Predictive Data Analysis on Clinical Data Using Generalized Linear Models (Logistic Regression & LASSO)

Arman Jahangiri
Mojtaba Kanani Sarcheshmeh
David Yang

February 11, 2024

Contents

1 Introduction

Heart disease continues to be the leading cause of death worldwide, thereby necessitating a better understanding of risk factors involving its occurrence. In attempts to describe the relationships between various risk factors and the likelihood of heart disease episodes, regression analysis is crucial for developing effective predictive models for heart disease. In our project, we utilize the "Heart Disease Dataset" from the public repository, Kaggle, to perform regression analysis for heart disease.

1.1 Data

The heart disease prediction data-set that we leveraged in this project consists of clinical data collected from patients undergoing diagnostic testing heart disease in the 1980s from four different institutions including the Cleveland Clinic in Cleveland USA, University Hospitals in Zurich and Basel Switzerland, Hungarian Institute of Cardiology in Budapest Hungary, and Veterans Administration Medical Center in Long Beach USA. This dataset contains 1026 samples collected via convenience sampling. Each sample comprises of 14 different attributes including both categorical and continuous predictors and a binary target response indicating the diagnosis of heart disease (Table 1). This comprehensive set of variables makes this data-set an invaluable resource for exploring the association between patient risk factors and the presence of heart disease.

Table 1: Description of Variables

Variable	**Explanation**	**Min**	**Max**
Age	Age of the patient	29	77
Sex	(1 = male, 0 = female)	0	1
cp	Chest Pain Type	0	3
trestbps	Resting Blood Pressure	94	200
chol	Serum Cholesterol	126	564
fbs	Fasting Blood Sugar (= 1 if ≥ 120)	0	1
restecg	Resting Electrocardiographic Results	0	2
thalach	Maximum Heart Rate Achieved	71	202
exang	Exercise-Induced Angina	0	1
oldpeak	ST Depression Induced by Exercise	0	6.2
slope	Slope of Peak Exercise ST Segment	0	2
ca	Number of Major Vessels	0	4
thal	Thalium Scintigraphy	0	3
target	Diagnosis of Heart Disease	0	1

This dataset was retrieved on Kaggle as a comma-separated values file (heart.csv). We selected this data-set for our project as it contained a diverse range of categorical and continuous variables that are important for creating a meaningful generalized linear model (GLM) for predicting heart disease. In addition, the dataset was also highly upvoted for usability and documentation on Kaggle further affirming its usability in regression analysis.

1.2 Objectives

The main objectives in our project is described below:

1. Our main objective in this project was to **develop a generalized linear model** for predicting the probability of heart disease given the relevant patient characteristic.
    - Since the response variable in the data-set is binary, we intuitively employed logistic regression.
    - Since this data-set is widely used for heart disease prediction, one of our primary focuses in this study was to obtain the best logistic regression model for the data-set.
2. Our second area of interest in this project involves **analyzing the log-odds ratio of each relevant variable** such as sex, age, and blood pressure for heart disease.
    - This is not only important for creating accurate regression modelling but also important for fostering advancements in preventative healthcare as it allows us to pinpoint select clinical features that are strongly predictive of heart disease diagnosis.
3. Finally, we are interested in exploring the associations and **correlations between various variables** in the data-set.

Ultimately, answering these questions in our study will enhance our understanding of the disease's determinants and also provide a valuable tool for early detection and intervention of heart disease.

2 Methods

2.1 Data pre-processing

Renaming Categories for Categorical Variables : To enhance the interpretability and clarity of our model, we addressed categorical variables that had been encoded numerically. Recognizing the importance of informative model coefficients and facilitating ease of interpretation, particularly in the visualization phase, we undertook the task of renaming these numerically encoded categorical variables with their corresponding descriptive strings.

Given that the initial numerical representation of categories lacked meaningful information, we conducted an in-depth analysis of the primary data sources, including the UCI and Kaggle pages. Drawing insights from these sources, we meticulously reassigned names to the categories, aiming to imbue them with greater informativeness and readability. This strategic renaming process not only contributed to the overall understanding of the model's outputs but also significantly improved the interpretability of visualizations associated with these categorical variables.

The table below provide a summary of the changes made:

Variable	**Previous Category Name**	**New Category Name**
Sex	0	female
Sex	1	male
cp	0	asymptomatic
cp	1	non-anginal pain
cp	2	atypical angina
cp	3	typical angina
fbs	0	<120mg/dl
fbs	1	>120mg/dl
restecg	0	normal
restecg	1	abnormality
restecg	2	hypertrophy
exang	0	No
exang	1	Yes
slope	0	upsloping
slope	1	flat
slope	2	downsloping
thal	1	normal
thal	2	fixed defect
thal	3	reversable defect
target	0	No disease
target	1	Disease

Table 2: Changes Made in the Names of Categories

Missing Values : We carefully looked through the data to see if any information was missing or if there were any "Not Available" (NA) values. In our study, which focuses on predicting heart disease, we're happy to report that there were no such gaps or missing pieces of information in our dataset. This is good news because it means that we have a complete set of data for each person in the study, with all the details we need. Having a complete dataset makes it easier for us to analyze and ensures that our predictions are based on solid information.

Duplicate Values : This process of handling duplicates is crucial for maintaining the accuracy and reliability of our data analysis. Duplicate entries can introduce biases and distort the results, leading to inaccurate conclusions. By systematically identifying and removing these duplicate instances, we ensure that each unique combination of values contributes appropriately to our analysis, preventing any unintentional skewing of results. This data refinement step enhances the overall quality of our dataset, setting the stage for more accurate and dependable insights as we delve into predicting heart disease. Removing duplicates is a common practice in data preprocessing, allowing us to work with a cleaner and more representative dataset. We took a close look at our dataset to identify and deal with any duplicate entries. Specifically, we found and removed a total of 723 duplicate rows. In our context, duplicates were determined by considering entire rows where all 14 columns contained precisely the same values.

Erroneous Values : We conducted a thorough examination of our dataset to identify and eliminate rows with erroneous entries, resulting in the removal of 25 observations. The criteria for identifying these inaccuracies included instances where values for categorical variables deviated from the expected grouping associated with their respective variables. In practical terms, this involved scrutinizing categorical variables and ensuring that their values aligned with the predefined categories. Any observations presenting values that did not correspond to the expected categories were considered erroneous and subsequently removed from the dataset. This proactive approach to data cleaning is essential for maintaining the integrity of our analysis, as inaccurate entries could potentially lead to misinterpretations and compromise the reliability of our findings.

Outlier Values : In the visualization phase, we identified outlier values. We removed one outlier observation for the cholesterol variable which was 564.
Following the completion of the data cleaning process, our dataset was refined to include a total of 295 observations (n = 295). This dataset was included in the D2L submission as heart-process.csv.

Z-score standardization : We applied Z-score standardization to all six continuous variables in our dataset, namely ”age” ”trestbps” ”chol” ”thalach” ”oldpeak” and ”ca” This standardization technique involves transforming the variables to have a mean of 0 and a standard deviation of 1. The purpose of this standardization was threefold:

1. Enhancing Numerical Stability: Standardizing variables contributes to improved numerical stability in the computations involved in regression analysis. By scaling the variables to a common scale, we mitigate issues related to the numerical precision of calculations, promoting more robust and reliable results.

2. Mitigating Multicollinearity Impact: Standardization assists in reducing the impact of multicollinearity, a phenomenon where independent variables in a regression model are highly correlated. Multicollinearity can introduce instability and inflated variance in regression coefficients. Standardizing variables helps alleviate this issue by ensuring that all variables are on a comparable scale, making it easier to discern the individual effects of each predictor.

3. Importance in Regularization Techniques: Scaling becomes crucial when employing regularization techniques such as Ridge or Lasso regression. In these techniques, the penalty term applied to the regression coefficients is influenced by the scale of the variables. By standardizing the variables, we ensure that the regularization penalty is applied uniformly, preventing any particular variable from dominating the regularization process solely based on its scale. This ensures a fair and effective application of regularization, leading to a more reliable and interpretable model.

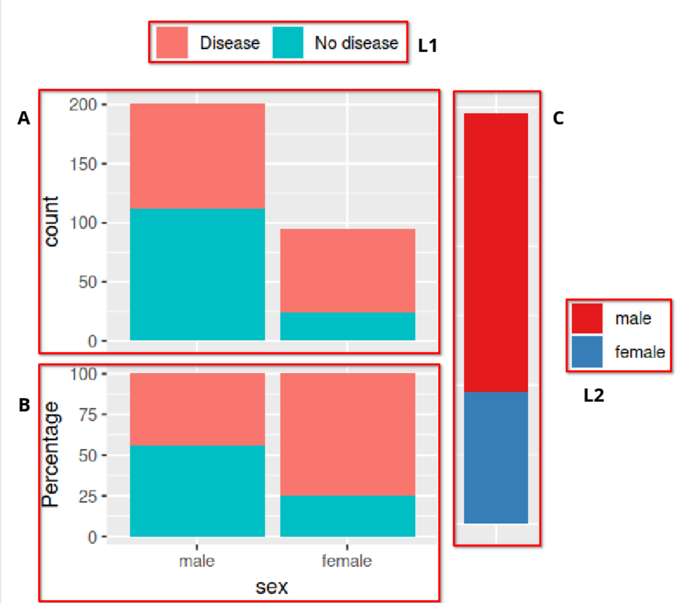
In summary, the Z-score standardization of continuous variables not only enhances the numerical stability of regression analysis but also addresses issues related to multicollinearity and ensures the effectiveness of regularization techniques, contributing to the overall robustness of our analytical approach.

