ANALYZING HEART DISEASE & CHOLESTEROL USING MULTIPLE LINEAR & LOGISTIC REGRESSION

By

Saleh Alshaya, Joaquin Sanchez Ibarra, Mark Sikder

STAT 632 Final Project

CALIFORNIA STATE UNIVERSITY,

EAST BAY

Department of Statistics and Biostatistics

Spring 2021

Contents

1 Abstract

For decades, research has indicated that cholesterol plays a role in heart health. An early diagnosis of heart disease can help health providers aid their patients in taking extra precautions and find a cure. Here, we provide a complete analysis and investigate how does cholesterol level affect individuals of certain ages based on the factor that an individual is suffering from heart disease using multiple linear regression. Furthermore, we develop a logistic regression model to classify and predict whether an individual is suffering from heart disease. The data used for the project was collected from Kaggle with 297 observations. Investigation of the cholesterol effect lead us to believe that cholesterol level does not depend on age but does have an effect on heart condition. Moreover, we found that an individual who had heart disease has a 5.34% greater serum cholesterol in mg/dl, on average, than a person who does not have heart disease problems. With this in mind, we developed a classifier with an AUC of 0.91 and concluded that our multiple logistic model is a good classifier for heart disease.

2 Problem and Motivation

In the United States, heart disease is one leading cause of death. Statistically, about 1 out every 4 deaths are caused by heart disease. There are several types of heart disease from coronary artery disease, heart Arrhythmias, heart valve disease, pericardial disease, heart values disease, just to name a few. But all types have something in common. As this heart disease becomes more extreme can lead to an increase in heart attacks. In this project, we build a statistical model that can classify if an individual has heart disease since early diagnosis of heart disease is a crucial task for many health care providers to help patients take extra precaution measures and reduce the effect of the disease. Besides heart disease, we will be investigating how cholesterol effects an individual suffering from a heart condition. Our body needs cholesterol, it makes cholesterol to make part of our body function correctly. The issue is when we add unnecessary cholesterol to our body by bad eating habits, smoking, or other unhealthy lifestyle. Consequently, we will look at the impact of certain factors we are interested on cholesterol based. Hence in the next sections, we will analyze the effects of cholesterol and classifying heart disease on an individual.

3 Data Description

We acquire the data "Heart Disease Cleveland UCI" from website Kaggle. This data set contains 14 variables where 5 are numeric and 9 categorical. There are no missing values in the data. The data only manipulation we performed was change the 9 categorical variables data that were defined as numeric in the data set into the factors. The relevant numeric variables that we used are serum cholesterol (chol) in mg/dl, age of individual in years, the maximum heart rate (thalach) and the oldpeak which is the ST depression (stage of a heart beat). The relevant categorical variables are the sex of the individual defined as one for males and zero for females. Resting electrocardiographic defined as a test that measures the electrical activity of the heart. It is a categorical variable with 3 levels: zero for normal results, one for ST-T wave abnormality, and two showing probable left ventricular hypertrophy. The

condition variable is defined for one is an individual has heart disease and zero an individual has no heart disease issues. (Note: The left ventricular hypertrophy is the pumping chamber of the heart that has become thicker.)

In our data set there are 297 observations. From which 201 are males and 96 are females. Also 160 of individual have no heart disease problems and the remaining 137 had heart disease.

Figure.1 shows the scatterplot for only numeric values only. The scatterplot in Figure.1 also shows the distribution where we can make the following conclusions. First, *age* and *trestbps* can be considered to be normally distributed as the mean and the median seems to be the same. The variables *chol* and *oldpeak* are positively skewed since their mean is greater than the median. Lastly, *thalach* is negatively skewed. However, the numeric variables do not seem to be highly correlated with each other. Figure.1 also shows a bar plot where we have 56.3% of the males have heart disease and about 25% of females have heart disease.

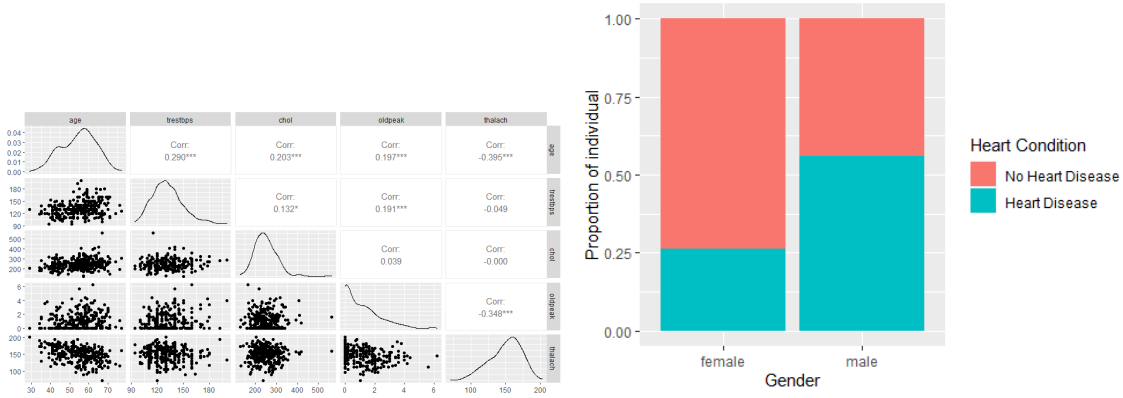


Figure 1: Scatterplots and Proportions of people who have heart disease

All variables and their descriptions can be found in Appendix 2 figure 10 & 11.

4 Questions of interest

To perform a proper analysis of our motivations, we can construct a series of questions that might be of interest to answer. By finding an answer to each of the questions, can aid us in finding associations and effects that each of the variables have on each other and ultimately answer our overall motive. Thus, we are interested in three questions in particular.

1. Does the cholesterol level for males and females depend on their age?
2. Given that the person has heart disease, what is the effect on cholesterol level?
3. How well can we classify heart disease of an individual?

5 Multiple Linear Regression Analysis, Results and Interpretation

The questions that we constructed in the previous section explain two motivations. However, the first two questions explain our first motivation whereas, the third question explains the second motive. From the questions itself, we can see that the first questions require Multiple Linear Regression and the third question is a Logistic Regression. To answer the first two questions, we have to construct a multiple linear regression model with *cholesterol* as the response. But before we can build the model, we need to perform an exploratory analysis to find the best predictors that are statistically significant to *cholesterol*.

5.1 Exploratory Analysis

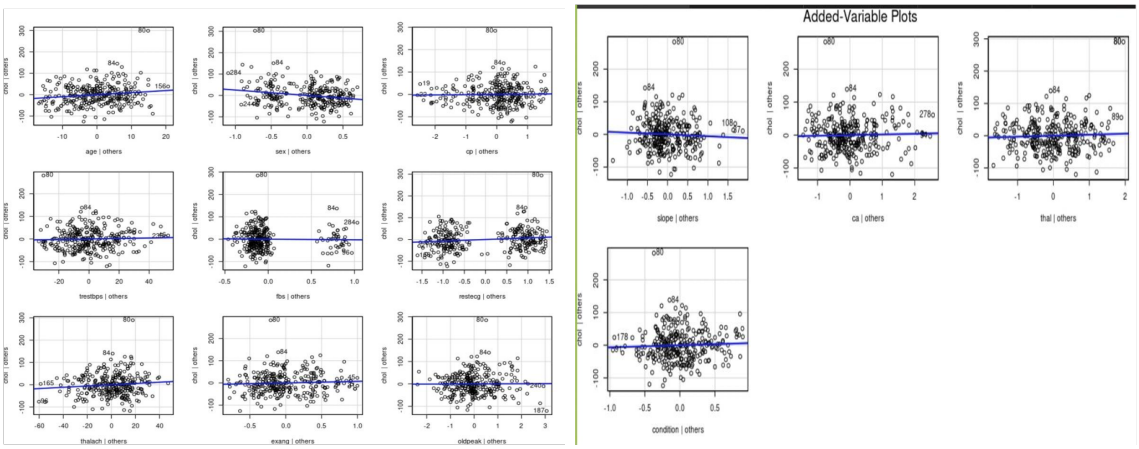


Figure 2: AVPlot

Added-Variable Plots in Figure.2 allows us to see the effect of a singular predictor with the response variable (serum cholesterol level) after controlling the effect of other predictors. The slopes are close to 0 and we do not see any significant linear relationships among the variables with cholesterol. But, that does not mean that none of the variables are not statistically significant with cholesterol. And, to prove this, we use the Global F-test. The Global F-test compares a model with no predictors (null model) to the model with all the predictors (full model), and provides an analysis whether any of the variables effect the response. We perform a hypothesis test with the null hypothesis referring to no predictors associated with *chol* and an alternative hypothesis referring to at least one variable associated with *chol*. Figure.12 reports the result of the hypothesis test.

