Trabajo Final Parte 2

Beloslava Dimitrova y Mariam Essamari

29.12.2022

Variable -> Descripción

M -> percentage of males aged 14–24 in total state population

So -> indicator variable for a southern state

Ed -> mean years of schooling of the population aged 25 years or over

Po1 -> per capita expenditure on police protection in 1960

Po2 -> per capita expenditure on police protection in 1959

LF -> labour force participation rate of civilian urban males in the age-group 14-24

M.F -> number of males per 100 females

Pop -> state population in 1960 in hundred thousands

NW -> percentage of nonwhites in the population

U1 -> unemployment rate of urban males 14–24

U2 -> unemployment rate of urban males 35–39

Wealth -> wealth: median value of transferable assets or family income

Ineq ->income inequality: percentage of families earning below half the median income

Prob -> probability of imprisonment: ratio of number of commitments to number of offenses

Time -> average time in months served by offenders in state prisons before their first release

Crime -> crime rate: number of offenses per 100,000 population in 1960

Observations are the 47 US states.

```
uscrime<-read.table("C:\\Users\\User\\Documents\\PREDICCION\\uscrime.txt",
header=TRUE)
head(uscrime)
```

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq
Prob
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1
0.084602
```

```
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4
0.029599
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0
0.083401
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7
0.015801
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4
0.041399
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6
0.034201
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

Para nuestro modelo hemos elegido como variable respuesta el nivel de crimen - número de delitos por 100,000 personas en 1960. Aunque no se especifica nada, estaría bien ver cuáles de las variables son más relacionadas con la respuesta.

```
cor(uscrime)
```

```
##                  M          So          Ed         Po1         Po2
LF
## M       1.00000000  0.58435534 -0.53023964 -0.50573690 -0.51317336 -
0.1609488
## So      0.58435534  1.00000000 -0.70274132 -0.37263633 -0.37616753 -
0.5054695
## Ed     -0.53023964 -0.70274132  1.00000000  0.48295213  0.49940958
0.5611780
## Po1    -0.50573690 -0.37263633  0.48295213  1.00000000  0.99358648
0.1214932
## Po2    -0.51317336 -0.37616753  0.49940958  0.99358648  1.00000000
0.1063496
## LF     -0.16094882 -0.50546948  0.56117795  0.12149320  0.10634960
1.0000000
## M.F    -0.02867993 -0.31473291  0.43691492  0.03376027  0.02284250
0.5135588
## Pop    -0.28063762 -0.04991832 -0.01722740  0.52628358  0.51378940 -
0.1236722
## NW      0.59319826  0.76710262 -0.66488190 -0.21370878 -0.21876821 -
0.3412144
## U1     -0.22438060 -0.17241931  0.01810345 -0.04369761 -0.05171199 -
0.2293997
## U2     -0.24484339  0.07169289 -0.21568155  0.18509304  0.16922422 -
0.4207625
## Wealth -0.67005506 -0.63694543  0.73599704  0.78722528  0.79426205
0.2946323
```

```
## Ineq    0.63921138  0.73718106 -0.76865789 -0.63050025 -0.64815183 -
0.2698865
## Prob    0.36111641  0.53086199 -0.38992286 -0.47324704 -0.47302729 -
0.2500861
## Time    0.11451072  0.06681283 -0.25397355  0.10335774  0.07562665 -
0.1236404
## Crime  -0.08947240 -0.09063696  0.32283487  0.68760446  0.66671414
0.1888663
##                M.F          Pop           NW           U1          U2
## M       -0.02867993 -0.28063762  0.59319826 -0.224380599 -0.24484339
## So      -0.31473291 -0.04991832  0.76710262 -0.172419305  0.07169289
## Ed       0.43691492 -0.01722740 -0.66488190  0.018103454 -0.21568155
## Po1      0.03376027  0.52628358 -0.21370878 -0.043697608  0.18509304
## Po2      0.02284250  0.51378940 -0.21876821 -0.051711989  0.16922422
## LF       0.51355879 -0.12367222 -0.34121444 -0.229399684 -0.42076249
## M.F      1.00000000 -0.41062750 -0.32730454  0.351891900 -0.01869169
## Pop     -0.41062750  1.00000000  0.09515301 -0.038119948  0.27042159
## NW      -0.32730454  0.09515301  1.00000000 -0.156450020  0.08090829
## U1       0.35189190 -0.03811995 -0.15645002  1.000000000  0.74592482
## U2      -0.01869169  0.27042159  0.08090829  0.745924815  1.00000000
## Wealth  0.17960864  0.30826271 -0.59010707  0.044857202  0.09207166
## Ineq    -0.16708869 -0.12629357  0.67731286 -0.063832178  0.01567818
## Prob    -0.05085826 -0.34728906  0.42805915 -0.007469032 -0.06159247
## Time    -0.42769738  0.46421046  0.23039841 -0.169852838  0.10135833
## Crime    0.21391426  0.33747406  0.03259884 -0.050477918  0.17732065
##               Wealth        Ineq         Prob          Time       Crime
## M       -0.6700550558  0.63921138  0.361116408  0.1145107190 -0.08947240
## So      -0.6369454328  0.73718106  0.530861993  0.0668128312 -0.09063696
## Ed       0.7359970363 -0.76865789 -0.389922862 -0.2539735471  0.32283487
## Po1      0.7872252807 -0.63050025 -0.473247036  0.1033577449  0.68760446
## Po2      0.7942620503 -0.64815183 -0.473027293  0.0756266536  0.66671414
## LF       0.2946323090 -0.26988646 -0.250086098 -0.1236404364  0.18886635
## M.F      0.1796086363 -0.16708869 -0.050858258 -0.4276973791  0.21391426
## Pop      0.3082627091 -0.12629357 -0.347289063  0.4642104596  0.33747406
## NW      -0.5901070652  0.67731286  0.428059153  0.2303984071  0.03259884
## U1       0.0448572017 -0.06383218 -0.007469032 -0.1698528383 -0.05047792
## U2       0.0920716601  0.01567818 -0.061592474  0.1013583270  0.17732065
## Wealth   1.0000000000 -0.88399728 -0.555334708  0.0006485587  0.44131995
## Ineq    -0.8839972758  1.00000000  0.465321920  0.1018228182 -0.17902373
## Prob    -0.5553347075  0.46532192  1.000000000 -0.4362462614 -0.42742219
## Time     0.0006485587  0.10182282 -0.436246261  1.0000000000  0.14986606
## Crime    0.4413199490 -0.17902373 -0.427422188  0.1498660617  1.00000000
```

Al final decidimos que estaría interesante intentar explicar el crimen por los gastos para policía el mismo año (1960), renta media de las familias, la probabilidad de que tras hacer un delito te encarcelan (num de delitos/num de encarcelamientos) y finalmente media de los años pasados estudiando de la población de >25 años.

```
mlr4var<-lm(Crime~Po2+Wealth+Prob+Ed, data=uscrime); mlr4var
```

```
##
## Call:
## lm(formula = Crime ~ Po2 + Wealth + Prob + Ed, data = uscrime)
##
## Coefficients:
## (Intercept)          Po2       Wealth         Prob           Ed
##     564.693      121.103       -0.202    -3874.422       57.988
```

```
summary(mlr4var)
```

```
##
## Call:
## lm(formula = Crime ~ Po2 + Wealth + Prob + Ed, data = uscrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -696.80 -139.73   30.48  147.00  518.06
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.647e+02  4.743e+02   1.191   0.2405
## Po2          1.211e+02  2.510e+01   4.824 1.88e-05 ***
## Wealth      -2.020e-01  9.667e-02  -2.090   0.0427 *
## Prob        -3.874e+03  2.208e+03  -1.755   0.0866 .
## Ed           5.799e+01  5.623e+01   1.031   0.3083
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 282.5 on 42 degrees of freedom
## Multiple R-squared:  0.5128, Adjusted R-squared:  0.4664
## F-statistic: 11.05 on 4 and 42 DF,  p-value: 3.256e-06
```

