

Trabajo3

```
## Loading required package: gplots

## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##      lowess

## Loading required package: raster

## Loading required package: sp

##
## Attaching package: 'raster'

## The following object is masked from 'package:e1071':
##      interpolate

## The following objects are masked from 'package:MASS':
##      area, select

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-5

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

## Loading required package: survival

## Loading required package: splines

##
## Attaching package: 'survival'

## The following object is masked from 'package:boot':
##      aml
```

```

## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:boot':
##
##     melanoma

## Loading required package: parallel

## Loaded gbm 2.1.1

```

Una de las bases de datos con las que vamos a trabajar tiene “más pinta de cuadrática” entonces nos podemos plantear hacerlo del siguiente modo:

modelo3 <- lm(y ~ I(x2^2)+x2) tenemos que poner la palabra reservada I para que interprete correctamente la potencia.

poly() para hacer combinaciones polinómicas.

Puede ocurrir que haya sinergia entre atributos, es decir, que no sean independientes unos de otros, entonces vamos a ver cómo escribimos una fórmula para que el modelo se ajuste como queremos a los datos que le pasamos.

modeloSinergico <- lm(y ~ x1*x2, data = datos) <=> lm(y ~ x1+x2+x1:x2, data = datos)

Despues de aprender un modelo podemos poner names(modelo) y nos dice los elementos que podemos consultar del modelo.

Lo convertimos a factor, asfactor

tune.knn va probando distintos parámetros y te devuelve el que mejor funciona.

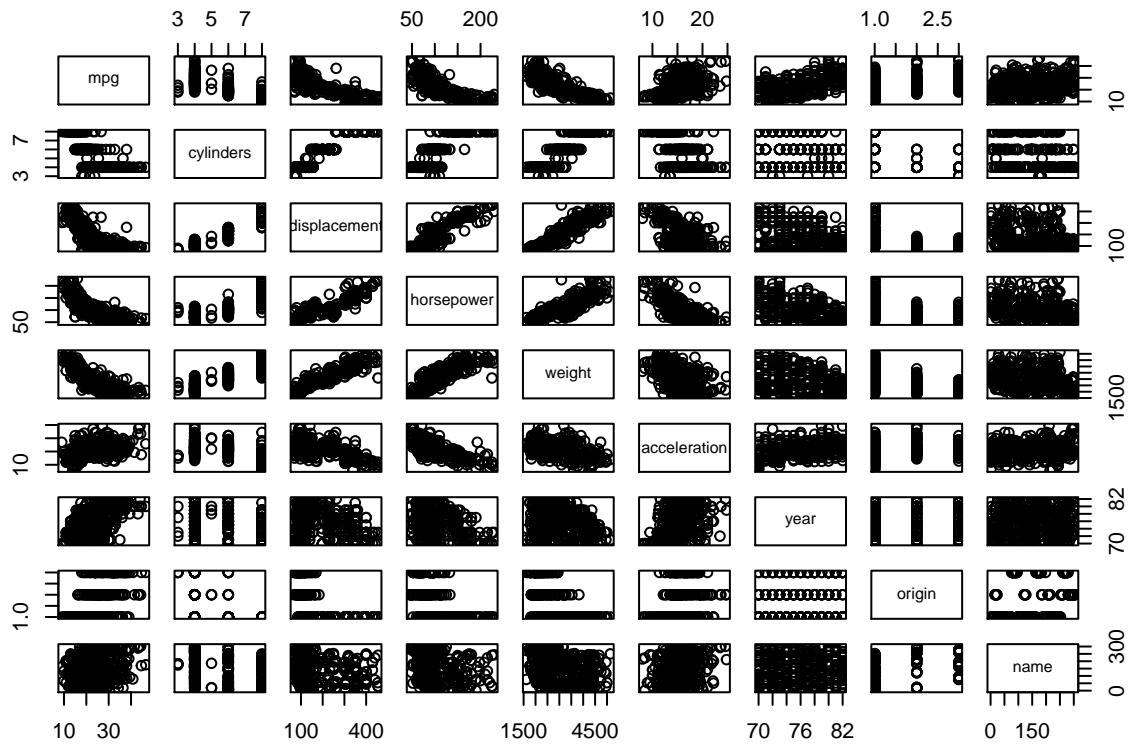
Ejercicio 1

a) Usar las funciones de R pairs() y boxplot() para investigar la dependencia entre mpg y las otras características. ¿Cuáles de las otras características parece más útil para predecir mpg? Justificar la respuesta.

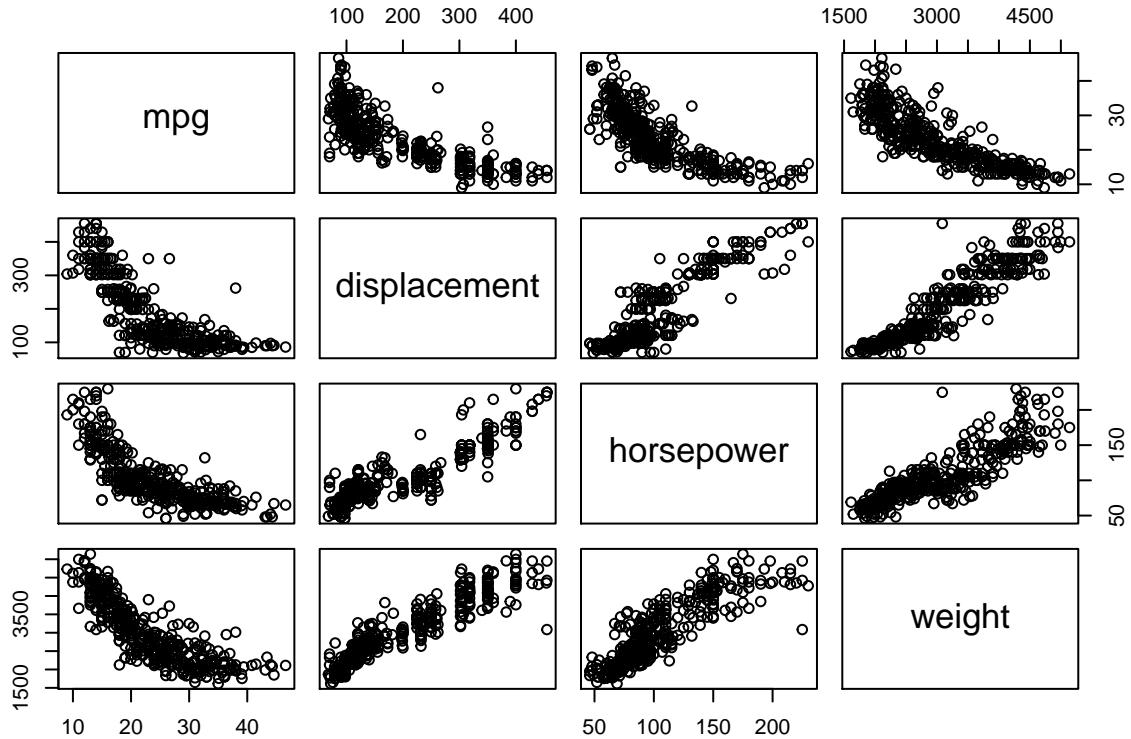
```

set.seed(41192)
attach(Auto)
pairs(Auto)