$H_0 : \beta_1 = \ldots = \beta_{13} = 0$ vs. $H_A : \beta_j \neq 0, j = 1, ..., 13$

From the above analysis we can see that the F-statistic is 2.17 with a p-value of 0.0031. Considering that our level of significance, $\alpha = 0.05$, we reject the null hypothesis as the p_value is less than α. Thus, this proves that there is at least one variable which is associated with *chol*.

We then check if their is multicollinearity issue between the variables. We found that the

variance inflation factors (VIFs) the VIFs are less than 5. Thus, this indicates that there are no issues with multicollinearity. As we have provided evidence that there exists at least one variable which is associated with *chol*, we can further proceed in finding the variables which strongly affect serum cholesterol level. We use stepwise regression to select our optimal variables which effects *chol*. Figure.13 shows the variables selected using AIC on the top and BIC on the bottom for the forward, backward and hybrid stepwise model. Selecting the best model just by looking at the summary statistics is kind of overwhelming and easy to lose track of the variables. Thus, Figure.3 provides a tabulation of all the AIC and BIC scores of the models. From Figure.3, we can see that all the models produce the same BIC score. But, backward AIC model and the stepwise AIC model report the lowest and the same AIC score amongst each other. We now have a choice in choosing any of the two models as they have the same number of variables with the same summary statistics. And, we decided to go forward with the *Backward Stepwise Model* which consists of 8 variables.

	df <dbl>	BIC <dbl>
chol.forward_BIC	4	3188.956
chol.backward_BIC	4	3188.956
chol.step_BIC	4	3188.956

	df <dbl>	AIC <dbl>
chol.forward_AIC	8	3170.778
chol.backward_AIC	8	3169.735
chol.step_AIC	8	3169.735

Figure 3: AIC and BIC summary

It is to be noted that, we also performed regression subset methodology to find the optimal model and Figure.14 shows the *r-squared, adjusted r-squared, Cp and BIC* of the model. Here we see that the *r-squared* model suggests 2 variables, *adjusted r-squared* suggests 8, *Cp* suggests 6 and *BIC* suggests 3. All the models except *adjusted r-squared* ignores most of the variables that we are interested in to answer our questions of interest. Thus, selecting the other models is pretty much going to lead us to a dead end.

5.2 Diagnostic Check for the first model

Backward Stepwise model is our chosen model with the optimal variables. But that does not mean that the model satisfies all the linearity assumptions. In the next few steps, we are going to discuss whether our model satisfies the **LINE** assumptions. **LINE** stands for *Linearity*, *Independence*, *Normality* and *Equal Variance*.

$$\widehat{Chol} = 151.491 + 1.048 \text{ age } - 23.754 \text{ sexmale} + 7.188 \text{ restecg1} + 13.848 \text{ restecg2} + 0.279 \text{ thalach } + 13.387 \text{ condition1}$$

5.3 Interpretation of the diagnostic

Normality assumption can be diagnosed using the *QQ-plot* as displayed in Figure.16. In the QQ-plot, we can see that almost all the distribution of the residuals are close to the

qq-line except for one residual. This indicates that it is a potential outlier and this breaks our normality assumption. Furthermore, the *Residual vs. Fitted* plot in Figure.16 shows that there is no obvious pattern within the residuals. This indicates that the equal variance is satisfied and the regression line is almost similar to the dotted line which is an indication that the Linearity assumption is also satisfied. We cannot particularly say that whether the independence assumption is satisfied or not since we can't check residuals vs time variable plot. So, assume that the independence assumption is satisfied. Continuing with our diagnostic check, we can use the *Standardized Residuals vs. Leverage* plot in Figure.16 to determine the outliers, bad and good leverage points and we can see that there exists one extreme outlier which is outside the threshold. And, this outlier is the data point 80. Relatively, we can see that there are no bad leverage points but 4 good leverage points over the threshold $h_{ii} > 4/n$. This leads us to believe that to resolve the normality assumption, we need to transform our data point since we do not want to remove data points. And, to do this, we can use the multivariate version of Box-Cox, powerTransform() function in the car package to get an analysis on how to transform the predictors. From Figure.17, we can see that the function suggests that we should square *thalach* and keep all the other predictors as it is. After squaring *thalach*, we re-run the powerTransform() function to find whether we need to transform the response or not and, the function suggests that we should take the log transformation of the response. In contrary, we refit the model and similarly check the assumptions for the new transformed model. The equation for the new transformed model is as follows:

$$\begin{aligned}\widehat{log(Chol)} = 4.97 + 0.0039 \text{ age } &- 0.0822 \text{ sexmale} + 0.0264 \text{ restecg1} + \\ &0.0548 \text{ restecg2} + 0.0031 \text{ thalach } + 0.000007 \text{ thalach }^2 + \\ &0.0528 \text{ condition1}\end{aligned}$$

5.4 Diagnostic and Interpretation of the transformed model

Using the suggested model above, we check the **LINE** assumptions. Similarly as before, we plot the *QQ-plot* to check the normality, *Residual vs. Fitted* plot to check equal variance and *Standardized Residuals vs. Leverage* find outliers, good and bad leverage points. Figure.17 provides us an with a diagnostic that the normality assumption was satisfied as the point 80 came closer to the qqline. Moreover, by evaluating the Shapiro-Wilk test, where we set up the null hypothesis, H_0 : as normality exists and the alternative hyptothesis, H_A : as the normality assumption is not satisfied, we can observe that the p-value is 0.2 which is greater than the $\alpha = 0.05$ and thus, suggest that the we fail to reject the null hypothesis, H_0. Hence, we conclude that the normality assumption is met. Figure.10 also shows that the equal variance assumption is still satisfied as no discernible trend is detected and the regression line is almost similar to the dotted line which is an indication that the Linearity assumption is also satisfied.

However, we decided to dig deeper into the analysis and decided to create an alternative model where we do not square *thalach* and performed the power transformations for the response and suggest the log transformation. Then we check the second model multivariate

linear assumptions. New Model:

$$\widehat{log(Chol)} = 5.108466 + 0.004015 \text{ age } - 0.083274 \text{ sexmale} + 0.026894 \text{ restecg1} + 0.055298 \text{ restecg2} + 0.001114 \text{ thalach } + 0.053402 \text{ condition1}$$

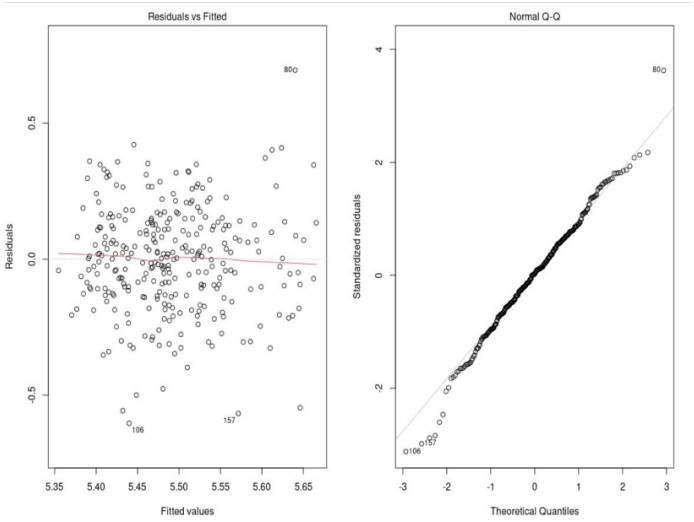


Figure 4: Diagnostics of the non-squared thalach model

When we analyze the diagnostics of the model, Figure.4, we can see that the assumptions are the same with squaring *thalach*. Truth be told, it looks more simpler with the p-values less than $\alpha = 0.05$, indicating that all the predictors are still significant. Hence, there is actually no valid reason to choose the complex model with the same significance and assumptions over the simple model. Choosing the new simple model as the final model is the best option and we can answer our questions of interest.

