sobreajuste

Cedric Prieels

16/11/2016

```
library(car)
library(boot)
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
library(ISLR)
```

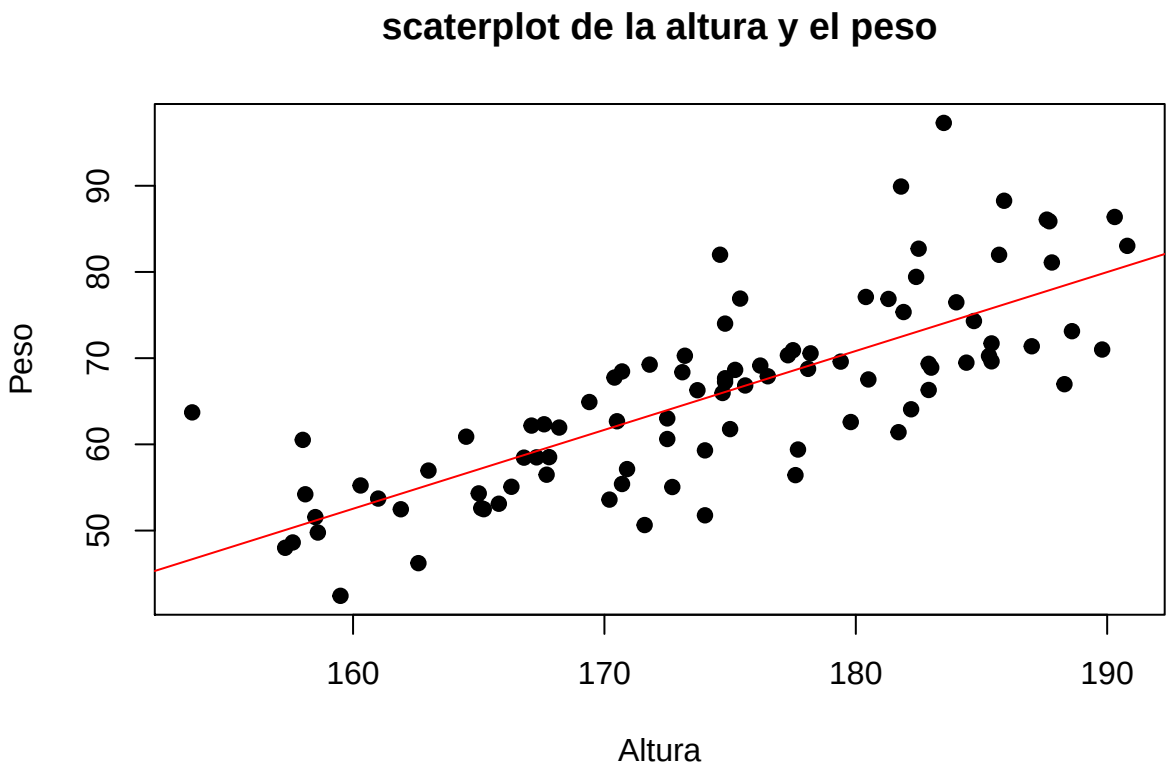
```
rm(list=ls())
setwd('/Users/ced2718/Documents/Universite/Modelizacion')
load('Pulsaciones.rda')
attach(Pulsaciones)
Pulsaciones
```

```
##    Pulse1 Pulse2    Correr   Fumar   Sexo Altura  Peso Actividad
## 1      64     88    corrio no fuma hombre  168.2 61.95     media
## 2      58     70    corrio no fuma hombre  182.9 66.31     media
## 3      62     76    corrio    fuma hombre  187.0 71.38      alta
## 4      66     78    corrio    fuma hombre  185.9 88.26      baja
## 5      64     80    corrio no fuma hombre  175.2 68.64     media
## 6      74     84    corrio no fuma hombre  184.7 74.31      baja
## 7      84     84    corrio no fuma hombre  183.0 68.90      alta
## 8      68     72    corrio no fuma hombre  187.7 85.88     media
## 9      62     75    corrio no fuma hombre  181.8 89.91     media
## 10     76    118    corrio no fuma hombre  181.7 61.41     media
## 11     90     94    corrio    fuma hombre  188.6 73.13      baja
## 12     80     96    corrio no fuma hombre  182.9 69.34     media
## 13     92     84    corrio    fuma hombre  177.3 70.34      alta
## 14     68     76    corrio no fuma hombre  169.4 64.91     media
## 15     60     76    corrio no fuma hombre  180.4 77.10      alta
## 16     62     58    corrio no fuma hombre  182.4 79.43      alta
## 17     66     82    corrio    fuma hombre  174.6 82.00     media
## 18     70     72    corrio    fuma hombre  184.0 76.48      alta
## 19     68     76    corrio    fuma hombre  187.8 81.10     media
## 20     72     80    corrio no fuma hombre  167.1 62.18      alta
## 21     70    106    corrio no fuma hombre  181.3 76.87     media
## 22     74     76    corrio no fuma hombre  178.2 70.56     media
## 23     66    102    corrio no fuma hombre  177.7 59.40     media
## 24     70     94    corrio    fuma hombre  190.8 83.03     media
## 25     96    140    corrio no fuma  mujer  153.6 63.71     media
## 26     62    100    corrio no fuma  mujer  165.8 53.11     media
```