2.2 Data visualization

Data visualization plays a crucial role in enhancing our understanding of complex datasets by providing a visual representation of patterns, trends, and relationships. It transforms raw data into accessible and meaningful insights, facilitating better decision-making and communication of findings.

Choosing the right plot is paramount in effective data visualization. The type of plot selected should align with the nature of the data and the insights intended for communication. For instance, a scatter plot might be ideal for showcasing relationships between numerical variables, while a bar chart is effective in comparing categorical data. The choice of plot significantly influences how well the audience comprehends and interprets the information presented. That is why we decided to design different plots for categorical and numerical variables.

Categorical Variables : To gain a more comprehensive understanding of the categorical variables, we opted to employ bar plots as our primary visualization method. Bar plots are inherently clear and easy to interpret. They use the length or height of bars to represent the frequency or proportion of each category, making it straightforward for viewers to grasp the distribution of the binary outcome across different categorical predictors. For these plots, Instead of employing a single bar plot we opted for a more comprehensive approach by stacking 3 distinct bar plots together. This strategy was implemented to thoroughly capture and address all the underlying information associated with each of the categorical variables.
Here is the visual representation illustrating information for the ”sex” variable. This example is provided to elucidate the interpretation of each component within the plot.



Examining Legend 1 (Labeled L1), you can discern that the representation of patients diagnosed with the disease is indicated by the salmon color, while information pertaining to patients not diagnosed with the disease is denoted by the sky blue color. It is important to note that we used consistent color coding for the outcome variable across all of our plots, Consistent color coding establishes a visual coherence throughout the plots. This cohesion helps viewers associate specific colors with distinct classes, creating a unified and easily recognizable representation.

- The top left chart (labeled A) is a count-based chart showing the absolute number of individuals with and without the disease, differentiated by sex. The salmon color portions represent individuals with the disease, while the sky blue portions represent those without the disease. From this chart, it's evident that there are more males with the disease than females, and also more males without the disease compared to females.
- The bottom left chart (labeled B) is a percentage-based chart showing the proportion of individuals with and without the disease relative to the total number of individuals within each sex. The proportions are represented similarly with salmon for the disease and sky blue for no disease. This chart suggests that a higher percentage of females are diagnosed with the disease compared to males when considering the proportion within each sex.
- The chart on the right (Labeled C) illustrates the count of observations for each category of the "sex" variable namely "male" and "female", allowing direct comparison. It is apparent that the number of "male" observations is nearly double that of "female" observations.

For interpreting these charts, our primary focus is on the bottom left chart (Labeled B), which emphasizes the percentages of disease or no disease in each category. This approach mitigates the impact of the number of observations in each group, enhancing the interpretability of the information.

Numerical Variables : When plotting numerical predictors against a binary outcome, several important aspects should be considered to ensure the clarity, accuracy, and interpretability of the visualization:

- Distribution Comparison : Understanding how the values of the numerical predictor vary between the two outcome groups provides insights into potential relationships.
- Central Tendency Differences : A noticeable difference in central tendency (Mean or Median) can suggest a potential association with the binary outcome.
- Consideration of Outliers : Outliers can disproportionately influence summary statistics, affecting the interpretation of the relationship between the predictor and the binary outcome.
- Appropriate Scaling : We should ensure that the axes are appropriately scaled to avoid distortion in the representation of the data. Inconsistent scaling can lead to misinterpretation of the plotted information.

Taking into account all these considerations, we opted to utilize the following types of plots to effectively represent the information contained in our numerical variables.

1. Violin Plot : We'll employ side-by-side half violin plots for each class of the target variable. This approach facilitates the comparison of observation distributions for each variable concerning their target class. Additionally, we can examine the mean of each distribution as a central tendency metric and compare them across target classes. The presence of long tails in violin plots can highlight potential outliers. To ensure comparability among different variables, we'll use their z-score standardized values instead of the original ones in the creation of distinct violin plots.
2. lollipop plot : It is used to delve deeper into the central tendency metrics of the scaled values for each numerical variable.
3. Parallel Coordinates Plot : It efficiently visualizes complex relationships among multiple numerical variables, enabling the simultaneous exploration of trends, patterns, and outliers
4. Pairs Plot : It provides a concise visual summary of pairwise relationships between numerical variables in a dataset, allowing for the identification of correlations, patterns, and potential outliers

2.3 Model selection

The model selection process performed in this study was an exhaustive and comprehensive process as one of the main objectives in this study was to obtain the **best** logistic regression model for predicting heart disease. Hence, in this portion of the project, we generated twelve total logistic regression models with different combinations of variables and two-way interaction terms. Afterwards, the best model was selected based on the comparison of various metrics that portray the model's fit for the data.

2.3.1 Model generation

Since the response variable in the GLM is the binary outcome, "Diagnosis of heart disease", intuitively **logistic regression** was employed. In logistic regression, the logit link function is typically used, however in certain instances, the use of the probit link function is better. Hence, in order to compare both link functions, the data was fitted with both the logit and probit link functions using all 13 predictors found in the data-set. Ultimately, we decided to use the **logit link function** when generating the logistic regression models in the later stages of the experiment. Before perform variable selection, an overall significance test for all predictors was first performed to verify that all coefficients were not equal to zero. This was done by comparing our full model consisting of all 13 predictors to the minimal model (intercept model) in a likelihood ratio test with $\alpha = 0.05$. In the test, the null hypothesis is that $\beta_j = 0$ for $j = 1, ..., 13$. The test statistic was defined as follows:

$$C = 2\left[l(\beta) - l(\beta_{min})\right] = Null\ deviance - Residual\ deviance$$

Note that, $C \sim \chi^2_{18}$ as the degrees of freedom is equal to the difference in the number of parameters in the regression models (some variables were dummy encoded).

After performing the overall significance test, the model generation process was partitioned into three main processes:

1. Model generation from the Full Model
2. Model generation from the Reduced Model
3. Model generation with Lasso Regression

These processes are explained in greater detail below.

1. Model generation from the Full Model:

The **full model** in the experiment was defined as the logistic regression model (logit link function) with all 13 predictors found in the in the data-set. In part 1 of model generation, we built various logistic regression models from the full model.

In order to explore the possibility of exponential relationships between the continuous variables in our data-sets with the the response outcome, we fitted various full models that included the quadratic term for our the continuous variable of interest. For example, in the case of the continuous variable ”ca”, our fitted model contained all 13 predictors plus the ca^2 term. Next, we used the Z-test for the quadratic term to see if it is signficant for our data-set. If so, we continued the process for higher power terms until there is no longer any significance in adding any exponential terms into our fitted logistic regression model. Through this process, we obtained **Model 1**, which includes all 13 predictors plus the term ca^2.

Next we explored which two-way interaction terms to include in our logistic regression model. In the first part, we added two-way interaction terms to Model 1 which aligned with current scientific literature. Adding these interaction terms to model 1 created **Model 2**. For example we found that:

- Sex was correlated resting blood pressure (Maranon and Reckelhoff, 2013), serum cholesterol levels (Fletcher, 2023) and fasting blood sugar levels (Ciarambino et al., 2022).
- Resting blood pressure was correlated with fasting blood sugar levels (Lv et al., 2018) and serum cholesterol levels (Sakurai et al., 2012).

In the second part of two-way interaction term selection, we used the backward step-wise selection function in R, step(), with the Akaike information criterion (AIC) criterion on the full model with all possible two-way interaction terms to obtain **Model 3**.

Next, we used the forward step-wise selection function in R, step(), with the AIC criterion to obtain **Model 4** and **Model 5**. For the step() function, we specified the maximum model as the full model with all possible two-way interaction terms. For **Model 4**, the starting point that we specified is Model 1, whereas for **Model 5** the starting that was the intercept model with only the intercept term.

2. Model generation from the Reduced Model:

The **reduced model** in the experiment was defined as the logistic regression model (logit link function) with all insignificant (Z-test test: $\alpha = 0.05$) coefficients removed from the full model. The reduced model was then used to build various logistic regression models. The reduced model included the following six predictors: sex, cp, trestbps, slope, ca, and thal.

First, exponential relationships between continuous variables in the reduced model the response outcome was investigated. The same approach was used as previously where exponential terms were added onto the reduced model and kept if they were significant according to the Z-test test. Through this process, we obtained **Model 6** which consisted of the six predictors in the reduced model plus the quadratic term ca^2.

In order to explore the effect of two-way interaction terms on our logistic regression model, we first created **Model 7** which consisted of all the predictors in Model 6 and also all possible two-way interaction terms for those predictors. Next, we explored the effect of the intercept on our logistic regression model by removing the intercept from Model 7 to form **Model 8**. Similar to the previous steps performed on the full model, we also used the backward and forward step-wise selection function, step(), in R with the AIC criterion to select which two-way interaction terms to retain in our model. We performed backward step-wise selection on Model 7 to obtain **Model 9**. On the other hand, we performed forward step-wise selection from the intercept model and Model 6 to obtain the same model, **Model 10**. In the forward step-wise selection process, we specified the maximum endpoint model as Model 7. Finally, to obtain **Model 11**, we added two-way interaction terms that aligned with current scientific knowledge. This included the correlation between sex and resting blood pressure.

