CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Predicting Heart Disease Outcomes Based on Logistic Regression

AUTHOR: Safia Khan

SUPERVISOR: Dr. Sanjoy Sinha

DATE: December 11th, 2020

Table of Contents

Acknowledgements 3
Abstract 4
1. Logistic Regression 5
_1.1 Introduction to Binary Logistic Regression Models 5
_1.2 Parameter Estimation Using Maximum Likelihood 8
_1.3 Significance of Parameters 11
2. Interpretations of Fitted Logistic Regression Models 13
_2.1 Odds Ratio 13
_2.2 Akaike Information Criterion 15
_2.3 Confusion Matrix and Accuracy 16
_2.4 Area Under the ROC Curve 18
_2.5 Backward Elimination 19
3. Heart Disease Data Set 21
_3.1 Introduction to the Data Set 21
_3.2 Manipulation of the Data Set 23
_3.3 Assessing the Residuals 24
4. Model Selection Using R 26
_4.1 Exploring the Full Model 26
_4.2 Selection of Best Regression Model 27
_4.3 Interpretations of Models 31
5. Conclusions 37
Appendix 39
References 49

Acknowledgements

I would like to acknowledge the principal investigators of the data collection that is used throughout this paper:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Abstract

This project is regarding the prediction of heart disease. Heart disease is the 2nd leading cause of death in Canada. Understanding the factors that can lead to an individual's diagnosis of heart disease can play an important role in the world of health care. We consider analyzing an observational data set that contains 13 independent variables as well as a binary outcome variable indicating the presence of the heart disease in a patient. In this project, logistic regression will be used to assess the variables to understand their significance and to obtain an accurate prediction model.

1. Logistic Regression

1.1 Introduction to Binary Logistic Regression Models

When dealing with categorical data from a target variable, logistic regression is often used to model the data. There are different types of logistic regression that can be used, such as simple logistic regression and multiple logistic regression. In the case of the data set that is used throughout this project, multiple logistic regression is used, as the outcome variable represents binary responses (indication of heart disease in a patient).

In the case of a binary logistic model, we will be setting the predictor variables (13 variables) as X and the dichotomous response variable as Y, defined by

$$\mathrm{Y} = \begin{cases} 0, \text{if absense of heart disease} \\ 1, \text{if presence of heart diease.} \end{cases}$$

Before discussing multiple logistic regression model, an introduction to simple logistic regression is needed. Simple logistic regression with a binary outcome variable will lead to a binomial distribution with parameters n_i and p_i. The parameter n_i represents the number of trials and the parameter p_i represents the probability of success in a given trial. Note that the probability of failure is expressed as $1 - p_i$.

To explore the relationship between each predictor variable, X, and outcome variable, Y, we can provide a scatterplot between the two variables. Figure 1 shows the scatterplot that represents the binary relationship between one of the predictor variable's, "Age", and the outcome variable, presence of the heart disease. Note that the scatterplot for the other predictor variables by the outcome variable look similar to the scatterplot shown in Figure 1.

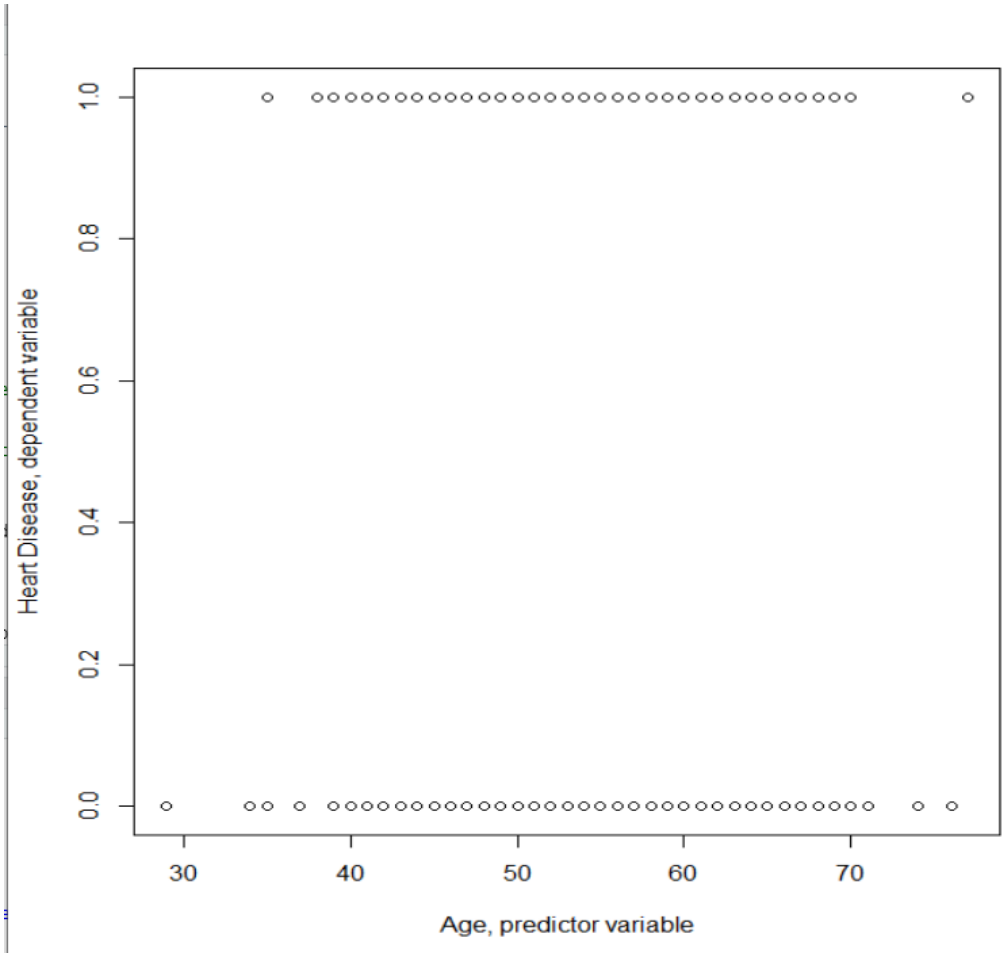


Figure 1: Scatterplot of the binary outcome of heart disease by age of 297 patients.

As shown in Figure 1, the points fall on two parallel lines. This plot depicts that the outcome variable is in fact binary, as one line represents Y=0 and the other Y=1. Suppose p_i denotes the probability that an individual has a heart disease. The probability p_i is defined as a function of the covariates x_i using the logistic regression model

$$\begin{aligned} \text{Logit}\left(\text{p}_{\text{i(x)}}\right) &= \log\left(\frac{\text{p}_{\text{i(x)}}}{1-\text{p}_{\text{i(x)}}}\right) \\ &= \log\left[e^{\beta_0+\beta_1 x_i}\right] \\ &= \beta_0 + \beta_1 x_i. \end{aligned} \tag{1.1}$$

The probability of success can be written as

$$E\left(\frac{y(x)}{n_i}\right) = p_i(x) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}. \quad (1.2)$$

The probability of success, p_i, will fall in the interval of [0,1]. Figure 2 exhibits the proportions of success $\frac{y_i}{n_i}$ at different values of the predictor variable x_i.

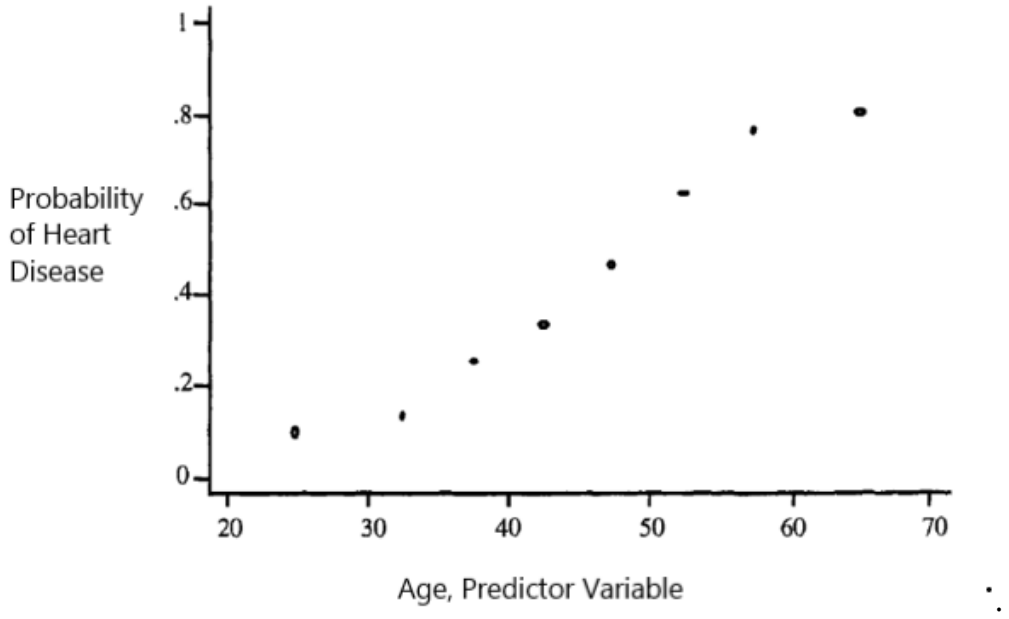


Figure 2: Plot of proportion of successes p_i by age.

As shown in the plot, the values of pi are between the range of [0,1]. The curve of the plot follows an "S-shape". In the case of a binary outcome variable, it is normal to have a plot that represents a typical plot of a cumulative distribution (S- shaped curve).

