Kwame Nkrumah University of Science and Technology

DETERMINING RISK FACTORS OF CARDIOVASCULAR DISEASES USING MULTIPLE LOGISTIC REGRESSION ANALYSIS

CASE STUDY: KOMFO ANOKYE TEACHING HOSPITAL(2011-2012)

By

Emmanuel Kwasi Opoku

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF M.SC INDUSTRIAL MATHEMATICS

October 14, 2015

Declaration

I hereby declare that this submission is my own work towards the award of the M.SC degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgement had been made in the text.

Emmanuel Kwasi Opoku (PG 8149212) ....

Student Signature Date

Certified by:

Mr. E. Harris ....

Supervisor Signature Date

Certified by:

Prof. S.k Amponsah ....

Head of Department Signature Date

Dedication

This thesis is dedicated to my wife, Mrs. Christiana Opoku and my lovely daughters, Nana Yaa Opoku Agyekumwaa and Maame Yaa Opoku Boatemaa.

It is also dedicated to my parents, Mr. Maxwell Opoku and Mad. Cecilia Antwi.

Abstract

Cardiovascular disease is one of the top two cause of death after diarrhoea disease. WHO(2011). The study analysed the risk factors of cardiovascular diseases using descriptive statistics and multiple logistic regression model. Multiple logistic regression was applied to assess the risk factors for the development of cardiovascular diseases in Ghana. Secondary data was obtained from the department of medicine of the komfo Anokye Teaching Hospital, in 2011 and 2012. The folders of 109 suspected cardiovascular patients made up of 78 males (71.6 %) and 31 females (28.4%) were available for the final analysis. The results showed that, risk factors such as family history, overweight, cholesterol level and hypertension were important predictors of cardiovascular diseases. However, it was found that the risk factors; gender, diabetes, smoking, alcohol consumption and age were not good predictors of cardiovascular diseases. Finally, there was sufficient evidence to show that hypertension poses the greatest risk to the development of cardiovascular diseases. More so, hypertensive female patients were more likely to develop cardiovascular diseases than hypertensive male patients.

Acknowledgements

I would like to express my profound gratitude to the Almighty God for His love, care and guidance and also seeing me through this work successfully.

I will again extend my deepest gratitude to my lecturer and supervisor, Mr. Emmanuel Harris for his direction, coordination and critics from the first to the last page of this work.

I wish to also show my sincere gratitude to my head of department, Prof. S.K Amponsah and my lecturers for their time and patience during the taught causes.

Again, I will say thank you to my parents for their commitment to my entire education.

My last appreciation goes to my wife for her understanding and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of study

1.1.1 Definition

Cardiovascular diseases are a group of disorders of the heart and blood vessels and include:

- coronary heart disease: disease of the blood vessels supplying the heart muscle;
- cerebrovascular disease: disease of the blood vessels supplying the brains;
- peripheral arterial disease: disease of blood vessels supplying the arms and legs;
- rheumatic heart disease: damages to the heart and heart valves from rheumatic fever, caused by streptococcal bacteria;
- congenital heart disease: malformation of heart structure existing at birth;
- deep vein thrombosis: blood clots in the leg veins, which can dislodge and move to the heart and lungs.

Coronary artery disease develops when the coronary arteries-the major blood vessel that supply the heart with blood, oxygen and nutrients become damaged or diseased.

Cerebrovascular disease ia a group of brain dysfunctions related to disease of the blood vessels supplying the brain. Cerebrovascular disease primarily affects people who are elderly or have a history of diabetes,smoking or ischemic heart

disease.

Peripheral artery disease is a common circulatory problem in which narrowed arteries reduce blood flow to the limbs. When peripheral is developed, the legs does not receive enough blood flow to keep up with the demand.

Rheumatic heart disease is a disease caused by acute rheumatic fever. Rheumatic fever is an illness that predominantly affects children with the highest rate occurring in the 5-14 year age group.

Congenital heart disease is the abnormalities in the heart and valves present at birth.The most common causes of congenital birth defect are genetic and environmental. One percent of children born with congenital heart problems have genetics as the main cause.A faulty gene or chromosomes are the main reasons for heart conditions.

Deep vein thrombosis refers to a condition of the blood vessels(veins) that is characterised by formation of blood clots. The condition mostly affects the deep veins of the leg or the pelvis.

The epidemic of cardiovascular diseases(CVDs), also known as Ischemic heart disease, a disease characterised by reduced blood supply of the heart muscle, is on the increase in developing countries across Africa.

That, according to medical experts was as a result of trend of cheaper lifestyle amongst the populace. An estimated 17.3 million people died of CVDs in 2008 and over 80 percent CVDs take place in low-and middle-income countries.WHO (2013c)

1.1.2 Signs and symptoms

The most common symptom of cardiovascular disease is angina or chest pain. Angina can be described as a discomfort, heaviness, pressure, aching, burning, fullness, squeezing or painful feeling in the chest.

Other symptoms of cardiovascular diseases include

- Shortness of breadth

- Weakness or dizziness
- Nausea
- Sweating
- Diminished hair and nail growth and affected limb.

1.1.3 Causes

Coronary heart disease begins with damage to lining and inner layers of the heart arteries. Several factors contribute to this damage . They include:

- Smoking, including second hand smoke
- High amounts of certain fats and cholesterol in the blood
- High blood pressure
- High amounts of sugar in the blood due to insulin resistance or diabetes
- Blood vessel inflammation

Cardiovascular diseases are diagnosed based on the patient's medical and family histories, risk factors, a physical exam and the results from tests. The following tests are used for clinical diagnosis of cardiovascular diseases.

- Electrocardiogram(EKG) detects and records the heart's electrical activity. The test shows how fast the heart is beating and it's rhythm.
- Echocardiography(Echo) uses sound waves to create a moving picture of the heart. The test provide information about the size and shape of the heart and how well the heart chambers and valves are working.
- Chest X Ray creates pictures of the organs and structures inside the chest, such as the heart,lungs and blood vessels. A chest x ray can reveal signs of heart failure.

- Blood tests check the levels of fats, sugar, cholesterol in the blood. Abnormal levels may be signs of cardiovascular disease.

1.2 Statement of problem

In Ghana, cardiovascular disease was the leading cause of deaths in 1991 and 2001. WHO(2011).

Available research in this area in Ghana employed only descriptive analysis in identifying the factors that affect cardiovascular diseases. The study therefore uses appropriate statistical technique(logistic regression) in identifying variables associated with cardiovascular diseases.

1.3 Objectives of the study

The general objectives of the study are:

1. To determine the variables that affect cardiovascular diseases.
2. To fit a multiple logistic regression model to assess the risk factors fordevelopment of cardiovascular diseases.

1.4 Methodology

A secondary data from the department of medicine of the Komfo Anokye Teaching Hospital was used for the study. Two years data from 2011 to 2012 was sampled out for the research.

The following informations was taken from the folders of cardiovascular disease patients: Gender, overweight, cholesterol levels, smoking, age, alcohol consumption, hypertensive , family history and diabetic.

The software package used for the analysis was R.

1.5 Significance of the study

This study proves to be important because;

1. The results or outcome would give information as to whether factors likegender, alcohol consumption, hypertension, overweight, smoking, diabetes, cholesterol level and family history are really the risk factors in being diagnosed of cardiovascular disease.
2. The findings will help individuals to know which of the factors stated abovecontribute more in the development of the disease. This will help us find a way of solving the problem. E.g. if high blood pressure or high cholesterol level contribute more to the risk of having the disease, then there will be the need to always take in food and do things that will prevent our blood pressure from rising up.
3. The findings will help people without the disease to know his/her estimatedrisk of having it. This will help non-patients to do everything possible to prevent them from getting the disease.

1.6 Limitations

The research was limited due to the following challenges;

1. Inadequate finance which would prevent us from visiting other hospitalsoutside Kumasi.
2. More people to help in the collection of the data and carrying out of the research

1.7 Organisation of study

The research was categorised under five chapters as follows:

