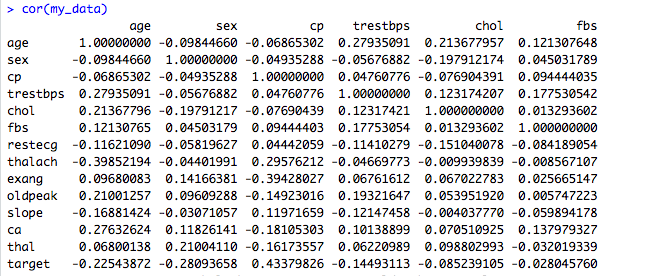
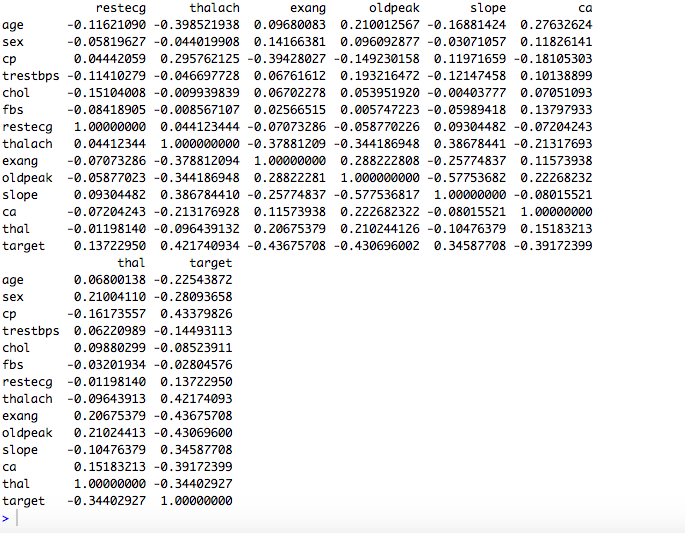
STATISTICAL ANALYSIS :

1) LOOKING FOR IMPORTANT CORRELATION :

The correlation matrix shows how strongly correlated the respective variables are to each other. As we define a strong correlation of being > |0.5|, there is only one pair of variables, which satisfies this criteria: “oldpeak” with “slope”.  
This was to be expected, since both of the variables deal with the ST-T segment of the ECG.  
Since “oldpeak” is a quantitatitive variable describing how much the effects of exercise effect the segment and “slope” is a qualitative variable describing how the segment is affected, we decided to keep both variables.

In the last row you can see how much “target” correlates with the other variables. Worth mentioning is that “age” and “sex” aren’t strong predictors for cardiovascular disease. It’s already measurable symptoms such as occurrence of an angina, ECG results and having a family history of thalassemia which are much more statistically relevant.





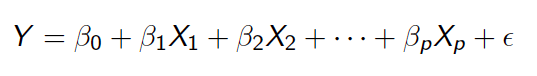
2) METHODS :

For the following methods we have chosen to apply all the predictors.

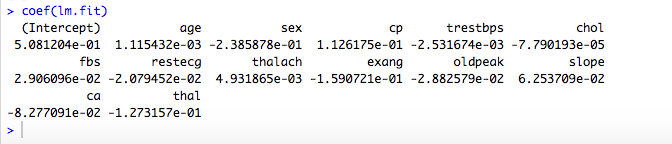
Multiple Linear Regression :

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

We based our approach to multiple linear regression on the formula:



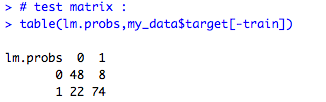
Where the Xi’s are the different predictors (age, sex…) and the βᵢ’s are the different coefficients:



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

We defined the target to be predicted as 1 if y 0.5, and as 0 otherwise.

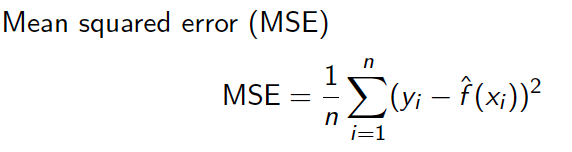
On the test set, we obtained these results :



This amounts to 122 correct predictions out of 152 in total.

And to test the accuracy of the model we also computed the MSE as 0.1973684.