1. Does the cholesterol level for males and females depend on their age?

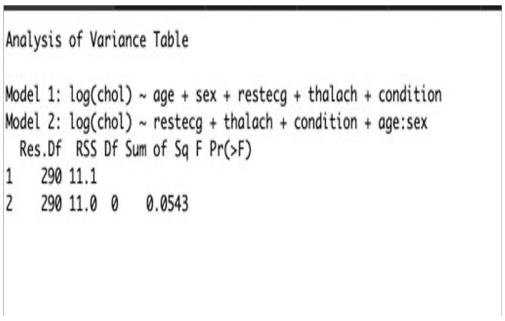


Figure 5: ANOVA Test

In order to understand whether there is any association with cholesterol for male and females based on their age, we conduct a ANOVA test, Figure.12. We set up a null and an alternative hypothesis to see the interaction between the terms.

$H_0 : \beta_6 = 0$ vs. $H_A : \beta_6 \neq 0$

After performing the above analysis, we can see that the p-value is greater than the level of significance, $\alpha = 0.05$. This leads us to fail to reject the null hypothesis. This means that there is no interaction effect between gender and age with cholesterol.

5.5 Interpretation For Question 1

So for answering our question, we have to conclude that the serum cholesterol level does not depend on their age considering all the other variables are held fixed.

2. Given that the person has heart disease, what is the effect on cholesterol level?

```
Call:
lm(formula = log(chol) ~ age + sex + restecg + thalach + condition,
    data = hd_new)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6104 -0.1128  0.0057  0.1211  0.7241

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.108466   0.138159   36.98   <2e-16 ***
age          0.004015   0.001397    2.87   0.0043 **
sexmale     -0.083274   0.025785   -3.23   0.0014 **
restecg1     0.026894   0.100781    0.27   0.7898
restecg2     0.055298   0.023333    2.37   0.0184 *
thalach      0.001114   0.000583    1.91   0.0571 .
condition1   0.053402   0.026598    2.01   0.0456 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2 on 290 degrees of freedom
Multiple R-squared:  0.102,     Adjusted R-squared:  0.0838
F-statistic: 5.51 on 6 and 290 DF,  p-value: 2.01e-05
```

Figure 6: Summary Statistic of Final Model

Similar to question 1, we can use the summary statistic in Figure. 6 to find whether cholesterol have any effect on a person suffering from heart disease.

$H_0 : \beta_6 = 0$ vs. $H_A : \beta_6 \neq 0$

From the summary statistics we can see that the p-value is less than the level of significance, $\alpha = 0.05$. Hence, we can reject the null hypothesis. This means that there is an effect between condition and cholesterol.

5.6 Interpretation For Question 2

We can conclude that an individual who had heart disease has a 5.34% greater serum cholesterol in mg/dl, on average, than a person who does not have heart disease problems when all other variables in the model are held constant. Computation for the interpretation is provided in Appendix 2.

6 Logistic Regression Analysis, Results and Interpretation

In our previous motivation, we investigated the interaction effect between gender, age, cholesterol and saw that age is not a significant factor which affects the cholesterol level for

male and female. However, cholesterol level itself, despite age and gender, acts as a significant factor and has an interaction effect in suggesting whether a person is suffering from heart disease or not. And we can use this analogy to see whether serum cholesterol level is a well suited predictor for our next motivation, which is predicting heart disease in a patient.

Now, from the above variable description we know that heart disease is binary, and multiple regression is not a good algorithm for solving classification problems. Hence, comes **Logistic Regression**, which can be used for classification as well as for regression problems, but is mainly used for classification problems. In our case, we are using **Logistic Regression** to predict the categorical dependent variable, *heart condition*, with the help of independent variables which is classifying and predicting whether an individual has a heart condition or not. But before building our classification model assuming that all the variables have a contribution in predicting an individual suffering from a heart condition, we need to find the most important predictors which has an affect in classifying the dependent variable which will allow us to ignore all the irrelevant variables and help us create a simple model. In order to do so, we can use *stepwise regression*.

6.1 Variable Selection

As we have previously mentioned in **Multiple Linear Regression**, *Stepwise Regression* can be classified into three categories, i.e. **Forward Stepwise**, **Backward Stepwise** and **Hybrid Stepwise**. And, from all these models, we can find out the *Akaike Information Criterion (AIC)* and *Bayesian Information Criterion*, which can help us in determining the best logistic model to be used for the classification problem. The **Forward Stepwise AIC Model** suggested that 9 predictors (thal, ca, cp, oldpeak, slope, sex, trestbps, exang and thalach) from the 14 predictors are estimated to be the most significant in classifying heart condition. **Forward Stepwise BIC Model** on the other hand concludes that only 4 predictors (thal, ca, cp, oldpeak) out of the 14 are significant enough to classify the dependent variable. Similarly, **Backward Stepwise AIC Model** and **Hybrid Stepwise AIC Model** reported the selection of the same 9 predictors out of the 14 predictor to be the most significant. However, the **Backward Stepwise BIC Model** and **Hybrid Stepwise BIC Model** reported the same selection of 6 predictors (sex, cp, trestbps, slope, ca, thal) but different from that of the 4 predictors selected in **Forward Stepwise BIC Model**.

6.2 Model Selection

MODEL\METRICS	AIC	BIC
Forward Stepwise Model	219.7	233.3
Backward Stepwise Model	219.7	223.3
Hybrid Model	219.7	223.3

Figure 7: AIC and BIC Summary Table

It is often quite difficult to interpret the best model just by observing the summary statistics. We can use the AIC and BIC scores generated from each of the models to make a judgment in selecting the best model. The above equation can be used to compute the AIC of the models and we can tabulate the results to compare and find the best model. From Figure.7, we can see that the AIC of all the three models are exactly the same. This might be the case due to the

fact that all selected predictors are same for all the AIC models. Further investigating by using the BIC, we can see that the **Backward Stepwise Model** and **Hybrid Stepwise Model** reports the same BIC but the **Forward Stepwise Model** reports a larger BIC. Hence, we can select either the **Backward Stepwise Model** and **Hybrid Stepwise Model** as the best model. We select **Hybrid Stepwise Model** as the optimal model for our classification problem. So, our final model in terms of logit form is as follows:

$$\begin{aligned}\log(\frac{\hat{p}(X)}{1-\hat{p}(X)}) = -8.8129 + 1.0834sexmale + 1.3131cp1 + 0.8cp2 + \\ 3.3781cp3 + 0.0246trestbps + 1.6077slope1 + \\ 1.6353slope2 + 2.1887ca1 + 3.6437ca2 + 1.1794ca3 + \\ 0.2844thal1 + 1.7423thal2\end{aligned}$$

One thing that can be noted from the above model is that, even though we anticipated that *cholesterol* is an important predictor to classify heart condition, our final model does not include *cholesterol*. This might be the case that from our summary statistics of the model from Question 2, we can observe that the p-value is actually very close to 0.05. Thus, even though it has an impact on the condition, it is still not the best predictor compared to all the other selected predictors. Hence, ignoring cholesterol in the final model is a good option.

6.3 Cross Validation and Evaluation

	actual		
prediction	0	1	Sum
0	46	10	56
1	4	30	34
Sum	50	40	90

Figure 8: Confusion Matrix

As we now have the final model, we can proceed to use the model in training our data set. But before rushing into using the whole data set to train the model, we use cross-validation in which we divide our entire data into two parts - training data and testing data. As the name, we train the model on training data and then evaluate on the testing set. Usually, the size of training data is set more than twice that of testing data, so we split the data in the ratio of 70:30. We use the 70% of the data to train and generate parameters to make a prediction/classification model. We then use this model and the rest 30% of the data to get the model's predictions. We might have created a model to predict whether an individual has a heart condition or not. But we do not know how well our model is predicting the dependent class variable. So, we need to evaluate our model and to do this we can use the Confusion Matrix. A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. Figure.8 shows the confusion matrix. From this matrix, we compute the *Accuracy* to be 0.84, *Sensitivity*

to be 0.75 and *Specificity* to be 0.92. From the three classification metrics, we can see that the model does produce a good accuracy. Alternatively, we can see that the *Specificity* is higher than the *Sensitivity*. This means that our model better at classifying a person not having heart disease rather than having heart disease. It is to be noted that to generate the confusion matrix metrics we had to set the threshold for the prediction of the model to 0.5 as logistic regression gives the probability of the outcome. In order to find the performance of the model on various thresholds, we generate a ROC curve (Figure.9). From this ROC, we find the Area Under the Curve, AUC, to be 0.91.