3. Model generation with Lasso Regression: (Model 12)

3.1. Introduction to LASSO

Lasso (Least Absolute Shrinkage and Selection Operator) regression is a regularization technique commonly used in linear regression to handle high-dimensional datasets. It introduces a penalty term based on the absolute values of the regression coefficients, encouraging sparsity in the model. This makes Lasso particularly useful for feature selection when dealing with datasets with many features.

3.2. Mathematical Formulation

The standard linear regression model aims to minimize the sum of squared differences between predicted and actual values. Lasso extends this by adding a regularization term:

$$\text{minimize}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right)$$

When λ is zero, Lasso reduces to standard linear regression. As λ increases, the penalty for non-zero coefficients also increases. This encourages some coefficients to become exactly zero, effectively performing feature selection.

3.3. Properties

Lasso regression has several key properties:

- **Feature Selection:** Lasso tends to shrink the coefficients of less important features to exactly zero, effectively performing automatic feature selection .This is why LASSO is chosen over other techniques such as Ridge or elasticnet (combination of LASSO and Ridge).
- **Sparsity:** As the regularization parameter λ increases, the number of non-zero coefficients decreases, leading to a sparse model.
- **Trade-off:** The choice of λ involves a trade-off between fitting the training data well and keeping the number of selected features to a minimum.

3.4. LASSO combined with Cross Validation

In the lasso technique, the optimal value of the regularization parameter, λ, is typically chosen through a process called cross-validation. Cross-validation involves partitioning the dataset into multiple subsets, training the model on some of these subsets, and then evaluating its performance on the remaining subsets. This process is repeated several times, and the average performance is used to assess how well the model generalizes to new data.

```
Data: $X$ (features), $y$ (target variable)
Result: Optimal $\lambda$ and coefficients $\beta_j$
for $\lambda$ in a range of candidate values do
    Train Lasso regression model with regularization parameter $\lambda$;
    Compute performance on validation set;
end
Identify the optimal $\lambda$ that minimizes validation error;
Train Lasso regression model on the entire dataset with the optimal $\lambda$;
```

Algorithm 1: Lasso Regression with Cross-Validation

3.5. LASSO and CV on our Dataset

We made use of lasso regression technique to perform variable selection on the data with all the 13 variables, ca^2, and all of their corresponding two-way interaction terms; we named it **(Model 12)**

The Cross-Validation method offered $\lambda = 0.06611$ to be optimal, which results in 18 nonzero coefficients, and a Standard Error of 0.04026.

The plot of $\log(\lambda)$ against the deviance in Lasso regression illustrates the trade-off between model complexity and goodness of fit. As λ increases, the deviance typically rises, reflecting increased regularization and model simplicity. The optimal λ is often identified at the point where the deviance is minimized, striking a balance between model sparsity and predictive performance.

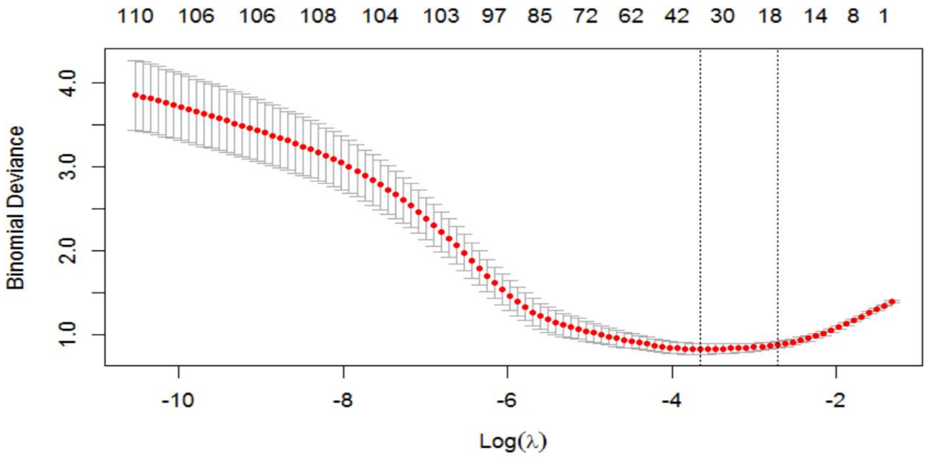


Figure 1: Lasso Regularization Path: log(λ) vs. Deviance

2.3.2 Model diagnostics

To gauge the performance of each model that we obtained in the model generation process, we calculated the following metrics for each model:

1. **Residual deviance** $D(Y, \hat{\mu})$: The residual deviance was either directly obtained from the GLM in R using deviance(model), or calculated by:

$$D(Y, \hat{\mu}) = \sum_{i=1}^{n} (r_i^D)^2$$

where r_i^D is the deviance residuals. Note that $D(Y, \hat{\mu}) \sim \chi^2_{n-(q+1)}$ where n is the number of observations and q is the number of predictors in the model. With this, we can calculate the p-value for the deviance where H_0 is that the model fits the data well.

2. **Akaike information criterion** (AIC): The AIC was either obtained directly in R with AIC(model), or calculated by:

$$AIC = 2k - ln(\hat{L}) = 2k + D(Y, \hat{\mu})$$

where k is the number of parameters in the model and $\hat{L}$ is the maximum value for the likelihood function for the model.

3. **McFadden's Pseudo-R^2:** This metric tells how well the model explains the variation in the data. It was calculated by:

$$R^2 = 1 - \frac{residual\ deviance}{null\ deviance}$$

4. **Likelihood ratio test** statistic: The likelihood ratio test statistic was calculated by:

$$C = 2[l(b) - l(b_{min})] = Null\ deviance - Residual\ deviance$$

where $l(b)$ is the likelihood of the model of interest and $l(b_{min})$ is the likelihood of the minimal model containing the intercept only. Note that $C \sim \chi^2_q$ where q is the number of predictors in the model of interest. With this, we can calculate the p-value for the likelihood ratio test where H_0 is that all coefficients except for the interest in the model of interest is equal to 0.

5. **Mean-squared error** (MSE) for prediction: The mean squared prediction error was either directly obtained through cross-validation with 10 folds via the cv.glm() function from the R library *boot*, or it was calculated by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where $\hat{Y_i}$ is the predicted outcome from the logistic regression model. Note that for the MSE calculated via cross-validation, the average of ten iterations was calculated.

2.4 Final model

As of result of the favourable metric calculations for Model 5, we ultimately selected Model 5 as the best model in our experiment. In this part of the study we obtained the 95 percent confidence intervals for our coefficient estimates using the confint() function in R as well as the fitted coefficient estimates using the coef() function. We also checked generalized linear model assumptions using residual plots for Model 5.

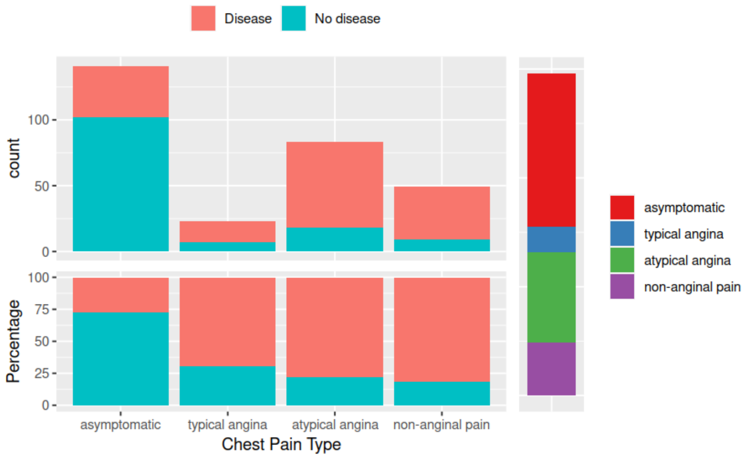
3 Results

3.1 Data visualization

In this section, we'll highlight key findings from our visualizations, featuring only the essential charts, while the remaining visuals are provided in the appendix to streamline the content.

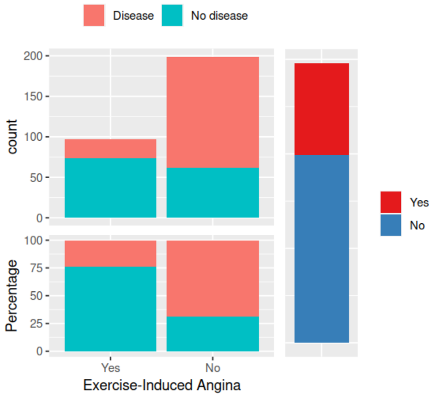
Categorical Variables

- Chest Pain Type



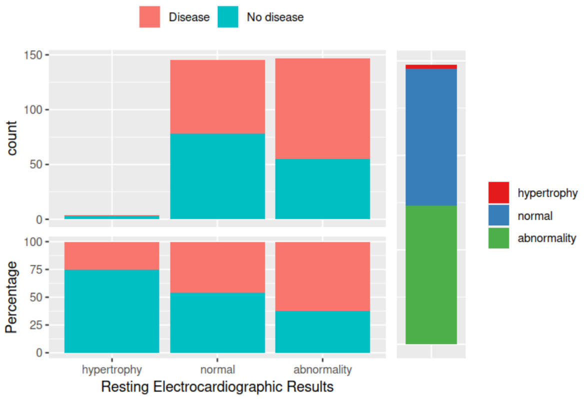
This plot reveals that the percentage of patients diagnosed with the disease is lowest among individuals without any symptoms of chest pain, followed by those with typical angina, atypical angina, and non-anginal pain.

