APPLICATION OF LOGISTIC REGRESSION IN PREDICTION OF CORONARY HEART DISEASE (CHD)

A Project report

Submitted in partial fulfillment for the award of

M.Sc. Degree Statistics

of Kannur University

2017-2019

BY

FATHIMATH SHAMRIN M A

Department of Statistics

Nehru Arts & Science College, Kanhangad

Kasaragod, 671314

CERTIFICATE

This is to certify that Ms.FATHIMATH SHAMRIN M A, has done the project work entitled“**APPLICATION OF LOGISTIC REGRESSION IN PREDICTION OF CORONARY HEART DISEASE (CHD)”** in partial fulfillment for the award of the M.sc Degree in Statistics of the Kannur University, under our supervision and guidance.

Name of the student	:	**FATHIMATH SHAMRIN M A**
Register. No	:	**B7PSST1002**
Year	:	**2019**

Geethu K. K.
Assistant Professor
Department of Statistics
(Project Guide)

Dr. K. Radhakrishnan Nair
Head of the Department
Dept. of Statistics

ACKNOWLEDGEMENT

I hereby acknowledge my sincere thanks to all the personals who helped me for conducting the project work. First of all, I thank Dr. K. Radhakrishnan Nair, the HOD of Department of Statistics, for his creative guidance and constructive suggestion during the entire course of this project.

I also express my sincere thanks to Assistant professor Geethu K. K., my project guide, who inspired me and guided me with valuable suggestions for all the stages of the project work. Also I would like to thank her for her consent guidance and valuable suggestions given to me for the improvement and finalization of this project.

I am grateful to Dr. T. Vijayan, Principal, Nehru Arts and Science College Kanhangad, for giving all support and encouragement.

I am also thankful to all the members of the Dept. of Statistics for their immense help.

I am also thankful to the staff of C K Nair Computer Center for helping us in making the project report. I once again thank all the personals who helped me in this work.

Place : Padnekad
date : 04-04-2019

Contents

Introduction

The term regression was first introduced by Sir Francis Galton (1822-1911) - a British anthropologist and meteorologist in his paper “Regression towards mediocrity in hereditary stature”in 1885 in which he explained the tendency of offspring to be smaller than large parents and larger than small parents referred to as “regression towards the mean”. In most model fitting situations today, there is no element of ‘regression’ in the original sense. Neverthless the word is so established that we continue to use it.

Regression analysis is a method for investigating functional relationships among variables. The relationship is expressed in the form of an equation or a model connecting the response or dependent variable and one or more explanatory or predictor variables. When the response variable is quantitative, the usual theory of Multiple Linear Regression (MLR) analysis holds good. However, situations where the response variable is qualitative are quite common and occur extensively in statistical applications.

For eg: to determine the risk factors for cancer in humans, data could be collected on several variables such as age, sex, smoking, diet, etc. The response variable here is dichotomous that either the person has cancer ($Y = 1$) or did not have cancer ($Y = 0$). In such cases, the usual MLR theory is not appropriate. Rather the statistical model preferred for the analysis of such binary (dichotomous) responses is the binary logistic regression model developed primarily by Cox and Walker and Duncan(1967). Thus logistic regression is a mathematical modelling approach that can be used to describe the relationship of several independent variables to a binary (dichotomous) dependent variable. Later on, the models to deal with polytomous (multinomial) responses evolved. Over the last decade the logistic regression model has become, in many fields the standard method of analysis in this situation.

Before beginning a study of logistic regression, it is important to understand that the goal of an analysis using this method is same as that of any model-building technique used in statistics to find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables. These independent variables are often called co-variates. The most common example of modelling is the usual linear regression model where the outcome variable is assumed to be continuous.

Logistic regression uses regression to predict the outcome of a categorical dependent variable on the basis of predictor variables. The probable outcomes of a single trial are modeled as a function of the explanatory variable using a logistic function. Logistic modelling is done on categorical data which may be of various types including binary and nominal. Another objective of logistic regression is to check if the probability of getting a particular value of the dependent variable is related to the independent variable. Multiple logistic regression is used when there are more than one independent variable under study.

Logistic regression‘s history can be traced back to the 19th century when it was first used to describe the growth rate of populations by Quetelet and Verhulst. Today logistic regression is widely used in the field of medicine and biology. Epidemiology is also an area where logistic regression is widely used for identification of risk factors for diseases and to plan for preventive medication. Studies concerned with public health and related policy decisions use logistic regression as an important statistical tool.

An important aspect of medical research is the prediction of various diseases and the analysis of factors that cause them. In this work, we focus on Heart disease, specifically the University of California (UCI) Heart Disease dataset. The present study uses logistic regression model to investigate factors that contribute significantly to enhancing the risk of Coronary Heart Disease (CHD). For analyzing this problem, we observe whether a person has or does not have Coronary Heart Disease (CHD) and investigate fourteen independent variables that may be factors affecting the Coronary Heart Disease risk. Logistic regression analysis allows one to predict probability of a binary dependent variable from a set of independent variables that may be continuous, discrete or a mix of them. Logistic regression method is a powerful technique because it is relatively free of restrictions and it allows analyzing a mix of all types of predictors.

Objectives

(i) To study and apply logistic regression for prediction of Coronary Heart Disease(CHD).

(ii) To investigate the factors that contribute significantly to enhancing risk of Heart Disease.

(iii) To study the effect of independent variables and their interactions.

(iv) To conduct the Hosmer and Lemeshow test.

(v) To study about the relationship between the variables using odds and odds ratio.

(vi) To test the normality of the data.

Data source and Data set Description

Coronary Heart Disease (CHD) datasets are taken from Data Mining Repository of University of California, Irvine (UCI). The CHD dataset contains 920 instances collected from Cleveland, Hungarian, VA Long Beach and Switzerland out of which Cleveland dataset is most complete and is used in this study. Coronary angiography determines the result of CHD diagnosis.

Cleveland dataset was collected by Dr. Robert Detrano, M.D and PhD Degree holder at V.A Medical centre and it dates from 1988.

It contains 303 cases, six of the cases have incomplete data which have been discarded and 297 examples were used during the experiments. Out of these 137 cases have heart disease and remaining 160 do not.

The 14 attributes of the Cleveland dataset along with the values and data types are as follows:

Feature	**Description**
Age	Age in years
Sex	Instance gender (1-male; 0-female)
Cp	Chest pain type (1-typical angina, 2-atypical angina, 3-non-anginal pain, 4-asymptomatic)
Trestbps	Resting blood pressure in mm/Hg
Chol	Serum cholestrol in mg/dl
Fbs	Fasting blood sugar>120 mg/dl (1-true, 0-false)
Restecg	Resting ECG results (0-normal, 1- ST-T wave abnormality, 2-Left ventricle hypertrophy)
Thalach	Maximum heart rate achieved during thallium stress test
Exang	Exercise induced angina (1-yes, 0-no)
Old peak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment (1-upsloping, 2-flat, 3-downsloping)
Ca	Number of major vessels coloured by flouroscopy (value 0 – 3)
Thal	Defect type (Values 3-normal, 6-fixed defect, 7-reversible defect)
Hd	Diagnosis of Heart disease (0-Healthy, 1-Unhealthy)

Chapter 1

INTRODUCTION TO LOGISTIC REGRESSION MODEL

1.1 Generalized Linear Model (GLM)

Generalized Linear Model (GLM) were introduced in a seminal paper by Nelder and Wedderburn (1972) in which a wide range of seemingly disparate problems of statistical modelling and inference (ANOVA, ANCOVA, multiple regression, Logistic regression, etc..) were set in an elegant unifying framework of great power and flexibility.

GLM assume that response variable has a probability distribution belonging to the exponential family of distributions that includes many distributions like normal, Bernoulli, binomial and Poisson distributions. The assumption specifies the random component of the model or it specifies the probabilistic mechanism by which the responses are assumed to be generated.

For these distributions, the variance of the response can be expressed in terms of the product of a single scale or dispersion parameter ϕ and a variance function denoted $V(\mu_i)$; the latter being a known function of the mean μ_i.

i.e; $V(Y_i)$=$\phi V(\mu_i)$,where ϕ>0.The variance function $V(\mu_i)$ describes how the variance of the response is related to mean of the response.

GLM is a general class of linear models that are made up of three components:

- **Random component** :-refers to the probability distribution of the response variable (Y); e.g. normal distribution for Y in the linear regression, or binomial distribution for Y in the binary logistic regression. Also called a noise model or error model.

- **Systematic component** :- specifies the explanatory variables $(X_1, X_2, ...X_k)$ in the model, more specifically their linear combination in creating the so called linear predictor;

e.g., $\beta_0 + \beta_1\ x_1 + \beta_2\ x_2$

- **Link Function** :-specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables; e.g., $\eta = g(E(Y_i)) = E(Y_i)$ for linear regression, or $\eta = logit(\pi)$ for logistic regression.

Assumptions

GLM operates correctly under following assumptions:

(i) The data $Y_1, Y_2, ..., Y_n$ are independently distributed, i.e., cases are independent.

(ii) The dependent variable Y_i does not need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal,...).

(iii) GLM does not assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the transformed response in terms of the link function and the explanatory variables; e.g., for binary logistic regression

$$logit(\pi) = \beta_0 + \beta_1 X$$

(iv) Independent (explanatory) variables can be even the power terms or some other nonlinear transformations of the original independent variables.

(v) The homogeneity of variance does not need to be satisfied. In fact, it is not even possible in many cases given the model structure, and over dispersion (when the observed variance is larger than what the model assumes) maybe present.

(vi) Errors need to be independent but not normally distributed.

(vii) It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.

(viii) Goodness-of-fit measures rely on sufficiently large samples, where a heuristic rule is that not more than 20% of the expected cells counts are less than 5.

1.2 Odds

Odds of an event are the ratio of the probability that an event will occur to the probability that it will not occur. If the probability of an event occuring is p, the probability of the event not occuring is $1 - p$. Then the corresponding odds is a value given by

$$odds(event) = \frac{p}{1-p}$$

Since logistic regression calculates the probability of an event occuring over the probability of an event not occuring, the impact of independent variables is usually explained in terms of odds. With logistic regression, the mean of the response variable p in terms of an explanatory variable x is modelled relating p and x through the equation $p = \alpha + \beta x$. Unfortunately, this is not a good model because extreme values of x will give values of $\alpha + \beta x$ that does not fall between 0 and 1. The logistic regression solution to this problem is to transform the odds using the natural algorithm. With logistic regression we model the natural log odds as a linear function of the explanatory variable.

$$logit(Y) = ln(odds) = ln\frac{p}{1-p} = \alpha + \beta x \tag{1.1}$$

where p is the probability of interested outcome and x is the explanatory variable. The parameters of logistic regression are α and β.

This is simple logistic model.

Taking the antilog of equation (1.1) on both sides, One can derive an equation for the prediction of the probability of the occurrence of interested outcome as

$$\begin{aligned} p &= p(Y = interested\ outcome/X = x, a\ specific\ value) \\ &= e^{\frac{\alpha+\beta x}{1+\alpha+\beta x}} \\ &= \frac{1}{1+e^{-(\alpha+\beta x)}} \end{aligned}$$

Extending the logit of the simple logistic regression to multiple predictors, one may construct a complex logistic regression as

$$logit(Y) = ln(\frac{p}{1-p}) = \alpha + \beta_1 x_1 + ... + \beta_k x_k \tag{1.2}$$

Therefore,

$$\begin{aligned} p =& p(Y = interested\ outcome/x_i = x_1, x_2, ..x_k) \\ =& \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}} \\ =& \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}} \end{aligned}$$

1.3 Odds Ratio (OR)

The odds ratio is a comparative measure of two odds relative to different events. For two events A and B, the corresponding odds of A occuring relative to B occuring is,

$$odds\ ratio[A\ v/s\ B] = \frac{odds(A)}{odds(B)} = \frac{P_A/1-P_A}{P_B/1-P_B} \tag{1.3}$$

An odds ratio is a measure of association between an exposure and an outcome will occur given a particular exposure compared to the odds of the outcome occuring in the absence of that exposure.

When a logistic regression is calculated, the regression coefficient (b_1) is the estimated increase in the logged odds of the outcome per unit increase in the value of the independent variable. In other words, the exponential function of the regression coefficient (e^{b_1}) is the odds ratio associated with a one unit increase in the independent variable.

