UNIVERSIDAD NACIONAL DE COLOMBIA

REGRESIÓN LINEAL MULTIPLE PARTE 1

Autor:

Edward Calderon Aristizabal

Juan Alejandro Espinosa Caceres

Carolina Rodriguez Ramirez

Kilmer Alejandro Peña Jimenez

Profesor:

Raul Alberto Perez Agamez

2022-01

La base de datos se encuentra conformada por 5 variables, de las cuales se tiene a Y como regresora y a X1,X2,X3 y X4 como variables predictoras. Donde las variables representan lo siguiente:

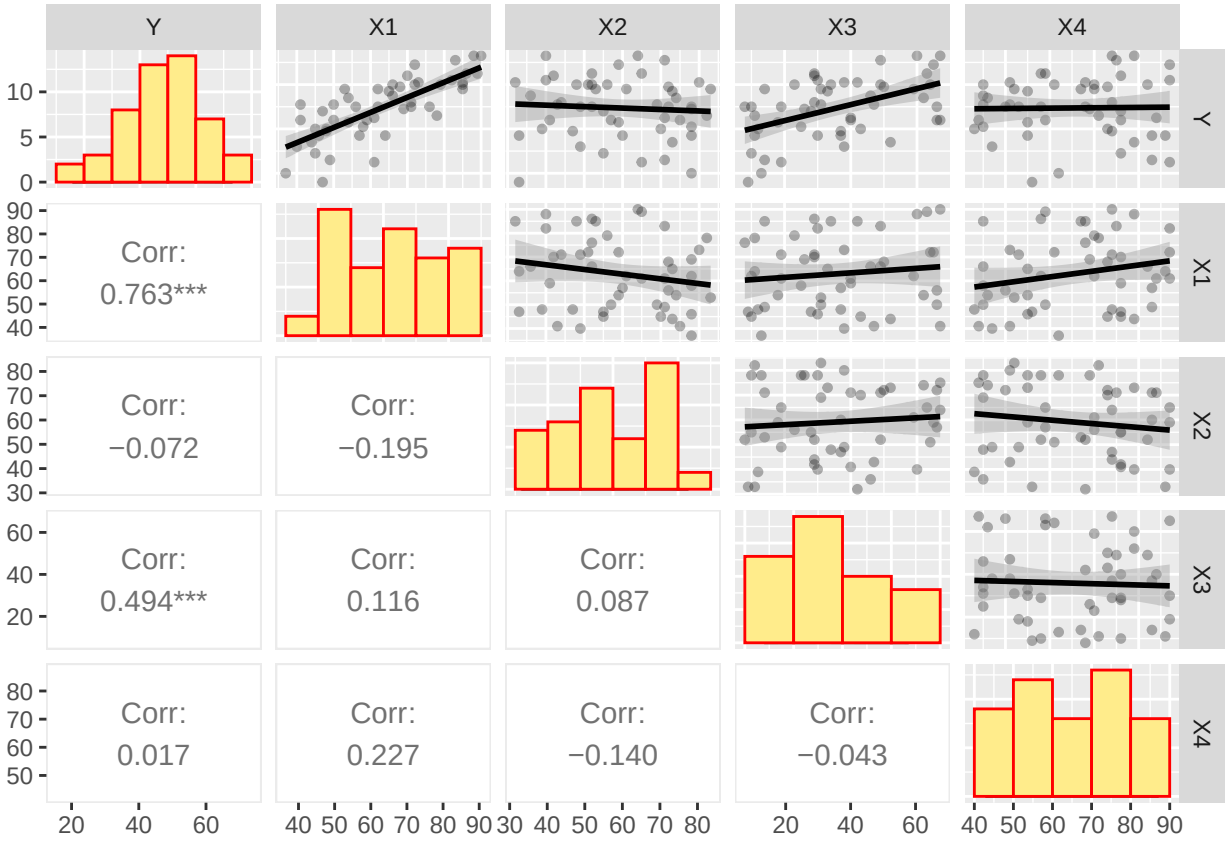
1) Y: Calificacion global de trabajo realizado por el supervisor
2) X1: Tasa de manejo de quejas de los empleados
3) X2: Tasa de no permision de privilegios especiales
4) X3: Tasa de oportunidad para aprender cosas nuevas
5) X4: Tasa de avance del supervisor a mejires puestos

Tiene la siguiente estructura.

```
##    Y X1 X2 X3 X4
## 1 50 71 49 19 53
## 2 60 72 51 64 61
## 3 58 62 61 59 70
## 4 71 83 71 49 75
## 5 44 41 75 67 44
## 6 73 88 40 60 79
```

La base de datos cuenta con una muestra de 50 observaciones. Antes de entender pretender ajustar cualquier modelo o realizar cualquier procedimiento, se debe de realizar un analisis descriptivo de las variables.

Analisis descriptivo



El anterioir grafico permite analizar tendencia lineal entre las variables respuestas y las predictoras. En este caso, interesa buscar relaciones lineales entre Y y las parejas X_j para j=1,. . . ,4

Se observa una relacion lineal positiva con X1 y X3. Las variables X2 y X4 parecen tener una relacion mas debil.

Ahora, la correlacion enter las variables ayuda a comprender si hay una relacion lineal para saber si es correcto o no ajustar un modelo.

Se tiene correlacion de Y con X1 de 0.763 y con X3 de 0494; respecto a las otras variables, presentan correlacion mas baja.

Estimacion del modelo:

Se procede a estimar un modelo de regresion lineal multiple con todas las variables predictoras. Ademas; se tiene en cuenta un analisis de significancia del modelo y de cada una de las variables.

