Tarea 5

Bernardo Mondragón 143743, Karen Delgado 142252, Juan Casas 141913, David Almog 136731

May 20, 2018

i) Obtenga todas las regresiones posibles y escoja una de acuerdo al criterio Cp de Mallow's

Corriendo todas las regresiones posibles se obtiene el siguiente output que indica cual es el mejor modelo dependiendo el numero de regresores:

```
## $which
##      POP    UR    IN    PR    CR   PI    PL    BL    SP    AI    MH    RP
## 1  FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 2  FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
## 3  FALSE  TRUE FALSE FALSE FALSE TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 4   TRUE  TRUE FALSE FALSE FALSE TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 5   TRUE  TRUE FALSE FALSE  TRUE TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 6  FALSE  TRUE  TRUE FALSE FALSE TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE
## 7  FALSE  TRUE  TRUE FALSE  TRUE TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE
## 8  FALSE  TRUE  TRUE FALSE  TRUE TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## 9  FALSE  TRUE  TRUE FALSE  TRUE TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
## 10 FALSE  TRUE  TRUE FALSE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 11 FALSE  TRUE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 12  TRUE  TRUE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##
## $label
##  [1] "(Intercept)" "POP"         "UR"          "IN"          "PR"
##  [6] "CR"          "PI"          "PL"          "BL"          "SP"
## [11] "AI"          "MH"          "RP"
##
## $size
##  [1]  2  3  4  5  6  7  8  9 10 11 12 13
##
## $Cp
##  [1] 45.4394441 24.4798358  3.1638170 -0.1981626  0.9392919  2.4819874
##  [7]  3.8050927  5.3910407  7.0725722  9.0077076 11.0040877 13.0000000
```

Segun el output obtenido, el mejor modelo que contiene 1 regresor esta dado por el modelo que contiene unicamente a la variable PI, el mejor modelo que contiene 2 regresores el modelo que contiene unicamente a las variables PI y SP y asi sucesivamente.

De acuerdo al criterio C_p de Mallows, el mejor modelo, de entre todos los posibles modelo es el modelo que contiene 4 variables e incluye unicamente a los regresores POP, UR, PI y PL.

ii) Diga si la que escogió sería tambien un buen candidato en terminos de R^2

Haciendo lo mismo que en la pregunta anterior, pero esta vez escojiendo como criterio el valor de R^2 se ontiene el siguinete output:

