Multiple logistic regression modeling on risk factors of diabetes.
Case study of Gitwe Hospital (2011-2013).

Sylvère Niyikora

A thesis submitted in partial fulfillment for the degree of Master of Science in applied statistics in Jomo Kenyatta University of Agriculture and Technology

2015
DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature: ………………………… Date: ……………………………

Sylvère Niyikora

This thesis has been submitted for examination with our approval as University Supervisors.

Signature: ………………………… Date: …………………

Dr. Joseph Mungátu

JKUAT, Kenya

Signature: ………………………… Date: ……………………………

Dr. Marcel Ndengo

JKUAT, Kigali
DEDICATION

This research project is dedicated especially to my Almighty God for His abundant blessings and protection, to my Wife NYIRAHABIMANA THERESIE and to my Children: ISINGIZWE NIYIKORA KEVIN, IHIMBAZWE NIYIKORA KEVINE and IKUZWE NIYIKORA CALVINE for the sacrifice they made for me to complete this project, their love, care, support, daily encouragement and enthusiasm that inspired me to achieve this goal.
ACKNOWLEDGEMENTS

I give the Almighty God glory and honor for bringing me this far. Thank you, God.

I am deeply indebted to my supervisors Dr. Joseph Mungátu and Dr. Marcel Ndengo for their guidance and inspiration. I am thankful for their technical support and encouragement.

I am deeply grateful to JKUAT for the initiative taken to open up the master’s level in applied statistics.

I am deeply grateful to Gitwe Hospital I visited to make this work a success and to the entire staff members of JKUAT who in diverse ways made this dream a reality.

I also wish to record my appreciation to my wife NYIRAHABIMANA Thérésie for her financial support and fruitful suggestions and encouragement.

Finally, special profound thanks go to all other concerned people who in one way or the other helped me in producing this work, I say thank you.
# TABLE OF CONTENTS

DECLARATION.................................................................................................................. i  
DEDICATION.................................................................................................................... ii  
ACKNOWLEDGEMENTS .................................................................................................. iii  
LIST OF TABLES ............................................................................................................. viii  
LIST OF FIGURES .......................................................................................................... ix  
LIST OF APPENDICES ................................................................................................... x  
ABBREVIATIONS ............................................................................................................ xi  
ABSTRACT ....................................................................................................................... xii  

CHAPTER ONE

INTRODUCTION................................................................................................................. 1  
  1.1. BACKGROUND OF THE STUDY ............................................................................... 1  
  1.2. STATEMENT OF THE PROBLEM ............................................................................ 2  
  1.3. RESEARCH OBJECTIVES ....................................................................................... 5  
     1.3.1. General objective ............................................................................................. 5  
     1.3.2. Specific objectives ........................................................................................... 6  
  1.4. RESEARCH HYPOTHESIS ...................................................................................... 6  
  1.5. JUSTIFICATION OF THE STUDY ......................................................................... 6  
  1.6. LIMITATIONS ...................................................................................................... 7
CHAPTER TWO

LITERATURE REVIEW .................................................................................................................. 8

2.0. INTRODUCTION .................................................................................................................. 8

2.1. DIABETES ............................................................................................................................... 8

2.1.1. Definition .......................................................................................................................... 8

2.1.2. Types .................................................................................................................................. 8

2.1.3. Clinical diagnosis ............................................................................................................. 10

2.1.4. Signs and symptoms ......................................................................................................... 11

2.1.5. Complications ................................................................................................................. 11

2.1.6. Risk factors ...................................................................................................................... 12

2.1.7. Conceptual frame work .................................................................................................. 23

CHAPTER THREE

METHODOLOGY .......................................................................................................................... 24

3.0. INTRODUCTION .................................................................................................................. 24

3.1. CATEGORICAL RESPONSE DATA ..................................................................................... 24

3.2. BINOMIAL DISTRIBUTION ............................................................................................... 24

3.3. CONTINGENCY TABLE ......................................................................................................... 25

3.3.1. Definitions ....................................................................................................................... 25

3.3.2. Sensitivity and Specificity .............................................................................................. 27

3.3.3. Odds ratio ....................................................................................................................... 27
3.4. GENERALISED LINEAR MODELS (GLMs)................................................................. 29

3.5. LOGISTIC REGRESSION MODEL............................................................................. 30

3.5.1. Introduction ....................................................................................................... 30

3.5.2. Binary Logistic Regression with single independent variable ......................... 30

3.5.3. Interpretation of Regression Coefficients......................................................... 31

3.5.4. The logistic curve ............................................................................................. 32

3.5.5. Assumptions of logistic regression................................................................. 33

3.5.6. Multiple Logistic Regression............................................................................ 34

3.5.7. Evaluation of a logistic regression model....................................................... 38

3.6. RESEARCH DESIGN ............................................................................................. 41

3.7. STUDY POPULATION .......................................................................................... 42

3.8. SAMPLE SIZE AND SAMPLING PROCEDURE.................................................... 42

3.9. DATA COLLECTION AND DATA ANALYSIS .................................................... 45

CHAPTER FOUR

RESEARCH RESULTS AND DISCUSSION.................................................................. 46

4.0: INTRODUCTION.................................................................................................... 46

4.1. CHECKING FOR MULTICOLLINEARITY BETWEEN THE INDEPENDENT
VARIABLES ................................................................................................................ 48

4.2. CHECKING FOR NON-LINEARITY BETWEEN THE DEPENDENT VARIABLE
AND INDEPENDENT VARIABLES AND FOR NON-NORMALITY OF ERRORS.......... 51
4.3. TESTING FOR NORMALITY ................................................................. 52

4.3. CHI-SQUARE TEST OF ASSOCIATION BETWEEN THE DEPENDENT AND INDEPENDENT VARIABLES................................................................. 53

4.4. MULTIPLE LOGISTIC REGRESSION MODEL FITTING. ...................... 57

  4.4.1. The fitted model with all the predictor covariates ................................ 57

  4.4.2. Testing for the significance of the individual parameters in the model. ................. 59

  4.4.3. Signs of coefficients analysis. .......................................................... 59

  4.4.4. The odds ratio results................................................................. 60

  4.4.5. The full model assessment......................................................... 61

  4.4.6. The model with significant parameters only ....................................... 64

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS.......................... 67

5.0. INTRODUCTION ............................................................................ 67

5.1. SUMMARY ON FINDINGS OF THE STUDY....................................... 66

5.2. CONCLUSION ............................................................................. 68

5.3. RECOMMENDATIONS ................................................................. 69

REFERENCES.................................................................................... 71

APPENDICES .................................................................................. App-1
LIST OF TABLES

Table 1.1: Number of people with diabetes by IDF regions, 2013 and projection in 2035 ....... 3
Table 3.1: Contingency table ........................................................................................................26
Table 3.2: Contingency table of joint and marginal probabilities .............................................28
Table 3.3: Number of patients by sector .....................................................................................43
Table 3.4: Calculation of sample size by Sector ...........................................................................43
Table 3.5: Number of people in the sample by each variable ......................................................44

Table 4.1: Checking for multicollinearity by Variance Inflation Factor (VIF) ....................... 49
Table 4.2: Correlations between the variables in the model .......................................................51
Table 4.3: Test of independence of age versus diabetes ...............................................................53
Table 4.4: Test of independence of Gender versus diabetes .........................................................54
Table 4.5: Test of independence of occupation status versus diabetes ......................................54
Table 4.6: Test of independence of smoking versus diabetes .......................................................55
Table 4.7: Test of independence of alcohol consumption versus diabetes ...............................55
Table 4.8: Test of independence of cholesterol level versus diabetes .........................................56
Table 4.9: Test of independence of hypertension versus diabetes ..............................................56
Table 4.10: Test of independence of family history of diabetes versus diabetes .....................57
Table 4.11: The Estimated coefficients, their S.E, and Wald test for the full model ............... 58
Table 4.12: The sign analysis ........................................................................................................60
Table 4.13: Odds Ratios and 95% Confidence Intervals for Covariates .....................................60
Table 4.14: Likelihood ratio test ..................................................................................................61
Table 4.15: Classification table for the model with all predictor variables ............................... 62
Table 4.16: Hosmer and Lemeshow Test ....................................................................................63
Table 4.17: Summarized results for the reduced model ............................................................. 64
Table 4.18: Classification table for the reduced model ............................................................. 65
LIST OF FIGURES

Figure 2. 1: Blood glucose regulation by insulin in healthy people and in people with type 1 or type 2 diabetes.......................................................................................................................... 9

Figure 3. 1: Logistic regression curve for $\beta_0 = 0, \beta_1 = 1$.......................................................... 33

Figure 4. 1: The state of the sample by each variable........................................................................ 47
Figure 4. 2: Diabetes state by each variable ..................................................................................... 47
Figure 4. 3: Normal Probability plots of errors .................................................................................. 53
Figure 4. 4: Receiver Operating Characteristic (ROC) curve for the full model......................... 63
Figure 4. 5: Receiver Operating Characteristic (ROC) curve for the reduced model ............ 66


LIST OF APPENDICES

Appendex 1: To whom it may concern letter ................................................................. App-1

Appendex 2: Data accessing request letter written to Gitwe hospital .......................... App-2

Appendex 3: Permission to access to data letter ............................................................. App-3
ABBREVIATIONS

BMI:  Body Mass Index

GDM:  Gestational Diabetes Mellitus

HDL:  High – Density Lipoprotein

IDF:  International Diabetes Federation

LDL:  Low – Density Lipoprotein

NCDs: Non-communicable diseases

ROC:  Receiver – Operator Characteristic

UN:   United Nations

VIF:  Variance Inflation Factor

WHO:  World Health Organization
ABSTRACT

The number of people with diabetes is increasing due to population growth, aging, urbanization. Increasing in prevalence of obesity and physical inactivity added to a misconception that diabetes is a disease for urban areas while rural areas are also concerned, this is the motivation of the study. In this study, a multiple logistic model was used to fit factors like age, gender, occupation status, smoking, alcohol consumption, Cholesterol level, hypertension and family history of diabetes as risk factors of diabetes. A three year period, that is from 2011 to 2013 data of Gitwe Hospital were used. The software package that has been used to process data is SPSS 15.0.

The test of independence between the dependent variable (diabetes), with the independent variables (age, gender, smoking, occupation status, alcohol consumption, cholesterol level, hypertension and family history of diabetes) was performed to verify whether they are statistically significant or not at 5% of level of significance to the outcome of diabetes. It is found that age, alcohol consumption, cholesterol level, occupation status and hypertension were associated with the outcome of having diabetes. The predictors like gender, smoking, family history of diabetes had negligible association with having diabetes. A multiple logistic regression model containing all the predictor variables was fitted and a test of significance on coefficients was performed. The Wald test reveals that, on one hand, the significant predictors are: Age, Occupation status, Alcohol consumption, Cholesterol level and Hypertension are statistically significant. On the other hand, the predictors which are not statistically significant are: Gender, smoking and family history of diabetes. Also, this study showed that older age, smoking, alcohol consumption, high cholesterol level and hypertension increases the risk of having diabetes, while being employed decreases the risk of having the illness.
From the odds ratio results, older age persons, patients who consume alcohol, patients with high cholesterol level and hypertensive persons are highly susceptible for diabetes occurrence. Finally, a multiple logistic regression with only significant parameters was fitted to compare it with that one with all predictor variables. Based on their respective Receiver Operator Characteristic (ROC) curve and their overall explanatory strength the conclusion is that the reduced model (which has the area under the ROC curve of 0.843 and its overall explanatory strength is 79.4% ) fits better the data than the model with all predictor variables (which has the area under the ROC curve of 0.825 and its overall explanatory strength is 77.7%). Finally, the recommendation about the use of the fitted model, in order to know the likelihood of getting diabetes was given. Other recommendations about helping to reduce the risk of getting diabetes were also given.
CHAPTER ONE

INTRODUCTION

1.1. BACKGROUND OF THE STUDY

The World Health Organization (WHO, 2011) defines diabetes as a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Diabetes is on the rise all over the world and countries are struggling to fight the disease. The misconception that diabetes is “a disease of the wealthy” is still held by some people; but the evidence published in the Diabetes Atlas of the International Diabetes Federation (IDF, 2013) disproves that delusion: a staggering 80% of people with diabetes live in low and middle-income countries, and the socially disadvantaged in any country are the most vulnerable to that illness.

Today’s emerging diabetes hotspots include countries in the Middle East, Western Pacific, sub-Saharan Africa and South-East Asia where economic development has transformed lifestyles. These rapid transitions are bringing previously unheard rates of obesity and diabetes; developing countries are facing a firestorm of ill health with inadequate resources to protect their population. Thus, it is necessary to increase awareness of the importance of a healthful diet and physical activity, especially for children and adolescents. Crucially though, environments have to be created that lay the foundations for healthy living. (IDF, 2013)

In the last two years, progress has been made toward driving political change for diabetes. Building on the momentum of the 2011 UN Political Declaration on non-communicable diseases (NCDs), the 66th World Health Assembly in May 2013 saw the unanimous adoption by Member States of a voluntary Global Action Plan for the prevention and control of NCDs. Diabetes is
now prominent on the global health agenda, with specific targets for access to essential medicines and for halting the growth of obesity and diabetes. (WHO, 2011)

According to Preis et al. (2009), population based studies of cardiovascular diseases (CVD) risk factor trends among subjects with and without diabetes show differing trend in disfavor of those with diabetes. Studies of adherence to guidelines for CVD prevention targets in patients with diabetes in general practice have shown that only 13% reach all the targets. Previous studies have found appropriate lifestyle intervention and/or drug treatment are effective in delaying or preventing both diabetes and its complications. (Jenssen, 2008)

1.2. STATEMENT OF THE PROBLEM

According to Sarah et Al. (2004), the number of people with diabetes is increasing due to population growth, aging, urbanization, and increasing prevalence of obesity and physical inactivity. Quantifying the prevalence of diabetes and the number of people affected by diabetes, now and in the future, is important to allow rational planning and allocation of resources.

According to Shaw et al. (2010), the world prevalence of diabetes in 2010 among adults aged 20-79 years was estimated to 6.4%, affecting 285 millions of adults. Between 2010 and 2030, there is an expected 70% increase in number of adults with diabetes in developing countries and a 20% increase in developed countries.

Each year more than 231,000 people in the United states and more than 3,96 million people worldwide die from diabetes and its complications and this number is expected to increase by more than 50 percent over next decade. Estimated global healthcare expenditures to treat and prevent diabetes and its complications is at least 376 billion US Dollar (USD) in 2010. By 2030, this number is projected to exceed some 490 billion USD. (IDF,2009)
Environmental and lifestyle factors are the main causes of the dramatic increase in type 2 diabetes prevalence. Genetic factors probably identify those most vulnerable to these changes.