- Exercise Induced Angina



This plot indicates that the percentage of patients diagnosed with the disease is lowest among those who experienced exercise-induced angina.

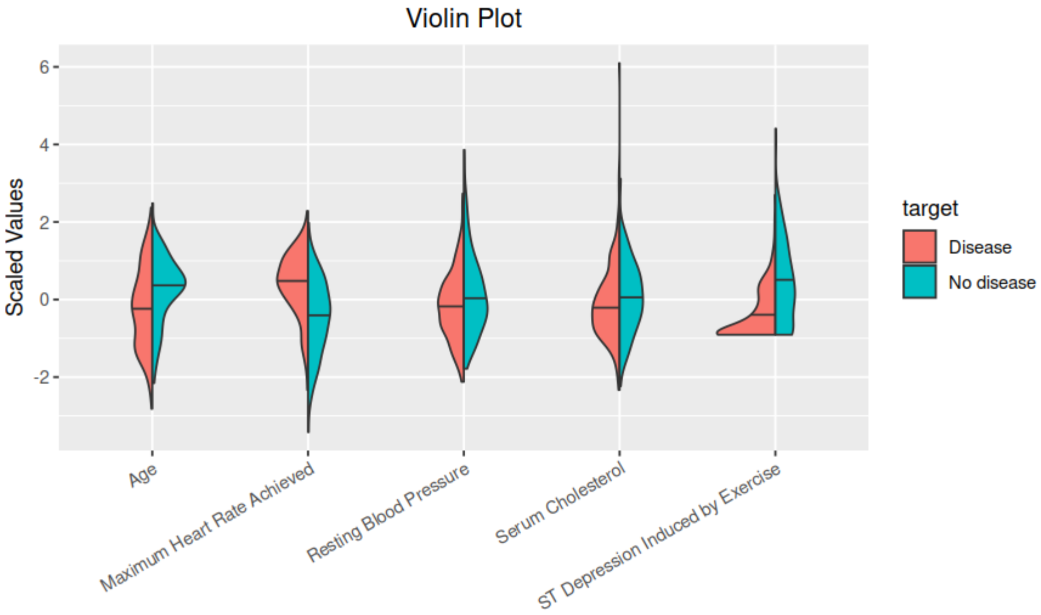
- Resting Electrocardiographic Results



In this visual representation, it becomes evident that the percentage of patients diagnosed with the disease is lowest among those whose resting electrocardiographic results indicate hypertrophy, followed by individuals with normal and abnormal results. However, it's important to note that the number of observations for individuals with hypertrophy is significantly lower than the other two categories, indicating that the diagnostic percentage for this category may not be reliable. Excluding this category, the percentage of patients diagnosed with the disease is lowest among individuals with normal resting electrocardiographic results, followed by those with abnormal results.

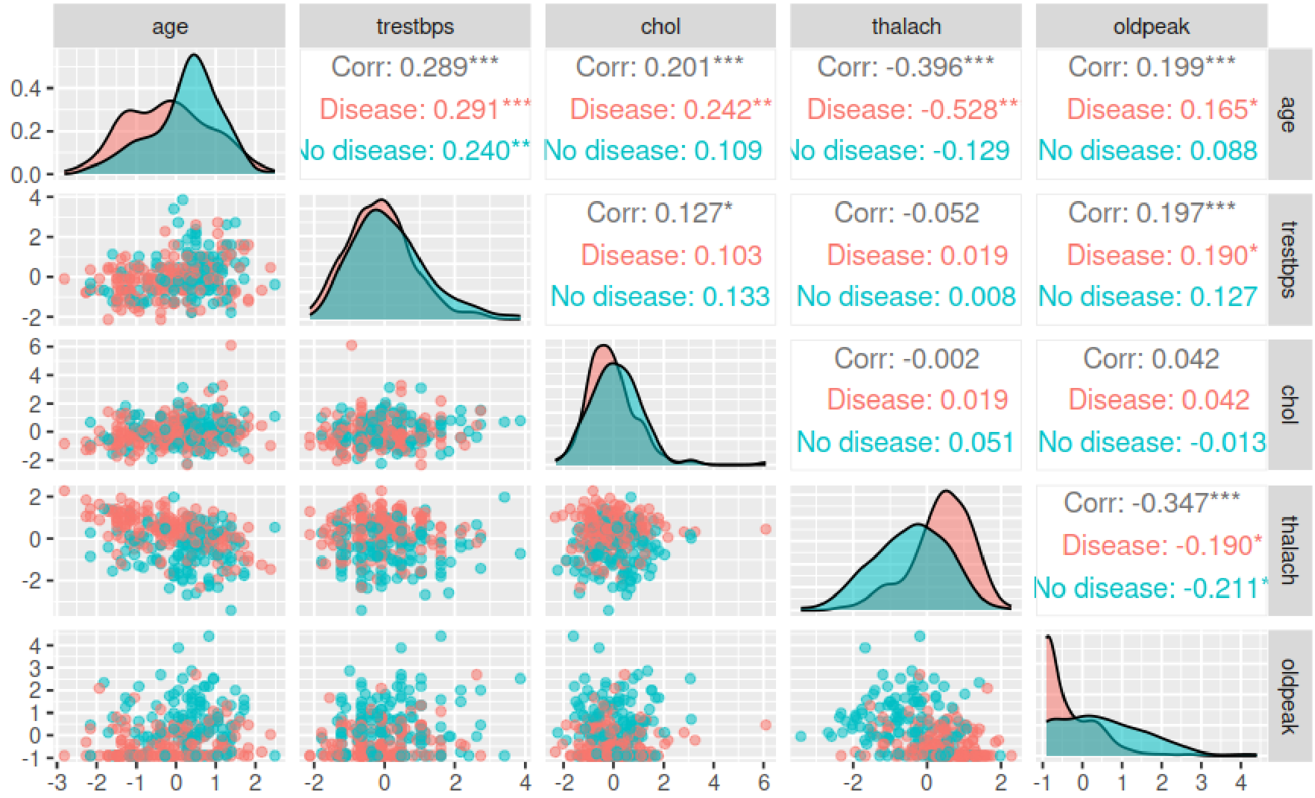
Numerical Variables

- Violin Plot



The extended tail in the Serum Cholesterol distribution suggests the presence of a potential outlier in this variable, warranting further investigation.

- Pairs Plot



presence of a possible outlier in serum cholesterol is also visible in the scatter plots. Upon a thorough examination of the dataset, a decision was made to exclude this observation reducing the number of our observations to 295. the observation was for a patient with cholesterol level of 564. Given that the mean for this variable is 247 with a standard deviation of 52, it is evident that this observation significantly deviates from the typical range. By removing this outlier, our dataset becomes more homogenous, allowing our models to better capture the general patterns in the observations without being unduly influenced by extreme values.

It is observed from the data that %68 of the patients in the sample were males and the average age among all the patients was 54.52. The average resting blood pressure was 131.60, which is higher than the normal (120). This is not surprising since the sample has been taken from the people who came to the hospital with heart-related symptoms However, the average cholesterol level is 247.16 with a standard deviation of 51.98. This is noticeable since a normal cholesterol level should be less than 200 mg/dL. A total cholesterol level of 200 to 239 mg/dL is borderline high; a total cholesterol level of 240 mg/dL (6.21 mmol/L) or greater is high. This provides further evidence of the observation with cholesterol level equal to 564 being an outlier.

Among our numerical predictors, the most substantial correlation is observed between thalach and age, yielding a correlation coefficient of -0.396. Additionally, a notable correlation exists between thalach and oldpeak, with a coefficient of -0.347. It is crucial to acknowledge these correlations, especially if multicollinearity becomes a concern in subsequent stages of analysis. Monitoring these relationships will be integral in ensuring the stability and reliability of our modeling efforts.

3.2 Model selection

In this section we discuss the main findings during the model selection process. For brevity, only key results are discussed, while supplementary results may be found in the appendix.

We first began the model selection process with an *overall significance test* to affirm that all of the coefficients in the logistic regression model were not all equal to zero ($H_0 : \beta_p = 0 \quad \text{for} \; p = 1, ..., 13$). From the overall significance test we obtained $C = 221.24$ and correspondingly a p-value of $5.442165 \times 10^{-37} << \alpha$. Hence, we rejected the null hypothesis and concluded that all coefficients were not equal to zero in the logistic regression model. Consequently, we searched for which coefficients to include in our model afterwards.

In the model selection process we explored the use of different link functions on the logistic regression model. We found that the logit link function performed slightly better than the probit link function in terms of AIC (224.18 vs 225.71) and residual deviance (186.18 vs 187.71) respectively when fitting the full model. Although the improvements in these metrics can be deemed negligible, the main advantage of the logit link function was that its coefficients are easily interpretable. Hence, we decided to use the logit link function to generate twelve different logistic regression models with various combinations of predictors and two-way interaction terms.