El modelo:

Se plantea un modelo de RLM para el problema:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_4 X_{i4} + \varepsilon_i, \quad i = 1, 2, \ldots, 50$$

Que tiene como supuesto:

$$\varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \quad i = 1, 2, \ldots, 50$$

También se puede especificar el modelo en términos matriciales, así:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{con} \quad \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

Especificación del modelo de RLM, ANOVA y parámetros estimados

	Estimate	std. Error	t value	Pr(> |t|)	
(Intercept)	4.17527	7.74452	0.539	0.5925	
X1	0.67105	0.06793	9.879	7.60e − 13	***
X2	0.01889	0.06770	0.279	0.7815	
X3	0.29223	0.05369	5.443	2.08e − 06	***
X4	−0.13123	0.07325	−1.792	0.0799	

Residual standard error: 6.859 on 45 degrees of freedom Multiple R-squared: 0.7657, Adjusted R-squared: 0.7449 F-statistic: 36.77 on 4 and 45 DF, p-value: $1.201e-13$

El modelo ajustado es:

$$Y_i = 4.17527 + 0.67105X_{i1} + 0.01889X_{i2} + 0.29223X_{i3} - 0.13123X_{i4} + \varepsilon_i$$
$$i = 1, 2, \ldots, 50$$

Prueba de Significancia de la regresión

Se quiere probar:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_4 = 0, \quad \text{vs.}$$
$$H_1 : \text{Algún } \beta_j \neq 0, j = 1, \ldots, 4.$$

Tabla ANOVA

```
myAnova(modelo)
```

```
##       Sum_of_Squares DF Mean_Square F_Value     P_value
## Model        6919.40  4   1729.8507 36.7703 1.20126e-13
## Error        2117.02 45     47.0448
```

Para ello se usa la tabla de análisis de varianza. De ella se obtienen los valores del estadístico de prueba $F_0 = 36.7703$ y su correspondiente valor-P vp $= 1.20126e-13$.

Dado que el valor p es menor que el nivel de significancia alfa dado, se rechaza la hipotesis nula a favor de la alterna y se concluye que el modelo es significativo.

Esto implica que, existe almenos una variable que es significativa y que permite explicar la relacion de la variable respuesta.

Cálculo e interpretación del coeficiente de determinación

Sabemos que $R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$, de manera que se puede calcular de la tabla ANOVA.

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{6919.40}{6919.40 + 2117.02} = 0.7657236$$

Este coeficiente de determinacion permite entender la proporcion de varianza que el modelo está explicando de la variable respuesta. DE esta manera, se entiende que aproximadamente el 76.6% de la varianza total es explicada por el modelo.

Por otra parte, el R^2 ajustado es el siguiente:

$$R^2_{\text{adj}} = 1 - \frac{(n-1)\,\text{MSE}}{\text{SST}} = 1 - \frac{(50-1)\,47.0448}{6919.40 + 2117.02} = 0.2551005$$

La anterior medida ayuda a tener una mejor idea de como se comporta la variabilidad explicada por el modelo puesto que el ajustado, si penaliza a multiples variables que no sean significativas. En ese orden de ideas, el R^2_{adj} es de 0.25551 aproximadamente. La variabilidad explicada por el modelo es de aproximadamente 25%. Es relativamente baja, esto puede darse por multiples motivos.

En otras palabras y teniendo en cuenta que R^2_{adj} penaliza la varianza a medida que se agregan covariables (factor que no tiene en cuenta por si solo R^2) se prefiere usar para el caso de Regresion Lineal Multiple (RLM) el ajustado.

Significancia de los parametros:

Se estandarizan las variables para tenerlas en la misma escala y que sean comparables.

```
## Coeficientes estimados y Coeficientes estimados estandarizados

##              Estimacion    Coef.Std
## (Intercept)  4.17527476  0.00000000
## X1           0.67104638  0.75073057
## X2           0.01889298  0.02074866
## X3           0.29223295  0.39869817
## X4          -0.13123142 -0.13366818
```

Estos son los coeficientes estandarizados del modelo puesto que, no se podria dictarminar si son comparables debido a su escala.

Segun la magnitud del valor absoluto de los coeficientes estandarizados, se entiende que la variable con mayor efecto sobre el modelo es X1 seguido de x3. Este resultado es consistente con el primer analisis descriptivo.

Prueba de significancia individual de los parametros usando la prueba t

Estas pruebas establecen el siguiente juego de hipótesis:

$$\begin{array}{l} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{array} \quad \text{para} \quad j = 1, 2, \ldots, 5.$$

De la tabla de parámetros estimados, a un nivel de significancia $\alpha = 0.05$ se rechaza H_0 si $|T_0| > T_{\frac{\alpha}{2}}, n-k-1$

Donde k representa el numero de variables, n el numero de la muestra.

Para este caso con la $T_{(1-\frac{0.05}{2},45)} = 2.014103$ basta comparar con los datos suministrados en la tabla anterior en la columna de t-values:

En ese orden de ideas, al analizar la columna "t value" se comparan los que su valor absoluto sea mayor al valor calculado. Segun esto, las variables significativas son X1 y X3. Nuevamente, esto es consistente con los analissi descriptivos realizados.

Se concluye que los parámetros individuales $\hat{\beta}_1, \hat{\beta}_3$ son significativos cada uno en presencia de los demás parámetros; por otro lado, se encuentra que $\hat{\beta}_0, \hat{\beta}_2, \hat{\beta}_4$ no son individualmente significativos en presencia de los demás parámetros.