Diabetes is now truly a pandemic, and its effects are particularly severe in low and middle income countries. (IDF, 2010)

The following table shows the situation of diabetes over the world in 2013 and the projected percentage of increase in people with diabetes in 2035:

Table 1. 1: Number of people with diabetes by IDF regions, 2013 and projection in 2035

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of people with diabetes Year 2013 (in millions)</th>
<th>Predicted percentage of increase in 2035</th>
</tr>
</thead>
<tbody>
<tr>
<td>NORTH AMERICA AND CARIBBEAN</td>
<td>37</td>
<td>37.3 %</td>
</tr>
<tr>
<td>SOUTH AND CENTRAL AMERICA</td>
<td>24</td>
<td>59.8 %</td>
</tr>
<tr>
<td>EUROPE</td>
<td>56</td>
<td>22.4%</td>
</tr>
<tr>
<td>MIDLE EAST AND NORTH AFRICA</td>
<td>35</td>
<td>96.2%</td>
</tr>
<tr>
<td>SUB-SAHRAN AFRICA</td>
<td>20</td>
<td>109.1%</td>
</tr>
<tr>
<td>SOUTH EAST ASIA</td>
<td>72</td>
<td>70.6%</td>
</tr>
<tr>
<td>WESTERN PACIFIC</td>
<td>138</td>
<td>46%</td>
</tr>
</tbody>
</table>

Source: IDF Atlas 2013, p.11

Here are some basic facts about diabetes worldwide, according to IDF (2013):

1. Each year the number of people with diabetes increases by 7 millions in the world.
2. By 2035, about 592 million people will have diabetes, a number which was 382 million in 2013.
4. During 2013, diabetes killed about 5.1 million adults worldwide.

5. Diabetes leads to complications and severe disabilities, including kidney disease, blindness, heart attack, stroke and neural damage leading to amputation and the need for chronic care.

6. The trend in 2013 revealed that there are three new cases every 10 seconds.

7. More than 80% of spending on medical care for diabetes is in the world’s richest countries, even though 80% of the people with diabetes live in low and middle income countries, where 76% of the burden lies.

8. The burden of illness caused by diabetes and the reduction in life expectancy in sub-Saharan Africa will hinder the region’s economic growth.

9. Diabetes caused at least 548 billion USD in health expenditure in 2013 (it means 11% of the total health spending on adults and this amount is predicted to be 627 USD in 2035.

10. More than 21 million live births were affected by diabetes during pregnancy in 2013

Concerning Sub-Saharan Africa, Mbanya (2009) says: “soon, four out of every five people with diabetes will live in developing countries. And the men and women most affected are of working age – the breadwinners of their families.”

According to Ayesha & Kaushik (2010), a child diagnosed with type 1 diabetes in sub-Saharan Africa has a life expectancy that varies between 7 months and 7 years, depending on the country, compared to 60 years in Western Europe; and over the next 20 years, it is predicted that sub-Saharan Africa will have the highest growth in the number of people with diabetes of any region in the world.
According to the same author, diabetes was once considered as a rare disease in sub-Saharan Africa. But in that part of the word, in 2010; 12.1 million adults were estimated to have diabetes and by 2030, it is estimated that 23.9 million adults in sub-Saharan Africa will have diabetes.

Data of 2010 on the condition of people with diabetes in sub-Saharan Africa and the complications of diabetes that they suffer is very scarce. However, it was estimated that at least:

1. 4.51 million people had eye complications.
2. 2.23 million people needed dialysis because of kidney damage.
3. 907,500 people had cardiovascular disease.
4. 423,500 people were blind because of diabetes.
5. 399,300 people had cerebrovascular disease.
6. 169,400 people had lost a foot because of amputation. (Ayesha et al., 2010)

Concerning Rwanda, the number of deaths due to diabetes in 2013 was estimated to be 5464.

In that same year, the prevalence in adults (20-79 years) was 4.38% and the total number of people living with diabetes was estimated to be 234000. (IDF, 2013)

Concerning the case study of Gitwe Hospital, number of people with diabetes during these five years has tripled, the number which was 43 in 2010 to become 143 in 2014.

All those facts on the disease explain the particular attention to be reserved to the disease.

1.3. RESEARCH OBJECTIVES

1.3.1. General objective

This study had the following general objective:

To provide the contribution to reducing the risk of having diabetes.
1.3.2. Specific objectives

The study had also the following specific objectives:

1. To test for the association between the risk factors (age, gender, smoking, occupation status, alcohol consumption, Cholesterol level, hypertension, and family history of diabetes) and diabetes.

2. To fit a multiple logistic model on the incidence of diabetes given the risk factors (age, gender, smoking, occupation status, alcohol consumption, Cholesterol level, hypertension, and family history of diabetes)

1.4. RESEARCH HYPOTHESIS

The following hypotheses were formulated in order to achieve the above objectives:

1. $H_0$: There is no association between having diabetes and risk factors like age, gender, smoking, occupation status, alcohol consumption, Cholesterol level, hypertension, and family history of diabetes $\text{VS } H_1$: There is an association between having diabetes and those risk factors. To test those hypotheses, the chi-square test of independence was used.

2. $H_0: \beta_i = 0$ (it means the coefficient $\beta_i$ in the fitted multiple logistic regression is not statistically significant) $\text{VS } H_1$: At least one of the coefficients is different from zero.

To test those hypotheses, the Wald test was used.

1.5. JUSTIFICATION OF THE STUDY

In Rwanda, many studies about communicable diseases are carried out and only few attempts are made for the study of non-communicable diseases. Older diabetics with cardiovascular and
micro vascular complications and disease unrelated to diabetes burden the hospital than the non-diabetic population.

This study may be proved to be important because:

- The results or outcome would give information as to whether factors like age, gender, smoking, occupation status, alcohol consumption, cholesterol level, hypertension and family history are really the risk factors in being diagnosed of diabetes.

- The findings will help individuals to know which of the factors stated above contribute 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 do things that will prevent the blood pressure from rising up.

The findings will help people without the disease to know their estimated risk 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:

Insufficiency in finance means and time has prevented us from visiting other hospitals. The research should also use the primary data from patients, but due to the above mentioned challenges, it was impossible to reach each patient. Some patients’ records should contain incomplete information. However, we rely on scientific methods to obtain the data and the analysis is based on superior analytical techniques, which we believe allow us to generalize our findings.
CHAPTER TWO

LITERATURE REVIEW

2.0. INTRODUCTION

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

2.1. DIABETES

2.1.1. Definition

According to FALL (2013), diabetes is a chronic disease that occurs when the body cannot produce enough insulin or cannot use insulin effectively. Insulin is a hormone produced in the pancreas that allows glucose from food to enter the body’s cells where it is converted into energy needed by muscles and tissues to function. According to the same author, person with diabetes does not absorb glucose properly, and glucose remains circulating in the blood (a condition known as hyperglycaemia) damaging body tissues over time. This damage can lead to disabling and life-threatening health complications.

2.1.2. Types

The World Health Organization (WHO, 2011) gives three main types of diabetes:

2.1.2. 1.Type 1 diabetes

Type 1 diabetes is caused by an autoimmune reaction, where the body’s defense system attacks the insulin-producing beta cells in the pancreas. As a result, the body can no longer produce the insulin it needs. Why this occurs is not fully understood. The disease can affect people of any age, but usually occurs in children or young adults. People with this form of diabetes need
insulin every day in order to control the levels of glucose in their blood. Without insulin, a person with type 1 diabetes will die.

2.1.2.2. Type 2 diabetes

Type 2 diabetes is the most common type of diabetes. It usually occurs in adults, but is increasingly seen in children and adolescents. In type 2 diabetes, the body is able to produce insulin but either this is not sufficient or the body is unable to respond to its effects (also known as insulin resistance), leading to a build-up of glucose in the blood.

Many people with type 2 diabetes remain unaware of their illness for a long time because symptoms may take years to appear or be recognized, during which time the body is being damaged by excess blood glucose. They are often diagnosed only when complications of diabetes have already developed.

The following figure explains type 1 and type 2 diabetes:

![Figure 2.1: Blood glucose regulation by insulin in healthy people and in people with type 1 or type 2 diabetes. (adapted from Nicholas et al., 1998)](image-url)
2.1.2. 3. Gestational diabetes

Women who develop a resistance to insulin and subsequent high blood glucose during pregnancy are said to have gestational diabetes (also referred to as gestational diabetes mellitus or GDM). Gestational diabetes tends to occur around the 24th week of pregnancy. The condition arises because the action of insulin is blocked, probably by hormones produced by the placenta.

As gestational diabetes normally develops later in pregnancy, the unborn baby is already well-formed but still growing. The immediate risk to the baby is therefore not as severe as for those whose mother had type1 diabetes or type2 diabetes before pregnancy (a condition known as diabetes in pregnancy). Nonetheless, uncontrolled gestational diabetes can have serious consequences for both the mother and her baby.

2.1.3. Clinical diagnosis

The following tests are used for clinical diagnosis of diabetes, according to the WHO (2011):

1. A **fasting plasma glucose test** measures blood glucose in a person who has not eaten anything for at least 8 hours. This test is used to detect diabetes and pre-diabetes.

2. An **oral glucose tolerance test** measures blood glucose after a person fasts at least 8 hours and 2 hours after the person drinks a glucose-containing beverage. This test can be used to diagnose diabetes and pre-diabetes.

3. A **random plasma glucose test**, also called a casual plasma glucose test, measures blood glucose without regard to when the person being tested last ate. This test, along with an assessment of symptoms, is used to diagnose diabetes but not pre-diabetes.
Test results indicating that a person has diabetes should be confirmed with a second test on a different day.

2.1.4. Signs and symptoms

Type 1 diabetes often develops suddenly and can produce symptoms such as: abnormal thirst and a dry mouth, frequent urination, lack of energy, extreme tiredness, constant hunger, sudden weight loss, slow-healing wounds, recurrent infections and blurred vision.

A number of skin rashes can occur in diabetes that is collectively known as diabetic dermadromes. (IDF,2013)

2.1.5. Complications

According to Ayesha & Kaushik (2010), here are some of the complications of diabetes:

2.1.5.1. Emotional distress

Diabetes is often associated with depression. As diabetes is a chronic illness, it requires continuous treatment and often leads to dependence on others. This can create many problems, including the breakdown of relationships and family abandonment.

2.1.5.2. Kidney disease

People with diabetes make up one third of patients admitted to dialysis units, and end-stage renal failure is the leading cause of hospital mortality in African diabetic patients.

2.1.5.3. Stroke

Stroke is up to four times more likely than in people without diabetes.

The mortality associated with stroke for type 2 diabetic patients in Africa is 3-6 times as high as in England. For almost 5% of diabetes patients in Tanzania, stroke is the first symptom for which they seek help.
2.1.5. 4. Blindness

Diabetes is a leading cause of blindness. In Africa, retinopathy affects up to 55% of diagnosed diabetes patients and 21-25% are already affected at the time of diagnosis.

2.1.5. 5. Heart attack

The risk of heart attack is three times greater for people with diabetes, and heart disease affects 20% of diabetic patients in Africa. In Africa, up to 50% of diabetics are affected by cardiomyopathy.

2.1.5. 6. Amputation

Diabetes is the leading cause of non-traumatic lower limb amputations. In Africa, about 12% of all diabetic patients have foot ulcers, and amputation occurs in up to 7% of all hospitalized diabetic patients.

2.1.6. Risk factors

2.1.6.1. Lifestyle associated risk factors

a. High Cholesterol level

Among the risk factors for type 2 diabetes, high cholesterol or increase in fat in the body may take the first place. Cholesterol is a fatty substance produced by the liver and carried by the blood to the rest of the body. The raised blood cholesterol increases the risk of having diabetes. Greater weight means a higher risk of insulin resistance, because fat interferes with the body's ability to use insulin. Different studies have shown that the number of overweight kids has tripled since 1980 and so children who are being diagnosed with type 2 diabetes has risen. (WHO, 2011)

The National Institute for Health (1999), also found in their study that the prevalence of type 2 diabetes has tripled in the last 30 years and much of the increase is related with increase in
cholesterol levels. People with Body mass index of 30kg or more have fivefold greater risk of diabetes than those with < 25kg BMI. A study done in Poland was concluded with the following results: The prevalence of diabetes or impaired glucose tolerance was found in 5.3% and 92.8% of subjects having diabetes or impaired glucose tolerance were either obese or have high cholesterol level and 32.4% had hypertension.

Also, according to the research of Mohan et al. (2004) increase in cholesterol level has become the emerging burden of risk factors for non-communicable disease, blood pressure, and is 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 activities.

The last but not the least findings found in Africa revealed the effect of high cholesterol on blood pressure is higher in males than in females (regression coefficient 0.64 and 0.38 respectively). Mafunda et al.(2006). The last recent study in Ghana has shown that adjusted odds ratio for developing hypertension for overweight or high cholesterol were 5.8 and 8.9 respectively (Addo et al.,2007). The findings from authors such as Olinto et al. (2003) in Brazil and Stella (2012) in Ghana arrived to conclude that high cholesterol level is associated with diabetes outcome.

Also, Okosun et al. (1998) correlated high cholesterol level and risk of diabetes in elderly population and also added the possibility of substantial reduction in diabetes is achievable by reduction in cholesterol levels through our way of eating and exercising.

b. Hypertension

According to Marvin (2005), hypertension is a medical term for elevated or high blood pressure. An elevated blood pressure increases the risk strokes, heart failure, kidney failure and heart
attacks and blood pressure plays an important part in the management of diabetes. Many people with high blood pressure also have diabetes. The same author states that blood pressure is the force exerted by the blood stream against the walls of arteries that carry blood to various parts of the body. Blood from larger arteries flows into smaller arteries or arterioles to various parts of the body. These can open wide (dilate) or close (constrict). If they are open, blood pressure remains low; if they are constrict, it gets high. Blood pressure is recorded in two numbers as millimeters of mercury (mm Hg). According to the same author, blood pressure levels for adults may be categorized as follows:

- **Hypertension:** $\frac{140 \text{ mm Hg and over Systolic (pumping pressure)}}{90 \text{ mm Hg and over diastolic (resting pressure)}}$
- **High Normal or Prehypertension:** $\frac{120-140 \text{ mm Hg and over Systolic (pumping pressure)}}{80-90 \text{ mm Hg and over diastolic (resting pressure)}}$
- **Ideal or Normal:** $\frac{110-120 \text{ mm Hg and over Systolic (pumping pressure)}}{75-80 \text{ mm Hg and over diastolic (resting pressure)}}$

According to Stella (2012), global assumptions are difficult due to heterogeneity between countries. According to a systematic review of studies reporting data from 1980 and 2004, the overall worldwide prevalence of high blood pressure was approximately 26% in adult population.