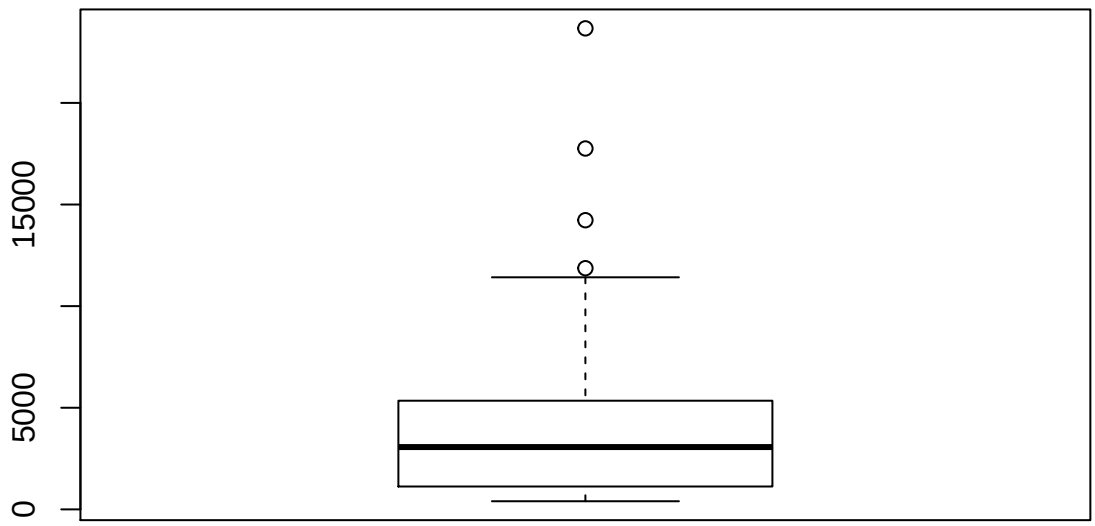
```
## $which
```

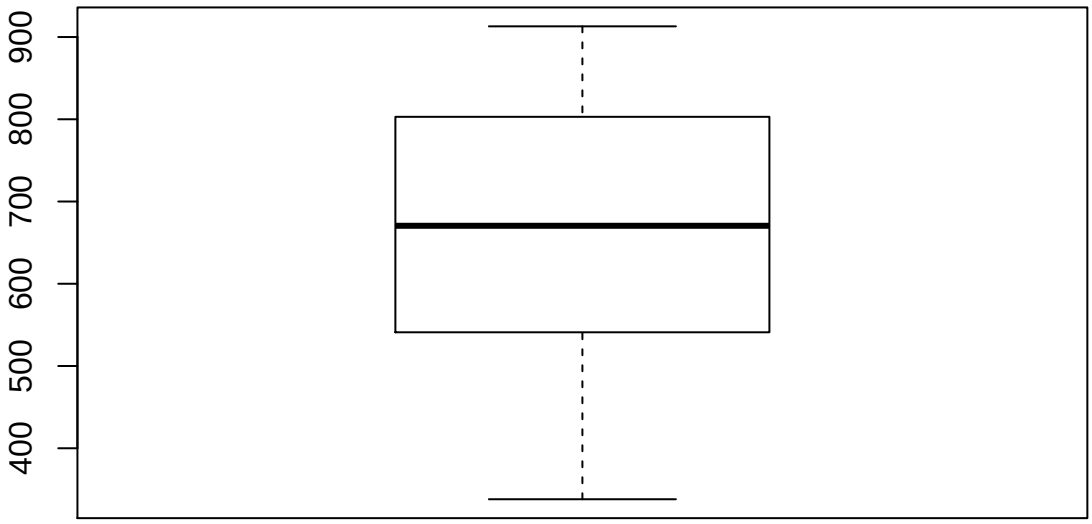
```
##      POP    UR    IN    PR    CR   PI    PL    BL    SP    AI    MH    RP
## 1  FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 2  FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
## 3  FALSE  TRUE FALSE FALSE FALSE TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 4   TRUE  TRUE FALSE FALSE FALSE TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 5   TRUE  TRUE FALSE FALSE  TRUE TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 6  FALSE  TRUE  TRUE FALSE FALSE TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE
## 7  FALSE  TRUE  TRUE FALSE  TRUE TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE
## 8  FALSE  TRUE  TRUE FALSE  TRUE TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## 9  FALSE  TRUE  TRUE FALSE  TRUE TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
## 10 FALSE  TRUE  TRUE FALSE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 11 FALSE  TRUE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 12  TRUE  TRUE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##
## $label
##  [1] "(Intercept)" "POP"         "UR"          "IN"          "PR"
##  [6] "CR"          "PI"          "PL"          "BL"          "SP"
## [11] "AI"          "MH"          "RP"
##
## $size
##  [1]  2  3  4  5  6  7  8  9 10 11 12 13
##
## $r2
##  [1] 0.5583456 0.6692410 0.7818578 0.8077563 0.8119224 0.8141312 0.8174006
##  [8] 0.8194005 0.8209387 0.8212520 0.8212695 0.8212892
```