Now that a simple binary logistic regression has been introduced, we will now discuss multiple logistic regression. Multiple logistic regression will be used throughout this project as there are 13 predictor variables, and we will account for the dependent variable representing a dichotomous outcome. The predictor, **x**, can be represented as a vector,

$$\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n})^T.$$

The multiple logistic regression model represented is defined by

$$\begin{aligned}\text{Logit}\left(p_{i(x)}\right) &= \log\left(\frac{p_{i(x)}}{1-p_{i(x)}}\right)\\ &= \log\left[e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}\right]\\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \qquad (1.3)\end{aligned}$$

The probability of success p_i for multiple logistic regression can be written as

$$E\left(\frac{y(x)}{n_i}\right) = p_i(\boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}. \qquad (1.4)$$

1.2 Parameter Estimation Using Maximum Likelihood

When discussing the parameter estimation in regression models, the methods are not the same for both linear regression and logistic regression. In linear regression, the least squares method is used to estimate the unknown parameters β_0 and β_1. The assumptions that can be made about the linear regression that allows us to use the least square method are not valid for the logistic regression. The method of maximum likelihood is used instead for the logistic regression. First, we will be discussing the maximum likelihood for simple logistic regression. The maximum likelihood method obtains the estimates of the regression parameters by maximizing the likelihood function. Recall as previously mentioned the $p_i(x)$ in Equation (1.2); this will be used for the likelihood function. Given a pair of (x_i, y_i) we can represent the contribution to the likelihood function as

$$\begin{cases} 1 - p_i(x), & \text{given } y_i = 0 \\ p_i(x), & \text{given } y_i = 1 \end{cases}.$$

The probability mass function of the binomial distribution with the given probabilities is defined by

$$f(y_i) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}.$$

Assuming the observations are independent, the likelihood function is obtained by evaluating the joint density of y_i as

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}\ . \tag{1.5}$$

The estimate of β is obtained by maximizing $\mathcal{L}(\beta)$. To do so, we take log on both sides, that is

$$l(\beta) = \log[\mathcal{L}(\beta)] = \sum_{i=1}^{n} \left\{\log\binom{n_i}{y_i} + y_i \log\left[\frac{p_i}{1-p_i}\right] + n_i \log(1-p_i)\right\}.$$

Since there are two unknown parameters, β_0 and β_1, to obtain their estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, we take the derivative of $l(\beta)$ and set it equal to 0. The estimate of β_0 is obtained by solving the estimating equation

$$\sum_{i=1}^{n} \{y_i - n_i p_i\} = 0,$$

with respect to β_0, and the estimate of β_1 is obtained by solving the estimating equation

$$\sum_{i=1}^{n} \{x_i[y_i - n_i p_i\} = 0,$$

with respect to β_1.

Now that the maximum likelihood method for simple logistic regression has been discussed, the maximum likelihood for multiple logistic regression can be introduced. The process is similar to the simple logistic regression except all regression estimates for vector **x** are accounted for. The parameters can be represented as a vector

$$\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T,$$

and the data can be represented as

$$(\mathbf{x_i}, \mathrm{y_i}) \text{ where } \mathrm{i} = (1, \dots, \mathrm{n}).$$

The maximum likelihood is written similar to Equation (1.5) except that now we introduce the vector $\boldsymbol{\beta}$ with the probability $p_i(\boldsymbol{x})$ from Equation (1.4). The likelihood equations for estimating β_0 and $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_\mathrm{p}\right)^T$ are given by

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} \{\mathrm{y_i} - \mathrm{n_i} p_i\} = 0,$$

and

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} \{\mathrm{x_{ij}}[\mathrm{y_i} - \mathrm{n_i} p_i]\} = 0,$$

for j=(1,…,p), respectively.

1.3 Significance of Parameters

After estimating the regression coefficients, the next step is to test the significance of the independent variables using hypothesis testing. Hypothesis testing will indicate how significant the independent variable is to predict the outcome variable, which leads to a model that performs well. Variables are tested to see if the model performs better with or without the inclusion of the insignificant variable in the model. To assess the significance of the variables, we first fit logistic regression models. This can be done using the statistical software "R", where the function "glm" produces a summary of the fitted logistic regression model. The summary of fit generates the estimate, standard error, z-values, and the p-values. In Section 1.2, we discussed how the estimates are found. The standard errors of the estimates are found for multiple logistic regression using the observed information matrix given below.

When estimating the variances of the estimators, $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$, the second partial derivative of the log-likelihood is used. This is expressed as,

$$\frac{\partial^2 l(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 p_i (1 - n_i p_i).$$

The observed information matrix is written as

$$\mathbf{I}(\boldsymbol{\beta}) = -[\partial^2 l(\beta) \,/\, \partial\beta \partial\beta^T].$$

To obtain the variance, we use the inverse of $\mathbf{I}(\boldsymbol{\beta})$, that is,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{I}^{-1}(\boldsymbol{\beta}),$$

where $\mathrm{Var}\left(\beta_j\right)$ is the j^{th} diagonal element of the matrix. The estimates of the variance of the $\hat{\beta}$'s is represented as $V\left(\hat{\beta}\right)$.

To obtain $\boldsymbol{I}(\hat{\boldsymbol{\beta}})$, which is used to fixed the estimated variance, we note that $\mathbf{I}(\hat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{V}\mathbf{X}$. The design matrix, **X**, is a (n x (p+1)) matrix of the covariates and **V** is a (n x n) diagonal matrix of diagonal points of $\hat{p}_i(1-\hat{p}_i)$. The **X** and **V** are denoted as,

$$\mathbf{X} = \begin{vmatrix} 1 & x_{11} & x_{12} \dots & x_{1p} \\ 1 & x_{21} & x_{22} \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n_1} & x_{n2} & x_{np} \end{vmatrix},$$

and

$$\mathbf{V} = \begin{vmatrix} \hat{p}_1(1-\hat{p}_1) & 0 & 0 \dots & 0 \\ 0 & \hat{p}_2(1-\hat{p}_2) & 0 \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \hat{p}_n(1-\hat{p}_n) \end{vmatrix},$$

respectively.

The standard error of $\hat{\beta}_j$ may be obtained as

$$SE(\hat{\beta}_j) = \left[V(\hat{\beta})\right]^{0.5}.$$

The z-values are obtained as

$$z_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)},$$

for j= 0,1,…p, and corresponding p-values are obtained by using the normal probability curve, that is, assuming that z follows the standard normal distribution, we have

$$p - value = P(|z| > z^*)$$

where $z^* = |calculated\ z_j|$. Any regression coefficient with a p-value greater than the level of significance α is considered not significant at level α.

2. Interpretations of Fitted Logistic Regression Models

Section 1 discusses how a logistic regression model is used to assess the significance of the predictors. Significance of variables tells how well each variable performs in a model. In this section, I will discuss how to assess a logistic regression model as a whole using the odds ratio, area under the ROC curve, accuracy and confusion matrices. I will introduce the backward elimination process to determine which model is the best regression using the assessment criteria stated.

2.1 Odds Ratio

Before introducing the odds ratio for the multiple logistic regression, I will first introduce the odds ratio for the simple binary logistic regression. In the case of the binary logistic regression, we can define a predictor variable to have a value of either 0 or 1. Expressing this in terms of a difference in logit model, we can write

$$= \text{logit}\big(\text{p}_\text{i}(\text{x} = 1)\big) - \text{logit}\big(\text{p}_\text{i}(x = 0)\big)$$

$$= \text{g}(1) - \text{g}(0)$$

$$= [\beta_0 + \beta_1] - \beta_0$$

$$= \beta_1.$$

To interpret the logit difference, we can use odds ratio. Odds ratio is defined as the ratio of odds of success at x=1 to the odds of success at x=0. That is, odds ratio,

$$\text{OR} = \frac{\frac{\text{p}(1)}{1-\text{p}(1)}}{\frac{\text{p}(0)}{1-\text{p}(0)}}. \qquad (2.1)$$

Given the outcome variable, Y, has two dichotomous outcomes, Y=0 and Y=1, the following table will represent the values of $p(x)$ and $1-p(x)$ at the two levels of x,

Outcome variable, Y	Independent variable, X	
	x=0	x=1
y=0	$1-\mathrm{p}(0)=\frac{1}{1+\mathrm{e}^{\beta_0}}$	$1-\mathrm{p}(1)=\frac{1}{1+\mathrm{e}^{\beta_0+\beta_1}}$
y=1	$p(0)=\frac{\mathrm{e}^{\beta_0}}{1+\mathrm{e}^{\beta_0}}$	$p(1)=\frac{\mathrm{e}^{\beta_0+\beta_1}}{1+\mathrm{e}^{\beta_0+\beta_1}}$
Total	1.0	1.0

Table 1: Equations for p and 1-p for simple binary logistic regression.

Then the odds ratio is,

$$\begin{aligned}\mathrm{OR} &= \frac{\left[\frac{\frac{\mathrm{e}^{\beta_0+\beta_1}}{1+\mathrm{e}^{\beta_0+\beta_1}}}{\frac{1}{1+\mathrm{e}^{\beta_0+\beta_1}}}\right]}{\left[\frac{\frac{\mathrm{e}^{\beta_0}}{1+\mathrm{e}^{\beta_0}}}{\frac{1}{1+\mathrm{e}^{\beta_0}}}\right]} \\ &= \frac{\mathrm{e}^{\beta_0+\beta_1}}{\mathrm{e}^{\beta_0}} \\ &= \mathrm{e}^{(\beta_0+\beta_1)-\beta_0} \\ &= \mathrm{e}^{\beta_1}.\end{aligned} \tag{2.2}$$

Thus, for a simple binary logistic regression with x=0 and x=1, the odds ratio is equal to e^{β_1}. To find the estimate of $\beta_1, \hat{\beta}_1$, one can take log on both sides of the odds ratio, that is,

$$\hat{\beta}_1 = \log\left(\widehat{\mathrm{OR}}\right) = \log\left(\mathrm{e}^{\hat{\beta}_1}\right).$$

Now, moving on to multiple logistic regression. Recall Equation (2.1). In the case of the multiple logistic regression, suppose the j^{th} predictor x_j is binary. Then we can write

$$\mathrm{OR} = \frac{\frac{\mathrm{p}(\mathrm{x_j}=1)}{1-\mathrm{p}(\mathrm{x_j}=1)}}{\frac{\mathrm{p}(\mathrm{x_j}=0)}{1-\mathrm{p}(\mathrm{x_j}=0)}}. \tag{2.3}$$

Equation (2.2) represents the algebra for OR in terms of the simple logistic regression. In case of the multiple logistic regression, we have

$$\mathrm{OR} = \frac{\exp\left(\beta_0+\beta_1 x_1+\cdots+\beta_{j-1}x_{j-1}+\beta_j+\beta_{j+1}x_{j+1}+\cdots+\beta_p x_p\right)}{\exp\left(\beta_0+\beta_1 x_1+\cdots+\beta_{j-1}x_{j-1}+\beta_{j+1}x_{j+1}+\cdots+\beta_p x_p\right)}$$

$$= \mathrm{e}^{\beta_\mathrm{j}},$$

for j=(1,…,p). This is used to estimate any coefficient at any j^{th} level.

