KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI

MULTIPLE LOGISTIC REGRESSION ANALYSIS TO DETERMINE RISK FACTORS FOR THE CLINICAL DIAGNOSIS OF DIABETES

CASE STUDY: KOMFO ANOKYE TEACHING HOSPITAL (2008-2009)

By

Stella Appiah

A thesis submitted to the Department of Mathematics, Kwame Nkrumah University of Science and Technology in partial fulfillment for degree of Masters Of Philosophy.

KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI

MULTIPLE LOGISTIC REGRESSION ANALYSIS TO DETERMINE RISK FACTORS FOR THE CLINICAL DIAGNOSIS OF DIABETES

CASE STUDY: KOMFO ANOKYE TEACHING HOSPITAL (2008-2009)

By

Stella Appiah

A thesis submitted to the Department of Mathematics, Kwame Nkrumah University of Science and Technology in partial fulfillment for degree of Masters Of Philosophy

College Of Science

June, 2012

CANDIDATES DECLARATION

I hereby declare that this submission is a true account of my own work towards the MPhil degree and that it contains no materials previously published by any other person nor material which has been accepted for the award of any other degree in the university except where due acknowledgement has been made in the text.

STELLA APPAIH (PG3941009) ……
……..……

Student name and ID (Signature)
(Date)

Certified by

NANA KENA FREMPONG ……..……..
……

SUPERVISOR (Signature)
(Date)

Certified by

MR. K.F. DARKWAH ……
……..……

HEAD OF DEPARTMENT (Signature)
(Date)

DEDICATION

This thesis is dedicated to my mother, Mrs. Barbara Appiah and late father Mr. S.C Appiah.

It is also dedicated to my lovely husband, Mr. Samuel Agyekum and my unborn lovely kids.

ACKNOWLEDGEMENT

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

I will again extend my deepest gratitude to my lecturer and supervisor, Nana Kena Frempong for his directions, coordination and critics from the first to the last page of this work.

I wish to also show my sincere gratitude to my former Head of Department, Dr. Amponsah and my lecturers for their time and patience during the taught courses.

Again, I will say thank you to my mum, Mrs. Barbara Oduro Appiah for her commitment to my entire education.

My last appreciation goes to my husband for his understanding and support.

ABSTRACT

The study was conducted to find out if factors like age, gender, occupation status, marital status, smoking, alcohol consumption, hypertensive, contribute to the clinical diagnosis of diabetic.

The data used for the study was a secondary data from the physiotherapy department of the Komfo Anokye Teaching Hospital in Kumasi in the Ashanti region of Ghana. Two years data was sampled out, that was, 2008 and 2009. The software package used for the analysis of the data was SAS, version 9.1.

The preliminary results which involved an independent test of our outcome variable, diabetes, with all the predictor variables (alcohol, smoking, hypertension, cholesterol level, gender, family history, occupation, and marital status) showed that alcohol, hypertension, cholesterol level, and occupation were statistically significant at 5% to the outcome of diabetes. Marital status, smoking, family history and gender were not statistically significant to diabetes. A test of association between diabetes and alcohol with gender as a confounder showed that one can get diabetes when he/she consumes alcohol and this can be influenced by the gender of the person.

The results from the statistical package showed that occupation, alcohol, cholesterol level and age were statistically significant to our outcome, diabetes.

The odds ratio of occupation (0.497), alcohol (1.958), cholesterol level (2.259) and age (1.021) showed that the unemployed, those who take alcohol, and people with high cholesterol levels have a higher risk in getting diabetes. Age increase also increases one's risk to having diabetes.

TABLE OF CONTENTS

CHAPTER TWO: REVIEW OF LITERAURE

CHAPTER THREE: METHODOLOGY

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background of Study

1.1.1 Definition

The term *diabetes*, without qualification, usually refers to diabetes mellitus, which roughly translates to excessive sweet urine (known as "glycosuria"). Several rare conditions are also named diabetes. The most common of these is diabetes insipidus in which large amounts of urine are produced (polyuria), which is not sweet (insipidus meaning "without taste" in Latin).

Criteria for the diagnosis and classification of diabetes have been revised several times and can differ between countries. The current WHO classification (2003), is as follows:

- type 1 (pancreatic beta-cell destruction leading to absolute insulin deficiency),
- type 2 (insulin resistance and relative insulin deficiency)
- other specific types of diabetes
- gestational diabetes

Type 1 diabetes mellitus is characterized by loss of the insulin-producing beta cells of the islets of Langerhans in the pancreas leading to insulin deficiency. This type of diabetes can be further classified as immune-mediated or idiopathic. The majority of type 1 diabetes is of the immune-mediated nature, where beta cell loss is a T-cell mediated autoimmune attack. There is no known preventive measure against type 1 diabetes, which causes approximately 10% of diabetes mellitus cases in North America

and Europe. Most affected people are otherwise healthy and of a healthy weight when onset occurs. Sensitivity and responsiveness to insulin are usually normal, especially in the early stages. Type 1 diabetes can affect children or adults but was traditionally termed "juvenile diabetes" because it represents a majority of the diabetes cases in children.

Type 2 diabetes mellitus is characterized by insulin resistance which may be combined with relatively reduced insulin secretion. The defective responsiveness of body tissues to insulin is believed to involve the insulin receptor. However, the specific defects are not known. Diabetes mellitus due to a known defect are classified separately. Type 2 diabetes is the most common type.

In the early stage of type 2 diabetes, the predominant abnormality is reduced insulin sensitivity. At this stage hyperglycemia can be reversed by a variety of measures and medications that improve insulin sensitivity or reduce glucose production by the liver.

Gestational diabetes mellitus (GDM) resembles type 2 diabetes in several respects, involving a combination of relatively inadequate insulin secretion and responsiveness. It occurs in about 2%–5% of all pregnancies and may improve or disappear after delivery. Gestational diabetes is fully treatable but requires careful medical supervision throughout the pregnancy. About 20%–50% of affected women develop type 2 diabetes later in life.

The following tests are used for clinical diagnosis of diabetes:

- A **fasting plasma glucose (FPG) 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.
- An **oral glucose tolerance test (OGTT)** 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.
- 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.