El segundo modelo (el submodelo) será la crimen explicada no tanto por factores sociológicos como la educación o riqueza, sino de la atención del estado al crimen (cuanto se paga a la policía, con que probabilidad te echan a la cárcel..)

```
mlr2var<-lm(Crime~Po2+Prob, data=uscrime)
```

```
library("plot3D")
```

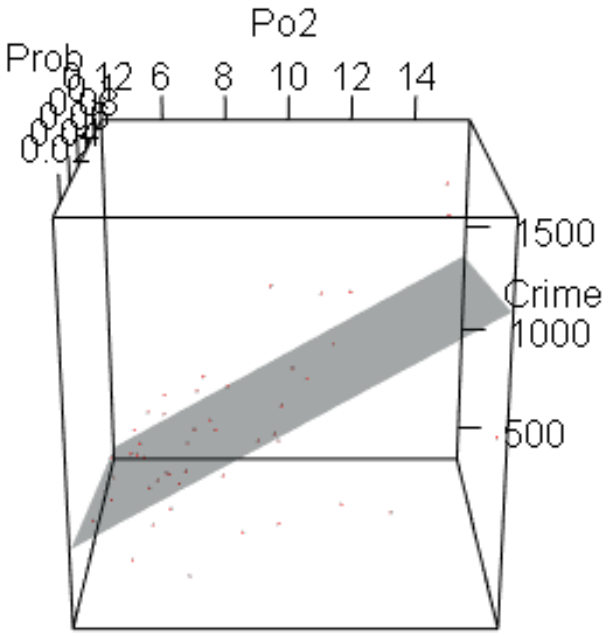
```
## Warning: package 'plot3D' was built under R version 4.1.3
```

```
grid.lines = 40
x.pred <- seq(min(uscrime$Po2), max(uscrime$Po2), length.out = grid.lines)
y.pred <- seq(min(uscrime$Prob), max(uscrime$Prob), length.out = grid.lines)
xy <- expand.grid( Po2 = x.pred, Prob = y.pred)
xy
```

```
##           Po2        Prob
## 1    4.100000 0.006900000
## 2    4.397436 0.006900000
## 3    4.694872 0.006900000
```

```
z.pred <- matrix(predict(mlr2var, newdata = xy),nrow = grid.lines, ncol =
grid.lines)

library(rgl)

library("plot3D")
data <- uscrime
mycolors<-c("red")

plot3d(
  x=uscrime$Po2, y=uscrime$Prob, z= uscrime$Crime,
  col = mycolors,
  type = 's',
  radius = 3,
  xlab = "Po2", ylab = "Prob", zlab = "Crime")

surface3d(x.pred, y.pred, z.pred, facets = NA, alpha=0.4, col="lightblue",
border="lightblue")
rglwidget()
```

```
## Warning in snapshot3d(scene = x, width = width, height = height): webshot
= TRUE
## requires the webshot2 package and Chrome browser; using rgl.snapshot()
instead
```



Ahora creamos el IC del valor medio de ambas variables. Estimando, vemos que el nivel del crimen estará entre 819.71 990.45 delitos por 100,000 personas y con predicción será entre 313.6 y 1496.57

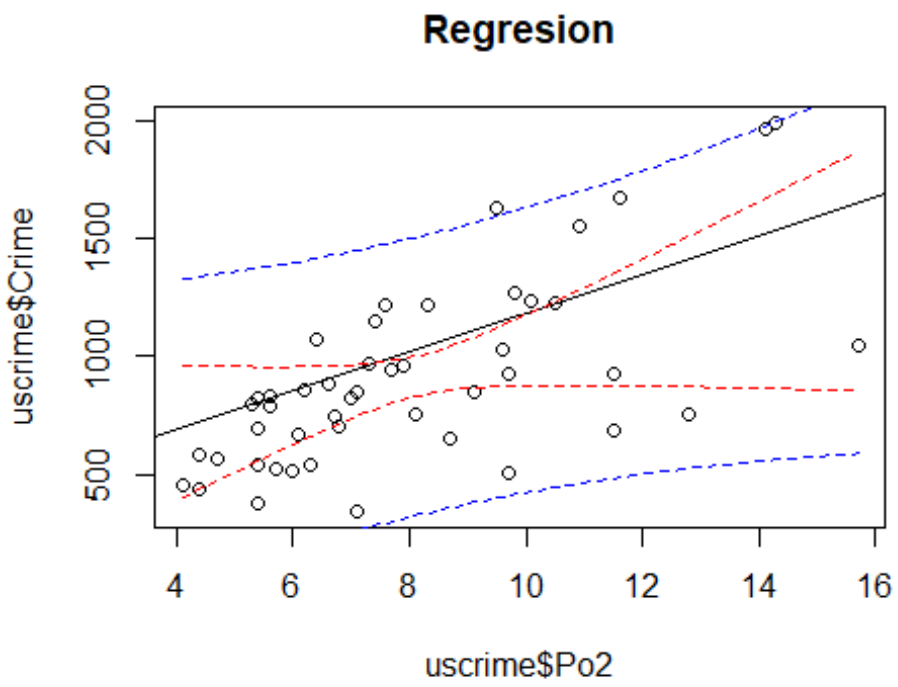
```
df<-data.frame(Po2=mean(uscrime$Po2), Prob=mean(uscrime$Prob))
predict(mlr2var, df, interval="prediction", level=0.95)

##        fit      lwr      upr
## 1 905.0851 313.5973 1496.573

predict(mlr2var, df, interval="confidence", level=0.95)

##        fit      lwr     upr
## 1 905.0851 819.7112 990.459
```

También creamos un intervalo de confianza para las variables regresoras.

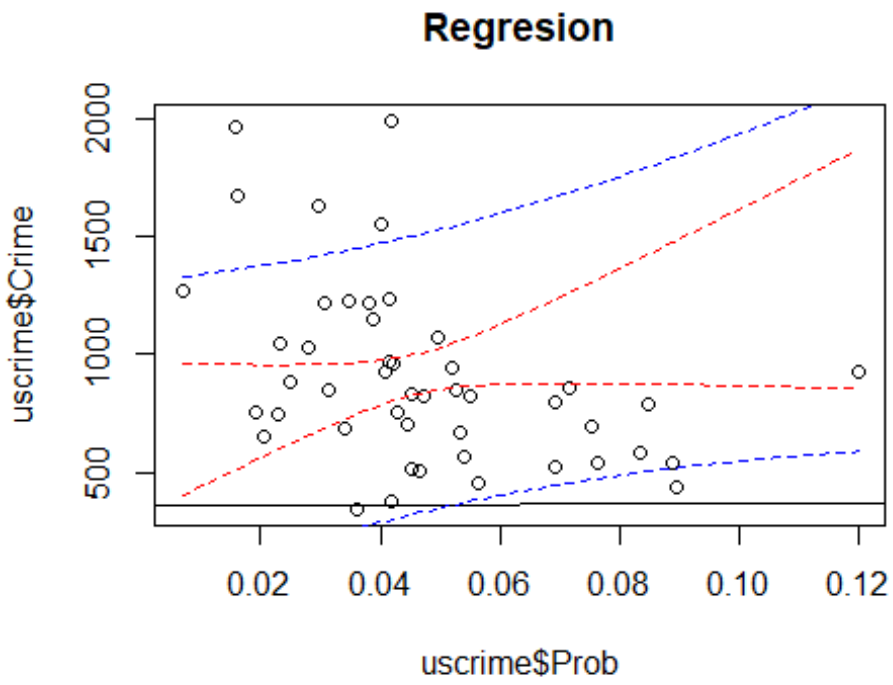
```
confint(mlr2var, level=0.95)

##                    2.5 %    97.5 %
## (Intercept)   -69.31558  782.4523
## Po2            47.74598  117.8056
## Prob        -6763.25039 1852.4993
```

