Punto 5 Trabajo 1

Jhonatan Smith Garcia

3/12/2021

```
# Librerias
require(olsrr)
```

```
## Loading required package: olsrr
```

```
## Warning: package 'olsrr' was built under R version 4.0.5
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

Ejercicio 5:

La base de datos a trabajar (SURGICAL) contiene informacion acerca de la supervivencia de pacientes con intervenciones quirurgicas hepaticas. Las variables presentadas son:

- liver_test : Puntuacion de la funcion de prueba hepatica
- enzyme_test : Resultado prueba de enzimas.
- pindex : Indice de pronostico.
- bcs : Puntuacion de coagulacion sanguinea.
- age : edad en años del paciente en cuestion
- gender : Genero del paciente. Variable indicadora que toma el valor de 1 para femenino y 0 para masculino
- alc_mod: Consumo de alcohol. Variable indicadora que toma el valor para 0 como no consume licor, 1 consumo moderado.
- alc_heavy : Variable indicadora del historial del consumo de alcohol. 0 para no consumo, 1 para consumo fuerte.
- y : Tiempo de supervivencia.

Se busca explicar el tiempo de supervivencia a traves de las variables disponibles. Es decir, ¿cuales de todas estas variables son las mejores para explicar el tiempo de supervivencia de un paciente sometido a una intervencion quirurgica hepatica?

Procediendo con validacion cruzada, se selecciona el 70% de la base de datos para entrenamiento y el 30% como datos de validacion o prueba (valores tomados empiricamente).

```
set.seed(1998)
porcentaje = 0.7 # Proporcion seleccioar base de datos
t_1 = sample(length(surgical$y), size = (length(surgical$y)*porcentaje))
train = surgical[t_1,]
test = surgical[-t_1,]
```

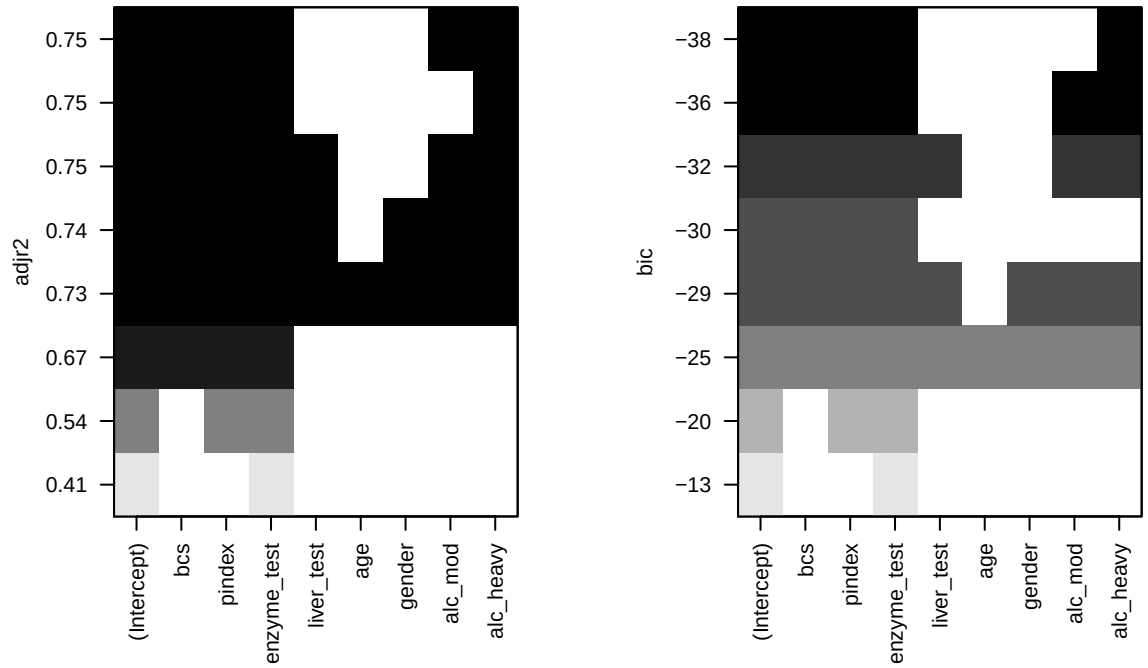
Una vez que la base de datos ha sido seleccionada, se utiliza la funcion *regsubset* para la seleccion de variables "hacia adelante".

