Research Project - Mathematical Statistic (30549)

January 2023

Introduction

Despite the advancement and discoveries which have been carried out by medicine in the past decades, cardiovascular diseases are still the number one cause of death globally, killing almost 18 million people per year, which amounts to more than 30% of all deaths worldwide. 80% of deaths caused by cardiovascular diseases are due to heart attacks and strokes, and one third of these occur prematurely in people under 70 years of age.
As prevention can really impact the lethality of such diseases, we decided to study this data set and to build a model in order to predict the onset of a heart disorder given some medical observations.

About the Dataset

The dataset [2] was built by Federico Soriano (September 2021) by combining five different databases available independently on the UCI Machine Learning Repository. More precisely, the datasets used were curated by Andras Janosi, (M.D - Hungarian Institute of Cardiology, Budapest), William Steinbrunn (M.D - University Hospital, Zurich, Switzerland), Matthias Pfisterer (M.D - University Hospital, Basel, Switzerland) and Robert Detrano (M.D and Ph.D., V.A. Medical Center, Long Beach and Cleveland Clinic Foundation). The final dataset contains 918 observations.

Moreover, each column specifies a different attribute:

1. `Age`: age of the patient in years;
2. `Sex`: sex of the patient (M stands for male, F stands for female);
3. `ChestPainType`: indicates one of four different chest pain types (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic);
4. `RestingBP`: resting blood pressure measured in mm Hg;
5. `Cholesterol`: serum cholesterol measured in mm/dl;
6. `FastingBS`: fasting blood sugar (1: if FastingBS $>$ 120 mg/dl, 0: otherwise);
7. `RestingECG`: resting electrocardiogram results (Normal: Normal, ST: having ST-T wave abnormality[1], LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria);
8. `MaxHR`: maximum heart rate achieved (numeric value between 60 and 202);
9. `ExerciseAngina`: exercise-induced angina (Y stands for Yes, N stands for No);
10. `Oldpeak`: oldpeak = ST depression value in the ECG;
11. `ST Slope`: the slope of the peak exercise ST segment (Up: upsloping, Flat: flat, Down: downsloping);
12. `HeartDisease`: output class (1 if the patient has a heart disease, 0 otherwise);

[1] T wave inversions and/or ST elevation or depression of $>$ 0.05 mV

We checked whether our dataset contained missing values using the `is.na()` function and we happily found out that there were none.

Furthermore, the only manipulation done on the dataset in order to make it more understandable and more tractable, is the addition of dummy variables for those for which it was applicable: `ChestPainType` was split in its four categories, as `RestingECG` and `ST Slope`; on the other hand, `Sex` has been made binary in terms of M (`SexM` 1 if male 0 if female), and the same for `ExerciseAngina`, but in terms of yes (`ExerciseAnginaY` 1 if yes 0 if no).

Finally, we identified outliers in the values of `Cholesterol` and `RestingBP`: some values of these columns were equal to zero, which is scientifically impossible. Therefore we guessed that this could be due to a measurement error and we set those values equal to the mean, in order not to lose otherwise interesting observations.

Exploratory Data Analysis

We began to explore the dataset by running a preliminary analysis using the `summary()` function. Then we decided to plot the distribution of some variables, such as `Age`, `RestingBP`, `Cholesterol`, `MaxHR` and `Oldpeak`. The plots are, of course, consistent with the analytical results given by the `summary()` function.

The sample of patients considered in the study ranges from 28 to 77 years old. Moreover, by running a Kolmogorov-Smirnov test, we see that `Age` is not normally distributed[2].
We can repeat the same procedure and we get that normality doesn't hold also for the distribution of `RestingBP`, `Cholesterol`, `MaxHR` and `Oldpeak`. However, the fact that our variables are not normally distributed will not affect our study. Moreover, we didn't deal with outliers, since the data is a result of a scientific research[3].

Note how the mean of the resting blood pressure is 132.4, which is a considerably high value[4]: this is probably due to the fact that we are dealing with patients who are potentially at risk of a heart disease, and therefore they might have higher blood pressure on average.