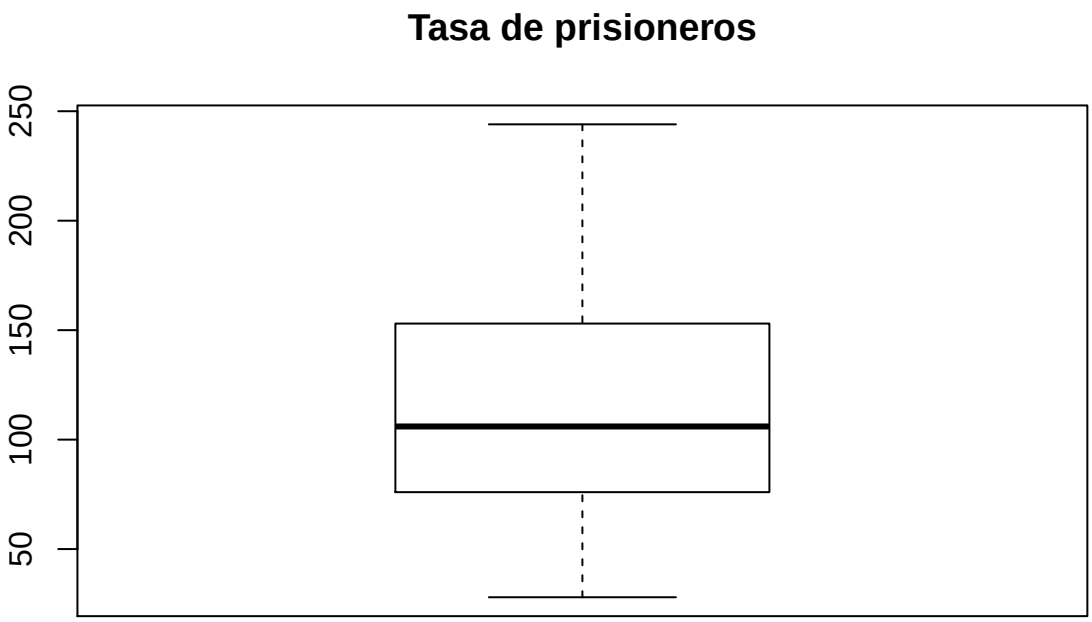
Esto indica que el mejor modelo es el que tiene el mayor numero de regresores. En este caso, el mejor modelo no coincide con el mejor modelo si se utilizara el criterio C_p de Mallows.

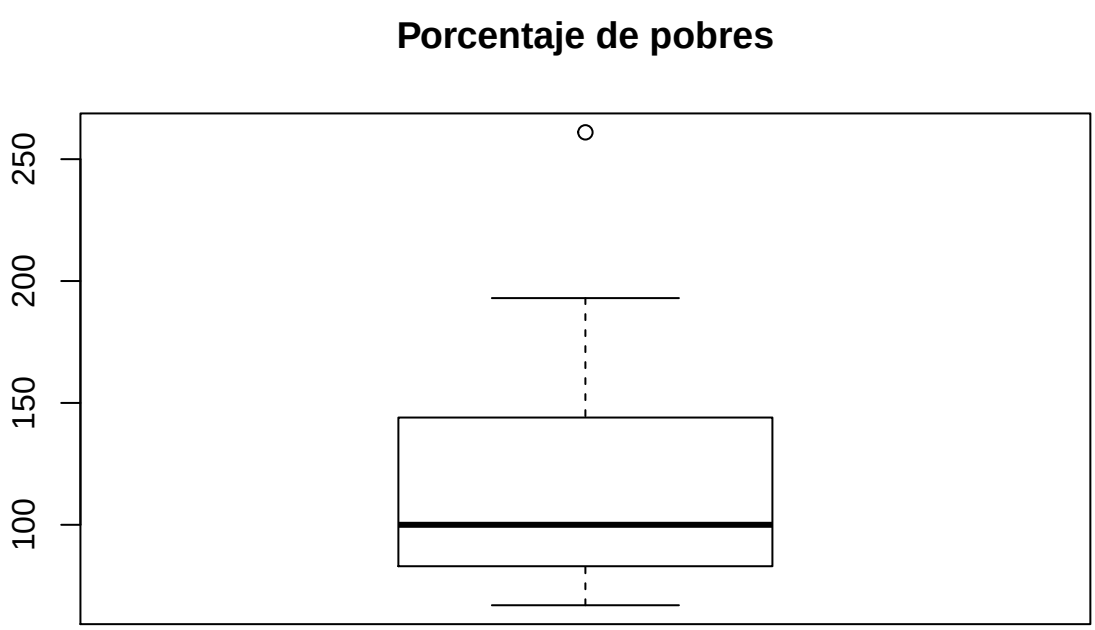
iii) Para el modelo escogido indique si existen:

Observaciones atípicas en el espacio de las x's:


Poblacion total
0
5000
15000


Poblacion urbana
400
500
600
700
800
900


Tasa de prisioneros
250
200
150
100
50



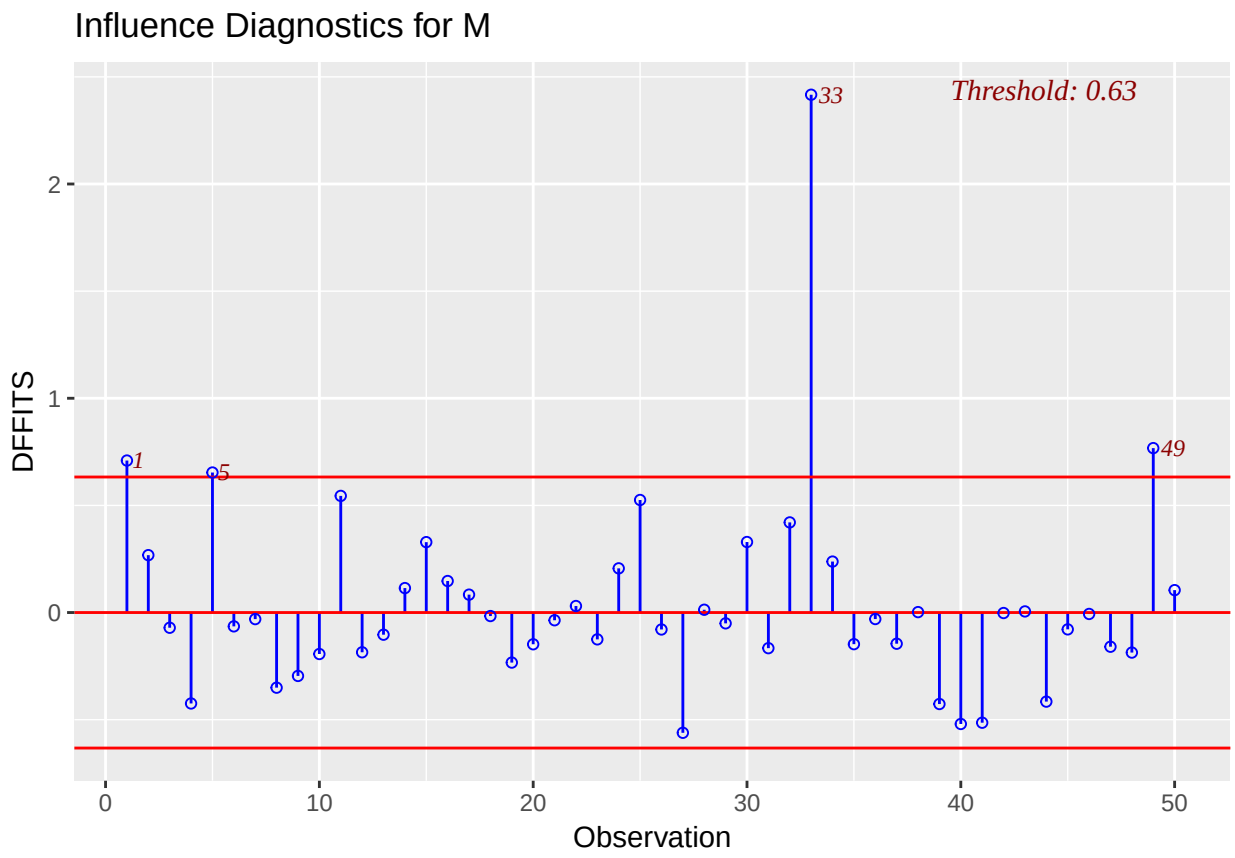
Observaciones atípicas en el espacio de las y's (observaciones a las que no se les ajusta bien un modelo que se ajustó a las otras, esto es, RSTUDENT alto):

Se obtienen los siguientes valores:

```
##    rstudent unadjusted p-value Bonferonni p
## 33 4.733941         2.3094e-05    0.0011547
```

Observaciones influyentes (DFFITS):