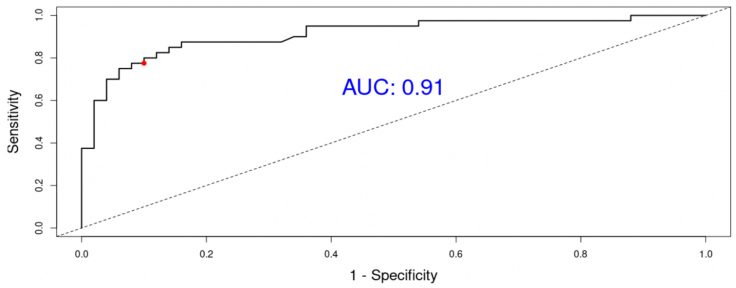


Figure 9: ROC Curve

6.4 Interpretation For Question 3

We can conclude that as the area under the curve is close to one, our multiple logistic model is a good model classifier for heart disease.

7 Conclusion

Many research suggests that cholesterol level tends to climb as an individual ages. But, our investigation reports that cholesterol level does not depend on age respective to gender. This does not mean that our analysis is unreliable. In terms of our exploration to find an effect, we had limitations with the data set. Our data was mostly designed for a binary classification of heart disease. Hence, the variables tend to focus on the factors which correlates with heart disease rather than cholesterol. However, we did find out that cholesterol level does have an effect on an individual who suffers from heart disease. Precisely, an individual who suffers from heart disease has a 5.34% greater serum cholesterol content in mg/dl, on average, than a person who does not have heart disease problems. Digging deeper into our investigation, it is crucial to address whether an individual is diagnosed with heart condition or not. And, with the help of logistic regression we developed a classifier which can predict whether an individual has a heart condition. This is applicable to many health care providers to detect patients having heart conditions and diagnose an early medication. Our classification model aims to have an accuracy of 0.91 over various thresholds and is better at predicting an individual not having heart condition rather than an individual having a heart condition. It is to be noted that our model was developed with only 297 observations and hence, a larger data set might produce different results. A possible extension might be working with a much

larger data set where we have more variables that can show a reliable effect on cholesterol and perform k-fold cross-validation rather than one-fold which might produce a much generalized result.

8 Appendix 1

8.0.1 R code

```
# Loading all the libraries
pacman::p_load(tidyverse, data.table, ggplot2, GGally, car, leaps,
               alr4, Amelia, caret, e1071, pROC, ggridges, bookdown,
               knitr, captioner, stringr)

# load heart disease data
heart_disease_data <- read.csv("heart_disease_data.csv")

# an overall overview of how the type of variables
glimpse(heart_disease_data)

# Data manipulations: change variables into factors
hd_new <-mutate(heart_disease_data,
         sex = as.factor(ifelse(sex==0, "female", "male")),
         cp = as.factor(cp),
         fbs = as.factor(fbs),
         restecg  = as.factor(restecg),
         exang = as.factor(exang),
         slope = as.factor(slope),
         ca = as.factor(ca),
         thal = as.factor(thal),
         condition = as.factor(condition))

attach(hd_new)

# Marginal Plot for the numerical variable
hd_new %>%
  select(age,trestbps , chol ,  oldpeak ,  thalach) %>%
  ggpairs()

# Marginal Plot for the categorical variable
thm <- theme(axis.title =element_text(size=28),
        axis.text = element_text(size=19))

ggplot(hd_new, aes(sex, y = chol)) + geom_boxplot() + thm
ggplot(hd_new, aes(cp, y = chol)) + geom_boxplot() + thm
ggplot(hd_new, aes(fbs, y = chol)) + geom_boxplot() + thm
ggplot(hd_new, aes(restecg, y = chol)) + geom_boxplot() + thm
ggplot(hd_new, aes(exang, y = chol)) + geom_boxplot()+ thm
ggplot(hd_new, aes(slope, y = chol)) + geom_boxplot() + thm
ggplot(hd_new, aes(thal, y = chol)) + geom_boxplot()+ thm
```

```r
ggplot(hd_new, aes(ca, y = chol)) + geom_boxplot() + thm
ggplot(hd_new, aes(condition, y = chol)) + geom_boxplot()+ thm


# Distributions of the variables for Numerical
ggplot(data = melt(hd_new), aes(x = value)) +
  geom_density() +
  facet_wrap(~variable, scales = "free") +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=14))

# summary statistics
summary(hd_new$age)
summary(hd_new$thalach)
summary(hd_new$chol)

# Explanatory Analysis of the data plots

# Comparing which gender suffers from heart disease the most
ggplot(hd_new, aes(sex, fill = condition)) + geom_bar() +
  scale_fill_discrete(name = "Heart Condition",
                      labels = c("No Heart Disease", "Heart Disease")) +
  labs(x = "Gender", y = "Number of individual")

# Comparing the number of patients with/without heart disease
# with different levels of chest pain
ggplot(hd_new, aes(x = cp, y = age, fill = sex)) +
  geom_boxplot() +
  labs(x = "Types of Chest Pain", y = "Age")

# Comparing the types of chest pain for different sex groups of different age
ggplot(hd_new, aes(x = cp, y = age, fill = sex)) +
  geom_boxplot() +
  labs(x = "Types of Chest Pain", y = "Age")

# Comparing the number of patients with greater than 120 mg/dl
# with/wihtout heart disease
ggplot(hd_new, aes(fbs, fill = condition)) + geom_bar() +
  scale_fill_discrete(name = "Heart Condition",
                      labels = c("No Heart Disease", "Heart Disease")) +
  labs(x = "Fasting Blood Sugar > 120mg/hl", y = "Number of Patients")

# Comparing the number of patients resting electographic results
# with/without heart disease
```

```r
ggplot(hd_new, aes(restecg, fill = condition)) + geom_bar(position = "dodge") +
  scale_fill_discrete(name = "Heart Condition",
                      labels = c("No Heart Disease", "Heart Disease")) +
  labs(x ="Resting Electrocardiographic Results", y = "Number of Patients")

# Boxplot of age vs sex with/without heart disease
ggplot(hd_new, aes(x = sex, y = age, fill = condition)) + geom_boxplot()

# Comparing the slope of the peak exercise ST segment for patients
# with/without heart condition
ggplot(hd_new, aes(slope, fill = condition)) + geom_bar(position = "dodge")


# Analysis

# Motivation 1: How does cholesterol level affect individuals of certain ages
# based on the factor that an individual is suffering from heart disease.

# Null and full model
chol.mod.null <- lm(chol~1, data = hd_new)
# summary(chol.mod.null)
chol.mod.full <- lm(chol~., data = hd_new)
# summary(chol.mod.full)

# Generating an avPlot to see the relationship between each of the predictors
# with the response considering the fact that the other variables are fixed.
avPlots(chol.mod.full)

# Global F-test:
anova(chol.mod.null, chol.mod.full)

# Check multicollinearity:
car::vif(chol.mod.full)

# Model selection using step wise function
n <- nrow(hd_new)
chol.forward_AIC <- step(chol.mod.null, scope = list(lower = chol.mod.null,
                                                     upper = chol.mod.full),
                         direction = "forward", trace=0)
summary(chol.forward_AIC) # age, sex, restecg

chol.forward_BIC <- step(chol.mod.null, scope = list(lower = chol.mod.null,
                                                     upper = chol.mod.full),
                         direction = "forward", trace = 0, k = log(n) )
```

```r
summary(chol.forward_BIC) # age, sex

chol.backward_AIC <- step(chol.mod.full, direction = "backward", trace = 0 )
summary(chol.backward_AIC) # age, sex, restecg, thalach, condition

chol.backward_BIC <- step(chol.mod.full, direction = "backward",trace = 0,
                          k = log(n) )
summary(chol.backward_BIC) # age, sex

chol.step_AIC <- step(chol.mod.full,scope = list(lower = chol.mod.null,
                                                 upper = chol.mod.full),trace= 0 )
summary(chol.step_AIC) # age, sex, restecg, thalach, condition

chol.step_BIC <- step(chol.mod.full, scope = list(lower = chol.mod.null,
                                                  upper = chol.mod.full),
                      trace = 0, k = log(n) )
summary(chol.step_BIC) # age, sex



# Model section using best regsubset
options(digits = 2)
mod.reg <- regsubsets(chol~., data = hd_new, nvmax = 13 )
summary.reg <- summary(mod.reg)
summary.reg$which

## Plot the statistics
par(mfrow = c(2, 2))
plot(summary.reg$rsq, xlab = "Number of Variables", ylab = "RSq", type = "b")
plot(summary.reg$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq",
     type = "b")
best_adj_r2 = which.max(summary.reg$adjr2)
points(best_adj_r2, summary.reg$adjr2[best_adj_r2],
col = "red",cex = 2, pch = 20)
plot(summary.reg$cp, xlab = "Number of Variables", ylab = "Cp", type = 'b')
best_cp = which.min(summary.reg$cp[-c(length(summary.reg$cp))])
points(best_cp, summary.reg$cp[best_cp], col = "red", cex = 2, pch = 20)
abline(1, 1, col = "blue")
plot(summary.reg$bic, xlab = "Number of Variables", ylab = "BIC", type = 'b')
best_bic = which.min(summary.reg$bic)
points(best_bic, summary.reg$bic[best_bic],
col = "red", cex = 2, pch = 20)

# Check the assumptions for the backward stepwise model
```

```r
par(mfrow=c(1,2))
plot(chol.backward_AIC, 1:2)

# fitted vs leverage points
p <- ncol(hd_new) - 1
n <- nrow(hd_new)
plot(hatvalues(chol.backward_AIC), rstandard(chol.backward_AIC),
     xlab = "Leverage", ylab="Standardized Residuals",cex.lab = 1.5, cex = 1.5)
abline(v=3*(p+1)/n, lty=2, lwd=2, col="red")
abline(h=c(-4,4), lty=2, lwd=2, col="blue")

index <- which(rstandard(chol.backward_AIC) > 4)
index

# Transforming the predictor and response
pt <- powerTransform(cbind(thalach,age) ~ 1 , data = hd_new)
summary(pt)

# The power transformation suggests that we square thalach.
chol.mod3 <- lm(chol ~ age + sex + restecg + thalach + I(thalach^2) +
                  condition, data = hd_new)

# transformed response
summary(powerTransform(chol.mod3))

# suggest model after transformation
chol.mod4<-lm(log(chol) ~ age + sex + restecg + thalach + I(thalach^2) +
                condition, data = hd_new)
summary(chol.mod4)

# Assumptions of transformed model
par(mfrow=c(1,2))
plot(chol.mod4, 1:2)
shapiro.test(rstandard(chol.mod4))

p <- ncol(hd_new) - 1
n <- nrow(hd_new)
plot(hatvalues(chol.mod4), rstandard(chol.mod4),
     xlab = "Leverage", ylab = "Standardized Residuals")
abline(v=3*(p+1)/n, lty=2, lwd=2, col="red")
abline(h=c(-4,4), lty=2, lwd=2, col="blue")


# We build a new model without squaring thalach
```

```r
chol.mod5 <-  lm(chol ~ age + sex + restecg + thalach + condition,
                 data = hd_new)
summary(chol.mod5)
# transfomed response
summary(powerTransform(chol.mod5))

chol.mod6 <-  lm(log(chol) ~ age + sex + restecg + thalach + condition,
                 data = hd_new)
summary(chol.mod6)

# Assumptions of second suggested model
par(mfrow=c(1,2))
plot(chol.mod6, 1:2)
shapiro.test(rstandard(chol.mod6))


# Question 1
## add interaction term between age and sex
chol.mod7 <-  lm(log(chol) ~  restecg + thalach + condition + age*sex ,
                 data = hd_new)
summary(chol.mod7)
## ANOVA test for interaction term
anova(chol.mod6, chol.mod7 )

# Question 2
summary(chol.mod6)


# Motivation 2: Classifying heart disease in a patient using logistic regression.

# Null and full logistic  model
model.null <- glm(condition ~ 1, hd_new, family = "binomial")
model.full <- glm(condition ~ ., hd_new, family = "binomial")

# model selection stepwise
forward_AIC <- step(model.null,scope =list(lower =model.null, upper=model.full),
                    direction = "forward", trace = 0 )
summary(forward_AIC) # AIC: 219.7

forward_BIC <- step(model.null, scope =list(lower =model.null,upper =model.full),
                    direction = "forward", trace = 0, k = log(n) )
summary(forward_BIC) # AIC: 233.3

backward_AIC <- step(model.full, direction = "backward", trace = 0 )
```

```r
summary(backward_AIC) # AIC: 219.7

backward_BIC <- step(model.full, direction = "backward", trace = 0, k = log(n))
summary(backward_BIC) # AIC: 223.9

step_AIC <- step(model.full,  scope =list(lower = model.null,upper =model.full),
                 trace = 0 )
summary(step_AIC) # AIC: 219.7

step_BIC <- step(model.full, scope =list(lower = model.null,upper = model.full),
                 trace = 0, k = log(n) )
summary(step_BIC) # AIC: 223.9
# Choose hybrid model

# Cross validation
set.seed(999)
index <- sample(1:nrow(hd_new), size =nrow(hd_new)*0.7, replace = FALSE )
train_df <- hd_new[index,]
test_df <- hd_new[-index,]

dim(train_df)
dim(test_df)
# train
final.model <- glm(condition ~ sex + cp + trestbps + slope + ca + thal,
                   family = "binomial", data = train_df )

# predict
pred <- predict(final.model, newdata = test_df , type = "response")

# threshold
predBinary <- ifelse(pred > .5, 1 , 0 )

# Confusion matrix
tb <- table(prediction = predBinary, actual = test_df$condition)
addmargins(tb)

# Accuracy:
(46+30)/90
# Sensitivity:
(30/40)
# Specificity:
(46/50)

# ROC
```

```
roc_obj <- roc(test_df$condition, pred)
plot(1 - roc_obj$specificities, roc_obj$sensitivities, type="l",
xlab = "1 - Specificity", ylab = "Sensitivity", cex.lab = 1.5 ,lwd =2)
text(.5,.6, "AUC: 0.91",cex=2.3, pos=3,col="blue")
# plot red point corresponding to 0.5 threshold:
points(x = 1-(45/50), y = 31/40, col="red", pch=19)
abline(0, 1, lty=2) # 1-1 line

# AUC Area under the curve
auc(roc_obj)
detach(hd_new)
```