1.1.2 Signs and symptoms

The classical symptoms of diabetes are polyuria (frequent urination), polydipsia (increased thirst) and polyphagia (increased hunger). Symptoms may develop rapidly (weeks or months) in type 1 diabetes while in type 2 diabetes they usually develop much more slowly and may be subtle or absent. Prolonged high blood glucose causes glucose absorption, which leads to changes in the shape of the lenses of the eyes, resulting in vision changes; sustained sensible glucose control usually returns the lens to its original shape. Blurred vision is a common complaint leading to a diabetes diagnosis; type 1

should always be suspected in cases of rapid vision change, whereas with type 2 change is generally more gradual, but should still be suspected.

People (usually with type 1 diabetes) may also present with diabetic ketoacidosis, a state of metabolic dysregulation characterized by the smell of acetone; a rapid, deep breathing known as Kussmaul breathing; nausea; vomiting and abdominal pain; and an altered states of consciousness. A rarer but equally severe possibility is hyperosmolar nonketotic state, which is more common in type 2 diabetes and is mainly the result of dehydration. Often, the patient has been drinking extreme amounts of sugar-containing drinks, leading to a vicious circle in regard to the water loss.

A number of skin rashes can occur in diabetes that is collectively known as diabetic dermadromes.

1.1.3 Causes

The cause of diabetes depends on the type. Type 2 diabetes is due primarily to lifestyle factors and genetics. Type 1 diabetes is also partly inherited and then triggered by certain infections, with some evidence pointing at Coxsackie B4 virus. There is a genetic element in individual susceptibility t o some of these triggers which has been traced to particular HLA genotypes (i.e., the genetic "self" identifiers relied upon by the immune system). However, even in those who have inherited the susceptibility, type 1 diabetes mellitus seems to require an environmental trigger.

1.2 Statement of the problem

Population is ageing and has become important development issue that requires urgent action.

The prevalence of diabetes rises with age, approximately 30% of the populations between age 60-65 and 40% of the 65 above people have diabetes. By comparism 11% of individuals between ages 35 and 45 are diagnose of diabetes. Fries (1980).

Diabetes is an important condition that poses a huge health care burden because of its associated high mortality and morbidity rate. The disease is the most common complication among Europeans and accounts for at least 60% of deaths and there is evidence that the mortality for pre-diabetes state such as Impaired Glucose Tolerance (IGT) is comparable to frank diabetes.

This overwhelming burden of disease suggests:

1. That if factors like age, gender, hypertension, occupation, smoking, alcohol consumption, family history and marital status are significant positive risk factors for the development of diabetes then critical attention would be given to these factors in other to avoid getting the disease.
2. That early detection of pre-diabetes and diabetes and treatment might reduce morbidity and mortality.

1.3 Objectives

The general objectives of this study are:

1. To find out if risk factors like age, gender, marital status, hypertension, smoking and family history, alcohol consumption, and occupation status are associated with having diabetes.
2. Fit a logistic model that determines the factors with high risk compared to others.

1.4 Methodology

The study would have considered all the main hospitals in the country but Komfo Anokye Teaching Hospital was sampled out for the research.The main data that was used for the study was a secondary data. This data was taken from the physiotherapy department of the Komfo Anokye Teaching Hospital. Two years data was sampled out for the research. (2008 and 2009)

The researcher visited the physiotherapy department of the Komfo Anokye Teaching Hospital in Kumasi to take the following information from the folders of Cerebral vascular Accident (CVA) and diabetes patients. Age, gender, occupation marital status, smoking, alcohol consumption, hypertensive, family history and diabetic.

The software package used for the analysis was SAS version 9.1.

1.5 Significance of the Study

Sedentary life style and lack of physical activity are associated with higher level of risk of morbidity and mortality in Ghana and in the World. Early diagnosis and treatment is an important step in the secondary prevention of disease and disability. Similarily, the regular screening can assess prevalence and actual health status of the community.

In Ghana, many studies about communicable disease are carried out and only few attempts are made for the study of non-communicable diseases.

Diabetes mellitus is a common health problem in old age. However, 50% of elderly diabetics are undiagnosed and more than 50% of all diabetic are over the age of 60. Older diabetics with cardiovascular and micro vascular complications and disease unrelated to diabetes burden the hospital than the non diabetic population. Aro et al. (1994).

Most studies focus mainly on different independent variables like demographic, socio-economic and body mass index as significantly associated with diabetes. Some of the studies in general population have explored the common factors like age, sex, and diet and body mass index.

This study proves to be an important because;

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

Secondly, 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 take in food and do things that will prevent our blood pressure from rising up.

Lastly, the findings will help people without the disease to know his/her 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;

1. Inadequate finance which would prevent us from visiting other hospitals outside Kumasi.

2. More people to help in the collection of the data and carrying out of the research.

1.7 Organisation of study

The research was categorised under five chapters as follows:

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

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

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

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

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

CHAPTER TWO

LITERATURE REVIEW

2.0 INTRODUCTION

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

2.0.1 Diabetes

Diabetes mellitus is defined as a group of metabolic diseases whose common feature is an elevated blood glucose level (hyperglycemia). Chronic hyperglycemia is associated with the long-term consequences of diabetes that include damage and dysfunction of the cardiovascular system, eyes, kidneys and nerves.

Diabetes mellitus (DM) was estimated to be the 29th leading cause of burden of disease in the world in 1990, accounting for 1.1% of total years lived with disability (YLD), around the same percentages respiratory infections or malignant neoplasms.

In the Version 1 estimates for the Global Burden of Disease (GBD) 2000 study, published in the World Health Report 2001, DM is the 20th leading cause of YLDs at global level, accounting for

1.4% of total global YLDs.

Criteria for the diagnosis and classification of diabetes have been revised several times and can differ between countries. The current WHO classification of disorders of glycaemia is as follows:

· type 1 (pancreatic beta-cell destruction leading to absolute insulin deficiency),

· type 2 (insulin resistance and relative insulin deficiency)

· other specific types of diabetes

· gestational diabetes

The onset of type 1 diabetes is most common in children or young adults and accounts for around 10% or less of the total number of people with diabetes. Type 2 diabetes accounts for almost all of the remaining cases of diabetes as the other forms are rare. Type 2 diabetes is a condition that predominantly affects middle-aged and older people but prevalence is increasing among children and young adults in countries with a high prevalence of obesity.