```

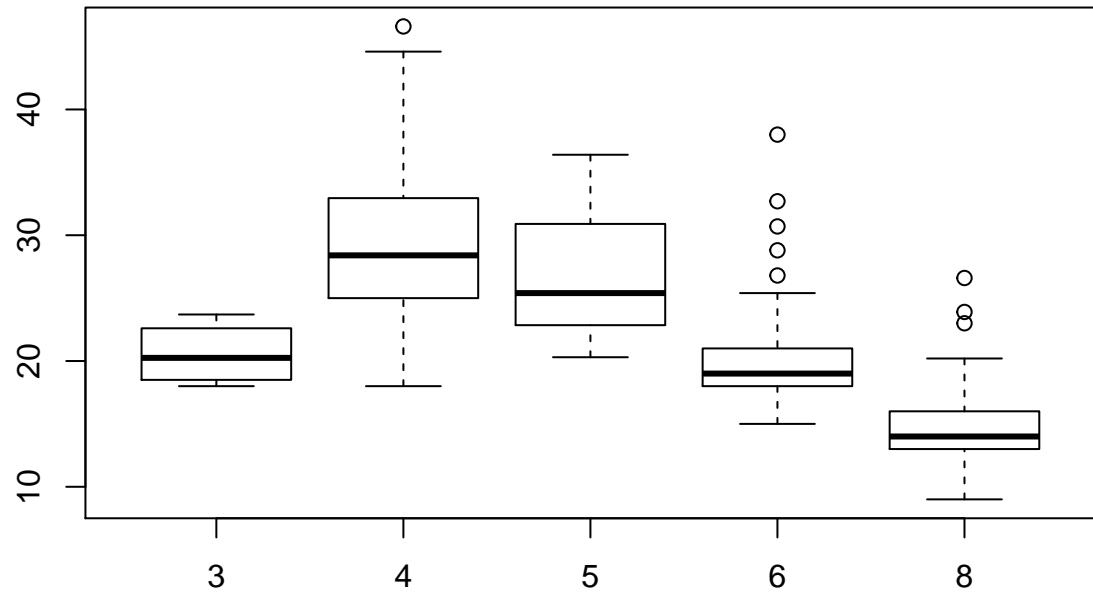


```
pairs(mpg ~ displacement + horsepower + weight)
```

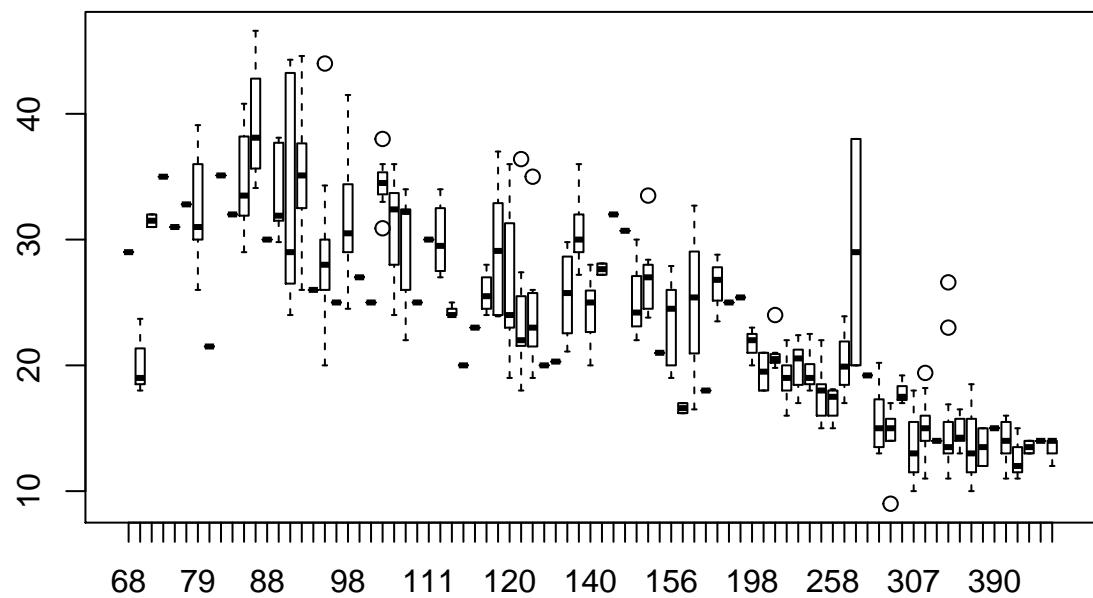


Como podemos ver las “gráficas de dependencias” de mpg con respecto a *displacement*, *horsepower* y *weight* son las gráficas que presentan un patrón más parecido entre ellas indicando que mpg tiene una relación fuerte con estas variables ya que se ajusta a ellas de un modo similar. Por ejemplo si vemos la gráfica con respecto a *acceleration* lo que tenemos es una nube de puntos mucho más dispersa. En cambio estas gráficas si que tienen un aspecto de ser ajustables linealmente.