Interpretación de los parámetros estimados

En este caso, el intercepto no tiene interpretabilidad. Por tanto, $\hat{\beta}_0$ no se interpreta

$\hat{\beta}_1 = 0.67105$ indica que por cada unidad de aumento en la tasa de numero de quejas de los empleados la califiacion global del trabajo bien hecho (Y) aumenta en 0.67105 unidades, cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_2 = 0.01889$ indica que por cada unidad de aumento en la tasa de no privilegios permitidos en la califiacion global del trabajo bien hecho (Y) aumenta en 0.01889 unidades, cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_3 = 0.29223$ indica que por cada unidad de aumento en la tasa de oportunidad de aprender en la califiacion global del trabajo bien hecho (Y) aumenta en 0.29223 unidades, cuando las demás variables predictoras se mantienen fijas.

$\hat{\beta}_4$ = -0.13123 indica que por cada unidad de aumento en la taa de avance de superioir a mejores puestos la califiacion global del trabajo bien hecho (Y) disminuye en 0.13123 unidades, cuando las demás variables predictoras se mantienen fijas.

Prueba F con sumas de cuadrados extras

Para esta prueba se elegiran convenientemente los parametros no significativos del modelo, en este caso $\beta_4, \beta_2, \beta_0$. Se plantean las siguientes hipotesis:

$$H_0: \ \hat{\beta}_4 = \ \hat{\beta}_2 = \ \hat{\beta}_0 = 0 \quad \text{vs.} \quad H_1: \ \text{Algún } \beta_j \neq 0, \ j = 0, 2, 4$$

Reescribiendo las hipotesis matricialmente:

$$\begin{cases} \mathrm{H}_0 : \mathbf{L}\underline{\beta} = \underline{\mathbf{0}} \\ \mathrm{H}_1 : \mathbf{L}\underline{\beta} \neq \underline{\mathbf{0}} \end{cases}$$

Donde la matriz L está dada por:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Esta prueba se desarrolla usando sumas de cuadrados extra y se requiere la tabla de todas las regresiones posibles como se presenta a continuación.

Modelo FULL

$$\hat{Y}_i = \hat{\beta}_0 + \beta_1 \hat{X}_{i1} + \beta_2 \hat{X}_{i2} + \cdots + \beta_4 \hat{X}_{i4} + \varepsilon_i, \quad i = 1, 2, \ldots, 50$$
$$\varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \quad i = 1, 2, \ldots, 80$$

Modelo reducido

$$\hat{Y}_i = \beta_1 \hat{x}_{i1} + \beta_3 \hat{x}_{i3} + + E_i$$
$$\varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \quad i = 1, 2, \ldots, 50$$

Estadistico de prueba F0

$$F_0 = \frac{(SSE(MR) - SSE(MF))/2}{MSE(MF)}$$
$$= \frac{(SSE(X_1, X_3) - SSE(X_1, X_2, X_3, X_4))/(n-3) - (n-5)}{MSE(X_1, X_2, X_3, X_4))}$$
$$= \frac{SSR(X_1, X_2, X_3, X_4 | X_0, , X_2, X_4)/2}{MSE(X_1, X_2, X_3, X_4)} \sim f_{2,45} \, bajo H_0$$

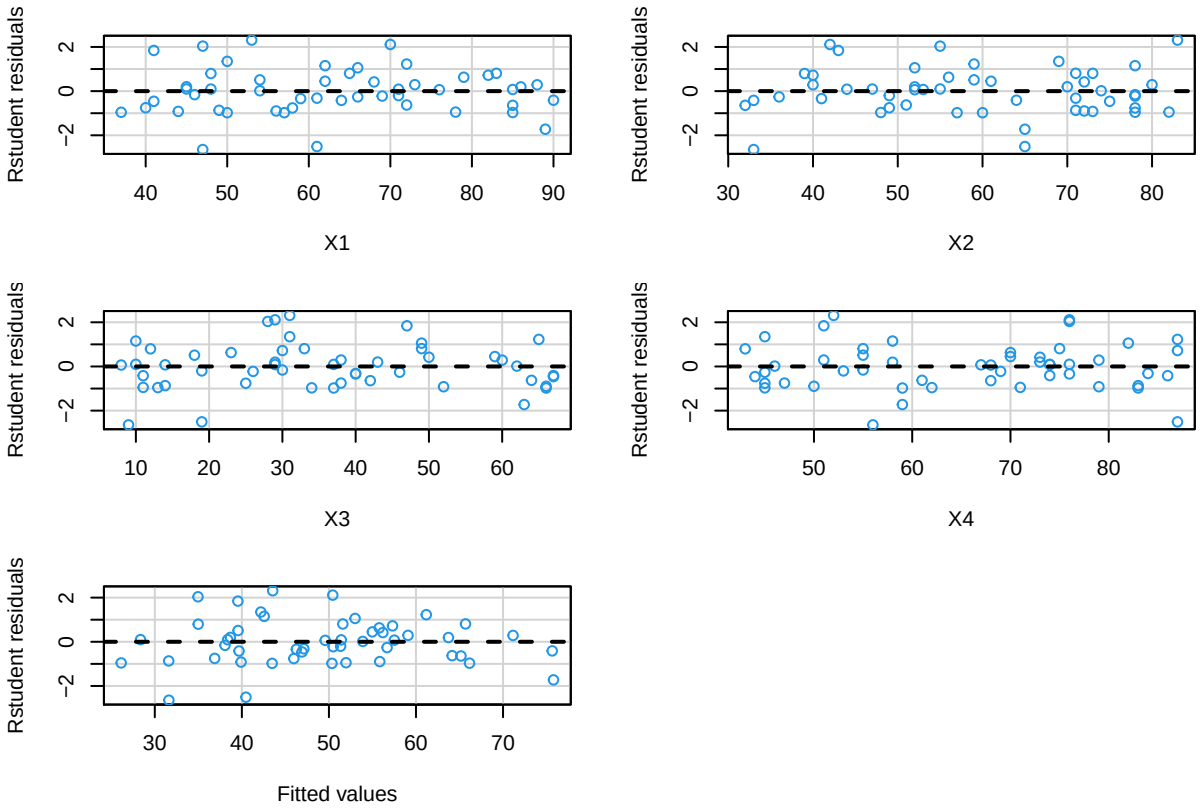
```
## Linear hypothesis test
##
## Hypothesis:
## X4 = 0
## X2 = 0
## (Intercept) = 0
##
## Model 1: restricted model
```

```
## Model 2: Y ~ X1 + X2 + X3 + X4
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     48 2283.3
## 2     45 2117.0  3     166.3 1.1783 0.3286
```