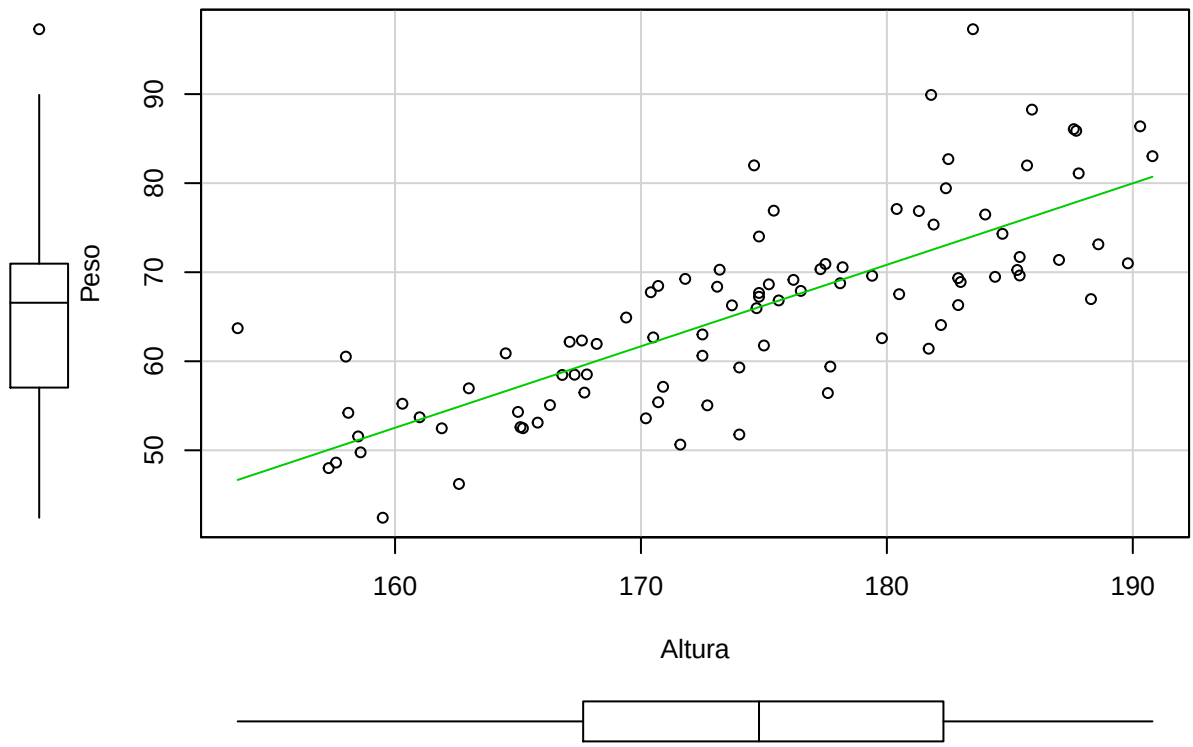
```
## 27     78    104    corrio    fuma  mujer  174.0 59.30     media
## 28     82    100    corrio no fuma  mujer  172.5 63.01     media
## 29    100    115    corrio    fuma  mujer  160.3 55.23     media
## 30     68    112    corrio no fuma  mujer  177.6 56.43     media
## 31     96    116    corrio no fuma  mujer  174.0 51.77     media
## 32     78    118    corrio no fuma  mujer  174.7 65.96     media
## 33     88    110    corrio    fuma  mujer  174.8 67.24     media
## 34     62     98    corrio    fuma  mujer  158.5 51.56     media
## 35     80    128    corrio no fuma  mujer  172.7 55.05     media
## 36     62     62 no corrio no fuma hombre  187.6 86.07      baja
## 37     60     62 no corrio no fuma hombre  179.4 69.60     media
## 38     72     74 no corrio    fuma hombre  175.4 76.91     media
## 39     62     66 no corrio no fuma hombre  177.5 70.92     media
## 40     76     76 no corrio no fuma hombre  183.5 97.29     media
## 41     68     66 no corrio    fuma hombre  170.4 67.75     media
## 42     54     56 no corrio    fuma hombre  175.6 66.83     media
## 43     74     70 no corrio no fuma hombre  185.4 71.71      alta
## 44     74     74 no corrio no fuma hombre  185.3 70.25     media
## 45     68     68 no corrio no fuma hombre  180.5 67.53      alta
## 46     72     74 no corrio    fuma hombre  171.8 69.25      alta
## 47     68     64 no corrio no fuma hombre  176.5 67.90      alta
## 48     82     84 no corrio    fuma hombre  185.7 81.99     media
## 49     64     62 no corrio no fuma hombre  189.8 71.00      alta
## 50     58     58 no corrio no fuma hombre  167.6 62.33      alta
## 51     54     50 no corrio no fuma hombre  174.8 74.01     media
## 52     70     62 no corrio    fuma hombre  166.8 58.46     media
## 53     62     68 no corrio    fuma hombre  184.4 69.48     media
## 54     48     54 no corrio    fuma hombre  173.1 68.37      alta
## 55     76     76 no corrio no fuma hombre  188.3 66.98      alta
## 56     88     84 no corrio no fuma hombre  185.4 69.63     media
## 57     70     70 no corrio no fuma hombre  178.1 68.76     media
## 58     90     88 no corrio    fuma hombre  170.5 62.68     media
## 59     78     76 no corrio no fuma hombre  182.5 82.70      alta
## 60     70     66 no corrio    fuma hombre  190.3 86.38     media
## 61     90     90 no corrio no fuma hombre  173.7 66.28      baja
## 62     92     94 no corrio    fuma hombre  174.8 67.67     media
## 63     60     70 no corrio    fuma hombre  181.9 75.35     media
## 64     72     70 no corrio no fuma hombre  179.8 62.59     media
## 65     68     68 no corrio no fuma hombre  182.2 64.07      