```
library("leaps")
```

```
## Warning: package 'leaps' was built under R version 4.0.5
```

```
reg_fitted = regsubsets(y~., data= train, nvmax = 8, method = "forward")
```

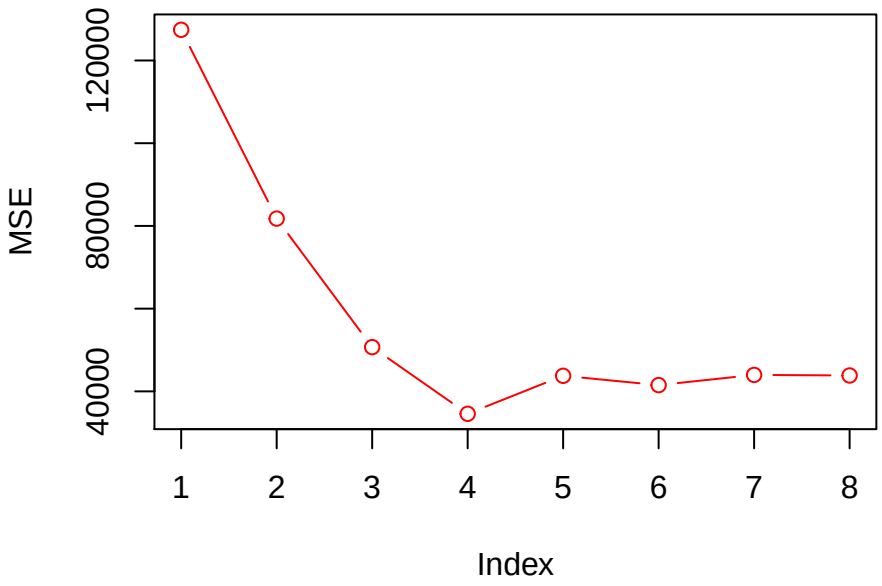
Con este metodo, se obtiene el siguiente grafico para la toma de decisiones segun diversos criterios.



Con este grafico, se busca seleccionar via el criterio de R^2_{adj} y bic los modelos más "consistentes" entre si, puesto que se busca seleccionar el valor mas alto en el primero teniendo siempre en cuenta el principio de parsimonia y con bic el valor mas pequeño.

- Segun R^2_{adj} el mejor modelo es el que tiene 6 variables; intercepto, bcs, pindex, enzym_test, alc_mod y alc_heavy
- Segun criterio bic el mejor modelo contiene 5 variables; intercept,bcs, pindex, enzyme_test y alc_heaby

Ahora, ¿Que modelo es mejor? Para responder a esta pregunta, se procede a analizar el MSE de cada modelo.



Con el grafico anterior, se observa los valores del MSE. El modelo con menor MSE es el de indice 4.

```
Smith <- which.min(MSE)
regfit.for <- regsubsets(y~.,data=surgical ,nvmax =8, method = "forward")
```

