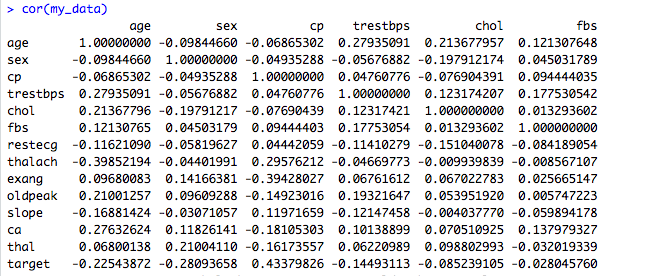
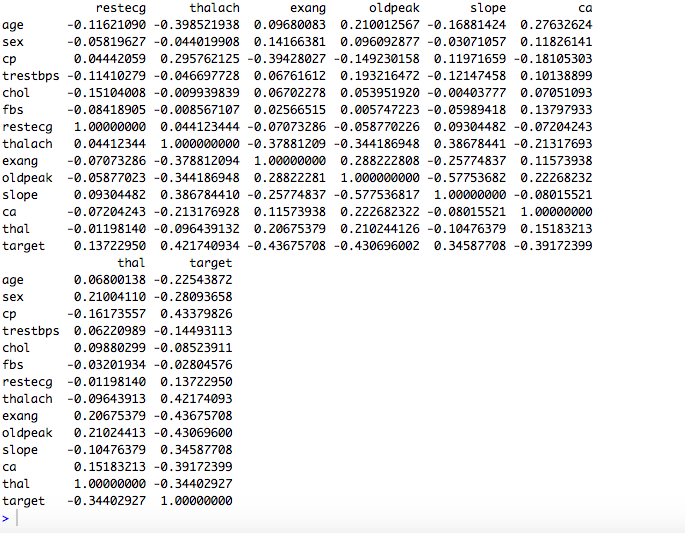
STATISTICAL ANALYSIS :

1) LOOKING FOR IMPORTANT CORRELATION :

Based the correlation matrix constructed with the 13 predictors, no strong (>0.5) correlation has been found.

So the analysis takes in count all the predictors.





We interpret the coefficients <0.5 as not strong, meaning that we could not erase one variable assuming that the others will replace it.

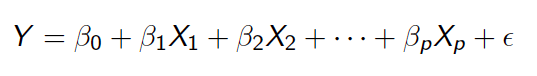
2) METHODS :

These next methods are done with all the predictors.

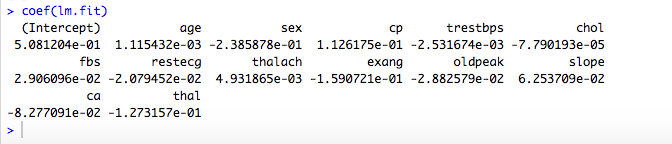
Multiple Linear Regression :

Linear regression is very simple approach but interesting to have as a base and comparaison for later.

We find a formual :



Where the Xis are the differents predictors (age, sex…) and the bi are the differents coefficients :

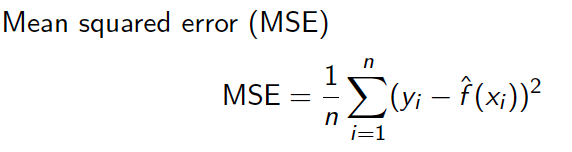


We used half of our datas as training set, and the other half as test set.

We set target=1 if we find >=0.5, and target=0 otherwise.

To test the accuracy of the model we computed the MSE and find 0.1578947.

*Recall*:

where yi = correct value

^f(xi) = estimated value

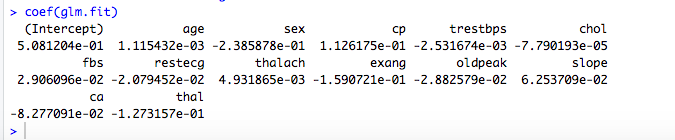
Logistic Regression :

Trying to predict a heart disease means that our response is qualitative instead of quantitative.

This means that we have to classify our « observations ».

Logistic regression is known to be an effective method in this case.

Here are the coefficients founds computing a logistic regression :



We used two way to estimate the error :

We computed the mean square error to compare with the linear regression :

Mean square error : MSE = 0.1578947

We find that computing the logistic regression model leads to the exact same model as the linear one.