alta
## 66     84     84 no corrio no fuma hombre  175.0 61.77     media
## 67     74     76 no corrio no fuma hombre  170.7 55.40     media
## 68     68     66 no corrio no fuma hombre  173.2 70.28     media
## 69     84     84 no corrio no fuma  mujer  167.8 58.52     media
## 70     61     70 no corrio no fuma  mujer  166.3 55.08     media
## 71     64     60 no corrio no fuma  mujer  167.3 58.49      alta
## 72     94     92 no corrio    fuma  mujer  158.0 60.52     media
## 73     60     66 no corrio no fuma  mujer  158.1 54.21     media
## 74     72     70 no corrio no fuma  mujer  161.0 53.71     media
## 75     58     56 no corrio no fuma  mujer  170.9 57.13     media
## 76     88     74 no corrio    fuma  mujer  164.5 60.89     media
## 77     66     72 no corrio no fuma  mujer  167.7 56.48     media
## 78     84     80 no corrio no fuma  mujer  165.2 52.48      baja
## 79     62     66 no corrio no fuma  mujer  165.0 54.31      alta
## 80     66     76 no corrio no fuma  mujer  165.1 52.60     media
```

```
## 81     80     74 no corrio no fuma  mujer  162.6 46.22     media
## 82     78     78 no corrio no fuma  mujer  170.2 53.59     media
## 83     68     68 no corrio no fuma  mujer  176.2 69.14     media
## 84     72     68 no corrio no fuma  mujer  171.6 50.64     media
## 85     82     80 no corrio no fuma  mujer  161.9 52.47      baja
## 86     76     76 no corrio    fuma  mujer  157.6 48.62      alta
## 87     87     84 no corrio no fuma  mujer  159.5 42.43      alta
## 88     90     92 no corrio    fuma  mujer  163.0 56.96      baja
## 89     78     80 no corrio no fuma  mujer  172.5 60.62      baja
## 90     68     68 no corrio no fuma  mujer  158.6 49.77     media
## 91     86     84 no corrio no fuma  mujer  170.7 68.46      alta
## 92     76     76 no corrio no fuma  mujer  157.3 48.00     media
```

```
plot(Altura, Peso, main='scaterplot de la altura y el peso', pch=19)
abline(lm(Peso~Altura), col="red")
```




```
scatterplot(Peso~Altura, smooth=FALSE)
```




```
# Salida de la regresion
Reg.1 <- lm(Peso~Altura) # regression model (y~x)
summary(Reg.1) #R proxima a 1 dice que el modelo es capaz de representar es capaz de representar la var
```

```
##
## Call:
## lm(formula = Peso ~ Altura)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5727  -4.7486  -0.0001   3.5080  23.2533
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -93.89492   13.88816  -6.761 1.33e-09 ***
## Altura        0.91516    0.07947  11.515  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.981 on 90 degrees of freedom
## Multiple R-squared:  0.5957, Adjusted R-squared:  0.5912
## F-statistic: 132.6 on 1 and 90 DF,  p-value: < 2.2e-16
```