Diabetes was considered as a problem of developed countries in the past, but now people of developing countries are also at equal risk of diabetes and migrants may often be more likely to develop diabetes than non migrant, Carballo and Frederik (2006).

Again, from the work of Cockram (2000), diabetes was recognized as a major global epidemic, with the prevalence of the disease rapidly increasing in many developed and/or developing Asian countries.

Both diabetes and hypertension are independent risk factors for cardiovascular disease which indicates that the co-existence of these conditions in a patient imposes a need for a significant blood pressure control (135/80mmHg) than the goal blood pressure recommend for patient who does not have diabetes (140/90mmHg).

Apart from the issue of high blood pressure as a risk factor for diabetes, Haffner and Heinrich (2003), also found that the most essential risk factors are obesity weight gain and sedentary life style.

Keil et al. (2001) and Kolenda and Muller (2005), grouped the risk factors of diabetes into non modifiable and lifestyle associated factors respectively.

Increased age, male sex and a positive family history count to the non modifiable Cardiovascular risk factors whiles smoking, overweight and obesity, diabetes mellitus, hypertension, physical inactivity, Hypercholesterolemia as well as hypertriglyceridemia belongs to the lifestyle associated risk factors.

2.1 LIFESTYLE ASSOCIATED RISK FACTORS

2.1.1 High Cholesterol level

The number one risk factor for type 2 diabetes is high cholesterol or increase in fat in the body. The American National Center for Health Statistics (2001), states that 30% of adults have their cholesterol level high and so are overweight. That is, repesenting 60 million people. Greater weight means a higher risk of insulin resistance, because fat interferes with the body's ability to use insulin. The same study, also saw that the number of overweight kids has tripled since 1980 and so children who are being diagnosed with type 2 diabetes has risen.

Hippocrates (2005), recognized that sudden death from diabetes is more common in those who are naturally fat than in the lean.

Bray (2004), compared whites and blacks on cholesterol levels. He suggested that control of overweight or high cholesterol level would eliminate 48% of the diabetes in whites and 28% in blacks.

Again, 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 five fold 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). Mufunda 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. (2008).

2.1.2 High Blood Pressure/ Hypertension

Blood pressure plays an important part in the management of diabetes. High blood pressure (hypertension) adds to the workload of the heart, arteries and kidneys. Damage to kidneys, eyes and feet are long-term complications that can go along with a diagnosis of diabetes. Other health risks include heart disease and strokes.

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.

It is very important to refer that the WHO MONICA Project (2004), sample mainly represents populations 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.

2.1.3 Smoking

Cigarette smoking was first pointed out as a risk factor for diabetes in men in the late 1980s and later confirmed in large cohort studies by Monica Augsburg (2001), in both men and women and most prominently amongst heavy smokers. A recent paper summarizes the 15 published prospective epidemiologic studies on this subject.

Recently, a large Swedish cross-sectional study showed that the prevalence of type 2 diabetes, diagnosed by an oral glucose tolerance test (OGTT), was equally increased for smokers and snus users with high tobacco consumption compared with non tobacco users.

2.1.4 Alcohol

A study by the American Diabetes Association (2009), showed that 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.

The exact nature of the dose-response relationship remains unclear. Several reviews have suggested a U-shaped relationship or a protective effect of moderate consumption with some question about the effect of higher levels of alcohol consumption. However, these reviews are narrative. Two quantitative reviews have been conducted. Carlsson et al. (2001) categorized consumption into predetermined moderate- and high-consumption groups and used current abstainers or low consumers as the reference group. In their analysis, moderate consumption was associated with a 30% reduced risk of diabetes among men (relative risk [RR] 0.72) and women (0.68).

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.

2.2 NON MODIFIEDABLE RISK FACTORS

2.2.1 Increased Age

It's a sad but true fact. The older we get, the greater our risk of type 2 diabetes. Even if an elderly person is thin, they still may be predisposed to getting diabetes. Scientists theorize that the pancreas ages right along with us, and doesn't pump insulin as efficiently as it did when we were younger. Also, as our cells age, they become more resistant to insulin as well. About.com (2004).

Again, the increase in prevalence of the disease has accelerated still due to aging population structures in developed countries and increasing obesity globally. High prevalence occurs in the Caribbean, West Africa, and among some communities in Britain, as in individuals of Indian sub continental origin, with heterogeneity in subgroups. Whether these occur as a result of genetic or environmental factors remains unclear; the case against a major genetic contribution has been made in other studies.

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.

An estimated 246 million people (6% of adults) worldwide had diabetes in 2007. The proportion is expected to rise to 7% by 2025. About 80% of the 246 million people with diabetes live in developing countries. An additional 7.5% of the adult population worldwide is estimated to have impaired glucose tolerance, which significantly increases the risk of developing type 2 diabetes. Tesfaye (2008).

Finally, l 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.

2.2.2 Sex/Gender

To our knowledge, the MONICA Augsburg Cohort Study (2007), is the first prospective population-based study to assess the sex-specific incidence of diabetes mellitus in a middle European population characterized by a relatively low risk of cardiovascular morbidity and mortality. The Augsburg study identified age, BMI, a parental history of diabetes, and low HDL cholesterol values as independent determinants of diabetes mellitus in both sexes. High systolic blood pressure, regular cigarette smoking, and high daily alcohol intake predicted diabetes in men only, whereas high uric acid values and physical inactivity during leisure time were associated with a higher risk of diabetes in women only. In this same study, the Framingham Study and the Finnmark Study, they demonstrated a strong positive association between BMI and diabetes in both sexes. However, in the Finnmark Study, the effect of a high BMI was much stronger in men. Concerning the predictive importance of low HDL cholesterol levels, contradictory results were reported from the Framingham Study, which showed significant effects in men only, and from the Finnmark Study, which demonstrated a significant predictive relevance in women only.