To get the estimate of the coefficient, we take log on both sides of the OR so that

$$\hat{\beta}_\mathrm{j} = \log\left(\widehat{\mathrm{OR}}\right) = \log\left(\mathrm{e}^{\hat{\beta}_\mathrm{j}}\right).$$

2.2 Akaike Information Criterion

When computing the summary fit using statistical software "R", produced in the summary is a value labelled "AIC". AIC stands for Akaike Information Criterion and is utilized in the interpretation of the fitted model. AIC follows the formula,

$$AIC = -2(maximum\ log-likelihood) + 2(number\ of\ free\ parameters).$$

The number of free parameters is the number of estimates used in the fitted model,

$\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_\mathrm{p}\right)^\mathrm{T}$.

The AIC evaluates how poorly the model fits the data set. This is based on the estimation provided using the maximum of log-likelihood. The AIC indicates the bias given from the log-likelihood by the number of free parameters.

When choosing the best fitted regression model, choose the model with the lowest AIC. The lower the AIC the lower the bias in the model is. The model with the lowest AIC will tend to produce a better cross-validation confusion matrix and accuracy predictor. The chosen estimates from the model with the lowest AIC should have estimates with competitive mean square errors, as AIC is based on the maximum log-likelihoods.

When comparing models used to fit the data set, aim for the model with the smallest AIC. The AIC will reduce when the former model has a small reduction of residual deviance compared to the following model. AIC is an informative measurement of a model for its overall goodness-of-fit. Combining AIC and more interpretation evaluations will lead to an easy process of selecting the best regression model.

2.3 Confusion Matrix and Accuracy

First, we will discuss the confusion matrix. Confusion matrices are a result of cross- classifying the predicted values against the actual values. In the case of this project, the confusion matrix is a binary classifier matrix, as there are only two classes. The possible values of the predicted class are “yes” and “no”. In order to classify the data, the classification must be set at a certain threshold of a value “c”, typically taken to be 0.5. Thus the table below will evaluate the model of the outcome variable by classifying subjects based on the threshold of “c”. The outline of the table is given as below

	Predicted (No)	**Predicted (Yes)**
Actual (No)	True Negative (TN)	False Positive (FP)
Actual (Yes)	False Negative (FN)	True Positive (TP)

Table 2: Confusion matrix table in general terms.

These values can be described as such,

- true positive is when the predicted value and the actual value are both positive,
- true negative is when the predicted value and the actual value are both negative,
- false positive is when the predicted value is positive, but the actual value is negative. This is also known as type 1 error,
- false negative is when the predicted value is negative, but the actual value is positive. This is also known as type 2 error.

The fitted model is considered an accurate representation of the data set if the true positive and negatives are maximized and the false positive and negatives are minimized.

Accuracy can be calculated using the confusion matrix. Accuracy is used to easily sum up what the confusion matrix says. Accuracy is calculated by taking the ratio of the true positives and true negatives against the total number, which gives a proportion of how accurate the fitted model is. This is denoted as,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}.$$

If accuracy is approximately 1, then the model fits the data well. If the accuracy is 1 then all the predicted values match the actual values.

2.4 Area Under the ROC Curve

The ROC curve is similar to the confusion matrix. The curve reflects on the model's classifiers performance over the threshold at the c=0.5. The plot consists of the true positive rate as the y-axis and the false positive rate as the x-axis. The formulas for the true positive rate and false positive rate are

$$\text{True Positive Rate} = \frac{\text{TP}}{\text{Actual number of "Yes"}},$$

and

$$\text{False Positive Rate} = \frac{\text{FP}}{\text{Actual number of "No"}},$$

respectively.

The area under the ROC curve (AUC) is measured as it gives and overall performance of the classifier over the entire range of the data set. The AUC measures if the model can discriminate between the individuals who have the outcome of interest (true positive rate) versus those who do not (false positive rate). The AUC ranges between 0 and 1. This is what the AUC measurement means in terms of how well the model can discriminate:

$$\text{AUC} = \begin{cases} 0.5, \text{ Sugests no discrimination} \\ (0.5,0.8), \text{ acceptable discrimination} \\ [0.8, 0.9), \text{ excellent discrimination} \\ [0.9,1], \text{ outstanding discrimination} \end{cases}.$$

Although the AUC is most ideal to be in the range of [0.9,1], it is very rare for the AUC to be in that range.

2.5 Backward Elimination

Backward elimination is a procedure to examine the independent variables that can lead to the "best" regression. Backward elimination is a part of stepwise regression. I will be discussing stepwise regression first.

Stepwise regression starts with one single independent variable and then building the model by adding one independent variable at a time. The variables are chosen using partial F-Test. The variable with the most significant F-value is chosen and will be added as an additional independent variable to the next model. As variables are added to the model, one should examine the overall model to see if the model is performing as well or better with the additional variables. Stepwise regression can then be processed using the following steps:

1. Provide a summary of fit that contains the z-values of the independent variables that are correlated to the dependent variable.

2. Check z-value for an estimate.

3. Check if the value is significant at level α.

4. If the value is significant, adopt the independent variable into the model and repeat the process with another independent variable. If the value is not significant, then remove the variable and declare the model as the best possible regression to fit the data set.

Now moving onto backward elimination. Backward elimination derives from stepwise regression as it attempts to find the best regression model to fit the data. Backward elimination has a similar process but commences will all independent variables included in the model and removes them based on their level of significance.

Backward elimination can then be processed using the following steps:

1. Provide a summary of fit with the z-values for all the possible independent variables that are correlated to the dependent variable.

2. Select the lowest F-value and compare it to the default F-value at a significance level of α.

3. If the F-value is smaller than the default F-value, remove the independent variable from the model and repeat steps with the adjusted model. If the F-value is larger than the default F-value then the model is deemed as the best regression model to fit the data set.

3. Heart Disease Data Set

3.1 Introduction to the Data Set

The data set selected for this project was attained from the UCI Machine Learning Repository website. The data set selected is referred to as the Heart Disease Data. This data set has been used for interpretations in countries all over the world and has been featured in numerous studies. The characteristic of the set is multivariate with a total of 75 attributes and 303 observations. There are multiple data sets on this topic, but I have selected the Processed Cleveland data set. The processed data set that contains the 303 patients from Cleveland, narrowed down from 75 attributes to 14 attributes. I will go in depth with each variable indicating which ones are the independent variables and which one is the dependent variable.

1. Attribute 1 is the age of the patient. This is measured in years. Age is a continuous independent variable.
2. Attribute 2 is the sex of the patient. The value "1" indicates that the patient is a male and "0" indicates female. Sex is a discrete binary independent variable.
3. Attribute 3 is CP. CP stands for chest pain where the patient's chest pain is described by 4 labels: "1" indicates typical angina, "2" indicates atypical angina, "3" indicates non-anginal pain and "4" indicates asymptomatic. CP is a discrete independent variable.
4. Attribute 4 is Trestbps. Trestbps is the patient's blood pressure at rest. The unit of measurement is mm Hg on admission to the hospital. Trestbps is a continuous independent variable.
5. Attribute 5 is Chol. Chol is the patient's serum cholesterol. The unit of measurement is mg/dl. Chol is a continuous independent variable.

6. Attribute 6 is Fbs. Fbs stands for fasting blood sugar. The value “1” represents when fbs>120 mg/dl and “0” represents when fbs≤120 mg/dl. This is a binary discrete independent variable.
7. Attribute 7 is Restecg. Restecg is the resting electrocardiographic results and is described under 3 values. The value “0” indicates normal, “1” indicates having ST-T wave abnormality where it has T wave inversions and/or ST elevation or depression of >0.05mV, and “2” indicates showing probable or definite left ventricular hypertrophy by Estes’ criteria. Restecg is a discrete independent variable.
8. Attribute 8 is Thalach. Thalach is the maximum heart rate a patient achieves. This is a continuous independent variable.
9. Attribute 9 is Exang. Exang is exercised induced angina described using 2 values. The value “0” indicates the patient has exercised induced angina and “1” indicates the patient does not have exercised induced angina. This is a binary discrete independent variable.
10. Attribute 10 is Oldpeak. Oldpeak is ST depression induced by exercise relative to rest. This is a continuous independent variable.
11. Attribute 11 is Slope. Slope is the slope of the peak exercise ST segment, described using 3 values. The value “1” indicates the slope is up-sloping, “2” indicates the slope is flat, and “3” indicates the slope is down-sloping.
12. Attribute 12 indicates CA. CA is the number of major vessels coloured by fluoroscopy that is described within a range of 0 to 3. CA is a continuous independent variable.
13. Attribute 13 is Thal. Thal is short for thalassemia which is a blood disorder that causes the body to produce less hemoglobin than usual. This is described under 3 values. The

value “3” indicates normal, “6” indicates fixed defect, and “7” indicates reversable defect. Thal is a discrete independent variable.

14. Attribute 14 is Num. Num represents the indication of heart disease, the angiographic disease status, represented under 5 values. The value “0” indicates no presence of heart disease, and “1”, “2”, “3” and “4” values indicate presence of heart disease at different levels. For “0” there is a <50% diameter narrowing, and for all other values there is >50% diameter narrowing. Num is a discrete response variable.

3.2 Manipulation of the Data Set

The Processed Cleveland data set contains a few missing values, specifically, for two independent variables, Thal and CA. The instances that contain a missing value are instance 88, 167, 193, 267, 288, and 303. This is a total of six values. I removed these instances completely from the data set, as they can affect the concluded statistical model. Keeping instances with missing variables can lead to these errors:

- missing values can reduce the accuracy of the model chosen. This is because it affects the significance testing of a variable as it contains illegitimate values,
- missing values will create a bias in the chosen estimated coefficients provided in the summary of fit,
- missing values will not represent the data set well. This makes it difficult to draw accurate conclusions.

I utilized Excel to create a new data set with these instances removed. The total instances in the clean data set is 297.

The data set was again manipulated by changing the independent variable from discrete to binary discrete. In the original data set, the presence of heart disease was described with 5 individual responses. Response “0” indicated no presence of heart disease and response 1-4 indicated different levels of heart disease. For the purpose of the model, we disregard the different levels of heart disease and interpret them as one individual level. This is expressed as shown,

$$\text{Num} = \begin{cases} 0, & \text{indicates } 0 \\ 1, & \text{indicates } 1 \\ 2, & \text{indicates } 1 \\ 3, & \text{indicates } 1 \\ 4, & \text{indicates } 1. \end{cases}$$

This way the data set will have a binary dependent variable. This issue was resolved by creating a new variable labelled “Num_2” on Excel. Num_2 was created with this code,

$$= \text{IF}([@\text{Num}] = 0, "0", "1"),$$

where the IF statement took in all Num values. If the value equals “0”, Num_2 would remain “0”, and if else, then Num_2 would change to “1”.