The odds ratio also can be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare various risk factors for the outcome.

OR $= 1 \Longrightarrow$ exposure does not affect outcome.
OR $> 1 \Longrightarrow$ exposure associated with higher odds of outcome.
OR $< 1 \Longrightarrow$ exposure associated with lower odds of outcome.

For eg: the variable smoking is coded as 0(=no smoking)and 1(=smoking)and the odds ratio for the variable is 3.2. Then, the odds for a positive outcome in smoking are 3.2 times higher than in non-smoking cases.

1.4 The logistic curve and logistic function

Logistic Regression is a method for fitting a regression curve $y = f(x)$, where y consist of binary coded (0, 1 – failure, success) data. When response is binary variable and x is numerical, logistic regression fits a logistic curve to the relationship between x and y.

Logistic curve is an S-shaped or sigmoid curve, often used to model population growth. A logistic curve starts with slow, linear growth, followed by exponential growth, which then slows again to stable state.

A simple logistic function is defined by the formula

$$Y = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}} \tag{1.4}$$

To provide flexibility, the logistic function can be extended to the form

$$Y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}} \tag{1.5}$$

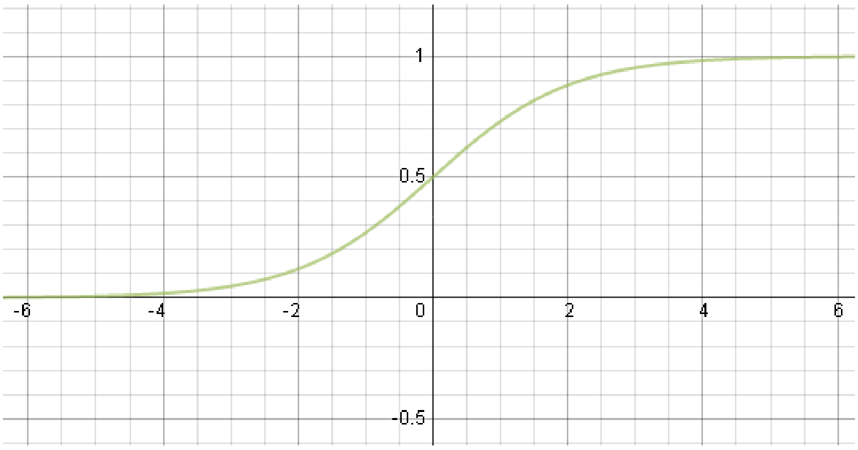
where α and β determine the logistic intercept and slope.

The function associated with the logistic distribution is called logistic function, also known as the logistic curve illustrates the properties of logistic distribution. The logistic function is given by the following formula :

$$f(x) = \frac{L}{1+ce^{-kx}} \tag{1.6}$$

Where, L, C and k are constant terms and e denotes well-known Euler's number representing the base of natural logarithm. Also, the values of x lie between $-\infty$ to $+\infty$, i.e, range of real numbers.

The curve of logistic function is shown below:



Here, the curve of logistic function f approaches L as x tends to $+\infty$ and it approaches to zero as x tends to$-\infty$

1.5 Inverse of logistic function

We can now define the inverse of the logistic function g, the logit(log odds):

$$g(F(Z)) = ln\frac{F(x)}{1-F(x)} = \beta_0 + \beta_1 x$$

and equivalently

$$\frac{F(z)}{1-F(z)} = e^{\beta_0+\beta_1 x}$$

1.6 Logistic Regression

Logistic regression is a model used for prediction of the probability of occurrence of an event. It makes use of several predictors variables that may be either numerical or categorical. Specifically Logistic regression can be used only with two types of target (response or dependent) variables. A categorical target variable that has exactly two categories i.e; a binary or a dichotomous variable and a continuous target variable that has values in the range 0 to 1 representing probability values or proportions.

Some of the assumptions of Logistic regression are:

i Logistic regression requires the dependent variable to be discrete mostly dichotomous.

ii Since Logistic regression estimates the probability of the event occurring (P(Y=1)), it is necessary to code the dependent variable accordingly that is the desired outcome should be coded to be 1.

iii Model should be fitted correctly. Neither it should be over fitted with the meaningless variables included nor it should be under fitted with the meaningful variable excluded.

iv Logistic regression requires each observation to be independent. Also the model should have little or no multicollinearity.

v While Logistic regression does not require a linear relationship between the dependent and independent variable, it requires that the independent variables are linearly related to the log odds of an event.

v Lastly Logistic regression requires large sample sizes because MLE are less powerful than OLS used for estimating unknown parameters in a linear regression model.

Now, we will note two important differences between logistic regression and linear regression:

The first difference concerns the nature of the relationship between the outcome and independent variables. In any regression problem, the key quantity is the mean value of the outcome variable, given the value of the independent variable.This quantity is called conditional mean and will be expressed as $E(Y/X)$ where Y denotes outcome variable and X denotes a value of independent variable. In linear regression, we assume that this mean may be expressed as an equation linear in X such as

$$E(Y/X) = \beta_0 + \beta_1 \tag{1.7}$$

This expression implies that it is possible for $E(Y/X)$ to take on any value as X ranges between $-\infty$ and ∞. With dichotomous data, the conditional mean must be greater than or equal to zero and less than or equal to 1, ($0 \leq E(Y/X) \leq 1$).

Many distribution functions have been proposed for use in the analysis of a dichotomous outcome variable. There are two primary reasons for choosing the logistic distribution. Firstly from a mathematical point of view,it is an extremely flexible and easily used function and second it lends itself to a clinically meaningful interpretation.

In order to simplify notation, we use the quantity $\pi(x) = E(Y/X)$ to represent the conditional mean of Y given X when the logistic regression is used.The specific form of the logistic regression model we use is,

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{1.8}$$

A transformation of $\pi(x)$ is central to our study of logistic regression is the logit transformation. This transformation is defined in terms of $\pi(x)$ as

$$g(x) = ln[\frac{\pi(x)}{1-\pi(x)}] = \beta_0 + \beta_1 x$$

The importance of this transformation is $g(x)$ has many of the desirable properties of a linear regression model. The logit $g(x)$, is linear in its parameters, maybe continuous, and may range from $-\infty$ to ∞, depending upon the range of x.

The second important difference between the linear and logistic regression model concerns the conditional distribution of the outcome variable. In the linear regression model, we assume that an observation of the outcome variable may be expressed as $Y = E(Y/X)+\epsilon$. The most common assumption is that ϵ follows a normal distribution with mean zero and some variance that is constant across levels of the independent variable. It follows that conditional distribution of the outcome variable given X will be normal with mean $E(Y/X)$ and a variance that is constant. This is not the case with a dichotomous outcome variable. In this situation, we may express the value of the outcome variable given x as $Y = \pi(x) + \epsilon$. Here the quantity ϵ may assume one of two possible values. If $Y = 1$, then $\epsilon = 1 - \pi(x)$ with probability $\pi(x)$ and if $Y = 0$, then $\epsilon = -\pi(x)$ with probability $1 - \pi(x)$. Thus ϵ has a distribution with mean 0 and variance equal to $\pi(x)[1 - \pi(x)]$. That is, the conditional distribution of the outcome follows a binomial distribution with probability given by conditional mean $\pi(x)$.

1.7 Logistic Regression for Binary Response

Logistic regression is most common model used to describe the relationship between a binary response variable and a set of co-variates. Typically, the response can be whether or not a patient is cured of a disease,whether or not an item in manufacturing process passes the quality control, whether or not a mouse is killed by toxic exposure in a toxicology experiment.

Let Y_i denote a binary response variable, assuming values for 0 and 1 .The probability distribution of Y_i is Bernoulli with $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$. For ease of exposition, first we shall assume that there is a single co-variate x_i. Expected value of the response variable is $E(Y_i) = \mu_i = \pi_i$. Therefore the analytical goal is to investigate the relationship between π_i and x_i. The linear regression,

$$E(Y_i/X_i) = \beta_0 + \beta_1 x_i$$

will yield predicted probabilities outside the range from 0 to 1 for sufficiently large x_i. Further, we cannot expect always a linear relationship between π_i and x_i. It turns out that one can take care of both these problems by transferring the probability using the logistic function. The logistic function will turn a probability into a quantity which can take any real value and which is often linearly related to the explanatory variables.

The function is specially useful when the response variable plotted against the co-variate gives the sigmoid. A sigmoid curve has the properties that the Y variable (the probability of success) is constrained to lie between 0 and 1 such that Y tends to 0 when x becomes small and Y tends to 1 when x becomes large (or the other way around) and the relationship between the Y- variable and the x-variable is linear from about $Y = 0.2$ to $Y = 0.8$. There are various types of functions that produce sigmoid curves. But the most specific one is the logistic function.

Also the usual assumption of homogenity of variance would be violated since the variance of the binary variable depends on a mean with

$$V(Y_i) = E(Y_i - \pi_i)^2 = \pi_i(1 - \pi_i) = \frac{exp(\beta_0 + \beta_1 x_i)}{(1 + exp(\beta_0 + \beta_1 x_i)^2}$$

Note that $\frac{\pi_i}{1-\pi_i}$ is the odds of success and that the odds of success, the log odds and the probability of success move along the same direction. That is, when π_i increases or decreases, so do the others.

If the logit or logistic function $log(\pi_i/1 - \pi_i)$ is adopted, the resulting model.

$$log(\frac{\pi_i}{1 - \pi_i}) = logit(\pi_i) = \beta_0 + \beta_1 x_i$$

is known as logistic regression model. If the predictor variable x_i is dichotomous taking values 0 or 1.

$$logit(\pi_i/x_i = 1) - logit(\pi_i/x_i = 0) = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$

β_1 is the change in log(odds) for a unit change in x_i, Equivalently, a unit change in x_i changes the odds of success multiplicatively by $exp(\beta_1)$. Thus $exp(\beta_1)$ has the interpretation as the odds ratio of the response for the possible values of the covariates. The logistic regression can also be expressed as

$$\begin{aligned} \pi_i &= \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)} \\ &= \frac{1}{1 + exp(-\beta_0 - \beta_1 x_i)} \end{aligned}$$

The logistic regression is a special case of GLM‘s where Y_i has a Bernoulli distribution and a logit link function, the canonical link function has been adopted. When x_i has j levels, the binary responses for N individuals can be grouped. Let n_j denote the number of individual with the j^{th} covariate pattern $(j = 1, 2, ...N)$ and Y_j denote the number of successes among n_j individuals. We may provisionally assume that all individual within a group respond independently with constant probability of success μ_j, depending only on group. Then Y_j has binomial distribution with probability of success μ_j, with $E(Y_j) = n_j\mu_j$ and $V(Y_j) = (n_j\mu_j)(1 - \mu_j)$

In many biomedical applications there is over dispersion in that, the count of the number of successes have variability that far exceeds that predicted by binomial distribution. To allow over dispersion or extra binomial variation a scale factor ϕ (with $\phi = 1$) is included in the specification of binomial variance,

$$V(Y_j) = \phi \, n_j \mu_j (1 - \mu_j) \tag{1.9}$$

impact on the logistic regression coefficients, if this over dispersion is not accounted for, is negligible as the estimates are still consistent and there is little loss of efficiency. Neglecting over dispersion, however, results in under estimation of standard errors, which in turn may result in misleading inferences, via shrinking the confidence intervals and reducing the p - values. If x_i is a Px1 vector of co-variates, the logistic regression model becomes,

$$\begin{aligned} logit(Y_i) &= log(\frac{\pi_i}{1 - \pi_i}) \\ &= \beta_0 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \end{aligned}$$

Fitting the logistic regression model where $x_{i1} = 1, \forall \, i = 1, 2, ...N$. Here β_k represents change in log odds for a unit change in x_{ik}, given that all other predictor variables remain constant. Equivalently, a unit change in x_{ik} changes the odds of success multiplicatively by a factor $exp(\beta_k)$. If β_i is negative, then $e^{\beta_1} < 1$, and the odds of success is actually reduced by $e^{-\beta_1}$.