*Recall*:

where yi = correct value

f̂ (xi) = estimated value

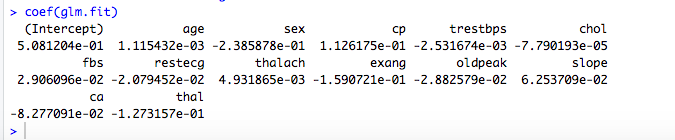
Logistic Regression :

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 = 0.1973684

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 :

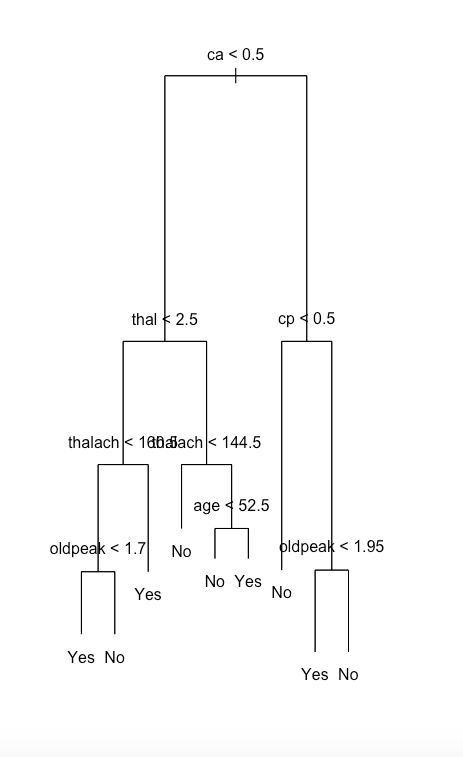
We are aware that « trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches seen ».

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 wants 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:

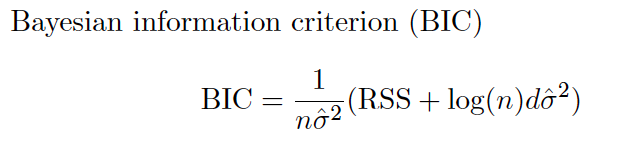


At this point of the analysis, even if the data set has already been pre-processed (selecting 14 variables instead of 76), we tried to find if we still couldn’t find an accurate subset, which would reduce the number of variables.

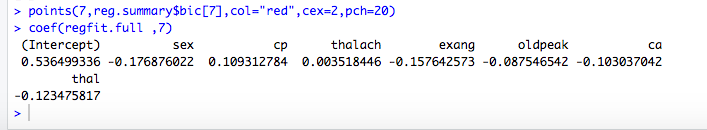
We performed : The Best Subset Selection.

We tried to find the optimal subset according to criterions :

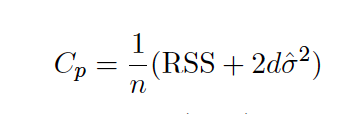
**BIC**:



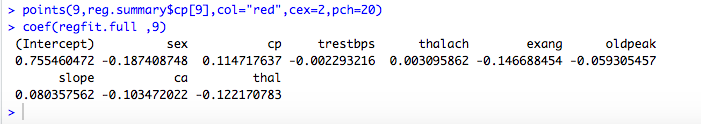
Subset of 7 selected predictors:



**Cp**:

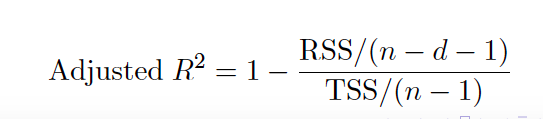


Subset of 9 predictors selected :

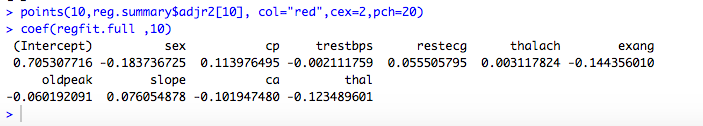


Cp and BIC are estimates of test MSE.

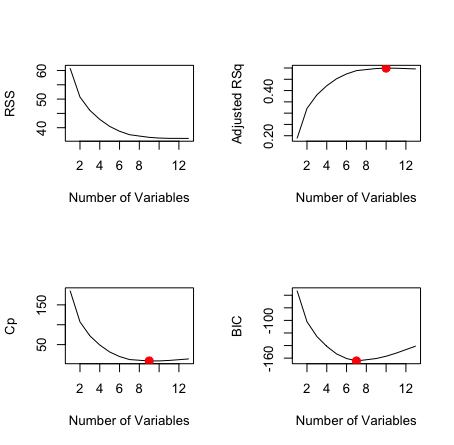
**Adjusted R square**:



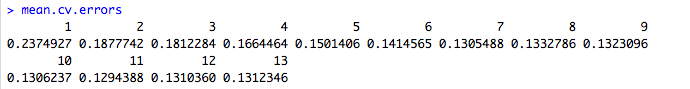
Subset of 10 predictors selected :



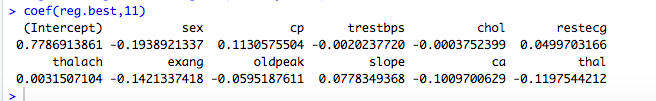
Plots of the results :