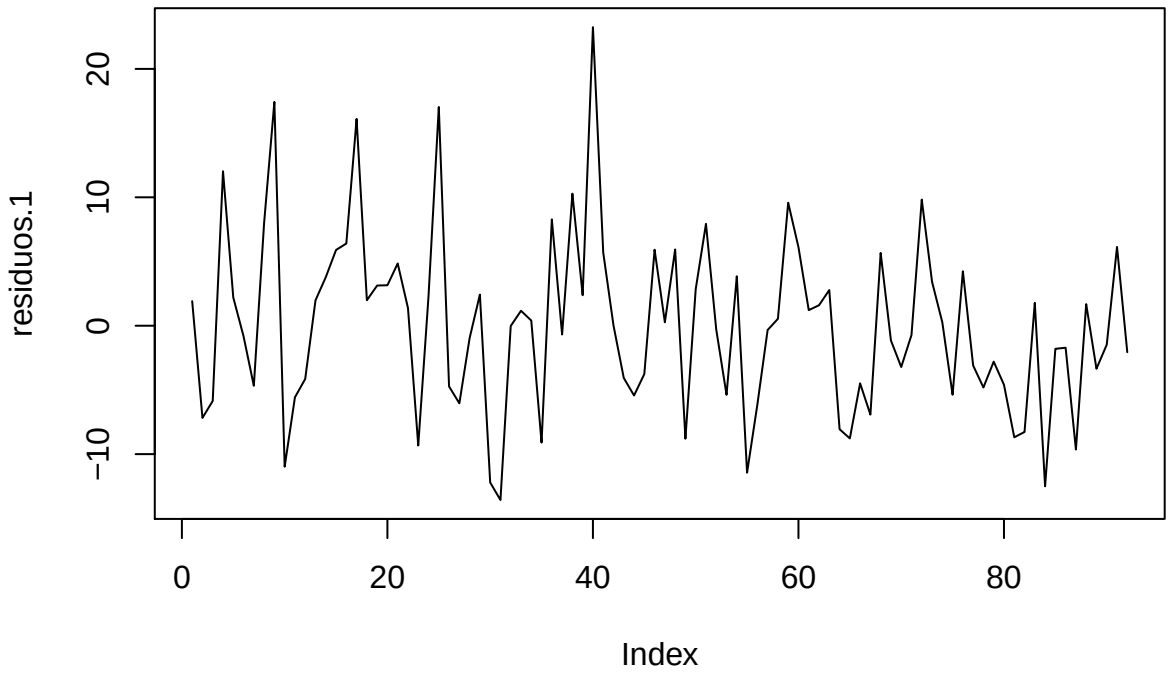
```
coef(Reg.1) # Para obtener los coeficientes
```

```
## (Intercept)      Altura
## -93.8949207   0.9151585
```

```r
fitted(Reg.1) #Devuelve los y estimados a partir del modelo de regression que vienen del lm
```

```
##        1        2        3        4        5        6        7        8
## 60.03474 73.48757 77.23972 76.23304 66.44085 75.13485 73.57908 77.88033
##        9       10       11       12       13       14       15       16
## 72.48089 72.38938 78.70397 73.48757 68.36268 61.13293 71.19967 73.02999
##       17       18       19       20       21       22       23       24
## 65.89175 74.49424 77.97185 59.02806 72.02331 69.18632 68.72874 80.71732
##       25       26       27       28       29       30       31       32
## 46.67342 57.83836 65.34266 63.96992 52.80499 68.63723 65.34266 65.98327
##       33       34       35       36       37       38       39       40
## 66.07478 51.15770 64.15295 77.78881 70.28451 66.62388 68.54571 74.03666
##       41       42       43       44       45       46       47       48
## 62.04809 66.80691 75.77546 75.68395 71.29119 63.32931 67.63055 76.05001
##       49       50       51       52       53       54       55       56
## 79.80216 59.48564 66.07478 58.75352 74.86031 64.51902 78.42942 75.77546
##       57       58       59       60       61       62       63       64
## 69.09481 62.13960 73.12151 80.25974 65.06811 66.07478 72.57241 70.65058
##       65       66       67       68       69       70       71       72
## 72.84696 66.25782 62.32263 64.61053 59.66868 58.29594 59.21110 50.70012
##       73       74       75       76       77       78       79       80
## 50.79164 53.44560 62.50567 56.64865 59.57716 57.28926 57.10623 57.19775
##       81       82       83       84       85       86       87       88
## 54.90985 61.86506 67.35601 63.14628 54.26924 50.33406 52.07286 55.27591
##       89       90       91       92
## 63.96992 51.24922 62.32263 50.05951
```

```r
#El valor real es el valor del modelo más un residuo. Este residuo tiene que tener una media 0 y una va
residuos.1<-residuals(Reg.1) # Pintamos el residuo
plot(residuos.1,type='l')
```



```r
mae<-function(obs,est){ return(mean(abs(obs-est)))
}
mse <-function(obs,est){ return(mean((obs-est)^2))
}
mean(residuos.1) # Residuos: media cero y varianza constante
```

```
## [1] 1.60106e-16
```

```r
yest.1 <- fitted(Reg.1)
mae.Reg.1 <- mae(Peso,yest.1); mae.Reg.1 # Error de validacion
```

```
## [1] 5.327923
```

```r
mse.Reg.1 <- mse(Peso,yest.1); mse.Reg.1
```

```
## [1] 47.66858
```

```r
new.x <- c(174, 156)
predict(Reg.1, data.frame(Altura=new.x)) #Devuelve un peso en kg
```

```
##        1        2
## 65.34266 48.86980
```