We then studied the percentage of male and female participants in the study: women are only 193 (21% of the whole) against the 725 male patients (79%). It is important to underline how this lack of proportionality in the sex of the participants might influence the results of our study, i.e. we might see higher probability for a man to develop a heart disease than in reality.

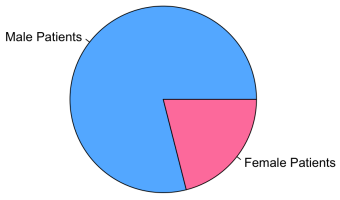


Figure 1: Sex Distribution of The Patients

[2] As a matter of fact, we get that p-value $< 2.2 \cdot 10^{-16}$, thus we reject the null hypothesis;
[3] In the result paragraph there is the analysis of influential outliers
[4] Ideally, blood pressure should be lower than `120` mmHg;

Finally, we plotted whether our patients had a heart disease or not[5] against `Sex`, `ExerciseAngina`, `ChestPain Type`, `RestingECG` and `Oldpeak`.
It is worth noting that the histogram where the presence of a heart disease is plotted by sex, Figure 2, gives us an interesting insight on how gender affects our study: even though it is reasonable to see a higher number of men in the column of patients affected by a heart disease, as we said that men are the majority of patients in our study, we can intuitively see that it is not proportional to the difference in male and female percentages in the dataset. Therefore, gender might actually play an important role in our analysis[6].

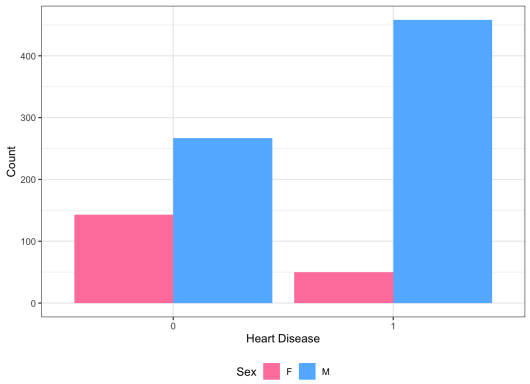


Figure 2: Heart Disease by Gender

Model

After the Exploratory Data Analysis, we can move on with the regression model. We used a logistic model, since we performed a categorical prediction (yes or no type of question), that is forecasting if a patient is predisposed to have a hearth disease based on some (or all) the characteristics of the individual present in the dataset. The model was selected using two different methods, step-wise and LASSO, of which we compared the results, together with the full model, in order to see which one has the best simplicity and accuracy. Moreover, we will analyze the fit of the model also thanks to the McFadden's pseudo-R^2, since the traditional R^2 is not adequate. The last thing that we did, was checking for the assumptions of the logistic regression

Firstly, we divided the set two parts, selected randomly: 80% of the dataset was designated to training, and the remaining 20% to testing. Note that the training and testing set used were the same for all the different models built, and because of randomness our results will change if our sets change. The 80/20 ratio was chosen by us and is purely arbitrary; when training a model, it is preferred to take a big enough portion of the data, without penalizing too much the testing set.

Model selection

As explained in the previous paragraph, the first method used for building our model is step-wise model selection. This method can be seen as a combination of step-up (forward selection) and step-down (backward elimination). Indeed, the step-wise method doesn't necessarily keep the parameters in the model once they have been added; each time a parameter is selected there is the possibility than one or more parameters in the model are taken out. It is like running forward selection, and after each step running backward elimination. As a criterion for the selection we used AIC[7], since it gives the best trade-off between simplicity and goodness-to-fit, while reducing potential overfitting.

[5] In the graphs, columns drawn above the 0 represent patients who didn't have a disease, while the columns above 1 represent the ones who did;
[6] See conclusions paragraph;
[7] Akaike Information Criterion, it is an estimator of prediction error widely used in model selection.