We also computed the cross validation error to estimate the test error and compare it with our further models :

Leave-one-out cross validation error : CV = 0.1319878

Classification Tree :

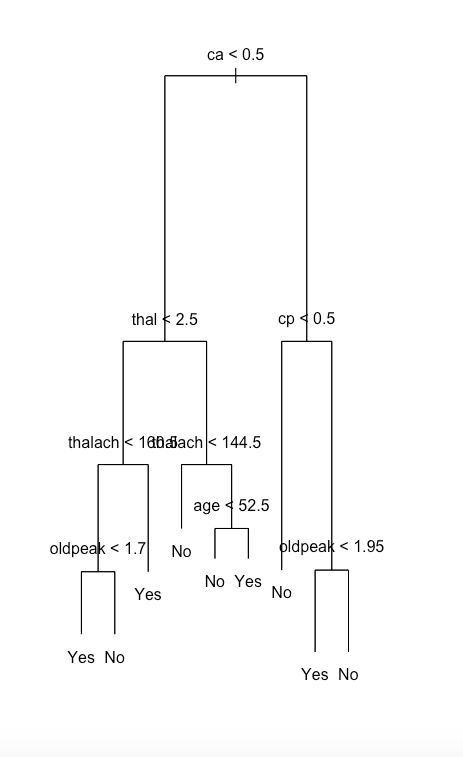
We are aware that « trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches seen ».

On the other hand, we also noticed that they provide a global view of the situation and can easily be interpreted.

This can be useful, for example in the case where a doctor want to explain the reason of the differents tests and conclusions to a patient.

This is the reason why we provided one model.

After computing the classification tree, we constructed the prunning tree of size 9, as our main goal is to have a general overview, and obtain :

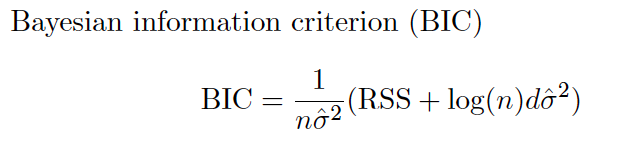


At this point of the analysis, even if the data set has already been (selecting 14 variables instead of 76), we tried to find if we still couldn’t find an accurate subset, which would reduce the number of variables.

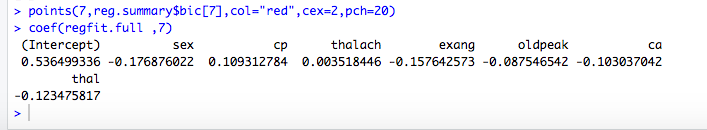
We performed : The Best Subset Selection.

We tried to find the optimal subset according to criterions :

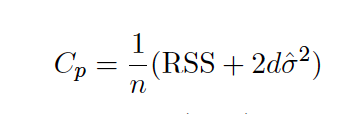
**BIC**:



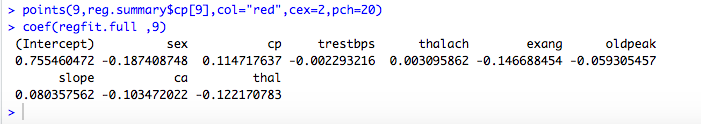
Subset of 7 selected predictors:



**Cp**:

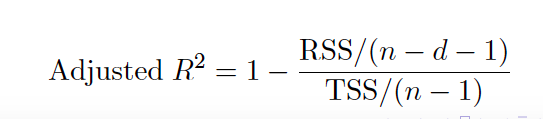


Subset of 9 predictors selected :

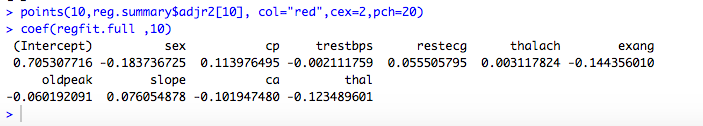


Cp and BIC are estimates of test MSE.