Since our data included various continuous variables (age, trestbps, chol, thalach, oldpeak, ca), we used the Z-test with $\alpha = 0.05$ to determine whether any of these variables contained higher order exponential relationships with the response outcome (target) in the full model. We found that the ca predictor had a quadratic relationship (ca^2) with the response variable as it had a p-value of 0.003136 associated with its z statistic (-2.954). Meanwhile, we did not find any other higher exponential relationships for the variable ca. This result involving ca^2 remained consistent when investigating the reduced model as well. Hence this quadratic term was included in both Model 1 and Model 6.

Through our preliminary search of scientific literature, we found proof that certain variables within our dataset had known correlations with each other. These correlations were added to Model 1 and Model 6 respectively to give Model 2 and Model 11 respectively. The inclusion of these interaction terms resulted in very minor improvements in terms of residual deviance but when penalizing for additional coefficients, it resulted in a decrease in performance as indicated by the AIC metric (Table 3).

In addition to manual interaction term selection, we also used automated algorithmic functions in R to select the best two-way interaction terms for Model 1 and Model 6 using the AIC criterion. In the process we used the backward and forward step-wise variable selection process which generated different logistic regression models. In all of the models produced, they included two-way interaction terms which affirmed the belief of correlations between predictors in our data-set. Interestingly, in our results for Model 3 which involved backward step-wise variable selection from the full model with all possible interaction terms, we obtained an residual deviance of 2450.968 which was much greater than the null deviance (407.162). As a result of this computed residual deviance, other metrics such as AIC and R^2 for Model 3 were both extremely unfavorable and erroneous. Correspondingly, Model 3 had the greatest MSE out of all the models included and was unexpectedly the worst model obtained during our experiment. This demonstrates the shortcomings of the backward variable selection function step() in R. On the contrary, Model 5 which was obtained from forward step-wise variable selection from the intercept model, was the best model obtained in the experiment. Not only was it's residual deviance by far the lowest amongst all models but it had a superior AIC metric and R^2 value (0.9235). Interestingly, this model however, did not perform the best in terms of MSE which suggests that model fit towards the data does not necessarily provide the best predictions.

Moreover, We also explored the effect of the intercept by fitting our reduced model with all possible two-way interaction terms without an intercept. The difference of including an intercept term in our model was shown through the constrast between Model 7 and Model 8 which only differ by the intercept term. In terms of performance, both models were practically equivalent in terms of all computed metrics. However, removing the intercept in a multiple regression model is an unconventional step and is generally not recommended unless provided with specific theoretical or practical reason for doing so. Hence, the removal of the intercept was not further explored in the other steps involving two-way interaction term selection.

Although Lasso regression penalizes for having additional coefficients, we found that Model 12 which was derived with Lasso regression actually had a moderately high AIC metric compared to the other models that we obtained in this study. In terms of other aspects such as MSE and R^2, Model 12 performed similarly to the other Models obtained from stepwise variable selection using the Reduced model.

Finally, we also computed the dispersion parameter in our logistic regression model to investigate whether overdispersion was confounding any of our results. We found that the overdispersion parameter $\phi = \frac{D(Y,\hat{\mu})}{n-q} = \frac{31.130}{(295-56)} = 0.1297 < 1$ for the full model. Hence our model was not suspect for overdispersion and quassi-logistic regression models were not further explored.

Table 3: Model Metrics for Model Comparison

Model	Deviance	AIC	NullDeviance	R2	lrt	lrt_pval	MSE
Model 1	177.576	217.576	407.162	0.5639	229.586	1.45×10^{-37}	0.1180
Model 2	174.178	224.178	407.162	0.5722	232.984	1.20×10^{-35}	0.1249
Model 3	2450.968	2694.968	407.162	-5.0196	-2043.807	1	0.3005
Model 4	56.604	160.604	407.162	0.8610	350.558	7.02×10^{-46}	0.1745
Model 5	31.130	143.130	407.162	0.9235	376.032	6.01×10^{-49}	0.2065
Model 6	193.105	217.105	407.162	0.5257	214.057	4.05×10^{-39}	0.1132
Model 7	140.077	264.077	407.162	0.6560	267.085	6.00×10^{-27}	0.1862
Model 8	140.077	264.077	408.957	0.6575	268.880	1.43×10^{-27}	0.1903
Model 9	181.754	213.754	407.162	0.5536	225.407	5.52×10^{-39}	0.1137
Model 10	181.754	213.754	407.162	0.5536	225.407	5.52×10^{-39}	0.1137
Model 11	191.613	217.613	407.162	0.5294	215.549	8.64×10^{-39}	0.1128
Model 12	169.7	251.762	407.1618	0.5832	172.925	3.93×10^{-18}	0.1202

3.3 Final model

Model 5 was the final model that was selected from the model selection process. Model had the lowest residual deviance indicating the best fit for the data. Moreover, Model 5 also had the lowest AIC which penalizes for including additional coefficients which affirmed its conciseness. Accordingly, in Model 5, $R^2 = 0.9235$, thus showing that Model 5 explained for 92 percent of the variation in the data. Moreover, the likelihood ratio test also indicated that all coefficients in Model 5 were significant. Despite having superior fit for the data, Model 5 had mediocre predictive performance as reflected by its sub-par MSE of 0.2065.

Model 5 included the following coefficients and two-way interaction terms:

- **Coefficients** (10): thal, ca, cp, slope, sex, ca^2, oldpeak, trestbps, thalach, chol
- **Two-way interactions** (7): thal:ca, slope:trestbps, thal;oldpeak, slope:oldpeak, slope:sex, sex:oldpeak, trestbps:chol

The coefficient estimates of the fitted model and 95 percent confidence intervals are shown in the Appendix. Note that since the logit link function was used, the coefficient estimates also represented the log-odds ratio for each risk factor in the predictive model.

The residual plots for Model 5 are shown in Figure 2 (Appendix). The top left of the figure shows that the linearity and equal variance assumption are satisfied and the residuals are evenly distributed around the horizontal line of $y = 0$.

4 Conclusion

Through the regression analysis of heart disease, we were able to identify key risk factors amongst clinical data for heart disease diagnosis. We also identified relationships and interactions between variables (objective 3) using data visualization which we further verified using model selection techniques (backward, forward, lasso, etc.). Based on our analysis of the heart disease dataset, conducted the process of data exploration, preprocessing, and model building. We considered various logistic regression models, including those with interaction terms and exponents to consider the relations of the predictors as well. The models were evaluated using different metrics such as deviance, AIC, pseudo-R^2, likelihood ratio test, and mean-squared prediction error. After comparison of these models, we selected the final model based on its performance metrics and the number of variables it had, but one can use other models based on different interests. The final model we selected, Model 5, fitted our data-set extraordinarily well as demonstrated by its R^2 and residual deviance (objective 1). However, we found that it's MSE was not optimal. This made us learn that although a regression model might fit the dataset best, it may not always be the most optimal model in terms of prediction.

Moreover, we found that the dispersion parameter for Model 5 was less than 1. Although we thoroughly discussed various methods of diagnosing and dealing with overdispersion, there was little talk in regards to under-dispersion, which in practice might be useful. We also found that the logit link function in logistic regression is particularly favorable, as coefficient estimates directly provide log-odds ratio for each variable of interest (objective 2). Meanwhile, on the other hand, through hands-on exploration of removing the intercept, we learned that the intercept serves an important role in regression models for baseline observations and should only be removed with caution.

In conclusion, the logistic regression model provided insights into the relationships between various patient attributes and the likelihood of heart disease. The selected model can serve as a tool for understanding and predicting heart disease based on a given new patient. Further research and validation may enhance the robustness of our findings.

References

[1] Dataset: Heart Disease Dataset (Public Health Dataset) https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/datahttps://www.kaggle.com/datasets/ disease-dataset/data

[2] Simmons B. (2021). Investigating on Heart Disease Datasets and Building Predictive Models, A Thesis submitted to the Graduate Faculty of Elizabeth City State University

[3] Gareth James, Daniela Witten, An introduction to Statistical Learning with Application R: Spring New York. Wickham and Grollemu

[4] Annette J. Dobson, Adrian G. Barnett (2018), An Introduction to Generalized Linear Models, Fourth Edition: Spring New York. Wickham and Grollemu; CRC Press

[5] Rodrigo Maranon, Jane F. Reckelhoff (2013), Sex and Gender Differences in Control of Blood Pressure, Clin Sci. 2013 Oct; 125(7): 311-318

[6] Jenna Fletcher (2023), What should my cholesterol level be at my age? Medical News Today.