3.3 Assessing the Residuals

A key feature of residual plots is the detection of outliers in the data set. Detecting outliers is of importance as it can determine if there are data entry errors which can lead to bias in the regression.

Residuals are calculated using the formula,

$$\text{residual} = \text{y} - \hat{y},$$

where y are the observed values and $\hat{y}$ are the predicted values given by the regression.

After fitting a model containing all independent variables, we fit the values of the data set using R. This is done using the code,

$$\mathrm{p.hat} <- \mathrm{fitted(fit)},$$

where fit is the fitted model.

The values that are produced are in the interval [0,1]. Appendix B contains the predicted values from the regression.

Now that the predicted values are calculated, they can be used in the residual equation. This is done on R using the code,

$$\mathrm{residuals} <- \mathrm{as.numeric}(\mathrm{Num}_2) - \mathrm{p.hat},$$

where Num_2 is the outcome variable, y, and p.hat are predicted values, $\hat{y}$.

There is a total of thirteen scatterplots that plot the residuals against each independent variable. The plots are shown in Appendix C(a) to Appendix C(d).

The plots indicate that there are no outliers shown in any of the plots, thus all 297 observational values will remain in the data set.

4. Model Selection Using R

4.1 Exploring the Full Model

As backward elimination is the stepwise process chosen, the first model should include all variables. The command glm() provided in statistical software R is used to perform the model. The command glm() will run a logistic regression that regresses the binary outcome of presence of heart disease on the 13 independent variables. The fit is written in the code as

$$\text{fit} < -\text{glm}(\text{as.numeric}(\text{Num\_2}) \sim \text{Age} + \text{Sex} + \text{CP} + \text{Trestbps} + \text{Chol} + \text{Fbs} + \text{Restecg} + \text{Thalach} + \text{Exang} + \text{Oldpeak} + \text{Slope} + \text{as.numeric}(\text{CA}) + \text{as.numeric}(\text{Thal}), \text{family} = \text{binomial}, \text{data} = \text{Cleveland\_2}).$$

The summary fit of this model is provided in Appendix D(a). The model follows a binomial distribution as the outcome variable is binary.

The model includes the independent variables “Age”, “Sex”, “CP”, “Trestbps”, “Chol”, “Fbs”, “Restecg”, “Thalach”, “Exang”, “Oldpeak”, “Slope”, “CA” and “Thal” along with the dependent variable “Num_2”. The summary of fit provides estimates for all 13 variables.

The logit is represented by the estimates of the model using Equation (1.3) as

$$\text{Logit}\big(\hat{\text{p}}(x)\big) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7 + \hat{\beta}_8 x_8 + \hat{\beta}_9 x_9 + \hat{\beta}_{10} x_{10} + \hat{\beta}_{11} x_{11} + \hat{\beta}_{12} x_{12} + \hat{\beta}_{13} x_{13}.$$

From the R output in Appendix C(a), the estimates $\hat{\beta} = \big(\hat{\beta}_1, \dots, \hat{\beta}_{13}\big)^{\text{T}}$ can be expressed as

$$\text{Logit}\big(\hat{\text{p}}(x)\big) = -7.37042 - 0.014164x_1 + 1.312073\,x_2 + 0.575898\,x_3 + 0.024044x_4 + 0.004995x_5 - 1.021918x_6 + 0.245153x_7 - 0.020665x_8 + 0.926104x_9 + 0.247386x_{10} + 0.570009x_{11} + 1.267719\,x_{12} + 0.343936x_{13}.$$

The AIC of the full model is 232.69.

Using Appendix D(a), the variable with the largest p-value in the model is "Age". "Age" is variable 1. As backward elimination is the used, age will be the variable that will be removed and assessed without its inclusion in the next model.

4.2 Selection of Best Regression Model

Now that the full model has been explored, selecting the most appropriate model can commence. In the full model, the variable "Age" was selected to be removed. As backward elimination is used, the second fit will consist of 12 variables as "Age" will no longer be included. The second fit is coded as

$$fit2 < -glm(as.numeric(Num\_2) \sim Sex + CP + Trestbps + Chol + Fbs + Restecg + Thalach + Exang + Oldpeak + Slope + as.numeric(CA) + as.numeric(Thal), family = binomial, data = Cleveland2).$$

The summary of fit 2 is provided in Appendix D(b).

The model includes the variables "Sex", "CP", "Trestbps", "Chol", "Fbs", "Restecg", "Thalach", "Exang", "Oldpeak", "Slope", "CA" and "Thal" against the dependent variable "Num_2". The summary of fit 2 provides estimates for these 12 variables. The logit equation with the inclusion of the estimates is expressed as

$$\text{Logit}\big(\hat{p}(x)\big) = -8.203028 + 1.346131x_1 + 0.583865x_2 + 0.022352x_3 + 0.004652x_4 - 1.031146x_5 + 0.243574x_6 - 0.018347x_7 + 0.942382x_8 + 0.260719\,x_9 + 0.561520x_{10} + 1.224859x_{11} + 0.340860x_{12}.$$

The AIC of the second model is 231.04. Recall that the AIC in the full model is 232.69. Clearly, the second model outperformed the full model. By the backward elimination process, we will deem the second model to be more accurate and continue to reduce the model.

According to Appendix D(b), the variable with the largest p-value is "Oldpeak". This variable will be removed in the next model.

The new model, fit 3, will consist of 11 variables. The model will be assessed with the elimination of "Age" and "Oldpeak" variable. The code to represent the fit is expressed as

$$fit3 <- glm(as.numeric(Num\_2) \sim Sex + CP + Trestbps + Chol + Fbs + Restecg + Thalach + Exang + Slope + as.numeric(CA) + as.numeric(Thal), family = binomial, data = Cleveland2).$$

The summary of fit 3 is provided in Appendix D(c).

The model includes the variables "Sex", "CP", "Trestbps", "Chol", "Fbs", "Restecg", "Thalach", "Exang", "Slope", "CA" and "Thal" against the dependent variable "Num_2". The summary of fit 3 provides estimates for these 11 variables. The logit equation with the inclusion of the estimates is expressed as,

$$\text{Logit}\big(\hat{p}(x)\big) = -8.474113 + 1.431549x_1 + 0.573804x_2 + 0.023568x_3 + 0.005059x_4 - 1.074331x_5 + 0.233448x_6 - 0.019485x_7 + 1.007935x_8 + 0.782011x_9 + 1.274038\,x_{10} + 0.344010x_{11}.$$

The AIC of this model is 230.59. Recall the AIC in the previous model is 231.04. Clearly, the further reduced model outperformed the previous model. By backward elimination process, we

will deem the third model to be the best model and attempt to further reduce it. According to Appendix D(c) the variable with the largest p-value is “Restecg”. This variable is removed in the next model.

The new model, fit 4, consists of 10 variables. The model is assessed with the elimination of “Restecg”, “Age” and “Oldpeak” variable. The code to represent the fit is expressed as

$$fit4 <- glm(as.numeric(Num\_2) \sim Sex + CP + Trestbps + Chol + Fbs + Thalach + Exang + Slope + as.numeric(CA) + as.numeric(Thal), family = binomial, data = Cleveland2).$$

The summary of fit 4 is provided in Appendix D(d).

The model includes the variables “Sex”, “CP”, “Trestbps”, “Chol”, “Fbs”, “Thalach”, “Exang”, “Slope”, “CA” and “Thal” against the dependent variable “Num_2”. The summary of fit 3 provides estimates for these 10 variables. The logit equation with the inclusion of the estimates is expressed as,

$$\text{Logit}\big(\hat{p}(x)\big) = -8.591527 + 1.498722x_1 + 0.569527x_2 + 0.024695x_3 + 0.005910x_4 - 1.068004x_5 - 0.019843x_6 + 1.021666x_7 + 0.822393x_8 + 1.280943x_9 + 0.330230x_{10}.$$

The AIC of this model is 230.22. Recall the AIC of the previous model was 230.59. The new model slightly outperformed the last model. As the model is still outperformed, we will deem the new model as the best regression model and further reduce it in the next model.

According to Appendix D(d) the variable with the largest p-value is “Chol”. This variable will be removed in the next model.

The new model, fit 5, will consist of 9 variables. The model will be assessed with the elimination of "Chol", "Restecg", "Age" and "Oldpeak" variable. The code to represent the fit is expressed as,

$$fit5 <- glm(as.numeric(Num\_2) \sim Sex + CP + Trestbps + Fbs + Thalach + Exang + Slope + as.numeric(CA) + as.numeric(Thal), family = binomial, data = Cleveland2).$$

The summary of fit 5 is provided in Appendix D(e).

The model includes the variables "Sex", "CP", "Trestbps", "Fbs", "Thalach", "Exang", "Slope", "CA" and "Thal" against the dependent variable "Num_2". The summary of fit 3 provides estimates for these 9 variables. The logit equation with the inclusion of the estimates is expressed as,

$$\text{Logit}\big(\hat{\text{p}}(\text{x})\big) = -7.248048 + 1.253367x_1 + 0.568646x_2 + 0.025231x_3 - 1.050740x_4 - 0.018587x_5 + 1.002225x_6 + 0.805342x_7 + 1.273935x_8 + 0.343888x_9.$$

The AIC of this model is 230.9. Recall the previous model's AIC as 230.22. The model with the elimination of variable "Chol" does not perform better according to it's AIC value. As fit 5 model did not perform better than it's previous model, the fit 5 will not be further reduced.

The model (fit 4) with the variables "Sex", "CP", "Trestbps", "Chol", "Fbs", "Thalach", "Exang", "Slope", "CA" and "Thal" against the dependent variable "Num_2" has performed the best according to the measurement of AIC. The models will be further interpreted before the conclusion of which model provides the best regression.

4.3 Interpretations of Models

Each model has been assessed using the Akaike Information Criterion measurement, but there are more ways to assess each model. Before discussing other methods to assess the models, the AIC value of the 5 models are provided below as a reference,

```
> AIC
    fit1     fit2     fit3     fit4     fit5
232.6887 231.0393 230.5924 230.2234 230.8958
```

Figure 3: The AIC values for each individual model.

Explained in Section 2 is the importance of measurements of odds ratio, confusion matrix, accuracy, and area under the ROC curve.

The full model provides information on all 13 variables. The full model had the worst performance according the AIC. Using the code provided in the Appendix A, this model will be further assessed.