The second method is LASSO, that penalizes the l_1 size of the parameters. To find the parameter λ, we used cross validation and decided to build two different models: one with `lambda.min`, that is the value that gives the least cross-validated error, and one with `lambda.1se`, that gives the most regularized model, such that aforementioned error is within one standard error of the minimum.

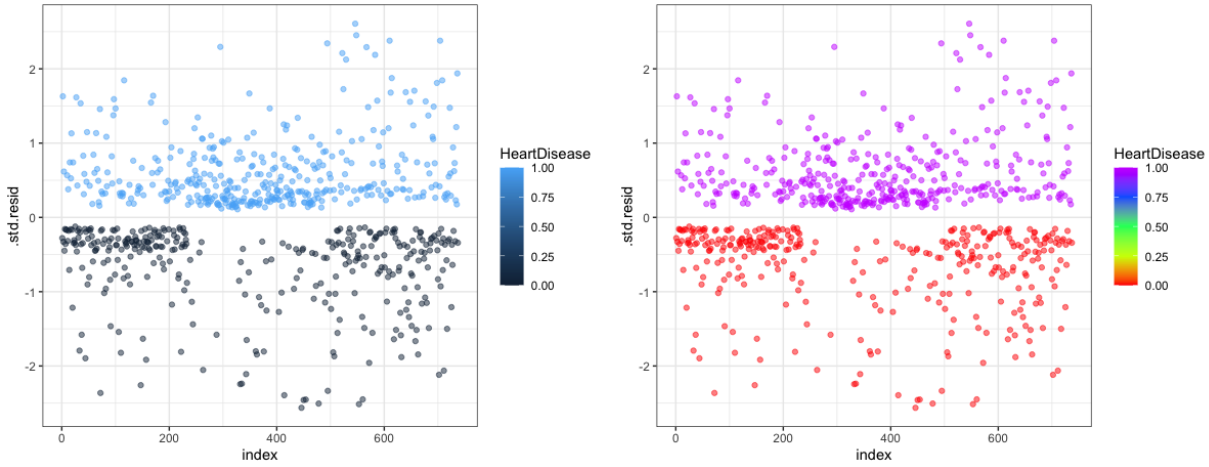
Results

Before explaining and comparing the results, we introduce the methods that were used. Once we built the models, we used them to get the probability for each of the patients in the test set to get a Heart Disease; then we converted all probabilities greater than 0.5 to 1 and all the others to 0, in order to make them comparable to the actual ones, and to compute the accuracy.

The full model, so the one with all 15 covariates, had an accuracy of 90%, meaning that for almost 89% of the test patients our model managed to identify if they had or hadn't an heart disease. The McFadden's pseudo-R^2 was equal to 0.51, that signals a very good fit of the model[8]. One thing to notice is that using the `vif()` function, used to analyze multicollinearity, emerged that the `ST_Slope` variables have a VIF score of about 4[9], while the others are very close to 1. Nonetheless, we decided to keep both covariates, since moderate correlation should not cause multicollinearity problems.

Passing to model selection, the step-wise method selected 11 covariates, dropping: `RestingBP`, `RestingECG Normal`, `RestingECGST`, `MaxHR`. An interesting thing to notice is that since the selection was brought out using AIC instead of the p-values, the `Age` has a p-value greater than 0.05 but wasn't ruled out. Despite the fact of having less covariates, this model had an accuracy of 91%, thus a bit higher than the full model and a lot more simple. Also for this model we had the same multicollinearity concern, that was treated in the exact same way as before. The McFadden's pseudo-R^2 was, as in the full model, equal to 0.51, thus implying a good fit. In this model, as in the full one, the p-value related to the F-test is a value really close to 0, and thus satisfies the condition <0.05, leading to the conclusion that in both cases model is better than no model.

Together with multicollinearity, we also checked the assumption of not having any influential outlier. Influential values are extreme individual data points that can lower the quality of the logistic regression model. The extreme values can be seen by plotting the Cook's distance values. To check if any of the outliers are influential, we inspected the standardized residual error. Generally, an absolute standardized residuals above 3 means that the data point should be removed.