9 Appendix 2

9.0.1 References:

1. Data Kaggle link: Heart Disease
2. Heat Disease Left Ventricular Hypertrophy (LVH).
3. Heartdisease & cholesterol risk
4. Heart disease rate deaths

9.0.2 Interpretation calculation

Calculations of interpretation of coefficients when predictors are not transformed and the the response variable is transformed by the logarithm.

$$\begin{aligned}
\beta_j =& \frac{\Delta \log(Y)}{\Delta X} \\
=& \frac{\log(Y_2) - \log(Y_1)}{\Delta X} \\
=& \frac{\log(\frac{Y_2}{Y_1})}{\Delta X} \\
=& \frac{\frac{Y_2}{Y_1} - 1}{\Delta X} \\
=& \frac{100}{100}\left(\frac{\frac{Y_2}{Y_1} - 1}{\Delta X}\right) \\
=& \frac{\%\Delta Y}{100(\Delta X)}
\end{aligned}$$

$$\therefore \beta_j \cdot 100\ \Delta X = \%\Delta Y$$

Interpretation: For every 1 unit increase in x there will be approximately of $(100)\hat{\beta}_j\%$ increase in the response when all other variables in the model are held constant.

9.0.3 Data description

	Variable Name	Information	Type of the Variable
Y_1	chol	serum cholesterol level	numeric
X_1	age	in years	numeric
X_2	sex	1 = male; 0 = female	categorical
X_3	restecg (resting electrocardiographic results)	Value 0: normal Value 1: abnormal Value 2: definite left ventricular hypertrophy	categorical
X_4	thalach	maximum heart rate achieved	numeric
X_5	slope: the slope of the peak exercise ST segment	Value 0: upsloping Value 1: flat Value 2: downsloping	categorical
X_6	trestbps	resting blood pressure	numeric
Y_2	condition	0 = no heart disease, 1 = heart disease	categorical

Figure 10: Data Description

	Variable Name	Information	Type of the Variable
x_8	cp (chest pain type)	Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic	categorical
x_9	oldpeak	ST depression induced by exercise	numeric
x_{10}	exang: exercise induced angina	1 = yes, 0= no	categorical
x_{11}	fbs: (fasting blood sugar > 120 mg/dl)	true = 1, false = 0	categorical
x_{12}	thal	0 = normal, 1 = fixed defect, 2 = reversible defect	categorical
x_{13}	ca	number of major vessels (0-3) colored by flourosopy	categorical

Figure 11: Data Description

9.0.4 Diagnostic plots for preliminary models

```
Analysis of Variance Table

Model 1: chol ~ 1
Model 2: chol ~ age + sex + cp + trestbps + fbs + restecg + thalach +
    exang + oldpeak + slope + ca + thal + condition
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1    296 800310
2    276 691548 20    108762 2.17 0.0031 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 12: ANOVA Test

```
Call:
lm(formula = chol ~ age + sex + restecg + thal, data = hd_new)