The odds ratio, confusion matrix, accuracy and AUC were calculated for the full model using R. The evaluations are given in Appendix E(a).

The OR indicates how well each predictor, *x* value, affects the outcome variable. The odds ratio should range between 0 and infinity and given that the OR>1, the variable is positively associated with the outcome. The higher the OR value, the higher it positively influences the outcome variable. Listed in Table 3 are the odds ratio for the 13 independent variables.

The variables with OR<1 are “Age”, “Fbs” and “Thalach”. Recall in Section 4.1, “Age” is the first variable removed from the full model. The OR value shows that “Age” does not positively influence the outcome variable. This indicates that the decision to remove “Age” from the first model is a good decision.

Variable	Odds Ratio
Age	0.9859361762
Sex	3.7138658483
CP	1.7787278261
Trestbps	1.0243354280
Chol	1.0050077206
Fbs	0.3599040993
Restecg	1.2778170041
Thalach	0.9795467088
Exang	2.5246544794
Oldpeak	1.2806736397
Slope	1.7682826563
CA	3.5527377615
Thal	1.4104886307

Table 3: OR values for each predictor variable, X, in the full model.

Additional information about the full model given in Appendix E(a) are:

- the model fits the data set with an accuracy of 84.8485%,
- the true negative value is 140. This indicates that 140 patients who are predicted to not have heart disease were predicted correctly,
- the true positive value is 112. This indicates that 112 patients who were predicted to have heart disease were predicted correctly,
- combining the true negative and true positive values indicates that 252 out of the 297 patients are correctly classified,

- the false positive (type 1 error) value is 20. This indicates that 20 of the patients were predicted to have heart disease when they do not have heart diseases,
- the false negative (type 2 error) is 25. This indicates that 25 of the patients were predicted to not have heart disease, when they do have heart disease,
- combining false negative and false positive value shows that 45 out of the 297 patients are misdiagnosed
- the AUC value is 84.6259%. This indicates that there is 84.6259% chance that the model can distinguish between the positive and negative class. Thus, the model can discriminate the outcomes well.

Now that the full model has been analyzed, we will compare the interpretations of the full model to the model with the best regression.

The model that provided the best regression in terms of the AIC value is fit 4. The model fit 4 includes the variables “Sex”, “CP”, “Trestbps”, “Chol”, “Fbs”, “Thalach”, “Exang”, “Slope”, “CA” and “Thal” against the dependent variable “Num_2”. Additional information about this models’ odds ratio, accuracy, confusion matrix and AUC are provided in appendix E(b).

The OR values for the reduced model are listed in Table 4.

Comparing the OR values of the reduced model in Table 4 to the OR values of the full model in Table 3, the OR values from the reduced model shows improvement. This indicates that the reduced model contains predictor values that have a higher chance of influencing the outcome. The only predictor value that decreased in value is the variable “Fbs”. In Section 4.2, fit 5 attempted removing the variable “Fbs” and it did not improve the model. Thus, the variable “Fbs” will remain in the reduced model.

Variable	Odds Ratio
Sex	4.4759652172
CP	1.7674304591
Trestbps	1.0250019451
Chol	1.0059273520
Fbs	0.3436937319
Thalach	0.9803529546
Exang	2.7778178638
Slope	2.2759401183
CA	3.6000318488
Thal	1.3912881957

Table 4: OR values for each predictor variable, X, in the reduced model.

Additional information about the reduced model given in Appendix E(b) are:

- the model fits the data set with an accuracy of 85.5219%,
- the true negative value is 142. This indicates that 142 patients who are predicted to not have heart disease were predicted correctly,
- the true positive value is 112. This indicates that 112 patients who were predicted to have heart disease were predicted correctly,
- combining the true negative and true positive values indicates that 254 out of the 297 patients are correctly classified,
- the false positive (type 1 error) value is 18. This indicates that 18 of the patients were predicted to have heart disease when they do not have heart diseases,

- the false negative (type 2 error) is 25. This indicates that 25 of the patients were predicted to not have heart disease, when they do have heart disease,
- combining false negative and false positive value shows that 43 out of the 297 patients were misdiagnosed,
- the AUC value is 85.2509%. This indicates that there is an 85.2509% chance that the model can distinguish between the positive and negative class. Thus, the model can discriminate the outcomes well.

The reduced model outperforms the full model. Thus, the model that is produced from fit 4 will be the chosen final model to represent the data set, Heart Disease. The summary of the full model fit with the inclusion of 95% upper and lower limits is given in Table 5.

An overview of the final model is listed below,

- the final model consists of 10 independent variables against 1 dependent variable. The chosen independent variables are the sex of the patient (Sex), the type of chest pain the patient has (CP), the patient's blood pressure at rest (Trestbps), the patient's serum cholesterol level (Chol), the patient's fasting blood sugar level (Fbs), the maximum heart rate the patient can achieve (Thalach), if the patient has exercised induced angina (Exang), the slope of the patient's ST segment at their peak while exercising (Slope), the colour of the patient's major vessel by fluoroscopy (CA), and if the patient has thalassemia (Thal). These independent variables are against the dependent variable, presence of heart disease in a patient (Num_2),

Covariate	Estimate	Std. Error	Z-Value	P-Value	Lower CI	Upper CI
Intercept	-8.591527	2.502261	-3.434	0.000596	-13.70398898	-3.852568111
Sex	1.498722	0.475877	3.149	0.001636	0.590016482	2.466344595
CP	0.569527	0.190198	2.994	0.002750	0.206288451	0.956636523
Trestbps	0.024695	0.010038	2.460	0.013888	0.005437453	0.045015807
Chol	0.005910	0.003607	1.638	0.101328	-0.001201232	0.013222871
Fbs	-1.068004	0.544658	-1.961	0.049894	-2.173276697	-0.027556442
Thalach	-0.019843	0.009351	-2.122	0.033840	-0.038881176	-0.001924691
Exang	1.021666	0.408926	2.498	0.012475	0.218141747	1.829420800
Slope	0.822393	0.308011	2.670	0.007585	0.224162769	1.438226862
CA	1.280943	0.245781	5.212	1.87e-07	0.822586332	1.790734860
Thal	0.330230	0.097895	3.373	0.000743	0.139802201	0.525268791

Table 5: The summary of fit of the final model chosen to represent the Heart Disease data set.

Note the upper and lower limits are at $\alpha = 0.05$.

- with the estimates given from the summary of the final fit, the logit equation from multiple logistic regression is expressed as:

$$\text{Logit}\big(\hat{p}(x)\big) = -8.591527 + 1.498722x_1 + 0.569527x_2 + 0.024695x_3 + 0.005910x_4 - 1.068004x_5 - 0.019843x_6 + 1.021666x_7 + 0.822393x_8 + 1.280943x_9 + 0.330230x_{10}.$$

This model is the final model as the predictor variables produce the most accuracy in predicting if a patient has heart disease.

5. Conclusions

Creating models that can aide in the medical industry for the prediction of a disease is crucial. This project demonstrated the prediction of a patient having heart disease based on numerous variables that influence the outcome, using the logistic regression. The importance of the model is that doctors can use the model to aide in their clinical decision making. The model can identify that if the patient is not yet diagnosed with heart disease, their health factors put them more at risk to developing heart disease in the future.

Section 1 discusses logistic regression and its importance. This Section introduced the logistic regression model for the binary outcome as well as its probability of success. Section 1 also discusses the method of estimating the unknown parameters, $\boldsymbol{\beta} = \left(\beta_1 , \dots , \beta_p\right)^T$, using the maximum likelihood method. After the parameters are estimated, they are tested for their significance to aide in the selection of the best regression model for the data set.

Section 2 discusses the interpretations that can be made about a model, and the selection process of said models. This section introduced the backward elimination; the process is used in the later sections to select the best regression model. Alongside assessing the significance of the parameters, Section 2 introduced other methods of assessing the regression model. The section discusses the importance of the interpretations of odd ratios, Akaike information criterion, confusion matrices, accuracy, and area under the ROC curve.

Section 3 discusses all 13 independent variables' meanings and their units of measurements. The data set originally contained 303 instances and was reduced to 297 instances as the instances with missing values were removed. The outcome variable was then manipulated to become a binary variable as it was originally a discrete, nonbinary variable. The residuals of the data set

were then assessed to determine if the logistic regression can be used. The residual plot for the 13 independent variables indicated that there are no outliers in the data set. Thus, all 297 instances should remain in the data set chosen for the analysis.

Section 4 explored models with backward elimination. Starting with the full model, the summary of fit showed that the variable “Age” did not impact the prediction of heart disease. This was concluded based on the predictor’s significance value and the overall model’s AIC value. Backward elimination was then repeated and all insignificant variables were removed. The model that was selected was model 4. Model 4 included 10 variables, “Sex”, “CP”, “Trestbps”, “Chol”, “Fbs”, “Thalach”, “Exang”, “Slope”, “CA” and “Thal”. Thus, the model removed the variables “Age”, “Restecg”, and “Oldpeak”. At the end of the Section 4, the full model was compared to the best regressed model. Based on the odds ratio, confusion matrix, accuracy, AIC and AUC, the reduced model outperformed the full model for all interpretation measurements. Thus, model 4 was chosen as the final model to predict the presence of heart disease.

In conclusion, although the full model does represent the data well, the model will predict heart disease in a patient at a higher level of accuracy if the age, the resting electrocardiographic results, and the ST depression of a patient when induced by exercise, are not included in the model.