In the USA, the prevalence of high cholesterol has increased from 50 million in 1990 to 65 million in 2000. Reported differences by gender and race are small. The increasing prevalence is primarily a consequence of trends for the population to become older and more obese of increasing survival of hypertensive patients as a result of improved lifestyles or more effective drug therapy. Also, the same author has given the results from the WHO MONICA Project (2004) from developed countries. Data from national surveys in six European countries, performed in the 1990s, using similar sampling and reporting techniques, estimated the
prevalence of high blood pressure/hypertension as 38% in Italy, 38% in Sweden, 42% in England, 47% in Spain, 49% in Finland and 55% in Germany. In Portugal, data suggest that 3,311,830 people have high blood pressure/hypertension (42.1%). As a result of progressive urbanization and westernization of their lifestyle, developing countries are now undergoing an epidemiological transition. These changes are leading to a new epidemiological situation with a decline in infectious diseases and emergence of diabetes and cardiovascular diseases.

However, the reported hypertension prevalence was 27.2 in India, 40.6% in Syria 23.9% among men and 13.7% among women in Vietnam and 27.1% among men and 30.2% among women in Tanzania. These values are lower compared with high blood pressure/hypertension prevalence in developed world, but the global tendency is for these values to increase. Differences in hypertension prevalence are not only present between countries, but also between racial or ethnic groups. The prevalence among U.S. Blacks is higher than in Whites and Mexican-Americans in both genders and all ages.

c. Smoking

There is evidence to show that smoking is a risk factor for Type 2 diabetes.

The studies of Hsin-Chieh et al. (2010), Ko & Cockram (2005), Rimm E, Chan J et al. (1995) have explained this link. Smoking has been identified as a possible risk factor for insulin resistance, which leads to diabetes. Smoking has also been shown to deteriorate glucose metabolism which may lead to the onset of type 2 diabetes. In his study, Willi et al. (2007) found that there is an association between active smoking and an increased risk of diabetes. On the basis of this study, it was estimated that 12% of all type 2 diabetes in the United States may be attributable to smoking. According to Montgomery (2005), women who smoke during pregnancy are at increased risk of developing gestational diabetes and also increase the risk of their
offspring developing diabetes later in life. Also, Haire-Joshu & Thomas (2005) stated that
Compared to non-smokers with diabetes, people with diabetes who smoke have twice the risk of
premature death. Furthermore, the risk of complications associated with tobacco use and diabetes
in combination is nearly 14 times higher than the risk of either smoking or diabetes alone.
Accordingly, Al-Delaimy et al. (2001) in a large prospective study of United States (US) nurses
stipulated that among those with diabetes the relative risks of mortality were 1.31 for past
smokers, 1.43 for current smokers of 1-14 cigarettes per day, 1.64 for smokers of 15-34
cigarettes per day, and 2.19 for current smokers of 35 or more cigarettes per day.
According to Mulhauser et al. (1996), smoking is associated with multiple complications of
diabetes. Nephropathy (kidney disease) has been shown to be common in type 1 diabetic patients
who smoke and smoking increases the risk of albuminuria in both types of diabetes. Here,
Albuminuria refers to the presence of protein in the urine and can indicate signs of kidney
disease.
The information given by Action on Smoking and Health (2012) is that in the British prospective
study of 7,735 men aged 40-59 years found that the benefit of giving up smoking was only
apparent after 5 years of smoking cessation and risk reverted to that of never-smokers only after
20 years. Men who gave up smoking during the first 5 years of follow up showed significant
weight gain and subsequently higher risk of diabetes than continuing smokers. However, the
authors concluded that in the long term, the benefits of giving up smoking outweigh the adverse
effects of early weight gain. Also, the US Cancer Prevention Study provided evidence that
stopping smoking for 10 years in men and five years in women could reduce the risk of diabetes
to that of nonsmokers.
d. Alcohol

Alcohol is used, among diabetics as well as nondiabetics. Then clinicians and researchers have tried to understand alcohol’s effect on the progression and complications of diabetes. A study made by Swade & Emanuele (1997) showed that in people with either type 1 or type 2 diabetes, single episodes of alcohol consumption (i.e., acute alcohol consumption) generally do not lead to clinically significant changes in blood sugar levels. In fact, some studies have indicated that isolated episodes of drinking with a meal may have a beneficial effect by slightly lowering blood sugar levels that tend to rise too high in diabetics and this potentially beneficial effect was observed in both men and women, regardless of age.

In their studies Arky and Freinkel (1964) have shown that the effect of alcohol may be observed in both type 1 and type 2 diabetics as well as in nondiabetics. According to the American Diabetes Association (2009), alcohol consumption is an influencing factor of diabetes. The biological mechanism is uncertain, but there are several factors that may explain the relationship, including increases in insulin sensitivity after moderate alcohol consumption, changes in levels of alcohol metabolites, increases in HDL cholesterol concentrations, or via the anti-inflammatory effect of alcohol. In a meta-analysis of epidemiological studies representing 13 cohorts, Carlsson et al. (2005) reported a protective effect of moderate alcohol consumption of about 32% (relative risk = 0.72). Both men and women demonstrate the same relationship, but the latter have a higher risk for diabetes at lower levels of average alcohol consumption. The risk associated with high consumption was described as being unclear. In the other meta-analysis, in which alcohol consumption was treated continuously, a U-shaped relationship was found for both men and women, with a more protective effect of moderate consumption observed for women. However, in both of these reviews, the reference group was composed of former drinkers and lifetime
abstainers. Because former drinkers may be inspired to abstain due to health concerns, they may actually be at increased risk of developing diabetes, known as the sick-quitter effect. Also in his study, Chao (2001) arrived at the fact that there is a significant positive correlation between blood pressure and all levels alcohol consumption.

Many studies such as one of Motala et al. (2008) in rural communities of South Africa, Burchfeil et al. (1995), Ajani et al. (2000) and Lapidus et al. (2005) have observed that subjects consuming alcohol were at higher risk of getting diabetes.

2.1.6.2. Non changeable risk factors

a. Age

The IDF (2013) stipulated that type 2 diabetes has long been regarded as a condition that affects older people. However, this perception requires modification as an increasing number of younger people (aged 20–40 years) are being diagnosed with the disease. The 2013 International Diabetes Federation report on the global prevalence of diabetes estimated that 382 million people had diabetes, with the greatest number aged between 40 years and 59 years. According to the same author, the incidence of type 1 diabetes among children is increasing in many countries, particularly in children under the age of 15 years. There are strong indications of geographic differences in trends but the overall annual increase is estimated to be around 3%. Evidence shows that incidence is increasing more steeply in some Central and Eastern European countries, where the disease is less common. Also, several European studies have suggested that, in relative terms, increases are greatest among younger children. There is also evidence that similar trends exist in many other parts of the world, but in sub-Saharan Africa incidence data are limited or non-existent. Special efforts must be made to collect
more data, especially in those countries where diagnoses may be missed. Some 79,100 children under 15 years are estimated to develop type 1 diabetes annually worldwide. Of the estimated 497,100 children living with type 1 diabetes, 26% live in the Europe Region, where the most reliable and up-to-date estimates of incidence are available, and 22% in the North America and Caribbean Region.

On the other hand, IDF in 2013 estimated the global prevalence of diabetes in people aged between 60 and 79 to be 18.6%, more than 134.6 million people, accounting for over 35% of all cases of diabetes in adults. By 2035, that number is projected to increase beyond 252.8 million.

Older people with diabetes have an increased rate of diabetes-related complications and they are at increased risk of some form of functional impairment resulting from those diabetes complications. In his study on subjects of at least 20 years old Kim et al. (2006), concluded that the prevalence of diabetes was 7.6%, and the prevalence of diabetes increased with age and peaked in the oldest age group in Africa. Usually, increased age is directly associated with diabetes since the functions of the organs in the body reduces as one grows, so there is development of all kinds of diseases especially diabetes and other cardiovascular diseases.

As a person older gets, as his risk of type 2 diabetes becomes greater. Even if an elderly person is thin, they still may be predisposed to getting diabetes. Scientists theorize that the pancreas ages right along with a person, and doesn't pump insulin as efficiently as it did when he was younger. Again, the increase in prevalence of the disease has accelerated still due to aging population structures in developed countries and increasing obesity globally. Muni (2008), also found that the prevalence of diabetes increases with advancing age to the point where more than half of people 60-69 years of age and approximately three-fourths of those 70 years of age and older are affected. Finally, I believe from the studies above that the biggest change will occur in the
developing world and developing countries if investment is made in both education on diabetes and on the elderly.

b. Gender

Many studies have concluded that there was no consistent difference in the percentage of men and women reported with diabetes independent from other characteristics.

In the study of Babwah, Baksh et al. (2006), an interesting finding was that the percentage of patients of East Indian descent was exactly the same for both genders: 62 (66.7%) men and 178 (66.7%) women ($p = 0.891$). Again, the same authors say findings from the United States suggest a disparity of approximately 10%–15% in favor of women, although an earlier study in Trinidad and Tobago reported no significant gender differences in the incidence of type 2 diabetes mellitus.

In Greece the prevalence was higher in men (7.6%) than in women (5.9%), but taken together, these findings suggest that there is no marked gender difference in the incidence and prevalence of type 2 diabetes mellitus. Additionally, in accordance with the study of Gordon and Dwight (1985), there was no consistent difference in percentage of men and women with reported diabetes, independent of other characteristics. However, the same study showed that there were interactions of sex and body mass, age or race. The prevalence of diabetes increases more rapidly with increased body mass index (BMI) for women than it does for men. Among white people, women in the most obese group were more likely than men to have reported diabetes, but women in the least obese groups were less likely than men to have reported diabetes. A similar relationship was not observed for black men and women.

Also, in the Nurses' Health Study, women who engaged in vigorous exercise at least once per week had a relative risk of type 2 diabetes mellitus of 0.8 ($p = 0.005$) compared with women who
did not exercise weekly. In contrast to the Augsburg study, the Physicians' Health Study documented that male physicians who exercised at least once per week had a relative risk of type 2 diabetes of 0.71 ($p = 0.006$) compared with those who exercised less frequently; the relative risk of diabetes decreased with increasing frequency of exercise. The physical activity measure used in the present study categorized individuals on the basis of their regular participation in leisure time activity. In general, men from a population-based study engage in more strenuous physical activity at work in comparison with women; thus, men are altogether more physically active, if both occupational and leisure time physical activity is taken into consideration. Therefore, the sex-related dissimilarities between physical activity and type 2 diabetes in the Augsburg study could be due to the fact that men who are classified as inactive are in fact quite active at work.

**c. Family History and Genetics**

Many studies show a strong family history among affected youth with 45-80% having at least one parent with diabetes and 74-100% having a first or second degree relative with type 2 diabetes. Children with diabetes are also more likely to have a family history of cardiovascular disease (CVD). Berenson et al.(1995) have shown that children of individuals with type 2 diabetes were more likely to be obese and have higher blood pressures (IDF 2003). Also, Hirschhorn et al. (2003) have stated that some genes contribute approximately 40-50% of the heritable risk for Type 1 diabetes. They said that people who have family members who have been diagnosed with type1 diabetes are at a greater risk for developing it themselves. African Americans, Hispanic-Americans and Native Americans all have a higher than normal rate of type 1 diabetes. Having a genetic disposition towards type 2 is not a guarantee of a diagnosis however. Lifestyle plays an important part in determining who gets diabetes.
According to Flores et al. (2003), it has long been known that type 2 diabetes is, in part, inherited. Family studies have revealed that first degree relatives of individuals with type 2 diabetes are about 3 times more likely to develop the disease than individuals without a positive family history of the disease. Those with family history of diabetes mellitus had 3.6 times odds of developing diabetes compared to those without. (Majgi et al., 2012)

d. Socioeconomic status

According to IDF (2013), characteristics of the social environment, including the socioeconomic status of the individuals (income and employment), social isolation or exclusion and access to local services and facilities (especially transport), are linked closely with population health. In particular, people with lower incomes, lower levels of education and lower status occupations have worse health than those with higher incomes, education levels or occupations.

In their study, Majgi et al. (2012) arrived to conclude that as the skill level of occupation increased, the prevalence of diabetes also increased.
2.1.7. Conceptual frame work

The following diagram illustrates the independent and dependent variables:

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Dependent variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Age</td>
<td></td>
</tr>
<tr>
<td>- Gender</td>
<td></td>
</tr>
<tr>
<td>- Smoking</td>
<td></td>
</tr>
<tr>
<td>- Occupation status</td>
<td></td>
</tr>
<tr>
<td>- Alcohol consumption</td>
<td></td>
</tr>
<tr>
<td>- Cholesterol level</td>
<td></td>
</tr>
<tr>
<td>- Hypertension</td>
<td></td>
</tr>
<tr>
<td>- Family History of diabetes</td>
<td></td>
</tr>
</tbody>
</table>

DIABETES

In this study, the risk factors (age, gender, smoking, occupation status, alcohol consumption, cholesterol level, hypertension and family history) are considered to have a contribution to the outcome of diabetes. The risk factors are the independent variables while the outcome of diabetes is the dependent variable.
CHAPTER THREE

METHODOLOGY

3.0. INTRODUCTION

This chapter is going to discuss about the methodology used, the study population, sample size and sampling procedure.

3.1. CATEGORICAL RESPONSE DATA

A categorical variable has a measurement scale consisting of a set of categories. For example, political philosophy may be measured as “liberal”, “moderate” or “conservative”; choice of accommodation might use categories “house” or “apartment”. Categorical scales also occur frequently in the health sciences, for measuring responses such as whether a patient survives an operation (yes, no), severity of an injury (none, mild, moderate, severe), and stage of a disease (initial, advanced). Categorical variables frequently occur in public health (e.g., categories “yes” and “no” for whether a given person has a illness), zoology (e.g., categories “fish”, “invertebrate” or “reptile”), education (e.g., categories “correct” and “incorrect” for students’ responses to an exam question), and marketing (e.g., categories “Brand A,” “Brand B,” and “Brand C” for consumers’ preference among three leading brands of a product). They even occur in highly quantitative fields such as engineering sciences and industrial quality control, when items are classified according to whether or not they conform to certain standards.