(a) Model (b) Stepmodel

Figure 3: Plot of the standardized residual error

As clear by Figure 3, and confirmed by an algorithmic check, no value is greater than 3 or lower than -3,

[8]Notice that anything above 0.2 for the McFadden's R^2 is considered a good fit.
[9]For reference, a VIF score of 1 means no correlation, 1-5 means moderate correlation, > 5 high correlation.

and thus we don't have to worry about influential outliers.

Lastly, we analyze the two LASSO models that we built. The first one, using `lambda.min`, generated a model with 13 out of 15 covariates, ruling out `RestingECGNormal` and `RestingECGST`. This model had an accuracy of 89.5%, and a pseudo-R^2 of 0.51, so identical to the previous two case. The second time, we used `lambda.1se`, and the model didn't include 3 covariates: `RestingECG Normal` and `RestingECGST`, as the other LASSO model, plus `RestingBP`. This time we achieved an accuracy of 89%, and a pseudo-R^2 of 0.48. What is evident, is that using LASSO penalization, the cost of removing a covariate is a bit higher than the previous proposed model, and the fitting remains almost unchanged.

A final remark to make is about the analysis of the residuals. Unlike linear regression, logistic regression does not require normality and homoscedasticity of the residuals. Nevertheless the plots of the residuals can be found in the appendix, for both the full model and the one built with step-wise selection.

Conclusion

In conclusion, we are overall satisfied with the results obtained by our models, both in terms of accuracy and goodness-to-fit. Between the tested models the best one is surely the one obtained using the step-wise selection method, since it gave the best accuracy, best fitting value, and also removed the most covariates, making it the simplest. A further development of this paper could be to check how this values change when we do a selection based on p-values instead of AIC, and how step-up e step-down perform with respect to it. Moreover, in order to possibly improve the obtained results, one could remove the variables that showed some correlation, in order to reduce multicollinearity.

Even if the results of our study are quite satisfactory, we need to consider that there might be some limitations due to some assumptions and features of our dataset. Firstly, as we briefly mentioned in the introduction, the population considered in our study comes mainly from three different countries: Hungary, Switzerland and USA. As different countries have different lifestyles, it could be interesting to enlarge our research to as many nations as possible, firstly to understand more generally the statistical behavior of heart diseases, but also to study how different habits and customs can affect them.

Moreover, as mentioned before, the sex of the participants is far from being homogeneously distributed between men and women. This might result in a misleading high correlation between being a man and developing a heart disease.
On the other hand, however, it is empirically proved that male patients have a higher probability of developing such a condition [3]. Although the scientific and medical causes of this phenomenon are still to be cleared up, studying how gender quantitatively affects our statistical analysis could be an interesting starting point to a further development of this research.

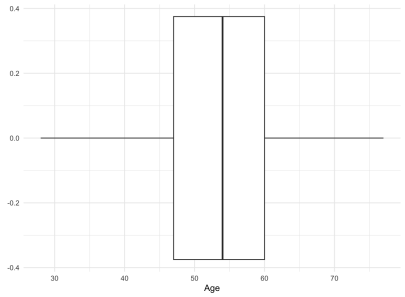
References

[1] F. Bijma, M. A. Jonker, A. van der Vaart, "*An Introduction to Mathematical Statistics*", Amsterdam University Press (2017)

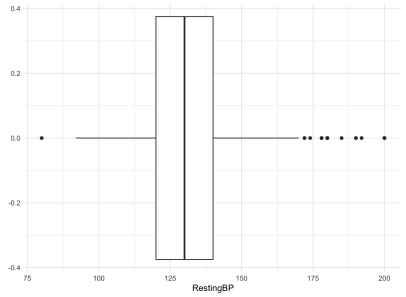
[2] F. Soriano, "*Heart Failure Prediction Dataset*" (2021), `https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction`

[3] G. Albreksten, I. Heuch, M. Lochen et al., "*Lifelong Gender Gap in Risk of Incident Myocardial Infarction, The Tromsø Study*", Jama Internal Medicine Research (2016)

