOBAL ~., datos, nbest = 2)
(bs_summary <- summary(bs))
```

```
## Subset selection object
## Call: regsubsets.formula(GLOBAL ~ ., datos, nbest = 2)
## 8 Variables  (and intercept)
##     Forced in Forced out
## PMA     FALSE      FALSE
## PAN     FALSE      FALSE
## ELE     FALSE      FALSE
## CLE     FALSE      FALSE
## MAT     FALSE      FALSE
## BIO     FALSE      FALSE
## LES     FALSE      FALSE
## ING     FALSE      FALSE
## 2 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          PMA PAN ELE CLE MAT BIO LES ING
## 1  ( 1 ) "*" " " " " " " " " " " " " "
## 1  ( 2 ) " " "*" " " " " " " " " " " "
## 2  ( 1 ) "*" " " "*" " " " " " " " " "
## 2  ( 2 ) " " "*" "*" " " " " " " " " "
## 3  ( 1 ) " " "*" " " "*" "*" " " " " "
## 3  ( 2 ) "*" " " "*" " " " " " " "*" " "
## 4  ( 1 ) " " "*" "*" "*" "*" " " " " " "
## 4  ( 2 ) "*" " " "*" "*" " " "*" " " " "
## 5  ( 1 ) "*" "*" "*" "*" " " " " " " "*"
## 5  ( 2 ) " " "*" "*" "*" "*" " " " " "*"
## 6  ( 1 ) "*" "*" "*" "*" " " " " "*" "*"
## 6  ( 2 ) "*" "*" "*" "*" " " "*" " " "*"
## 7  ( 1 ) "*" "*" "*" "*" " " "*" "*" "*"
## 7  ( 2 ) "*" "*" "*" "*" "*" " " "*" "*"
## 8  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*"
```

```
cbind(
  Cp = bs_summary$cp,
  R2 = bs_summary$rsq,
  adj_R2 = bs_summary$adjr2
)
```

```
##               Cp        R2    adj_R2
##  [1,] 149408.110 0.6868269 0.6852042
##  [2,] 151733.651 0.6819586 0.6803107
##  [3,]  80747.634 0.8305660 0.8288010
##  [4,]  82089.618 0.8277567 0.8259625
##  [5,]  42053.681 0.9115727 0.9101838
```