Residuals:
    Min      1Q  Median      3Q     Max
-119.81  -31.19   -2.75   28.19  275.72

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  203.498     18.692   10.89  < 2e-16 ***
age            0.895      0.330    2.71  0.00711 **
sexmale      -22.838      6.829   -3.34  0.00093 ***
restecg1      10.886     25.530    0.43  0.67013
restecg2      15.830      5.873    2.70  0.00744 **
thal1        -15.034     12.817   -1.17  0.24175
thal2          9.003      6.588    1.37  0.17280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50 on 290 degrees of freedom
Multiple R-squared:  0.108,     Adjusted R-squared:  0.09
F-statistic: 5.88 on 6 and 290 DF,  p-value: 8.38e-06


Call:
lm(formula = chol ~ age + sex, data = hd_new)

Residuals:
    Min      1Q  Median      3Q     Max
-129.52  -34.28   -4.76   29.33  289.75

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  202.672     18.773   10.80   <2e-16 ***
age            1.068      0.324    3.30   0.0011 **
sexmale      -20.079      6.257   -3.21   0.0015 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50 on 294 degrees of freedom
Multiple R-squared:  0.0735,    Adjusted R-squared:  0.0672
F-statistic: 11.7 on 2 and 294 DF,  p-value: 1.33e-05
```

```
Call:
lm(formula = chol ~ age + sex + restecg + thalach + condition,
    data = hd_new)

Residuals:
    Min      1Q  Median      3Q     Max
-120.11  -30.45   -2.77   27.33  283.85

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  151.491     35.014    4.33  2.1e-05 ***
age            1.048      0.354    2.96  0.00333 **
sexmale      -23.754      6.535   -3.63  0.00033 ***
restecg1       7.188     25.541    0.28  0.77860
restecg2      13.848      5.913    2.34  0.01987 *
thalach        0.279      0.148    1.89  0.06020 .
condition1    13.387      6.741    1.99  0.04798 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50 on 290 degrees of freedom
Multiple R-squared:  0.112,     Adjusted R-squared:  0.0932
F-statistic: 6.07 on 6 and 290 DF,  p-value: 5.31e-06


Call:
lm(formula = chol ~ age + sex, data = hd_new)

Residuals:
    Min      1Q  Median      3Q     Max
-129.52  -34.28   -4.76   29.33  289.75

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  202.672     18.773   10.80   <2e-16 ***
age            1.068      0.324    3.30   0.0011 **
sexmale      -20.079      6.257   -3.21   0.0015 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50 on 294 degrees of freedom
Multiple R-squared:  0.0735,    Adjusted R-squared:  0.0672
F-statistic: 11.7 on 2 and 294 DF,  p-value: 1.33e-05
```

```
Call:
lm(formula = chol ~ age + sex + restecg + thalach + condition,
    data = hd_new)

Residuals:
    Min      1Q  Median      3Q     Max
-120.11  -30.45   -2.77   27.33  283.85

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  151.491     35.014    4.33  2.1e-05 ***
age            1.048      0.354    2.96  0.00333 **
sexmale      -23.754      6.535   -3.63  0.00033 ***
restecg1       7.188     25.541    0.28  0.77860
restecg2      13.848      5.913    2.34  0.01987 *
thalach        0.279      0.148    1.89  0.06020 .
condition1    13.387      6.741    1.99  0.04798 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50 on 290 degrees of freedom
Multiple R-squared:  0.112,     Adjusted R-squared:  0.0932
F-statistic: 6.07 on 6 and 290 DF,  p-value: 5.31e-06


Call:
lm(formula = chol ~ age + sex, data = hd_new)

Residuals:
    Min      1Q  Median      3Q     Max
-129.52  -34.28   -4.76   29.33  289.75

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  202.672     18.773   10.80   <2e-16 ***
age            1.068      0.324    3.30   0.0011 **
sexmale      -20.079      6.257   -3.21   0.0015 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50 on 294 degrees of freedom
Multiple R-squared:  0.0735,    Adjusted R-squared:  0.0672
F-statistic: 11.7 on 2 and 294 DF,  p-value: 1.33e-05
```

Figure 13: Stepwise Regressions

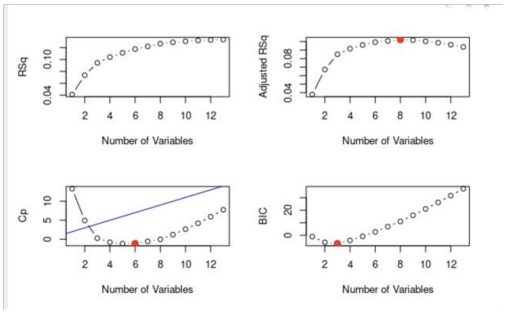


Figure 14: Regression Subset

```
Call:
lm(formula = chol ~ age + sex + restecg + thalach + condition,
    data = hd_new)

Residuals:
    Min      1Q  Median      3Q     Max
-120.11  -30.45   -2.77   27.33  283.85

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  151.491     35.014    4.33  2.1e-05 ***
age            1.048      0.354    2.96  0.00333 **
sexmale      -23.754      6.535   -3.63  0.00033 ***
restecg1       7.188     25.541    0.28  0.77860
restecg2      13.848      5.913    2.34  0.01987 *
thalach        0.279      0.148    1.89  0.06020 .
condition1    13.387      6.741    1.99  0.04798 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50 on 290 degrees of freedom
Multiple R-squared:  0.112,     Adjusted R-squared:  0.0932
F-statistic: 6.07 on 6 and 290 DF,  p-value: 5.31e-06
```

Figure 15: Summary Statistic of the initial model

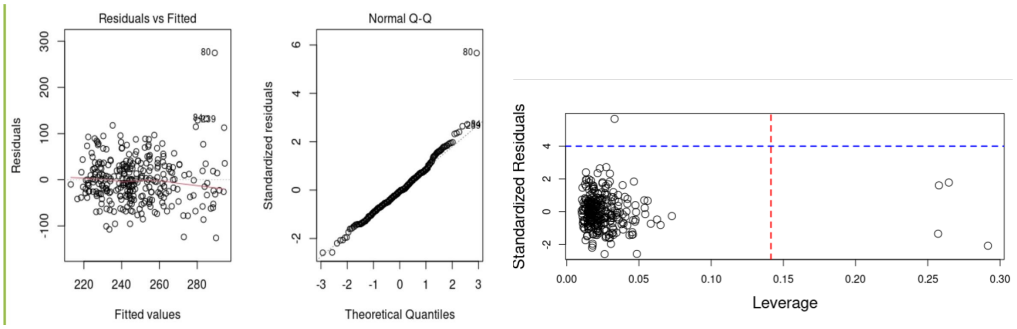


Figure 16: Diagnostics

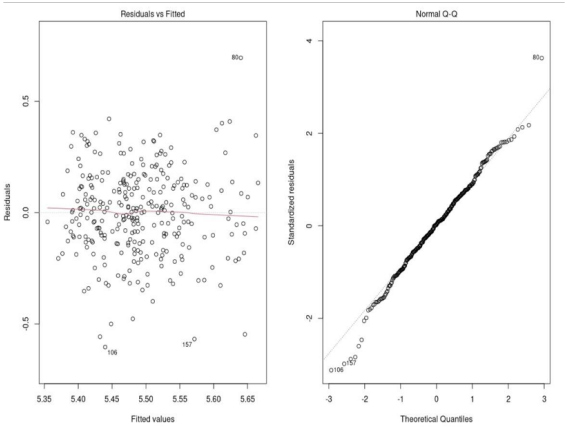
```
## bcPower Transformations to Multinormality
##         Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## thalach       2.5           2         1.80          3.1
## age           1.3           1         0.71          1.9
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                          LRT df   pval
## LR test, lambda = (0 0)  83  2 <2e-16
##
## Likelihood ratio test that no transformations are needed
##                          LRT df  pval
## LR test, lambda = (1 1)  22  2 2e-05
```

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1     0.009          0       -0.38         0.4

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)

Likelihood ratio test that no transformation is needed
```