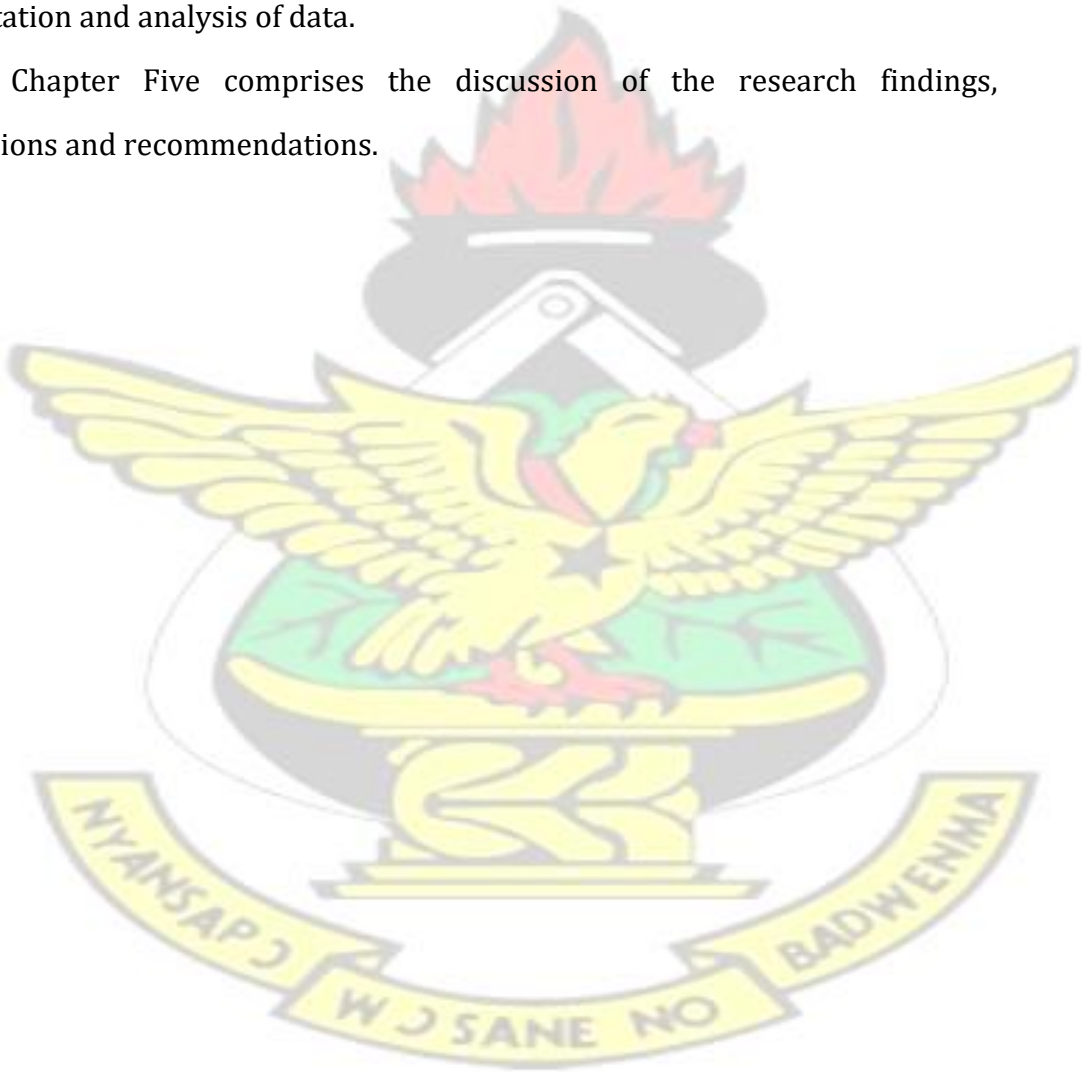
Chapter One dealt with the background of the study, statement of the problem, purpose of the study/objectives, methodology and significance of the study. The other segment comprises of limitations.

Chapter Two examined the related literature review, the place of the research and other research findings.

Chapter Three describes the mathematical methods used for the research, type of data, source of data and data analysis procedures.

Chapter Four provided the results of the study. This includes the presentation and analysis of data.

Chapter Five comprises the discussion of the research findings, conclusions and recommendations.



Chapter 2

Literature Review

2.1 Introduction

This chapter reviews the work of other researchers in relation to the study.

Cardiovascular disease refers to any disease that affects the cardiovascular system, principally cardiac disease, vascular diseases of the brain and kidney, and peripheral artery disease. Cardiovascular disease is the leading cause of deaths worldwide, though, since the 1970's, cardiovascular mortality rates have declined in many high-income countries. At the same time, cardiovascular deaths and disease have increased at fast rate in low-and middle income-countries and occurs almost equally in men and women .

2.2 Literature Review on Variables Associated with Cardiovascular Diseases

According to Assareh et al. (2013), in a study, found the prevalence of different risk factors using descriptive statistics. The prevalence of coronary artery disease risk factors were: hypertension(45.3%) ; high cholesterol (34.5%); diabetic mellitus(27.6%) ; family history(20.7%) and smokers(19.9%) .

Dominguez-Rodriguez et al. (2013), also found in a study at Tenerife,Spain using multiple logistic regression analysis that, smoking was the only variable associated independently with Acute Coronary Syndrome in young women.
The findings of this study indicated that, smoking is an independent predictor of Acute Coronary Syndrome in women less than 40 years.

Ira and Nancy (1997), used descriptive statistics in their study and found that, as many as 30% of all coronary heart disease deaths in the United States each

year are attributed to cigarette smoking, with high risk being strongly doserelated. Smoking nearly doubles the risk of Ischemic stroke

Vithanage et al. (2013), in their study, used a case control design method to identify the important risk factors of myocardial infarction(MI) prevailing in the Kandy district of Sri Lanka. In the analysis, hypertension, type II diabetes, smoking, and high cholesterol levels were identified as the independent risk factors of myocardial infarction. However, the anthropometric measurements, waist hip ratio and body mass index did not show an association with myocardial in-

farctions.

Longjian and Howard (2012), in their study used descriptive statistics to identify the risk factors of Heart Failure in the United States. The results are Diabetes (3.1%), overweight(8.0%), cigarette smoking(17.1%), coronary heart disease (60%), hypertension(10.1 %), male sex (8.9%), low physical activity (9.2%) , less than high school education(8.9%) and valvular heart disease(2.2%).

Grundtvig et al. (2009) also found that, smoking increases the risk of a first acute myocardial infarction (AMI) relatively more in females than in men. At younger ages(less than 50 years), smoking is deleterious in women than in men, with a larger negative impact of the total number of cigarette smoked per day.

Olutobi et al. (2001), also used descriptive statistics to examine pattern of cardiovascular disease mortality autopsy in Accra. They found out that, proportionate mortality ratio (PMR) for Cardiovascular disease increased with age, rising steeply in the mid-life to peak in the very old, accounting for almost 50% of deaths examined by age 85 years. They however indicated that the age pattern of mortality shown in their study does not follow the normal U-shaped pattern of mortality pattern in a population.

Abbey et al. (1999) again found using descriptive statistics that, at younger age, the relative risk of hypercholesterolemia is lower in women compared with men. Above 65 years of age mean low-density lipoprotein(LDL)

levels rise by 10% and 14 % respectively whereas high-level lipoprotein)(HDL) remained unchanged. Above 65 years of age mean IDL cholesterol is higher in women compared with men. At all ages HDL cholesterol levels are 0.26 to 0.36mmol/l higher in women but a low level HDL cholesterol implicates a higher coronary heart disease risk in women than in men.

Also, according to Longo-Mbenza et al. (2014), in a study to assess the associations of high density lipoprotein (HDL-C),cardiovascular disease with diabetic retinopathy in the Kinshasha Region used descriptive statistics and logistic regression model. There was a significant u-shaped relationship between diabetic retinopathy and high density lipoprotein(HDL-C). Also, smoking status, diabetic retinopathy were significant determinants for cardiovascular diseases.

Shaikh et al. (2014), in a study also identified the risk factors for anthracycline induced cardiac dysfunction in pediatric patients. Multiple logistic regression model was applied to assess the risk factors for development of cardiac dysfunction. 110 pediatric oncology patients were available for final analysis.15 (14%) children had cardiac dysfunction within a month; out of them 10/15 (67%) had isolated diastolic dysfunction, while 28 (25%) developed dysfunction within a year. 19 (17%) had pericardial effusion. 11 expired and out of them, 7 had significant cardiac dysfunction. Cumulative dose , radiation therapy and sepsis were found to be independent risk factors associated anthracycline induced cardiac dysfunction.

Also, according to the research of Tanmay and Arnab (2013) in Asian India,the cause of cardiovascular is multi factorial and no single factor is an absolute cause. However, hypertension and diabetes are highly prevalent among the Asian population which are also risk factors of cardiovascular diseases.

Eloamany et al. (2011) in a study found again that diabetes mellitus as a major risk factor for the development of coronary artery disease and that patients with diabetes have increased cardiovascular morbidity and mortality.

A research in a gender-related differences in the management of hypertension by cardiologist in Finland, found using descriptive statistics that,it

was significantly more common for female patients than male patients to have firstdegree relatives with coronary artery disease before the age of 65 (76% vs 62%, P= 0.0026). For the sisters of the female patients the cumulative risk of coronary heart disease by the age of 65 years was almost twice that of the sisters of the male patients (25.9% vs 15.8%, P=0.0123). The risk for the brothers of the females did not significantly differ from that of the brothers of the male patients, but it was 3.5 times that of the brothers of the controls. Thus, while a history of coronary heart disease in first-degree relatives is a risk factor for the disease, the risk is greater in women than in men. Pohjola-Sintonen et al. (1998).

Simon and Rosolova (2002) indicated in a study using descriptive statistics that, innate susceptibility to coronary heart disease showed that family history of premature coronary heart disease conferred excess risk. A history of death due to coronary heart disease in parents of the cohort was found to be associated with a 30% increased risk of coronary heart disease, a risk which was not mediated by other risk factors.

Maas and Appelman (2010) in a study found that ,cardiovascular disease develops 7 to 10 years later in women than in men and it is still the major cause of death in women.The risk of heart disease in women is often underestimated due to the misconception that females are 'protected' from cardiovascular diseases.

According to Yun-Mi et al. (1998), the risk of death from coronary heart disease increased significantly in men with the highest cholesterol level. In their study, they found that, increase in cholesterol level has become the burden of risk factors of non-communicable disease, blood pressure, and now a major public health problem for all age groups. They said blood pressure is frequently elevated in children with high cholesterol level or increase in fat as compared to lean subjects. This they said is possible related to their sedentary lifestyle, altered eating habits, increased fat content of diets and decreased physical activity.

Djousse et al. (2007), used descriptive statistics to categorise alcohol consumption into predetermined moderate-and high-consumption groups and

used current abstainers or low consumers as the reference group. In their analysis, moderate consumption was associated with a 30% reduced risk of diabetes among men.

Towfighi et al. (2009) in a surveys found that, over the two past decades,the prevalence of myocardial infarctions has increased in mid life (35 to 54 years) women, while declining in similarly aged men.

Sesso et al. (2001) in an epidemiological studies indicated that, family or parental history of myocardial infarction is a risk factor for coronary heart disease.