Gráfico con las bandas de predicción y estimación:

```
po2.s<-seq(min(uscrime$Po2), max(uscrime$Po2), length=100)
prob.s<-seq(min(uscrime$Prob), max(uscrime$Prob), length=100)

crime.pred<-predict(mlr2var, data.frame(Po2=po2.s, Prob=prob.s),
interval="prediction", level=0.95)
crime.conf<-predict(mlr2var, data.frame(Po2=po2.s, Prob=prob.s),
interval="confidence", level=0.95)

plot(uscrime$Po2, uscrime$Crime, main="Regresión")
abline(mlr2var, col="black")

## Warning in abline(mlr2var, col = "black"): only using the first two of 3
## regression coefficients

lines(po2.s, crime.pred[,2], col="blue", lty=2)
lines(po2.s, crime.pred[,3], col="blue", lty=2)

lines(po2.s, crime.conf[,2], col="red", lty=2)
lines(po2.s, crime.conf[,3], col="red", lty=2)
```




```
plot(uscrime$Prob, uscrime$Crime, main="Regresión")
abline(mlr2var, col="black")

## Warning in abline(mlr2var, col = "black"): only using the first two of 3
## regression coefficients

lines(prob.s, crime.pred[,2], col="blue", lty=2)
lines(prob.s, crime.pred[,3], col="blue", lty=2)

lines(prob.s, crime.conf[,2], col="red", lty=2)
lines(prob.s, crime.conf[,3], col="red", lty=2)
```



De estas dos graficas se ven las bandas de confianza (de predicción y de estimación), es decir los límites de la recta de regresión. Se observa, ya que no hemos hecho gráficos hasta ahora que la pendiente de la variable Prob es nula (horizontal), es decir se puede dudar que aunque la variable esta correlada con la variable respuesta, no aporta para el modelo cuando tenemos Po2 ya en ello.

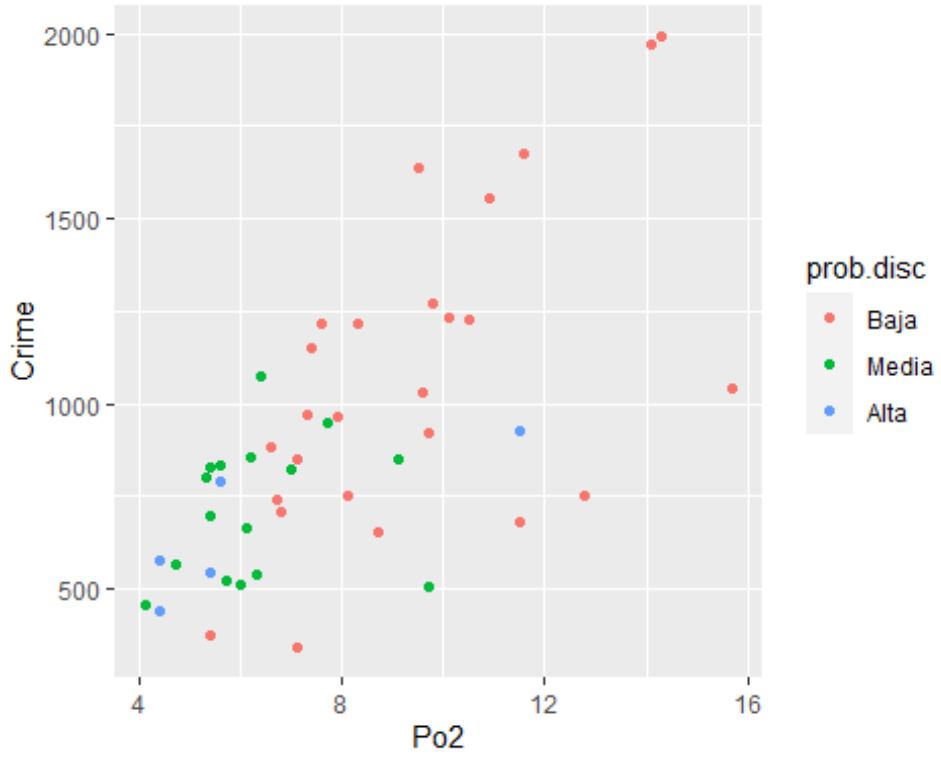
```
anova(mlr2var)

## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Po2        1 3058626 3058626 36.2649 3.132e-07 ***
## Prob       1  111290  111290  1.3195    0.2569
## Residuals 44 3711012   84341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Efectivamente, con un F=1.3195 y un p-valor=0.2569 no tenemos evidencias suficientes para rechazar H0 que la pendiente sea 0 y por esto se puede concluir que la variable Prob no aporta información relevante en este modelo.

Ahora generamos un diagrama de dispersión donde etiquetamos con colores las observaciones para indicar a qué categorías de una variable categórica pertenecen.

Vamos a discretizar la variable Prob en 3 niveles - probabilidad de encarcelar Baja, Media y Alta.

```r
prob.disc<-cut(uscrime$Prob, breaks=seq(min(uscrime$Prob), max(uscrime$Prob),
length.out=4), include.lowest = TRUE)
levels(prob.disc)<-c("Baja", "Media", "Alta")
table(prob.disc)

## prob.disc
##  Baja Media  Alta
##    26    16     5

uscrime$prob.disc<-prob.disc

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.2

ggplot(data =  uscrime) +
  geom_point(mapping = aes(x = Po2, y =Crime, colour=prob.disc))
```



Aquí observamos algo muy interesante. Estábamos esperando (en general) que la relación de los gastos de policía y nivel del crimen estén relacionados negativamente, es decir a mayor gasto para la protección por parte de la policía que haya menos crimen. Sin embargo, observamos todo lo contrario. Antes de hacer conclusiones, podemos suponer que esta relación entre las dos variables es puramente asociativa (sabiendo un valor de Po2 podemos predecir la Crimen) y no causativa. Es decir, no podemos concluir que cuanto más gastamos para policía tanto más crimen habrá.

Observando cómo hemos categorizado las observaciones, parece que los estados donde se gasta más para protección por parte de la policía hay una probabilidad baja de que se

encarcela una persona habiendo cometido un delito y probabilidad alta en los estados donde se paga menos a la policía.

Añadimos una variable categorica:

```
uscrime$So[which(uscrime$So==0)]<-"Northern"
uscrime$So[which(uscrime$So==1)]<-"Southern"
uscrime$So<-factor(uscrime$So, levels = c("Northern","Southern"))
mlr2var.alt<-lm(Crime~Po2+Prob+So, data=uscrime)
summary(mlr2var.alt)
```

```
##
## Call:
## lm(formula = Crime ~ Po2 + Prob + So, data = uscrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -633.73  -99.62   10.27  122.02  597.01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   329.51     201.03   1.639   0.1085
## Po2            89.51      16.75   5.346 3.24e-06 ***
## Prob        -4794.26    2251.50  -2.129   0.0390 *
## SoSouthern    244.20     101.63   2.403   0.0207 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 275.8 on 43 degrees of freedom
## Multiple R-squared:  0.5245, Adjusted R-squared:  0.4914
## F-statistic: 15.81 on 3 and 43 DF,  p-value: 4.496e-07
```

Vamos a dejar la categoría por defecto - 0 que significa que no es de un estado del sur. 1 será respectivamente estado del sur. Como se tiene la preocupación que en los estados del sur hay menos control de armas (por lo menos recientemente), tiene sentido que sepamos cuanto más crimen hay en los estados del sur si se paga igual a la policía que en el norte. Esto será que en los estados de sur con esta condición hay 244.2 delitos por 100,000 personas más que en los estados del norte.

Queremos comprobar a través de las sumas parciales cuál de las tres variables es la mejor para entrar como ultima. Esto será la variable que tiene menor SSR cuando el resto de las variables regresoras ya están en el modelo. Creamos 2 modelos con una secuencia de entrada de variables diferente:

```
modelo1<-lm(Crime~Po2+So+Prob, data=uscrime)
modelo2<-lm(Crime~Prob+So+Po2, data=uscrime)

anova(mlr2var.alt)
```

```
## Analysis of Variance Table
##
```

```
## Response: Crime
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Po2        1 3058626 3058626 40.1993 1.169e-07 ***
## Prob       1  111290  111290  1.4627   0.23311
## So         1  439293  439293  5.7736   0.02066 *
## Residuals 43 3271718   76086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modelo1)
```

```
## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Po2        1 3058626 3058626 40.1993 1.169e-07 ***
## So         1  205595  205595  2.7021   0.10751
## Prob       1  344989  344989  4.5342   0.03899 *
## Residuals 43 3271718   76086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modelo2)
```

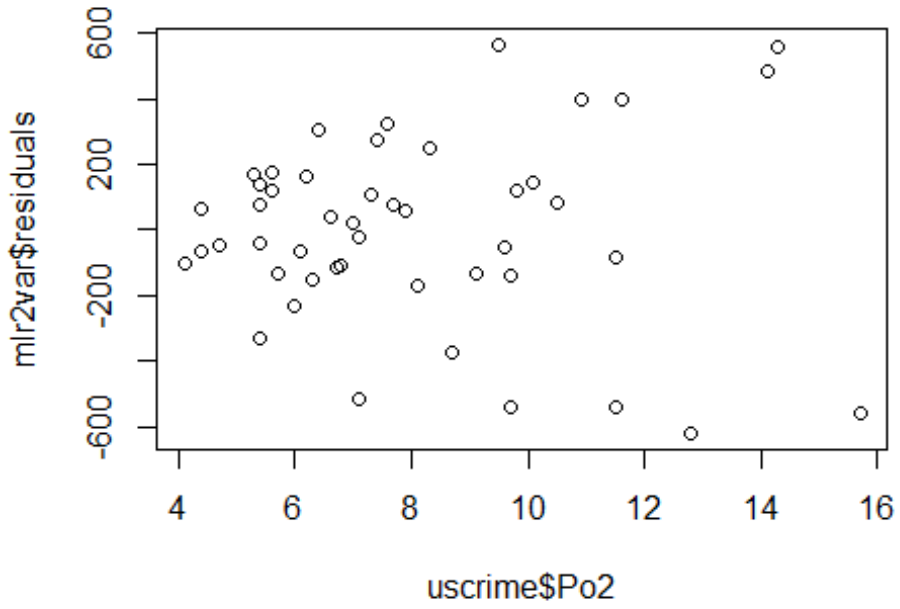
```
## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Prob       1 1257075 1257075 16.5217 0.0002008 ***
## So         1  177902  177902  2.3382 0.1335629
## Po2        1 2174233 2174233 28.5758 3.241e-06 ***
## Residuals 43 3271718   76086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se observa que el mejor modelo será el modelo1 en que Prob es la variable que entra ultima con SSR=344989. La que tiene mayor SSR es en modelo2, donde Po2 entra como ultima con SSR=2174233.

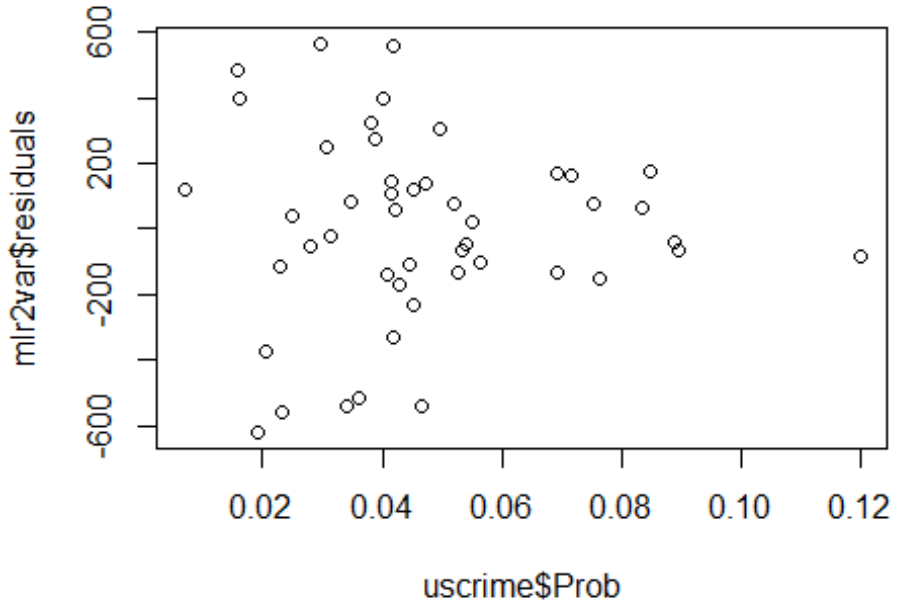
LINEALIDAD EN EL MODELO

Una cuestión importante es ver como entran las variables en el modelo. Considerando que vamos introduciendo una a una las variables cuantitativas predictoras, no se observa falta la linealidad. Sin embargo, este grafico de la variable regresora frente a los residuos no ensena exactamente la realidad, ya que hay otras variables en el modelo que pueden influir. Por esto creamos un gráfico de residuos parciales.

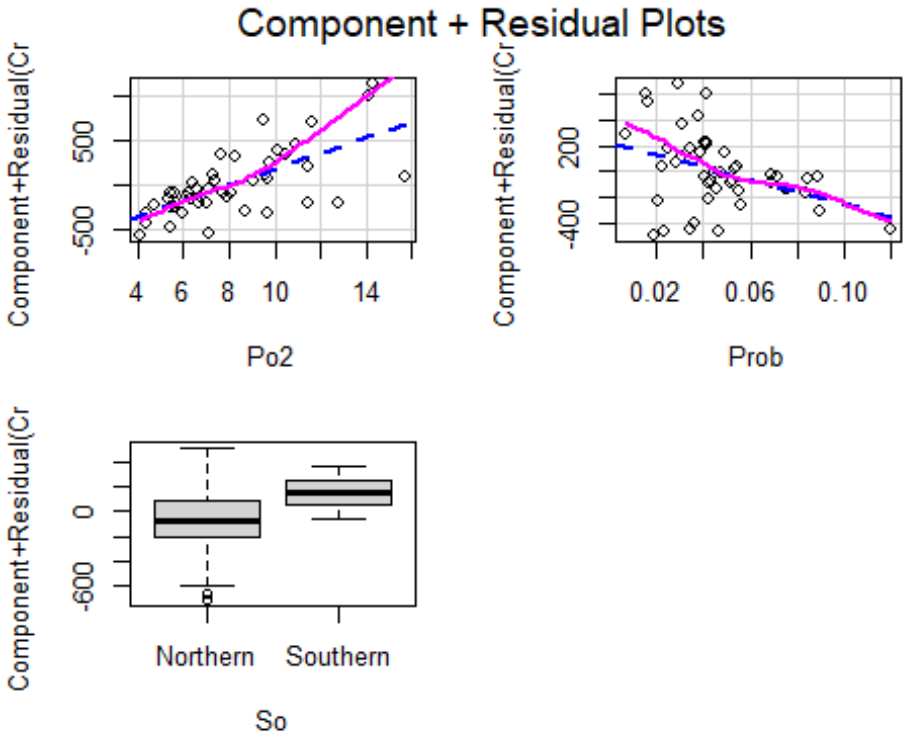
```
plot(uscrime$Po2, mlr2var$residuals)
```




```
plot(uscrime$Prob, mlr2var$residuals)
```