Again, the same study showed that systolic blood pressure was positively associated with diabetes in men, but not in women. Similar results have been observed in other prospective cohort studies in men. In the Framingham Study, systolic blood pressure was

associated with the incidence of diabetes in the univariate analysis, but this effect lost significance after multiple adjustment in both sexes. In contrast to systolic blood pressure, in the present study, actual hypertension was strongly associated with diabetes in men and women. Thus, the low impact of systolic blood pressure in women might be attributed to the fact that more women with actual hypertension had controlled blood pressure values in comparison with men.

Additionally, in accordance with the Rancho Bernardo Cohort (2007), a significant association between heavy alcohol intake and incident diabetes in men was observed in the Augsburg study. Other prospective studies suggested an inverse association between alcohol consumption and risk of type 2 diabetes mellitus in men and women or reported no significant impact on disease development in men and women. The apparent lack of association between alcohol intake and type 2 diabetes in women may be due to the fact that there were too few heavy drinkers among women or the fact that women consumed wine more often than men did in the present study.

Lastly in Monica's cohort study, physical inactivity during leisure time was associated with an 80% increased risk of type 2 diabetes only in women. In general, a positive association between physical inactivity and type 2 diabetes independent of obesity has been noted in many prospective studies in men and women. 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 = .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 = .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.

2.2.3 Family History and Genetics

It appears 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.

Also, in accordance with present study, several prospective studies conducted in men—the Honolulu Heart Program (2007), the Pennsylvania Alumni Health Study, (2007) also observed a high impact of parental history on diabetes incidence. No prospective studies investigating the association between a parental history of diabetes and the incidence in their female offspring were found in the literature; thus, the Augsburg results of an approximately 2.5 times higher risk of incident diabetes in women with a positive parental history of diabetes compared with women without such parental history have to be confirmed by further studies.

2.2.4 Socioeconomic status

Usually the socioeconomic status (SES) is measured by education, income and status or all three levels combined. Nocon et al. (2007). In the assessment of socioeconomic disparities within a population, education has been shown to have the most important impact caused by the fact that educational attainment might also affect occupation and income and can influence a person's access and understanding of nutritional information. Trichopoulo et al. (2002). In general, a higher socioeconomic level has been associated with a healthier dietary behaviour.

Thus, in a meta-analysis a higher socioeconomic level was associated with higher intakes of fruit and vegetables. Irala-Estevez et al,(2000). Moreover, individuals of lower SES reported to consume more total fat, saturated fat, meat and meat products compared to their higher social class counterparts. Hulshof et al. (2003) and Lopez-azpiazu et al. (2003). In the *EPIC-Study*, participating European citizens who were higher educated had lower intake values of added fats and oils. Linseisen et al. (2002).

Furthermore, lower SES examined across educational level and annual income has been attributed to increased prevalence of cardiovascular risk factors such as hypertension, obesity, diabetes mellitus and hypercholesterolemia. Panagiotakos et al. (2008). Additionally, an inverse relationship between SES and smoking and physical inactivity has been observed. In summary, SES is linked to a variety of health determinants, thus influencing the development of cardiovascular events. Nocon et al. (2007).

CHAPTER THREE

METHODOLOGY

3.0 INTRODUCTION

The study would have considered all the main hospitals in the country but Komfo Anokye Teaching Hospital was sampled out for the research.

The data used for the study was secondary data. This was taken from the physiotherapy department of the Komfo Anokye Teaching Hospital. Two years data was sampled out, that is, 2008 and 2009.

The researcher visited the physiotherapy department of the Komfo Anokye Teaching Hospital in Kumasi to take the following information from the folders of patients with Cerebral Vascular Accident (CVA) and diabetes. Age, gender, occupation marital status, smoking, alcohol consumption, family history, hypertension and diabetes.

The software package used for the data analysis was SAS version 9.1.

3.1 Categorical variable

These are variables that places individuals literally into categories and cannot be quantified in a meaningful way. Example would be diseases like diabetes, occupation, cholesterol level, gender etc.

Analysis of categorical data involves the use of data tables. A *two-way* table present categorical data by counting the number of observations that fall into each group for two variables one divided into rows and the other into columns.

3.1.1 CHI-SQUARE TEST FOR INDEPENDENCE

TABLE 3.1: I×J CONTINGENCY TABLE

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

From table 1.1, *I* is the number of levels for one categorical variable, and *J* is the number of levels for the other categorical variable.

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

O_{ij} is the observed frequency count at level O_{i+} of Variable *I* and level O_{+j} of Variable *J*.

N is the total sample size.

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

For example, in this thesis, disease might be classified by diabetic and not diabetic and would be associated with occupation which would be classified as employed and unemployed. We would use a chi-square test for independence to determine whether occupation is associated with the outcome of having diabetes. Other independent variables that may be associated with diabetes are gender, hypertension, cholesterol level, marital status, smoking, alcohol and family history.

Age would not be used in this test because it is a continuous variable and not categorical.

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

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

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

3.1.2 State the Hypotheses

Suppose that from the table 3.1 Variable J has 0_{+j} levels, and Variable I has 0_{i+} levels. The null hypothesis states that knowing the level of Variable J does not help you predict the level of Variable I. That is, the variables are independent.

H_0: Variable J and Variable I are independent.

H_a: Variable J and Variable I are not independent.

The alternative hypothesis is that knowing the level of Variable J can help you predict the level of Variable I.

3.1.3 Analyze Sample Data

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

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

  $$DF = (I - 1) \times (J - 1) \quad \textbf{(3.1)}$$

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

- Expected frequencies. The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute $O_{i+} \times O_{+j}$ expected frequencies, according to the following formula.

  $$E_{ij} = \frac{O_{i+} \times O_{+j}}{N} \quad \textbf{(3.2)}$$

In equation (32.), E_{ij} is the expected frequency count for level O_{i+} of Variable I and level O_{+j} of Variable J.

- Test statistic. The test statistic is a chi-square random variable (χ^2) defined by the following equation.

$$\chi^2 = \sum_i^I \sum_j^J \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)} \quad \textbf{(3.3)}$$

In equation (3.3), $\boldsymbol{O_{ij}}$ is the observed frequency count at level $\boldsymbol{O_{i+}}$ of Variable J and level $\boldsymbol{O_{+j}}$ of Variable I and $\boldsymbol{E_{ij}}$ is the expected frequency count at level $\boldsymbol{O_{i+}}$ of Variable J and level $\boldsymbol{O_{+j}}$ of Variable I.