Figure 17: Power Transformation



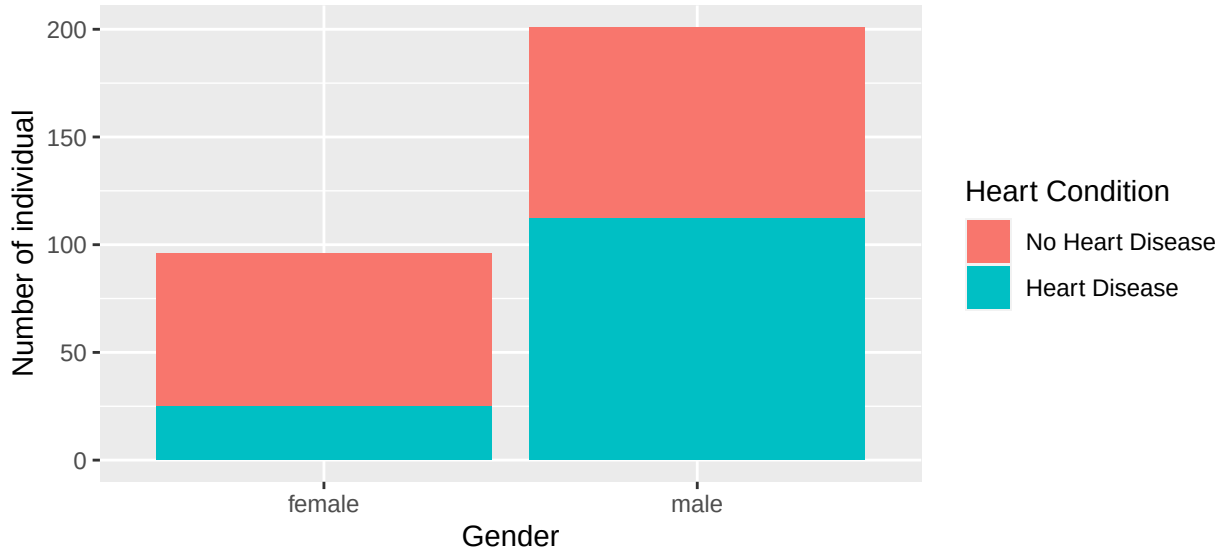

```
        Shapiro-Wilk normality test

data:  rstandard(chol.mod6)
W = 1, p-value = 0.2
```

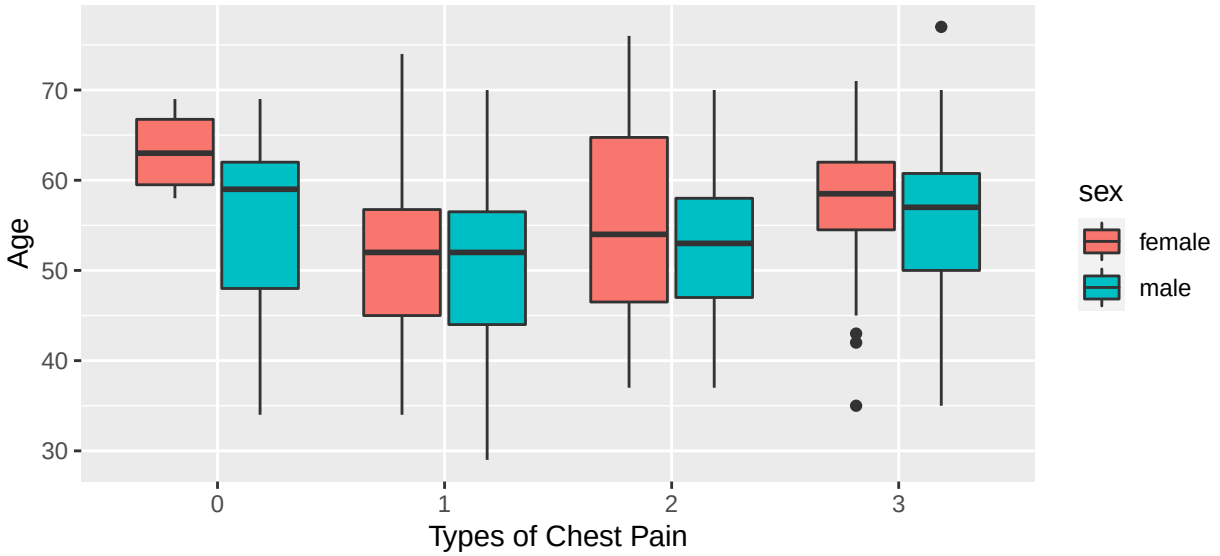
Figure 18: Diagnostics of the transformed model

9.0.5 Exploratory data analysis figures

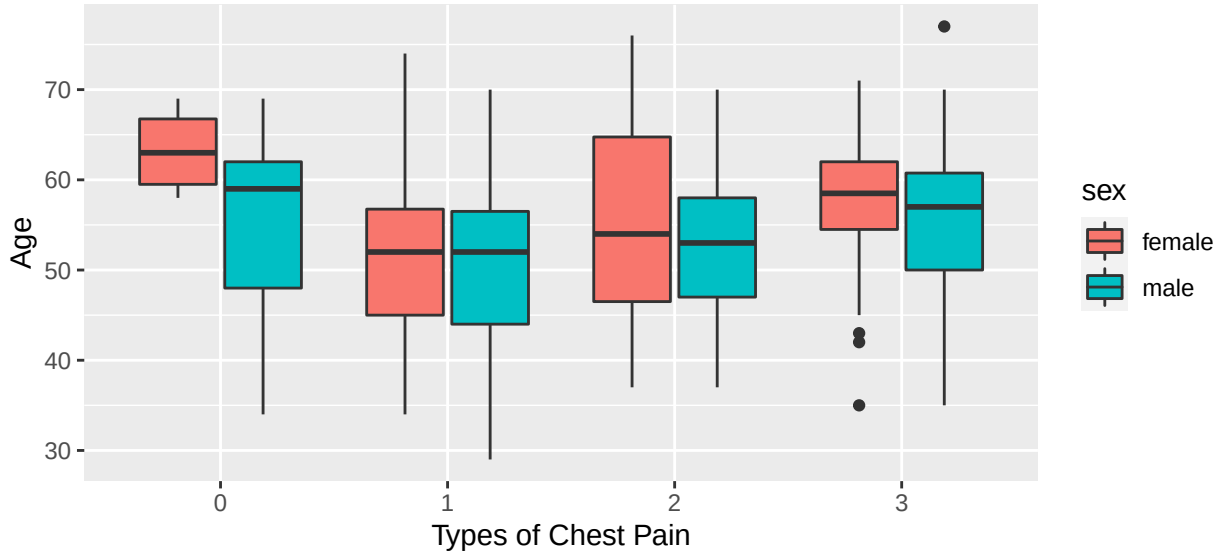
```
# Comparing which gender suffers from heart disease the most
ggplot(hd_new, aes(sex, fill = condition)) + geom_bar() +
  scale_fill_discrete(name = "Heart Condition",
                      labels = c("No Heart Disease", "Heart Disease")) +
  labs(x = "Gender", y = "Number of individual")
```



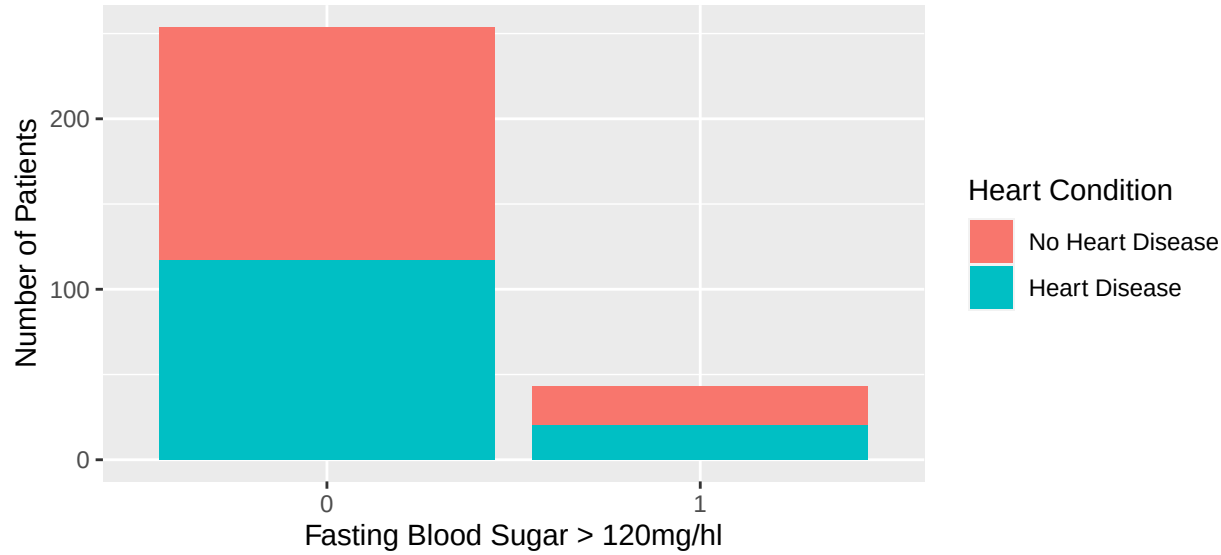

```
# Comparing the number of patients with/without heart disease with different
# levels of chest pain
ggplot(hd_new, aes(x = cp, y = age, fill = sex)) +
  geom_boxplot() +
  labs(x = "Types of Chest Pain", y = "Age")
```



```
# Comparing the types of chest pain for different sex groups of different age
ggplot(hd_new, aes(x = cp, y = age, fill = sex)) +
  geom_boxplot() +
  labs(x = "Types of Chest Pain", y = "Age")
```