El estadistico F tiene valor de 1.1783 y su valor p asociado a dicha prueba es de 0.3286 por tanto no se rechaza la hipotesis nula y se concluye que el parametro β_j no son significativos para el modelo para j =0,2,4

Se concluye que el conjunto de predictoras simultaneamente no son significativas, en presencia de los demás parámetros lo que implica que las variable X2, X4 y el intercepto no son significativas para explicar la variable respuesta Y. Notese que este resultado coincide con la prueba de significancia individual de los parametros. (Ver summary del modelo en las variables)

Graficos de residuales estudentizados vs valores ajustados

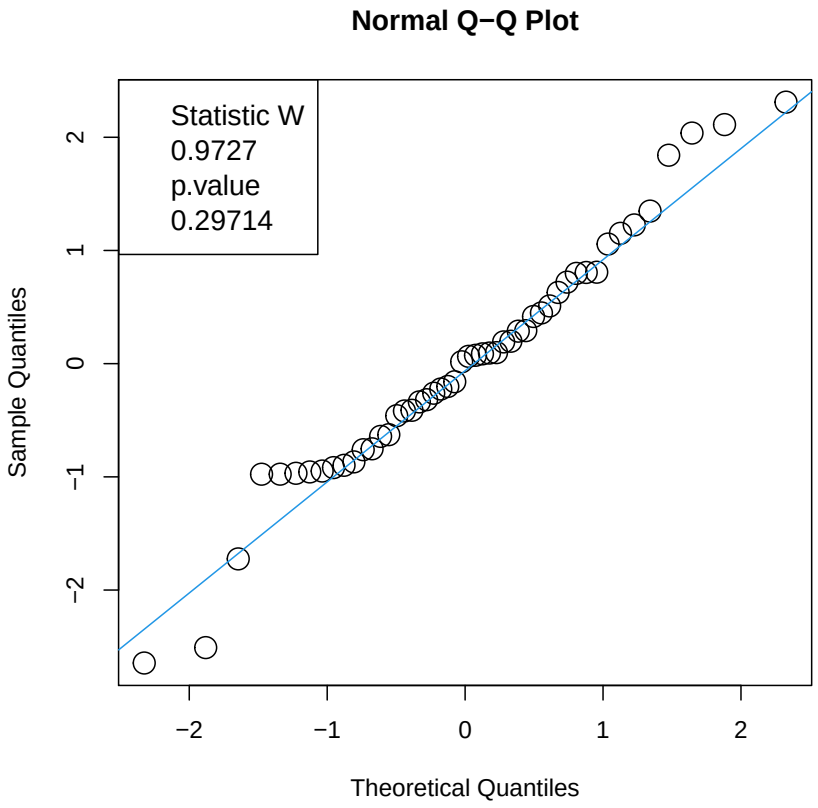


1) Se observa que los datos son aleatorios alrededor de 0, no se identifican patrones dentro de las graficas de estudentizados vs valores ajustados lo que indica que no hay problema de varianza constante.
2) No se evidencian puntos por encima de +- 3 desviaciones. En un principio, parece que no hay evidencia de datos atipicos.
3) En apariencia en el grafico de valores ajustados pareciese haber un punto mas alejado de la nube de puntos. Quizas pueda tratarse de un punto de balanceo.
4) No se observan patrones y se ve aleatoreamente distribuidos. Se asume que no hay carencia de ajuste (Lack of fit) en el modelo.

Gráfico de normalidad

$$H_0 : \varepsilon_i \sim N(0, \sigma^2)$$

$$H_1 : \varepsilon_i \nsim N(0, \sigma^2)$$



Dado el grafico qqnorm y la prueba de Shaphiro Wilk con el juego de hipotesis ya nombrado, se observa que el valor p es de 0.34 (grande) por tanto no se rechaza la hipotesis nula y se concluye que los residuales distribuyen normal estandar

```
##    Y      ri      ei  se.yhat residuals hii.value
## 1 50 -0.2016 -0.2038 1.906101   -1.3425    0.0772
## 2 60 -0.6274 -0.6317 1.959819   -4.1519    0.0816
## 3 58  0.4482  0.4522 1.638522    3.0118    0.0571
## 4 71  0.8080  0.8111 2.048109    5.3094    0.0892
## 5 44 -0.4598 -0.4639 2.771796   -2.9106    0.1633
## 6 73  0.2869  0.2899 2.512772    1.8502    0.1342
```

Se asume que la observación i es atipica si un ei grande ($|ei| > 3$) y Se considera potencialmente atipica con ri grande ($|ri| > 3$).