WHO (2013a) in the Global health estimates(2013),also indicated that 65% of all cardiovascular disease patients are 70 years and above, 50-69 years(26%), 30-49 years(5%), 15-29 years(1.2%), 5-14 years(0.2%), 1-59 months(0.3%) and 027 days(0.06%).

A study from the World Health Organisation(WHO) on Tobacco Free Iniative(T.F.I) indicated that 57.7% of persons smoking 30 cigarettes per day had died as compared to only 36.3% of non-smokers.The world Health report on Smoking and Tobacco also indicated that smoking is estimated to cause over 22% of cardiovascular diseases.

WHO (2014) in their global report on alcohol indicated that, alcohol killed 3.3 million people in 2012. The report also points to the fact that a higher percentage of deaths among men(7.6%) than among women(4%) are from alcoholrelated causes.

WHO (2013b) found that, high blood pressure can lead to hypertension which is a major risk factor for overall mortality on a global scale. By changing the structure of arteries, high (also known as raised or elevated) blood pressure increases the risk of stroke, heart disease, and kidney failure, as well as other diseases.

WHO (2013d) in their annual report said that,about 347 million people have diabetes. In 2004, an estimated 3.4 million people died from consequences

of high fasting blood sugar with more than 80% of deaths occurring in low-and middle-income countries.

Chapter 3

Methodology

3.1 Introduction

The study would have considered all the main hospitals in the country but Komfo Anokye Teaching Hospital was sampled out for the research. The data used for the study was a secondary data.This was taken from the department of medicine of the Komfo Anokye Teaching Hospital. Two years data was sampled out, that is, 2010 and 2012. The researcher visited the department of medicine of the Komfo Anokye Teaching Hospital in Kumasi to take the following information from the folders of Cardiovascular disease patients. gender, overweight, cholesterol level , smoking, age, alcohol consumption, hypertensive, family history and diabetic. The software package used for the analysis was R.

3.2 Categorical variable

These are variables that places individuals literally into categories and cannot be quantified in a meaningful way. Examples are diabetes, occupation, cholesterol level, gender etc. Analysis of categorical data involves the use of data tables. A two-way table presents categorical data by counting the number of observations that fall into each group for two variables one divided into rows and the other into columns.

3.3 Chi-square Test of Independence

	J				
	1	2	3	...	*j*

	1	O_{11}	O_{12}	O_{13}	...	O_{1j}	O_{1+}
I	2	O_{21}	O_{22}	O_{23}	...	O_{2j}	O_{2+}
	...		...	...	...	...	
	i	O_{i1}	O_{i2}	O_{i3}	...	O_{ij}	O_{i+}
		O_{+1}	O_{+2}	O_{+3}	...	O_{+j}	N

From Table 3.3, I is the number of levels for one categorical variable, and J is the number of levels for the other categorical variable.

O_{i+} is the total number of sample observations at level i and O_{+j} is the total number of sample observations at level j.

O_{ij} is the observed frequency count at level O_{i+} of variable I and level O_{+j} of variable J. N is the total sample size.

A chi-square test for independence was used to determine whether there was a significant association between the independent variables in the data set and the outcome of having cardiovascular disease.

For example, in these thesis, a disease might be classified by cardiovascular disease and no cardiovascular disease and would be associated with gender which would be classified as male and female. We would use a chi-square test for independence to determine whether gender is associated with the outcome of having cardiovascular disease. Other independent variables that may be associated with cardiovascular disease are family history, cholesterol level, overweight, diabetes, smoking, alcohol consumption and hypertension. Age would not be used in this test because it is a continuous variable and not categorical variable.

The test procedure described is appropriate when the following conditions are met:

- The sampling method is simple random sampling.

- The variables under study are each categorical.
- If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

The test procedure consists of three steps: (1) state the hypotheses, (2) analyse sample data, and (3) interpret results.

3.3.1 Statement of the Hypotheses

Suppose that from table 3.3, variable J has O_{+j} levels and variable I has O_{i+} level. The null hypothesis states that knowing the level of variable J does not help you predict the level of variable I. That is, the variables are independent.

H_0 : variable J and variable I are independent

H_1 : variable J and variable I are not independent

The alternative hypothesis states that knowing the level of variable J can help you predict the level of variable I.

3.3.2 Analysis of Sample Data

Using sample data, find the degrees of freedom expected frequencies, test statistic, and p-value associated with the test statistic.

- Degrees of freedom: The degrees of freedom(DF)is equal to:

$$DF = (I - 1)(J - 1) \tag{3.1}$$

In equation 3.1, I is the number of levels for one categorical variable, and J is the number of levels for the other categorical variable.

- Expected frequencies: The expected frequencies counts are computed separately for each level of one categorical variable at each level of other

categorical variable. Compute $O_{i+} \times O_{+j}$ expected frequencies according to the following formula,

$$E_{ij} = \frac{O_{i+} \times O_{+j}}{N} \tag{3.2}$$

In equation 3.2, E_{ij} is the expected frequency count for level O_{i+} of variable I and level O_{+j} of variable J

- Test statistic: The test statistic is a chi-square random variable(χ) defined by the following equation,

$$\chi^2 = \sum_{i}^{I} \sum_{j}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-i)} \tag{3.3}$$

In equation 3.3, O_{ij} is the observed frequency count at level O_{i+} of variable I and level O_{+j} of variable J and E_{ij} is the expected frequency count for level O_{i+} of variable I and level O_{+j} of variable J.

- The $P - value$ is the probability of observing sample statistic as extreme as the test statistics.

3.3.3 Interpretation of Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis.

Typically, this involves comparing the $p - value$ to the significance level, and rejecting the null hypothesis when the $p - value$ is less than the significance level.

3.4 Logistic Regression

3.4.1 Introduction

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables.

It is often the case that the outcome variable is discrete taking on two or more possible values. Over the last decade, the logistic regression model has become, in many fields, the standard method of analysis in this situation.Hosmer et al. (1989).

What distinguishes a logistic regression model from the linear regression model is that the outcome variable in logistic regression is binary or dichotomous. This difference between logistic and linear regression is reflected both in the choice of a parametric model and it assumptions. Once this difference is accounted for, the methods employed in an analysis using logistic regression follow the same general principles used in linear regression.

3.5 Generalised Linear Models and Logistic Regression

The logistic regression model is an example of a broad class of models known as Generalized Linear Models (GLM). For example, GLMs also include linear regression, ANOVA, Poisson regression, etc. There are three components to a GLM:

Random component: refers to the probability distribution of the response variable (Y); e.g. binomial distribution for Y in the binary logistic regression.

Systematic component: refers to the explanatory variables (X_1,X_2,X_3,...,X_k) as a combination of linear predictors of explanatory variables,

e.g $\eta = logit(\pi)$ for logistic regression.

3.5.1 Binary Logistic Regression

Models how binary response variable depend on a set of explanatory variable.

Random Component: The distribution of Y is Binomial.

Systematic Component: X_s are explanatory variables (can be both continuous, discrete, or both) and are linear in the parameter $\beta_0 + \beta_1 x_i + ... + \beta_0 + \beta x_k$

$$\eta = logit(\pi) = \log(\frac{\pi}{1-\pi}) \tag{3.4}$$

3.6 Logistic Regression with Single Independent Variable

The general formula for the logistic regression model with single variable is

$$\pi(X) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{3.5}$$

The transformation of $\pi(X)$ that is central to the study of logistic regression is the logit transformation. This transformation is defined in terms of $\pi(X)$, as

$$g(x) = \left[\frac{\pi(x)}{1-\pi(x)}\right] \tag{3.6}$$

$$g(x) = \ln\left[\frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{1-\left[\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right]}\right] \tag{3.7}$$

$$g(x) = \ln\left[\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \times \frac{1+e^{\beta_0+\beta_1 x}}{1}\right] \tag{3.8}$$

$$g(x) = \ln\left(e^{\beta_0+\beta_1 x}\right) \tag{3.9}$$

$$g(x) = \log_e e^{\beta_0+\beta_1 x} \tag{3.10}$$

$$g(x) = \beta_0 + \beta_1 x \tag{3.11}$$

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

3.7 Fitting the Single Logistic Regression Model

The method of estimation used in fitting the logistic regression model is the maximum likelihood.

In order to apply this method we must first construct a function called the likelihood function. This function expresses the probability of observed data as a function of the unknown parameters. The maximum likelihood estimators of these parameters are chosen to be those values that maximizes this function. Thus, the resulting estimators 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 $\pi(x)$ given in equation (3.5) provides (for an arbitrary value of $\beta = (\beta_0,\beta_1)$, the vector of parameters) the conditional probability that Y is equal to 1 given X.

This will be 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 $\pi(X_i)$ denote the value of $\pi(X_i)$ computed at X_i

A way to express the contribution to the likelihood function for the pair (X_i,Y_i) is through expression

$$\pi(X_i)_{y_i}[1 - \pi(X_i)]_{1-y_i} \tag{3.12}$$

Since the observation are assumed to be independent, the likelihood function is obtained as the product of the terms given in equation 3.12 as

$$L(\Pi : y) = \prod_{i=1}^{n} \pi(X_i)^{y_i}[1 - \pi(X_i)]^{1-y_i} \tag{3.13}$$

$$= \pi(x_i)^{\sum_{i=1}^{n} y_i}(1 - \pi(x_i))^{n-\sum_{i=1}^{n} y_i} \tag{3.14}$$

For estimation, we will work with the log-likelihood. This expression, the log likelihood is given as

$$l(\pi : y) = \ln[l(\beta)] = \sum_{i=1}^{n}\{y_i \ln[\pi(X_i)] + (n - y_i)\ln[1 - \pi(X_i)]\} \tag{3.15}$$

The maximum likelihood estimate (MLE) of π is that value that maximizes l (equivalent to maximizing L).