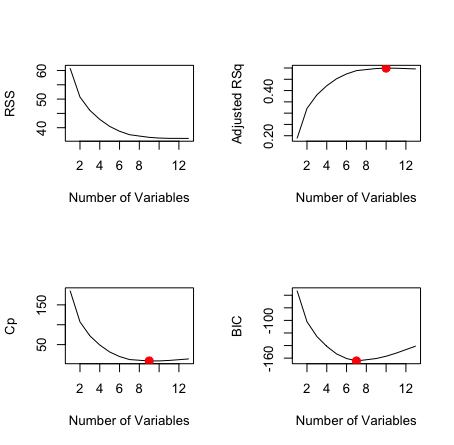
**Adjusted R square**:



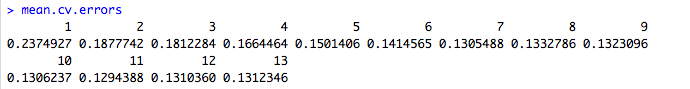
Subset of 10 predictors selected :



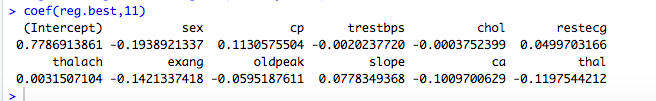
Plots of the results :



**Applying the Cross** **Validation**:



we find the best subset has 11 variables, with coefficients :



We conclued that there is no clear subset selection, which is not surprising because the selection has already be done previously.

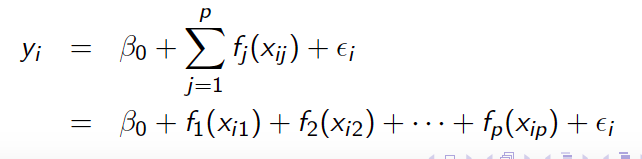
After this subset selection, we started again with differents methods, using the last 11 variables found :

Logistic Regression :

We got results, but logistic regression is not hard to comlpute, so it didn’t make sense to remove variable in this general model.

Generalize Additive Model (GAM) :

Instead of allowing coefficents to the variables, we provide non-linear functions of these variables :

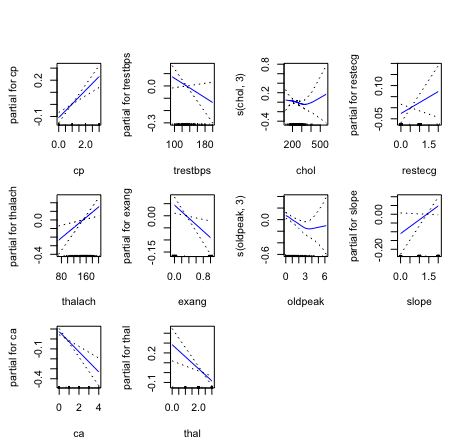


We compared 5 diffrents models.

When we computed the best subset selection, we notices that even if the 11-variables (cv = 0.1294388) was the best one, the 7-variable was also very close (cv = 0.1305488). So we’ll use this difference between the 7 and others 3 variables of the 11-variables model.

We started giving cubique functions to 7 variables of the best subset selection and quadratics functions to the 3 left (the ones in the 11-variables best subset but not in the 7-variable).

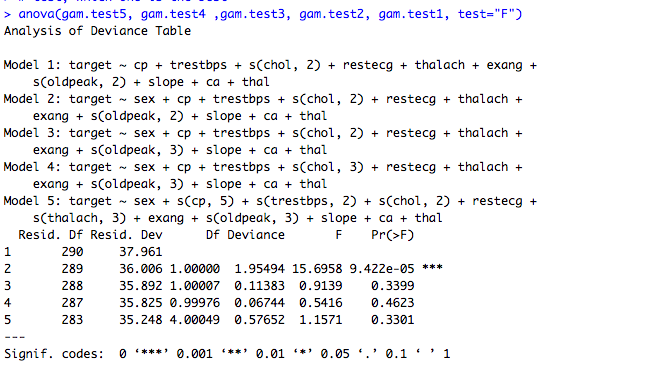
To approximate each functions for each variables we made a plot of this first try :



We notices that except for chol and oldpeak, all function can actually be approximated by linear function.

The next moove was then to identify wich were the best functions for these two variables left.

We tried cubique and quadratic functions and compared the models :



We find that the second model is the best, using cubics functions for our variables chol and oldpeak.

We finish and obtain our best model, with MSE = 0.1121472