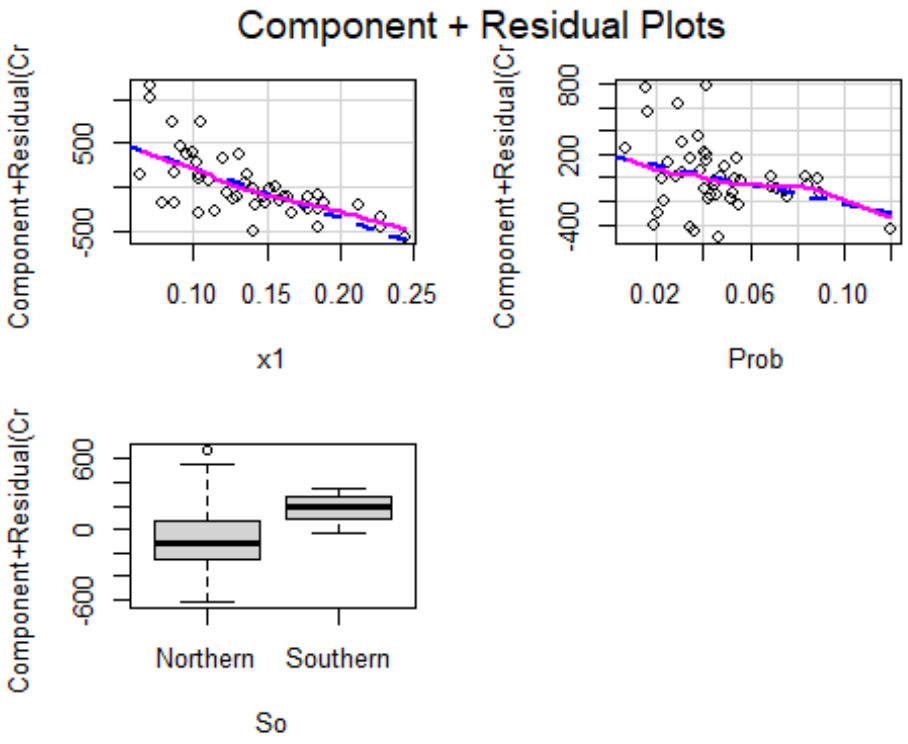


```
library(car)

crPlots(mlr2var.alt)
```



Podemos cuestionar si Po2 entra de forma lineal, la pendiente parece un poco curvada. Lo que podemos hacer es transformarla (con el riesgo de empeorar la interpretabilidad).

```
#modelo alternativo que pueda resolver el problema de que Po2 no entra de
forma lineal
x1<-1/uscrime$Po2
mlr2var.alt2<-lm(Crime~x1+Prob+So, data=uscrime)
crPlots(mlr2var.alt2)
```



Podemos ver que hemos conseguido corregir la linealidad con la transformación. X1 será la inversa de la Po2.

CAPACIDAD PREDICTIVA DEL MODELO

Luego, para comprobar cual modelo tiene la mejor capacidad para predecir nuevos datos, hacemos una comparación de los Rpress.

Creamos una función para calcular los valores PRESS

```
PRESS <- function(model) { i <- residuals(model)/(1 -
lm.influence(model)$hat); sum(i^2)}

PRESS(mlr4var)

## [1] 4426891

PRESS(mlr2var)

## [1] 4386477

PRESS(mlr2var.alt)

## [1] 3990358
```

Como buscamos el menor valor de PRESS, suponemos que el mejor modelo será el mlr2var.alt

```
a1<-anova(mlr4var)
scta1<-sum(a1$`Sum Sq`)
1-PRESS(mlr4var)/scta1

## [1] 0.3566433

a2<-anova(mlr2var)
scta2<-sum(a2$`Sum Sq`)
1-PRESS(mlr2var)/scta2

## [1] 0.3625167

a3<-anova(mlr2var.alt)
scta3<-sum(a3$`Sum Sq`)
1-PRESS(mlr2var.alt)/scta3

## [1] 0.4200842
```

Efectivamente, aunque es un valor relativamente pequeño (buscamos valores lo más altos posibles de Rpress) el modelo mlr2var.alt tiene la mejor capacidad predictiva.

CASOS INFLUYENTES

Vamos a ver cuál es el caso más influyente de ente modelo. Nuestro criterio será que una observación que tiene un leverage (hat value) mayor que 2*(3+1)/47 se considera influyente.

```
influence(mlr2var.alt)$hat

##          1          2          3          4          5          6
7
## 0.08804989 0.03726857 0.09098724 0.12953114 0.04064612 0.05688484
0.09021210
##          8          9         10         11         12         13
14
## 0.13335552 0.06623264 0.04489187 0.07314507 0.05032483 0.05713648
0.05938876
##         15         16         17         18         19         20
21
## 0.06805513 0.07164324 0.11767717 0.45681135 0.09020996 0.04197967
0.07573917
##         22         23         24         25         26         27
28
## 0.10489515 0.03841151 0.03930088 0.09723605 0.15167341 0.04432770
0.03744373
##         29         30         31         32         33         34
35
## 0.18829119 0.07183438 0.07089230 0.03411552 0.07693117 0.03651999
0.05385222
##         36         37         38         39         40         41
42
## 0.09152052 0.09506966 0.08756404 0.09271209 0.09926524 0.07165889
```

```
0.18471575
##         43         44         45         46         47
## 0.06704586 0.03903669 0.09524662 0.04238578 0.04788293

criterio<-2*(3+1)/47
unname(which(influence(mlr2var.alt)$hat>criterio))

## [1] 18 29 42
```

Observaciones 18, 29 y 42 son influyentes en el modelo mlr2var.alt.

```
uscrime.alt<-uscrime[-c(18,29,42),]
mlr2var.sin<-lm(Crime~Po2+Prob+So, data=uscrime.alt)



anova(mlr2var.alt)

## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Po2        1 3058626 3058626 40.1993 1.169e-07 ***
## Prob       1  111290  111290  1.4627   0.23311
## So         1  439293  439293  5.7736   0.02066 *
## Residuals 43 3271718   76086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(mlr2var.sin)

## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Po2        1 3455667 3455667 49.9460 1.512e-08 ***
## Prob       1    2130    2130  0.0308   0.86162
## So         1  503263  503263  7.2738   0.01019 *
## Residuals 40 2767525   69188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Quitando las observaciones influyentes, podemos ver que el MSE ha mejorado de 76086 a 69188.

PROBLEMAS DE MULTICOLINEALIDAD

Para comprobar si hay multicolinealidad, primero vamos a ver la matriz de correlaciones:

```
cor(uscrime[,c("Po2", "Prob")])
```

```
##              Po2       Prob
## Po2   1.0000000 -0.4730273
## Prob -0.4730273  1.0000000
```

También podemos ver para el modelo con 4 variables:

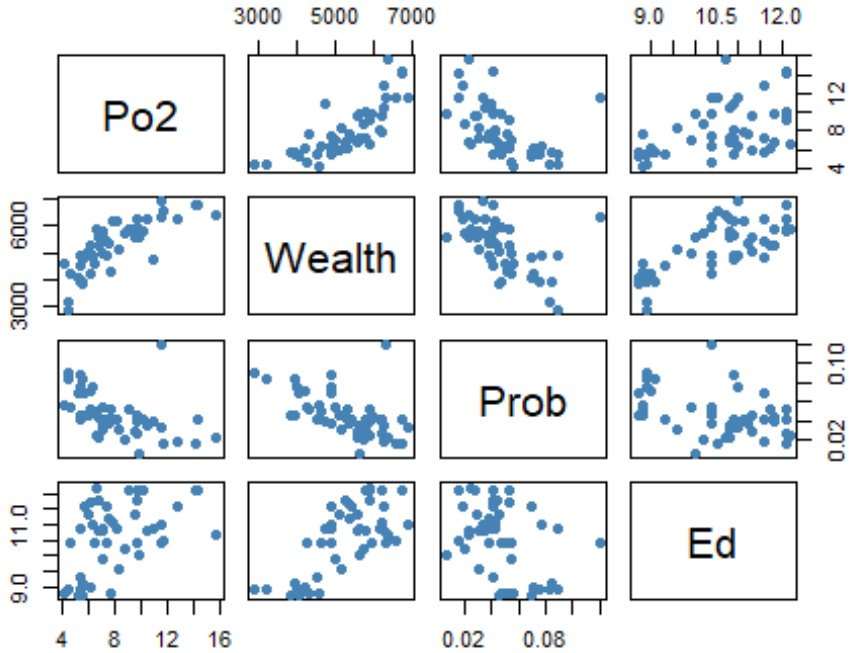
```
cor(uscrime[,c("Po2", "Wealth", "Prob", "Ed")])
```