The first derivative of l with respect to π is

$$\frac{\partial l}{\partial \pi} = \frac{\sum_{i=1}^{n} y_i}{\pi} - \frac{(n - \sum_{i=1}^{n} y_i)}{1 - \pi} \tag{3.16}$$

and is referred to as the score function.

The information function is the negative of the curvature in $l = logL$. For the likelihood considered previously,the information is

$$I(\Pi) = [-\frac{\partial^2 l}{\partial \pi^2}] \tag{3.17}$$

$$= [\frac{\sum_{i=1}^{n} y_i}{\pi^2} + \frac{(n - \sum_{i=1}^{n} y_i)}{(1 - \pi)^2}] \tag{3.18}$$

To find the value of β that maximizes $L(\beta)$ with respect to β_0, β_1, partially and set the resulting expression equate to zero.

These equations, known as the likelihood equation, are;

X

$$[y_i - \pi(X_i)] = 0 \tag{3.19}$$

and

X

$$X_i[y_i - \pi(X_i)] = 0 \tag{3.20}$$

3.8 Testing for the Significance of the Single Independent Variable

In logistic regression, comparison of observed to predicted values is based on the log likelihood function defined in equation (3.15).

The comparison of observed to predicted values using the likelihood function are based on the following expression:

$$D = -2\ln\left[\frac{\textit{likelihood of the fitted model}}{\textit{likelihood of the saturated model}}\right] \tag{3.21}$$

The quantity inside the large bracket in the expression above is called the likelihood ratio test. A saturated model is one that contains as many parameters as there are data points.

Using equation (3.15) and (3.21) becomes

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left[\frac{\pi_i}{y_i}\right] + (1-y_i)\ln\left[\frac{1-\pi_i}{1-y_i}\right]\right] \tag{3.22}$$

From equation (3.22), $\pi_i = \pi(X_i)$. The statistic, D,in the equation is called deviance. This plays the same role as the residual sum of square plays in linear regression. It is identically equal to the sum of square error(*SSE*). In an instance where the values of our outcome variable are 0 and 1 just as in this study, the likelihood of our saturated model is 1. Specifically it follows from the definition of a saturated model that $\pi_i = y_i$ and the likelihood is

$$l(\textit{saturated model}) = \prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{1-y_i} = 1$$

Thus, ie follows from equations (3.21) that the deviance is

$$D = -2(\textit{likelihood of fitted model}) \tag{3.23}$$

Assessing the significance of an independent variables require that we compare the value of *D* with and without the independent variables in the equation. The change in *D* due to the inclusion of the independent variable in the model is obtained as;

$$G = D(\textit{model without the variable}) - D(\textit{model with the variable}) \tag{3.24}$$

This statistic plays the same role in logistic regression as the numerator of the partial *F test* does in linear regression. Because the likelihood of the saturated model is common to both values of *D* being differenced to compute *G*, it can be expressed as;

$$G = -2\ln\left[\frac{\textit{likelihood without the variable}}{\textit{likelihood with the variable}}\right] \tag{3.25}$$

For cases of a multiple independent variable, ie is easy to show when the variables are not in the model, the maximum likelihood estimate of β_0 is $\ln\left(\frac{n_1}{n_0}\right)$ where $n_1 = \sum y_i$ and $n_0 = \sum(y_i - 1)$ and the predicted value is constant, $\frac{n_1}{n}$. In this case, the value of *G* is

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n_0}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}(\hat{\pi}_i)^{y_i}(1-\hat{\pi}_i)^{(1-y_i)}}\right] \tag{3.26}$$

or

$$G = 2\left(\sum_{i=1}^{n}\left[y_i\ln(\hat{\pi}_i) + (1+y_i)\ln(1-\hat{\pi}_i)\right] - \left[n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)\right]\right) \tag{3.27}$$

Under the hypothesis that β_1 is equal to zero, the statistics G follows a Chi-squared distribution with 1 degree of freedom. Two other similar, statistically equivalent test known is the Wald test and Score test. The assumption needed for these tests are the same as those of the likelihood ratio test in equation (3.26). The Wald test is obtained by comparing the maximum likelihood estimate of the slope parameter $\hat{\beta}_1$, to an estimate of its standard error. The resulting ratio, under the hypothesis that $\beta_1 = 0$, will follow a standard normal distribution

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \tag{3.28}$$

Another test use in testing for the significance of a variable is the Score test. This test is based on the distribution theory of the deviation of the loglikelihood. The test statistics for the Score test(ST) is

$$ST = \frac{\sum_{i=1}^{n} X_i(y_i - \bar{y_i})}{\sqrt{(y_i - \bar{y_i})\sum_{i=1}^{n}(X_i - \bar{X}_i)}} \tag{3.29}$$

3.9 Confidence Interval Estimation of Single Logistic Regression Variable

The confidence interval estimators for slope and intercept are based on their respective Wald Tests. The endpoints of a $100(1-\alpha)\%$ confidence interval for the slope coefficient are

$$\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_1) \tag{3.30}$$

and for the intercept they are

$$\hat{\beta}_0 \pm Z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_0) \tag{3.31}$$

In equation (3.31), $Z_{1-\frac{\alpha}{2}}$ is the upper 100(1 − α)% point from the standard normal distribution and $SE()$ denotes a model-based estimator of the standard error of the respective parameter estimator. The estimated values are provided in the output following the fit of a model and, in addition, many statistical software packages provide the endpoints of the interval estimates. The standard error is calculated using the logit of the linear part of the logistic regression model and, as such, is most like the fitted line in a linear regression model. The estimator of the logit is

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3.32}$$

The estimator of the variance of the estimator of the logit requires obtaining the variance of a sum. In this case is

$$\hat{Var}[\hat{g}(x)] = \hat{Var}(\hat{\beta}_0) + x^2 \hat{Var}(\hat{\beta}_1) + 2x\hat{Cov}(\hat{\beta}_0, \hat{\beta}_1) \tag{3.33}$$

In general, the variance of the estimator of the logit requires obtaining the variance of a sum is equal to the sum of the of the variances of each term and twice the covariance of each possible pair of terms formed from the components of sum. The endpoints of a 100(1 − α)% Wald-based confidence interval for the logit are

$$\hat{g}(x) \pm Z_{1-\frac{\alpha}{2}} SE[\hat{g}(x)] \tag{3.34}$$

where $SE[\hat{g}(x)]$ is the positive square of the variance estimator in equation (3.32)

3.10 The multiple Logistic Regression Model

The general form of the multiple logistic regression model is;

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}} \tag{3.35}$$

From equation (3.35), p = the number of independent variables and

$P(Y = 1/x) = \pi(x)$ = the conditional probability that the outcome is present.

The logit of the multiple logistic regression model is given by

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{3.36}$$

in which case the regression model is

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{3.37}$$

3.11 Fitting the Multiple Logistic Regression Model

The method of estimation used in fitting a multiple logistic regression model is the maximum likelihood estimation method. The likelihood function is nearly identical to that given in equation (3.12) with only a change being that $\pi(x)$ is now defined as in equation (3.36). The will be $p + 1$ coefficients likelihood equations that are obtained by differentiating the log likelihood function with respect to $p + 1$ coefficients. The likelihood equation that results are expressed as

$$\sum_{i=1}^{n} \left[y_i - \pi(x_i) \right] = 0 \tag{3.38}$$

and

$$\sum_{i=1}^{n} x_{ij} \left[y_i - \pi(x_i) \right] = 0 \quad \text{for } j=1,2,\ldots,p \tag{3.39}$$

As in univariate model, the selection of the likelihood equation requires special statistical software packages, In calculating the standard error, we will have to find the estimates of the variance and covariance of our coefficients. The method of estimating variances and covariances of the estimated coefficients follows from the theory of maximum likelihood estimation which states that the estimates are obtained from the matrix of second partial derivatives of the log likelihood function. The general form of these partial derivatives is

$$\frac{\partial^2 L(B)}{\partial \beta_j^2} = -\sum_{i=1}^{n} X_{ij}^2 \pi_i (1 - \pi_i) \quad (3.40)$$

and

$$\frac{\partial^2 L(B)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} X_{ij} X_{il} \pi_i (1 - \pi_i) \quad (3.41)$$

for $j, l = 0,1,2,...,p$ where π_i denote $\pi(X_i)$ and p denotes the number of covariates in the model. If $(p + 1) \times (p + 1)$ matrix containing the negative of the terms given in equations (3.40) and (3.40) be denoted as $I(\beta)$. This matrix is called the observed information matrix. The variances and covariances of the estimate coefficients are obtained from the inverse of this matrix which is denoted as;

$$Var[I(\beta)] = I^{-1}(\beta)$$

The estimated standard errors of the estimated coefficient can also be used. This is denoted as;

$$SE(\hat{\beta}_j) = [\hat{Var}(\hat{\beta}_j)]^{\frac{1}{2}} \quad (3.42)$$

for $j = 0,1,2,...,p$. A formulation of the information matrix which is useful for the model fitting and assessment of the fit is $\hat{I}(\hat{\beta}) = X^0 V X$ where X is n by $p + 1$ matrix containing the data for each subject, and V is an $n \times j$ diagonal matrix with general element $\pi_i(1 - \pi_i)$. That is , the matrix X is

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ 1 & X_{11} & X_{12} & \dots & X_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ \vdots & & & & \vdots \\ \vdots & & & & \vdots \\ 1 & X_{11} & X_{12} & \dots & X_{1p} \end{bmatrix} \quad (3.43)$$

The matrix V is

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & & & \\ & \hat{\pi}_2(1-\hat{\pi}_2) & & 0 \\ & & \ddots & \\ 0 & & & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix} \quad (3.44)$$

3.12 Testing for the Significance of the Multiple Logistic Regression Parameters

As in the univariate, the first step in this process is usually to assess the significance of the variable in the model. The likelihood ratio test for overall significance of the p coefficients for independent variables in the model is performed in exactly the same manner as in the univariate case. The test is based on the statistic G given in equation (3.25). The only difference is that the fitted values, π under the model are based on the vector containing $p + 1$ parameters, $\hat{\beta}$.