Método holdout

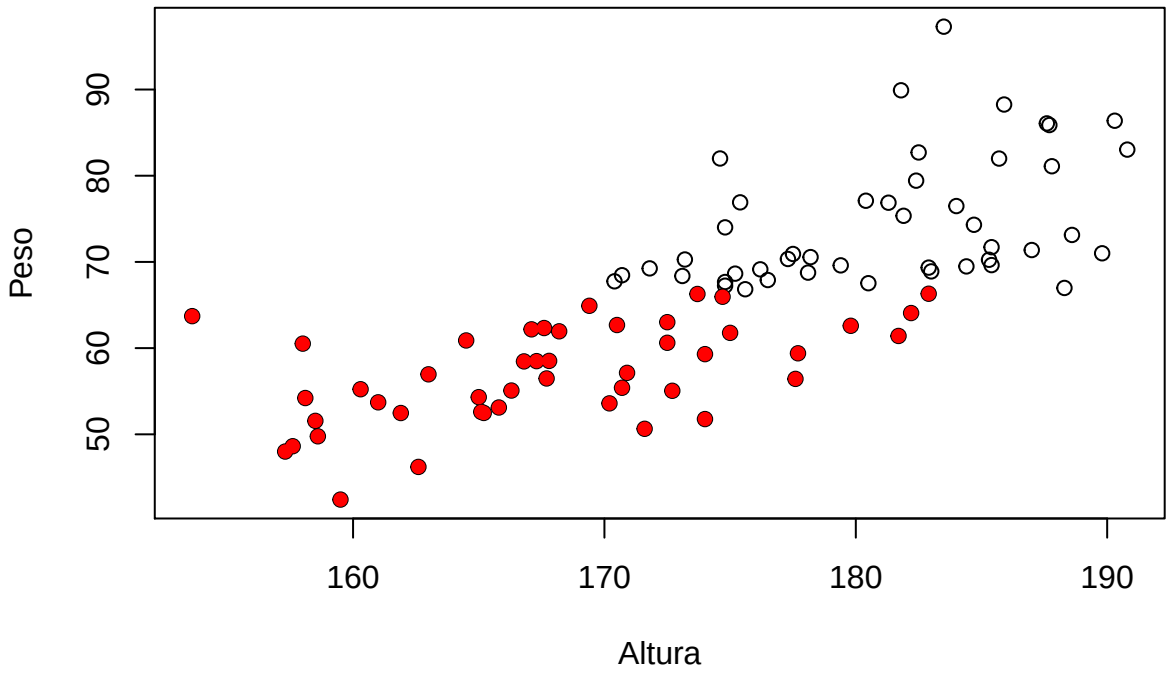
Devidimos la muestra en dos : muestra de entranamiento y de test. Ordenamos los datos del peso y como train, cogemos los valores de peso mas bajos. Vamos a comprobar que no es una buena selección. Vamos además a suponer que la y es constante (no regressión lineal en este caso) y igual a la media => si elegimos las alturas pequeñas en la muestra de test, el valor de y no va a ser correction.

```r
# Validation train-test
plot(Altura, Peso)
n <- length(Altura)

train <- 1:ceiling(n/2)
order.index <- order(Peso)

Peso.sort <- Peso[order.index]
Altura.sort <- Altura[order.index]

#Pintamos en rojo las muestras de entranamiento
points(Altura.sort[train], Peso.sort[train], pch=16, col="red")
```




```
mean.peso <- mean(Peso.sort[train]) #y.est=cte, la cte es la media de y, selecionada en tr abline(h=mea
```

```
mse.train <- mse(Peso.sort[train], mean.peso); mse.train
```

```
## [1] 31.42861
```