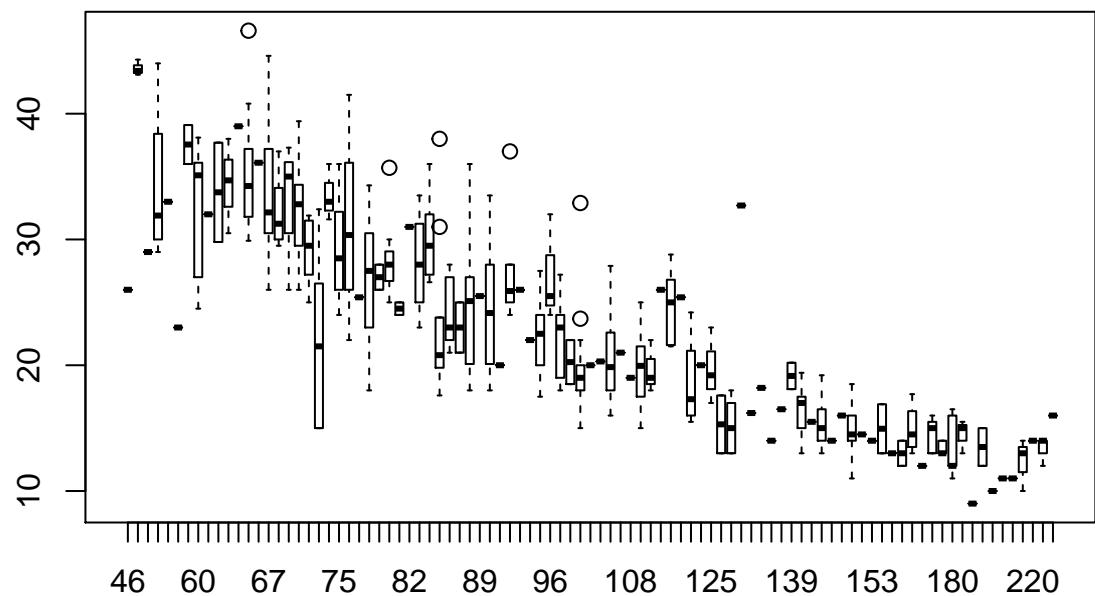
```
boxplot(mpg ~ cylinders)
```



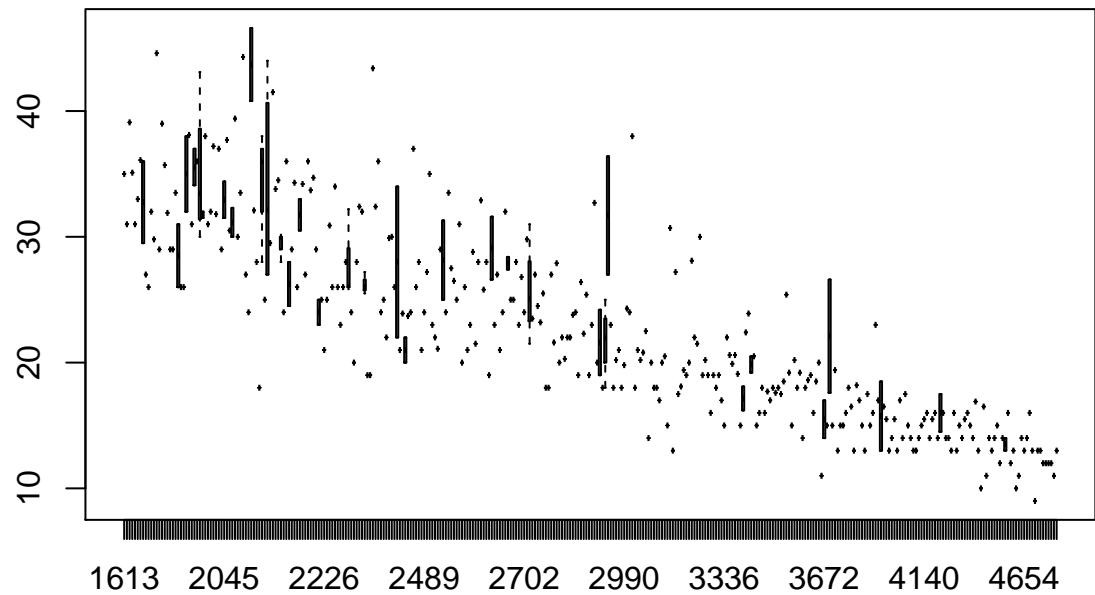
```
boxplot(mpg ~ displacement)
```



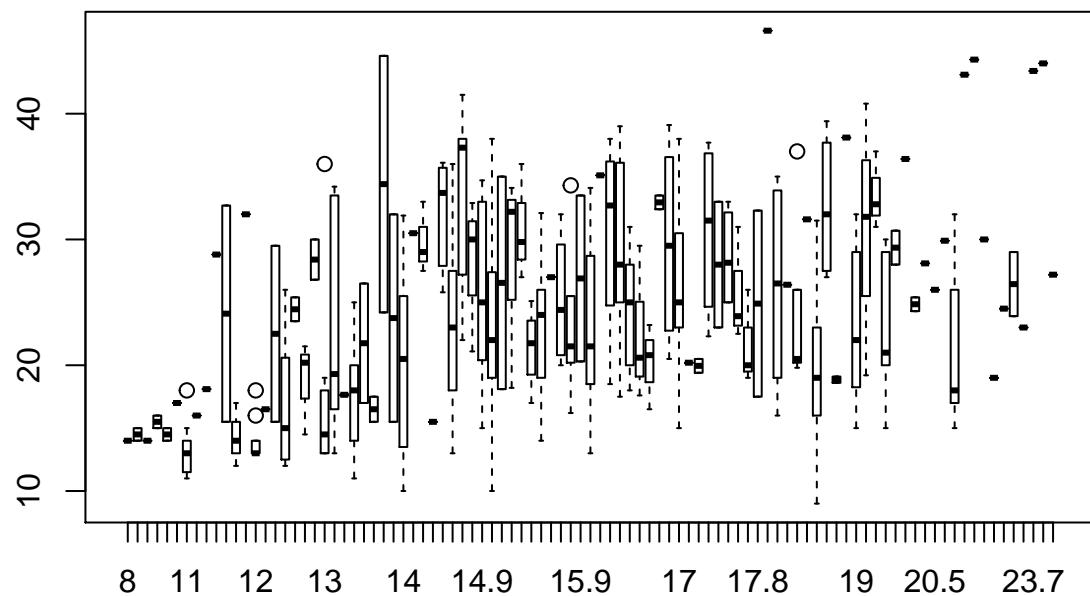
```
boxplot(mpg ~ horsepower)
```



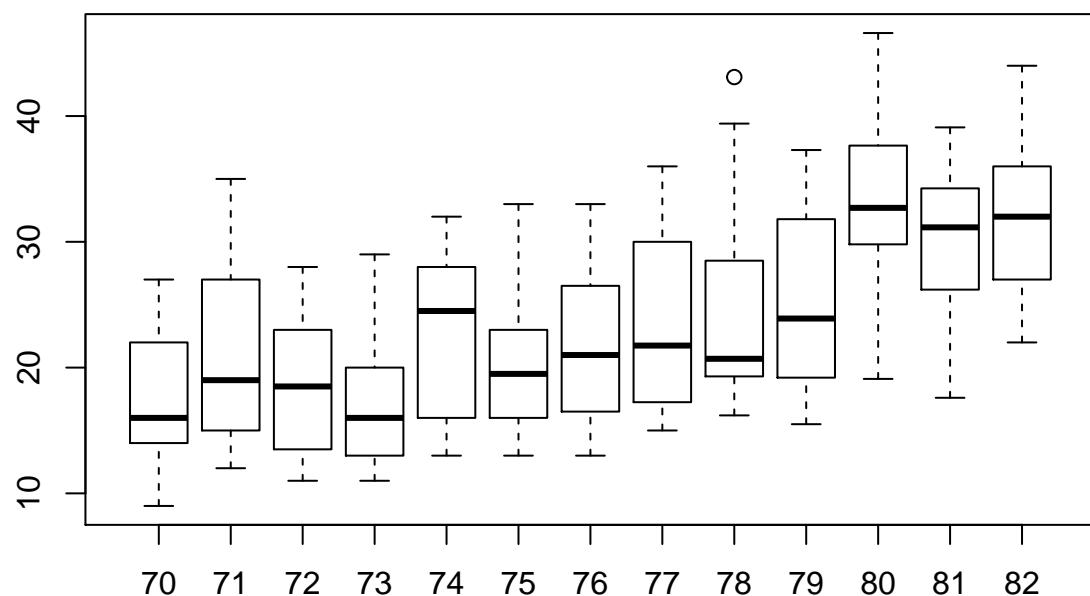
```
boxplot(mpg ~ weight)
```



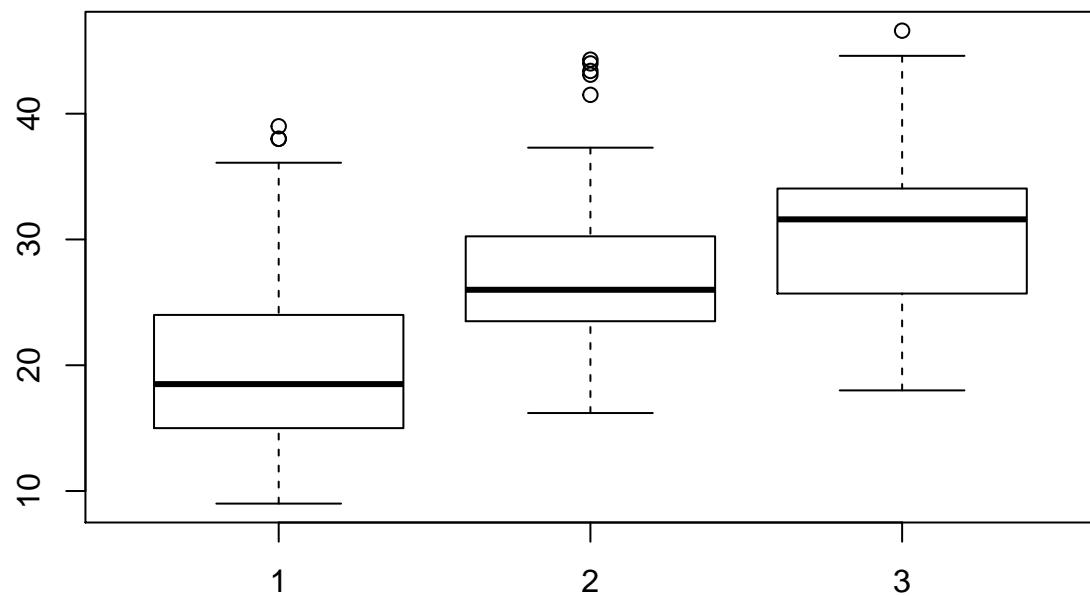
```
boxplot(mpg ~ acceleration)
```



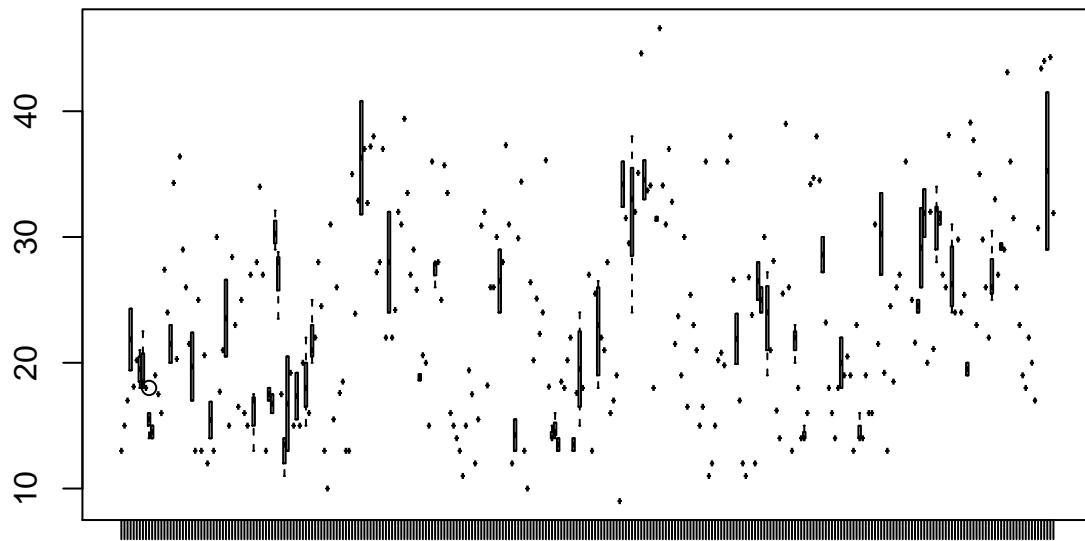
```
boxplot(mpg ~ year)
```



```
boxplot(mpg ~ origin)
```



```
boxplot(mpg ~ name)
```



amc ambassador brougham dodge colt ford torino plymouth valiant vw rabbit

b) Seleccionar las variables predictoras que considere más relevantes.

Vamos a seleccionar aquellas que hemos visto en el apartado anterior que parecen seguir un patrón similar:

```
datos = Auto[, c("mpg", "displacement", "horsepower", "weight")]
```

c) Partitionar el conjunto de datos en un conjunto de entrenamiento (80%) y otro de test (20%). Justificar el procedimiento usado.