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

3.1.4 Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

3.2 LOGISTIC REGRESSION

3.2.1 Introduction

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. It is often the case that the outcome variable is discrete, taking on two or more possible values. Over the last decade the logistic regression model has become, in many fields, the standard method of analysis in this situation. Hosmer and Lemeshow, (2000).

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

Example 1.0

Table 1.2 lists the age categories in years and presence and absence of evidence of diabetes for the 225 subjects whose folders were selected for the study. The outcome variable is diabetes which is coded with a value of 0, to indicate diabetes is absent, or 1 to indicate that the individual is diabetic. The table 1.2 also contains, for each age group, the frequency of occurrence of each outcome as well as the mean (or proportion with diabetes present) for each group.

By examining the table, a clearer picture of the relationship begins to emerge. It appears that as age increases, the proportion of individuals with evidence of diabetes increases. In any regression problem the key quantity is the mean value of the outcome variable, given the value of the independent variable. This quantity is called the *conditional mean* and is expressed as "*E(Y/x)*" where *Y* denotes the outcome variable and *x* denotes a value of the independent variable. The quantity *E(Y/x)* is read "the expected value of *Y*, given the value *x*."

The column labeled "Mean" in Table 1.2 provides an estimate of *E(Y/x)*. With dichotomous data, the conditional mean must be greater than or equal to zero and less than or equal to 1 [i.e., $0 \leq E(Y/x) \leq 1$]. This can be seen in Figure 1.0.

In addition, the plot shows that this mean approaches zero and 1 "gradually." The change in the *E(Y/x)* per unit change in *x* becomes progressively smaller as the conditional mean gets closer to zero or 1. The curve presented in the graph is said to be *S – shaped*. The model we would use is that of the logistic regression model.

Table 3.2: Age and diabetes status of 225 patients from data.

AGE	No. of subjects	Diabetic	Non diabetic	Mean (proportion)
20 – 29	20	2	18	0.10
30 – 35	25	4	21	0.16
36 – 39	22	6	16	0.27
40 – 45	25	10	15	0.40
45 – 49	23	12	11	0.52
50 – 54	18	10	8	0.55
55 – 59	27	17	10	0.63
60 – 69	20	14	6	0.70
70 – 79	30	25	5	0.83
80 – 89	15	13	2	0.87

The choice of age categories are based on literature from Homers and Lemeshow (2000).

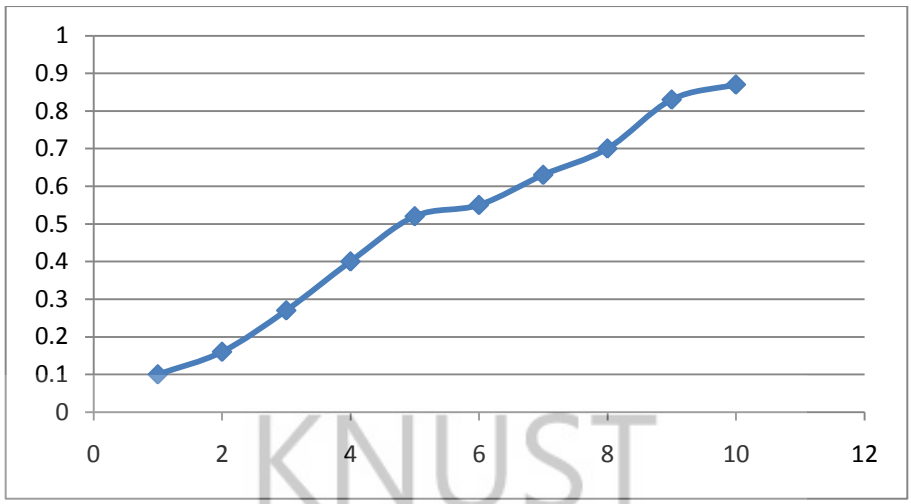

Figure 1.0. Plot of the percentage of 225 patients with diabetes in each age group.

3.3 Generalized Linear Models (GLMs) and Logistic Regression.

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

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

Systematic Component - refers to the explanatory variables (x_1, x_2,...,x_k) as a combination of linear predictors; e.g. $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ as we have seen in logistic regression.

Link Function, η or $g(\mu)$ - specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables;

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

3.3.1 Binary Logistic Regression

Models how binary response variable depends on a set of explanatory variable

Random component: The distribution of Y is *Binomial*

Systematic component: *X*s are explanatory variables (can be continuous, discrete, or both) and are linear in the parameters $\beta_0 + \beta x_i + ... + \beta_0 + \beta x_k$

Link function: *Logit*

$$\eta = \text{Logit}(\pi) = \text{Log}\left(\frac{\pi}{1-\pi}\right) \quad \textbf{(3.4)}$$

3.4 Logistic Regression with single independent variable.

The general formular for the logistic regression model with single variable is;

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

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

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

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

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

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

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

$$g(x) = \beta_0 + \beta_1 x \qquad \textbf{(3.7)}$$

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

3.5 Fitting the Single Logistic Regression Model

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

In order to apply this method we must first construct a function, called the *likelihood function.* This function expresses the probability of the observed data as a function of the unknown parameters. The *maximum likelihood estimators* of these parameters are chosen to be those values that maximize this function. Thus, the resulting estimators are those which agree most closely with the observed data. We now describe how to find these values from the logistic regression model.

If Y is coded as 0 or 1 then the expression π (x) given in equation (3.0) provides (for an arbitrary value of $\boldsymbol{\beta} = (\beta_0, \beta_1)$, the vector of parameters) the conditional probability that

Y is equal to 1 given x. This will be denoted as P(Y = 1 | x). It follows that the quantity 1 - π (x) gives the conditional probability that Y is equal to zero given x, P(Y = 0 | x). Thus, for those pairs (x_i, y_i), where $y_i = 1$, the contribution to the likelihood function is $\pi(x_i)$ and for those pairs where y_i=0, the contribution to the likelihood function is 1-$\pi(x_i)$, where the quantity $\pi(x_i)$ denotes the value of π (x) computed at x_i.