Chapter 2

ESTIMATION AND TESTING OF HYPOTHESES

2.1 Fitting the logistic regression model

Suppose we have a sample of n independent observations of the pair $(x_i, y_i), i = 1, 2, ..., n$ where y_i denotes the value of a dichotomous outcome variable and x_i is the value of the independent variable for the i^{th} subject. Furthermore, assume that the outcome variable has been coded as 0 or 1, representing the absence or the presence of the characteristic, respectively.

To fit the logistic regression model in equation (1.8) to asset of data requires that we estimate the value of β_0 and β_1, unknown parameters.

In linear regression, the method of least square is most often used for estimating unknown parameters. In this method, we choose those values of β_0 and β_1 which minimize the sum of squared deviations of the observed values of Y from the predicted values based upon the model. Under the usual assumptions for linear regression, the method of least square yields estimator with a number of desirable statistical properties. Unfortunately, when the method of least squares is applied to a model with a dichotomous outcome, the estimators no longer have these properties.

In logistic regression, the method is more complicated. It is called maximum likelihood method. The general method of estimation that leads to the least squares function under the linear regression model is called maximum likelihood. Maximum likelihood will provide values of β_0 and β_1 which maximize the probability of obtaining the observed data set. It requires iterative computing and is easily done with most computer software.

We use likelihood function to estimate the probability of observed data as a function of unknown parameters, given the unknown parameters (β_0 and β_1)."Likelihood" is a probability, specifically the probability that the observed values of the dependent variable may be predicted from the observed values of the independent variables. Like any probability, the likelihood varies from 0 to 1. The maximum likelihood estimators of these parameters are chosen to be those values that maximize this function. Thus the resulting are those which agree most closely with the observed data.

We now describe how to find these values from the logistic regression model.

If Y is coded as 0 or 1 then the expression for $\pi(x)$ given in equation (1.8) provides (for an arbitrary value of , the vector of parameters) the conditional probability that Y is equal to 1 given x . This is denoted as $P(Y = 1/x)$. It follows that the quantity $1 - \pi(x)$ gives the conditional probability that Y is equal to zero given x, $P(Y = 0/x)$. Thus, for those pairs (x_i, y_i) , where $y_i = 1$, the contribution to the likelihood function is $\pi(x_i)$, and for those pairs where $y_i = 0$, the contribution to the likelihood function is $1 - \pi(x_i)$, where the quantity $\pi(x_i)$ denotes the value of $\pi(x)$ computed at x_i.

A convenient way to express the contribution to the likelihood function for the pair (x_i, y_i) is through the expression

$$(\pi(x_i))^{Y_i}[1 - \pi(x_i)]^{1-Y_i} \tag{2.1}$$

Since, the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in expression (2.1) as follows;

$$l(\beta) = \prod_{i=1}^{n}(\pi(x_i))^{Y_i}[1 - \pi(x_i)]^{1-Y_i} \tag{2.2}$$

The principal of maximum likelihood states that we use as our estimate of β, the value which maximizes the expression in equation (2.2). However, it is easier mathematically to work with the log if equation (2.2) termed as log likelihood is defined as,

$$L(\beta) = ln((l(\beta))) = \sum_{i=1}^{n} Y_i ln[\pi(x_i)] + (1 - Y_i)ln(1 - \pi(x_i)) \tag{2.3}$$

To find the value of β that maximizes $L(\beta)$, we differentiate $L(\beta)$ with respect to β_0 and β_1 and set the resulting expressions equal to zero, these equations, known as the likelihood equations, are:

$$\sum_i [Y_i - \pi(x_i)] = 0 \tag{2.4}$$

$$\sum_i X_i[Y_i - \pi(x_i)] = 0 \tag{2.5}$$

In equations (2.4) and (2.5), the summation is over i varying from 1 to n.

In linear regression, the likelihood equations, obtained by differentiating the sum of squared deviation function with respect to β are linear in the unknown parameters and thus are easily solved.

For logistic regression, the expression in equations (2.4) and (2.5) are non linear in β_0 and β_1, and thus require special methods for their solution. These methods are iterative in nature and have been programmed into available logistic regression software.

An interesting consequence of equation (2.4) is

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \tag{2.6}$$

That is sum of the observed values of Y is equal to the sum of the predicted values.

2.2 Testing of Hypotheses

In linear regression, where the response variable is normally distributed , we can use t or F statistics for testing significance of explanatory variables. But in logistic regression, the response variables are Bernoulli distributed. So we have to use different test statistic, with exact distributions unknown. Fortunately, there exist fairly good approximations to the distributions of the test statistic.

We shall use two different types of test statistics: the (log) likelihood ratio statistic (often referred to as the $-2logQ$ statistic) and the Wald statistic. In general, the likelihood statistic is superior to the Wald statistic (it gives more reliable results). The Wald statistic has the advantage that it is computationally easy and is given automatically in the output of most statistical computer package (eg: SAS).

To calculate the likelihood ratio statistic for testing of an explanatory variable, we compare $-2logL$ for the full model, containing all explanatory variables, with the that for the reduced model, containing all explanatory variables except the one of interest. $-2logL$ is a measure of model fit, that measures how well the given model explains the data (the lower the better). If the full model explains the data 'much better' than the reduced model, the difference D, between $-2logL_{red}$ and $-2logL_{full}$ will be large. In this case, we reject the null hypothesis that the variable is non-significant. The statistic D, known as deviance, has chi-square distribution with 1 d.f. Therefore we compare D with suitable quantities of $\chi^2(1)$ to make the conclusion.

The Wald statistic can be used to assess the contribution of individual predictors or the significance of individual coefficients in a given model. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient. The Wald statistic is asymptotically distributed as a chi-square distribution.

$$W_j = \frac{\beta_j^2}{SE\beta_j^2} \tag{2.7}$$

Each Wald statistic is compared with a chi-square with 1 degree of freedom.

Hauck and Donner examined the performance of the Wald test and found that it behaved in an aberrant manner, often failing to reject null hypothesis when the coefficient was significant. They recommend that likelihood ratio test be used. Jennings has also looked at adequacy of inferences in logistic regression based on Wald statistic and made similar conclusions.

2.3 Confidence Interval Estimation

The basis for construction of the interval estimators is the same statistical theory we used to formulate the tests for significance of the model. In particular the confidence interval estimators for the slope and intercept are based on their respective Wald tests. The endpoints of a $100(1-\alpha)\%$ confidence interval for the slope coefficient are:

$$\hat{\beta}_1 \pm Z_{1-\alpha/2}\hat{SE}(\hat{\beta}_1)$$

and for intercept they are

$$\hat{\beta}_0 \pm Z_{1-\alpha/2}\hat{SE}(\hat{\beta}_0)$$

where, $Z_{1-\alpha/2}$ is the upper $100(1-\alpha/2)\%$ point from the standard normal distribution and $S\hat{E}(.)$ denotes a model-based estimator of the standard error of the respective parameter estimator.

2.4 Interpretation of fitted logistic regression model

The interpretation of any fitted model requires that we be able to draw practical inferences from the estimated coefficients in the model.

The interpretation involves two issues: determining the functional relationship between the dependent variable and the independent variable, and approximately defining the unit of change for the independent variable.

The first step is to determine what function of the dependent variable yields a linear function of the independent variables.This is called link function.

2.5 Link Function

A link function is a function of the dependent variable which yields a linear function of the independent variable.

Examples of link function

1. Identity function idN,in linear regression.
2. In logistic regression,its the logit transformation

$$g(x) = ln(\tfrac{\pi(x)}{1-\pi(x)}) = \beta_0 + \beta_1 x \text{ , with } \pi(x) = \tfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

2.6 Logit Function

The logistic equation is stated in terms of the probability that $Y = 1$, which is π, and the probability that $Y = 0$, which is $1 - \pi$

$$ln(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x \tag{2.8}$$

The left-hand side of the equation represents the logit transformation, which takes the natural log of the ratio of the probability that Y is equal to 1 compared to the probability that it is not equal to 1. As we know, the probability π is just the mean of the Y-values assuming 0,1 coding which is often expressed as μ. The logit transformation could then be written in terms of the mean rather than the probability

$$ln(\frac{\mu(x)}{1-\mu(x)}) = \alpha + \beta x \tag{2.9}$$

The transformation of the mean represents a link to the central tendency of the distribution sometimes called the location, one of the important defining aspects of any given probability distribution. The log transformation represents a kind of link function that is sometimes given more generally as $g(.)$. For logistic regression, this is known as the logit link function.

2.7 Stepwise Algorithm

Employing a stepwise selection procedure can provide a fast and effective means to screen a large number of variables and to fit a number of logistic regression equations simultaneously.

Stepwise regression is a way to build a model by adding or removing predictor variables, usually via a series of F-tests or T-tests. The variables to be added or removed are chosen based on the test statistics of the estimated coefficients. While the technique does have its benefits, it requires skill on the part of the researcher so should be performed by people who are very familiar with statistical testing. In essence, unlike most regression models, the models created with stepwise regression should be taken with a grain of salt; they require a keen eye to detect whether they make sense or not.

There are two types of Stepwise Regression

1. Forward
2. Backward

Forward stepwise(FSTEP)

(1) if FSTEP is the first method requested, estimate the parameter and likelihood function for the initial model. Otherwise, the final model from the previous method is the initial model for FSTEP. Obtain the necessary information: MLEs of the parameter for the current model, predicted probability $\hat{\pi}_1$, likelihood function for the current model, and so on.

(2) Based on the MLEs of the current model, calculate the score statistic for every variable eligible for inclusion and find its significance.

(3) Choose the variable with the smallest significance. If that significance is less than the probability for a variable to enter, then go to step 4; otherwise, stop FSTEP.

(4) Update the current model by adding a new variable. If this results in a model which has already been evaluated, stop FSTEP.

(5) Calculate LR or Wald statistic or conditional statistic for each variable in the current model. Then calculate its corresponding significance.

(6) Choose the variable with the largest significance. If that significance is less than the probability for variable removal, then go back to step (2); otherwise, if the current model with the variable deleted is the same as a previous model, stop FSTEP; otherwise, go to next step.

(7) Modify the current model by removing the variable with the largest significance from the previous model. Estimate the parameters for the modified model and go back to step (5).

Backward stepwise(BSTEP)

(1) Estimate the parameters for the full model which includes the final model from the previous method and all eligible variables. Only variables listed on the BSTEP variable list are eligible for entry and removal. Let the current model be the full model.

(2) Based on the MLEs of the current model, calculate the LR or Wald statistic or conditional statistic for every variable in the model and find its significance.

(3) Choose the variable with the largest significance. If that significance is less than the probability for a variable removal, then go to step (5); otherwise, if the current model without the variable with the largest significance is the same as the previous model, stop BSTEP; otherwise, go to the next step.

(4) Modify the current model by removing the variable with the largest significance from the model. Estimate the parameters for the modified model and go back to step (2).

(5) Check to see any eligible variable is not in the model. If there is none, stop BSTEP; otherwise, go to the next step.

(6) Based on the MLEs of the current model, calculate the score statistic for every variable not in the model and find its significance.

2.8 Dichotomous Independent Variable

We assume that the independent variable X is nominal scaled and dichotomous. In most cases that means its coded as either 0 or 1. Now calculating the slope β_1 is straight forward and is done in few steps.

$$g(1) - g(0) = (\beta_0 + \beta_1 * 1) - (\beta_0 + \beta_1 * 0) = (\beta_0 + \beta_1) - (\beta_0) = \beta_1$$

(1) Define two values of the covariate to be compared.

(2) Substitute those two in the logit $(g(1), g(0))$

(3) Calculate the difference $g(1) - g(0)$

Inorder to interpret the result, we need to discuss in terms of odds ratio.

The possible values of the logistic probabilities may be conviniently displayed in a 2x2 table. The odds of the outcome being present among individuals with $x = 1$ is defined as $\frac{\pi(1)}{1-\pi(1)}$. Similarly odds of outcome being present among individuals with $x = 0$ is defined as $\frac{\pi(0)}{1-\pi(0)}$. The odds ratio of the odds for $x = 1$ to the odds for $x = 0$, and is given by the equation

$$OR = \frac{\pi(1)/1 - \pi(1)}{\pi(0)/1 - \pi(0)} \tag{2.10}$$

Substituting the logistic regression model probabilities into the OR, we obtain that the relationship between the odds ratio and the regression coefficient is

$$OR = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \Big/ \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} \Big/ \frac{1}{1+e^{\beta_0}}} \tag{2.11}$$

$$= e^{\beta_0+\beta_1} \big/ e^{\beta_0} = e^{\beta_0+\beta_1-\beta_0} = e^{\beta_1} \tag{2.12}$$

Hence for logistic regression with a dichotomous independent variable coded 1 and 0, the relationship between the odds ratio and regression coefficient is

$$OR = e^{\beta_1} \tag{2.13}$$

This relation between coefficient and odds ratio is the reason why logistic regression has proven to be such a powerful analytic research tool.