3.2. BINOMIAL DISTRIBUTION

Often, categorical data result from \( n \) independent and identical trials with two possible outcomes for each, referred to as “success” and “failure”. Identical trials mean that the probability of success is the same for each trial. Independent trials means the response outcomes are
independent random variables. In particular, the outcome of one trial does not affect the outcome of another. These are often called Bernoulli trials. Let $\pi$ denote the probability of success for a given trial. Let $Y$ denote the number of successes out of the $n$ trials. Under the assumption of $n$ independent, identical trials, $Y$ has the binomial distribution with index $n$ and parameter $\pi$ and it is denoted by $Y \sim B(n, \pi)$. The probability of outcome $y$ for the random variable $Y$ is given by:

$$P(y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \quad \text{for } y = 0, 1, 2, \ldots, n$$

$$= \frac{n!}{y!(n-y)!} \pi^y (1 - \pi)^{n-y} \quad \text{for } y = 0, 1, 2, \ldots, n \quad (3.1)$$

If $Y$ is random variable such that $Y \sim B(n, \pi)$, then $Y$ has mean and standard deviation:

$$E(Y) = \mu = n\pi \quad \text{and} \quad \sigma = \sqrt{n\pi(1 - \pi)} \quad (3.2)$$

### 3.3. CONTINGENCY TABLE

#### 3.3.1. Definitions

Suppose there are two categorical variables, denoted by $X$ and $Y$. Let $I$ denote the number of categories of $X$ and $J$ the number of categories of $Y$. A rectangular table having $I$ rows for the categories of $X$ and $J$ columns for the categories of $Y$ has cells that display the $IJ$ possible combinations of outcomes. A table of this form that displays counts of outcomes in the cells is called a contingency table. A table that cross classifies two variables is called a two-way contingency table. A two-way table with $I$ rows and $J$ columns is called an $I \times J$ (read $I$-by-$J$ table).
Table 3.1: Contingency table

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>...</th>
<th>j</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>n_{11}</td>
<td>n_{12}</td>
<td>n_{13}</td>
<td>n_{14}</td>
<td>...</td>
<td>n_{1j}</td>
</tr>
<tr>
<td>2</td>
<td>n_{21}</td>
<td>n_{22}</td>
<td>n_{23}</td>
<td>n_{24}</td>
<td>...</td>
<td>n_{2j}</td>
</tr>
<tr>
<td>3</td>
<td>n_{31}</td>
<td>n_{32}</td>
<td>n_{33}</td>
<td>n_{34}</td>
<td>...</td>
<td>n_{3j}</td>
</tr>
<tr>
<td>i</td>
<td>n_{i1}</td>
<td>n_{i2}</td>
<td>n_{i3}</td>
<td>n_{i4}</td>
<td>...</td>
<td>n_{ij}</td>
</tr>
</tbody>
</table>

\[ n_{1.} \quad n_{2.} \quad n_{3.} \quad n_{.} \quad n \]

In the table 3.1 above, we denote the cell counts by \( n_{ij} \). The marginal frequencies are the row totals \( \{ n_{i.} \} \) and the column totals \( \{ n_{.j} \} \), while \( n = \sum_{i,j} n_{ij} \) denotes the total sample size.

Probabilities for contingency tables can be of three types: joint, marginal, or conditional. Suppose first that a randomly chosen subject from the population of interest is classified on \( X \) and \( Y \). Let \( \pi_{ij} = p_{ij} = P(X = i, Y = j) = \frac{n_{ij}}{n} \) denote the probability that \((X,Y)\) falls in the cell in row \( i \) and column \( j \). The probabilities \( \{ \pi_{ij} \} \) form the joint distribution of \( X \) and \( Y \) and they satisfy \( \sum_{i,j} \pi_{ij} = 1 \). The marginal distributions \( \pi_{i.} = \frac{n_{i.}}{n} \) and \( \pi_{.j} = \frac{n_{.j}}{n} \) are the row and column totals of the joint probabilities. When a separate probability distribution for \( Y \) at each level of \( X \) is constructed, such a distribution consists of conditional probabilities for, given the level of \( X \) and it is called a conditional distribution.
3.3.2. Sensitivity and Specificity

The result of a diagnostic test is said to be positive if it states that the disease is present and negative if it states that the disease is absent. The accuracy of diagnostic tests is often assessed with two conditional probabilities: Given that a subject has the disease, the probability the diagnostic test is positive is called the sensitivity. Given that the subject does not have the disease, the probability the test is negative is called the specificity. It means, let $X$ denotes the true state of a person, with categories $1 =$ diseased, $2 =$ not diseased, and let $Y =$ outcome of diagnostic test, with categories $1 = $ positive, $2 = $ negative. Then we have:

\[
\text{sensitivity} = \Pr(Y = 1|X = 1) \quad \text{and} \quad \text{specificity} = \Pr(Y = 2|X = 2)
\]

The higher the sensitivity and specificity, the better the diagnostic test.

3.3.3. Odds ratio

If an event $A$ has probability $p(A)$ of occurring, the odds of $A$ occurring are defined as:

\[
\text{Odds} \ (A) = \frac{p(A)}{1-p(A)} \quad (3.3)
\]

which implies that:

\[
p(A) = \frac{\text{Odds}(A)}{1+\text{Odds}(A)} \quad (3.4)
\]

Now, suppose that $X$ denotes the event that an individual is exposed to a risk of having a disease and that $D$ denotes the event that the individual becomes diseased. We denote the complementary events as $\bar{X}$ and $\bar{D}$. The odds of an individual contracting the disease given that he is exposed are:

\[
\text{Odds} \ (D/X) = \frac{p(D/X)}{1-p(D/X)}
\]
And the odds of an individual contracting the disease given that he is not exposed are:

\[ \text{Odds} \left( D / \bar{X} \right) = \frac{p(D/\bar{X})}{1 - p(D/\bar{X})} \]

The odds ratio is:

\[ \Delta = \frac{\text{Odds} \left( D/X \right)}{\text{Odds} \left( D/\bar{X} \right)} \quad (3.5) \]

is a measure of the influence of exposure on subsequent disease.

Now consider the following contingency table of joint and marginal probabilities:

Table 3.2: Contingency table of joint and marginal probabilities

<table>
<thead>
<tr>
<th></th>
<th>( D )</th>
<th>( D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{X} )</td>
<td>( \pi_{00} )</td>
<td>( \pi_{01} )</td>
</tr>
<tr>
<td>( X )</td>
<td>( \pi_{10} )</td>
<td>( \pi_{11} )</td>
</tr>
<tr>
<td>( \pi_0 )</td>
<td>( \pi_1 )</td>
<td>1</td>
</tr>
</tbody>
</table>

In this notation,

\[ p(D/X) = \frac{\pi_{11}}{\pi_{10} + \pi_{11}} \]

\[ p(D/\bar{X}) = \frac{\pi_{01}}{\pi_{00} + \pi_{01}} \]

So that:

\[ \text{Odds}(D/X) = \frac{\pi_{11}}{\pi_{10}} \quad (3.6) \]

\[ \text{Odds}(D/\bar{X}) = \frac{\pi_{01}}{\pi_{00}} \quad (3.7) \]
And the odds ratio is:

\[
\Delta = \frac{\pi_{11} \pi_{00}}{\pi_{01} \pi_{10}}
\]  

(3.8)

3.4. GENERALISED LINEAR MODELS (GLMs)

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

There are three components to a Generalized Linear Model:

- **Random Component**: The *random component* of a Generalized Linear Model identifies the response variable \( Y \) and selects a probability distribution for it. Denote the observations on \( Y \) by \( (Y_1, Y_2, \ldots, Y_n) \). Standard GLMs treat \( Y_1, Y_2, \ldots, Y_n \) as independent.

- **Systematic Component**: The *systematic component* of a GLM specifies the explanatory variables. These enter linearly as predictors on the right-hand side of the model equation. That is, the systematic component specifies the variables that are the \( \{x_j\} \) in the expression:

\[
\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k
\]  

(3.9)

This linear combination of the explanatory variables is called the *linear predictor*.

- **Link Function**: Let us denote the expected value of \( Y \), the mean of its probability distribution, by \( \mu = E(Y) \).

\[
\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k
\]  

(3.10)

The third component of a GLM, the *link function*, specifies a function \( g(\cdot) \) that relates \( \mu \) to the linear predictor as:

\[
g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k
\]  

(3.11)

The link function \( g(\cdot) \) connects the random and systematic components.
3.5. LOGISTIC REGRESSION MODEL

3.5.1. Introduction

Given the nature of the problem at hand and of data used in this work, multiple logistic regression model is used. In general, this model is employed to model the outcomes of a categorical dependent variable. For categorical variables, it is inappropriate to use linear regression model because the response values are not measured on a ratio scale and the error terms are not normally distributed. In addition, the linear regression model can generate as predicted values any real number ranging from negative to positive infinity, whereas a categorical variable can only take on a limited number of discrete values within a specified range.

The crucial limitation of linear regression is that it cannot deal with dependent variables that are dichotomous and categorical. Many interesting variables in business and medical world are dichotomous. For example, consumers make a decision to buy or not buy, a product may pass or fail quality control; there are good or poor credit risks; an employee may be promoted or not, etc.

A range of regression techniques have been developed for analyzing data with categorical dependent variables. These techniques include logistic regression analysis. Logistic regression determines the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories. The logistic regression is the most popular multivariable method used in health science (Tetrault et al., 2008).

3.5.2. Binary Logistic Regression with single independent variable

Many categorical response variables have only two categories. Denote a binary response variable by $Y$ and its two possible outcomes by 1 (“success”) and 0 (“failure”). The distribution of $Y$ is specified by probabilities:
\( P(Y = 1) = \pi \) of success and \( P(Y = 0) = (1 - \pi) \) of failure. Its mean is \( E(Y) = \pi \).

For \( n \) independent observations, the number of successes has the binomial distribution specified by the index \( n \) and parameter \( \pi \). The formula was given in equation (3.2). Each binary observation is a binomial variate with \( n = 1 \). Although Generalized Linear Models can have multiple explanatory variables, for simplicity we introduce them using a single \( x \).

The value of \( \pi \) can vary as the value of \( x \) changes, and \( \pi \) is replaced by \( \pi(x) \) when we want to describe its dependence on that value. Relationships between \( \pi(x) \) and \( x \) are usually nonlinear rather than linear. In the logistic regression model, the random component for the (success, failure) outcomes has a **binomial distribution**. The link function is the logit function

\[
\ln[\pi/(1 - \pi)] \text{ of } \pi, \text{ which is defined as the log of odds of success and symbolized by } \logit(\pi)\text{.}
\]

Logistic regression models are often called *logit models*. Whereas \( \pi \) is restricted to the range \([0,1]\), the logit can be any real number.

The model:

\[
\ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 X
\]  
(3.12)

From equation (3.12), we deduce:

\[
\frac{\pi(x)}{1 - \pi(x)} = e^{\beta_0 + \beta_1 X}
\]

\[
\pi(x) = e^{\beta_0 + \beta_1 X} - \pi(x)e^{\beta_0 + \beta_1 X}
\]

\[
\pi(x)(1 + e^{\beta_0 + \beta_1 X}) = e^{\beta_0 + \beta_1 X}
\]

\[
\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}
\]  
(3.13)

3.5.3. Interpretation of regression coefficients

Consider the case in which the dependent variable may take only the values 1 (for success) and 0 (for failure) and a single independent variable X.
In this case, the logistic regression equation is:

\[ \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 X \]

as given in equation (3.12)

Now, suppose we consider an impact of a unit increase in \( X \). The logistic regression equation becomes:

\[ \ln \left( \frac{\pi'(x)}{1 - \pi'(x)} \right) = \beta_0 + \beta_1 (X + 1) \]

\[ \ln \left( \frac{\pi'(x)}{1 - \pi'(x)} \right) = \beta_0 + \beta_1 X + \beta_1 \] (3.14)

Subtracting equation (3.9) from (3.11) we get:

\[ \ln \left( \frac{\pi'(x)}{1 - \pi'(x)} \right) - \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 X + \beta_1 - \beta_0 - \beta_1 X \]

to arrive at:

\[ \beta_1 = \ln \left( \frac{\frac{\pi'(x)}{1 - \pi'(x)}}{\frac{\pi(x)}{1 - \pi(x)}} \right) \]

\[ \beta_1 = \ln \left( \frac{\text{odds}'}{\text{odds}} \right) \] (3.15)

That is, \( \beta_1 \) is the log of the ratio of the odds at \( X + 1 \) and \( X \).

which may be also written as:

\[ e^{\beta_1} = \frac{\text{odds}'}{\text{odds}} \] (3.16)

The regression coefficient \( \beta_1 \) is interpreted as the log of the odds ratio comparing the odds after a one unit increase in \( X \) to the original odds.

3.5.4. The logistic curve

Logistic regression fits a logistic curve to the relationship between \( x \) and \( y \). Logistic curve has an S-shaped or sigmoid curve). A logistic curve starts with slow, linear growth, followed by exponential growth, which then arrives again to a stable rate.
3.5.5. Assumptions of logistic regression

- Logistic regression does not assume a linear relationship between the dependent and independent variables.

- The dependent variable must be a dichotomy (2 categories).

- The independent variables are not normally distributed, nor linearly related, nor of equal variance within each group.

- The categories (groups) must be mutually exclusive and exhaustive; a case can only be in one group and every case must be a member of one of the groups.

- Larger samples are needed than for linear regression because maximum likelihood coefficients are large sample estimates.
3.5. 6. Multiple Logistic Regression

3.5.6.1. The model

Let us consider the general logistic regression model with multiple explanatory variables. Denote the $k$ predictors for a binary response $Y$ by $X_1, X_2, ..., X_k$.

We use $\pi(x)$ to represent the probability that $Y = 1$ for success, and $1 - \pi(x)$ to represent the probability that $Y = 0$.

These probabilities are written in the following form:

\[
\pi(x) = P(Y = 1/X_1, X_2, ..., X_k) \quad (3.17)
\]

\[
1 - \pi(x) = P(Y = 0/X_1, X_2, ..., X_k) \quad (3.18)
\]

The model for the log odds is:

\[
\text{logit} \left( \pi(x) \right) = \ln \frac{P(Y=1/X_1, X_2, ..., X_k)}{P(Y=0/X_1, X_2, ..., X_k)}
\]

which gives

\[
\ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_k + \varepsilon
\]

and

\[
\ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_k + \varepsilon
\]

\[
\therefore \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon \quad (3.19)
\]

which yields to:

\[
\pi(x) = P(Y = 1/X_1, X_2, ..., X_k) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon}} \quad (3.20)
\]

The parameter $\beta_j$ refers to the effect of $X_j$ on the log odds that $Y = 1$, controlling the other predictor variables. For example, $\exp(\beta_j)$ is the multiplicative effect on the odds of a one-unit increase in $X_j$, at fixed levels of the other predictor variables.
3.5.6.2. The Parameters estimation

The goal of logistic regression is to estimate the $K + 1$ unknown parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$. This is done with maximum likelihood estimation which entails finding the set of parameters for which the probability of the observed data is greatest. The maximum likelihood equation is derived from the binomial distribution of the dependent variable.