A priori había pensado en calcular en primer lugar la variable mpg1 del siguiente apartado para así poder realizar un particionamiento más homogéneo de las etiquetas positivas y negativas en los conjuntos de

entrenamiento y test. El problema es que para hacer tal cosa tendría que calcular la mediana de todos los datos de Auto y lo que queremos es calcular la mediana, y por tanto el valor de mpg1 sólo en base a los datos de entrenamiento puesto que se dijo en teoría que para no contaminar el aprendizaje no podíamos usar los datos de test para calcular una mediana, sería como mirar los datos antes de aprender. Entonces he realizado simplemente un submuestreo de los datos aleatorio para dividirlos en entrenamiento y test.

```
n = nrow(datos)
idx_train = sample(seq(n), ceiling(0.8*n))

datos.train = datos[idx_train,]
datos.test = datos[-idx_train,]
```

d) Crear una variable binaria, mpg01, que será igual 1 si la variable mpg contiene un valor por encima de la mediana, y -1 si mpg contiene un valor por debajo de la mediana. La mediana se puede calcular usando la función median(). (Nota: puede resultar útil usar la función data.frames() para unir en un mismo conjunto de datos la nueva variable mpg01 y las otras variables Auto).

```
mediana = median(datos.train$mpg)
mpg1.train = sapply(datos.train$mpg, function(x) if (x < mediana) return(0) else return(1))
mpg1.test = sapply(datos.test$mpg, function(x) if (x < mediana) return(0) else return(1))
```

- Ajustar un modelo de regresión logística a los datos de entrenamiento y predecir mpg01 usando las variables seleccionadas en b). ¿Cuál es el error de test del modelo? Justificar la respuesta.

```
trainRL = data.frame(mpg01 = mpg1.train, datos.train)
testRL = data.frame(mpg01 = mpg1.test, datos.test)

RL = glm(mpg01 ~ displacement+horsepower+weight, data = trainRL, family = binomial)

prediccion = predict(RL, newdata = testRL, type = "response")
RL.pred = rep(0, length(testRL$mpg01))
RL.pred[prediccion > .5] = 1

cat("El % de errores en test con RL es: ", sum(testRL$mpg01 != RL.pred)/length(testRL$mpg01)*100)

## El % de errores en test con RL es: 15.38462
```