Appendix

Appendix A : Code

```
install.packages("readxl")
library("readxl")
install.packages("ISLR")
library(ISLR)
install.packages("Metrics")
library("Metrics")
install.packages("tidyverse")
library(tidyverse)
library(broom)

#import data set using files
Cleveland2 <- read_excel("Carleton University/Carleton University Year 5/MATH
4905(honours project)/Data Sets and Code/Cleveland2.xlsx")
View(Cleveland2)

#this attaches the variables
attach(Cleveland2)

#start using glm function
# with all variables
fit<-
glm(as.numeric(Num_2)~Age+Sex+CP+Trestbps+Chol+Fbs+Restecg+Thalach+Exang
+Oldpeak+Slope+as.numeric(CA)+as.numeric(Thal), family=binomial, data=Cleveland2)
summary(fit)
#Compute the fit odds ratio, confusion matrix, accuracy, AUC
tidy_fit<- tidy(fit)
Odds_Ratio1<- exp(tidy_fit$estimate)
Odds_Ratio1
Num_2<- as.numeric(Num_2)
pred<-predict(fit,Cleveland2,type="response")
Cleveland2$pred<- ifelse(pred>=0.5,1,0)
AUC1<- auc(Cleveland2$Num_2, Cleveland2$pred)
AUC1
accuracy1<- accuracy(Cleveland2$Num_2, Cleveland2$pred)
accuracy1
confusion_matrix1<- table(Cleveland2$Num_2, Cleveland2$pred,
dnn=c("True","Predicted"))
confusion_matrix1
```

```
#Introduce fit 2
#remove variable age
fit2<-
glm(as.numeric(Num_2)~Sex+CP+Trestbps+Chol+Fbs+Restecg+Thalach+Exang+Oldp
eak+Slope+as.numeric(CA)+as.numeric(Thal), family=binomial, data=Cleveland2)
summary(fit2)

#introduce fit 3
#remove variable oldpeak
fit3<-
glm(as.numeric(Num_2)~Sex+CP+Trestbps+Chol+Fbs+Restecg+Thalach+Exang+Slop
e+as.numeric(CA)+as.numeric(Thal), family=binomial, data=Cleveland2)
summary(fit3)

#introduce fit 4
#remove variable restecg
fit4<-
glm(as.numeric(Num_2)~Sex+CP+Trestbps+Chol+Fbs+Thalach+Exang+Slope+as.num
eric(CA)+as.numeric(Thal), family=binomial, data=Cleveland2)
summary(fit4)
#Compute the fit odds ratio, confusion matrix, accuracy, AUC
tidy_fit4<- tidy(fit4)
Odds_Ratio4<- exp(tidy_fit4$estimate)
Odds_Ratio4
Num_2<- as.numeric(Num_2)
pred4<-predict(fit4,Cleveland2,type="response")
Cleveland2$pred4<- ifelse(pred4>=0.5,1,0)
AUC4<- auc(Cleveland2$Num_2, Cleveland2$pred4)
AUC4
accuracy4<- accuracy(Cleveland2$Num_2, Cleveland2$pred4)
accuracy4
confusion_matrix4<- table(Cleveland2$Num_2, Cleveland2$pred4,
dnn=c("True","Predicted"))
confusion_matrix4

#introduce fit 5
#remove variable chol
fit5<-
glm(as.numeric(Num_2)~Sex+CP+Trestbps+Fbs+Thalach+Exang+Slope+as.numeric(C
A)+as.numeric(Thal), family=binomial, data=Cleveland2)
summary(fit5)

#compare all AIC's
AIC<- c(fit=AIC(fit), fit2=AIC(fit2), fit3=AIC(fit3), fit4=AIC(fit4), fit5=AIC(fit5))
AIC
```

```
#the residuals of the variables
p<-fitted(fit)
fitted(fit)
residuals<-as.numeric(Num_2)-p #plot against covariates
residuals

#plot of the residuals
par(mfrow=c(2,2))
plot(x=Age,y=residuals, ylab="residuals", xlab="age", main="Age vs. Residuals
plot",abline(0,0))
plot(x=Sex,y=residuals, ylab="residuals", xlab="sex", main="Sex vs. Residuals
plot",abline(0,0))
plot(x=CP,y=residuals, ylab="residuals", xlab="CP", main="CP vs. Residuals
plot",abline(0,0))
plot(x=Trestbps,y=residuals, ylab="residuals", xlab="trestbps", main="Trestbps vs.
Residuals plot",abline(0,0))
#Chol, Fbs, Restecg, Thalach vs. Residuals
par(mfrow=c(2,2))
plot(x=Chol,y=residuals, ylab="residuals", xlab="chol", main="Chol vs. Residuals
plot",abline(0,0))
plot(x=Fbs,y=residuals, ylab="residuals", xlab="fbs", main="Fbs vs. Residuals
plot",abline(0,0))
plot(x=Restecg,y=residuals, ylab="residuals", xlab="restecg", main="Restecg vs.
Residuals plot",abline(0,0))
plot(x=Thalach,y=residuals, ylab="residuals", xlab="thalach", main="Thalach vs.
Residuals plot",abline(0,0))
#Exang, Oldpeak, Slope, CA vs. Residuals
par(mfrow=c(2,2))
plot(x=Exang,y=residuals, ylab="residuals", xlab="Exang", main="Exang vs. Residuals
plot",abline(0,0))
plot(x=Oldpeak,y=residuals, ylab="residuals", xlab="oldpeak", main="Oldpeak vs.
Residuals plot",abline(0,0))
plot(x=Slope,y=residuals, ylab="residuals", xlab="slope", main="Slope vs. Residuals
plot",abline(0,0))
plot(x=CA,y=residuals, ylab="residuals", xlab="CA", main="CA vs. Residuals
plot",abline(0,0))
#Thal vs. Residuals
plot(x=Thal,y=residuals, ylab="residuals", xlab="thal", main="Thal vs. Residuals
plot",abline(0,0))
```

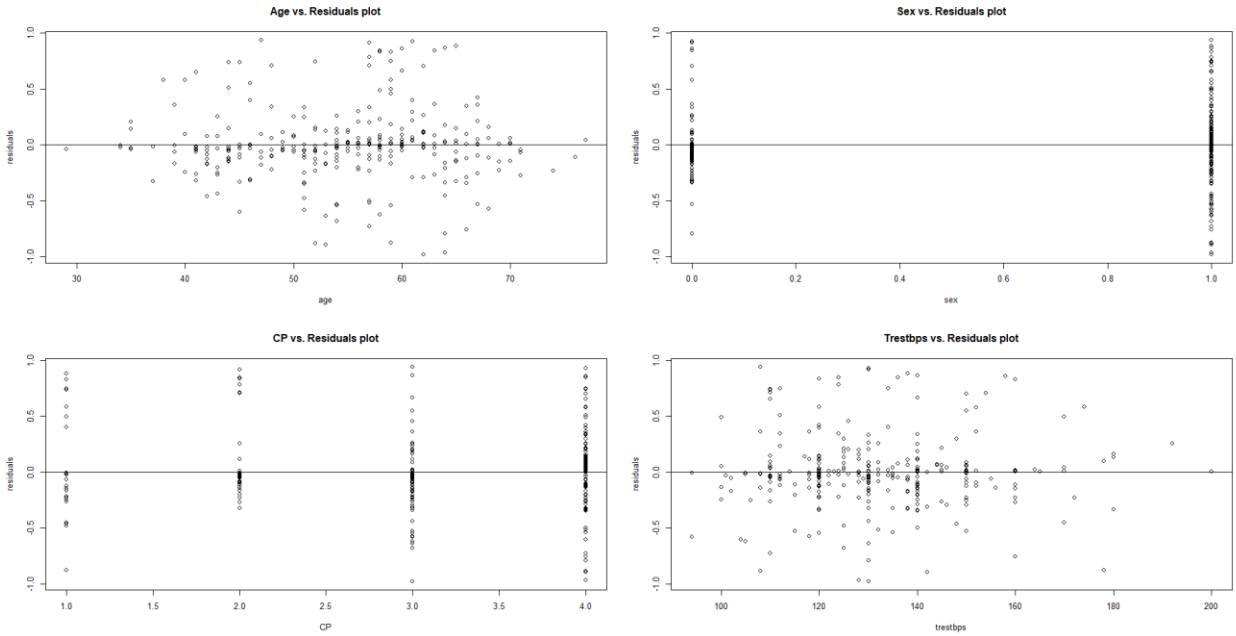
Appendix B:

The image below are the predicted values of the 297 observational values given by the regression model containing all thirteen independent variables and the one outcome variable.

```
> fitted(fit)
          1           2           3           4           5           6           7           8           9          10          11
0.266023854 0.997764279 0.991842631 0.323325153 0.022484651 0.028863589 0.891871550 0.128500498 0.913722320 0.875336134 0.501415413
         12          13          14          15          16          17          18          19          20          21          22
0.086533628 0.795574613 0.126762917 0.230124330 0.091793115 0.287800504 0.202787191 0.045391327 0.050629579 0.163284916 0.009683734
         23          24          25          26          27          28          29          30          31          32          33
0.164548789 0.913632160 0.992709495 0.044329073 0.021148689 0.122794426 0.252397688 0.905426504 0.147879399 0.859715517 0.134412993
         34          35          36          37          38          39          40          41          42          43          44
0.541638029 0.149556382 0.126443366 0.924605405 0.984376926 0.976158129 0.292439373 0.988727066 0.246228796 0.272526908 0.058907168
         45          46          47          48          49          50          51          52          53          54          55
0.071850591 0.768941864 0.088895099 0.916200131 0.138902203 0.166976337 0.020045125 0.326467354 0.490004874 0.042915575 0.910945675
         56          57          58          59          60          61          62          63          64          65          66
0.986775924 0.747413464 0.347438015 0.680160399 0.478061234 0.666974259 0.310167271 0.997939951 0.010171906 0.893748866 0.996113932
         67          68          69          70          71          72          73          74          75          76          77
0.333888526 0.530534371 0.990105697 0.450490342 0.059906776 0.787691535 0.993260619 0.851867259 0.259494364 0.148035389 0.975841043
         78          79          80          81          82          83          84          85          86          87          88
0.339492238 0.095640838 0.935563961 0.602949468 0.172399397 0.170230471 0.837438951 0.045690568 0.114754336 0.178317685 0.072662268
         89          90          91          92          93          94          95          96          97          98          99
0.055961022 0.341477758 0.996557221 0.979124477 0.013889378 0.029476768 0.845899626 0.944838826 0.950223939 0.165698905 0.104747741
        100         101         102         103         104         105         106         107         108         109         110
0.112957226 0.023592402 0.229606172 0.046090610 0.977491767 0.144431324 0.814857373 0.800063033 0.964515131 0.639611346 0.782741359
        111         112         113         114         115         116         117         118         119         120         121
0.701586831 0.063759833 0.745249233 0.735865512 0.320146602 0.044226980 0.041292323 0.993426687 0.961264627 0.941362868 0.995164946
        122         123         124         125         126         127         128         129         130         131         132
0.249434931 0.982875047 0.117173265 0.033282479 0.996578952 0.958819210 0.030518396 0.024416705 0.543878354 0.581359235 0.037095974
        133         134         135         136         137         138         139         140         141         142         143
0.346529177 0.030933042 0.056086252 0.937698468 0.883502979 0.858784745 0.116933776 0.105749359 0.501987023 0.008446257 0.820846099
        144         145         146         147         148         149         150         151         152         153         154
0.622379612 0.061416901 0.995279834 0.044424346 0.087037384 0.052644432 0.126393910 0.168302653 0.529041593 0.985056495 0.968859198
        155         156         157         158         159         160         161         162         163         164         165
0.991779318 0.751269353 0.916501465 0.972897136 0.573872763 0.031819353 0.958935199 0.032519856 0.134892497 0.221320694 0.518992775
        166         167         168         169         170         171         172         173         174         175         176
0.083600125 0.795153823 0.017201046 0.982212750 0.894376276 0.417305136 0.289507690 0.979744148 0.990800716 0.882206136 0.977704163
        177         178         179         180         181         182         183         184         185         186         187
0.437747697 0.638730394 0.655892546 0.986251527 0.461437974 0.876452317 0.138147181 0.088332626 0.173859622 0.990860337 0.743251443
        188         189         190         191         192         193         194         195         196         197         198
0.990472974 0.061436470 0.999231806 0.888216182 0.118000514 0.950536321 0.227949918 0.329094632 0.013411320 0.167977516 0.044930473
        199         200         201         202         203         204         205         206         207         208         209
0.336687559 0.230433676 0.207082299 0.265780778 0.998154887 0.995267763 0.929538573 0.054474305 0.297475541 0.015977971 0.418128775
        210         211         212         213         214         215         216         217         218         219         220
0.262472872 0.903810315 0.252000298 0.219642897 0.005062863 0.322794891 0.793346701 0.169493843 0.061248146 0.021088901 0.006433250
        221         222         223         224         225         226         227         228         229         230         231
0.995502370 0.637823567 0.006637393 0.115199308 0.096228926 0.855294451 0.655567153 0.050336480 0.867641667 0.885395340 0.230686273
        232         233         234         235         236         237         238         239         240         241         242
0.113345468 0.987369412 0.940988171 0.601365619 0.028899130 0.052643140 0.038047054 0.017819123 0.044802130 0.597899634 0.015926660
        243         244         245         246         247         248         249         250         251         252         253
0.578296678 0.510409243 0.905620247 0.865321752 0.029522982 0.727537469 0.959368087 0.966089785 0.046899071 0.204478562 0.023817843
        254         255         256         257         258         259         260         261         262         263         264
0.253918686 0.111190588 0.143324246 0.215459862 0.143455146 0.152676332 0.010458629 0.055624676 0.936451998 0.923019697 0.543665043
        265         266         267         268         269         270         271         272         273         274         275
0.421368119 0.083152296 0.955200517 0.756922854 0.995874120 0.065842933 0.251350357 0.453642798 0.290835607 0.060142554 0.292535296
        276         277         278         279         280         281         282         283         284         285         286
0.102747769 0.963839148 0.062636096 0.885849722 0.029159015 0.704818223 0.996886646 0.958136492 0.202256319 0.085660647 0.638399886
        287         288         289         290         291         292         293         294         295         296         297
0.025361490 0.854024323 0.989996143 0.153449713 0.018367679 0.978654041 0.662745084 0.260910059 0.936749005 0.946755042 0.084023319
```

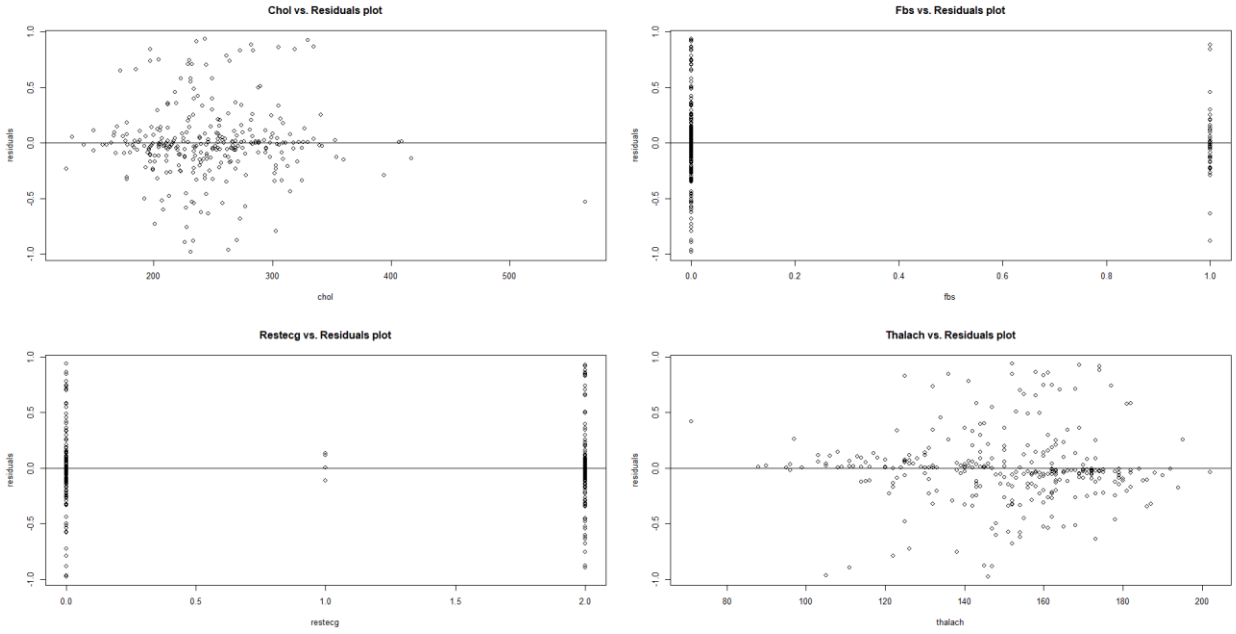
Appendix C(a):

Residual plots for independent variables Age, Sex, CP and Trestbps.



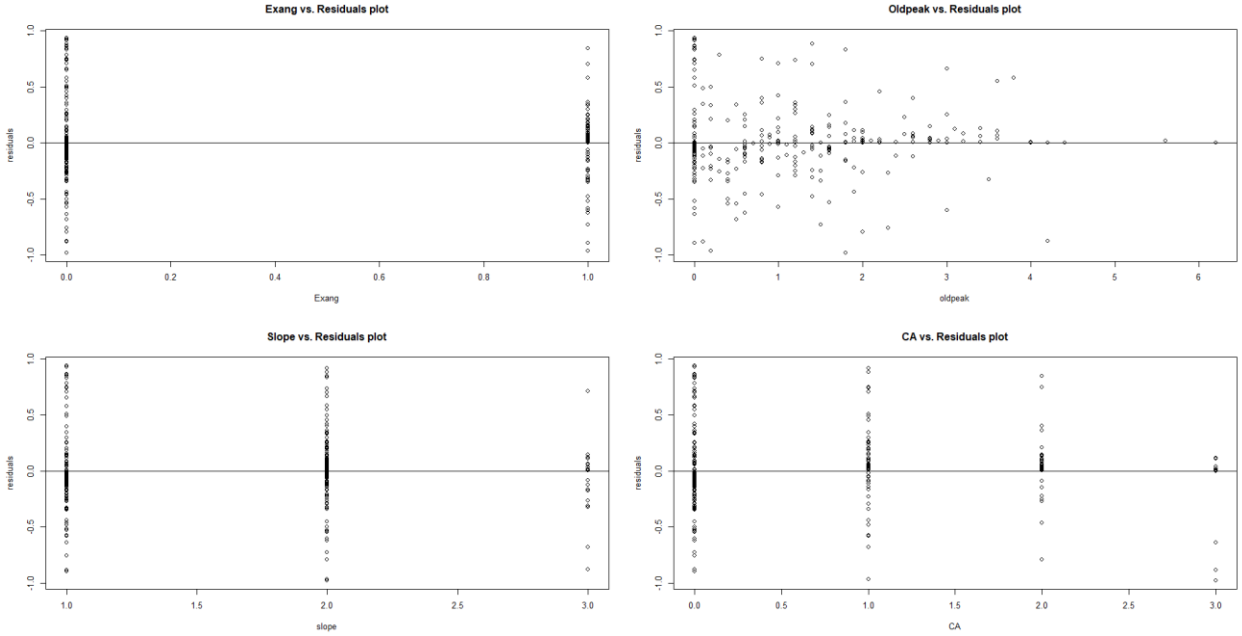
Appendix C(b):

Residuals plot for independent variables Chol, Fbs, Restecg and Thalach



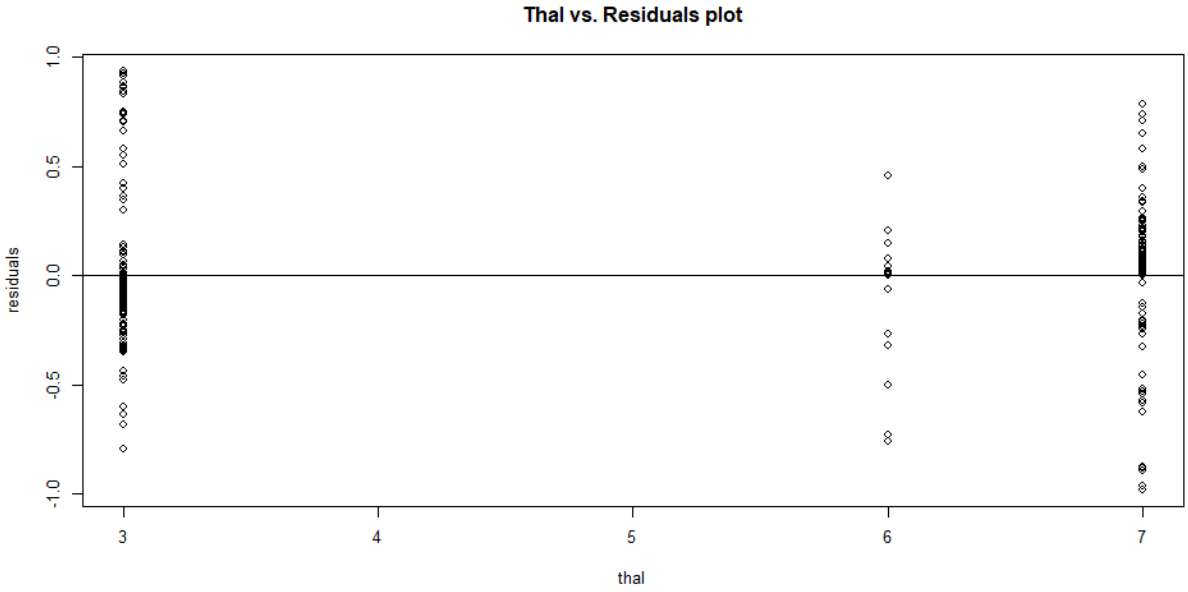
Appendix C(c):

Residuals plot for independent variables Exang, Oldpeak, Slope and CA.