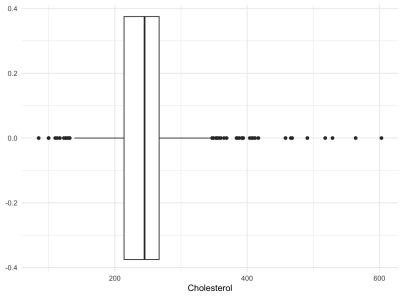
Appendix - Plots



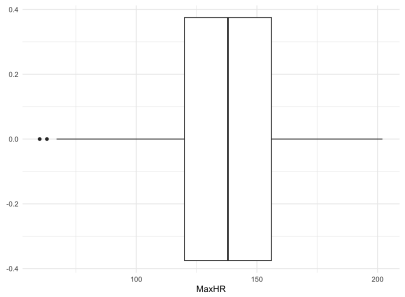
(a) Age distribution



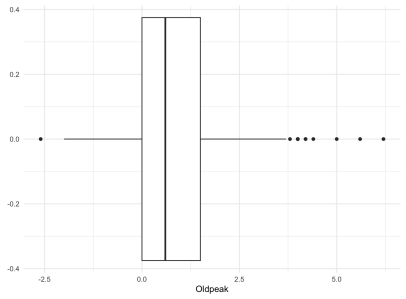
(b) Blood Pressure distribution



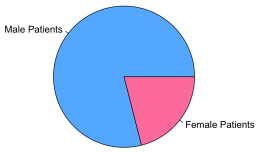
(c) Cholesterol distribution



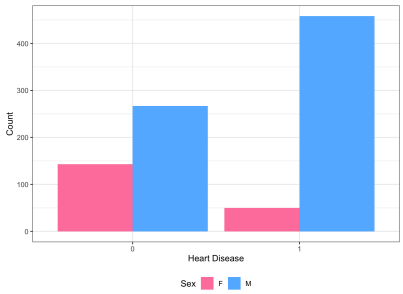
(d) Maximum Heart Rate achieved distribution



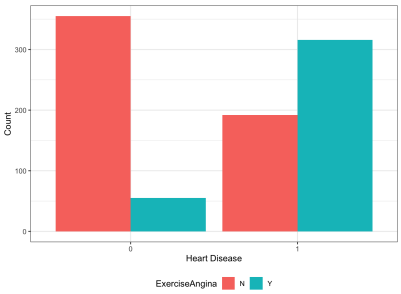
(e) ST Depression (`Oldpeak`) distribution



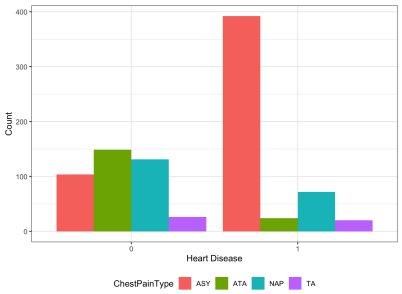
(f) Male VS Women Partecipants



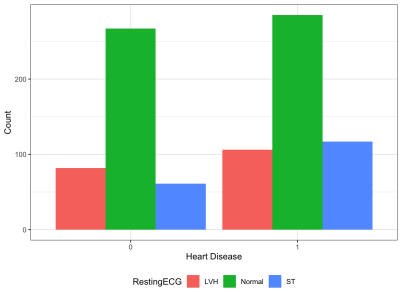
(g) Heart Disease across Gender



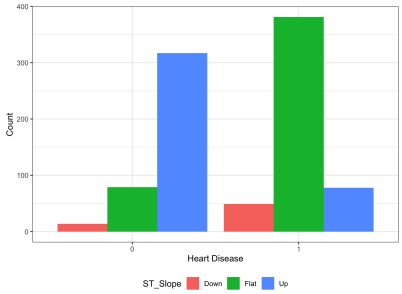
(h) Heart Disease across Exercise Induced Angina