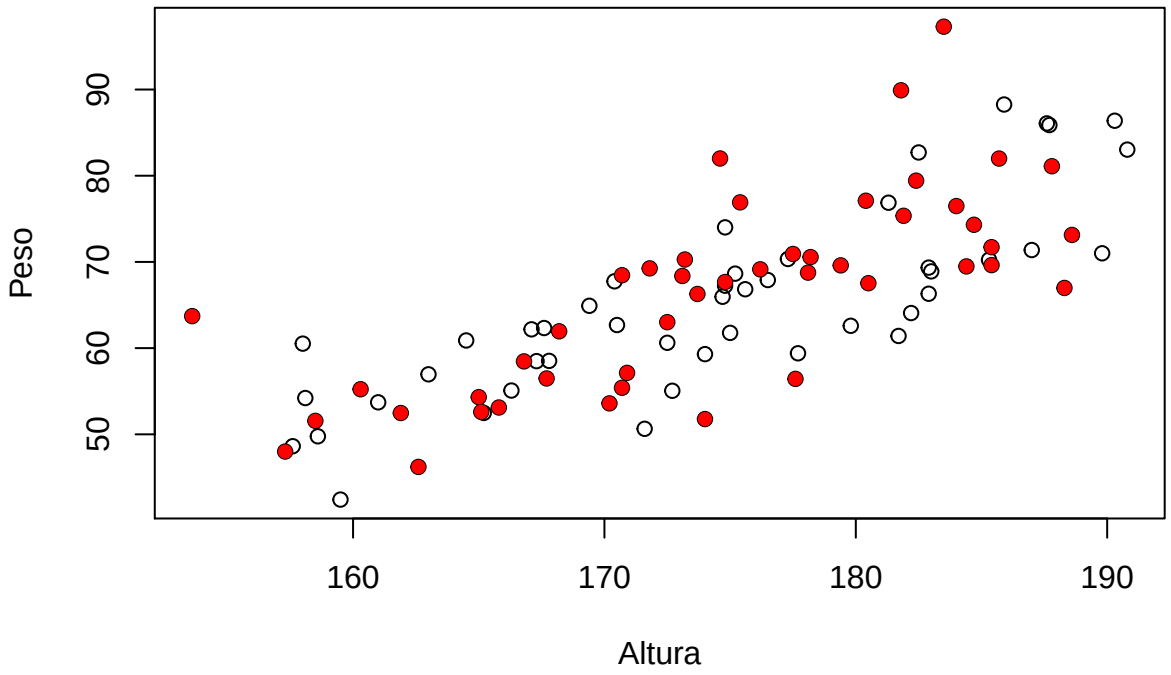
```
mse.test <- mse(Peso.sort[-train], mean.peso); mse.test
```

```
## [1] 354.6602
```

El error de los individuos que tienen el peso grande es mucho mayor, porque hemos elegidido mal los datos de entranamiento. Queremos obtener dos errores similares, aquí estamos haciendo un sobreajuste.

Vamos a elegir datos aleatorios en el train, y los otros van al test, lo que tienen que ser mejor.

```
# Para obtener todos los mismos resultados cuando utilizamos un generador de números aleatorios
set.seed(1)
train <- sample(n,ceiling(n/2))
plot(Altura, Peso)
points(Altura[train], Peso[train], pch=16, col="red")
```




```
mean.peso <- mean(Peso[train]) #y.est=cte abline(h=mean.peso)

mse.train <- mse(Peso[train],mean.peso); mse.train
```