Under the null hypothesis that P slope" coefficient for the covariates in the model are equal to zero, the distribution of G will be Chi-squared with p degree-of-freedom.

The Wald test is obtained by comparing the maximum likelihood estimate of the slope parameter, β_j to an estimate of the standard error. The resulting ratio, under the hypothesis that $H_0 : \beta_j = 0$, for $j = 0,1,2,...,p$ will follow a standard normal distribution

$$W = \frac{-\beta j}{SE(\beta j)} \quad (3.45)$$

3.13 Confidence Interval Estimation in Multiple Logistic Regression

The confidence interval estimate for the logit are a bit more complicated for the multiple variable model than the results in equation (3.33). The basic idea is the same only there are now more terms involved in the summation. It follows from equation (3.35) that the general expression for the estimator of the logit for a model containing p covariates.

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p \tag{3.46}$$

An alternative way to express the estimator of the logit in the equation (3.37) is through the use of the vector notation as $\hat{g}(x) = X^0\hat{\beta}$, where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ denote the estimator of the $p+1$ covariates and the vector $X^0 = (x_0, x_1, x_2, \ldots, x_p)$ represent the constant and a set of values of the $p-$ *covariates* in the model, where $x_0 = 1$.

The expression for the estimator of the variance of the estimator of the logit in equation(3.46) is

$$\hat{Var}[\hat{g}(x)] = \sum_{j=0}^{p} X_j^2 Var(\hat{\beta}_j) + \sum_{j=0}^{p}\sum_{k=j+1}^{p} 2X_jX_kCov(\hat{\beta}_j, \hat{\beta}_k) \tag{3.47}$$

This can be express much more concisely by using the matrix expression for the estimator of the variance of the estimator of the coefficients. From the expression for the observed information matrix, we have that,

$$\hat{Var}(\hat{\beta}) = (X^0VX)^{-1} \tag{3.48}$$

It follows from equation (3.39) that an equivalent expression for the estimator in equation (3.47) is

$$\hat{Var}[\hat{g}(x)] = X^0 \hat{Var}(\hat{\beta}) X$$

$$= X^0 (X^0 V X)^{-1} X \tag{3.49}$$

3.14 Odds Ratio

The odds of the outcome being present among individuals with $Y = 1$ is defined as $\pi(1)/[1 - \pi(1)]$. Similarly, the odds of the outcome being present among individuals with $Y = 0$ is defined as $\pi(0)/[1 - \pi(0)]$. The odds ratio, denoted by OR, is defined as the ratio of the odds for $Y = 1$ to the odds for $Y = 0$, and is given by the equation

$$OR = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]} \tag{3.50}$$

Outcome	Independent (X)	
Variable (Y)	$X = 1$	$X = 0$
$y = 1$	$\pi(1) = \frac{e^{\beta 0+\beta 1}}{1+e^{\beta 0+\beta 1}}$	$\pi(0) = \frac{e^{\beta 0}}{1+e^{\beta 0}}$
$y = 0$	$1-\pi(1) = \frac{1}{1+e^{\beta 0+\beta 1}}$	$1-\pi(0) = \frac{1}{1+e^{\beta 0}}$
Total	1.0	1.0

$$OR = \frac{\left[\frac{e^{\beta 0+\beta 1}}{1+e^{\beta 0+\beta 1}}\right] / \left[\frac{1}{1+e^{\beta 0+\beta 1}}\right]}{\left[\frac{e^{\beta 0}}{1+e^{\beta 0}}\right] / \left[\frac{1}{1+e^{\beta 0}}\right]}$$

$$= \frac{e^{\beta 0+\beta 1}}{e^{\beta 0}}$$

$$= e^{(\beta 0+\beta 1)-\beta 0}$$

$$OR = e^{\beta 1}$$

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

$$OR = e^{\beta 1} \tag{3.51}$$

The interpretation given for the odds ratio is based on the fact that in many instances it approximates a quantify called the relative risk. This parameter is equal to the ratio $\pi(1)/\pi(0)$. It follows from equation (3.49) that the odds ratio approximate the relative risk if $[1 - \pi(0)]/[1 - \pi(1)] \approx 1$. This holds when $\pi(x)$ is small for both $x = 1$ and $x = 0$

3.15 Confidence Limits for odds Ratio

This is obtained by finding the confidence limits for the log odds ratio. In general, the limits for a $100(1-\alpha)\%$ confidence interval for the coefficient are of the form

$$\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} \times SE(\hat{\beta}_1) \tag{3.52}$$

The corresponding limits for the odds ratio obtained by exponentiating these limits are

$$\exp[\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} \times SE(\hat{\beta}_1)] \tag{3.53}$$

3.16 Specifying the Multiple Logistic Model

The logit of the multiple logistic regression model is given by the equation (3.36)

$$g(x) = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 Family\ History + \beta_4 Overweight + \beta_5 Diabetes + \beta_6 Smoking + \beta_7 Cholesterol\ level + \beta_8 Hypertension + \beta_9 Alcohol\ consumption$$

in which case the logistic regression model is

$$P(Y = 1/x) = \pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{3.54}$$

$$\pi(x) = \frac{\exp^{\beta_0+\beta_1 Age+\beta_2 Gend+\beta_3 Fami+\beta_4 Over+\beta_5 Diab+\beta_6 Smok+\beta_7 Chol+\beta_8 Hyp+\beta_9 Alcohol}}{1 + \exp^{\beta_0+\beta_1 Age+\beta_2 Gend+\beta_3 Fami+\beta_4 Over+\beta_5 Diab+\beta_6 Smok+\beta_7 Chol+\beta_8 Hyp+\beta_9 Alcohol}}$$

$$Y_i = \begin{cases} 1 & \text{, } \textit{Cardiovascular disease Present} \\ 0 & \textit{Otherwise} \end{cases}$$

Gender	(*male* = 1, *female* = 0)
Family history	(*Yes* = 1, *No* = 0)
Overweight	(*Yes* = 1, *No* = 0)
Diabetes	(*Yes* = 1, *No* = 0)
Smoking	(*Yes* = 1, *No* = 0)
Hypertension	(*Yes* = 1, *No* = 0)
Cholesterol Level	(*High* = 1 , *Low* = 0)
Alcohol consumption	(*Yes* = 1, *No* = 0)

3.17 Pearson Chi-square statistic and deviance

In logistic regression there are several possible ways to measure the difference between the observed and fitted values. To emphasize the fact that the fitted values in logistic regression are calculated for each covariates pattern and depend on the estimated probability for that covariate pattern, we denote the fitted value for the *jth* covariate pattern as $\hat{y}_j$ where

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(x_j)}}{1 + e^{\hat{g}(x_j)}} \tag{3.55}$$

and in equation(3.55), $\hat{g}(x_j)$ is the estimated logit. We consider two measures of the difference between the observed and fitted values, that is the Pearson residual and the deviance residual. For a particular covariate pattern, the

Pearson residual is defined as

$$r(y_j\hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}} \tag{3.56}$$

The summary statistic based on these residuals is the Pearson Chi-square statistic

$$\chi^2 = \sum_{j=1}^{J} r(y_j,\hat{\pi}_j)^2 \tag{3.57}$$

The deviance residual is defined as

$$d(y_j\hat{\pi}_j) = \pm\left\{2y_j \ln\left[\frac{y_j}{m_j\hat{\pi}_j}\right] - (m_j - y_j)\ln\left[\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right]\right\}^{\frac{1}{2}} \tag{3.58}$$

where the sign ± is the same as the sign of $(y_j - m_j\hat{\pi}_j)$. For covariates patterns with $y_j = 0$ the deviance residual is

$$d(y_j\hat{\pi}_j) = -\sqrt{2m_j\,|\ln(1-\hat{\pi}_j)\,|} \tag{3.59}$$

The summary statistic based on the deviance residuals is the deviance

$$D = \sum_{j=1}^{J} d(y_j,\hat{\pi}_j)^2 \tag{3.60}$$

3.18 The Hosmer-Lemeshow Test

The Hosmer-Lemeshow goodness-of-fit test statistic, $\hat{C}$ is obtained by calculating the Pearson Chi-squared statistic from $g \times 2$ table of observed and estimated expected frequencies. A formula defining the calculation of $\hat{C}$ is as follows:

$$\hat{C} = \sum_{i=1}^{n} \frac{(O_k - n_k\bar{\pi}_k)^2}{n_k\bar{\pi}_k(1-\bar{\pi}_k)} \tag{3.61}$$

where n_k is the total number of subject in the k^{th} group, c_k denotes the number of covariate patterns in the k^{th} decile.

$$O_k = \sum_{j=1}^{c_k} y_i \tag{3.62}$$

is the number of response among the c_k covariate patterns, and

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n_k} \tag{3.63}$$

is the average estimated probability. Using an extensive set of simulations, Hosmer and Lemeshow(1980) demonstrated that, when $J = n$ and the fitted logistic regression model is the corrected model, the distribution of the statistic $\hat{C}$ is well approximated by the Chi-squared distribution with $g - 2$ degree of freedom χ^2_{g-2}

3.19 Selecting the best model for cardiovascular Disease

In selecting the best model for cardiovascular disease, the calculated Alkaike Information Criterion (AIC) of each model is to be considered. The smaller these values, the better the model fits the data.