```
##                Po2     Wealth       Prob         Ed
## Po2     1.0000000  0.7942621 -0.4730273  0.4994096
## Wealth  0.7942621  1.0000000 -0.5553347  0.7359970
## Prob   -0.4730273 -0.5553347  1.0000000 -0.3899229
## Ed      0.4994096  0.7359970 -0.3899229  1.0000000
```

Aquí en el segundo modelo nos puede preocupar una posible falta de independencia entre Po2 y Wealth o entre Wealth y Educación.

Podemos graficar las relaciones entre las variables.

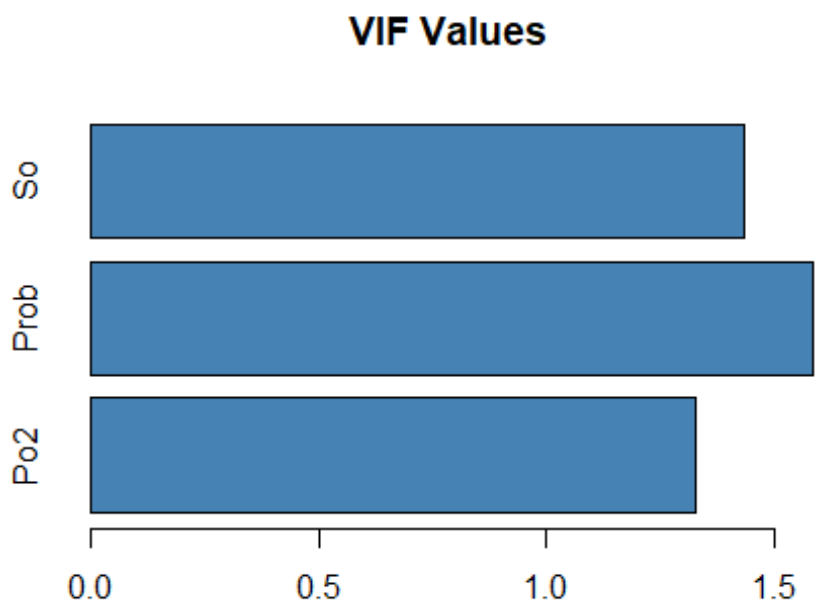
```
plot(uscrime[,c("Po2", "Wealth", "Prob", "Ed")], pch=20, cex=1.5,
col='steelblue')
```



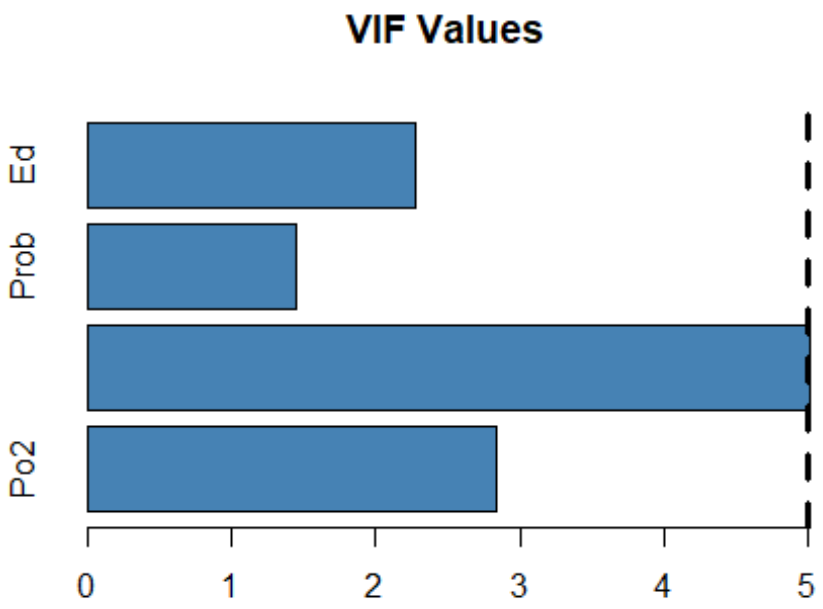
Y finalmente, un buen criterio para comprobar la multicolinealidad es a través del VIF. Valores mayores que 5 nos ensenaran si alguna variable tiene una correlación alta entre ella y el resto de predictoras.

```
vif(mlr2var.alt)
```

```
##      Po2     Prob       So
## 1.325434 1.584385 1.432585

barplot(vif(mlr2var.alt), main = "VIF Values", horiz = TRUE, col =
"steelblue")
abline(v = 5, lwd = 3, lty = 2)
```




```
vif(mlr4var)

##      Po2   Wealth     Prob       Ed
## 2.839697 5.013840 1.452347 2.280655

barplot(vif(mlr4var), main = "VIF Values", horiz = TRUE, col = "steelblue")
abline(v = 5, lwd = 3, lty = 2)
```



En el segundo modelo, la variable Wealth nos puede causar problemas de multicolinealidad (que sea correlada con el resto de variables predictoras), aunque es muy poco mayor que 5.

ELEGIR VARIABLES PARA EL MODELO

Otra buena manera de elegir variables que pueden formar parte en nuestro modelo es a través del método forward y backward que consiste en encontrar un modelo con que introduciendo (o quitando en el caso de backward) variables la F sea máxima posible (F=MSModel/MSError), es decir que el modelo explica la mayor proporción de la variabilidad posible.

```
#con backward
library(MASS)
full.model<-lm(Crime~., data=uscrime)
modback <- stepAIC(full.model, trace=TRUE, direction="backward")

## Start:  AIC=510.27
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob + Time + prob.disc
##
##             Df Sum of Sq     RSS    AIC
## - LF         1        50 1133739 508.27
## - NW         1        54 1133744 508.27
## - So         1       755 1134444 508.30
## - Pop        1      7234 1140923 508.57
## - Time       1     24172 1157862 509.26
## - Po2        1     26425 1160114 509.35
```

```
## - M.F         1     29904 1163593 509.49
## - Wealth      1     36149 1169838 509.74
## - Prob        1     39485 1173174 509.88
## <none>                    1133689 510.27
## - Po1         1    110568 1244257 512.64
## - U1          1    150429 1284118 514.13
## - prob.disc   2    221257 1354946 514.65
## - M           1    250342 1384031 517.65
## - U2          1    254546 1388235 517.79
## - Ed          1    308188 1441877 519.57
## - Ineq        1    497075 1630764 525.36
##
## Step:  AIC=508.27
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob + Time + prob.disc
##
##             Df Sum of Sq     RSS    AIC
## - NW         1       116 1133855 506.28
## - So         1       746 1134484 506.30
## - Pop        1      7199 1140937 506.57
## - Time       1     24346 1158084 507.27
## - Po2        1     29500 1163239 507.48
## - Wealth     1     37154 1170893 507.79
## - Prob       1     40130 1173868 507.91
## - M.F        1     40961 1174700 507.94
## <none>                    1133739 508.27
## - Po1        1    117781 1251520 510.92
## - prob.disc  2    230124 1363862 512.96
## - U1         1    175587 1309326 513.04
## - U2         1    254575 1388314 515.79
## - M          1    256777 1390515 515.87
## - Ed         1    352156 1485895 518.98
## - Ineq       1    510600 1644339 523.75
##
## Step:  AIC=506.28
## Crime ~ M + So + Ed + Po1 + Po2 + M.F + Pop + U1 + U2 + Wealth +
##     Ineq + Prob + Time + prob.disc
##
##             Df Sum of Sq     RSS    AIC
## - So         1      1037 1134892 504.32
## - Pop        1      7135 1140990 504.57
## - Time       1     24877 1158732 505.30
## - Po2        1     29965 1163820 505.50
## - Wealth     1     37337 1171192 505.80
## - Prob       1     40333 1174187 505.92
## - M.F        1     41010 1174865 505.95
## <none>                    1133855 506.28
## - Po1        1    118610 1252465 508.95
## - U1         1    175736 1309591 511.05
## - prob.disc  2    242148 1376003 511.37
```