No se encuentran observaciones atipicas al analizar los datos. Algo consistente con el analisis descriptivo.

NOTA: Para esta seccion, se muestra solo un fragmento de la tabla de datos. De manera unicamente ilustrativa para ejemplificar que se está observando y en que criterios se puede realizar los analisis de interes. Para verificar si hay o no observaciones que sean atipicas, se usa la funcion filter del paquete dplyr

Observaciones de balanceo

Se asume que la observación i es un punto de balanceo si hii > 2p/n. En esta práctica tenemos como criterio que: hii > 2(k+1)/n = 2(5/50) = 0.2 De acuerdo a la columna hii.value la observación

```
## [1] Y         ri        ei        se.yhat   residuals hii.value
## <0 rows> (or 0-length row.names)
```

Segun esto, las observaciones 32 y 47 son observaciones de balanceo. Ahora, el valor por el cual superan el criterio de la matriz hat no es muy elevado. Es decir, no sobrepasa muy por encima los valores exigidos entonces incluso aqui, se podria analizar dichos puntos para tenerlos presentes y mirar si se descartan o no.

Observaciones influenciales

Para identificar estos valores utilizaremos 3 criterios, que son:

- Se dice que la observacion sera influencial si Di > 1.
- una observacion sera influencial si |DFFITS| > 2(p/n)ˆ0.5
- observaciones con un covratio tal que |COVRATIO-1| > 3(p/n) son cadidatas a ser influenciales donde p es el numero de variables

```
head(prb)
```

```
##   dfb.1_ dfb.X1 dfb.X2 dfb.X3 dfb.X4 dffit cov.r cook.d   hat
## 1  FALSE  FALSE  FALSE  FALSE  FALSE FALSE FALSE  FALSE FALSE
## 2  FALSE  FALSE  FALSE  FALSE  FALSE FALSE FALSE  FALSE FALSE
## 3  FALSE  FALSE  FALSE  FALSE  FALSE FALSE FALSE  FALSE FALSE
## 4  FALSE  FALSE  FALSE  FALSE  FALSE FALSE FALSE  FALSE FALSE
## 5  FALSE  FALSE  FALSE  FALSE  FALSE FALSE FALSE  FALSE FALSE
## 6  FALSE  FALSE  FALSE  FALSE  FALSE FALSE FALSE  FALSE FALSE
```

NOTA: Con una funcion de usuario, se genera un data frame donde para cada criterio anterior mencionado, evalua cada observacion. Si sobre pasa la cota (es decir, es influencial, de balanceo o a tipico) lo marca como "TRUE". Se ilutra, como se puede observar de manera general dicho data frame.

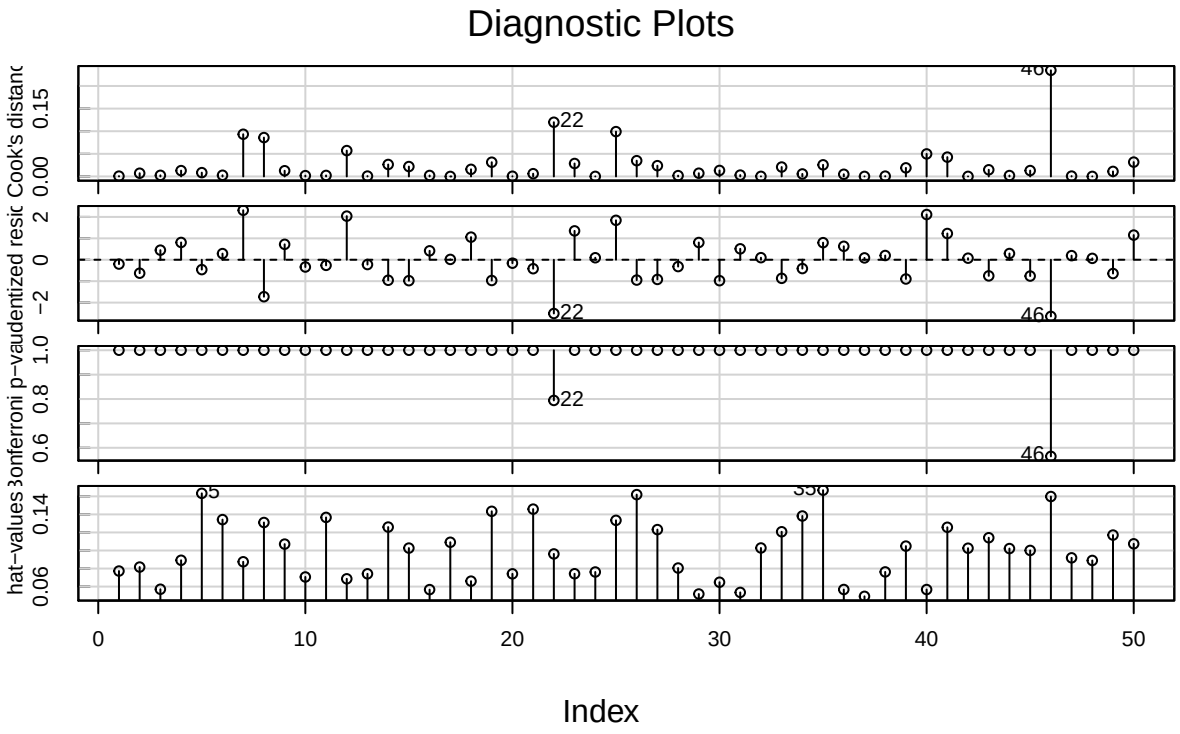
```
filter(prb, prb$cov.r=="TRUE")
```

```
##    dfb.1_ dfb.X1 dfb.X2 dfb.X3 dfb.X4 dffit cov.r cook.d   hat
## 22  FALSE  FALSE  FALSE  FALSE  FALSE FALSE  TRUE  FALSE FALSE
## 46   TRUE  FALSE  FALSE  FALSE  FALSE  TRUE  TRUE  FALSE FALSE
```