Chapter 4

Data Analysis and Results

4.1 Introduction

The chapter describes in detail the results of our research. The results are in two parts. The preliminary results which involve our independent test of cardiovascular disease with some of our variables in the data and test of multicollinearity among the variables.

The second results involves the logistic regression model with our predictor variables. The data used for the research consist of nine variables; age, gender, family history, cholesterol level, overweight, diabetes, smoking, alcohol consumption and hypertension.

There were 78 males and 31 females. 50 people had a family history of cardiovascular disease whiles 59 people did not. 34 people consume alcohol and 75 did not. 52 were into smoking and 57 were not. 75 were hypertensive and 34 were not hypertensive. 62 were diabetic and 47 were not. 47 people were overweight and 62 were not, and 66 people had high cholesterol level and 43 people had low cholesterol level.

4.2 Preliminary Results

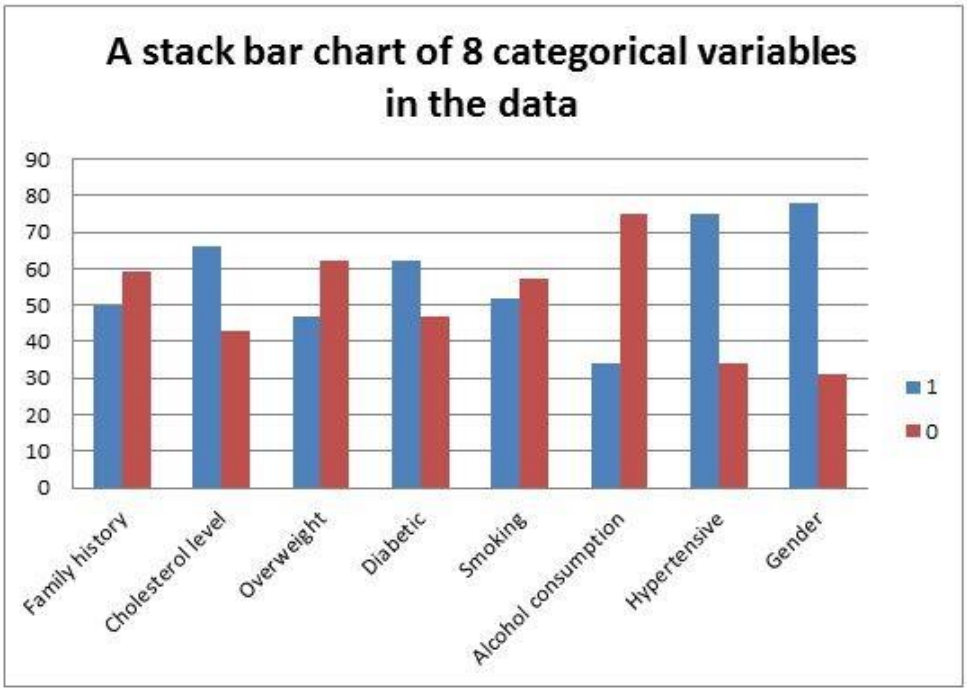


Figure 4.1: A stack bar chart of categorical variables in the data

The Figure 4.1 shows stack bar chart of the frequencies of the 8 categorical variables. The 1 or blue bars represent 'yes' and male and 0 or the brown bars represent 'no' and females.

There were 78 males(1) and 31 females(0). 50 people had a family history of cardiovascular disease(1) while 59 people did not(0). 34 people consume alcohol(1) and 75 did not(0). 75 people were hypertensive(1) and 34 were not hypertensive(0). 62 people were diabetic(1) and 47 were not (0). 47 people were overweight(1) and 62 were not(0) and 66 people had high cholesterol(1) and 43 people had low cholesterol(0).

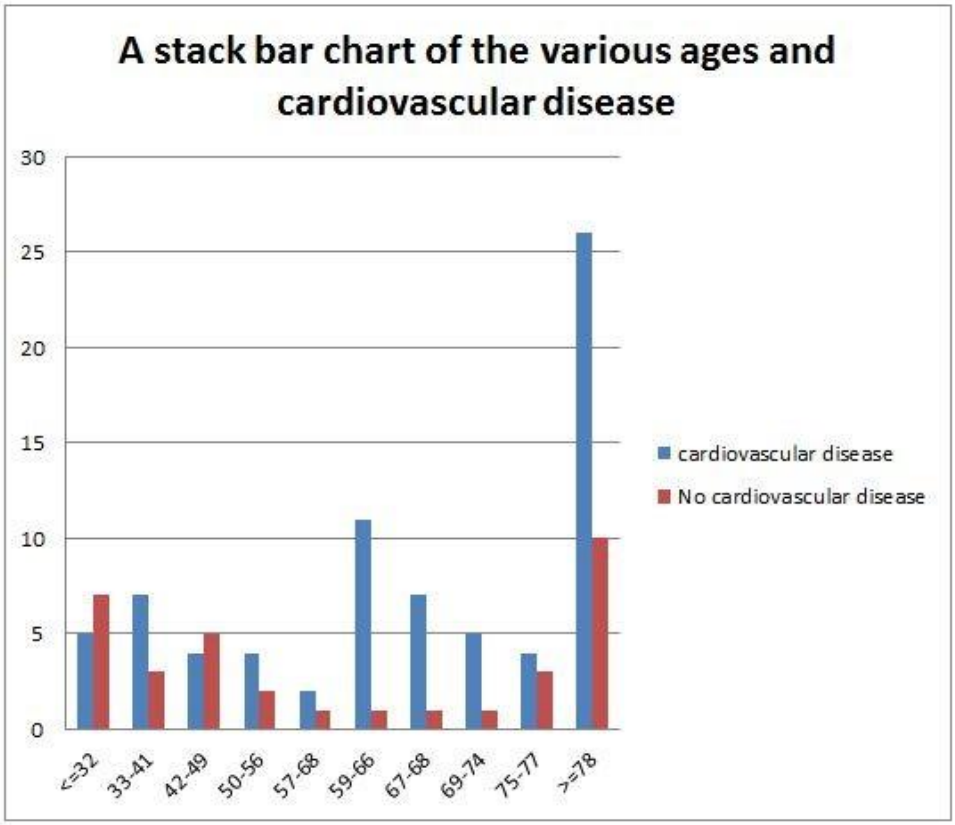


Figure 4.2: A stack bar chart of the various ages and cardiovascular status of all 109 subjects from our data

From the Figure 4.2, age category of 32 years and below and 42-49 years, show that patients without cardiovascular disease dominates in terms of frequency counts whilst patients with cardiovascular disease increased with frequency counts for ages of 50-70 years and above.

Table 4.1: Independent test for cardiovascular disease versus overweight

		Overweight		
		No	Yes	
Cardiovascular.disease				
	No	27	7	
	Yes	35	40	
	$\chi^2 = 10.228$	$df = 1$		$p - value = 0.001384$

H_0: Outcome of cardiovascular disease is not associated with overweight.

H_1: Outcome of cardiovascular disease is associated with overweight.

The p-value (0.001384) is less than the significant level (0.05). Thus, we conclude that there is statistical evidence between overweight and outcome of cardiovascular disease.

Table 4.2: Independent test for cardiovascular disease versus hypertension

		Hypertension		
		No	Yes	
Cardiovascular.disease				
	No	25	9	
	Yes	9	66	
	$\chi^2 = 41.27$	$df = 1$		$p - value = 1.329e - 10$

H_0: Outcome of cardiovascular disease is not associated with hypertension.

H_1: Outcome of cardiovascular disease is associated with hypertension. The p-value (0.000) is less than the significant level (0.05). Thus, we conclude that there is statistical evidence between hypertension and outcome of cardiovascular disease.

Table 4.3: Independent test for cardiovascular disease versus diabetes

		Hypertension		
		No	Yes	
Cardiovascular.disease				
	No	20	14	
	Yes	27	48	
	$\chi^2 = 4.969$	$df = 1$		$p - value = 0.02581$

H_0: Outcome of cardiovascular disease is not associated with diabetes.

H_1: Outcome of cardiovascular disease is associated with diabetes.
The p-value (0.02581) is less than the significant level (0.05). Thus, we conclude that there is statistical evidence between diabetes and outcome of cardiovascular disease.

Table 4.4: Independent test for cardiovascular disease versus smoking

		Hypertension		
		No	Yes	
Cardiovascular.disease				
	No	50	9	
	Yes	32	43	
	$\chi^2 = 8.932$	$df = 1$		$p - value = 0.002802$

H_0: Outcome of cardiovascular disease is not associated with smoking.

H_1: Outcome of cardiovascular disease is associated with smoking.

The p-value (0.0028) is less than the significant level (0.05). Thus, we conclude that there is statistical evidence between smoking and outcome of cardiovascular disease.

Table 4.5: Independent test for cardiovascular disease versus alcohol consumption

		Hypertension		
		No	Yes	
Cardiovascular.disease				
	No	22	12	
	Yes	53	22	
	$\chi^2 = 0.3873$	$df = 1$		$p - value = 0.5337$

H_0: Outcome of cardiovascular disease is not associated with alcohol consumption.

H_1: Outcome of cardiovascular disease is associated with alcohol consumption.

The p-value (0.537) is greater than the significant level (0.05). Thus, we conclude that there is no statistical evidence between alcohol consumption and outcome of cardiovascular disease.