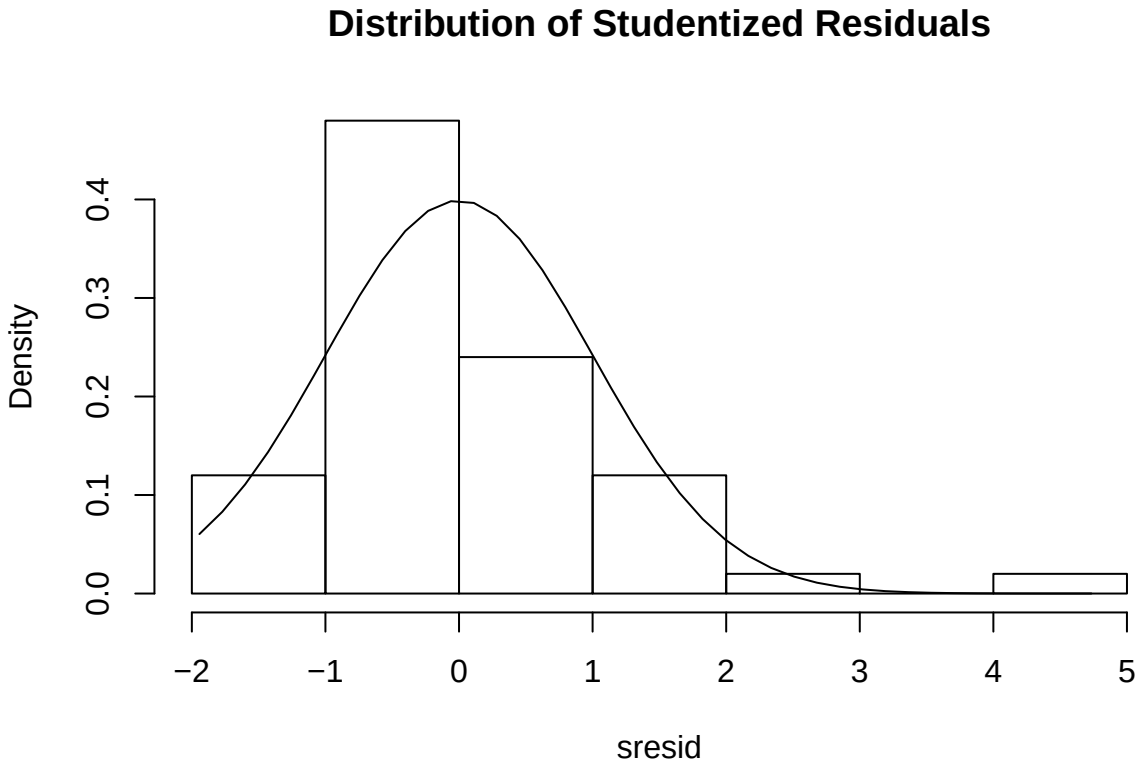
Para el modelo seleccionado a travez del criterio de Mallows se obtiene el siguinete grafico:



iv) Intente explicar porqué son atípicos los estados que encontró en el punto anterior, si es que encontró alguno. Por ejemplo: gran población, etc.

Creemos que la observacion 33 correspondiente al estado de Nevada es un valor atipico debido a que en este estado hay una alta tasa de prisioneros.

V) Obtenga una gráfica de probabilidad normal de los RSTUDENT y coméntela



Es muy parecida a la distribucion normal estandar.

vi) Deje fuera un estado tomado al azar, corra su modelo y obtenga un intervalo de predicción (95%) para su tasa de asesinatos. ¿Es satisfactorio este intervalo? ¿Cuál hubiera sido su predictor si no hubiera usado los regresores y sólo las tasas de los otros edos.? ¿se ganó algo con la regresión?

vii) Con el objeto de contrastar la hipotesis de que ser un estado fronterizo con Mexico influye en la tasa de asesinatos se introduce en el modelo una variable indicadora que vale 1 si el estado es fronterizo ¿Es su coeficiente significativo?. Compare este resultado con la simple prueba de dos muestras aplicada a los estados fronterizos versus los que no lo son. Comente los resultados