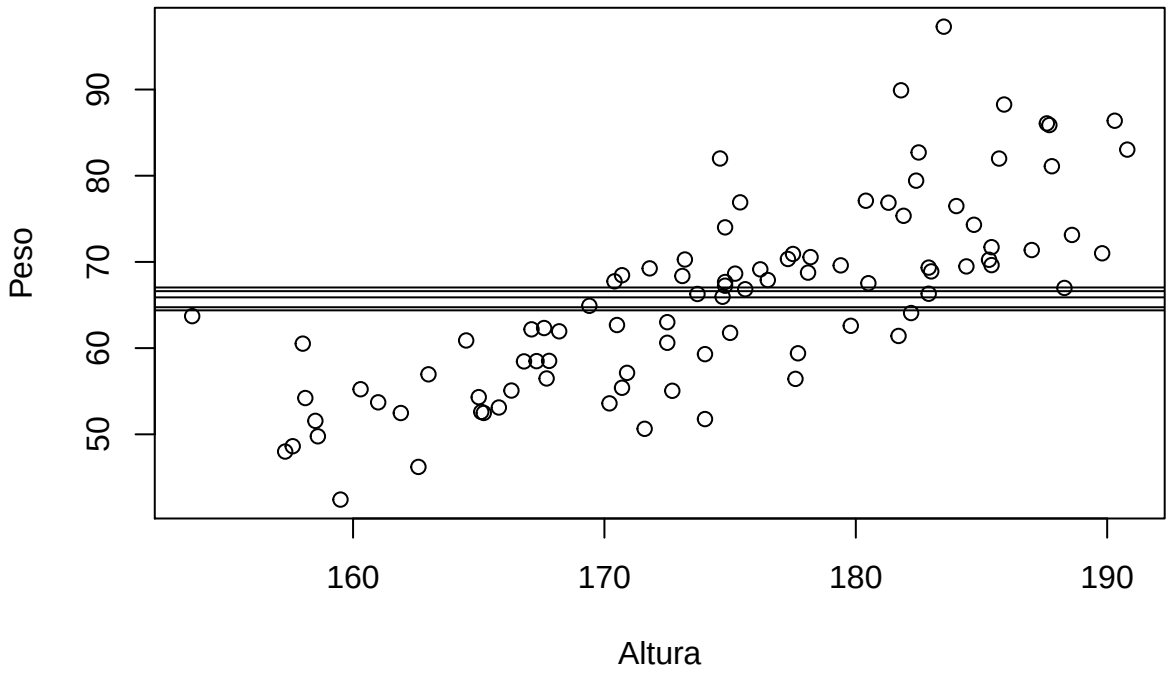
```
## [1] 124.3255
```

```
mse.test <- mse(Peso[-train],mean.peso); mse.test
```

```
## [1] 112.5511
```

Este método tiene dos incovenientes explicado en los apuntes. Si quitamos en set.seed el entranamiento va a tener conjuntos de entrenamiento distintos. Vemos que el error cambia mucho en funciòn de la muestra de entranamiento elegida.

```
plot(Altura, Peso)
for (i in c(1:5)){
train <- sample(n,ceiling(n/2))
mean.peso <- mean(Peso[train]) #y.est=cte esa cte es la media de la variable y selecciona
abline(h=mean.peso)
print(mse(Peso[-train],mean.peso))
}
```




```
## [1] 141.9859
## [1] 149.7081
## [1] 105.106
## [1] 105.3994
## [1] 100.6536
```

