METHODOLOGY

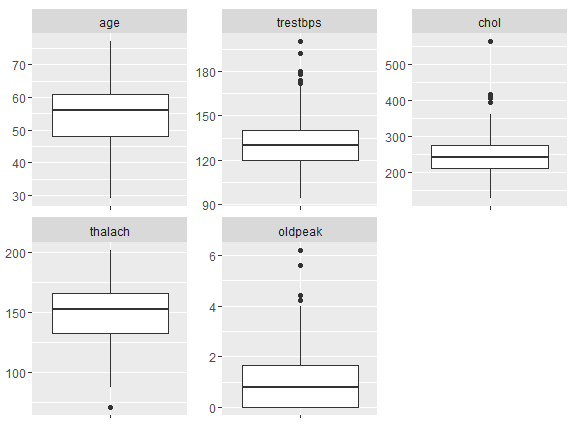
1. Data Cleaning

The original dataset contained 303 observations, of these 2% were invalid. These rows were removed without significant impact to data integrity or model strength as an appropriate classification method for the missing data was not identified. Further, some data transformation was implemented to improve data usability. Categorical variables were mistakenly classified as numeric datatype and were appropriately updated to the factor datatype. Lastly, dimensions were renamed and reordered for convenience.

1. Exploratory Analysis

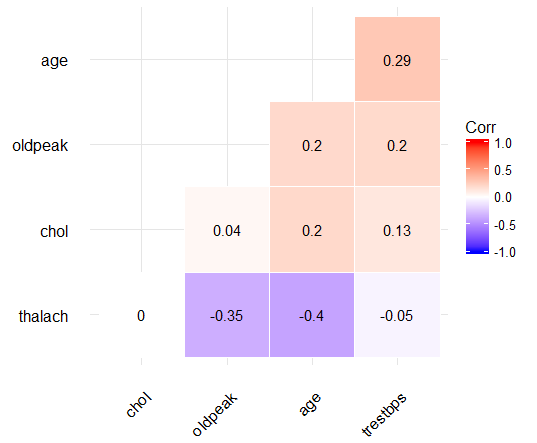
A boxplot was constructed, Figure 1, to observe the presence of outliers and skewness in the numerical variables. We observe that chol, trestbps and oldpeak have a number of outlying data while age has some left skewness, indicating this is a relatively younger sample.

**Figure 1**

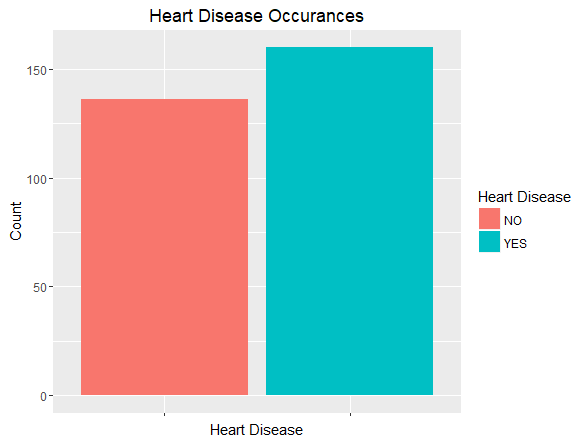


Next, a correlogram was made, Figure 2, to explore the correlation of numerical covariates on each other. A strong negative correlation between age and thalach and oldpeak and thalach was observed, while a strong positive correlation between age and trestbps was observed. Other notable positive correlations were seen for age and oldpeak, age and chol, and oldpeak and trestbps.

**Figure 2**



The dependent variable is checked for imbalance. There are 160 observations that present with heart disease and 136 that do not (Fig 3). A ratio 20:17 indicates the data is fairly balanced.



MODEL COMPARISON

It was observed that the variables fbs and chol did not have a significant relationship with the dependent variable target and that chol did not have strong correlations with the other numerical variables. As a result, these two variables were not included in our initial model (Fig 4).

Figure 4, initial model

After the initial model was created, backwards selection and stepwise selection were employed to determine variable selection for our final two reduced models. At each step of the backwards selection, a drop in deviance test was conducted between each successive reduced model and the previous model, beginning with the initial model. The backwards selection process ended when the drop in deviance test did not select for the next iteration (Fig 5).

Figure 5, backwards selection model

For the stepwise selection, a minimum Akaike information criteria (AIC) approach was used when performing model comparisons. Beginning with covariate that produces the model with the smallest AIC, we continue to add or remove the next covariates in the same way. After each meaningful combination is explored, we select for the model that has the lowest AIC (Fig 6).

Figure 6, stepwise model

After performing the backwards selection and stepwise selection, a final model comparison was done using both minimum AIC and drop in deviance. The stepwise model was found to be necessary.

INTERACTION TERMS

In the course of exploring the data and developing our model we considered adding interaction terms. After determining the strength of correlation between the covariates and testing the linearity of the model, no specific interaction term appeared as an obvious choice. To be complete, we explored the population data from where our sample was derived. Given the disparity in location, geographic location – Budapest, Zurich, Basel and Cleveland, no obvious demographic interactions were made apparent either. Despite our preliminary findings, interaction terms are an area that warrants further exploration.

GOODNESS OF FIT and R^2

Our stepwise model was tested for goodness of fit using both the deviance as a test statistic and finding the pseudo R^2. Our model has a residual deviance of 182.16 on 280 degrees of freedom (p=0.999) which provides strong evidence against rejecting the null. While pseudo R^2 values for count data are not commonly reported, we endeavoured to use the most common one, McFadden’s R^2 which provided a value of 0.554. Taking into consideration both of these tests, we concluded that our model is an adequate fit.

SUMMARY

Using regression modelling techniques, in particular, logistic regression allows one to model independent categorical and/or numeric variables on a binary dependent variable. In this dataset, the presence of heart disease is a binary response. We propose the use of regression modeling and associated statistical techniques to determine which attributes may help to predict heart disease. We apply our methods to determine the most significant covariates to include in a final model and show a reduced model, one that focuses on key features, has a better overall fit than a full model for this dataset.