- Ajustar un modelo K-NN a los datos de entrenamiento y predecir mpg01 usando solamente las variables seleccionadas en b). ¿Cuál es el error de test del modelo? ¿Cuál es el valor de K que mejor ajusta los datos?

```
getErrorKNN <- function(datos.train, datos.test, et.train, et.test ){
  #normalizamos los datos
  train.norm = scale(datos.train[,c("weight","displacement","horsepower")])
  medias = attr(train.norm, "scaled:center")
  escalados = attr(train.norm, "scaled:scale")
  test.norm = scale(datos.test[,c("weight","displacement","horsepower")], medias, escalados)
```

```

datos.full = rbind(train.norm, test.norm)
et.full = as.factor(c(et.train, et.test))

set.seed(75570417)
mknns = tune.knn(datos.full, et.full, k=1:20, tunecontrol = tune.control(sampling = "cross"), cross = TRUE)

mejor_k = mknns$best.model$k

KNN = knn(train.norm, test.norm, et.train, k = mejor_k, prob = TRUE)
cat("Se ha elegido como mejor k el: ", mejor_k, "\n")
err.test = sum(et.test != KNN)/length(et.test)*100
cat("El % de errores que obtenemos en el test es: ", err.test, "\n")

KNN
}

KNN = getErrorKNN(trainRL, testRL, mpg1.train, mpg1.test)

## Se ha elegido como mejor k el: 7
## El % de errores que obtenemos en el test es: 16.66667

• Pintar las curvas ROC (instalar paquete ROCR en R) y comparar y valorar los resultados obtenidos por ambos modelos.

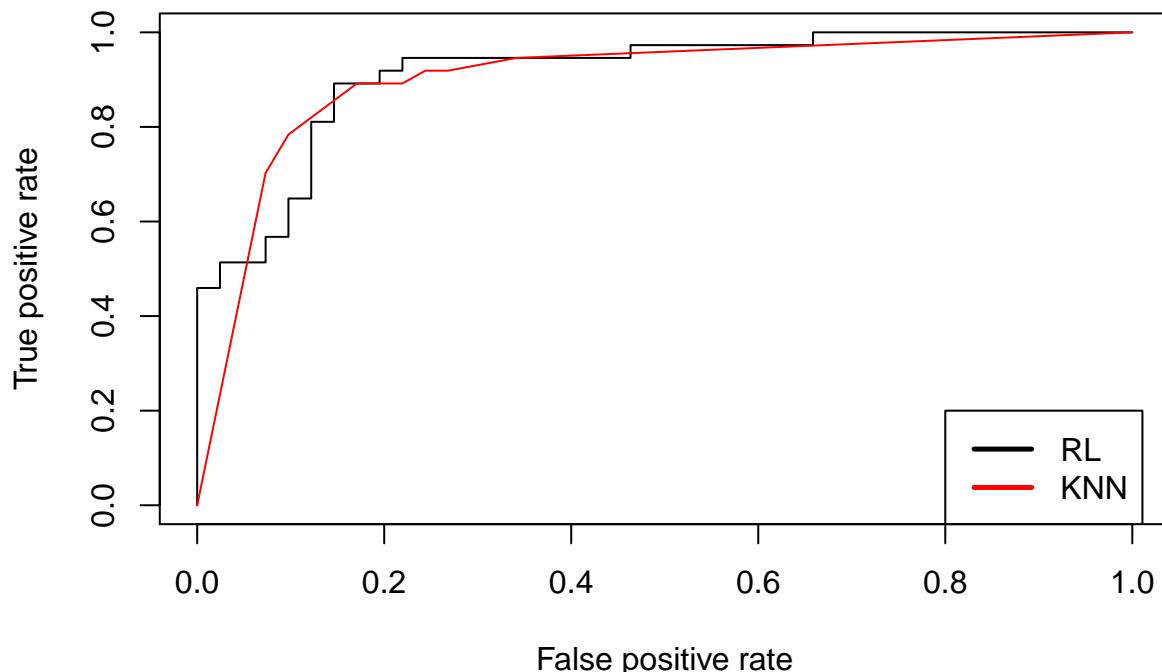
pRL = prediction(prediccion, testRL$mpg01)
perf = performance(pRL, "tpr", "fpr")
plot(perf, main = "Curvas ROC")

prob = attr(KNN, "prob")
prob = ifelse(KNN == "0", 1-prob, prob)
pKNN = prediction(prob, testRL$mpg01)
perf = performance(pKNN, "tpr", "fpr")

plot(perf, add = TRUE, col = "red")
legend(0.8, 0.2, c("RL", "KNN"), lty=c(1,1), lwd=c(2.5,2.5), col=c("black", "red"))

```

Curvas ROC



Al ini-

cio RL parece mas estable puesto que pasa más tiempo teniendo un tasa nula de falso positivos, ahora bien cuando esta tasa aumenta vemos como antes la misma pérdida o el mismo error de clasificación es el KNN el que acierta un mayor número de veces, es decir, que aunque cometemos el mismo error al menos clasificamos bien más muestras.

Cuando la tasa de falsos positivos aumenta es nuevamente la RL la que mejor se comporta puesto que tiene una tasa de verdaderos positivos más elevada aunque no difiere en gran medida de la del KNN. Con lo cual en mi opinión es el KNN el que tiene un mejor comportamiento puesto que es el que da mejores resultados ante un error igual.

e) Estimar el error de test de ambos modelos pero usando Validación Cruzada de 5-particiones. Comparar con los resultados obtenidos en el punto anterior.

```

fullRL = rbind(trainRL, testRL)
RL = glm(mpg01 ~ displacement+horsepower+weight, data = fullRL, family = binomial)
pp = cv.glm(fullRL, RL, K = 5)

cat("El error de test obtenido haciendo VC en RL ha sido: ", pp$delta[1]*100, "\n")

## El error de test obtenido haciendo VC en RL ha sido: 8.477098

'
folds = kfold(full.data, k = 5)
err = 0

for (i in seq(5)) {
  getErrorKNN(full.data[folds != i], full.data[folds == i], et.train, et.test)
  KNN = knn(full.data[folds != i], full.data[folds == i], full.labels[folds != i], k = mejor_k, prob =

```

```

    err = err + sum(full.labels[folds == i] != KNN)/length(full.labels[folds == i])*100
}

cat("El error de test obtenido haciendo VC en KNN ha sido: ", err/5)
'

## [1] "\nfolds = kfold(full.data, k = 5)\nerr = 0\n\nfor (i in seq(5)) {\n  getErrorKNN(full.data[fold

```

Ejercicio 2

Usar la base de datos Boston (en el paquete MASS de R) para ajustar un modelo que prediga si dado un suburbio este tiene una tasa de criminalidad (crim) por encima o por debajo de la mediana. Para ello considere la variable crim como la variable salida y el resto como variables predictoras.

- a) Encontrar el subconjunto óptimo de variables predictoras a partir de un modelo de regresión-LASSO (usar paquete glmnet de R) donde seleccionamos sólo aquellas variables con coeficiente mayor a un umbral prefijado.
- b) Ajustar un modelo de regresión regularizada con “weight-decay” (ridge-regression) y las variables seleccionadas. Estimar el error residual del modelo y discutir si el comportamiento de los residuos muestran algún indicio de “underfitting”.

Lo que vamos a hacer en primer lugar es emplear validación cruzada para obtener el mejor s para nuestros datos, este s lo que hace es regular el impacto de la restricción de “contracción”, la cual lleva a coeficientes más pequeños cuanto mayor sea su impacto, es decir, el valor de s .