Thus as a fourth and final step we‘ve,

1. Define two values of the covariate to compared $(x = 1, x = 0)$
2. Substitue those two into the logit $(g(1), g(0))$
3. Calculate the difference $(g(1) - g(0))$
4. Exponentiate the logit difference to obtain an odds ratio.

The odds ratio approximates how much more likely or unlikely it is for the outcome to be present among those subjects with $x = 1$ compared to those subjects with $x = 0$.

For eg: assume that Y is the presence or absence of Heart disease and x denote whether or not a person engages in regular strenuous physical exercise. If the odds ratio is $OR = 0.5$ ($OR = 2$), then odds of heart disease among those subjects who exercise is one-half (twice) the odds of heart disease for those subjects who do not exercise in the study population.

$\hat{OR}$ tends to have a distribution that is highly skewed to the right, due to the fact that its range is between 0 and ∞ only for extremely large sample sizes, the distribution would be normal. Hence, inferences are usually based on the sampling distribution of $ln(\hat{OR}) = \beta_1$ which tends to follow a normal distribution for much smaller sample sizes.

A confidence $100(1 - \alpha)\%$ intervals for the $\hat{OR}$ is given by

$$exp[\hat{\beta}_1 \pm Z_{1-\alpha/2} \hat{SE(} \hat{\beta}_1)] \tag{2.14}$$

In summary for dichotomous variable the parameter of interest is the odds ratio. An estimate of parameter may be obtained from the estimated logistic regression coefficient, regardless of how the variable is coded. The relationship between the logistic regression coefficient and the odds ratio provide the foundation for our interpretation of all logistic regression results.

2.9 The Hosmer-Lemeshow Test

The Hosmer-Lemeshow test is a statistical test for goodness of fit for logistic regression models. It is used frequently in risk prediction models. The test assesses whether or not the observed event rates match expected event rates in the subgroups of the model population. The Hosmer-Lemeshow test is specifically identifies subgroups as the deciles of fitted risk values. Models for which expected and observed event rates in sub groups are similar are called calibrated.

Hosmer and Lemeshow (1980) and Lemeshow and Hosmer proposed grouping based on the values of the estimated probabilities. Suppose for sake of discussion that $J = n$. In this case we think of the n columns as corresponding to the n values of the estimated probabilities, with the first column corresponding to the smallest value, and the n^{th} column to the largest value. Two grouping strategies were proposed as follows:

1. Collapse the table based on percentiles of the estimated probabilities.
2. Collapse the table based on fixed values of the estimated probability.

With the first method, use of $g = 10$ groups results in the first group containing the $n_1 = n/10$ subjects having the smallest estimated probabilities, and the last group containing $n_{10} = n/10$ subjects having the largest estimated probabilities. With the second method, use of $g = 10$ groups results in cut points defined at the values $k/10, k = 1, 2, ...9$ and the group contain all subjects with estimated probabilities between adjacent cut points.

For eg: the first group contains all subjects whose estimated probability is less than or equal to 0.1, while the tenth group contains those subjects whose estimated probability is greater than 0.9. For the $Y = 1$ row, estimates of the expected values are obtained by summing the estimated probabilities over all subjects in a group. For the $Y = 0$ row, the estimated expected value is obtained by summing over all subjects in the group, one minus the estimated probability. For either grouping strategy, the Hosmer-Lemeshow goodness of fit statistic, $\hat{C}$ is obtained by calculating the Pearson chi-square statistic from the gx2 table of observed and estimated expected frequencies. A formula defining the calculation of C is as follows

$$\hat{C}{=}\textstyle\sum_{k=1}^{g}\frac{(O_k - \acute{n}_k\tilde{\pi}_k)^2}{\acute{n}_k\pi_k(1-\pi_k)}$$

where, $\acute{n}_k$ is the total number of subjects in the k^{th} group, C_k denotes the number of co-variate patterns in the k^{th} decile,

$$O_k = \textstyle\sum_{j=1}^{c_k} Y_j$$

is number of responses among C_k co-variate patterns, and

$$\bar{\pi}_k = \textstyle\sum_{j=1}^{c_k}\frac{m_j\hat{\pi}_j}{\acute{n}_k}$$

is the average estimated probability.

The distribution of the statistic $\hat{C}$ is well approximated by the chi-square distribution with $g - 2$ d.f.

The advantage of a summary goodness of fit statistic like $\hat{C}$ is that it provides a single, easily interpretable value that can be used to assess fit. The great disadvantage is that in the process of grouping we may miss an important deviation from fit due to small number of individual data points.

2.10 Classification Tables

Another way of evaluating the fit of a given logistic regression model is via a Classification Table. The Real Statistics Logistic Regression data analysis tool produces this table.This table is a result of cross-classifying the outcome variable Y, with a dichotomous variable whose values are derived from estimated logistic probabilities. The table shows a comparison of the number of successes ($Y = 1$) predicted by the logistic regression model compared to the number actually observed and similarly the number of failures ($Y = 0$) predicted by the logistic regression model compared to the number actually observed.

We have four possible outcomes:

True Positives (TP) = the number of cases which were correctly classified to be positive, i.e. were predicted to be a success and were actually observed to be a success.

False Positives (FP) = the number of cases which were incorrectly classified as positive, i.e. were predicted to be a success but were actually observed to be a failure.

True Negatives (TN) = the number of cases which were correctly classified to be negative, i.e. were predicted to be a failure and were actually observed to be a failure.

False Negatives (FN) = the number of cases which were incorrectly classified as negative, i.e. were predicted to be a negative but were actually observed to be a success.

2.11 Area under ROC curve

Sensitivity and specificity rely on a single cut point to classify a test result as positive. A more complete description of classification accuracy is given by area under the ROC (Receiver Operating Characteristic) curve. This curve originates from signal detection theory, shows how the receiver operates the existance of signal in the presence of noise. It plots the probability of detecting true signal(sensitivity) and false signal(1-specificity) for an entire range of possible cut points. The area under the ROC curve which ranges from zero to one, provides a measure of the model‘s ability to discriminate between those subjects who experience outcome of interest versus those who do not.

Chapter 3

ANALYSIS OF HEART DISEASE

3.1 Missing Data Exploration

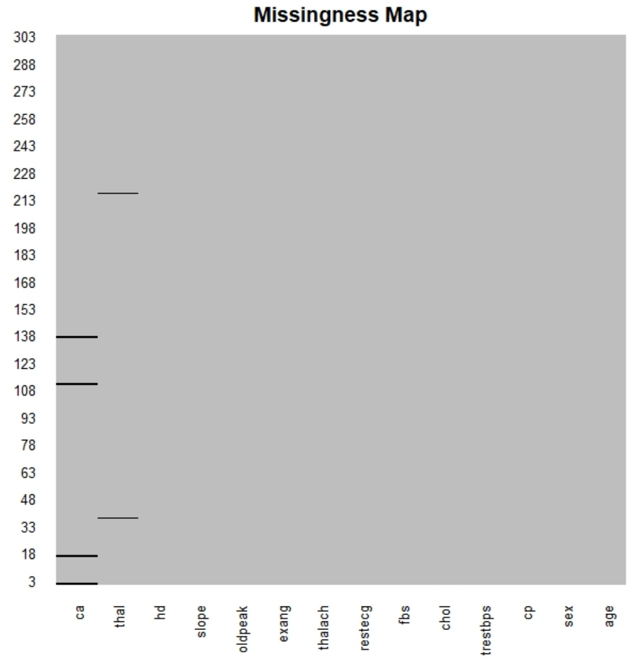


Figure 3.1: Missingness Map

The above Figure 3.1 represents the proportion of missing data in the dataset. As we can see from the Figure above, there are few missing values in the values of the variables 'ca'(number of major vessels coloured by fluoroscopy) and 'thal'(thal defect). The black bars denote the percentage of missing values.

3.2 Data Exploration

3.2.1 Multicollinearity

Corrrelations

	age	trestbps	chol	thalach	oldpeak
age	1.00000	0.29048	0.20264	-0.39456	0.19712
trestbps	0.29048	1.00000	0.13154	-0.04911	0.19124
chol	0.20264	0.13154	1.00000	-0.00007	0.03860
thalach	-0.39456	-0.04911	-0.00007	1.00000	-0.34764
oldpeak	0.19712	0.19124	0.03860	-0.34764	1.00000

Table 3.1: Correlation between variables

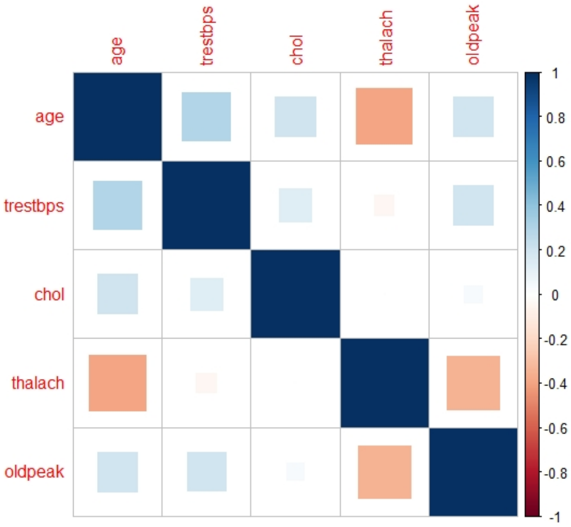


Figure 3.2: Correlation plot

One of the assumptions for a linear model is that the predictor variables should be independent. So, we check linear relationship between our predictor variables. For checking multicollinearity, in Figure 3.2 a square representation was used where blue represents positive correlation and red negative. The larger the square the larger the correlation. We can see that the matrix is symmetrical and that the diagonal are perfectly positively correlated because it shows the correlation of each variable with itself.

We need our independent variables to have as little collinearity as possible. From the correlation matrix, we can see that the maximum colinearity is between maximum heart rate achieved and age which is -0.39456. Since, it is greater than -0.5, we can conclude that the variables have a very little correlation.

3.3 Normality Testing

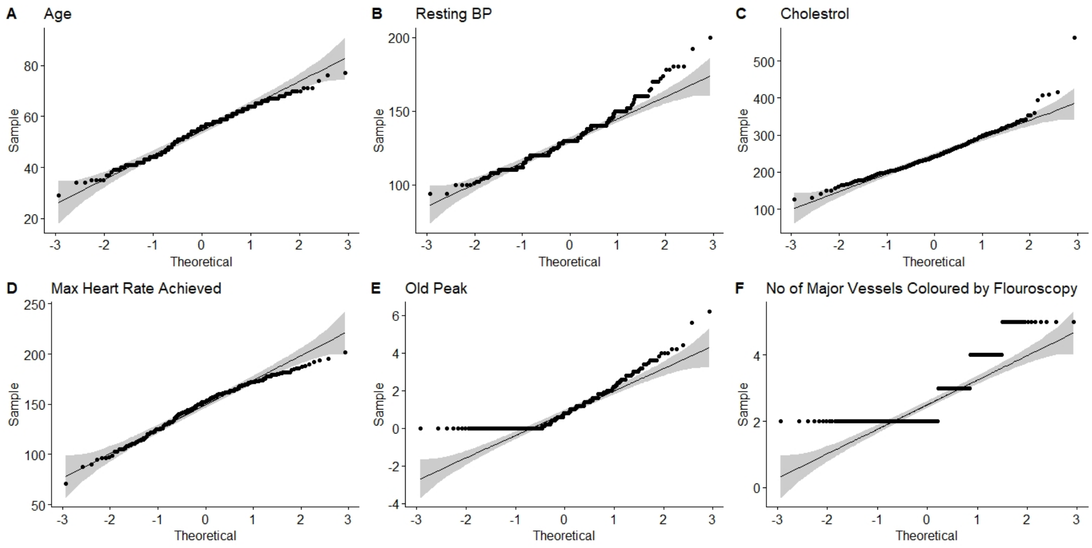


Figure 3.3: Q-Q plot for Numerical Variables

Shapiro-Wilk normality test	
Variables	**P-value**
Age	0.005424
Resting BP	$2.416e^{-06}$
Cholestrol	$1.019e^{-08}$
Maximum Heart Rate Achieved	$9.044e^{-}05$
Old Peak	$< 2.2e^{-16}$
No. of Major Vessels coloured by Flouroscopy	$< 2.2e^{-16}$

Table 3.2: Results of Shapiro-Wilk Normality test

From table 3.2, it is clear that none of the p-values is greater than 0.05 implying that the distribution of the data are significantly different from normal distribution. In other words, we can assume that there is no normality in any of the distribution. Figure 3.3 shows the same. However, Maximum Heart Rate Achieved shows a negative skewness and Old Peak is positively skewed.

