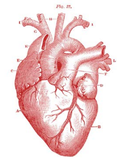
Heart Disease Prediction
COMP 3354 – Statistical Analysis



MOTIVATION

According to the World Health Organization heart diseases, also known as Cardiovascular Diseases (CVDs) represent the number one cause of death globally. As deaths of CVDs can be prevented through appropriate treatment, we hope to ensure that patients with a high risk of CVDs can get identified using a statistical approach, even before the first symptoms start to show.

VARIABLES

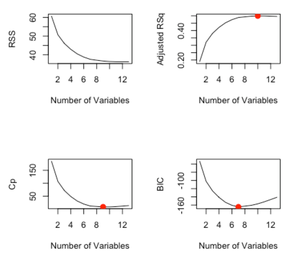
The dataset consists of 303 individuals, with 165 cases of cardiovascular diseases and 136 members of the control group.

It originally contained a total of 76 variables, many of which were deemed statistically irrelevant by previous researchers, to end up with :

- Age
- Sex
- Chest pain (cp)
- Resting blood pressure (trestbps)
- Serum cholesterol levels (chol)
- Resting electrocardiographic results (restecg)
- Maximum heartrate achieved (thalach)
- Exercise induced angina (exang)
- ST depression induced by exercise (oldpeak)
- Peak exercise ST segment slope (slope)
- Number of major blood vessels coloured by fluoroscopy (ca)
- Hereditary disease Thalassemia (thal)
- Target

Best Subset Selection

The 13 variables we worked on were already a subset of 76 variables but we verified that we couldn't find a smaller subset.
We have found, that there was no one distinct way, for which to select the predictors to use in the statistical model and this is the reason why we kept the 11 variables obtained with cross-validation*.
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 R^2 and eventually through Cross Validation.



* sex, cp, trestbps, chol, restecg, thalach, exang, oldpeak, slope, ca, thal

METHODS :

- Multiple linear regression
- Logistic regression
- Classification tree :

accessible to non-expert and easily intepretable model :

- Generalized Additive Model :

Providing functions as coefficients for each variables,
this is the most accurate model found according to the tests errors.

Results :

MSE = 0.1121472

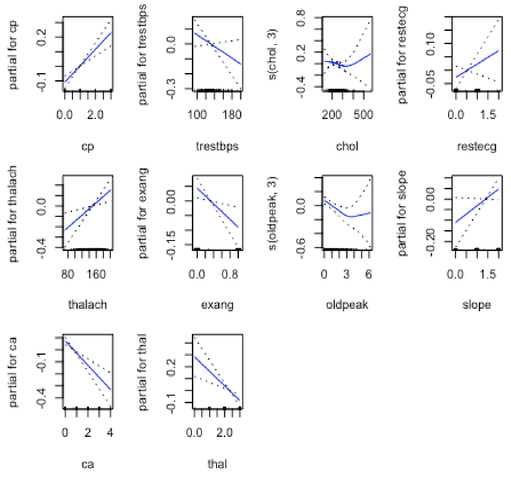
Best functions found for each variable :

linear functions for :

- sex
- cp
- trestbps
- restecg
- thalach
- exang
- slope
- ca
- thal

quadratic functions for :

- oldpeak
- chol



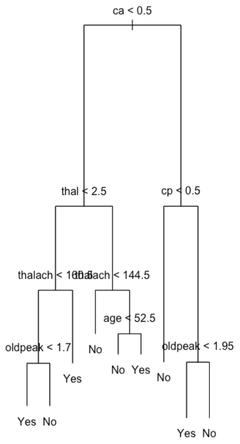
How to get this model :

training function (in R) :

```
predict(gam.test2,my_data)
```

prediction function (in R) :

```
gam.test2=gam(target~sex+cp+trestbps+s(chol, 3)+restecg+thalach+exang+s(oldpeak, 3)+slope+ca+thal,data=my_data)
```



DISCUSSION:

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.