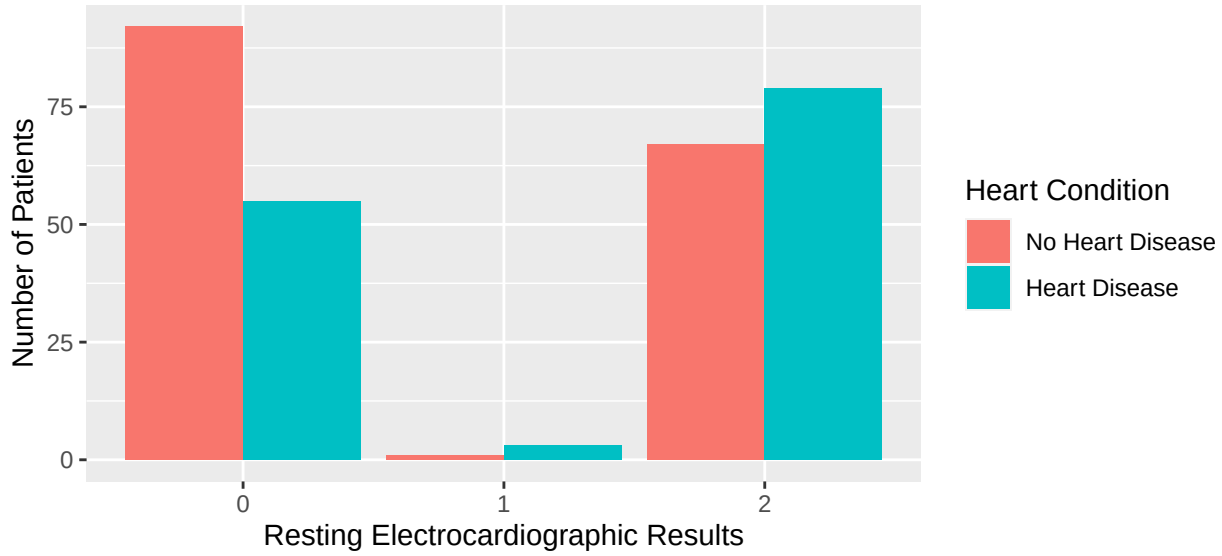



```
# Comparing the number of patients with greater than 120 mg/dl with/wihtout
#heart disease
ggplot(hd_new, aes(fbs, fill = condition)) + geom_bar() +
  scale_fill_discrete(name = "Heart Condition",
                      labels = c("No Heart Disease", "Heart Disease")) +
  labs(x = "Fasting Blood Sugar > 120mg/hl", y = "Number of Patients")
```

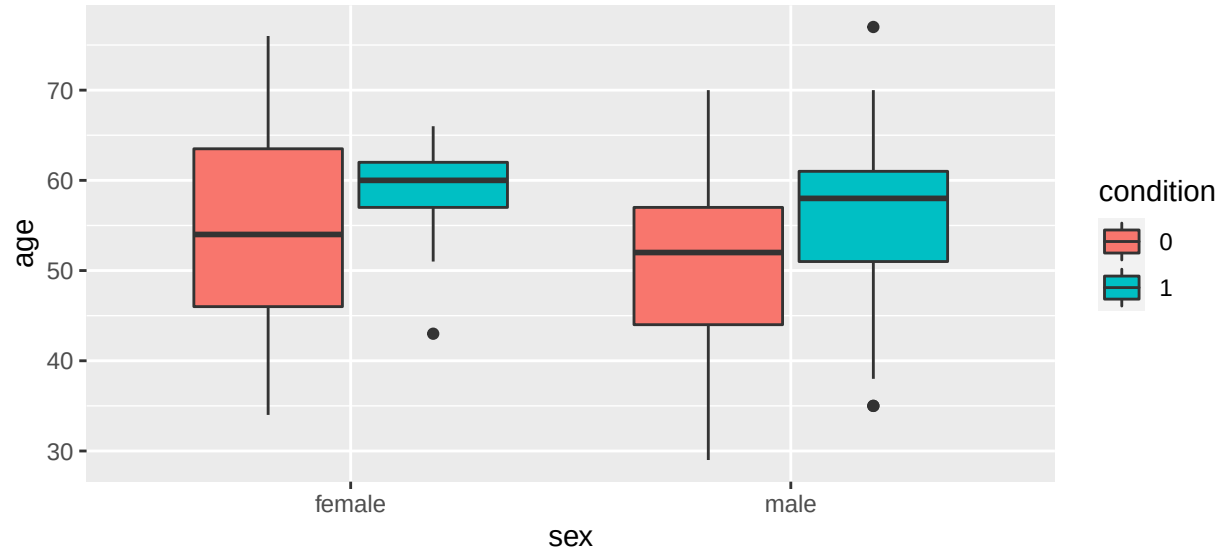



```
# Comparing the number of patients resting electographic results with/without
# heart disease
ggplot(hd_new, aes(restecg, fill = condition)) + geom_bar(position = "dodge") +
```

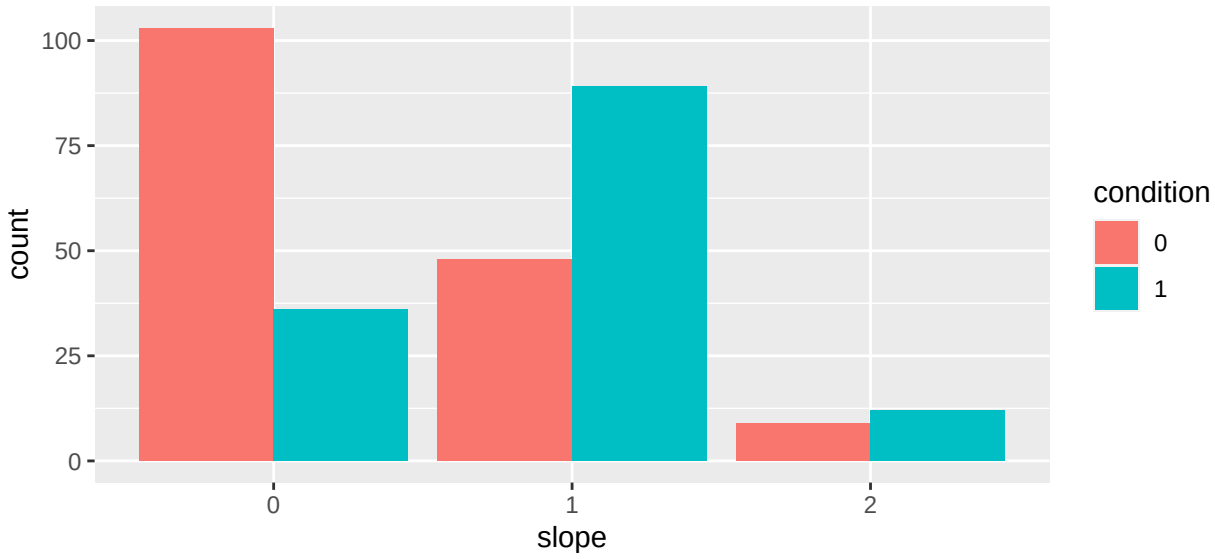
```
  scale_fill_discrete(name = "Heart Condition",
                      labels = c("No Heart Disease", "Heart Disease")) +
  labs(x ="Resting Electrocardiographic Results", y = "Number of Patients")
```




```
# Boxplot of age vs sex with/without heart disease
ggplot(hd_new, aes(x = sex, y = age, fill = condition)) + geom_boxplot()
```




```
# Comparing the slope of the peak exercise ST segment for patients with/without
#heart condition
ggplot(hd_new, aes(slope, fill = condition)) + geom_bar(position = "dodge")
```


100
75
50
25
0
count
0
1
2
slope
condition
0
1