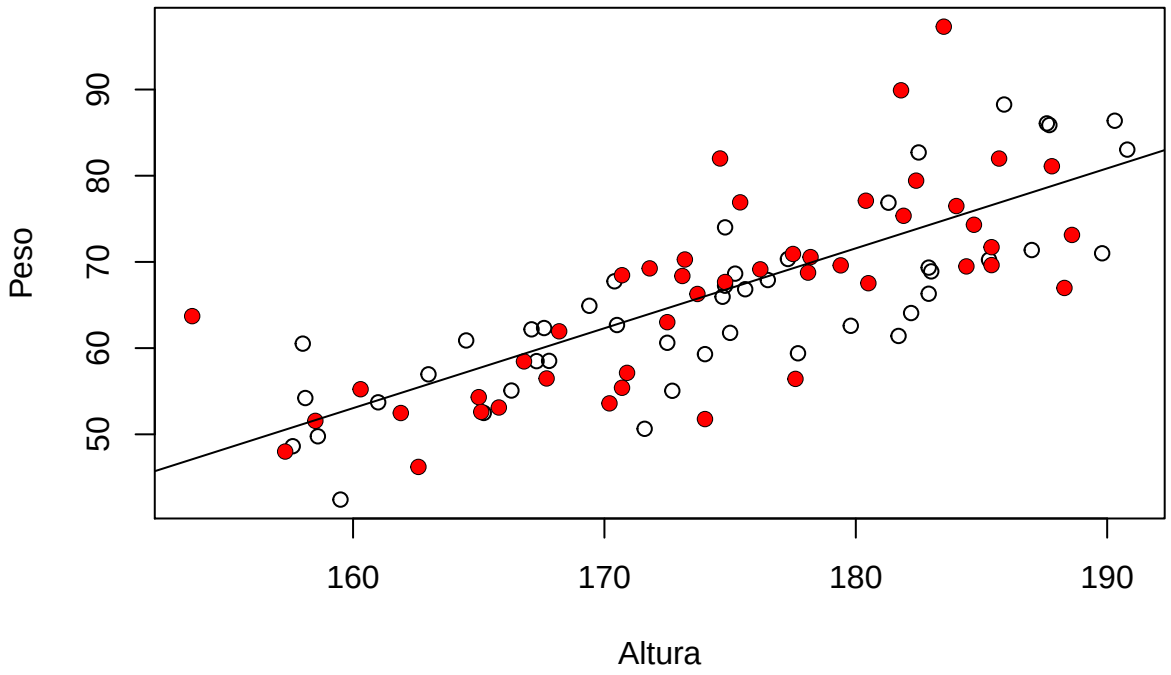
Ahora hacemos lo mismo pero cambiando el modelo usado : en lugar de una constante, poner un modelo lineal. Tenemos que ver si este modelo es mejor, y si el error cambia o no si cambiamos la muestra de entrenamiento (entonces, ya no usamos mean.peso y lo cambiamos por el modelo lineal). También no funciona bien si tenemos pocos datos en la muestra, porque hay que dividirla en dos.

```
set.seed(1)

train <- sample(n,ceiling(n/2))

Reg.train <- lm(Peso~Altura, data=Pulsaciones, subset=train)

plot(Altura, Peso)
points(Altura[train], Peso[train], pch=16, col="red")
abline(a=Reg.train$coefficients[1], b=Reg.train$coefficients[2])
```




```
mse.train <- mse(Peso[train], fitted(Reg.train))
mse.train
```

```
## [1] 57.37094
```

```
mse.test <- mse(Peso[-train], predict(Reg.train, data.frame(Altura=Altura[-train])))
mse.test
```

```
## [1] 38.91509
```

Leave one out

```
train <- 1:n
yest.3 <- rep(NA, n) #Definimos la dimension del vector para ahorrar memoria
for (i in train){
  #Ejecutamos lm para toda la base de datos excepto el elemento -i
  Reg.i <- lm(Peso~Altura, data=Pulsaciones, subset=train[-i])
  yest.3[i] <- predict(Reg.i, data.frame(Altura=Altura[i])) #Cada elemento de yest se ha generado con u
}
mse.Reg.3<-mse(Peso,yest.3); mse.Reg.3
```

```
## [1] 50.02996
```

Inconvenientes del método : si n es muy grande, tarda mucho en correr el código.

```
Reg.3 <- glm(Peso~Altura, data=Pulsaciones)
cv.err <- cv.glm(Pulsaciones, Reg.3) # cv.glm() refit the model all the n times
cv.err$delta #delta gives the mse error (first value) and the bias corrected version (seco # The bias c
```

```
## [1] 50.02996 50.01690
```

k-fold

Menos datos para el train en este caso pero es cuesta menos correr este código.

```
idx.aleatorios <- sample(1:n,n,replace=F)
K <- 10 #Número de intervalos
tam <- ceiling(n/K) #Número de elementos en cada uno de los intervalos
yest4 = rep(NA, n)
for (i in 0:(K-1)){
idx.test <- idx.aleatorios[(i*tam+1):((i+1)*tam)] #Reordenamos los datos de manera aletoria para estar
idx.test <- idx.test[!is.na(idx.test)] #tam vale 10 y tenemos 92 datos. El ultimo intervalo tiene 2 ele
lm4 <- lm(Peso~Altura, subset=-idx.test)
yest4[idx.test] <- predict(lm4, data.frame(Altura=Altura[idx.test])) #Predicta los valores para la mues
}
mse4 <- mse(Peso,yest4); mse4
```

```
## [1] 50.3897
```

```
# We can also use glm()
Reg.4 <- glm(Peso~Altura, data=Pulsaciones)
cv.err <- cv.glm(Pulsaciones, Reg.4, K=K) # cv.glm() refit the model considering the k-fol
cv.err$delta # As mentioned before, here both numbers are not the same. It turns out
```

```
## [1] 50.72746 50.56223
```

```
# that has more of an effect for k-fold cross-validation. Do not using the leave-one-out # cross-valida
```

Da un error un poco más alto que en el caso precediente pero hemos simplificado batsante los cálculos hechos. El segundo método de cálculo nos devuelve dos valores más diferentes que antes (porque estamos penalizando el método, no estamos usando el mayor número posible para test como antes).