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

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \quad \textbf{(3.8)}$$

Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in equation (3.8) as

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \quad \textbf{(3.9)}$$

The principle of maximum likelihood states that we use as our estimate of $\boldsymbol{\beta}$ the value which maximizes the expression in equation (3.8) However, we will work with the log of equation (3.9). This expression, the *log likelihood* is given as:

$$L(\boldsymbol{\beta}) = \text{In}[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \left\{ y_i \text{In}[\pi(x_i)] + (1-y_i)\text{In}[1- \pi(x_i)] \right\} \quad \textbf{(3.10)}$$

To find the value of $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$ we differentiate $L(\boldsymbol{\beta})$ with respect to β_0, β_1 partially and set the resulting expressions equal to zero.

These equations, known as the *likelihood equations*, are;

$$\sum[\ y_i - \pi(x_i)] = 0 \qquad \textbf{(3.11)}$$

and

$$\sum \mathrm{x_i}[\ y_i - \pi(x_i)] = 0 \qquad \textbf{(3.12)}$$

3.6 Testing for the significance of the single independent variable

In logistic regression, comparison of observed to predicted values is based on the log likelihood function defined in equation (3.10).

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

$$D = -2\mathrm{In}\left[\frac{(\text{likelihood of the fitted model})}{(\text{likelihood of the saturated model})}\right] \qquad \textbf{(3.13)}$$

The quantity inside the large brackets in the expression above is called the **likelihood ratio**. Such a test is called the **likelihood ratio test.** A saturated model is one that contains as many parameters as there are data points. Using equation (3.10) and (3.13) becomes

$$D = -2\sum_{i=1}^{n}\left[y_i\ \mathrm{In}\ \left(\frac{\pi_i}{y_i}\right) + (1 - y_i)\ \mathrm{In}\ \left(\frac{1-\pi_i}{1-y_i}\right)\right] \qquad \textbf{(3.14)}$$

From equation (3.14), $\pi_i = \pi(x_i)$**.** The statistic, *D*, in the equation is called the *deviance.* This plays the same role as the residual sum of square plays in linear regression. It is identically equal to the sum of square error (SSE).

In an instance where the values of our outcome variable are 0 and 1 just as in this study, the likelihood of our saturated model is 1. Specifically it follows from the definition of a saturated model that $\pi_i = y_i$ and the likelihood is

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

Thus, it follows from equation (3.13) that the deviance is

$$D = -2\text{In (likelihood of fitted model)} \qquad \textbf{(3.15)}$$

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

$$G = D(\text{model without the variables}) - D(\text{model with the variables}) \qquad \textbf{(3.16)}$$

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

$$G = -2\text{In}\left[\frac{(\text{likelihood without the variable})}{(\text{likelihood with the variable})}\right] \qquad \textbf{(3.17)}$$

For cases of a multiple independent variable, it is easy to show when the variables are not in the model, the maximum likelihood estimate of β_0 is $\text{In}(n_1/n_0)$ where $n_1 = \sum y_i$

and $n_0 = \sum(y_i - 1)$ and the predicted value is constant, n_1/n. In this case, the value of G is;

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n_0}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\left(\hat{\pi}_i\right)^{y_i}\left(1-\hat{\pi}_i\right)^{\left(1-y_i\right)}}\right] \qquad \textbf{(3.18)}$$

Or

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i \ln(\hat{\pi}_i) + (1+y_i)\ln(1-\hat{\pi}_i)\right] - \left[n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n)\right]\right\} \qquad \textbf{(3.19)}$$

Under the hypothesis that $\boldsymbol{\beta_1}$ is equal to zero, the statistic G follows a chi-square distribution with 1 degree of freedom.

Two other similar, statistically equivalent tests known is the Wald test and the Score test. The assumption needed for these test are the same as those of the likelihood ratio test in equation (3.18).

The Wald test is obtained by comparing the maximum likelihood estimate of the slope parameter, $\hat{\beta}_1$, to an estimate of its standard error. The resulting ratio, under the hypothesis that $\beta_1 = 0$, will follow a standard normal distribution.

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \qquad \textbf{(3.20)}$$

Another test use in testing for the significance of a variable is the Score test. This test is based on the distribution theory of the derivatives of the log likelihood. The test statistic for the Score test (ST) is

$$\mathbf{S}T = \frac{\sum_{i=1}^{n} x_i\left(y_i - \bar{y}_i\right)}{\sqrt{\left(y_i - \bar{y}_i\right)\sum_{i=1}^{n}\left(x_i - \bar{x}_i\right)}} \qquad \textbf{(3.21)}$$

3.7 Confidence Interval Estimation of Single Logistic Regression Variable

The confidence interval estimators for slope and intercept are based on their respective Wald tests. The endpoints of a 100(1 – α)% confidence interval for the slope coefficient are

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\mathrm{SE}(\hat{\beta}_1) \qquad \textbf{(3.22)}$$

and for the intercept they are

$$\hat{\beta}_0 \pm z_{1-\alpha/2}\mathrm{SE}(\hat{\beta}_0) \qquad \textbf{(3.23)}$$

In equation (3.23), $z_{1-\alpha/2}$ is the upper 100(1- α/2)% point from the standard normal distribution and SE(.) denotes a model-based estimator of the standard error of the respective parameter estimator.

The estimated values are provided in the output following the fit of a model and, in addition, many statistical software packages provide the endpoints of the interval estimates.

The standard error is calculated using the logit of the linear part of the logistic regression model and, as such, is most like the fitted line in a linear regression model. The estimator of the logit is;

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \textbf{(3.24)}$$

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

$$\hat{\text{Var}}[\hat{g}(x)] = \text{Var}(\hat{\beta}_0) + x^2 \text{Var}(\hat{\beta}_1) + 2x \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \qquad \textbf{(3.25)}$$

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

$$\hat{g}(x) \pm z_{1-\alpha/2} \text{SE}[\hat{g}(x)] \qquad \textbf{(3.26)}$$

where $\text{SE}[\hat{g}(x)]$ is the positive square root of the variance estimator in (3.24).