Con el objetivo de tener un buen s lo que vamos a hacer es obtenerlo mediante validación cruzada en la cual emplearemos todos los datos, ya vimos en teoría que con validación cruzada si empleamos todos los datos podemos tener un buen estimador sin riesgo de que usar los datos de test suponga un problema de sobreajuste.

```

attach(Boston)

set.seed(41192)
cv.out = cv.glmnet(as.matrix(Boston[,-1]), Boston[,1], alpha=1)
bestlam = cv.out$lambda.min #el mejor ese obtenido

```

Ahora aplicamos un modelo lasso a nuestros datos empleando el s que hemos obtenido, esto nos dará una serie de coeficientes para las distintas características de la muestra, aquellos que, en valor absoluto, estén por encima de un cierto umbral (que iremos modificando hasta obtener un buen resultado) serán los de las variables que seleccionemos para el siguiente apartado.

```

out = glmnet(as.matrix(Boston[,-1]),Boston[,1], alpha = 1)
lasso.coef = predict(out, type="coefficients", s=bestlam)

umbral = 0.1
var_seleccionadas = which(abs(lasso.coef) > umbral)[-1]

```

```

#generamos un submuestreo aleatorio 80-20
set.seed(41192)
train = sample(1:nrow(Boston), nrow(Boston)*0.8)
test = (-train)

ridge.mod = glmnet(as.matrix(Boston[train, var_seleccionadas]), Boston[train,1], alpha = 0)
ridge.pred = predict(ridge.mod, s=bestlam,newx = as.matrix(Boston[test, var_seleccionadas]))

error = sqrt(mean((Boston[test,1] - ridge.pred)^2)/(nrow(Boston[test,])-2))
print(error)

## [1] 0.3129668

```

El error que calculamos es el RSE que es la raíz cuadrada del error cuadrático medio, a continuación mostramos una tabla con los distintos parámetros que hemos probado. Los parámetros que hemos ido modificando han sido el umbral que marca la selección de las características y el s para el weight decay, ya que aunque tenemos uno óptimo para el Lasso no tiene por qué ser el mismo para esta nueva regresión, aunque veremos que sí. En la siguiente tabla recogemos los parámetros junto con el RSE obtenido con ellos (para la partición 80-20 que hacemos de los datos con la semilla seleccionada):

umbral	s	RSE
0.1	bestlam	0.3129668
0.1	100	0.4107367
0.1	200000bestlam	0.4997851
0	bestlam	0.3214119
0.45	bestlam	0.47722858

Como vemos el bestlam es el que mejores resultados da con el mismo umbral, también hemos observado que si dividimos el bestlam por prácticamente cualquier número el error no varía, no obstante y dado que veo que lo que marca el s es el peso de una condición que podríamos llamar de regularización opto por dejarlo lo “más grande posible”, para evitar en la medida de lo posible sobreajuste. Cuando amplio demasiado el s, dándole mucho peso a la regularización entonces el algoritmo se olvida de ajustar los datos y da peores errores.

Por otro lado si para el umbral hacemos que se consideren todas las variables el error empeora aunque no demasiado, indicando que no hay tanto ruido que afecte al algoritmo como podríamos pensar. En cambio lo que sí que empeora más el resultado es ser demasiado elitistas con la selección de variables ya que según vemos estamos perdiendo información para un buen ajuste.

Por tanto en mi opinión observamos underfitting cuando le damos demasiado peso a la condición de regularización, cuando el s es demasiado grande, para los otros parámetros el error fuera de la muestra es relativamente bueno.

c) Definir una nueva variable con valores -1 y 1 usando el valor de la mediana de la variable crim como umbral. Ajustar un modelo SVM que prediga la nueva variable definida (usar el paquete e1071 de R). Describir con detalle cada uno de los pasos dados en el aprendizaje del modelo SVM. Comience ajustando un modelo lineal y argumente si considera necesario usar algún núcleo. Valorar el resultado del uso de distintos núcleos.

Anteriormente ya hemos generado las muestras de entrenamiento y test, ahora vamos a tomar la mediana de los datos de entrenamiento (ya hemos dicho por qué esto se hace así) y luego a generar la variable que

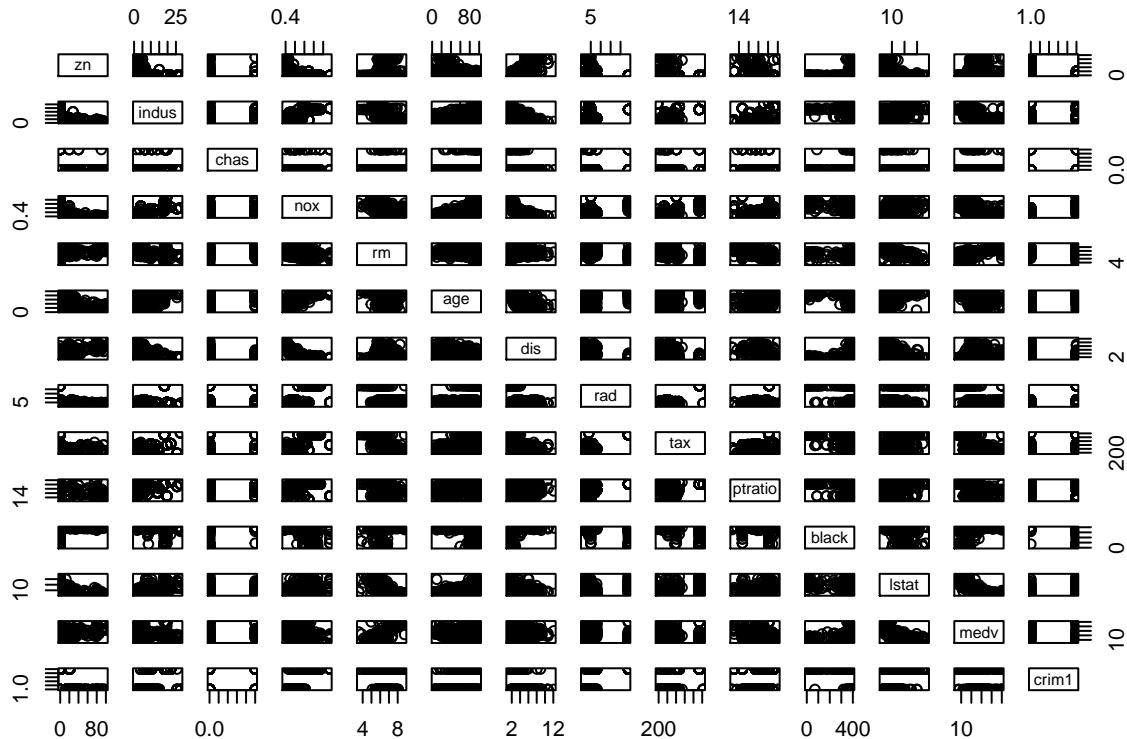
llamaremos, *crim1* a partir de ella:

```
mediana = median(Boston[train,1])
crim1.train = sapply(Boston[train,1], function(x) if (x < mediana) return(-1) else return(1))
crim1.test = sapply(Boston[test,1], function(x) if (x < mediana) return(-1) else return(1))

#generamos un dataframe con estos datos + la nueva var.
data.train = data.frame(Boston[train,-1], crim1 = as.factor(crim1.train))
data.test = data.frame(Boston[test, -1], crim1 = crim1.test)
```