```
coef(regfit.for,Smith)
```

```
## (Intercept)      pindex enzyme_test  liver_test   alc_heavy
## -789.012038    7.876493    7.547724  125.473761  359.874634
```

Segun esto, las variables que mejor explica la variable respuesta están dadas por las anteriores.

El modelo es, finalmente:

$Y = -789.912038 + 7876493 * pindex + 7.547724 * enzyme.test + 125.473761 * liver.test + 359.87464 * alc.heavy$

Sin embargo, ¿esto cambiará radicalmente si se utiliza otro metodo de seleccion? Lo recomendable seria verificar varios metodos de seleccion de variables y observar segun el problema, cual seria el modelo más optimo.

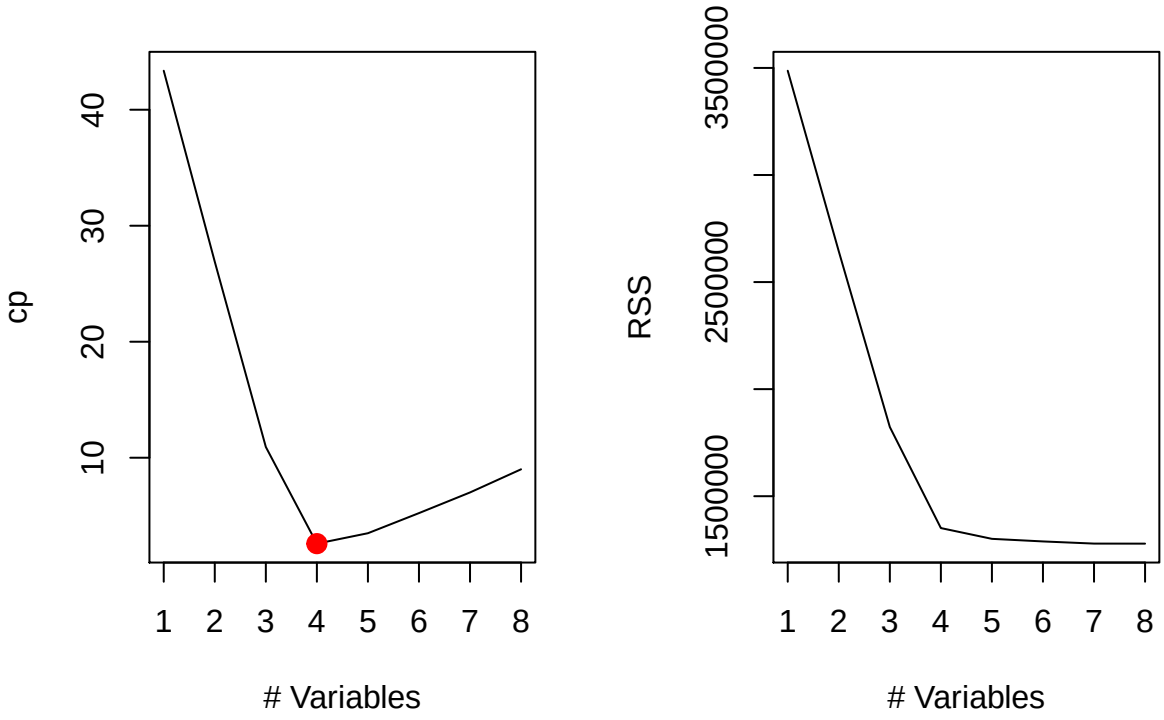
Suponga que hay un experto que está a cargo de la investigacion. Esta persona podría ayudar a seleccionar cuales de todos los modelos resultantes de todos los posibles metodos de seleccion de variables (llamese cp, backward, step-wise, por nombrar algunos) serian mas utiles acorde a su conocimiento en el campo.

¿Por qué? Sin importar que metodo se utilice, se espera cierta consistencia entre metodos. Puedan diferir quizas en algunas variables pero de manera general, deberian seleccionar mas o menos las mismas. Un ejemplo para ilustrar esto con otros criterios de seleccion en forward.

```
reg_fitted_cp = regsubsets(y~., data= train, nvmax = 8)
reg.summary = summary(reg_fitted_cp)
```