3.8 The Multiple Logistic Regression Model

The general form of the multiple logistic regression model is;

$$\pi(\mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}} \qquad \textbf{(3.27)}$$

From equation (3.27), p = the number of independent variables and $P(Y=1/\mathbf{x}) = \pi(\mathbf{x})$ = the conditional probability that the outcome is present.

The logit of the multiple logistic regression model is given by;

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \qquad \textbf{(3.28)}$$

in which case the regression model is

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \qquad \textbf{(3.29)}$$

3.9 Fitting the Multiple Logistic Regression Model

The method of estimation used in fitting a multiple logistic regression model is the maximum likelihood estimation method. The likelihood function is nearly identical to that given in equation(3.8) with only a change being that π **(x)** is now defined as in equation(3.28).There will be p+1 likelihood equations that are obtained by differentiating the log likelihood function with respect to the p+1 coefficients. The likelihood equation that results is expressed as;

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0 \qquad \textbf{(3.30)}$$

and

$$\sum_{i=1}^{n} \mathbf{x}_{ij} [y_i - \pi(\mathbf{x_i})] = 0 \qquad \text{for j = 1, 2, …, } p \qquad \textbf{(3.31)}$$

As in the univariate model, the solution of the likelihood equation requires special statistical software packages. In calculating the standard error, we will have to find the

estimates of the variance and covariance of our coefficients. The method of estimating variances and covariance of the estimated coefficients follows from the theory of maximum likelihood estimation which states that the estimators are obtained from the matrix of second partial derivatives of the log likelihood function.

The general form of these partial derivatives is;

$$\frac{\partial^2 L(\beta)}{\partial \beta_j{}^2} = -\sum_{i=1}^{n} x_{ij}{}^2 \pi_i (1 - \pi_i) \quad \textbf{(3.32)}$$

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \, \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} \, x_{il} \pi_i (1 - \pi_i) \quad \textbf{(3.33)}$$

for j,l = 0,1,2,..,p where π_i denotes $\pi(\mathbf{x}_i)$ and p denotes the number of covariates in the model. If $(p+1) \times (p+1)$ matrix containing the negative of the terms given in equations (3.32) and (3.33) be denoted as $\mathbf{I}(\boldsymbol{\beta})$. This matrix is called the *observed information matrix.* The variances and covariances of the estimated coefficients are obtained from the inverse of this matrix which is denoted as $\mathbf{Var\ [I}(\boldsymbol{\beta})\mathbf{]} = \mathbf{I}^{-1}(\boldsymbol{\beta})$

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

$$\mathrm{SE}(\hat{\beta}_j) = [\widehat{\mathrm{Var}}(\hat{\beta}_j)]^{1/2} \quad \textbf{(3.34)}$$

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

That is, the matrix **X** is

$$\mathbf{X} \quad \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{11} & x_{12} & \dots & x_{1p} \end{bmatrix} \qquad \textbf{(3.35)}$$

The matrix **V** is

$$\mathbf{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ & 0 & & \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix} \qquad \textbf{(3.36)}$$

3.10 Testing for the Significance of the Multiple Logistic Regression Parameters

As in the univariate, the first step in this process is usually to assess the significance of the variables in the model. The likelihood ratio test for overall significance of the p coefficients for independent variables in the model is performed in exactly the same manner as in the univariate case. The test is based on the statistic **G** given in equation (3.17). The only difference is that the fitted values, π, under the model are based on the vector containing p + 1 parameters, $\hat{\boldsymbol{\beta}}$**.** Under the null hypothesis that p "slope"

coefficients for the covariates in the model are equal to zero, the distribution of **G** will be chi-square with *p* degrees-of-freedom.

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

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \qquad \textbf{(3.37)}$$

3.11 Confidence Interval Estimation in Multiple Logistic Regression

The confidence interval estimators for the logit are a bit more complicated for the multiple variable model than the results in equation (3.25). The basic idea is the same only there are now more terms involved in the summation. It follows from equation (3.27) that the general expression for the estimator of the logit for a model containing p covariates

$$\hat{g}(\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p \qquad \textbf{(3.38)}$$

An alternative way to express the estimator of the logit in the equation (3.29) is through the use of the vector notation as $\hat{g}(\mathbf{x}) = \mathbf{x}' \hat{\boldsymbol{\beta}}$, where the vector $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ denotes the estimator of the *p*+1 coefficients and the vector $\mathbf{x}' = (x_0, x_1, x_2, \ldots, x_p)$ represent the constant and a set of values of the *p*- covariates in the model, where x_0=1.

The expression for the estimator of the variance of the estimator of the logit in equation (3.38) is

$$\hat{\text{Var}}[\hat{g}(\mathbf{x})] = \sum_{j=0}^{p} x_j^2 \text{Var}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2\, x_j\, x_k \text{Cov}(\hat{\beta}_j, \hat{\beta}_k) \qquad \textbf{(3.39)}$$

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

$$\hat{\text{Var}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}' \mathbf{V} \mathbf{X})^{-1} \qquad \textbf{(3.40)}$$

It follows from equation (3.31) that an equivalent expression for the estimator in equation (3.39) is

$$\begin{aligned} \hat{\text{Var}}[(\hat{g}(\mathbf{x}))] &= \mathbf{x}' \hat{\text{Var}}(\hat{\boldsymbol{\beta}}) \mathbf{x} \\ &= \mathbf{x}' (\mathbf{X}' \mathbf{V} \mathbf{X})^{-1} \mathbf{x} \end{aligned} \qquad \textbf{(3.41)}$$

3.12 ODDS RATIO

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

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \qquad \textbf{(3.42)}$$

Table 3.3: Values of the Logistic Regression Model When the Independent Variable is Dichotomous

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

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) \Big/ \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) \Big/ \left(\frac{1}{1+e^{\beta_0}}\right)}$$

$$= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}$$

$$= e^{(\beta_0+\beta_1)-\beta_0}$$

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

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

$$OR = e^{\beta_1} \qquad \textbf{(3.43)}$$

The interpretation given for the odds ratio is based on the fact that in many instances it approximates a quantity called the relative risk. This parameter is equal to the ratio $\pi(1)/\pi(0)$. It follows from equation (3.41) that the odds ratio approximates the relative risk if [1- $\pi(0)$] / [1- $\pi(1)$] ≈1. This holds when $\pi(\mathbf{x})$ is small for both x =1and x = 0.