A continuación vamos a aplicar un modelo de support vector machine a los datos de entrenamiento, a priori el kernel que vamos a usar va a ser lineal. El principal problema que tenemos es que nuestros datos no tienen solo las propiedades con las cuales el hecho de hacer un plot de la clasificación dada no nos dirá qué tipo de núcleo obtener. Lo que vamos a hacer es usar la función *pair* para ver si encontramos algún tipo de patrón:

```
data.full = rbind(data.train, data.test)
pairs(data.full)
```



Dado que es una clasificación binaria da la sensación de que el clasificador lineal va a ir, así que lo que vamos a hacer es entrenar un clasificador svm con cada uno de los núcleos disponibles y ver qué error fuera de la muestra obtenemos: s

```
svmfitL = svm(crim1~., data=data.train[,-1], kernel = "linear")
svm.pred = predict(svmfitL, newdata = data.test[,-1])
cat("Obtenemos un error de (lineal):", sum(svm.pred != data.test$crim1))

## Obtenemos un error de (lineal): 12
```

```



```

Por lo tanto el que mejor resultados da es el clasificador con un núcleo lineal como sospechábamos.

Como vemos consideramos todas las variables excepto la variable crim ya que esta es la que queremos predecir, con lo que estaríamos contaminando el aprendizaje. Indicar que como comentó la profesora no tenemos que usar solamente aquellas variables seleccionadas en apartados anteriores del ejercicio, sino que consideraremos todas. Esto se debe a que según qué modelo de aprendizaje estemos empleando unas variables tendrán más importancia que otras; cada modelo tiene sus criterios y “preferencias”, entonces a priori, sin un análisis más exhaustivo de los datos, no podemos descartar ninguna de las características.

Ejercicio 3

Usar el conjunto de datos Boston y las librerías randomForest y gbm de R.

1. Dividir la base de datos en dos conjuntos de entrenamiento (80%) y test (20%).

```
attach(Boston)

## The following objects are masked from Boston (position 3):
##   age, black, chas, crim, dis, indus, lstat, medv, nox, ptratio,
##   rad, rm, tax, zn

train = sample(nrow(Boston), nrow(Boston)*0.8)
test = -train
```

2. Usando la variable medv como salida y el resto como predictores, ajustar un modelo de regresión usando bagging. Explicar cada uno de los parámetros usados. Calcular el error de test.

```
set.seed(41192)
bag = randomForest(medv ~., data = Boston, subset = train, mtry = ncol(Boston)-1, importance = TRUE)
pred = predict(bag, newdata = Boston[test,])
cat("El MSE con bagging es: ", mean((pred - Boston[test,]$medv)^2), "\n")

## El MSE con bagging es:  9.038211
```

3. Ajustar un modelo de regresión usando Random Forest. Obtener una estimación del número de árbol necesario. Justificar el resto de parámetros usados en el ajuste. Calcular el error de test y compararlo con el obtenido con bagging.

```
randomfcv = rfcv(Boston[,1:13], Boston$medv)
#uso 6 porque es el que menos MSE da quitando el 13 que seria hacer el bagging
rf = randomForest(medv ~., data = Boston, subset = train, mtry = 6, importance = TRUE)
```

4. Ajustar un modelo de regresión usando Boosting (usar gbm con distribution = ‘gaussian’). Calcular el error de test y compararlo con el obtenido con bagging y Random Forest.

Ejercicio 4

Usar el conjunto de datos OJ que es parte del paquete ISLR.

1. Crear un conjunto de entrenamiento conteniendo una muestra aleatoria de 800 observaciones, y un conjunto de test conteniendo el resto de observaciones. Ajustar un árbol a los datos de entrenamiento, con “Purchase” como la variable respuesta y las otras variables como predictores (paquete tree de R).
2. Usar la función summary() para generar un resumen estadístico acerca del árbol y describir los resultados obtenidos: tasa de error de “training”, número de nodos del árbol, etc.
3. Crear un dibujo del árbol e interpretar los resultados.
4. Predecir la respuesta de los datos de test, y generar e interpretar la matriz de confusión de los datos de test. ¿Cuál es la tasa de error de test? ¿Cuál es la precisión del test?
5. Aplicar la función cv.tree() al conjunto de “training” y determinar el tamaño óptimo del árbol. ¿Qué hace cv.tree?

para el vaging mtray = ncol-1 de los datos que tenemos (es un random forest)