```
## - U2         1    254965 1388820 513.81
## - M          1    280262 1414117 514.66
## - Ed         1    353244 1487099 517.02
## - Ineq       1    530180 1664034 522.31
##
## Step:  AIC=504.32
## Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + U1 + U2 + Wealth + Ineq +
##     Prob + Time + prob.disc
##
##             Df Sum of Sq     RSS    AIC
## - Pop        1      7661 1142552 502.64
## - Time       1     24677 1159569 503.33
## - Po2        1     28929 1163820 503.50
## - Wealth     1     39080 1173971 503.91
## - M.F        1     39982 1174874 503.95
## - Prob       1     41714 1176606 504.02
## <none>                   1134892 504.32
## - Po1        1    118147 1253039 506.97
## - prob.disc  2    248097 1382989 509.61
## - U1         1    194543 1329435 509.76
## - U2         1    285562 1420454 512.87
## - M          1    317039 1451931 513.90
## - Ed         1    353936 1488828 515.08
## - Ineq       1    663404 1798295 523.95
##
## Step:  AIC=502.64
## Crime ~ M + Ed + Po1 + Po2 + M.F + U1 + U2 + Wealth + Ineq +
##     Prob + Time + prob.disc
##
##             Df Sum of Sq     RSS    AIC
## - Po2        1     29258 1171811 501.82
## - Time       1     35506 1178058 502.07
## - Wealth     1     36817 1179369 502.13
## - Prob       1     43344 1185896 502.39
## <none>                   1142552 502.64
## - M.F        1     63335 1205887 503.17
## - Po1        1    113814 1256367 505.10
## - prob.disc  2    259702 1402254 508.26
## - U1         1    214040 1356593 508.71
## - U2         1    291991 1434543 511.33
## - M          1    337734 1480287 512.81
## - Ed         1    349097 1491650 513.17
## - Ineq       1    674136 1816688 522.43
##
## Step:  AIC=501.82
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Wealth + Ineq + Prob +
##     Time + prob.disc
##
##             Df Sum of Sq     RSS    AIC
## - Time       1     18989 1190800 500.58
```

```
## - Prob       1     31712 1203523 501.08
## - Wealth     1     34451 1206261 501.19
## <none>                   1171811 501.82
## - M.F        1     91922 1263733 503.37
## - prob.disc  2    250850 1422661 506.94
## - U1         1    212000 1383810 507.64
## - U2         1    292840 1464651 510.31
## - M          1    320420 1492231 511.18
## - Ed         1    337694 1509505 511.73
## - Ineq       1    693644 1865455 521.68
## - Po1        1    775145 1946956 523.69
##
## Step:  AIC=500.58
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Wealth + Ineq + Prob +
##     prob.disc
##
##             Df Sum of Sq     RSS    AIC
## - Prob       1     16326 1207126 499.22
## - Wealth     1     28330 1219129 499.68
## <none>                   1190800 500.58
## - M.F        1    124410 1315210 503.25
## - prob.disc  2    235775 1426575 505.07
## - U1         1    201214 1392014 505.92
## - U2         1    280886 1471686 508.53
## - M          1    301491 1492291 509.19
## - Ed         1    365050 1555849 511.15
## - Ineq       1    675108 1865907 519.69
## - Po1        1    801581 1992380 522.77
##
## Step:  AIC=499.22
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Wealth + Ineq + prob.disc
##
##             Df Sum of Sq     RSS    AIC
## - Wealth     1     27100 1234226 498.26
## <none>                   1207126 499.22
## - M.F        1    111942 1319067 501.39
## - U1         1    214451 1421577 504.91
## - U2         1    300952 1508077 507.68
## - M          1    302870 1509996 507.74
## - prob.disc  2    418020 1625145 509.20
## - Ed         1    378192 1585318 510.03
## - Ineq       1    678354 1885479 518.18
## - Po1        1    810937 2018063 521.37
##
## Step:  AIC=498.26
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + prob.disc
##
##             Df Sum of Sq     RSS    AIC
## <none>                   1234226 498.26
## - M.F        1    125836 1360061 500.83
```

```
## - U1         1    254485 1488710 505.07
## - M          1    279024 1513250 505.84
## - U2         1    362441 1596666 508.36
## - prob.disc  2    466820 1701046 509.34
## - Ed         1    441246 1675471 510.63
## - Ineq       1    850821 2085047 520.91
## - Po1        1   1223100 2457325 528.63
```

```
horizonte<-(Crime~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
Wealth + Ineq + Prob + Time)
empty.model<-lm(Crime~1,data=uscrime)
modfor <- stepAIC(empty.model, trace=TRUE, direction="forward",
scope=horizonte)
```

```
## Start:  AIC=561.02
## Crime ~ 1
##
##          Df Sum of Sq     RSS    AIC
## + Po1     1   3253302 3627626 532.94
## + Po2     1   3058626 3822302 535.39
## + Wealth  1   1340152 5540775 552.84
## + Prob    1   1257075 5623853 553.54
## + Pop     1    783660 6097267 557.34
## + Ed      1    717146 6163781 557.85
## + M.F     1    314867 6566061 560.82
## <none>                6880928 561.02
## + LF      1    245446 6635482 561.32
## + Ineq    1    220530 6660397 561.49
## + U2      1    216354 6664573 561.52
## + Time    1    154545 6726383 561.96
## + So      1     56527 6824400 562.64
## + M       1     55084 6825844 562.65
## + U1      1     17533 6863395 562.90
## + NW      1      7312 6873615 562.97
##
## Step:  AIC=532.94
## Crime ~ Po1
##
##          Df Sum of Sq     RSS    AIC
## + Ineq    1    739819 2887807 524.22
## + M       1    616741 3010885 526.18
## + M.F     1    250522 3377104 531.57
## + NW      1    232434 3395192 531.82
## + So      1    219098 3408528 532.01
## + Wealth  1    180872 3446754 532.53
## <none>                3627626 532.94
## + Po2     1    146167 3481459 533.00
## + Prob    1     92278 3535348 533.72
## + LF      1     77479 3550147 533.92
## + Time    1     43185 3584441 534.37
```