For a set of observations in the data $(x_i; y_i)$, the contribution to the likelihood function is $\pi(x_i)$, where $y_i = 1$, and $1 - \pi(x_i)$, where $y_i = 0$. The following equation results for the contribution (call it $\varphi(x_i)$) to the likelihood function for the observation $(x_i; y_i)$:

$$\varphi(x_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (3.21)$$

The equation 3.18 accounts for only one set of observations. The observations are assumed to be independent of each other so we can multiply their likelihood contributions to obtain the complete likelihood function. The result is given in equation (3.22):

$$l(\beta) = \prod_{i=1}^{k} \varphi(x_i) = \prod_{i=1}^{k} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

$$\therefore \quad l(\beta) = \pi(x_i)^{\sum_{i=1}^{k} y_i} [1 - \pi(x_i)]^{k - \sum_{i=1}^{k} y_i}$$

$$\therefore \quad l(\beta) = \pi(x_i)^{\sum_{i=1}^{k} y_i} [1 - \pi(x_i)]^{k} [1 - \pi(x_i)]^{\sum_{i=1}^{k} y_i}$$

$$\therefore \quad l(\beta) = \left[ \frac{\pi(x_i)}{1 - \pi(x_i)} \right]^{\sum_{i=1}^{k} y_i} [1 - \pi(x_i)]^k \quad (3.22)$$

Note that the equation (3.16) and (3.17) give respectively:

$$\left( \frac{\pi(x_i)}{1 - \pi(x_i)} \right) = e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j} \quad \text{and} \quad \pi(x_i) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j}}$$

This leads us to write equation (3.19) as:

$$l(\beta) = \left( e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j} \right)^{\sum_{i=1}^{k} y_i} \left( 1 - \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j}} \right)^k$$
\[ l(\beta) = \left( e^{\beta_0 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i \sum_{j=1}^{k} \beta_j x_{ij}} \right) \left( \frac{1 + e^{\beta_0 \sum_{j=1}^{k} \beta_j x_j} - e^{\beta_0 \sum_{i=1}^{k} y_i \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\beta_0 \sum_{j=1}^{k} \beta_j x_j}} \right)^k \]

\[ l(\beta) = \left( e^{\beta_0 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i \sum_{j=1}^{k} \beta_j x_{ij}} \right) \left( \frac{1}{1 + e^{\beta_0 \sum_{j=1}^{k} \beta_j x_j}} \right)^k \]

\[ l(\beta) = \left( e^{\beta_0 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i \sum_{j=1}^{k} \beta_j x_{ij}} \right) \left( 1 + e^{\beta_0 \sum_{j=1}^{k} \beta_j x_j} \right)^{-k} \] (3.23)

In the equation (3.20), \( \beta \) is the collection of parameters \( \beta_0, \beta_1, \ldots, \beta_k \), and \( l(\beta) \) is the likelihood function of \( \beta \). The Maximum likelihood estimates (MLE’s) \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \) can be obtained by calculating the \( \beta \) which maximizes \( l(\beta) \). However, to simplify the mathematics, let us take the logarithm of equation (3.20). As shown in equation (3.21), \( L(\beta) \) denotes the log likelihood expression.

\[ L(\beta) = \ln(l(\beta)) = \ln \left[ \left( e^{\beta_0 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i \sum_{j=1}^{k} \beta_j x_{ij}} \right) \left( 1 + e^{\beta_0 \sum_{j=1}^{k} \beta_j x_j} \right)^{-k} \right] \]

\[ L(\beta) = \ln(l(\beta)) = \left( \beta_0 \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i \sum_{j=1}^{k} \beta_j x_{ij} \right) - k \ln \left( 1 + e^{\beta_0 \sum_{j=1}^{k} \beta_j x_j} \right) \] (3.24)

The critical points of a function (maxima and minima) occur when the first derivative equals 0. If the second derivative evaluated at that point is less than zero, then the critical point is a maximum. Thus, finding the maximum likelihood estimates requires computing the first derivative of the log likelihood function \( L(\beta) \).

Thus, differentiating equation (3.21) with respect to \( \beta_0 \), we get:

\[ \frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{k} y_i - k \frac{e^{\beta_0 + \sum_{i=0}^{k} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{i=0}^{k} \beta_j x_{ij}}} \]

\[ \frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{k} y_i - k \pi(x_i) \]
Also, differentiating equation (3.21) with respect to \( \beta_j \), we get:

\[
\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{k} y_i \sum_{j=1}^{k} x_j - k \sum_{j=1}^{k} x_j \left( \frac{e^{\beta_0 + \sum_{i=0}^{k} \beta_j x_i}}{1 + e^{\beta_0 + \sum_{i=0}^{k} \beta_j x_i}} \right)
\]

\[
\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{k} y_i \sum_{j=1}^{k} x_j - k \sum_{j=1}^{k} x_j \pi(x_i)
\]

\[
\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{k} x_{ik} \left[ y_i - \pi(x_i) \right]
\]

The maximum likelihood estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_j \) for \( \beta_0 \) and \( \beta_j \) can be found by setting each of the equations respectively (3.25) and (3.26) equal to zero and solving for each \( \beta_j \).
3.5.7. Evaluation of a logistic regression model

Evaluations of logistic regression model include the overall evaluations, statistical test of individual predictors, goodness-of-fit statistics, and validations of predicted probabilities. Each is illustrated next for the logistic model.

3.5.7.1. Overall model evaluations: The likelihood ratio test

A logistic model is said to provide a better fit to the data if it demonstrates an improvement over the intercept only model (also called the null model, which has no predictors).

The likelihood ratio test for overall significance of the coefficients for the independent variables in the model is used. The test is based on the statistic" G" under the null hypothesis:

\[ H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0. \]

and G statistic is calculated as:

\[ G = \chi^2 = (-2 \ln \text{likelihood of null model}) - (-2 \ln \text{likelihood of model with the variables}) \]

(3.29)

The distribution of "G" is a chi-square with k degree-of-freedom, where k is the number of covariates in the logistic regression equation. This is a measure of how well all of the independent variables affect the response variable. (Bewick, Cheek, & Ball, 2005).

If the \( p - value \) for the overall model fit statistic is less than the conventional 0.05, then reject \( H_0 \) at \( \alpha = 0.05 \) and the conclusion will be that there is evidence that at least one of the independent variables contributes to the prediction of the outcome.
3.5.7. 2. Statistical significance of individual regression coefficients:

Wald test and Confidence Interval

a) Wald statistic

To assess the significance of the logistic regression coefficients, the **Wald statistic** is used.

( Afifi et al., 2004) and (Bewick et al., 2005).

The Wald statistic $W_j$ is is calculated as:

$$ W_j = \frac{\hat{\beta}_j^2}{[SE(\hat{\beta}_j)]^2} \quad (3.30) $$

Where $\hat{\beta}_j$ represents the estimated coefficient of $\beta$ and $SE(\hat{\beta}_j)$ is its standard error. Under the null hypothesis $H_0 : \beta_j = 0$, the quantity (3.27) follows a chi-square distribution with one degree of freedom. If the estimated value of the slope is small and its estimated variability is large, then we cannot conclude that the slope is significantly different from zero and vise versa (Afifi et al., 2004).

b) Confidence Interval

Odds ratio with 95% confidence interval (CI) can be used to test for the contribution of individual predictors (Katz, 1999). Note that however, that unlike the p value, the 95% CI does not report a measure’s statistical significance.

The 95% Confidence Interval is used to estimate the precision of the Odds Ratio (OR). A large Confidence Interval indicates a low level of precision of the Odds Ratio, whereas a small Confidence Interval indicates a higher precision of the Odds Ratio.

This is computed as follows:

- A 95% Confidence intervals for $\hat{\beta}_j$, is given by: $\hat{\beta}_j \pm 1.96 \times SE(\hat{\beta}_j)$ \quad (3.31)

- A 95% CI for log Odds Ratio $= ln (OR) \pm 1.96 \times \{SE ln (OR)\}$ \quad (3.32)
where \( \ln(OR) \) is the sample log odds ratio, and \( SE \ln(OR) \) is the standard error of the log odds ratio.

\[ -A \ 95\% \ CI \ for \ Odds \ Ratio = e^{\ln(OR) \pm 1.96 \times (SE \ln(OR))} \]  

(Morris & Gardner, 1988).

3.5.7.3. Goodness-of-fit statistics: Hosmer - Lemeshow test

The Hosmer-Lemeshow test helps to examine whether the observed proportions of events are similar to the predicted probabilities of occurrence in subgroups of the model population.

The Hosmer-Lemeshow test is assessed by dividing the predicted probabilities into deciles (10 groups based on percentile ranks) and then computing a Pearson Chi-square that compares the predicted to the observed frequencies in a 2-by-10 table.

The value of the test statistics is:

\[ H = \sum_{g=1}^{10} \left( \frac{O_g - E_g}{E_g} \right)^2 \]  

(3.34)

In the above formula, \( O_g \) and \( E_g \) denote the observed events, and expected events for the \( g^{th} \) risk deciles group respectively. The test statistic asymptotically follows a chi-square distribution with \( 8 \) (number of groups minus two) degrees of freedom.

Small values (with large \(-value closer to 1\)) indicate a good fit to the data, therefore, good overall model fit. Large values (with \( p - value < 0.05 \)) indicate a poor fit to the data.

3.5.7.4. Graphing prediction accuracy: Receiver Operating Characteristic (ROC) curve

One primary goal of performing logistic regression is to generate an equation that can reliably classify observations into one of two outcomes. The degree to which predictions agree with the data may be shown graphically by a receiver operating characteristic (ROC) curve.

According to Hosmer & Lemeshow (2000), the ROC curve is a plot of sensitivity versus
1-specificity. Sensitivity is defined as the proportion of observations correctly classified as an event. Specificity is defined as the proportion of observations correctly classified as nonevent. Hence, 1-specificity is the proportion of observations misclassified as an event; which is also called the false positive fraction.

ROC (Receiver Operating Characteristic) analysis is being used as a method for evaluation and comparison of classifiers (Ferri et al., 2002). The ROC gives complete description of classification accuracy as given by the area under the ROC curve.

The model with a larger area below the ROC curve is considered as better model. Alternatively, the one with the greatest height on the ROC curve at a desirable probability cutoff should be chosen. In other words, the best model is the one associated with the greatest sensitivity and the lowest 1-specificity.

3.6. RESEARCH DESIGN

Research design used for this study is descriptive survey. A study of the risk factors of diabetes is described through a multiple logistic regression model.

The choice of the design survey was considered appropriate because it allows verifying whether the studied factors are statistically significant or not. The description through a multiple logistic regression model was preferred because the dependent variable is dichotomous (having diabetes and not having diabetes) and the independent variables are either continuous or categorical. The use of survey therefore was considered to be more appropriate in terms of resources, time and the overall objective of the study.
3.7. STUDY POPULATION

This study was conducted to find out whether factors like older age, gender, occupation status, smoking, alcohol consumption, Cholesterol level, hypertension and family history of diabetes cause jointly or partially the presence of diabetes illness.

A three year period secondary data from Gitwe Hospital are used. This period extends from 2011 to 2013.

Data considered in this study are relevant for the following reasons:

- First, the area is considered to be more appropriate in terms of financial resources, time and the overall objective of the study.
- Second, the Hospital is located in rural area and works with the people living in different lifestyle and all economic categories are represented in the sample.
- Last, the data are recent and may reveal current situation of diabetes.

3.8. SAMPLE SIZE AND SAMPLING PROCEDURE

Sampling is a process of selecting a number of individuals or objects from a population such that the selected group contains elements representative of the characteristics sought in the entire population. Gitwe Hospital and the three years 2011, 2012 and 2013 were purposively selected according to the objectives of the study. The target population of the study includes in total 311 patients from Gitwe Hospital (2011-2013) dispatched in the following six different sectors as:

Table 3.3: Number of patients by sector

<table>
<thead>
<tr>
<th>Sector</th>
<th>Ruhango</th>
<th>Kabagali</th>
<th>Mukiing</th>
<th>Kinihira</th>
<th>Bwerama</th>
<th>Busoro</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>18</td>
<td>46</td>
<td>43</td>
<td>35</td>
<td>166</td>
<td>3</td>
<td>311</td>
</tr>
</tbody>
</table>

The sample size for patients is determined using the Yamane (1967) formula. That is:
where \( N \) is the population size and \( e \) is the precision level.

Concerning our case study; the total number of patients’ folders (\( N \)) is 311. Then, by the equation (3.19), the sample size is given as

\[
(3.35)\quad n = \frac{N}{1+N(e)^2} = \frac{311}{1+311*(0.05)^2} = 174.9 \approx 175
\]

Table 3.4: Calculation of sample size by Sector

<table>
<thead>
<tr>
<th>PERCENTAGE FOR EACH SECTOR</th>
<th>SAMPLE SIZE BY SECTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUHANGO: ( \frac{18+100}{311} = 6% )</td>
<td>( n_{Ruhango} = \frac{175 \times 6}{100} = 10.5 \approx 11 )</td>
</tr>
<tr>
<td>KABAGALI: ( \frac{46+100}{311} = 15% )</td>
<td>( n_{Kabagali} = \frac{175 \times 15}{100} = 26.25 \approx 26 )</td>
</tr>
<tr>
<td>MUKINGO: ( \frac{43+100}{311} = 14% )</td>
<td>( n_{Mukingo} = \frac{175 \times 14}{100} = 24.5 \approx 25 )</td>
</tr>
<tr>
<td>KINIHIRA: ( \frac{35+100}{311} = 11% )</td>
<td>( n_{Kinihira} = \frac{175 \times 11}{100} = 19.25 \approx 19 )</td>
</tr>
<tr>
<td>BWERAMANA: ( \frac{166+100}{311} = 53% )</td>
<td>( n_{Bweramana} = \frac{175 \times 53}{100} = 92.75 \approx 93 )</td>
</tr>
<tr>
<td>BUSORO: ( \frac{3+100}{311} = 0.9% )</td>
<td>( n_{Busoro} = \frac{175 \times 0.9}{100} = 1.5 \approx 1 )</td>
</tr>
<tr>
<td>TOTAL SAMPLE SIZE</td>
<td>( n=175 )</td>
</tr>
</tbody>
</table>

However, systematic random sampling has been used to select the patients’ folders to be included in the sample size of each Sector. With this sampling technique, each element has an equal probability of being selected, but combinations of elements have different probabilities. In a population of size \( N \), if the sample desired is of size \( n \), the sampling interval \( k=\frac{N}{n} \); randomly, a number \( j \) between 1 and \( k \) is selected, and then every \( k^{th} \) element thereafter is taken. It means elements \( j+k \), \( j+2k \), \( j+3k \), \( j+4k \), etc until the sample of size \( n \) is completed.