(i) Heart Disease by Chest Pain Type



(j) Heart Disease by ECG at rest



(k) Heart Disease by ST Slope

(l) Correlation Plot

Appendix - Results

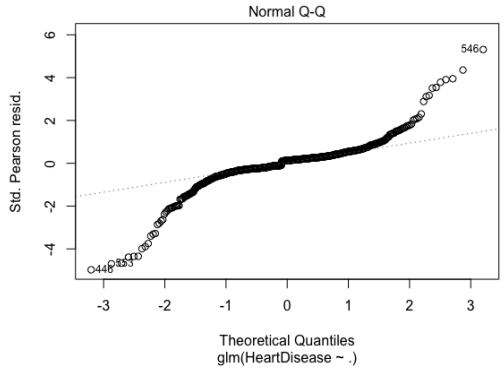
```r
model_data <- augment(model) %>%
  mutate(index = 1:n())

model_data %>% top_n(3, .cooksd)
ggplot(model_data, aes(index, .std.resid)) +
  geom_point(aes(color = HeartDisease), alpha = .5) +
  theme_bw()

model_data %>%
  filter(abs(.std.resid) > 3)
```

(m) Model

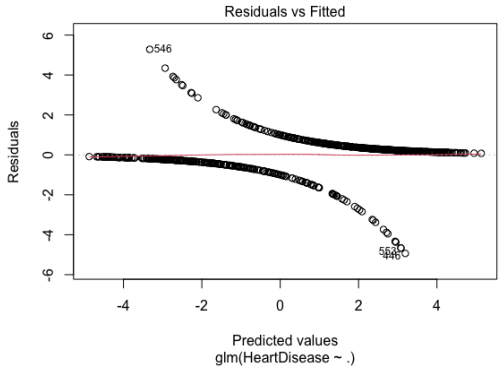
```r
step_model_data <- augment(step_model) %>%
  mutate(index = 1:n())

step_model_data %>% top_n(3, .cooksd)

ggplot(model_data, aes(index, .std.resid)) +
  geom_point(aes(color = HeartDisease), alpha = .5) +
  theme_bw() +
  scale_color_gradientn(colours = rainbow(5))

step_model_data %>%
  filter(abs(.std.resid) > 3)
```

(n) Stepmodel

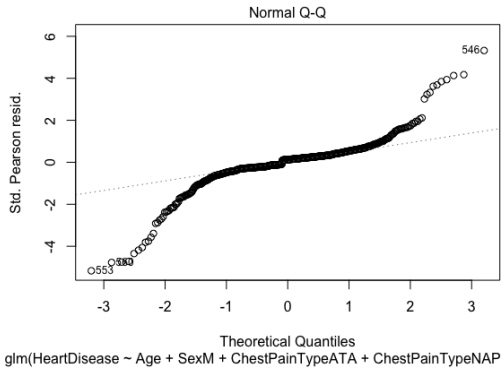
Figure 4: Snippets of the code to plot the graphs of the standardized residual error. The filter function didn't give any result, confirming that there are no data points with absolute distance >3



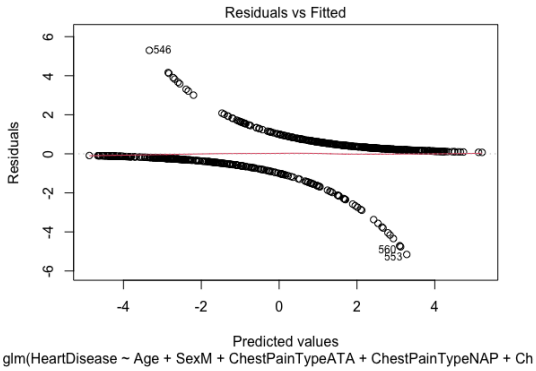
(a) Q-Q plots of the residuals for the model



(b) Residuals vs Fitted for the model



(c) Q-Q plots of the residuals for the step-wise model



(d) Residuals vs Fitted for the step-wise model