3.4 Correlation Analysis

3.4.1 Age v/s Resting Blood Pressure

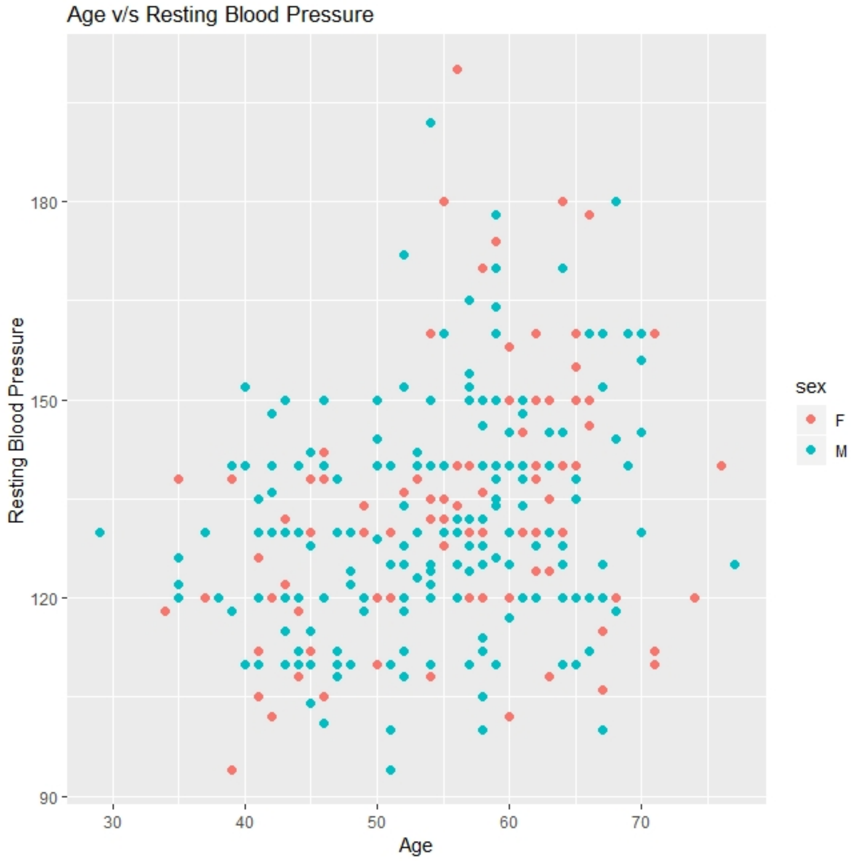


Figure 3.4: Age v/s Resting Blood Pressure

As we can see from the Figure 3.4, the resting blood pressure increases as the age increases for both males and females. This tells us that there is positive relationship between Resting blood pressure and Age.

3.4.2 Age v/s Cholestrol

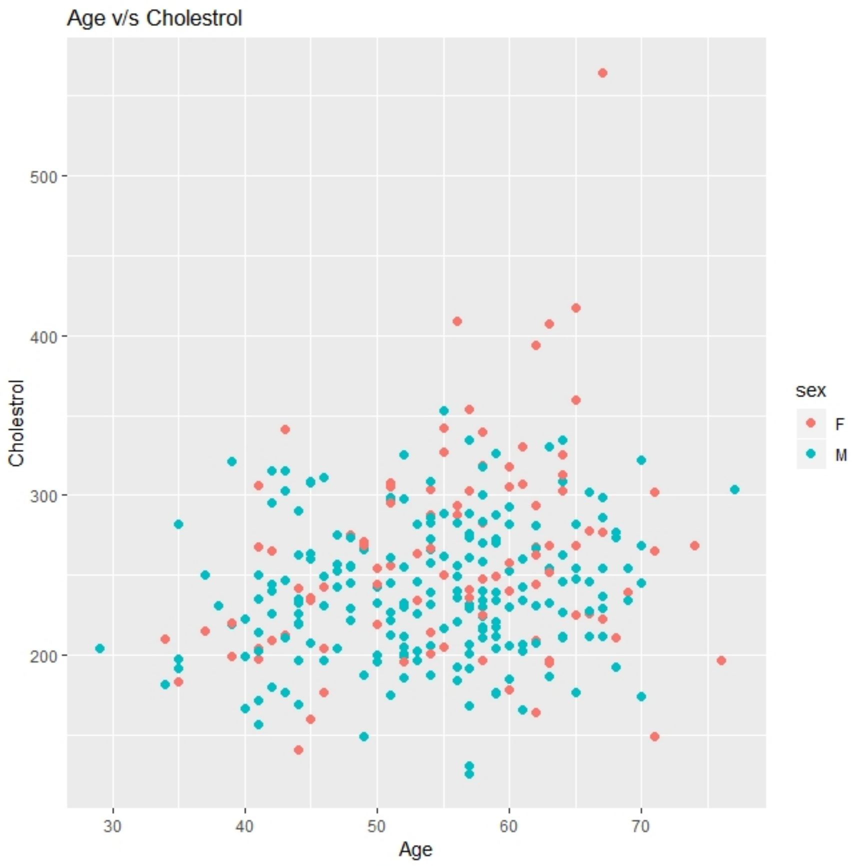


Figure 3.5: Age v/s Cholestrol

In Figure 3.5, the variable age and cholesterol shows very little positive relationship. As the age increases the cholesterol level increases a little bit. There are almost no outliers except one, which represents for a female with very high cholesterol level. We will remove any such outliers that might affect the results drastically.

3.5 Analysis on Individual Variables

3.5.1 Age

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
29.00	48.00	56.00	54.54	61.00	77.00

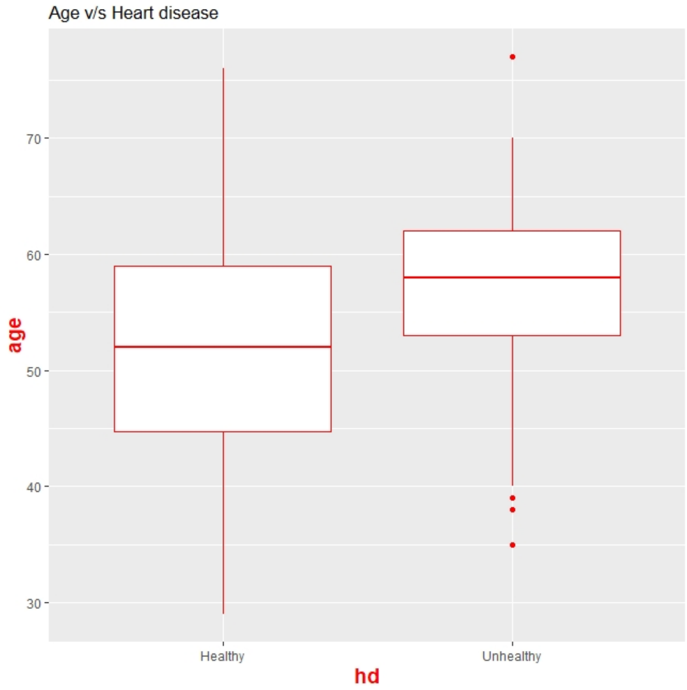


Figure 3.6: Age v/s Heart Disease

Patients from age 29 years to 77 years were included in this data set. Around 50% of patient's age was in between 45-65 years. The median of age for people with heart disease is 58 and without heart disease is 52. There is no visualized difference in ages for patients with or without heart disease. People having heart disease shows some outliers.

Applying Logistic Regression

	Estimate	**Std. Error**	**z value**	**Pr(>$\|z\|$)**
(Intercept)	-3.0512	0.76862	-3.97	7.2e-05 ***
age	0.05291	0.01382	3.829	0.000128 ***

Table 3.3: Age v/s Heart Disease

Here the p-value of age in Table 3.3 is 7.2e^{-05} which is less than 0.05. Therefore we reject the null hypothesis. Hence there is a significant impact between age and heart disease. The coefficient for intercept=-3.05122 and age=0.05291. Here the simple logistic regression model that relates the age to log odds of heart disease is

$$ln\frac{P}{1-P} = -3.05122 + 0.05291 * age\ of\ the\ person$$

Therefore the probability of having heart disease is

$$\frac{1}{1 + exp(-(0.05291 * age - 3.05122))}$$

For eg: For a patient having of age 58, entering the value age=58 in the equation give the estimated probability of having heart disease of 0.99:
Therefore the probability of having heart disease= $\frac{1}{1+exp(-(0.05291*58-3.05122))} = 0.99$

3.5.2 Sex

hd	Female	Male
Healthy	71	89
Unhealthy	25	112

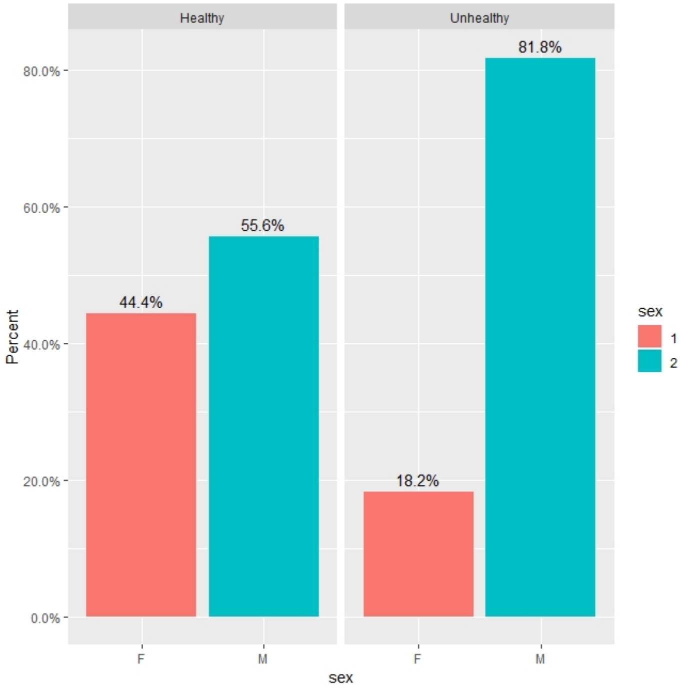


Figure 3.7: Sex v/s Heart Disease

From Figure 3.7, it is clear that level wise comparison for target feature shows men are more likely to have a heart disease than women. In the Figure, 1 represent female and 2 represent male.

Applying Logistic Regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.0438	0.2326	-4.488	7.18e-06 ***
sexM	1.2737	0.2725	4.674	2.95e-06 ***

Table 3.4: Sex v/s Heart Disease

Here from Table 3.4, the model for heart disease becomes

$$Heart\ disease = (-1.0438) + 1.2737 * patient\ is\ a\ male$$

where we assign the value 0 if the patient is female and the value 1 if the patient is male Therefore here we get an increase in log(odds) of males having heart disease and 1.2737 is the log(odds ratio) of males having heart disease over odds of females having heart disease. Therefore it is clear that males are having more heart disease compared to females in the Cleveland heart disease data.