[7] Tiziana Ciarambino, Pietro Crispino, Gaetano Leto, Erika Mastrolorenzo, Ombretta Para, Mauro Giordano (2022), Influence of Gender in Diabetes Mellitus and Its Complication, Int J Mol Sci. 2022 Aug; 23(16): 8850

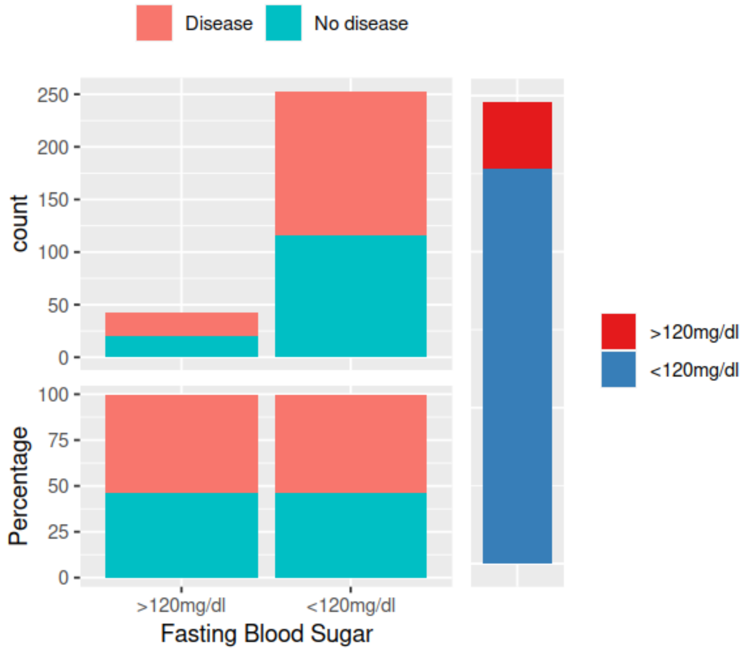
[8] Masaru Sakurai, Jeremiah Stamler, Katsuyuki Miura, Ian J Brown, Hideaki Nakagawa, Paul Elliott, Hirotsugu Ueshima, Queenie Chan, Ioanna Tzoulaki, Alan R Dyer, Akira Okayama, Liancheng Zhao (2011), Relationship of Dietary Cholesterol to Blood Pressure: The INTERMAP Study, J Hypertens. 2011 Feb; 29(2): 222–228.

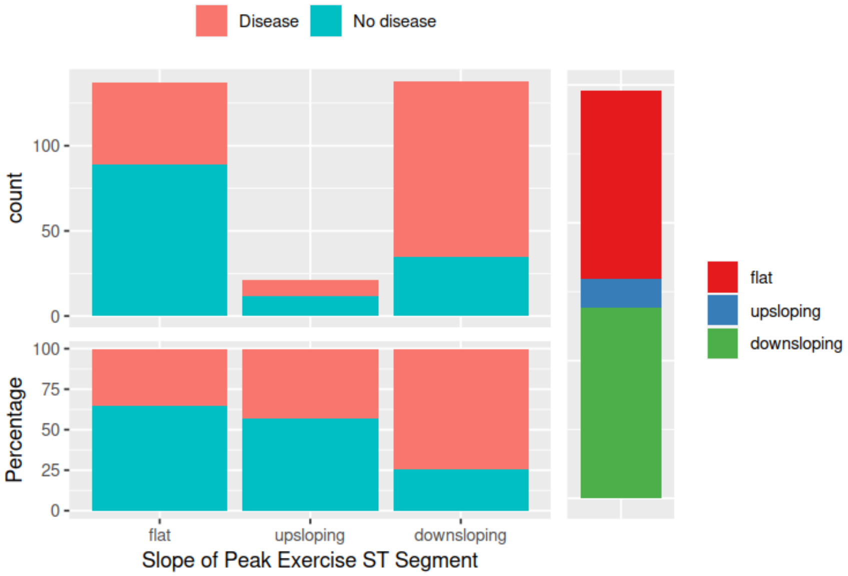
[9] Yaogai Lv, Yan Yao, Junsen Ye, Xin Guo, Jing Dou, Li Shen, Anning Zhang, Zhiqiang Xue, Yaqin Yu, Lina Jin (2018), Sci Rep. 2018; 8: 7917.