Table 4.6: Independent test for cardiovascular disease versus all the predictor variables

Variable	Chi-Square Value	P-value
Family history	1.9944	0.1579
Cholesterol level	3.485	0.06194
Overweight	10.228	0.001384
Diabetes	4.969	0.02581
Smoking	8.932	0.002802
Alcohol Consumption	0.3873	0.5337
Hypertension	41.27	$1.329e^{-10}$
Gender	8.146	0.003718

In table4.6, the variables overweight, diabetes, smoking, hypertension and gender were associated with the outcome of cardiovascular disease since their respective p-values are less than the significant level (0.05).

4.2.1 Test of Association

A confounding variable (confounding factor, lurking variable, etc) is defined as an extraneous variable in a statistical model that correlates (positively or

negatively) with both the dependent and the independent variable. ORs in our calculation represents odds ratio.

Table 4.7: The relationship between cardiovascular disease and hypertension for males

Male		cardiovascular disease		
		Yes	No	Total
Hypertension	Yes	53	7	60
	No	7	11	18
Total		60	18	78

Computing the odds ratio in table 4.7, O

$$R = \frac{53(11)}{7(7)} = 11.898$$

From the results in table 4.7, the probability of a hypertensive male patient being diagnosed of cardiovascular disease is 12 times that of non-hypertensive male patients.

Table 4.8: The relationship between cardiovascular disease and hypertension for females

Female		cardiovascular disease		
		Yes	No	Total
Hypertension	Yes	13	2	15
	No	2	14	16
Total		15	16	31

Computing the odds ratio in table 4.8 O

$$R = \frac{13(14)}{2(2)} = 45.5$$

From the results in table 4.8, the probability of a female hypertensive patients being diagnosed of cardiovascular disease is 20 times that of non-hypertensive patients.

Table 4.9: Association of cardiovascular disease and hypertension ignoring gender as the confounder

		cardiovascular disease Yes	No	Total
Hypertension	Yes	66	9	75
	No	9	25	34
Total		75	34	109

Computing the odds ratio in table 4.9 O

$$R = \frac{66(25)}{9(9)} = 20.370$$

From the results in table 4.9, the probability of hypertensive patients being diagnosed of cardiovascular disease is 20 times that of non-hypertensive patients.

4.3 Collinearity Diagnostic Test

Table 4.10: Correlation Diagnostics

Predictor variables	Tolerance	Variance Inflation Factor
Family History	0.727	1.376
Cholesterol level	0.770	1.298
Overweight	0.648	1.543
Diabetes	0.602	1.662
Smoking	0.495	2.022
Alcohol consumption	0.790	1.266
Hypertension	0.654	1.529
Gender	0.251	3.991
Age	0.325	3.081

Before building the model for cardiovascular disease, the set of independent variables must be tested to see if they are fit to be included in the model using Tolerance and the Variance Inflation Factor (VIF). The tolerance of a variable is defined as 1 minus the squared multiple correlation of this variable in the regression equation. The smaller the tolerance of the variable ,the more redundant is its contribution to the regression(shows presence of multicollinearity). The variance inflation factor quantifies the severity of multicollinearity in the regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient is increased

because of collinearity. A VIF of 10 and above shows the presence of multicollinearity.

From table 4.10, the variance inflation factor(VIF) values of all the predictor variables are less than 10 indicating lack of collinearity among themselves.

4.4 Logistic Modeling with Categorical Predictors

The data set contains eight categorical predictor variables: Family history, cholesterol level, overweight, diabetes, smoking, alcohol consumption, hypertension and gender. The response variable was the cardiovascular disease status.

Table 4.11: Testing Global Null Hypothesis: $\beta = 0$

χ^2	90.96172
degree of freedom	9
$p - value$	1.043822e-15

From table 4.11,the test statistic is the likelihood ratio test which is the difference between the residual deviance for the model with predictors and the null model. The test statistic is distributed chi-square with degrees of freedom equal to the differences in degrees of freedom between the current model and the null model.

The chi-square value of 90.962 with 9 degrees of freedom and an associated p-value(0.000) less than the significant level(0.001) indicate that the model as a whole fits significantly better than an empty model.

Thus, we reject our null hypothesis and conclude that at least one of the regression coefficients in the model is not equal to zero.

Table 4.12: Model 1-Analysis of Maximum Likelihood Estimates for model with all the predictor variables

Coefficients:	Estimate	Std. Error	$z-value$	$Pr(>\|z\|)$
(Intercept)	-4.24934	2.15682	-1.970	0.048817 *
Family History	-4.27943	1.45778	-2.936	0.003329 **
Cholesterol level	-2.37345	1.12748	-2.105	0.035283 *
Overweight	6.88352	1.69214	4.068	4.74e-05 ***
Diabetes	1.99615	1.20477	1.657	0.097545 .
Smoking	1.07441	1.12339	0.956	0.338869
Alcohol consumption	1.46456	1.04064	1.407	0.159320
Hypertension	6.40021	1.70692	3.750	0.000177 ***
Gender	-1.71541	2.01302	-0.852	0.394128
Age	0.03360	0.04911	0.684	0.493936

Parameter estimates are displayed in table 4.12.The maximum likelihood estimate output shows the coefficients, their standard errors, the z-statistic(sometimes called a Wald Z-statistic), and the associated p-values. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

The fit indices, including the null, deviance residuals and the AIC are found in table 6.1 at appendix A. This determines whether the model with predictors fits significantly better than a model with just an intercept.(i.e. a null model). Hence, the significant predictors were family history, cholesterol level, overweight and hypertension.

4.4.1 The fitted multiple logistic regression model with all parameters

So we used the estimates from the multiple logistic regression for making our interpretation of the model from 4.12 $\text{logit}(P(Y = 1/x)) = -4.249 + 0.034 \times \text{age} - 1.715 \times \text{gender} - 4.279 \times \text{family history} + 6.884 \times \text{overweight} + 1.996 \times \text{diabetic} + 1.074 \times \text{smoking} - 2.373 \times \text{cholesterol level} + 6.400 \times \text{hypertensive} + 1.465 \times \text{alcohol consumption}$

Table 4.13: Model 2-Analysis of Maximum Likelihood Estimates After Stepwise variable Selection

Coefficients:	Estimate	Std. Error	*zvalue*	*Pr*(> |*z*|)
(Intercept)	-2.7892	1.2689	-2.198	0.02794 *
Family History	-4.1933	1.3387	-3.132	0.00173 **
Cholesterol level	-1.9693	1.0418	-1.890	0.04872 *
Overweight	6.2767	1.4819	4.236	2.28e-05 ***
Diabetes	1.7783	1.0846	1.640	0.10109
Smoking	1.3755	0.9644	1.426	0.15377
Hypertension	5.3956	1.3138	4.107	4.01e-05 ***

The maximum likelihood estimate output from From table 4.13 shows the coefficients, their standard errors, the z-statistic(sometimes called a Wald Z-statistic), and the associated p-values after stepwise selection of the variables. The AIC and the indices of model fit are found in table 6.2 at appendix A.

4.4.2 The fitted multiple logistic regression model after stepwise selection of variables

So we used the estimates from the multiple logistic regression for making our interpretation of the model from 4.12 $\text{logit}(P(Y = 1/x)) =$

$-2.7892 - 4.1933 \times$ family history $- 1.9693 \times$ cholesterol level $+ 6.2767 \times$ overweight $+ 1.7783 \times$ diabetic $+ 1.3755 \times$ smoking $+ 5.3956 \times$ hypertensive

Table 4.14: Model 3-Analysis of the Maximum Likelihood Estimates of only significant Predictors

Coefficients:	Estimate	Std. Error	$z - value$	$Pr(>\|z\|)$	OR
(Intercept)	-2.2016	1.0615	-2.074	0.03808 *	0.110624
Family History	-2.8963	0.9080	-3.190	0.00142 **	0.055225
Cholesterol level	-1.6001	0.9318	-3.117	0.00202 **	0.201886
Overweight	5.9227	1.4686	4.033	5.51e-05 ***	373.429129
Hypertension	6.1440	1.3886	4.425	9.66e-06 ***	465.909030

Table 4.14,is the maximum likelihood estimates output of the model with only significant predictors. The AIC and the indices of model fit are found in table 6.3 at appendix A.

4.4.3 The fitted multiple logistic regression model with only significant parameters.

$\text{logit}(P(Y = 1/x)) = -2.202 - 2.896 \times$ family history $+ 5.992 \times$ overweight $+ 6.144 \times$ hypertension $- 1.600 \times$ cholesterol level.

Table 4.15: Odds Ratios and Confidence Intervals for the coefficient estimates of the Significant Predictors

coefficients	OR	2.5 %	97.5 %
(Intercept)	0.110624	0.005952625	6.002168e-01
Family.History	0.055225	0.006942385	2.771869e-01
Cholesterol.level	0.201886	0.024358944	1.085960e+00
Overweight	373.429129	31.911770435	1.251549e+04
Hypertension	465.909030	49.503095434	1.432011e+04

From table 4.15, For a one unit increase in hypertension, the odds of being diagnosed of cardiovascular disease versus not being diagnosed increases by a factor 465.91. Hence, there exist a strong association between hypertension and cardiovascular disease.

Table 4.16: Assessing Model Fit by Akaike Information Criterion (AIC)

Model	AIC
1	64.336
2	60.85
3	60.125

Table 4.16 shows the Akaike Information Criterion (AICs) of the three models.In selecting the best model for cardiovascular disease, the calculated Alkaike Information Criterion (AIC) of each model is to be considered. The smaller these values, the better the model fits the data. Hence, the third model which is the model with only the significant predictors is selected.