```
par(mfrow = c(1,2))
plot(reg.summary$cp, xlab = "# Variables", ylab = "cp", type = "l")
a1 = which.min(reg.summary$cp)
```

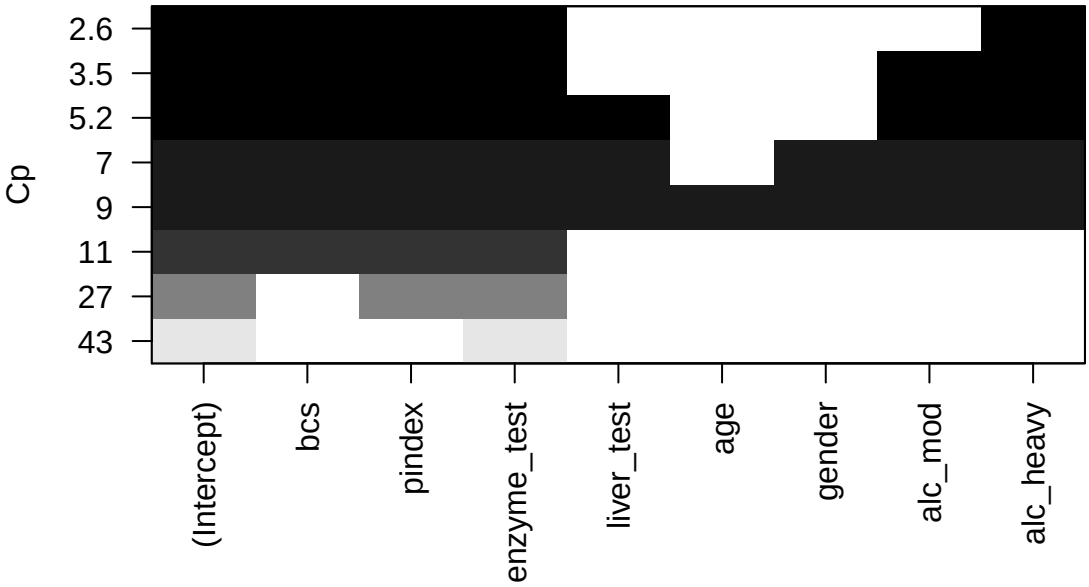
```
points(a1, reg.summary$cp [a1], col = "red", cex = 2, pch = 20)
plot(reg.summary$rss, xlab = "# Variables", ylab = "RSS", type = "l")
```



Al mirar el metodo de Cp y RSS ambos coinciden en que el mejor modelo es aquel con 4 variables. ¿Serán estos los mismos modelos?

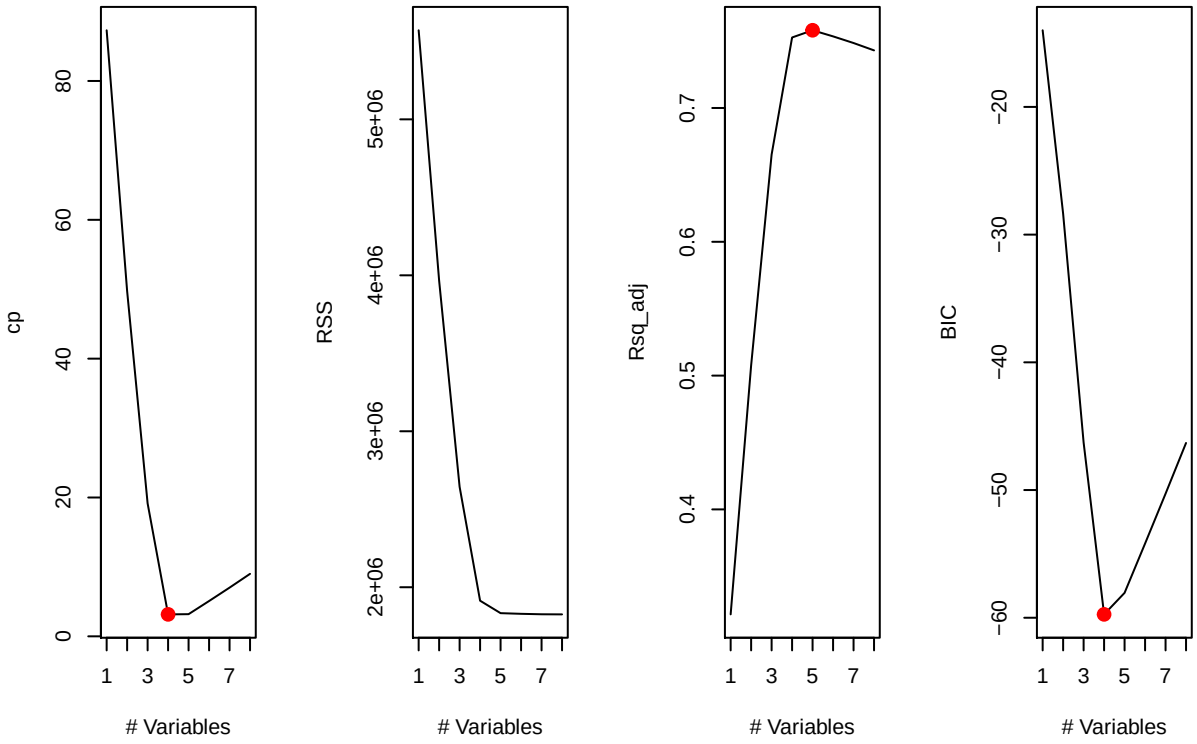
Nuevamente, analizando graficamente.