Por criterio de cov.r se tiene que estas observaciones son observaciones influenciables. Por tanto, se recomienda analizar la observacion 22 y 46 para verificarlar, mirar si se pueden remover por algun error a la hora de tomar los datos.

Tenga presente que cada uno de los criterios utilizados tiene metodologias diferentes. Idealmente se espera consistencia entre todos los criteriores entre si. Si un metodo las categoriza como observaciones influenciales y otro no, depende del investigador mirar detenidamente las observaciones y tomar decisiones.

Gráficas de chequeos y diagnosticos

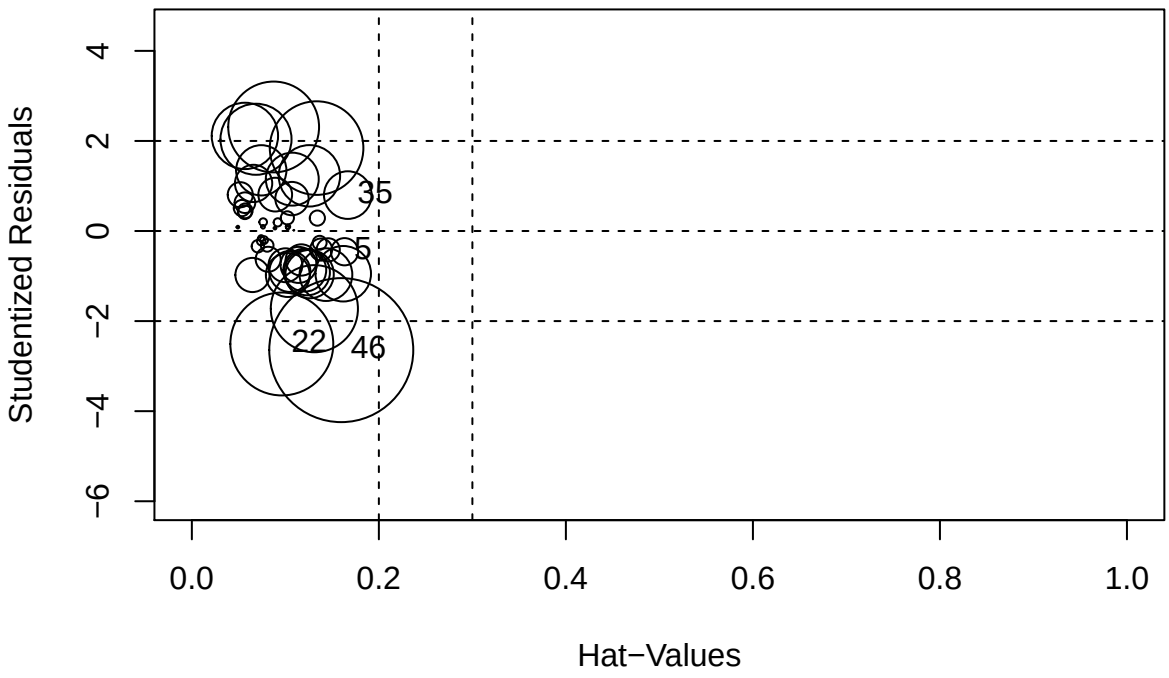


1) El primer grafico de la distancia de Cook, no hay ningun valor superior a 1. Entonces no se identifica puntos de balanceo.

2) La segunda grafica señala a las observaciones 22 y 46 por tener valores altos. Pero no supera el criterio de +-3 respecto a la linea del grafico, se asumen no hay observaciones atipicas.

3) El tercer grafico tambien ayuda a identificar si existen valores atipicos. Nuevamente, no existen entonces valores atipicos aunque si señala a la observacion 46 con un valor particularmente alto.

4) En la gráfica de hat.values no se ven puntos de balanceo (no superan el valor del criterio)




```
##       StudRes        Hat       CookD
## 5  -0.4598345 0.16330915 0.008401492
## 22 -2.5085060 0.09621092 0.119874272
## 35  0.7975782 0.16679000 0.025675449
## 46 -2.6445664 0.15977957 0.234726801
```

Dado este grafico, no se encuentran observaciones atipicas ni ninguna otra observacion que presente problemas.

Multicolinealidad para modelo full

Se debe realizar un analisis de multicolinealidad entre las variables.

Matriz de correlación de variables predictorias

```
##              Y         X1          X2          X3          X4
## Y   1.00000000  0.7627430 -0.07181392  0.49375049  0.01689227
## X1  0.76274299  1.0000000 -0.19461349  0.11649229  0.22739047
## X2 -0.07181392 -0.1946135  1.00000000  0.08745703 -0.13967990
## X3  0.49375049  0.1164923  0.08745703  1.00000000 -0.04326672
## X4  0.01689227  0.2273905 -0.13967990 -0.04326672  1.00000000
```

Al observar la matriz de correlacion de las variables, no se percibe una correlacion fuerte entre las predictoras. Entonces se podria sospechar que no hay problemas de multicolinealidad.