3.5.3 Chest Pain Type

	Chest Pain Type			
hd	1	2	3	4
Healthy	16	40	65	39
Unhealthy	7	9	18	103

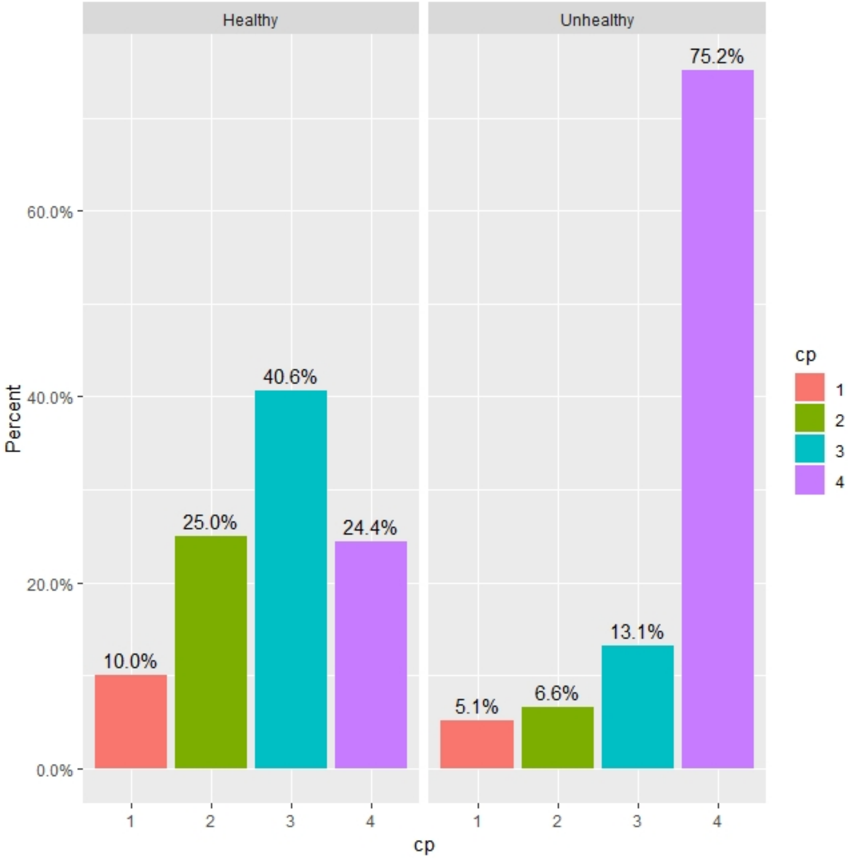


Figure 3.8: Chest Pain Type v/s Heart Disease

From the Figure 3.8, It is clear that the most Unhealthy patients have had a chest pain of type of 4(asymptomatic). However, the most of the healthy people have had a Type 3 pain (non-anginal pain). In the Figure 3.8, 1, 2, 3 & 4 represent typical angina, atypical angina, non-anginal pain & asymptomatic pain.

Applying Logistic Regression

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-0.8267	0.4532	-1.824	0.068116
cp2	-0.6650	0.5844	-1.138	0.255133
cp3	-0.4573	0.5256	-0.870	0.384269
cp4	1.7978	0.4906	3.664	0.000248 ***

Table 3.5: Chest pain type v/s Heart Disease

Here in Table 3.5, the p-value of cp4 is less than 0.05. Hence it has significant association with heart disease. Therefore chest pain can be taken as a necessary indicator of heart disease . Patients with chest pain of type 4 are having more risk of Heart disease.

3.5.4 Resting Blood Pressure

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
94.0	120.0	130.0	131.7	140.0	200.0

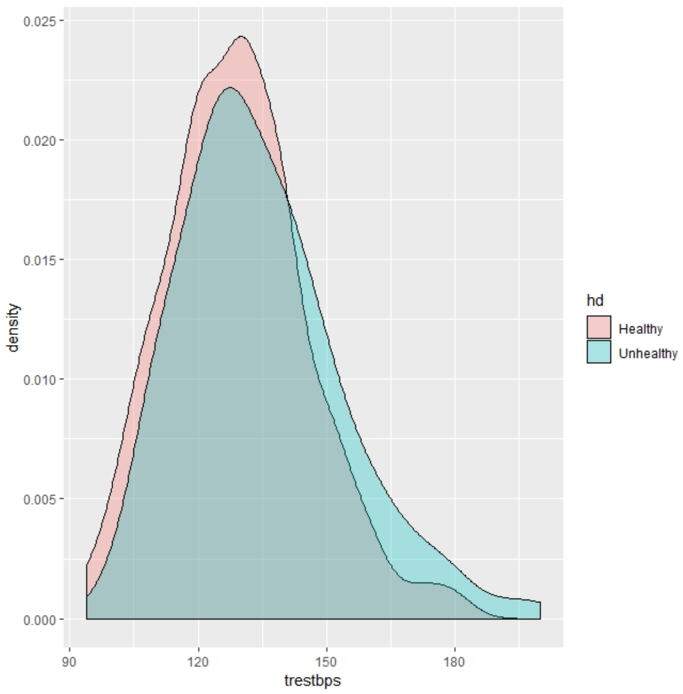


Figure 3.9: Resting Blood Pressure v/s Heart Disease

The aggregated resting blood pressure for the entire cohort from Figure 3.9, exhibited a median value of 130 which was similar to that for the diseased and non-diseased groups (i.e. 130 for both) .The distribution of resting blood pressure values was slightly larger for the diseased compared to the non-diseased group.In addition, there was a slight difference in skewness between the diseased and non-diseased cohorts. Therefore, resting blood pressure was not a good predictive feature.

Applying Logistic Regression

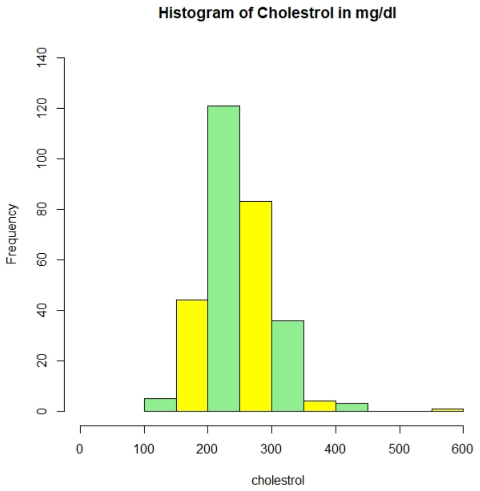
	Estimate	**Std. Error**	**z value**	**Pr(>$\|z\|$)**
(Intercept)	-2.494132	0.905112	-2.756	0.00586 **
trestbps	0.017745	0.006807	2.607	0.00914 **

Table 3.6: Resting Blood Pressure v/s Heart Disease

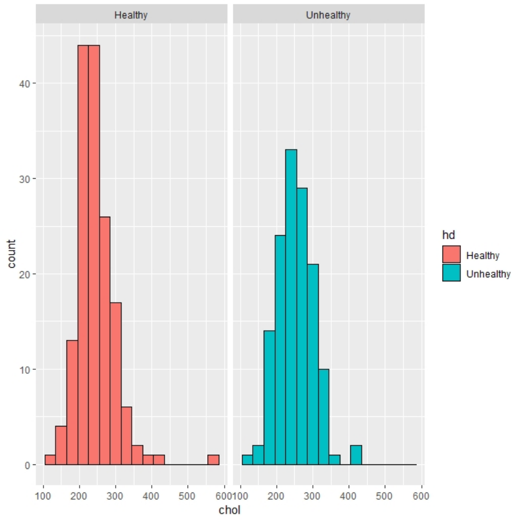
The p-value for Resting BP in Table 3.6 is greater than 0.05. Therefore the null hypothesis is accepted. Hence, Resting BP has no significant association with heart disease. The coefficient for intercept=2.494132 and trestbps=0.017745.

3.5.5 Cholestrol

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
126.0	211.0	243.0	247.4	276.0	564.0



(a) Cholestrol v/s Heart Disease for the entire cohort



(b) Cholestrol v/s Heart Disease for healthy and diseased

The distribution of patient‘s cholesterol level is highly right skewed(Fig (a)), showing that few patients have had extremely high cholesterol levels. When we compare this distribution separately for patients with a heart disease and patients without a heart disease (Fig (b)), the healthy patients‘s distribution is leptokurtic. That means, there were many healthy people who had there cholesterol level around 200-220 mg/dl than patients with a heart disease.

Applying Logistic Regression

	Estimate	Std. Error	z value	Pr(>$\|z\|$)
(Intercept)	-0.929987	0.576336	-1.614	0.107
chol	0.003130	0.002279	1.373	0.170

Table 3.7: Cholestrol v/s Heart Disease

The p-value of cholestrol in Table 3.7 is greater than 0.05. Therefore the null hypothesis is accepted which interprets that cholestrol has no significant association with heart disease in the Cleveland data. The coefficient for intercept=-0.929987 and cholestrol=0.003130.

3.5.6 Fasting Blood Sugar

hd	0	1
Healthy	137	23
Unhealthy	117	20

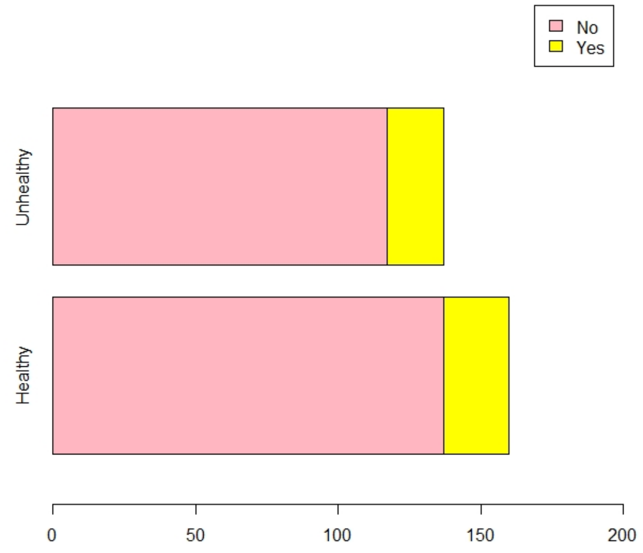


Figure 3.11: Fasting Blood Sugar v/s Heart Disease

From Figure 3.11, It is clear that most of the patients have had fasting blood sugar level less than 120mg/dl (normal). Level wise bar charts for target feature show the same pattern for both patients and suggest that the fasting blood sugar level may not be a deciding factor for having a heart disease or not.

Applying Logistic Regression

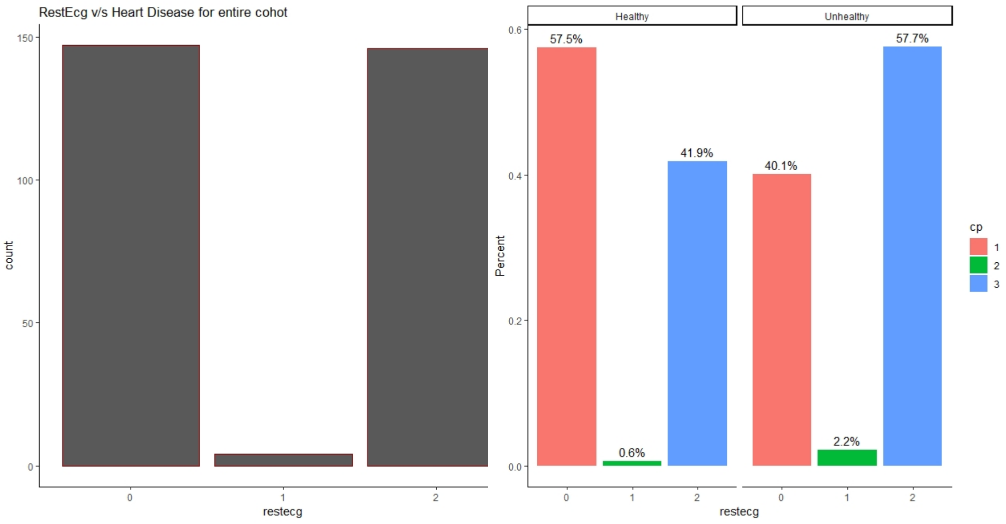
	Estimate	Std. Error	z value	Pr(>$|z|$)
(Intercept)	-0.15781	0.12588	-1.254	0.210
fbs1	0.01805	0.33064	0.055	0.956

Table 3.8: Fasting Blood Sugar v/s Heart Disease

The p-value of cholestrol in Table 3.8 is greater than 0.05, therefore the null hypothesis is accepted. Hence fasting blood sugar has no significant association with heart disease. The coefficient for intercept=-0.15781 and fasting blood sugar=0.01805.