```
plot(reg_fitted_cp, scale = "Cp")
```

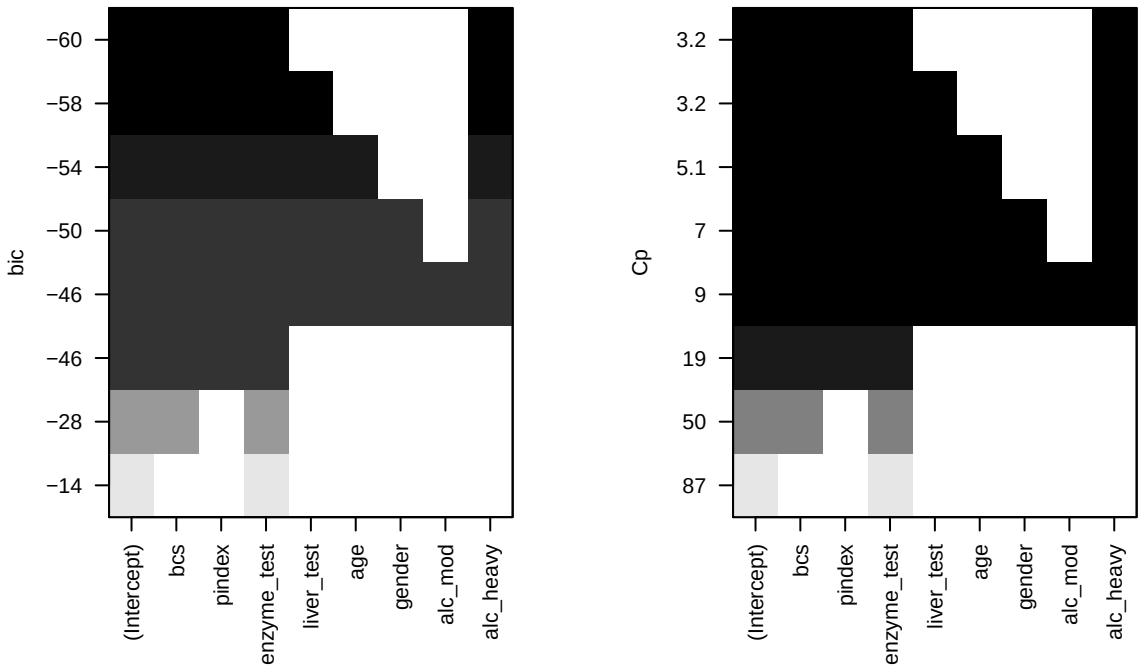


Se muestra solo un metodo mas seleccionado y se observa que, dada una seleccion forward por el criterio de Cp de Mallows, los metodos son muy parecidos entre si pues selecciona practicamente las mismas variables pero, ¿Y si se parte de una seleccion hacia atras? ¿Es igual la seleccion?

```
regfit.back = regsubsets(y~.,data=surgical ,nvmax =8, method = "backward")
reg.back.summary = summary(regfit.back)
par(mfrow = c(1,4))
plot(reg.back.summary$cp, xlab = "# Variables", ylab = "cp", type = "l")
a1 = which.min(reg.back.summary$cp)
points(a1, reg.back.summary$cp [a1], col = "red", cex = 2, pch = 20)
plot(reg.back.summary$rss, xlab = "# Variables", ylab = "RSS", type = "l")
plot(reg.back.summary$adjr2, xlab = "# Variables", ylab = "Rsq_adj", type = "l")
a4 = which.max(reg.back.summary$adjr2)
points(a4, reg.back.summary$adjr2 [a4], col = "red", cex = 2, pch = 20)
plot(reg.back.summary$bic, xlab = "# Variables", ylab = "BIC", cex = 2, pch = 20, type = "l")
a5 = which.min(reg.back.summary$bic)
points(a5, reg.back.summary$bic [a5], col = "red", cex = 2, pch = 20)
```



Al observar rapidamente los metodos de seleccion, todos coinciden en que el modelo mas optimo es de 4 vvariables, por cualquier criterio de los mostrados dado los valores. ¿Será nuevamente los modelos con las mismas variables? Se ilustran solo 2 de los criterios, a modo de ejemplo.



Finalmente, se observa que al aplicar metodo de selccion forward y backward, dado los criterios probados como bic, R^2_{adj}, cp y rss, se observa que los modelos seleccionados son, en terminos generales, el mismo; exeptuando cambios muy puntuales pero que en ultimas, no son cambios significativos.

De esta manera, se concluye que el modelo final será el primero hallado, teniendo en cuenta los comentarios anteriores y una posible verificacion del experto ya que estos modelos casi que convergen al mismo.