```
## + U2      1     17848 3609778 534.70
## + Pop     1      5666 3621959 534.86
## + U1      1      2878 3624748 534.90
## + Ed      1       767 3626859 534.93
##
## Step:  AIC=524.22
## Crime ~ Po1 + Ineq
##
##          Df Sum of Sq     RSS    AIC
## + Ed      1    587050 2300757 515.53
## + M.F     1    454545 2433262 518.17
## + Prob    1    280690 2607117 521.41
## + LF      1    260571 2627236 521.77
## + Wealth  1    213937 2673871 522.60
## + M       1    181236 2706571 523.17
## + Pop     1    130377 2757430 524.04
## <none>                2887807 524.22
## + NW      1     36439 2851369 525.62
## + So      1     33738 2854069 525.66
## + Po2     1     30673 2857134 525.71
## + U1      1      2309 2885498 526.18
## + Time    1       497 2887310 526.21
## + U2      1       253 2887554 526.21
##
## Step:  AIC=515.53
## Crime ~ Po1 + Ineq + Ed
##
##          Df Sum of Sq     RSS    AIC
## + M       1    239405 2061353 512.37
## + Prob    1    234981 2065776 512.47
## + M.F     1    117026 2183731 515.08
## <none>                2300757 515.53
## + Wealth  1     79540 2221218 515.88
## + U2      1     62112 2238646 516.25
## + Time    1     61770 2238987 516.26
## + Po2     1     42584 2258174 516.66
## + Pop     1     39319 2261438 516.72
## + U1      1      7365 2293392 517.38
## + LF      1      7254 2293503 517.39
## + NW      1      4210 2296547 517.45
## + So      1      4135 2296622 517.45
##
## Step:  AIC=512.37
## Crime ~ Po1 + Ineq + Ed + M
##
##          Df Sum of Sq     RSS    AIC
## + Prob    1    258063 1803290 508.08
## + U2      1    200988 1860365 509.55
## + Wealth  1    163378 1897975 510.49
## <none>                2061353 512.37
```

```
## + M.F     1     74398 1986955 512.64
## + U1      1     50835 2010518 513.20
## + Po2     1     45392 2015961 513.32
## + Time    1     42746 2018607 513.39
## + NW      1     16488 2044865 513.99
## + Pop     1      8101 2053251 514.19
## + So      1      3189 2058164 514.30
## + LF      1      2988 2058365 514.30
##
## Step:  AIC=508.08
## Crime ~ Po1 + Ineq + Ed + M + Prob
##
##          Df Sum of Sq     RSS    AIC
## + U2      1    192233 1611057 504.79
## + Wealth  1     86490 1716801 507.77
## + M.F     1     84509 1718781 507.83
## <none>                1803290 508.08
## + U1      1     52313 1750977 508.70
## + Pop     1     47719 1755571 508.82
## + Po2     1     37967 1765323 509.08
## + So      1     21971 1781320 509.51
## + Time    1     10194 1793096 509.82
## + LF      1       990 1802301 510.06
## + NW      1       797 1802493 510.06
##
## Step:  AIC=504.79
## Crime ~ Po1 + Ineq + Ed + M + Prob + U2
##
##          Df Sum of Sq     RSS    AIC
## <none>                1611057 504.79
## + Wealth  1     59910 1551147 505.00
## + U1      1     54830 1556227 505.16
## + Pop     1     51320 1559737 505.26
## + M.F     1     30945 1580112 505.87
## + Po2     1     25017 1586040 506.05
## + So      1     17958 1593098 506.26
## + LF      1     13179 1597878 506.40
## + Time    1      7159 1603898 506.58
## + NW      1       359 1610698 506.78
```

```
summary(mlr2var.alt)
```

```
##
## Call:
## lm(formula = Crime ~ Po2 + Prob + So, data = uscrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -633.73  -99.62   10.27  122.02  597.01
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   329.51     201.03   1.639   0.1085
## Po2            89.51      16.75   5.346 3.24e-06 ***
## Prob        -4794.26    2251.50  -2.129   0.0390 *
## SoSouthern    244.20     101.63   2.403   0.0207 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 275.8 on 43 degrees of freedom
## Multiple R-squared:  0.5245, Adjusted R-squared:  0.4914
## F-statistic: 15.81 on 3 and 43 DF,  p-value: 4.496e-07

summary(modback)

##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + prob.disc,
##     data = uscrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -419.33 -104.67  -25.01  106.87  473.76
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6596.25    1134.73  -5.813 1.12e-06 ***
## M                 90.65      31.34   2.892 0.006373 **
## Ed               179.03      49.22   3.637 0.000835 ***
## Po1               91.54      15.12   6.055 5.28e-07 ***
## M.F               25.44      13.10   1.942 0.059748 .
## U1             -9040.82    3273.20  -2.762 0.008891 **
## U2               225.82      68.51   3.296 0.002168 **
## Ineq              66.65      13.20   5.050 1.21e-05 ***
## prob.discMedia  -279.02      77.40  -3.605 0.000915 ***
## prob.discAlta   -257.56     103.57  -2.487 0.017521 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 182.6 on 37 degrees of freedom
## Multiple R-squared:  0.8206, Adjusted R-squared:  0.777
## F-statistic: 18.81 on 9 and 37 DF,  p-value: 2.84e-11

summary(modfor)

##
## Call:
## lm(formula = Crime ~ Po1 + Ineq + Ed + M + Prob + U2, data = uscrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -470.68  -78.41  -19.68  133.12  556.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
## Po1           115.02      13.75   8.363 2.56e-10 ***
## Ineq           67.65      13.94   4.855 1.88e-05 ***
## Ed            196.47      44.75   4.390 8.07e-05 ***
## M             105.02      33.30   3.154  0.00305 **
## Prob        -3801.84    1528.10  -2.488  0.01711 *
## U2             89.37      40.91   2.185  0.03483 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200.7 on 40 degrees of freedom
## Multiple R-squared:  0.7659, Adjusted R-squared:  0.7307
## F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

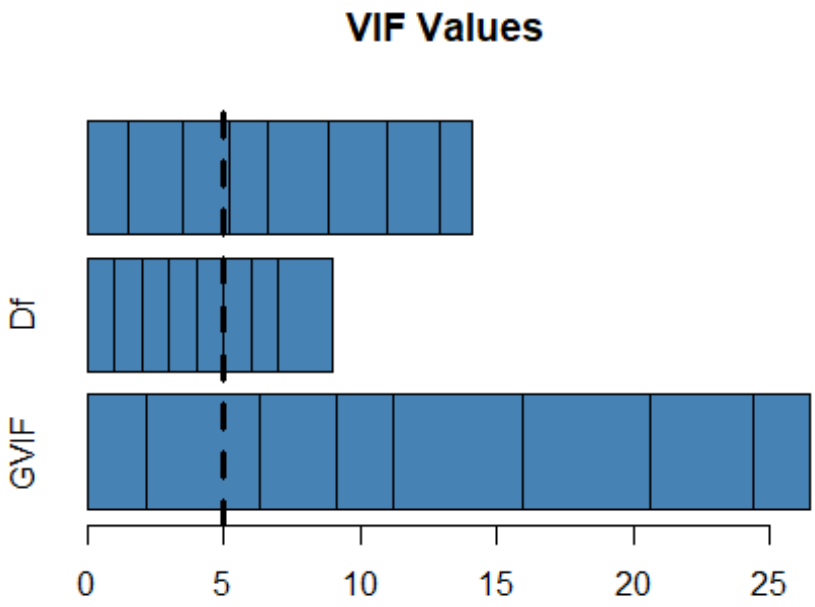
Aplicando los 3 criterios de bondad de ajuste para ver cuál es el mejor modelo, aunque con el test de F todos los modelos salen significativos (el modelo lineal explica bien la respuesta) el modelo backward es el que tiene mayor R-ajustada (parte de la varianza explicada por la recta de regresión, teniendo en cuenta las variables ya incluidas) con un valor de 0.7444 y tiene el menor error estandarizado.

Si queremos ver si hay problemas de multicolinealidad:

```
vif(modback)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## M          2.139783  1        1.462800
## Ed         4.181611  1        2.044899
## Po1        2.783602  1        1.668413
## M.F        2.054654  1        1.433406
## U1         4.802241  1        2.191402
## U2         4.616073  1        2.148505
## Ineq       3.822818  1        1.955203
## prob.disc 2.058556  2        1.197818
```

```
barplot(vif(modback), main = "VIF Values", horiz = TRUE, col = "steelblue")
abline(v = 5, lwd = 3, lty = 2)
```



No hay problemas de multicolinealidad, hay una sola variable (M.F) que puede ser redundante (no se puede rechazar que su pendiente sea nula), pero sobre todo es el mejor modelo que hemos sido capaces de encontrar.