3.5.7 Resting ECG results

	Resting ECG result		
hd	0	1	2
Healthy	92	1	67
Unhealthy	55	3	79



(a) RestECG v/s Heart Disease for the entire cohort

(b) RestECG v/s Heart Disease for healthy and diseased

0- Normal
1- Having ST–T wave abnormality
2- Showing definite left ventricular Hypertrophy

Most of the patients (Fig(a)) exhibited normal Resting Electrocardiograhic results. However, a higher proportion of diseased patients had abnormal ST wave patterns (Fig(b)) suggesting that this feature may contribute some predictive power.

Applying Logistic Regression

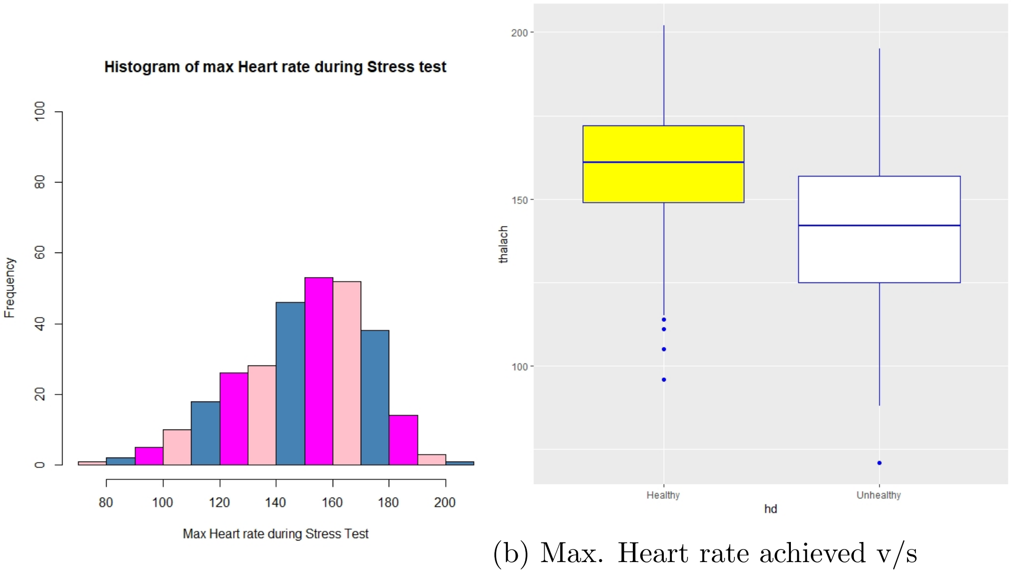
	Estimate	Std. Error	z value	Pr(>$\|z\|$)
(Intercept)	-0.5145	0.1704	-3.018	0.00254 **
restecg1	1.6131	1.1672	1.382	0.16698
restecg2	0.6792	0.2380	2.854	0.00432 **

Table 3.9: Resting ECG Results v/s Heart Disease

From Table 3.9, the p-value of restecg2 is less than 0.05. Hence it has significant association with heart disease. Therefore Resting ECG can be taken as an indicator of heart disease. Patients with restecg 2 are having more risk of heart disease.

3.5.8 Maximum Heart Rate Achieved During Stress Test

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
71.0	133.0	153.0	149.6	166.0	202.0



(a) Max. Heart rate achieved v/s Heart Disease for the entire cohort

(b) Max. Heart rate achieved v/s Heart Disease for healthy and diseased

The histogram (Fig(a)) for maximum heart rate achieved during stress test by patients is left skewed as few patients showed a comparatively low heart rate. The separate Boxplot

(Fig((b))) for two levels of target feature show healthy people have had a quite higher maximum heart rate (around 160) compared to the maximum heart rate (150) of patients with a heart disease.

Applying Logistic Regression

	Estimate	**Std. Error**	**z value**	**Pr(>$\|z\|$)**
(Intercept)	6.472142	1.002574	6.456	1.08e-10 ***
thalach	-0.044312	0.006627	-6.687	2.28e-11 ***

Table 3.10: Maximum Heart rate during stress test v/s Heart Disease

Here the p-value of thalach (maximum heart rate during stress test) in Table 3.10 is less than 0.05.Therefore we reject the null hypothesis. Hence there is a significant impact between Maximum Heart Rate achieved and heart disease. The coefficient for intercept=6.472142 and thalach=-0.044312. Here the simple logistic regression model that relates the age to log odds of heart disease is

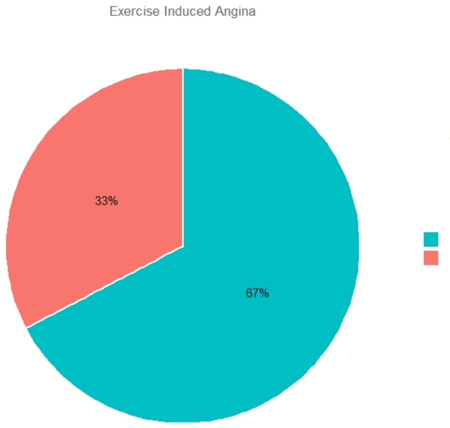
$$ln\frac{P}{1-P} = 6.472142 + -0.044312 * thalach$$

Therefore probability of having heart disease=$\frac{1}{1+exp(-(-0.044312*thalach+6.474142))}$

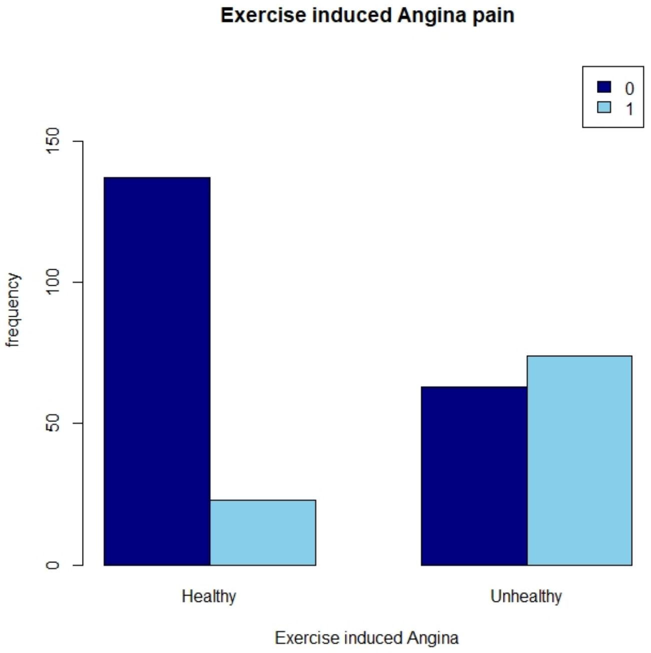
3.5.9 Exercise Induced Angina

	Exercise Induced Angina	
hd	0	1
Healthy	137	23
Unhealthy	63	74

Table 3.11: Exercise Induced Angina v/s Heart Disease



(a) Exercise Induced Angina v/s Heart Disease for the entire cohort



(b) Exercise Induced Angina v/s Heart Disease for healthy and diseased

Out of all patients around 67% (Fig(a)) haven't had an exercise induced angina. But having a exercise induced angina for a patient with a heart disease is more prominent than a patient without a heart disease (Fig(b)). This shows having a exercise induced angina may be a deciding factor for having a heart disease.

Applying Logistic Regression

	Estimate	**Std. Error**	**z value**	**Pr(>$\|z\|$)**
(Intercept)	-0.7768	0.1522	-5.103	3.34e-07 ***
exang1	1.9454	0.2831	6.871	6.37e-12 ***

Table 3.12: Exercise Induced Angina v/s Heart Disease

Here the model for heart disease in Table 3.12 becomes $heart\,disease = (-0.7768) + 1.9454 * exang1$, where we assign the value 0 if the patient is exang0(no pain) and the value 1 if the patient is exang1(there is a pain). Therefore here we get an increase in log(odds) of exang1 having heart disease and 1.9454 is the log(odds ratio) of exang1 having heart disease over odds of exang0 having heart disease. Therefore it is clear that exang1 are having more heart disease compared to exang0 in the Cleveland heart disease data.

3.5.10 Thal Defect

	thal Defect		
hd	**Normal**	**Fixed Defect**	**Reversible Defect**
Healthy	127	6	27
Unhealthy	37	12	88

Table 3.13: Thal Defect v/s Heart Disease

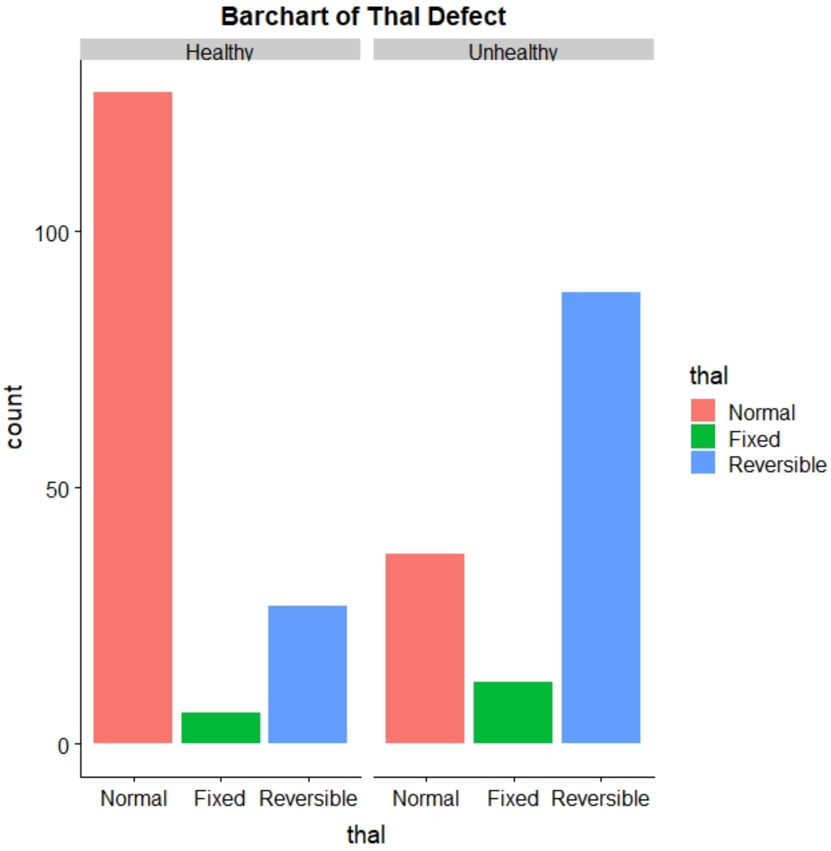


Figure 3.15: Thal Defect v/s Heart Disease

When considering the bar chart for all patients levels of heart status in Figure 3.15, commonly seen are Normal and Reversible Defect. However, two bar charts drawn for target feature are not the same for two levels of target. Healthy people mostly showed a normal heart status while sick people showed mostly a reversible defect condition.

Applying Logistic Regression

Here in Table 3.14, the p-value of thal6 and thal7 are less than 0.05. Hence these has significant association with heart disease. Therefore thal defect can be taken as an indicator of heart disease. Patients with thal6 and thal7 are having more risk of heart disease.

	Estimate	**Std. Error**	**z value**	**Pr(>$|z|$)**
(Intercept)	-1.2333	0.1868	-6.601	4.07E-11 ***
thal6	1.9264	0.5338	3.609	0.000307 ***
thal7	2.4148	0.2886	8.367	$<$ 2.00E-16 ***

Table 3.14: Thal Defect v/s Heart Disease

3.6 Application of Logistic Regression

3.6.1 Proportion of Heart Disease

Heart Disease	
Healthy	**Unhealthy**
160	137

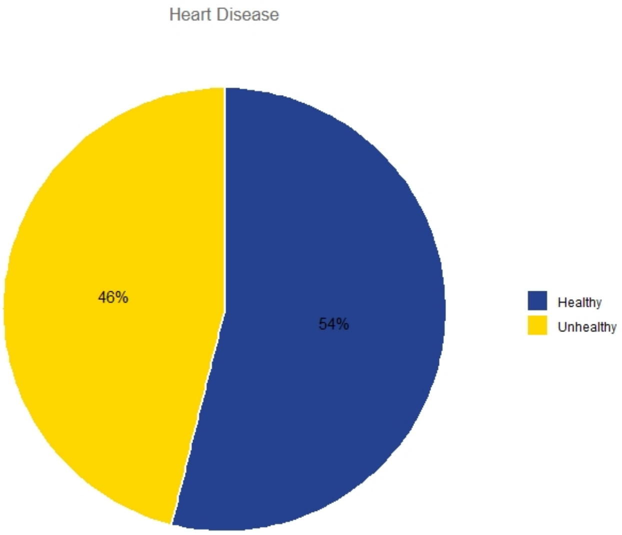


Figure 3.16: Pie Diagram of Heart Disease

Figure 3.16 shows that among all the total number of patients 46% have Heart Disease and 54% does not have heart Disease.