The following table displays the number of patients within the sample by each variable:
Table 3. 5: Number of people in the sample by each variable

<table>
<thead>
<tr>
<th>The Patient is diabetic</th>
<th>Number of patients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>66</td>
<td>37.7%</td>
</tr>
<tr>
<td>Yes</td>
<td>109</td>
<td>62.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age of patient</th>
<th>Number of patients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>From 20 to 39 years</td>
<td>44</td>
<td>25.1%</td>
</tr>
<tr>
<td>From 40 to 59 years</td>
<td>55</td>
<td>31.4%</td>
</tr>
<tr>
<td>60 years and over</td>
<td>76</td>
<td>43.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender of patient</th>
<th>Number of patients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>105</td>
<td>60.0%</td>
</tr>
<tr>
<td>Female</td>
<td>70</td>
<td>40.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Occupation status of the patient</th>
<th>Number of patients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employed</td>
<td>93</td>
<td>53.1%</td>
</tr>
<tr>
<td>Unemployed</td>
<td>82</td>
<td>46.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoking</th>
<th>Number of patients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>14</td>
<td>8.0%</td>
</tr>
<tr>
<td>No</td>
<td>161</td>
<td>92.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alcohol consumption</th>
<th>Number of patients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>93</td>
<td>53.1%</td>
</tr>
<tr>
<td>No</td>
<td>82</td>
<td>46.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cholesterol level of the patient</th>
<th>Number of patients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>101</td>
<td>57.7%</td>
</tr>
<tr>
<td>Low</td>
<td>74</td>
<td>42.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The patient has hypertension</th>
<th>Number of patients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>115</td>
<td>65.7%</td>
</tr>
<tr>
<td>No</td>
<td>60</td>
<td>34.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The Family history of diabetes</th>
<th>Number of patients</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>45</td>
<td>25.7%</td>
</tr>
<tr>
<td>No</td>
<td>130</td>
<td>74.3%</td>
</tr>
</tbody>
</table>
3.9. DATA COLLECTION AND DATA ANALYSIS

The study would have considered all the district hospitals in the country but GITWE Hospital was sampled out purposively for the research. The secondary data were used. Those data were taken from GITWE Hospital patients’ database record related to Cardiovascular diseases and diabetes, covering a three-year period (2011 to 2013).

The following information from the folders of registered patients was taken: age, gender, occupation status, smoking, alcohol consumption, Cholesterol level, hypertension, family history and having diabetes or not. The SPSS 15.0 software package is used to process data.
CHAPTER FOUR

RESEARCH RESULTS AND DISCUSSION

4.0: INTRODUCTION

This chapter describes three steps about the results of this study. The first step verifies some important assumptions (non multicollinearity, non-linearity and non-normality) underlying the logistic regression. The second step discusses the chi-square test of dependence and the third step discusses the results from the fitted logistic regression model with the predictor variables.

The data used for the research consist of eight independent variables: Age, Gender (Gen), Occupation status (Occ.Stat), Smoking (Smok), Alcohol Consumption (Alcoh), Cholesterol level (chol), Hypertension (Hyper) and Family history of diabetes (Famhist). The dependent variable of our study is diabetes.

Thus, from the equation (3.17) the model to fit is of the following form:

\[
\pi_i = \frac{\exp(\beta_0 + \beta_1 \text{age} + \beta_2 \text{gen} + \beta_3 \text{Occ.Stat} + \beta_4 \text{Smok} + \beta_5 \text{Alcoh} + \beta_6 \text{Chol} + \beta_7 \text{Hyper} + \beta_8 \text{Famhist} + \varepsilon)}{1 + \exp(\beta_0 + \beta_1 \text{age} + \beta_2 \text{gen} + \beta_3 \text{Occ.Stat} + \beta_4 \text{Smok} + \beta_5 \text{Alcoh} + \beta_6 \text{Chol} + \beta_7 \text{Hyper} + \beta_8 \text{Famhist} + \varepsilon)}
\]  

(4.1)

The study consisted of 93 people who were employed and 82 people who were unemployed. 115 people were hypertensive and 60 were not hypertensive. 93 people consumed alcohol and 82 did not, 14 were smokers and 161 were not, 45 people had a family history of diabetes whiles 130 people did not. There were 105 males and 70 females, while 101 people were with high cholesterol level and 74 with low cholesterol level. 109 people were diabetics while 66 were not.

The following figures show the state of the sampled patients:
Figure 4.1: The state of the sample by each variable (Source: This figure is drawn from the data of the table 3.2 above)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20-39 years</td>
<td>Males</td>
<td>Employed</td>
<td>Smokers</td>
<td>Non Smokers</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>40-59 years</td>
<td>Females</td>
<td>Unemployed</td>
<td></td>
<td></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>60 years and over</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Figure 4.2: Diabetes state by each variable
4.1. Checking for multicollinearity between the independent variables

If one variable is a perfect linear function of another in the model, standard errors become infinite and the solution to the model becomes indeterminate. To the extent that one independent is a near but not perfect linear function of another independent, the problem of multicollinearity will occur in logistic regression. As the independents increase in correlation with each other, the standard errors of the logit (effect) coefficients will become inflated. Multicollinearity does not change the estimates of the coefficients, only their reliability. To avoid the misleading results, we have used the Variance Inflation Factor (VIF) to check for multicollinearity between the independent variables. According to Robert M. (2007), the Variance Inflation Factor (VIF) and tolerance are both widely used measures of the degree of multi-collinearity of the $i^{th}$ independent variable with the other independent variables in a regression model. Practitioners often inappropriately apply rules or criteria that indicate when the values of VIF or tolerance have attained unacceptably high levels. Neter et al. (1989) state that a maximum VIF value in excess of 10 is often taken as an indication that multicollinearity may be unduly influencing the estimates. Also, Hair et al. (1995) suggest that a VIF of less than 10 are indicative of inconsequential collinearity. Marquardt (1970) uses a VIF greater than 10 as a guideline for serious multicollinearity.

According to Kennedy (1992), a VIF greater than 10 indicates harmful collinearity. When the VIF reaches these threshold levels, researchers may feel compelled to reduce the collinearity by eliminating one or more variables from their analysis; combining two or more independent variables into a single index; resorting to a biased regression technique that can reduce the variance of the estimated regression coefficients; or, in rejecting a paper because VIF exceeds a threshold value. (Belsley et al., 1980)
The following tables show the results of the checking from SPSS 15.0:

Table 4.1: Checking for multicollinearity by Variance Inflation Factor (VIF) (from a to h)

<table>
<thead>
<tr>
<th></th>
<th>Collinearity Statistics</th>
<th></th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tolerance</td>
<td>VIF</td>
<td></td>
</tr>
<tr>
<td>a) Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender of patient</td>
<td>.971</td>
<td>1.030</td>
<td>Occupation status of the patient</td>
</tr>
<tr>
<td>Occupation status of the patient</td>
<td>.951</td>
<td>1.051</td>
<td>Cigarette taking</td>
</tr>
<tr>
<td>Cigarette taking</td>
<td>.974</td>
<td>1.027</td>
<td>Alcohol consumption</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>.975</td>
<td>1.026</td>
<td>Cholesterol level of the patient</td>
</tr>
<tr>
<td>Cholesterol level of the patient</td>
<td>.974</td>
<td>1.026</td>
<td>The patient has hypertension</td>
</tr>
<tr>
<td>The patient has hypertension</td>
<td>.974</td>
<td>1.026</td>
<td>The Family history of diabetes</td>
</tr>
<tr>
<td>The Family history of diabetes</td>
<td>.975</td>
<td>1.025</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Collinearity Statistics</th>
<th></th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tolerance</td>
<td>VIF</td>
<td></td>
</tr>
<tr>
<td>b) Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age of patient</td>
<td>.947</td>
<td>1.056</td>
<td>Occupation status of the patient</td>
</tr>
<tr>
<td>Occupation status of the patient</td>
<td>.977</td>
<td>1.024</td>
<td>Cigarette taking</td>
</tr>
<tr>
<td>Cigarette taking</td>
<td>.977</td>
<td>1.024</td>
<td>Alcohol consumption</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>.975</td>
<td>1.026</td>
<td>Cholesterol level of the patient</td>
</tr>
<tr>
<td>Cholesterol level of the patient</td>
<td>.975</td>
<td>1.026</td>
<td>The patient has hypertension</td>
</tr>
<tr>
<td>The patient has hypertension</td>
<td>.974</td>
<td>1.026</td>
<td>The Family history of diabetes</td>
</tr>
<tr>
<td>The Family history of diabetes</td>
<td>.980</td>
<td>1.020</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Collinearity Statistics</th>
<th></th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tolerance</td>
<td>VIF</td>
<td></td>
</tr>
<tr>
<td>c) Occupation status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age of patient</td>
<td>.879</td>
<td>1.138</td>
<td>Gender of patient</td>
</tr>
<tr>
<td>Gender of patient</td>
<td>.879</td>
<td>1.138</td>
<td>Cigarette taking</td>
</tr>
<tr>
<td>Cigarette taking</td>
<td>.977</td>
<td>1.023</td>
<td>Alcohol consumption</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>.980</td>
<td>1.021</td>
<td>Cholesterol level of the patient</td>
</tr>
<tr>
<td>Cholesterol level of the patient</td>
<td>.961</td>
<td>1.040</td>
<td>The patient has hypertension</td>
</tr>
<tr>
<td>The patient has hypertension</td>
<td>.934</td>
<td>1.071</td>
<td>The Family history of diabetes</td>
</tr>
<tr>
<td>The Family history of diabetes</td>
<td>.980</td>
<td>1.020</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Collinearity Statistics</th>
<th></th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tolerance</td>
<td>VIF</td>
<td></td>
</tr>
<tr>
<td>d) Smoking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age of patient</td>
<td>.876</td>
<td>1.142</td>
<td>Gender of patient</td>
</tr>
<tr>
<td>Gender of patient</td>
<td>.876</td>
<td>1.142</td>
<td>Cigarette taking</td>
</tr>
<tr>
<td>Cigarette taking</td>
<td>.952</td>
<td>1.051</td>
<td>Alcohol consumption</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>.985</td>
<td>1.015</td>
<td>Cholesterol level of the patient</td>
</tr>
<tr>
<td>Cholesterol level of the patient</td>
<td>.942</td>
<td>1.062</td>
<td>The patient has hypertension</td>
</tr>
<tr>
<td>The patient has hypertension</td>
<td>.928</td>
<td>1.077</td>
<td>The Family history of diabetes</td>
</tr>
<tr>
<td>The Family history of diabetes</td>
<td>.979</td>
<td>1.022</td>
<td></td>
</tr>
</tbody>
</table>
### e) Alcohol consumption

<table>
<thead>
<tr>
<th></th>
<th>Tolerance</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of patient</td>
<td>.876</td>
<td>1.142</td>
</tr>
<tr>
<td>Gender of patient</td>
<td>.899</td>
<td>1.113</td>
</tr>
<tr>
<td>Occupation status of the patient</td>
<td>.953</td>
<td>1.049</td>
</tr>
<tr>
<td>Cigarette taking</td>
<td>.985</td>
<td>1.016</td>
</tr>
<tr>
<td>Cholesterol level of the patient</td>
<td>.944</td>
<td>1.060</td>
</tr>
<tr>
<td>The patient has hypertension</td>
<td>.928</td>
<td>1.078</td>
</tr>
<tr>
<td>The Family history of diabetes</td>
<td>.975</td>
<td>1.025</td>
</tr>
</tbody>
</table>

### f) Cholesterol level

<table>
<thead>
<tr>
<th></th>
<th>Tolerance</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of patient</td>
<td>.877</td>
<td>1.141</td>
</tr>
<tr>
<td>Gender of patient</td>
<td>.904</td>
<td>1.106</td>
</tr>
<tr>
<td>Occupation status of the patient</td>
<td>.969</td>
<td>1.032</td>
</tr>
<tr>
<td>Cigarette taking</td>
<td>.975</td>
<td>1.026</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>.978</td>
<td>1.023</td>
</tr>
<tr>
<td>The patient has hypertension</td>
<td>.943</td>
<td>1.060</td>
</tr>
<tr>
<td>The Family history of diabetes</td>
<td>.975</td>
<td>1.025</td>
</tr>
</tbody>
</table>

### g) Hypertension

<table>
<thead>
<tr>
<th></th>
<th>Tolerance</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of patient</td>
<td>.919</td>
<td>1.088</td>
</tr>
<tr>
<td>Gender of patient</td>
<td>.902</td>
<td>1.109</td>
</tr>
<tr>
<td>Occupation status of the patient</td>
<td>.954</td>
<td>1.048</td>
</tr>
<tr>
<td>Cigarette taking</td>
<td>.974</td>
<td>1.027</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>.974</td>
<td>1.026</td>
</tr>
<tr>
<td>Cholesterol level of the patient</td>
<td>.955</td>
<td>1.047</td>
</tr>
<tr>
<td>The Family history of diabetes</td>
<td>.976</td>
<td>1.025</td>
</tr>
</tbody>
</table>

### h) Family history of diabetes

<table>
<thead>
<tr>
<th></th>
<th>Tolerance</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of patient</td>
<td>.876</td>
<td>1.142</td>
</tr>
<tr>
<td>Gender of patient</td>
<td>.907</td>
<td>1.103</td>
</tr>
<tr>
<td>Occupation status of the patient</td>
<td>.953</td>
<td>1.049</td>
</tr>
<tr>
<td>Cigarette taking</td>
<td>.978</td>
<td>1.023</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>.975</td>
<td>1.026</td>
</tr>
<tr>
<td>Cholesterol level of the patient</td>
<td>.941</td>
<td>1.063</td>
</tr>
<tr>
<td>The patient has hypertension</td>
<td>.929</td>
<td>1.076</td>
</tr>
</tbody>
</table>

The tables above (tables 4.1) reveal that we have no problem of multicollinearity among our independent variables, since in all cases, the VIF < 10. Thus we may proceed with all our independent variables to fit the multiple logistic model.
4.2. Checking for non-linearity between the dependent variable and independent variables and for non-normality of errors

The logistic regression does not assume a linear relationship between the dependents and the independents normally distributed error terms are not assumed. The following table is the output of SPSS 15.0 on the linearity between the dependent variable (diabetes outcome) and the independent variables (age, gender, occupation status, smoking, hypertension, alcohol consumption, cholesterol level and family history).

Table 4.2: Correlations between the variables in the model

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The Patient is diabetic</td>
<td>1.000</td>
<td>.337</td>
<td>.010</td>
<td>.262</td>
<td>.031</td>
<td>.167</td>
<td>.241</td>
<td>.307</td>
<td>.080</td>
</tr>
<tr>
<td>Age of patient</td>
<td>1.000</td>
<td>.277</td>
<td>.057</td>
<td>-.038</td>
<td>.043</td>
<td>.050</td>
<td>-.193</td>
<td>.052</td>
<td></td>
</tr>
<tr>
<td>Gender of patient</td>
<td>1.000</td>
<td>.028</td>
<td>-.060</td>
<td>.051</td>
<td>.104</td>
<td>.025</td>
<td>.107</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupation status of the patient</td>
<td>1.000</td>
<td>.066</td>
<td>.082</td>
<td>.170</td>
<td>.094</td>
<td>.081</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cigarette taking</td>
<td>1.000</td>
<td>-.103</td>
<td>.039</td>
<td>.035</td>
<td>.067</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>1.000</td>
<td>.077</td>
<td>.021</td>
<td>-.024</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cholesterol level of the patient</td>
<td>1.000</td>
<td>.137</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The patient has hypertension</td>
<td>1.000</td>
<td>.039</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Family history of diabetes</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From the table 4.2, it can be observed that the largest correlation coefficient is 0.337 which is recorded between age of patient and the dependent variable. This shows that there is no variable which exhibits any stronger linear association with the dependent variable.