```
##  [6,]  47274.365 0.9006436 0.8990831
##  [7,]  25098.568 0.9470710 0.9459567
##  [8,]  26247.561 0.9446657 0.9435007
##  [9,]  14157.721 0.9699789 0.9691847
## [10,]  15064.266 0.9680811 0.9672367
## [11,]   5569.138 0.9879626 0.9875784
## [12,]   6946.224 0.9850798 0.9846036
## [13,]   2332.962 0.9947414 0.9945446
## [14,]   2489.921 0.9944128 0.9942037
## [15,]      9.000 0.9996106 0.9995939
```

Como es de suponer, la ecuación con todas las variables regresoras tiene un menor coeficiente de Mallows y una mayor R^2, sin embargo, bajo ciertas circunstancias un modelo con una menor cantidad de variables regresoras podría ser más útil.

Por cuestiones académicas escogí el onceavo modelo, el cual considera PMA, PAN, ELE, CLE, LES e ING.

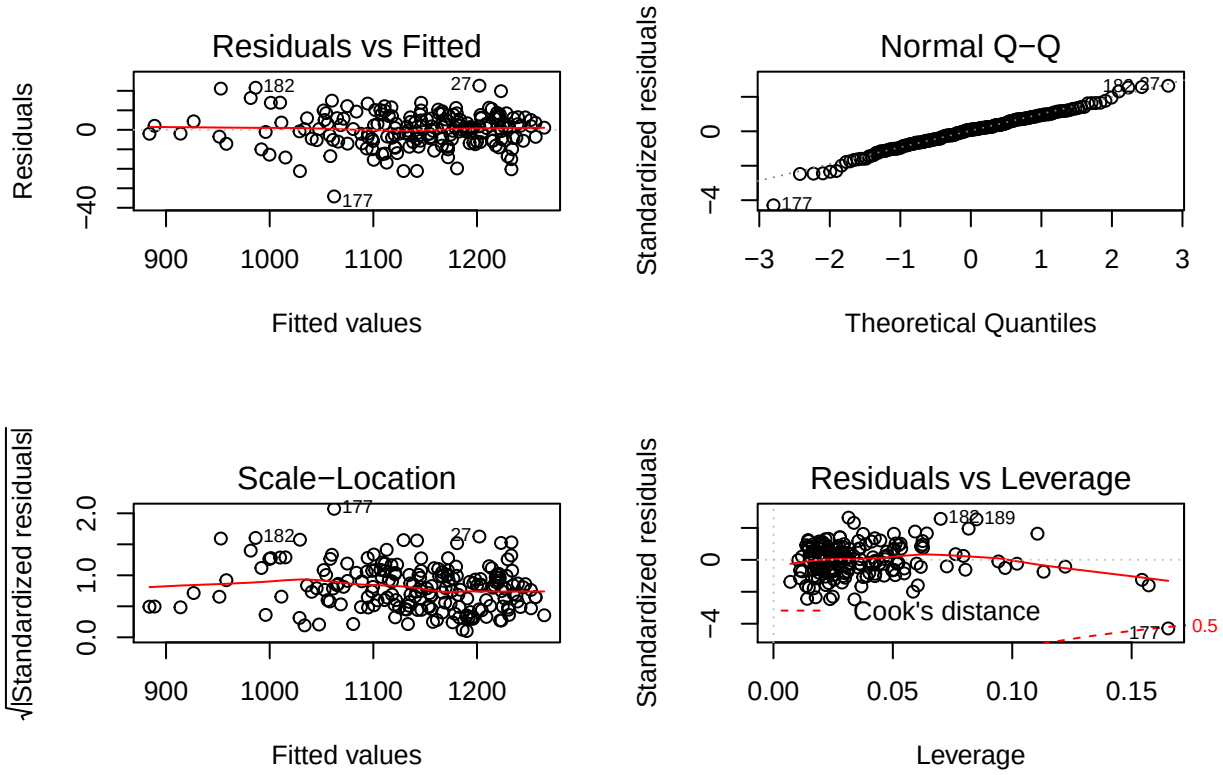
```
modelo2 <- lm(GLOBAL ~ PMA + PAN + ELE + CLE + LES + ING, datos)
summary(modelo2)
```

```
##
## Call:
## lm(formula = GLOBAL ~ PMA + PAN + ELE + CLE + LES + ING, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.191  -5.220   0.778   5.924  22.651
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.378493   9.382346   2.065   0.0403 *
## PMA          0.226623   0.010171  22.281   <2e-16 ***
## PAN          0.183735   0.009609  19.120   <2e-16 ***
## ELE          0.188384   0.010317  18.259   <2e-16 ***
## CLE          0.167405   0.008742  19.150   <2e-16 ***
## LES          0.110423   0.006589  16.759   <2e-16 ***
## ING          0.103315   0.005364  19.260   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.71 on 188 degrees of freedom
## Multiple R-squared:  0.988,  Adjusted R-squared:  0.9876
## F-statistic:  2572 on 6 and 188 DF,  p-value: < 2.2e-16
```

La ecuación viene dada por

$$GLOBAL = 19.3785 + 0.2266PMA + 0.1837PAN + 0.1884ELE + 0.1674CLE + 0.1104LES + 0.1033ING$$

y al igual que la primera ecuación, todas sus variables regresoras son positivas. Además, el modelo se ajusta un 98.79%.



La gráfica muestra una correcta distribución de residuos, por lo que los supuestos del modelo se cumplen para esta ecuación. Se puede observar que no se presentan datos atípicos bajo este modelo.