VIF'S Para analizar problemas de multicolinealidad con los VIF's, se analiza la siguiente tabla:

```
##       X1       X2       X3       X4
## 1.109243 1.061922 1.030782 1.069205
```

Para VIF's con valores >10 indica que hay problemas de multicolinealidad. Según este criterio no se detecta problemas de multicolinealidad.

Proporciones de varianza

Como en los datos β_0 no tiene interpretabilidad, se trabaja con los datos centrados. Para ello:

```
##   Val.propio cond.index       Pi.X1      Pi.X2        Pi.X3       Pi.X4
## 1  1.3767104   1.000000 0.253564448 0.19453685 5.535736e-06 0.223077961
## 2  1.0776766   1.130257 0.092116688 0.08307667 7.100570e-01 0.005244458
## 3  0.8540064   1.269670 0.003867106 0.52174258 2.706285e-02 0.546870004
## 4  0.6916066   1.410885 0.650451757 0.20064389 2.628746e-01 0.224807577
```

Segun el indice de condicion, existe problemas graves de multicolinealidad si dicho indice es mayor de 31. En ninguna de estas variables hay problemas graves de multicolinealidad ni siquiera problemas ligeros. Esto era de esperarse puesto que en la matriz de correlacion se encontró que la covarianza entre las predictoras era muy pequeña.

En terminos generales, no se detecta ningun probelma de multicolinealidad entre las variables predictoras.

Selección de variables

Bajo el modelo ajustado sin las observaciones, se procede a realizar el analisis correspondiente al mejor modelo a ajustar acorde a los criterios solicitados.

comparando todos los posibles modelos y teniendo siempre presente el principio de parsimonia:

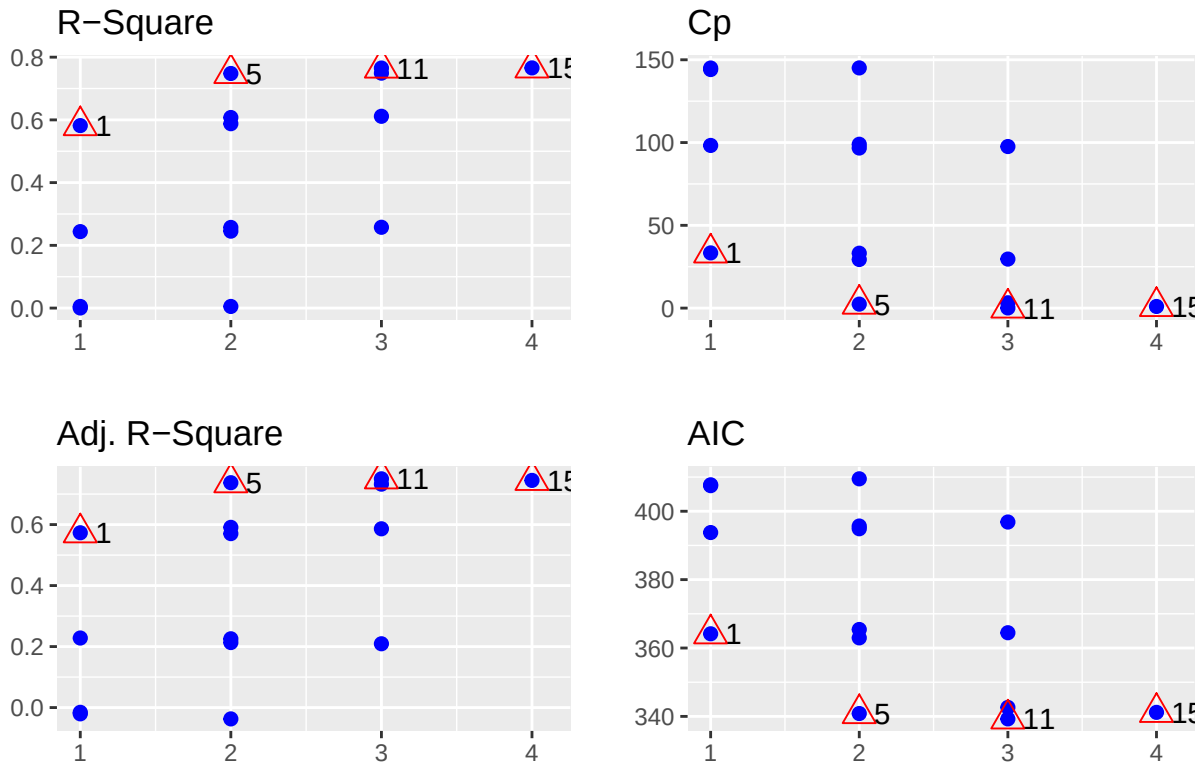
- Por criterio de R^2_{adj} y R^2, se selecciona el modelo con el valor mas alto
- Por criterio de MSE se selecciona el modelo con menor valor de este estadistico, aunque es equivalente con el anterior (A menor MSE mayor R^2_{adj} y R^2 asi que se esperan resultados similares)
- Por criterior de C_p para el valor mas pequeño de dicho estadistico; estadistico que está dado por:

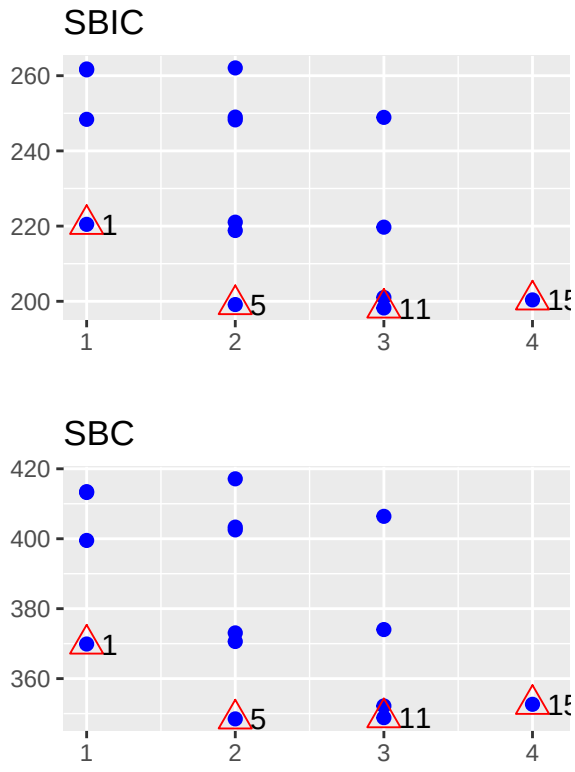
$$C_p = \frac{SSE_p}{MSE(X_1, X_2, ..., X_k)} - (n - 2p)$$

Donde SSE_p es el SSE del moderlo de regresion con $p-1 \leq k$ variables predictoras y el MSE del denominador es el SSE con todas las k predictoras.

```
##   Index N Predictors     R-Square Adj. R-Square Mallow's Cp
## 1     1 1         X1 0.5817768739    0.57306389   34.332738
## 3     2 1         X3 0.2437895468    0.22803516   99.253699
## 2     3 1         X2 0.0051572388   -0.01556865  145.090443
## 4     4 1         X4 0.0002853486   -0.02054204  146.026241
## 6     5 2      X1 X3 0.7479736691    0.73724914    4.409483
## 7     6 2      X1 X4 0.6076205025    0.59092350   31.368666
```

Con esto en mente, se seleccionaria acorde a los criterios mencionados; sin embargo, revisar uno a uno los criterios de la tabla puede ser un trabajo poco efectivo y engorroso. Para ello, se apoya en la siguiente grafica:


R−Square
0.8
0.6
0.4
0.2
0.0
1
2
3
4
1
5
11
15
Cp
150
100
50
0
1
2
3
4
1
5
11
15
Adj. R−Square
0.6
0.4
0.2
0.0
1
2
3
4
1
5
11
15
AIC
400
380
360
340
1
2
3
4
1
5
11
15



Al observar todos los criterios, se concluye queel mejor modelo a seleccionar es el 5. Puesto que si bien el 11 tiene valores "mejores" segun el resto de criterios, dicha diferencia no es significativa y por tanto, se decide utilizar el modelo 5.

```
##   Index N Predictors  R-Square Adj. R-Square Mallow's Cp
## 6     5 2      X1 X3 0.7479737     0.7372491    4.409483
```

Este modelo tiene a las variables X1 y X3.

Con las funciones suministradas por el docente

```
##    k  R_sq adj_R_sq     SSE      Cp Variables_in_model
## 1  1 0.582    0.573 3779.24  34.333                X1
## 2  1 0.244    0.228 6833.44  99.254                X3
## 3  1 0.005   -0.016 8989.82 145.090                X2
## 4  1 0.000   -0.021 9033.84 146.026                X4
## 5  2 0.748    0.737 2277.42   4.409             X1 X3
## 6  2 0.608    0.591 3545.71  31.369             X1 X4
## 7  2 0.588    0.570 3724.09  35.161             X1 X2
## 8  2 0.257    0.226 6713.02  98.694             X2 X3
## 9  2 0.245    0.213 6820.19 100.972             X3 X4
## 10 2 0.005   -0.037 8989.38 147.081             X2 X4
## 11 3 0.765    0.750 2120.68   3.078          X1 X3 X4
## 12 3 0.749    0.733 2268.02   6.210          X1 X2 X3
## 13 3 0.612    0.586 3510.55  32.621          X1 X2 X4
## 14 3 0.258    0.209 6708.34 100.595          X2 X3 X4
## 15 4 0.766    0.745 2117.02   5.000       X1 X2 X3 X4
```

La anterior tabla muestra todas las posibles regresiones a realizar con una funcion de usuario. Cabe destacar

que para resolver el ejercicio se han usado funciones personalizadas. Sin embargo, esta funcion posee informacion mas concisa y, solo por si acaso, se presenta la tabla de posibles regresiones puesto que un inciso del taller asi lo exige.

Ajuste del nuevo modelo

```
##
## Call:
## lm(formula = Y ~ X1 + X3, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1350  -4.1512   0.0051   4.3998  16.3670
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.56295    4.47972  -0.349    0.729
## X1           0.63904    0.06590   9.697 8.57e-13 ***
## X3           0.30086    0.05404   5.567 1.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.961 on 47 degrees of freedom
## Multiple R-squared:  0.748,  Adjusted R-squared:  0.7372
## F-statistic: 69.74 on 2 and 47 DF,  p-value: 8.59e-15
```

De este modelo, se muestran los parametros estimados y demas valores. Claramente todas las variables son significativas y es el que, siguiendo un principio de parsimonia, tiene menos variables y acoge una mayor proporcion de varianza explicada por el modelo siguiendo el principio de parsimonia.

Por esto, se recomienda finalmente este modelo para trabajar.