Also, the following model Summary from SPSS 15.0: R=0.553; R Square =0.306 and Adjusted R Square = 0.273 leads to conclude that the multiple linear regression model cannot be proposed.
4.3. Testing for normality

The following are the Q-Q plots of standardized residuals from SPSS 15.0:
Figure 4. 3: Normal Probability plots of errors

It is observable from above Q- Q Plots of regression standardized residuals that errors are not normally distributed. Thus the logistic regression may be used instead of ordinary least squares regression.

4.3. Chi-square test of association between the dependent and independent variables.

Table 4. 3 : Test of independence of age versus diabetes

<table>
<thead>
<tr>
<th></th>
<th>Age of patient</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>From 20 to 39 years</td>
<td>From 40 to 59 years</td>
</tr>
<tr>
<td>The Patient is diabetic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Frequency</td>
<td>25</td>
<td>27</td>
</tr>
<tr>
<td>Expected Frequency</td>
<td>16.6</td>
<td>20.7</td>
</tr>
<tr>
<td>Yes Frequency</td>
<td>19</td>
<td>28</td>
</tr>
<tr>
<td>Expected Frequency</td>
<td>27.4</td>
<td>34.3</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>55</td>
</tr>
<tr>
<td>Frequency</td>
<td>44.0</td>
<td>55.0</td>
</tr>
</tbody>
</table>

H₀: Outcome of diabetes is not associated with age.
H$_1$: Outcome of diabetes is associated with age.

The Pearson Chi-square test statistic is:

$\chi^2_{calculated} = 21.909$ with 2 as degree of freedom and the p-value=$0.000 < 0.05$

Thus, we may conclude that there is statistical evidence of the association between age and the outcome of diabetes.

Table 4.4: Test of independence of Gender versus diabetes

<table>
<thead>
<tr>
<th>Gender of patient</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>40</td>
</tr>
<tr>
<td>Female</td>
<td>26</td>
</tr>
<tr>
<td>Total</td>
<td>66</td>
</tr>
</tbody>
</table>

H$_0$: Outcome of diabetes is not associated with gender.

H$_1$: Outcome of diabetes is associated with gender.

The Pearson Chi-square test statistic is:

$\chi^2_{calculated} = 0.016$ with 1 as degree of freedom and the p-value=$0.899 > 0.05$

Thus, we may conclude that there is no statistical evidence of the association between gender and the outcome of diabetes.

Table 4.5: Test of independence of occupation status versus diabetes

<table>
<thead>
<tr>
<th>Occupation status of the patient</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employed</td>
<td>24</td>
</tr>
<tr>
<td>Unemployed</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>66</td>
</tr>
</tbody>
</table>

H$_0$: Outcome of diabetes is not associated with occupation status.
H$_1$: Outcome of diabetes is associated with occupation status.

The Pearson Chi-square test statistic is:

\[ \chi^2_{calculated} = 11.981 \] with 1 as degree of freedom and the p-value=0.001 < 0.05

Thus, we may conclude that the outcome of diabetes is associated with the occupation status.

Table 4.6: Test of independence of smoking versus diabetes

<table>
<thead>
<tr>
<th>Smoking</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>The Patient is diabetic</td>
<td>Frequency</td>
</tr>
<tr>
<td>No</td>
<td>6</td>
</tr>
<tr>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
</tr>
</tbody>
</table>

H$_0$: Outcome of diabetes is not associated with smoking.

H$_1$: Outcome of diabetes is associated with smoking.

The Pearson Chi-square test statistic is:

\[ \chi^2_{calculated} = 0.171 \] with 1 as degree of freedom and the p-value=0.679 > 0.05

Thus, we may conclude that the outcome of diabetes is not associated with the smoking.

Table 4.7: Test of independence of alcohol consumption versus diabetes

<table>
<thead>
<tr>
<th>Alcohol consumption</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>The Patient is diabetic</td>
<td>Frequency</td>
</tr>
<tr>
<td>No</td>
<td>28</td>
</tr>
<tr>
<td>Yes</td>
<td>65</td>
</tr>
<tr>
<td>Total</td>
<td>93</td>
</tr>
</tbody>
</table>

H$_0$: Outcome of diabetes is not associated with alcohol consumption.

H$_1$: Outcome of diabetes is associated with alcohol consumption.
The Pearson Chi-square test statistic is:
\[ \chi^2_{calculated} = 4.889 \] with 1 as degree of freedom and the p-value=0.027 < 0.05
Thus, we may conclude that the outcome of diabetes is associated with the alcohol consumption.

Table 4. 8 : Test of independence of cholesterol level versus diabetes

<table>
<thead>
<tr>
<th></th>
<th>Cholesterol level of the patient</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>The Patient is diabetic</td>
<td>No Frequency</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Expected Frequency</td>
<td>38.1</td>
</tr>
<tr>
<td></td>
<td>Yes Frequency</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>Expected Frequency</td>
<td>62.9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>Frequency</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>Expected Frequency</td>
<td>101.0</td>
</tr>
</tbody>
</table>

H\(_0\): Outcome of diabetes is not associated with cholesterol level.

H\(_1\): Outcome of diabetes is associated with cholesterol level.

The Pearson Chi-square test statistic is:
\[ \chi^2_{calculated} = 10.151 \] with 1 as degree of freedom and the p-value=0.001 < 0.05
Thus, the conclusion is that the outcome of diabetes is associated with the cholesterol level.

Table 4. 9: Test of independence of hypertension versus diabetes

<table>
<thead>
<tr>
<th></th>
<th>The patient has hypertension</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>The Patient is diabetic</td>
<td>No Frequency</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Expected Frequency</td>
<td>43.4</td>
</tr>
<tr>
<td></td>
<td>Yes Frequency</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>Expected Frequency</td>
<td>71.6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>Frequency</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>Expected Frequency</td>
<td>115.0</td>
</tr>
</tbody>
</table>

H\(_0\): Outcome of diabetes is not associated with hypertension.
H$_1$: Outcome of diabetes is associated with Hypertension.

The Pearson Chi-square test statistic is:

\[ \chi^2_{calculated} = 16.525 \] with 1 as degree of freedom and the p-value = 0.000 < 0.05

Thus, we may conclude that the outcome of diabetes is associated with the hypertension.

Table 4. 10: Test of independence of family history of diabetes versus diabetes

<table>
<thead>
<tr>
<th>The Patient is diabetic</th>
<th>The Family history of diabetes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td>Frequency</td>
</tr>
<tr>
<td>No</td>
<td>14</td>
<td>52</td>
</tr>
<tr>
<td>Expected Frequency</td>
<td>17.0</td>
<td>49.0</td>
</tr>
<tr>
<td>Yes</td>
<td>Frequency</td>
<td>Frequency</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>78</td>
</tr>
<tr>
<td>Expected Frequency</td>
<td>28.0</td>
<td>81.0</td>
</tr>
<tr>
<td>Total</td>
<td>Frequency</td>
<td>Frequency</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>130</td>
</tr>
<tr>
<td>Expected Frequency</td>
<td>45.0</td>
<td>130.0</td>
</tr>
</tbody>
</table>

H$_0$: Outcome of diabetes is not associated with hypertension.

H$_1$: Outcome of diabetes is associated with Hypertension.

The Pearson Chi-square test statistic is:

\[ \chi^2_{calculated} = 1.124 \] with 1 as degree of freedom and the p-value = 0.289 > 0.05

Thus, we may conclude that the outcome of diabetes is not associated with the family history of diabetes.

4.4. MULTIPLE LOGISTIC REGRESSION MODEL FITTING.

4.4.1. The fitted model with all the predictor covariates

The data set used contains eight predictor variables: Age, Gender (Gen), Occupation status (Occ.Stat), Smoking (Smok), Alcohol Consumption (Alcoh), Cholesterol level (chol), Hypertension (Hyper) and Family history of diabetes (Famhist). The dependent variable remains the same: the diabetes status.

The model to fit is as follows:
For the diabetes, Let $Y_i$ be the binary outcome of diabetes (Yes or No) for the individual $i$. $Y_i \sim Bernoulli (\pi_i)$.

Now, the model will be:

$$\ln \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{gen} + \beta_3 \cdot \text{Occ.stat} + \beta_4 \cdot \text{smok}$$

$$+ \beta_5 \cdot \text{Alco} + \beta_6 \cdot \text{Chol} + \beta_7 \cdot \text{hyper} + \beta_8 \cdot \text{Famhist} + \varepsilon$$

(4.2)

Or

$$\pi = \frac{\exp(\beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{gen} + \beta_3 \cdot \text{Occ.stat} + \beta_4 \cdot \text{smok})}{1 + \exp(\beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{gen} + \beta_3 \cdot \text{Occ.stat} + \beta_4 \cdot \text{smok})}$$

(4.3)

SPSS software package is used to process the data. The maximum likelihood method is used to estimate the coefficients and its standard error. Table 4.11 shows the SPSS output.

Table 4.11: The Estimated coefficients, their S.E, and Wald test for the full model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>B</th>
<th>Std. Error</th>
<th>Wald</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-47.549</td>
<td>13.853</td>
<td>11.781</td>
<td>.001</td>
</tr>
<tr>
<td>Age of patient</td>
<td>1.142</td>
<td>.259</td>
<td>19.459</td>
<td>.000</td>
</tr>
<tr>
<td>Gender of patient</td>
<td>.143</td>
<td>.408</td>
<td>.123</td>
<td>.726</td>
</tr>
<tr>
<td>Occupation Status</td>
<td>-1.208</td>
<td>.397</td>
<td>9.252</td>
<td>.002</td>
</tr>
<tr>
<td>Smoking</td>
<td>-0.640</td>
<td>.678</td>
<td>0.891</td>
<td>.345</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>.818</td>
<td>.387</td>
<td>4.466</td>
<td>.035</td>
</tr>
<tr>
<td>Cholesterol level</td>
<td>.991</td>
<td>.394</td>
<td>6.324</td>
<td>.012</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1.028</td>
<td>.391</td>
<td>6.914</td>
<td>.009</td>
</tr>
<tr>
<td>Family history of diabetes</td>
<td>.482</td>
<td>.458</td>
<td>1.106</td>
<td>.293</td>
</tr>
</tbody>
</table>

The table 4.11 displays parameter estimates in the B column, the standard error and the Wald test.

Thus, using the estimates of the parameters in table 4.11, we get the following model:

$$\pi = \frac{\exp(-47.549+1.142 \cdot \text{age}+0.143 \cdot \text{gen} - 1.208 \cdot \text{Occ.stat} - 0.640 \cdot \text{smok} + 0.818 \cdot \text{Alco} + 0.991 \cdot \text{Chol} + 1.028 \cdot \text{Hyper} + 0.482 \cdot \text{Famhist} + \varepsilon)}{1 + \exp(-47.549+1.142 \cdot \text{age}+0.143 \cdot \text{gen} - 1.208 \cdot \text{Occ.stat} - 0.640 \cdot \text{smok} + 0.818 \cdot \text{Alco} + 0.991 \cdot \text{Chol} + 1.028 \cdot \text{Hyper} + 0.482 \cdot \text{Famhist} + \varepsilon)}$$

(4.4)
4.4.2. Testing for the significance of the individual parameters in the model.

To test the hypothesis:

\[ H_0 : \beta_j = 0 \quad (\text{for the individual parameter } \beta_j) \]

Versus

\[ H_1 : \beta_j \neq 0 \quad (\text{for the individual parameter } \beta_j) \]

Consider Wald and Sig. column of the table 4.11. The information given by the table reveals that the significant predictors are: Age (p-value = 0.000 < 0.05), Occupation status (p-value = 0.002 < 0.05), Alcohol consumption (p-value=0.035 < 0.05), Cholesterol level (p-value=0.012 < 0.05) and Hypertension (p-value=0.009 < 0.05).

On the other hand, the predictors which are not statistically significant are:

Gender (p-value = 0.726 > 0.05), smoking (p-value = 0.345 > 0.05) and family history of diabetes (p-value = 0.293 > 0.05).

4.4.3. Signs of coefficients analysis.

The sign of the coefficients of the estimated logistic function in Table 4.11 above gives an explanation of the explanatory variables used, as given in Table 4.12
Table 4.12: The sign analysis

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Codes</th>
<th>Sign</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of patient</td>
<td>Quantitative</td>
<td>Positive</td>
<td>Older age increases the probability of having diabetes.</td>
</tr>
<tr>
<td>Gender</td>
<td>1 Male 0 Female</td>
<td>Positive</td>
<td>Male increases the probability of having diabetes.</td>
</tr>
<tr>
<td>Occupation status</td>
<td>1 Employed 0 Unemployed</td>
<td>Negative</td>
<td>To be employed decreases the probability of having diabetes.</td>
</tr>
<tr>
<td>Smoking</td>
<td>1 No 0 Yes</td>
<td>Negative</td>
<td>Not Smoking decreases the probability of having diabetes.</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>1 Yes 0 No</td>
<td>Positive</td>
<td>Consumption of alcohol increases the probability of having diabetes.</td>
</tr>
<tr>
<td>Cholesterol level</td>
<td>1 High 0 Low</td>
<td>Positive</td>
<td>High cholesterol level increases the probability of having diabetes.</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1 Yes 0 No</td>
<td>Positive</td>
<td>Being hypertensive increases the probability of having diabetes.</td>
</tr>
<tr>
<td>Have a Family History of diabetes</td>
<td>1 Yes 0 No</td>
<td>Positive</td>
<td>Having a Family History of diabetes increases the probability of getting the disease.</td>
</tr>
</tbody>
</table>

4.4.4. The odds ratio results.

The Exp(B) column contains the exponential of parameter estimates. These values represent odds ratios for the corresponding predictor variables. In the table 4.13 bellow, the 95% Wald Confidence Limit shows the confidence interval (CI) for the odds ratio.

Table 4.13: Odds Ratios and 95% Confidence Intervals for Covariates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Exp(B)</th>
<th>95% Confidence Interval for Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Age of patient</td>
<td>3.133</td>
<td>.192</td>
</tr>
<tr>
<td>Gender</td>
<td>1.154</td>
<td>.519</td>
</tr>
<tr>
<td>Occupation status</td>
<td>0.299</td>
<td>1.537</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.527</td>
<td>.140</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>2.266</td>
<td>1.061</td>
</tr>
<tr>
<td>Cholesterol level</td>
<td>2.694</td>
<td>1.244</td>
</tr>
<tr>
<td>Hypertension</td>
<td>2.796</td>
<td>1.299</td>
</tr>
<tr>
<td>Have a Family History of diabetes</td>
<td>1.619</td>
<td>.660</td>
</tr>
</tbody>
</table>
From Table 4.13, it is evident that patients of older age, patients who consume alcohol, persons with high cholesterol level and hypertensive persons are highly susceptible for diabetes occurrence.

4.4.5. The full model assessment

Table 4.14: Likelihood ratio test

<table>
<thead>
<tr>
<th>MODEL</th>
<th>Model Fitting Criteria</th>
<th>Likelihood Ratio Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept only</td>
<td>-2 Log Likelihood</td>
<td>Chi-Square</td>
</tr>
<tr>
<td>Final</td>
<td>204.670</td>
<td>60.461</td>
</tr>
<tr>
<td></td>
<td>144.209</td>
<td></td>
</tr>
</tbody>
</table>

The table 4.14 displays the Likelihood Ratio test.