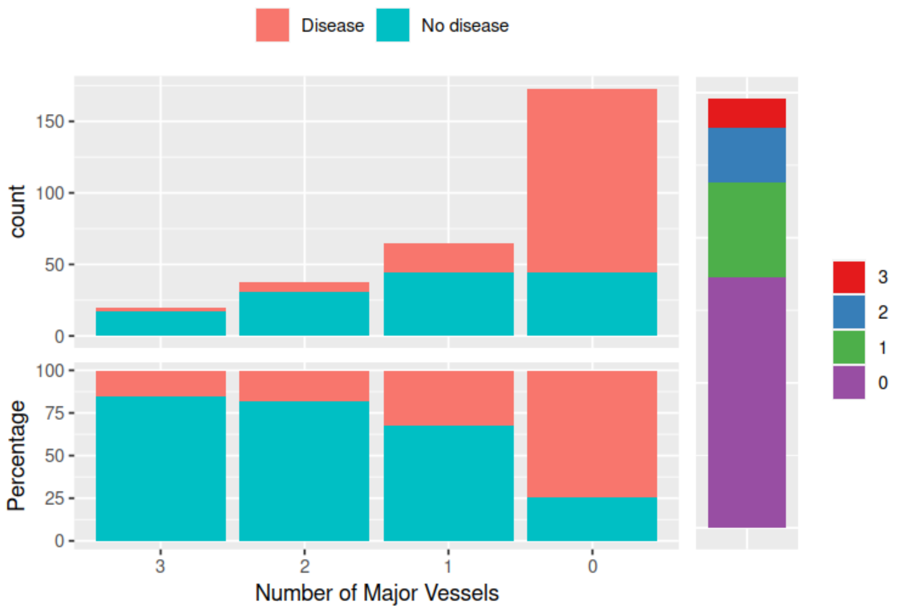
5 Appendix

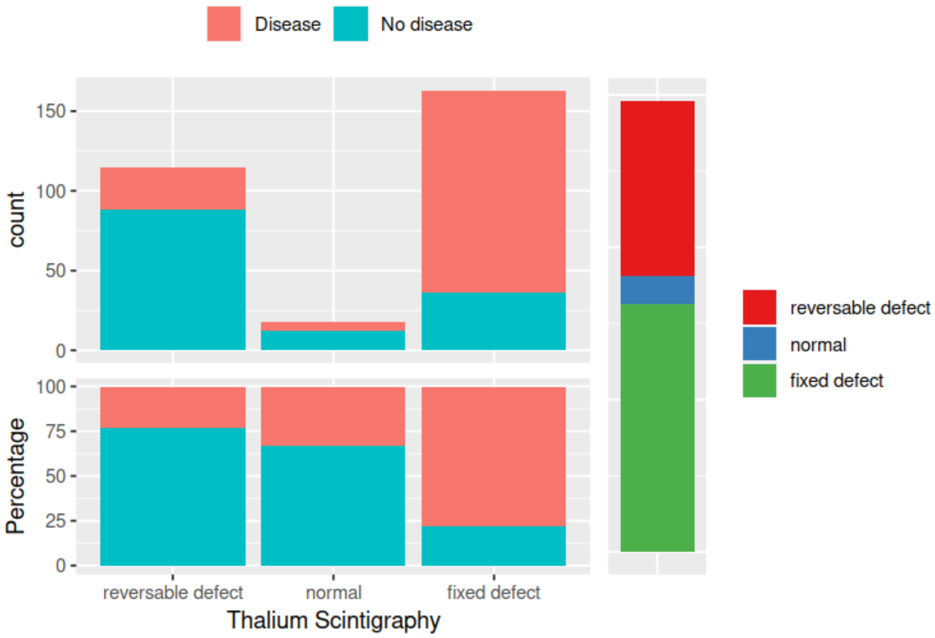
5.1 Data Visualization

Categorical Variables

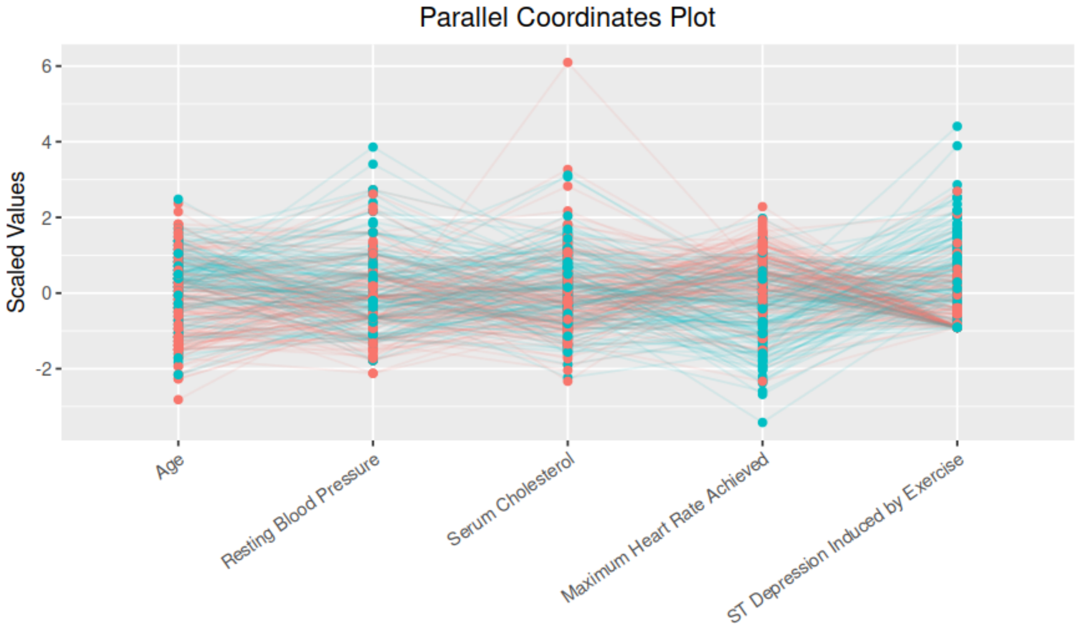


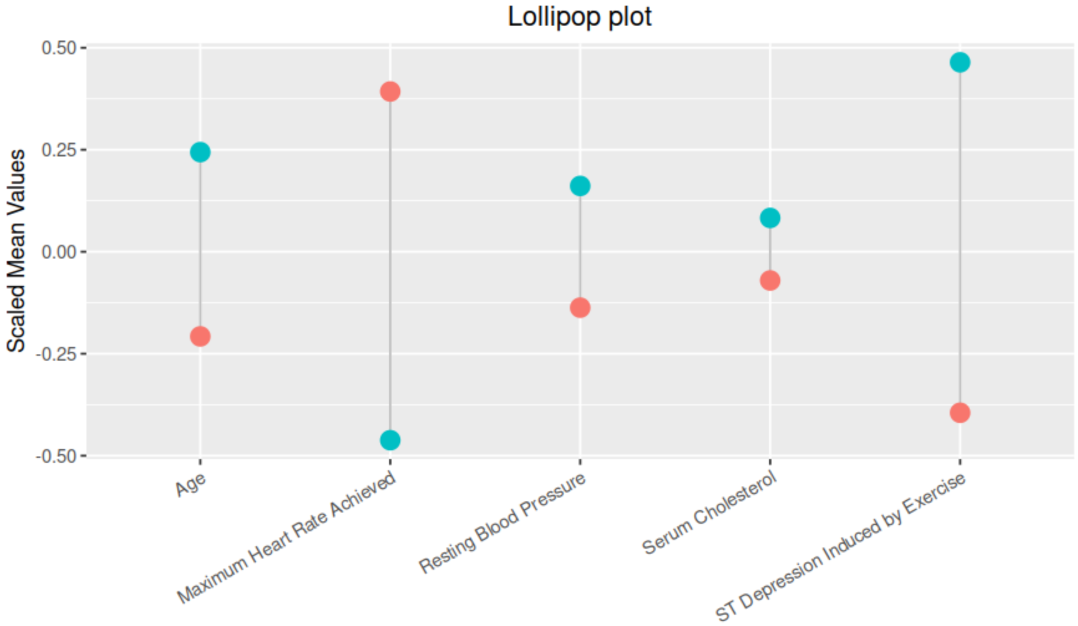



Disease
No disease
count
150
100
50
0
Percentage
100
75
50
25
0
3
2
1
0
Number of Major Vessels
3
2
1
0


Disease
No disease
count
150
100
50
0
Percentage
100
75
50
25
0
reversable defect
normal
fixed defect
Thalium Scintigraphy
reversable defect
normal
fixed defect

Numerical Variables





5.2 Final model

The coefficient estimates and 95 percent confidence for Model 5 are shown in the next two tables.

Coefficient	Estimate	Std_Error
(Intercept)	-224.33	109.14
factor(thal)normal	-302.42	56661.49
factor(thal)reversable defect	163.49	69.29
ca	-10.47	33.23
factor(cp)atypical angina	-51.86	26.95
factor(cp)non-anginal pain	54.84	29.84
factor(cp)typical angina	-61.73	33.46
factor(slope)flat	242.96	111.33
factor(slope)upsloping	375.30	20843.86
factor(sex)male	135.20	60.56
I(ca^2)	72.27	51.37
oldpeak	-167.42	74.35
trestbps	7.86	4.45
thalach	52.09	25.41
chol	-10.29	6.07
factor(fbs)¿120mg/dl	45.53	31.39
factor(thal)normal:ca	65.89	9168.60
factor(thal)reversable defect:ca	2.19	31.81
factor(slope)flat:trestbps	6.61	2.99
factor(slope)upsloping:trestbps	-198.06	29060.78
factor(thal)normal:oldpeak	-16.47	14220.80
factor(thal)reversable defect:oldpeak	112.16	52.80
factor(slope)flat:oldpeak	154.23	69.21
factor(slope)upsloping:oldpeak	189.75	41470.01
factor(slope)flat:factor(sex)male	-94.72	38.78
factor(slope)upsloping:factor(sex)male	-320.98	27981.99
factor(sex)male:oldpeak	63.13	26.68
I(ca^2) : *thalach*	-38.33	21.91
ca:thalach	21.60	12.83
factor(thal)normal:factor(sex)male	393.03	60827.28
factor(thal)reversable defect:factor(sex)male	-59.36	21.67
factor(sex)male:factor(fbs)¿120mg/dl	-119.68	68.45
factor(sex)male:thalach	-30.41	11.10

Continued on next page

Table 4 – continued from previous page

Coefficient	Estimate	Std_Error
factor(thal)normal:thalach	34.33	9279.84
factor(thal)reversable defect:thalach	22.13	8.28
factor(sex)male:chol	18.25	6.38
factor(slope)flat:I(ca^2)	-69.13	37.00
factor(slope)upsloping:I(ca^2)	-118.25	10391.37
oldpeak:trestbps	11.16	5.18
ca:I(ca^2)	-31.87	16.52
factor(cp)atypical angina:thalach	-19.35	8.36
factor(cp)non-anginal pain:thalach	-81.69	57.94
factor(cp)typical angina:thalach	-24.78	10.97
ca:factor(sex)male	78.83	46.83
I(ca^2) : $chol$	6.33	8.00
ca:factor(slope)flat	77.57	44.48
ca:factor(slope)upsloping	60.26	33451.88
trestbps:chol	4.21	2.17
factor(cp)atypical angina:oldpeak	-12.11	7.76
factor(cp)non-anginal pain:oldpeak	69.02	41.88
factor(cp)typical angina:oldpeak	-119.95	58.55
ca:factor(cp)atypical angina	-50.29	29.65
ca:factor(cp)non-anginal pain	64.66	34.88
ca:factor(cp)typical angina	3.36	14.87
thalach:chol	10.12	3.81
oldpeak:factor(fbs)¿120mg/dl	96.22	47.32
factor(slope)flat:factor(slope)upsloping	98.29	28938.82
oldpeak:thalach	-48.51	12.42
chol:factor(fbs)¿120mg/dl	30.01	27.36
factor(slope)flat:trestbps:chol	-40.32	18.33
factor(slope)upsloping:trestbps:chol	33.92	3785.09
factor(cp)atypical angina:oldpeak:trestbps	29.07	11.76
factor(cp)non-anginal pain:oldpeak:trestbps	-7.48	6.40
factor(cp)typical angina:oldpeak:trestbps	22.22	13.85
ca:factor(sex)male:thalach	49.12	26.98
I(ca^2) : $factor(sex)male : thalach$	32.34	21.09
ca:factor(thal)normal:factor(sex)male	100.11	32757.25
ca:factor(thal)reversable defect:factor(sex)male	-62.15	18.35

Continued on next page

Table 4 – continued from previous page

Coefficient	Estimate	Std_Error
factor(sex)male:factor(fbs)¿120mg/dl:thalach	-3.47	14.26
factor(sex)male:factor(thal)normal:thalach	-36.20	9.29
factor(sex)male:factor(thal)reversable defect:thalach	48.05	17.51
factor(sex)male:factor(sex)male:chol	13.92	5.44
factor(sex)male:factor(slope)flat:I(ca^{2})	25.66	22.68
factor(sex)male:factor(slope)upsloping:I(ca^{2})	11.09	11072.94
factor(sex)male:oldpeak:trestbps	-45.52	19.79
factor(sex)male:I(ca^{2}) $: thalach$	-4.69	9.11
factor(sex)male:ca:thalach	-8.89	4.92
factor(sex)male:factor(thal)normal:factor(sex)male	142.12	22415.99
factor(sex)male:factor(thal)reversable defect:factor(sex)male	14.61	7.85
factor(sex)male:factor(sex)male:factor(fbs)¿120mg/dl	5.18	15.49
factor(sex)male:factor(sex)male:thalach	-6.19	1.38
factor(sex)male:factor(thal)normal:thalach	2.17	8.29
factor(sex)male:factor(thal)reversable defect:thalach	-18.22	6.68
factor(sex)male:factor(sex)male:chol	-16.68	8.06
factor(sex)male:factor(slope)flat:I(ca^{2})	4.75	15.68
factor(sex)male:factor(slope)upsloping:I(ca^{2})	15.74	9983.88
factor(sex)male:oldpeak:trestbps	-1.69	9.97
factor(sex)male:I(ca^{2}) $: thalach$	-23.81	12.12
factor(sex)male:ca:thalach	-7.02	8.43
factor(sex)male:factor(thal)normal:factor(sex)male	-14.80	8.55
factor(sex)male:factor(thal)reversable defect:factor(sex)male	-23.09	13.65
factor(sex)male:factor(sex)male:factor(fbs)¿120mg/dl	-4.47	6.59
factor(sex)male:factor(sex)male:thalach	2.43	2.91
factor(sex)male:factor(thal)normal:thalach	10.55	9.99
factor(sex)male:factor(thal)reversable defect:thalach	9.36	6.87
factor(sex)male:factor(sex)male:chol	13.46	10.24
factor(sex)male:factor(slope)flat:I(ca^{2})	-7.43	5.96
factor(sex)male:factor(slope)upsloping:I(ca^{2})	-5.15	28.64
factor(sex)male:oldpeak:trestbps	-14.22	6.58
factor(sex)male:I(ca^{2}) $: thalach$	-6.99	11.88
factor(sex)male:ca:thalach	15.03	3.89