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



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



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

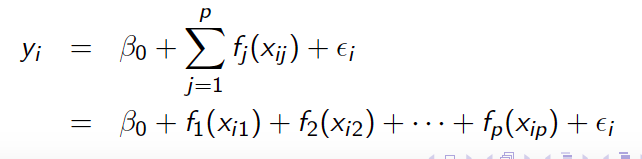
After this subset selection, we started again with different methods, using the last 11 variables found:

Logistic Regression:

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

Generalize Additive Model (GAM) :

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

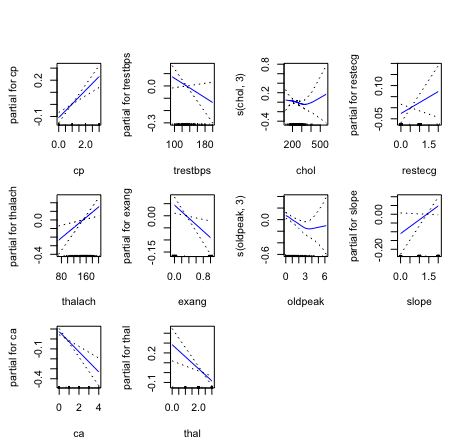


We compared 5 different models.

When we computed the best subset selection, we noticed 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 cubic functions to 7 variables of the best subset selection and quadratic functions to the 3 left (the ones in the 11-variables best subset but not in the 7-variable).

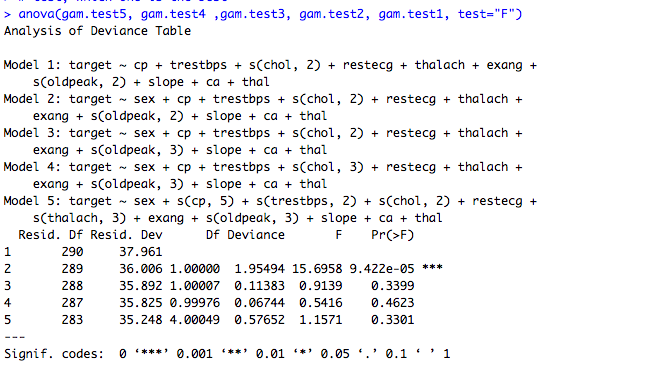
To approximate each functions for each variables we made a plot of this first try :



We notice that except for chol and oldpeak, all function can actually be approximated by using a 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 cubic functions for our variables chol and oldpeak.

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

RESULTS AND DISCUSSION :

We have found, that there was no one distinct way, for which to select the predictors to use in the statistical model. Interesting here is that the 7 variables found through the BIC Criterion are a subset of the 9 predictors of the Cp Criterion, which in turn is a subset of the predictors we found trough the adjusted R2 and eventually through Cross Validation.

Common to all four subsets are the predictors “sex”, “cp”, “thalach”, “exang”, “oldpeak”, “ca” and “thal”.

Worth noting here is, that “cp”, “thalach”, “exang”, “oldpeak” can all be directly linked to the heart and it measurably not functioning correctly. “thal” is similar in that it indirectly affects how much oxygen reaches the heart. We didn’t expect the gender of the person to be included in all the four subsets. It’s especially surprising considering it only has an approximate -0.28 coefficient, whereas slope had an approximate 0.35 coefficient and was only included in three of the subsets. Our guess is that the data is generally slightly screwed towards male subjects, since they made up 68.3% of the population.

Also interesting is the fact, that age and fasting blood sugar were included in none of the data subsets. As the first quartile of the population starts at 47.5 years of age and the third at 61 years, most of the there isn’t that much of an age gap in between most of the individuals. This shows, that being healthy in general is more relevant to preventing heart disease, than just being young.

The fbs value is typically used for predicting diabetes and its pre-stage. Since the rest of the dataset’s variables are purely heart focussed or general variables such as sex and age, it comes to no surprise, that fbs isn’t as relevant as the others.

We still decided to go with the full set of 11 variables, since the population of our dataset is only 303 entries, which wasn’t too much for our computer to handle. For a considerably bigger dataset we would have chosen a variation with only the 7 variables or would have tried to reduce the dimensions with Principle Component Analysis.

In order to predict heart disease, the most accurate model found, comparing tests errors, is the generalize additive model. It uses non-linear functions as coefficients for the predictors.

A shown in the method part, cubic functions must be assigned to the variables chol and oldpeak, while all the other variables can sufficiently have linear function.

This project wasn’t meant to be ground-breaking research in and on its own. We hope to support current existing research in the field and ideally supplement it with our own insights and perspectives. At the very least we’d like to confirm past results and use the opportunity to learn about public health.  
It is our goal to raise awareness on heart diseases as they are the number one cause of death worldwide and as can be seen from our data can’t just easily be predicted by just sex and age. They can often only be predicted when the first symptoms start to show, which can fatal.

We are happy to have had the opportunity to explore the dataset and learn something about heart disease and how it can be prevented and are looking forward to receiving feedback.