The -2 log likelihood for the constant only model obtain by fitting the constant only model is 204.670; and the -2 log likelihood for the overall model was 144.209.

Thus the value of the likelihood ratio test is:

\[ G = 204.670 - 144.209 = 60.461 \]

The null hypothesis is:

\[ H_0 : \beta_1 = \beta_2 = \ldots = \beta_8 = 0. \]
\[ H_1 : \exists \beta_j \neq 0, j = 1, 2, \ldots, 8 \]

The results show that at least one of the predictors ' regressions coefficient is not equal to zero because of the small p-values =0.000 which is less than 0.05. This would lead us to reject \( H_0 \) in favor of \( H_1 \) and we conclude that at least one and perhaps all beta's coefficient are different from zero.
Table 4.15: Classification table for the model with all predictor variables.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>Percent correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>43</td>
<td>23</td>
</tr>
<tr>
<td>Yes</td>
<td>16</td>
<td>93</td>
</tr>
<tr>
<td>Overall percentage</td>
<td>33.7%</td>
<td>66.3%</td>
</tr>
</tbody>
</table>

From table 4.15, we conclude that:

- 65.2% of all patients who do not have diabetes are correctly classified and 34.8% are incorrectly classified.

- 85.3% from all patients who have diabetes are correctly classified and 14.7% are incorrectly classified.

- The overall correct percentage was 77.7% which reflects the model’s overall explanatory strength.
By use of the ROC curve in figure 4.4 for the classification accuracy, it is found that the area under the ROC curve, which ranges from 0 to 1 provides the measure of the model’s ability to discriminate between those subject who experience the response of interest versus those who do not. The area under the ROC curve for the full model is 0.825 which may be considered as reasonable discrimination.

Table 4.16: Hosmer and Lemeshow Test

<table>
<thead>
<tr>
<th>Step</th>
<th>Chi-square</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.037</td>
<td>8</td>
<td>.533</td>
</tr>
</tbody>
</table>

By Hosmer and Lemeshow test, the table 4.16 gives the output from SPSS 15.0. Our Hosmer-Lemeshow statistic has a significance of 0.533 which means that it is not statistically significant.
and we fail to reject the null hypothesis that there is no difference between observed and model-predicted values, implying that the model’s estimates fit the data at an acceptable level.

4.4.6 The model with significant parameters only

The following step is the fitting of model with statistically significant parameters only (age, occupation status, alcohol consumption, cholesterol level and hypertension). Results are summarized in Table 4.17.

Table 4.17: Summarized results for the reduced model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>B</th>
<th>Std. Error</th>
<th>Wald</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95% Confidence Interval for Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>Intercept</td>
<td>-44.679</td>
<td>9.170</td>
<td>23.742</td>
<td>.000</td>
<td>2.971</td>
<td>.007</td>
</tr>
<tr>
<td>Age of patient</td>
<td>1.089</td>
<td>.249</td>
<td>19.217</td>
<td>.000</td>
<td>.207</td>
<td>.548</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>.825</td>
<td>.381</td>
<td>4.682</td>
<td>.030</td>
<td>1.081</td>
<td>4.818</td>
</tr>
<tr>
<td>Cholesterol level</td>
<td>.962</td>
<td>.386</td>
<td>6.223</td>
<td>.013</td>
<td>1.229</td>
<td>5.570</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1.043</td>
<td>.386</td>
<td>7.318</td>
<td>.007</td>
<td>1.333</td>
<td>6.041</td>
</tr>
</tbody>
</table>

From Table 4.17, the reduced model is written as follows:

$$\pi = \frac{\exp(-44.679+1.089\times age - 1.215\times Occ. Stat + 0.825\times Alcoh + 0.962\times chol + 1.043\times Hyper)}{1+\exp(-44.679+1.089\times age - 1.215\times Occ. Stat + 0.825\times Alcoh + 0.962\times chol + 1.043\times Hyper)}$$ (4.5)

The results above indicates that: patients with older age are more susceptible to develop diabetes; An employed person is less susceptible to develop diabetes; consuming alcohol increases the susceptibility; persons with high cholesterol level are more susceptible than those with low cholesterol level and hypertensive patients are more likely to develop diabetes than those who are not hypertensive.

The exponent (Exp (B)) in Table 4.17 is the odds ratio. Thus, for example:

- The odds for patients who consume alcohol to those patients who do not take it to develop diabetes is 2.282.
- The odds for patients with high cholesterol level to patients with low cholesterol level to develop the illness is 2.616.

- The odds for hypertensive person to that one who is not hypertensive to develop diabetes is 2.838.

Table 4. 18: Classification table for the reduced model

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The Patient is diabetic</td>
<td>No</td>
</tr>
<tr>
<td>The Patient is diabetic</td>
<td>No</td>
<td>41</td>
</tr>
<tr>
<td>Yes</td>
<td>11</td>
<td>98</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td>29.7%</td>
<td>64.6%</td>
</tr>
</tbody>
</table>

Table 4.18 gives the classification table. The information from the same table is that observations are classified as follows:

(i) 62.1% of all patients who do not have diabetes are correctly classified, and 37.9% are incorrectly classified.

(ii) 89.9% from all patients who have diabetes are correctly classified, 10.1% are incorrectly classified.

(iii) The overall correct percentage was 79.4%, which reflects the model's overall explanatory strength.

Plotting sensitivity versus (1–specificity) over all possible cut-points is shown in the Figure 4.5 below. The area under the ROC curve for the full model is 0.843 this is considered reasonable discrimination.
Comparing the two models (model with all predictor covariates and the reduced model), area under the ROC curve has become a particularly important measure for evaluating models’ performance because it is the average sensitivity over all possible specificities. The larger the area, the better the model performs. (Bradley, 1997).

We conclude that the reduced model (which has the area under the ROC curve of 0.843 and its overall explanatory strength is 79.4%) fits better the data than the model with all predictor variables (which has the area under the ROC curve of 0.825 and its overall explanatory strength is 77.7%).
CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.0. INTRODUCTION

This chapter talks of the summary of findings, conclusion and recommendation of the entire study.

5.1. SUMMARY ON FINDINGS OF THE STUDY

In this study, the risk factors of developing diabetes using logistic regression were studied. The risk factors used are: age, gender, occupation status, smoking, alcohol consumption, cholesterol level, hypertension, family history of diabetes. The study consisted of 93 people who were employed and 82 people who were unemployed. 115 people were hypertensive and 60 were not hypertensive. 93 people consumed alcohol and 82 did not, 14 were smokers and 161 were not, 45 people had a family history of diabetes whiles 130 people did not. There were 105 males and 70 females, while 101 people were with high cholesterol level and 74 with low cholesterol level. 109 people were diabetics while 66 were not. The test of association of diabetes with all the predictor variables (age, gender, smoking, occupation status, alcohol consumption, cholesterol level, hypertension, and family history of diabetes) has been done. It was found that age, alcohol consumption, cholesterol level, occupation status and hypertension were associated with the outcome of having diabetes. The predictors like gender, smoking, family history of diabetes had negligible association with having diabetes.

Considering age, the bar chart representation showed an increase in number of diabetics as the age increases. In the model with all predictor variables, the Wald test reveals that, on one hand, the significant predictors are: Age (p-value = 0.000), Occupation status (p-value = 0.002),
Alcohol consumption (p-value=0.035), Cholesterol level (p-value=0.012) and Hypertension (p-value=0.009 < 0.05). On the other hand, the predictors which are not statistically significant are: Gender (p-value = 0.726), smoking (p-value = 0.345) and family history of diabetes (p-value = 0.293). Concerning the sign of coefficients analysis, this study showed that older age, smoking, alcohol consumption, high cholesterol level and hypertension increases the risk of having diabetes, while being employed decreases the risk of having the illness.

From the odds ratio results, older age, patients who consume alcohol, persons with high cholesterol level and hypertensive persons are highly susceptible for diabetes occurrence.

In addition to the model with all independent variables, the model with only significant predictor variables was fitted. By comparing the two models, the conclusion is that the reduced model (which has the area under the ROC curve of 0.843 and its overall explanatory strength is 79.4% ) fits better the data than the model with all predictor variables (which has the area under the ROC curve of 0.825 and its overall explanatory strength is 77.7%).

5.2. CONCLUSION

In this study, risk factors of developing diabetes using logistic regression model were studied. The risk factors used are age, gender, occupation status, smoking, alcohol consumption, cholesterol level, hypertension, family history of diabetes. The binary logistic regression model is used to estimate the probability of having diabetes. Firstly, the chi-square test of association between diabetes and all the predictor variables showed that age, occupational status, alcohol consumption, cholesterol level and hypertension are statistically significant. Secondly, significance testing for the logistic coefficients using Wald test show that factors like age, occupational status, alcohol consumption, cholesterol level and hypertension are significant as predictor variables of diabetes. The model fitted showed that getting diabetes does not depend
significantly on the gender of a person, having a family history of diabetes and smoking. Instead, there is an increased risk of getting the diabetes as a person gets older. To assess the fitness of the model the maximum likelihood test and Hosmer and Lemeshow test were used.

5.3. RECOMMENDATIONS

Based on the findings from this study, the following recommendations were formulated in order to give our contribution in fighting the most disabling disease like diabetes in people:

- Use the fitted model to predict the likelihood of getting the disease. For example, a 35 years old person, who is employed, who consumes alcohol with high cholesterol level and who is not hypertensive has the probability:

$$\pi = \frac{\exp(-44.679+1.089+41-1.215+1 + 0.825+1 + 0.962+1 + 1.043+0)}{1+\exp(-44.679+1.089+41-1.215+1 + 0.825+1 + 0.962+1 + 1.043+0)} = 0.315$$ of getting diabetes.

The use of the model should be associated to other advices to reduce the risk of getting diabetes.

- Physical activities: these are necessary in everyday life. Many researchers have shown that the risk of diabetes and associated insulin resistance can be reduced significantly by trying to lose weight, especially for those who are severely obese (BMI > 35 kg/m²).

Physical exercises remain extremely important for control of the fats in the body and thus are important for the control of blood cholesterol, diabetes type 2, and cardiovascular disease which are the most dangerous consequences of obesity on health.

- The diet: this should also play a vital positive role on health. By eating sufficient fruits and vegetables, one gets access to several health benefits, due to an assumed complex interaction of containing biological active compounds.

The consumption of vegetables is likely to be a particular value due to their fibre content, low energy density and relative beneficial effect on blood sugar.
The joint WHO-FAO consultation on carbohydrates concluded that foods rich in slowly digested or resistant starch, or high in soluble fibre, might be protective against diabetes; while the people with a high intake of dietary fat, in particular saturated fats, are at increased risk of insulin resistance and type 2 diabetes.

In general, fruits and vegetables are characterized by a favored low energy density, a lack of cholesterol, a low fat content with beneficial fatty acid profile as well as a high content of vitamins, minerals, secondary plant compounds and fibres. A higher consumption of fruit and vegetables can be associated with reduced consumption of foods from animal origin and therefore for instance Saturated fatty acids.

In 2003, the WHO identified convincing associations for reduced risk of cardiovascular disease and diabetes and consumption of fruit and vegetables. (WHO, 2003)

- The data used for the study were secondary data from Gitwe Hospital in RUHANGO district, southern province. We therefore recommend that researchers who will want to do their research should find primary data from patients, because of the disadvantages associated with secondary data.

More follow up studies should be done to assess the benefits of different treatment modalities on control of cardiovascular risk factors such as blood pressure and lipids in diabetes patients to prevent further Cardio Vascular Diseases and other complications in Rwanda. Especially the focus should be on assessing the effect of the interventions based on healthy lifestyle such as increased physical activity, smoking cessation, weight loss and a healthy dietary pattern, and the rural area should be focused on.
REFERENCES


Chao, W. (2001). The Effects of Alcohol Consumption in People with Diabetes Mellitus:
A study based on a cohort of 7,418 patients seen at St Thomas’ Hospital between 1979 and 1998.


Fall (2013). Hints for health: Tips to keep you healthy. Healthcare Ministries, the international medical outreach of AGWM.


Available from http://www.diabetesatlas.org/content/economic-impactofdiabetes.


World Health Organization (WHO, 2014) . *Fact Sheet No.312: What is Diabetes? Available at:*


APPENDICES

Appendix 1: To whom it may concern letter

JOMO KENYATTA UNIVERSITY
OF
AGRICULTURE AND TECHNOLOGY
KIGALI CAMPUS

P.O. BOX 3373, KIGALI, RWANDA. Email: deputydir.kigalicampus@jkuat.ac.ke

TO WHOM IT MAY CONCERN

SUBJECT: SC384/C10/4000/13 NIYIKORA SYLVERE

This is to confirm that Niyikora Sylvere Registration number SC384/C10/4000/13 is a bona fide student pursuing a Master Degree in Applied Statistics our esteem institution currently in his second year semester one.

Any assistance accorded to his is highly appreciated

Yours Faithfully,

JOSEPH ODUOR
DEPUTY DIRECTOR

[Signature]

[Logo of Jomo Kenyatta University of Agriculture and Technology]

JCUAT IS ISO 9001:2008 CERTIFIED
Setting Trends in Higher Education, Research and Innovation
Appendix 2: Data accessing request letter written to Gitwe hospital

NIYIKORA SYLVERE  
Jomo Kenyata University of Agriculture and Technology (JCUAT/ Kigali Campus)  
Tel: 0783103638  

September 1, 2014

Doctor, Director of Gitwe Hospital  
RUHANGO/ SOUTHERN PROVINCE

Re: Request for permission of accessing to data needed in my masters’ research

Dear sir,

I humbly write this letter in order to request for the permission to access the data within Gitwe Hospital, necessary in my masters’ research.

In fact, I am student at Jomo Kenyata University of Agriculture and Technology in second year in Masters of Applied statistics. My research is oriented to the application of statistics, especially on modeling multiple logistic regression on risk factors of diabetes, and I would like to collect the necessary information from the free years folders of diabetic patients within Gitwe Hospital.

I, hopefully, Dear Sir, am waiting for your positive reaction to my request.

Yours faithfully,

NIYIKORA SYLVERE
Appendix 3: Permission to access to data letter

GITWE HOSPITAL

Gitwe 20th September 2014

To NIYIKORA SYLVERE
JKUAT/Kigali Campus

Re: PERMISSION TO ACCESS TO DATA

Dear Sir,

The Gitwe Hospital appreciates the opportunity to respond positively to your request in your letter written on September 1, 2014 to our institution. Our institution understands its interest on having access to files containing information on patients. That research is considered as necessary for the institution to fulfill its regulatory functions.

Yours Faithfully,

Administrator and Manager of Gitwe Hospital

TWAGIRAMUNGU Zacharie