3.6.2 Logistic Curve for Heart Disease

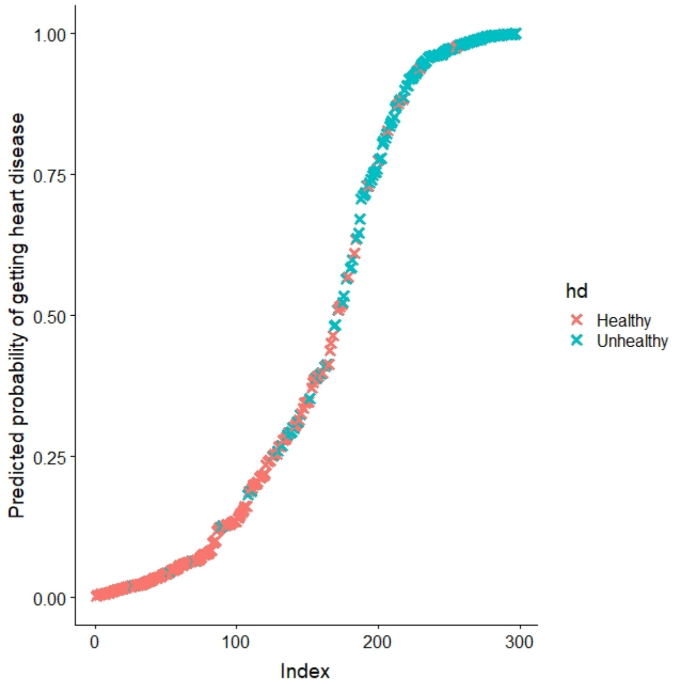


Figure 3.17: Logistic Curve for Heart Disease

The Logistic curve in Figure 3.17 shows predicted probabilities that each patient has heart disease along with their actual heart disease status. Most of the patients with heart disease are predicted to have high probability of having heart disease(in turquoise colour) And Most of the patients without heart disease(salmon colour) are predicted to have low probability of having heart disease.

3.6.3 Fitting of Logistic Regression Model

Now, we perform logistic regression on the model including all the 14 variables of Heart Disease. Then we find the most significant variables and exclude them from the model and again conduct logistic regression on the reduced model. We also calcuate goodness of fit for the model using Hosmer and Lemeshow test.

	Estimate	**Std. Error**	**z value**	**Pr(>$\|z\|$)**
(Intercept)	-8.652487	3.006158	-2.878	0.00400 **
age	-0.013763	0.024745	-0.556	0.57808
sexM	1.546014	0.529995	2.917	0.00353 **
cp2	1.239566	0.770874	1.608	0.10784
cp3	0.245959	0.663312	0.371	0.71078
cp4	2.086480	0.666547	3.130	0.00175 **
trestbps	0.024364	0.011269	2.162	0.03062 *
chol	0.004448	0.003993	1.114	0.26526
fbs1	-0.596246	0.607848	-0.981	0.32664
restecg1	0.810202	2.435102	0.333	0.73935
restecg2	0.473895	0.383518	1.236	0.21659
thalach	-0.017723	0.011109	-1.595	0.11065
exang1	0.709456	0.440018	1.612	0.10689
oldpeak	0.357875	0.230070	1.556	0.11983
slope2	1.155286	0.473794	2.438	0.01475 *
slope3	0.525147	0.919661	0.571	0.56798
ca	1.311510	0.279276	4.696	2.65e-06 ***
thal6	-0.010974	0.790210	-0.014	0.98892
thal7	1.392715	0.425194	3.275	0.00105 **

Table 3.15: Logistic Regression conducted on original model

From the p-values for the regression coefficients, we can see that sex, cp, trestbps, slope, ca and thal (p-value<0.05) are significant. On the other hand age, trestbps, chol, fbs, restecg, thalach and exang (p-value>0.05) are not significant. Now we will remove the variables that are not significant and form a new model.

Now, Lets fit a second equation without the insignificant variables and test whether this reduced model fits the data as well:

	Estimate	**Std. Error**	**z value**	**Pr(>$\|z\|$)**
(Intercept)	-10.83356	1.97214	-5.493	3.94e^{-}08 ***
sexM	1.4239	0.47448	3.001	0.002691 **
cp2	1.00877	0.75051	1.344	0.178911
cp3	0.20579	0.6602	0.312	0.755265
cp4	2.47223	0.64521	3.832	0.000127 ***
trestbps	0.0246	0.01013	2.427	0.015205 *
slope2	1.8318	0.41358	4.429	9.46E-06 ***
slope3	1.3645	0.7007	1.947	0.051494
. **ca**	1.32207	0.24861	5.318	1.05E-07 ***
thal6	0.06788	0.71713	0.095	0.924585
thal7	1.54348	0.4011	3.848	0.000119 ***

Table 3.16: Logistic Regression conducted on reduced model

From the p-values for the regression coefficients from the reduced model, we can see that all the regression coefficients make significant contribution to the equation.

3.6.4 Testing Goodness of Fit using Hosmer And Lemeshow Test

Goodness of Fit for the final model

Hosmer and Lemeshow test (binary model)
data: data$hd, fitted(m2)
X-squared = 7.3681, df = 8, p-value = 0.4975

Table 3.17: Hosmer and Lemeshow Test for final model

Hosmer and Lemeshow test produces a chi-square of 7.3681 with 8 df, yielding a p-value of 0.4975 which is insignificant. This suggests that the model is a satisfactory fit to the data, and that interactions and non-linearities are not needed.

3.6.5 Stepwise Regression

Backward

Start: AIC=-613.17
hd $\sim$ age + sex + cp + trestbps + chol + fbs + restecg + thalach + exang + oldpeak + slope + ca + thal

	Df	Sum of Sq	RSS	AIC
- age	1	0.0197	34.311	-615.00
- chol	1	0.0817	34.373	-614.47
<none>			34.292	-613.17
- oldpeak	1	0.2388	34.530	-613.11
- restecg	1	0.2844	34.576	-612.72
- slope	1	0.2994	34.591	-612.59
- fbs	1	0.3251	34.617	-612.37
- trestbps	1	0.3621	34.654	-612.05
- thalach	1	0.6865	34.978	-609.29
- exang	1	0.8670	35.159	-607.76
- sex	1	1.1235	35.415	-605.60
- cp	1	1.3808	35.672	-603.45
- thal	1	2.7913	37.083	-591.93
- ca	1	3.9390	38.231	-582.88

Step: AIC=-615 hd $\sim$ sex + cp + trestbps + chol + fbs + restecg + thalach + exang + oldpeak + slope + ca + thal

	Df	Sum of Sq	RSS	AIC
- chol	1	0.0720	34.383	-616.38
<none>		34.311	-615.00	
- oldpeak	1	0.2459	34.557	-614.88
- restecg	1	0.2766	34.588	-614.62
- slope	1	0.2986	34.610	-614.43
- fbs	1	0.3352	34.646	-614.12
- trestbps	1	0.3424	34.654	-614.05
- thalach	1	0.6934	35.005	-611.06
- exang	1	0.8919	35.203	-609.38
- sex	1	1.1637	35.475	-607.10
- cp	1	1.3978	35.709	-605.14
- thal	1	2.7870	37.098	-593.81
- ca	1	4.0814	38.393	-583.62

Step: AIC=-616.38
hd$\sim$sex + cp + trestbps + fbs + restecg + thalach + exang + oldpeak + slope + ca + thal

	Df	Sum of Sq	RSS	AIC
<none>			34.383	-616.38
- oldpeak	1	0.2416	34.625	-616.30
- slope	1	0.2870	34.670	-615.91
- restecg	1	0.3302	34.714	-615.54
- fbs	1	0.3413	34.725	-615.45
- trestbps	1	0.3721	34.755	-615.18
- thalach	1	0.6729	35.056	-612.62
- exang	1	0.9242	35.307	-610.50
- sex	1	1.0925	35.476	-609.09
- cp	1	1.4152	35.798	-606.40
- thal	1	2.8637	37.247	-594.62
- ca	1	4.2277	38.611	-583.94

In backward stepwise method we use step function to eliminate the variables from model that have higher p-values and are insignificant one by one. Finally we will get the model that has lowest AIC value. The lower the AIC value the better the model fit. Here the variable age got removed first and then chol. The final AIC value obtained is -616.38 which interprets that stepwise regression also gives a better fit of the model.

3.6.6 ROC Curve for the Logistic Regression Model

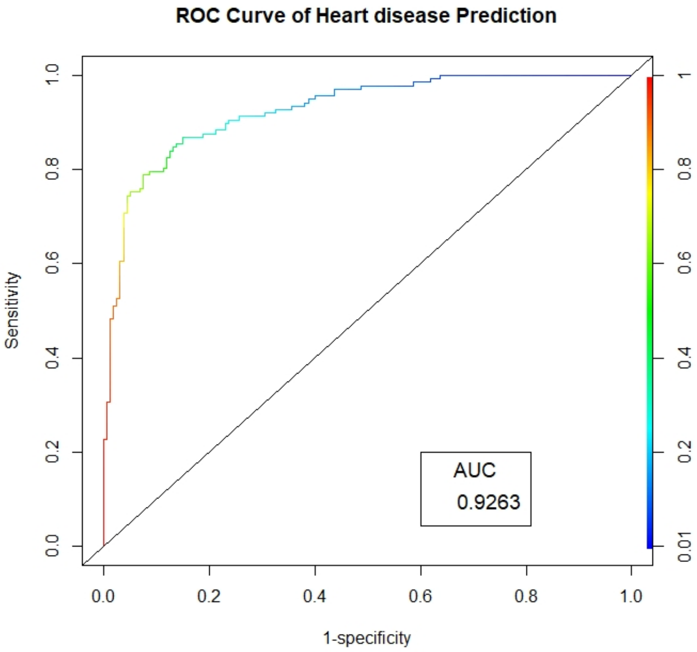


Figure 3.18: ROC Curve for Logistic Regression Model

The Receiver Operating characteristic (ROC) curve is the graphical representation of all the possible outcome between 0 and 1. It is the plot of sensitivity vs specificity. It gives us range of cuttoff value or the threshold value that we can use to predict the accuracy of model. Here we look at optimal threshold value to predict the accuracy of model. Here from the ROC curve in Figure 3.18, the accuracy of the above logistic regression model is found to be 0.9263 which also suggest that logistic regression is a better model for prediction.

Bibliography

1. Belsley, David. A., Edwin. Kuh, and Roy. E. Welsch. 1980. Regression Diagnostics: Identifying Influential Data and Sources ofCollinearity, New York: John Wiley and Sons.

2. Draper, N. & Smith, H. (1998). Applied regression analysis. 3rd edn. Wiley-Blackwell. 736 pp.

3. Dr. Robert Detrano (1988), UCI Repository Website: Heart Disease Data. Available at : https://archive.ics.uci.edu/ml/datasets/heart+Disease

4. Gujarati, N. Damodar 2003. Basic Econometrics' fourth edition (international), The McGraw-Hill Companies, Inc

5. Hosmer, D.W. and Lemeshow, S. (2000), Applied Logistic Regression, 2nd. ed., John Wiley and Sons, New York.

6. John Verzani. (2014). Using R for Introductory Statistics (Chapman & Hall/CRC The R Series) 1st Edition

7. Joseph M. Hilbe, Practical Guide To Logistic Regression,2015 by Taylor & Francis Group, LLC.

8. K. Dietz, M. Gail, K. Krickeberg, A. Tsiatis, J. Samet:Statistics for Biology and Health, 2^{nd} edition, SpringerVerlg, Newyork Inc.

9. Kleinbaum, D. G.(1994):Logistic Regression. A Self Learning Text. SpringerVerlg, Newyork Inc.