Table 4.17: Assessing Model Fit by Hosmer and Lemeshow Test

Model	chi-square	D.F	sig
1	5.047	8	0.753

The table 4.17 shows the Hosmer and Lemeshow goodness of fit test. How well our model fits depends on the difference between the model and the observed data. Our model appears to fit well because we have no significant difference between the model and the observed data (i.e. the p-value is above 0.05).

4.5 Selecting the best model for cardiovascular Disease

In selecting the best model for cardiovascular disease, the calculated Akaike Information Criterion (AIC) of each model was considered. The smaller these values, the better the model fits the data. From the calculated AIC values presented table 4.16, the model with only the significant parameters has the lowest AIC. Hence, the model with only significant predictors will be used for prediction.

Example 4.0

Calculating the Log odds of some individuals with our model

- A person who is hypertensive, overweight and has high cholesterol level and also has family history of cardiovascular disease

$$\text{logit}(P(Y = 1/x)) = -2.202 - 2.896(1) + 5.992(1) + 6.144(1) - 1.600)(1)$$

$$= \mathbf{5.368}$$

This person has a high risk of getting cardiovascular disease.

- A person who is hypertensive but not overweight and has high cholesterol level with no family history of cardiovascular disease

$$\text{logit}(P(Y = 1/x)) = -2.202 - 2.896(0) + 5.992(0) + 6.144(1) - 1.600)(1)$$
$$= \mathbf{2.342}$$

This person has 2.342 risk of getting cardiovascular disease.

- A person who is overweight and has family history of cardiovascular disease but is not hypertensive and has low cholesterol level

$$\text{logit}(P(Y = 1/x)) = -2.202 - 2.896(1) + 5.992(1) + 6.144(0) - 1.600(0)$$
$$= \mathbf{0.824}$$

This person has low risk of getting cardiovascular disease.

Chapter 5

Conclusions and Recommendations

5.1 Introduction

This chapter talks of the conclusions and recommendations of the entire study.

5.2 Conclusions

From the study, the following conclusions were drawn from both our preliminary results and the results from our model in achieving our objectives;

1. The results showed that risk factors such as family history, cholesterol level,overweight and hypertension were significant to the diagnosis of cardiovascular diseases. However, risk factors; diabetes, smoking, gender, age and alcohol consumption were not good predictors of cardiovascular diseases.

2. Hypertension poses the greatest risk to the diagnosis of cardiovascular disease.

5.3 Recommendations

Based on the findings from this study, the following recommendations are made to reduce cardiovascular diseases.

1. Regular health screening to check hypertension and cholesterol levels andhealth promotion programs to check overweight may help to prevent car-
   diovascular diseases.

2. Access to and use of health services need to be increased in our communitiesparticularly for cardiovascular diseases. .

REFERENCES

Abbey, M., Owen, A., Suzakawa, M., Roach, P., and Nestel, P. (1999). Effects of menopause and hormone replacement therapy on plasma lipids, lipoproteins and ldl-receptor activity. *PUBMED*, 259-69.

Assareh, A. R., Cheraghi, M., Nourizadeh, M., anadHabib Haybar, F. D., and Kiarsi, M. R. (2013). Distributions of ischemic heart disease risk factors in patients who were admitted for angioplasty in iran. *World Journal of Cardiovascular Diseases*, 3:45–49.

Djousse, L., Biggs, M., Mukamal, k. J., and Siscovick, D. (2007). Alcohol consumption and type 2 diabetes among older adults: The cardiovascular health study. pages 1758–1765.

Dominguez-Rodriguez, A., Arroyo-Ucar, E., and Pedro Abreu-Gonzalez and, G. B.-P. (2013). Smoking and the risk of acute coronary syndrome in young women treated in an emergency department. *World Journal of Cardiovascular Diseases*, 3:9–12.

Eloamany, M. F., Badran, H. M., Salah, T., and Kamal, A. M. (2011). Diagnostics value of dobutamine stress doppler tissue imaging in diabetic patients with suspected coronary artery diseaase. *World Jounal of Cardiovascular Diseases*, 1:1–12.

Grundtvig, M., Hagen, T., German, M., and Reikvam, A. (2009). Sex-based differences in premature first myocardial infarction caused by smoking: twice as many years lost by women as by men. *PUBMED*, pages 16:174–9.

Hosmer, D. W., Jovanovic, B., and Lemeshow, S. (1989). Best subsets logistic regression. biometrics. pages 1265–1270.

Ira, S. and Nancy, H. (1997). Cigarette smoking, cardiovascular disease, and stroke. *American Heart Association*.

Longjian, L. and Howard, J. (2012). Epidemiology of heart failure and the scope and the scope of the problem. *World Jounal of Cardiovascular Diseases*, 5.

Longo-Mbenza, B., Moise, M., Thiery, G., Igor, L. P., Stephen, C., and Emmanuel, M. (2014). Association of high density lipoprotein cholesterol and framingham cardiovascular risk with diabetic retinopathy in african type 2 diabetics. *World Journal of Cardiovascular Diseases*, pages 179–188.

Maas, A. and Appelman, Y. (2010). Gender difference in coronary heart disease. *Netherlands Heart Journal*, 18:598–602.

Olutobi, A., John, k. A., Ama, d.-g. A., and Kwadwo, A. K. (2001). Patterns of cardiovascular disease mortality in ghana: A 5-year review of autopsy cases at korle-bu teaching hospital.

Pohjola-Sintonen, S., Rissanent, A., Liskolat, P., and Luomanma, K. (1998). Family history as a risk factor of coronary heart disease in patients under 60 years of age. *European Heart Journal*, pages 235–239.

Sesso, H., Lee, I., Gaziano, J., Rexrode, K., Glynn, R., and Buring, J. (2001). Maternal and paternal history of myocardial infarction and risk of cardiovascular disease in men and women. *PUBMED*, pages 393–398.

Shaikh, A. S., Muhammad, M. A., Mohsin, S., Qalab, A., Zehra, F., and Atiq, M. (2014). Risk factors associated with anthracycline induced cardiac dysfunction in pediatric patients. *World Jounal of Cardiovascular Diseases*, 4:377–383.

Simon, J. and Rosolova (2002). Family history- and independent risk factors for coronary heart diseaase, it is time to be practical. *European Heart Journal*, pages 1637–1638.

Tanmay, N. and Arnab, G. (2013). Cardiovascular risk factors in asian indian population: Systematic review. *Jounal of Cardiovascular Diseases Research*, 4:222–228.

Towfighi, A., Zheng, L., and Ovbiagele, B. (2009). Sex-specific trends in midlife coronary heart diseaase risk and prevalence. *PUBMED*, pages 169:1762–6.

Vithanage, P., Ranhith Kumarasiri and, S. A., andRohini Tennakoon, M. K., Gunawardana, N., and andSrinath Illeperuma, U. P. (2013). Among the risk factors of myocardial infarction, anthropometry has no association: A case control study in the central region of sri lanka. *World Journal of Cardiovascular Diseases*, 3:1–5.

WHO (2013a). Global health estimates: Deaths by age, sex and cause.

WHO (2013b). High blood pressure - country experiences and effective interventions utilized across the european region and effective interventions utilized across the european region.

WHO (2013c). World health organisation statistics.

WHO (2013d). World health report on diabetes.

WHO (2014). Alcohol kills 3.3 million people in 2012. *Ghana News Agency*.

Yun-Mi, S., Joohon, S., and Joung, S. K. (1998). Which cholesterol level is related to the lowest mortality in population with low men cholesterol level: A 6.4 year follow-up study of 482.472 korean men. *American Journal of Epidemiology*, 151:739–747.

Chapter 6

Appendix A

Table 6.1: Summary of Output for model 1

Deviance Residuals	Min	1Q	Median	3Q
	-4.0893	-0.1078	0.0460	0.1979

Dispersion parameter for binomial family taken to be 1
Null deviance is 135.298 on 108 degrees of freedom
Residual deviance is 44.336 on 99 degrees of freedom
AIC is 64.336

Number of Fisher Scoring iterations: 7

Table 6.2: Summary of Output for model 2

Deviance Residuals	Min	1Q	Median	3Q	M
	-3.9608	-0.1307	0.0793	0.2113	1.5

Null deviance: 135.30 on 108 degrees of freedom
Residual deviance: 46.85 on 102 degrees of freedom
AIC: 60.85
Number of Fisher Scoring iterations: 7

Table 6.3: Summary of Output for model 3

Deviance Residuals	Min	1Q	Median	3Q
	-3.7336	-0.2102	0.0964	0.4284

Dispersion parameter for binomial family taken to be 1
Null deviance is 135.298 on 108 degrees of freedom
Residual deviance is 51.817 on 104 degrees of freedom
AIC is 60.125
Number of Fisher Scoring iterations is 7

Table 6.4: Independent test for cardiovascular disease versus gender

		Hypertension		
		No	Yes	
Cardiovascular.disease				
	No	16	18	
	Yes	15	60	
	$\chi^2 = 8.416$	$df = 1$		$p - value = 0.003718$

Table 6.5: Independent test for cardiovascular disease versus cholesterol

		Hypertension		
		No	Yes	
Cardiovascular.disease				
	No	9	25	
	Yes	34	41	
	$\chi^2 = 3.485$	$df = 1$		$p - value = 0.06194$

Table 6.6: Independent test for cardiovascular disease versus family history

	Hypertension	

		No	Yes
Cardiovascular.disease			
	No	15	19
	Yes	44	31
	$\chi^2 = 1.9944$	$df = 1$	$p - value = 0.1579$