Table 4: Coefficient estimates

Table 5: Coefficients with 95% Confidence Intervals

Coefficient	**2.5%**	**97.5%**
(Intercept)	-242.42	-304.47
factor(thal)normal	-1456.78	870.39
factor(thal)reversable defect	91.45	250.63
ca	-11.14	-9.80
factor(cp)atypical angina	-44.63	-59.44
factor(cp)non-anginal pain	48.92	31.98
factor(cp)typical angina	-62.43	-61.03
factor(slope)flat	321.27	255.70
factor(slope)upsloping	-36.67	781.35
factor(sex)male	134.05	136.38
I(caˆ2)	71.20	73.41
oldpeak	-181.01	-187.86
trestbps	1.13	13.38
thalach	55.37	45.05
chol	-12.61	-9.73
factor(fbs)¿120mg/dl	68.31	48.41
factor(thal)normal:ca	-118.32	249.41
factor(thal)reversable defect:ca	1.53	2.87
factor(cp)typical angina:oldpeak	NA	-13.96
ca:factor(sex)male	77.78	79.83
I(caˆ2):chol	6.16	6.51
ca:factor(slope)flat	81.00	99.63
ca:factor(slope)upsloping	-664.75	778.94
trestbps:chol	0.81	6.79
factor(cp)atypical angina:oldpeak	-19.49	0.22
factor(cp)non-anginal pain:oldpeak	65.89	38.00
factor(cp)typical angina:oldpeak	-123.53	-109.62
ca:factor(cp)atypical angina	-55.94	-73.02
ca:factor(cp)non-anginal pain	53.27	63.72
ca:factor(cp)typical angina	2.55	12.35
thalach:chol	5.74	14.45
oldpeak:factor(fbs)¿120mg/dl	85.36	100.66

Table 5: 95 percent coefficient confidence intervals

Table 6: Coefficients with 95% Confidence Intervals

Coefficient	**2.5%**	**97.5%**
(Intercept)	-242.42	-304.47
factor(thal)normal	-1456.78	870.39
factor(thal)reversable defect	91.45	250.63
ca	-11.14	-9.80
factor(cp)atypical angina	-44.63	-59.44
factor(cp)non-anginal pain	48.92	31.98
factor(cp)typical angina	-62.43	-61.03
factor(slope)flat	321.27	255.70
factor(slope)upsloping	-36.67	781.35
factor(sex)male	134.05	136.38
I(caˆ2)	71.20	73.41
oldpeak	-181.01	-187.86
trestbps	1.13	13.38
thalach	55.37	45.05
chol	-12.61	-9.73
factor(fbs)¿120mg/dl	68.31	48.41
factor(thal)normal:ca	-118.32	249.41
factor(thal)reversable defect:ca	1.53	2.87
factor(slope)flat:trestbps	2.08	14.33
factor(slope)upsloping:trestbps	-792.23	400.83
factor(thal)normal:oldpeak	-339.71	297.74
factor(thal)reversable defect:oldpeak	37.42	196.61
factor(slope)flat:oldpeak	183.41	211.71
factor(slope)upsloping:oldpeak	-642.16	1051.51
factor(slope)flat:factor(sex)male	-141.41	-37.86
factor(slope)upsloping:factor(sex)male	-975.45	326.86
factor(sex)male:oldpeak	54.46	94.61
I(caˆ2):thalach	-40.22	-53.30
ca:thalach	26.33	17.89
factor(thal)normal:factor(sex)male	-989.57	1737.02
factor(thal)reversable defect:factor(sex)male	-79.82	-26.37
factor(sex)male:factor(fbs)¿120mg/dl	-124.95	-168.10
factor(sex)male:thalach	-43.43	-13.89
factor(thal)normal:thalach	-145.37	209.16
factor(thal)reversable defect:thalach	9.44	32.22

Table 6: (continued)

Coefficient	**2.5%**	**97.5%**
factor(sex)male:chol	8.26	23.16
factor(slope)flat:I(caˆ2)	-77.05	-96.18
factor(slope)upsloping:I(caˆ2)	-329.45	93.77
oldpeak:trestbps	12.13	9.68
ca:I(caˆ2)	-36.24	-44.44
factor(cp)atypical angina:thalach	-32.93	-7.62
factor(cp)non-anginal pain:thalach	-40.41	-75.64
factor(cp)typical angina:thalach	NA	-13.96
ca:factor(sex)male	77.78	79.83
I(caˆ2):chol	6.16	6.51
ca:factor(slope)flat	81.00	99.63
ca:factor(slope)upsloping	-664.75	778.94
trestbps:chol	0.81	6.79
factor(cp)atypical angina:oldpeak	-19.49	0.22
factor(cp)non-anginal pain:oldpeak	65.89	38.00
factor(cp)typical angina:oldpeak	-123.53	-109.62
ca:factor(cp)atypical angina	-55.94	-73.02
ca:factor(cp)non-anginal pain	53.27	63.72
ca:factor(cp)typical angina	2.55	12.35
thalach:chol	5.74	14.45
oldpeak:factor(fbs)¿120mg/dl	85.36	100.66

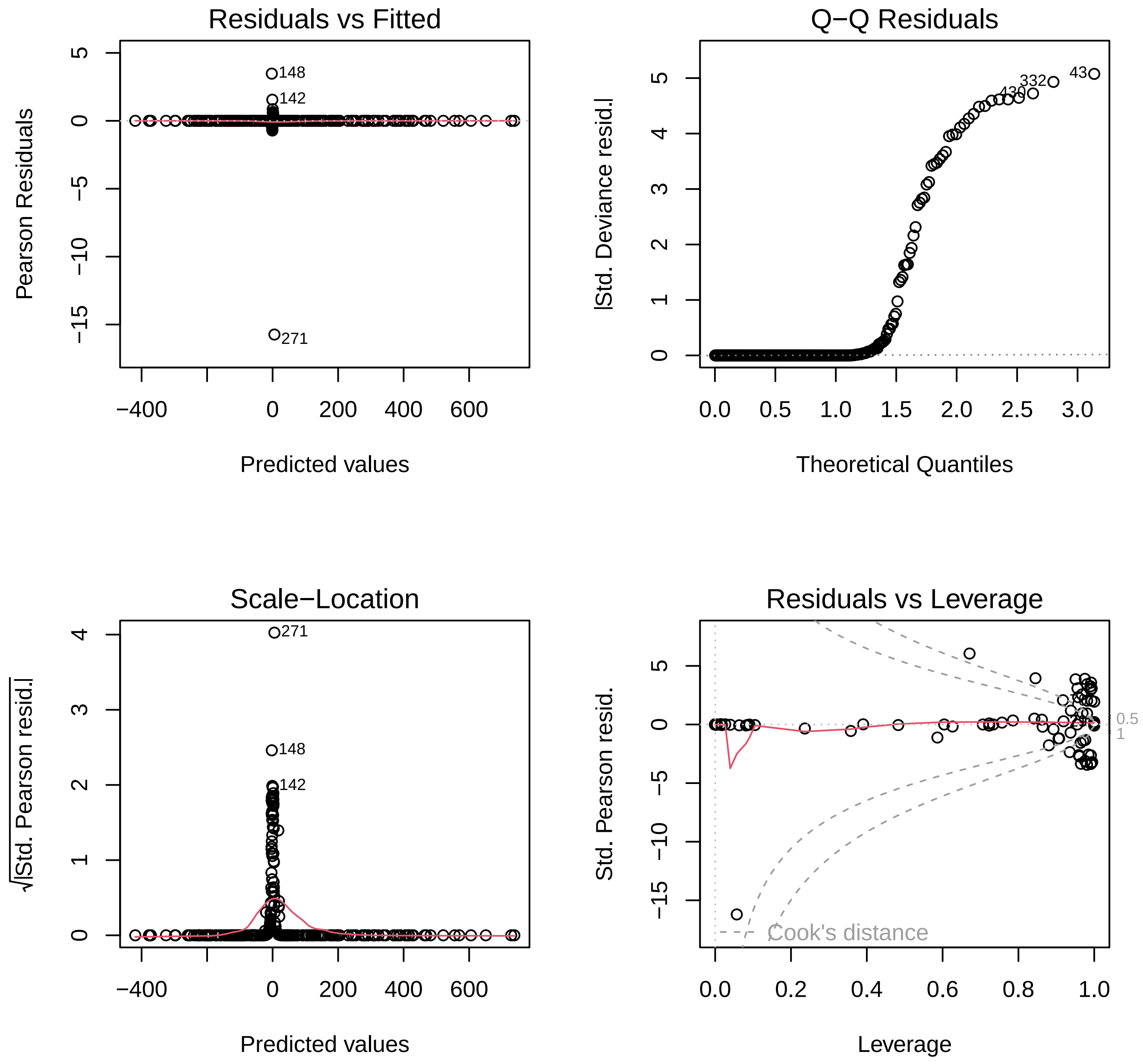


Figure 2: Residual plots for Model 5

5.3 Code

The code used to conduct the experiment is provided in a separate Rmarkdown file and the knit file is attached below with the code.