Appendix C(d):

Residuals plot for the independent variable, Thal.



Appendix D(a): Fit 1

The image below represents the glm() command function for the full model with all 13 independent variables against the dependent variable.

```
Call:
glm(formula = as.numeric(Num_2) ~ Age + Sex + CP + Trestbps +
    Chol + Fbs + Restecg + Thalach + Exang + Oldpeak + Slope +
    as.numeric(CA) + as.numeric(Thal), family = binomial, data = Cleveland2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7818  -0.5207  -0.1863   0.4248   2.3622

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.372042   2.879476  -2.560  0.01046 *
Age             -0.014164   0.023970  -0.591  0.55459
Sex              1.312073   0.488474   2.686  0.00723 **
CP               0.575898   0.191197   3.012  0.00259 **
Trestbps         0.024044   0.010730   2.241  0.02504 *
Chol             0.004995   0.003774   1.324  0.18561
Fbs             -1.021918   0.555330  -1.840  0.06574 .
Restecg          0.245153   0.185005   1.325  0.18513
Thalach         -0.020665   0.010225  -2.021  0.04327 *
Exang            0.926104   0.413343   2.241  0.02506 *
Oldpeak          0.247386   0.211832   1.168  0.24287
Slope            0.570009   0.363085   1.570  0.11644
as.numeric(CA)   1.267719   0.265384   4.777 1.78e-06 ***
as.numeric(Thal) 0.343936   0.100361   3.427  0.00061 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 204.69  on 283  degrees of freedom
AIC: 232.69

Number of Fisher Scoring iterations: 6
```

Appendix D(b): Fit 2

The image below represents the glm() command function for the reduced model with 12 independent variables against the dependent variable.

```
> summary(fit2)

Call:
glm(formula = as.numeric(Num_2) ~ Sex + CP + Trestbps + Chol +
    Fbs + Restecg + Thalach + Exang + Oldpeak + Slope + as.numeric(CA) +
    as.numeric(Thal), family = binomial, data = Cleveland2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7823  -0.5345  -0.1748   0.4222   2.3776

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.203028   2.518247  -3.257 0.001124 **
Sex              1.346131   0.483665   2.783 0.005383 **
CP               0.583865   0.190616   3.063 0.002191 **
Trestbps         0.022352   0.010309   2.168 0.030149 *
Chol             0.004652   0.003708   1.255 0.209569
Fbs             -1.031146   0.552085  -1.868 0.061800 .
Restecg          0.243574   0.184921   1.317 0.187779
Thalach         -0.018347   0.009402  -1.951 0.051023 .
Exang            0.942382   0.411836   2.288 0.022123 *
Oldpeak          0.260719   0.210961   1.236 0.216509
Slope            0.561520   0.363142   1.546 0.122036
as.numeric(CA)   1.224859   0.253198   4.838 1.31e-06 ***
as.numeric(Thal) 0.340860   0.099949   3.410 0.000649 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 205.04  on 284  degrees of freedom
AIC: 231.04

Number of Fisher Scoring iterations: 6
```

Appendix D(c): Fit 3

The image below represents the glm() command function for the reduced model with 11 independent variables against the dependent variable.

```
> summary(fit3)

Call:
glm(formula = as.numeric(Num_2) ~ Sex + CP + Trestbps + Chol +
    Fbs + Restecg + Thalach + Exang + Slope + as.numeric(CA) +
    as.numeric(Thal), family = binomial, data = Cleveland2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7863  -0.5246  -0.1751   0.4468   2.3492

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -8.474113   2.517806  -3.366 0.000764 ***
Sex               1.431549   0.479022   2.988 0.002804 **
CP                0.573804   0.190700   3.009 0.002622 **
Trestbps          0.023568   0.010160   2.320 0.020362 *
Chol              0.005059   0.003677   1.376 0.168861
Fbs              -1.074331   0.547571  -1.962 0.049763 *
Restecg           0.233448   0.183402   1.273 0.203062
Thalach          -0.019485   0.009342  -2.086 0.037010 *
Exang             1.007935   0.409139   2.464 0.013757 *
Slope             0.782011   0.313735   2.493 0.012682 *
as.numeric(CA)    1.274038   0.248963   5.117  3.1e-07 ***
as.numeric(Thal)  0.344010   0.099790   3.447 0.000566 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 206.59  on 285  degrees of freedom
AIC: 230.59

Number of Fisher Scoring iterations: 6
```

Appendix D(d): Fit 4

The image below represents the glm() command function for the reduced model with 10 independent variables against the dependent variable.

```
> summary(fit4)

Call:
glm(formula = as.numeric(Num_2) ~ Sex + CP + Trestbps + Chol +
    Fbs + Thalach + Exang + Slope + as.numeric(CA) + as.numeric(Thal),
    family = binomial, data = Cleveland2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8670  -0.5566  -0.1748   0.4547   2.3489

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -8.591527   2.502261  -3.434 0.000596 ***
Sex               1.498722   0.475877   3.149 0.001636 **
CP                0.569527   0.190198   2.994 0.002750 **
Trestbps          0.024695   0.010038   2.460 0.013888 *
Chol              0.005910   0.003607   1.638 0.101328
Fbs              -1.068004   0.544658  -1.961 0.049894 *
Thalach          -0.019843   0.009351  -2.122 0.033840 *
Exang             1.021666   0.408926   2.498 0.012475 *
Slope             0.822393   0.308011   2.670 0.007585 **
as.numeric(CA)    1.280943   0.245781   5.212 1.87e-07 ***
as.numeric(Thal)  0.330230   0.097895   3.373 0.000743 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 208.22  on 286  degrees of freedom
AIC: 230.22

Number of Fisher Scoring iterations: 6
```

Appendix D(e): Fit 5

The image below represents the glm() command function for the reduced model with 9 independent variables against the dependent variable.

```
> summary(fit5)

Call:
glm(formula = as.numeric(Num_2) ~ Sex + CP + Trestbps + Fbs +
    Thalach + Exang + Slope + as.numeric(CA) + as.numeric(Thal),
    family = binomial, data = Cleveland2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8752  -0.5573  -0.1878   0.4730   2.4683

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.248048   2.312335  -3.135 0.001721 **
Sex              1.253367   0.443325   2.827 0.004696 **
CP               0.568646   0.188831   3.011 0.002600 **
Trestbps         0.025231   0.009945   2.537 0.011178 *
Fbs             -1.050740   0.540131  -1.945 0.051734 .
Thalach         -0.018587   0.009144  -2.033 0.042076 *
Exang            1.002225   0.403071   2.486 0.012902 *
Slope            0.805342   0.305314   2.638 0.008346 **
as.numeric(CA)   1.273935   0.241789   5.269 1.37e-07 ***
as.numeric(Thal) 0.343888   0.097930   3.512 0.000445 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 210.90  on 287  degrees of freedom
AIC: 230.9

Number of Fisher Scoring iterations: 6
```

Appendix E(a): Full Model

Output below represents the odds ratio, AUC, confusion matrix and accuracy for the full model
with the 13 independent variables

```
> Odds_Ratio1
 [1] 0.0006285834 0.9859361762 3.7138658483 1.7787278261 1.024335428C
 1.0050077206 0.3599040993 1.2778170041 0.9795467088 2.5246544794
[11] 1.2806736397 1.7682826563 3.5527377615 1.4104886307
> AUC1
[1] 0.8462591
> accuracy1
[1] 0.8484848
> confusion_matrix1
     Predicted
True   0   1
   0 140  20
   1  25 112
```

Appendix E(b): Best Regressed Reduced Final Model

Output below represents the odds ratio, AUC, confusion matrix and accuracy for the best regressed reduced model with the 10 independent variables, “sex”, “CP”, “trestbps”, “chol”, “fbs”, “thalach”, “exang”, “slope”, “CA”, and “thal”.

```
> Odds_Ratio4
 [1] 0.0001856724 4.4759652172 1.7674304591 1.0250019451 1.0059273520 0.3436937319 0.9803529546
 2.7778178638 2.2759401183 3.6000318488 1.3912881957
> Num_2<- as.numeric(Num_2)
> pred4<-predict(fit4,Cleveland2,type="response")
> Cleveland2$pred4<- ifelse(pred4>=0.5,1,0)
> AUC4<- auc(Cleveland2$Num_2, Cleveland2$pred4)
> AUC4
[1] 0.8525091
> accuracy4<- accuracy(Cleveland2$Num_2, Cleveland2$pred4)
> accuracy4
[1] 0.8552189
> confusion_matrix4<- table(Cleveland2$Num_2, Cleveland2$pred4, dnn=c("True","Predicted"))
> confusion_matrix4
    Predicted
True   0   1
   0 142  18
   1  25 112
```

References

[1] Government of Canada. (February 9, 2017). *Heart Disease in Canada.* Retrieved from https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html

[2] UCI Machine Learning. *Heart Disease Data Set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/heart+disease

[3] Hosmer, David W., and Stanley Lemeshow. (2000). *Applied Logistic Regression, Second Edition*. Wiley- Interscience Publication, New York.

[4] Frank E. Harrel, Jr. (2015). *Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Second Edition.* Springer, New York.

[5] Aniruddha Bhandari. (2020). Analytics Vidhya. *Everything you Should Know about Confusion Matrix for Machine Learning.* Retrieved from https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

[6] Data School. (2014). *ROC curves and Area Under the Curve explained (video).* Retrieved from https://www.dataschool.io/roc-curves-and-auc-explained/

[7] Mayo Clinic. (2019). *Thalassemia- Symptoms and Causes.* Retrieved from https://www.mayoclinic.org/diseases-conditions/thalassemia/symptoms-causes/syc-20354995

[8] Hyan Kang. (2013). NCBI. *The prevention and handling of missing data.* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/

[9] Jason T. Newsom. (2016). Categorical Data Analysis. *Diagnostics For Logistic Regression.* Retrieved from http://web.pdx.edu/~newsomj/cdaclass/ho_diagnostics.pdf

[10] Sadanori Konishi, Genshior Kitagawa. (2008). *Information Criteria and Statistical Modeling*. Springer, New York.