3.13 Confidence Limits for Odds Ratio

This is obtained by finding the confidence limits for the log odds ratio. Then, exponentiate these limit to obtain limits for the odds ratio. In general, the limits for a 100(1-α) % confidence interval for the coefficient are of the from

$$\hat{\beta}_1 \pm z_{1-\alpha/2} \text{ x } SE(\hat{\beta}_1) \quad \textbf{(3.44)}$$

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

$$\exp[\hat{\beta}_1 \pm z_{1-\alpha/2} \text{ x } SE(\hat{\beta}_1)] \quad \textbf{(3.45)}$$

3.14 Specifying the Multiple Logistic Regression Model

The logit of the multiple logistic regression model is given by the equation (3.46)

$g(\mathbf{x}) = \beta_0 + \beta_1$age + $\beta_2$gender + $\beta_3$occupation + $\beta_4$hypertensive + $\beta_5$alcohol + $\beta_6$smoking + $\beta_7$cholesterol level + $\beta_8$marital status + $\beta_9$family history.

in which case the logistic regression model is

$$P(Y=1/\mathbf{x}) = \pi(\mathbf{x}) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad \textbf{(3.47)}$$

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 age+\beta_2 gend+\beta_3 0ccu+\beta_4 hypten+\beta_5 alco+\beta_6 smk+\beta_7 choles+\beta_8 mari+\beta_9 fami}}{1+e^{\beta_0+\beta_1 age+\beta_2 gend+\beta_3 0ccu+\beta_4 hypten+\beta_5 alco+\beta_6 smk+\beta_7 choles+\beta_8 mari+\beta_9 fami}}$$

$$Y_i = \begin{cases} 1 & \text{diabetic} \\ 0 & \text{non diabetic} \end{cases}$$

Gender (male = 1; female = 0),

Occupation (employed = 1; unemployed = 0)

Alcohol consumption (Yes = 1; No = 0)

Smoking (Yes = 1; No = 0)

Cholesterol level (High = 1; Low = 0)

Hypertensive (Yes = 1; No = 0)

Family history (Yes = 1; No = 0)

Marital status (married = 1; single = 0)

3.15 Assessing the fit of the Logistic Model

The term used in describing how effective our model is, is referred to as the *goodness – of-fit.* To assess the goodness-of-fit of the model, then we should understand what it means to say that a model fits. Suppose we denote the observed sample values of the outcome variable in vector form as **y** where $\mathbf{y}' = (y_1, y_2, y_3 \ldots y_n)$. We denote the values predicted by the model, or fitted values, as **y** where $\hat{\mathbf{y}}' = (\hat{y}_1, \hat{y}_2, \hat{y}_3 \ldots \hat{y}_n)$We conclude that the model fits if;

1). Summary measures of the distance between **y** and $\hat{\mathbf{y}}$ are small.

2). The contribution of each pair $(y_i, \hat{y}_i)$, i= 1,2,3,..,n to these summary measures is unsystematic and is small relative to the error structure of the model.

Thus, a complete assessment of the fitted model involves both the calculation of summary measures of the distance between **y** and $\hat{\mathbf{y}}$, and a thorough examination of the individual components of these measures.

Goodness-of-fit is assessed over the constellation fitted values determined by the covariates in the model, not the total collection of covariates. For instance, suppose that our fitted model contains p independent variables, $\mathbf{x}' = (x_1, x_2, x_3, \ldots, x_p)$, and let J denote the number of distinct values of **x** observed. If some subjects have the same value of **x** then $J<n$. We denote the number of subjects with $\mathbf{x} = \mathbf{x}_j$ by m_j, j=1,2,3,…,J. It follows that $\sum m_j = n$. Let y_j denote the number of positive responses, y=1 among the m_j subjects with $\mathbf{x} = \mathbf{x}_j$. It follows that $\sum y_j = n_1$, the total number of subjects with y=1. The distribution of the goodness-of-fit statistics is obtained by letting n become large. If the number of covariate patterns (this describe a single set of values for the covariates in a model.) also increases with n then each value of m_j tends to be small.

3.16 Pearson chi-square statistic and deviance

In logistic regression there are several possible ways to measure the difference between the observed and fitted values. To emphasize the fact that the fitted values in logistic regression are calculated for each covariate pattern and depend on the estimated

robability for that covariate pattern, we denote the fitted value for the jth covariate pattern as $\hat{y_j}$ where

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(\mathbf{x}_j)}}{1+e^{\hat{g}(\mathbf{x}_j)}} \tag{3.48}$$

and in equation (3.48), $\hat{g}(\mathbf{x}_j)$ is the estimated logit.

We consider two measures of the difference between the observed and fitted values, that is the Pearson residual and the deviance residual. For a particular covariate pattern the Pearson residual is defined as

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1-\hat{\pi}_j)}} \tag{3.49}$$

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

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2 \tag{3.50}$$

The deviance residual is defined as

$$d(y_j, \hat{\pi}_j) = \pm \left\{ 2 \left[y_j In\left(\frac{y_j}{m_j \hat{\pi}_j}\right) + (m_j - y_j) In\left(\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right) \right] \right\}^{1/2} \tag{3.51}$$

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

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j|\ln(1-\hat{\pi}_j)|} \quad \textbf{(3.52)}$$

and the deviance residual when $y_j = m_j$, is

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j|\ln(\hat{\pi}_j)|} \quad \textbf{(3.53)}$$

The summary statistic based on the deviance residuals is the deviance

$$D=\sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2 \quad \textbf{(3.54)}$$

3.17 The Hosmer-Lomeshow tests

The Hosmer-lemeshow goodness-of-fit statistic, $\hat{C}$, is obtaining by calculating the Pearson chi-square statistic from g×2 table of observed and estimated expected frequencies. A formular defining the calculation of $\hat{C}$ is as follows:

$$\hat{C} = \sum_{k=1}^{g} \frac{\left(o_k - n_k'\bar{\pi}_k\right)^2}{n_k'\bar{\pi}_k\left(1-\bar{\pi}_k\right)} \quad \textbf{(3.55)}$$

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

$$o_k = \sum_{j=1}^{c_k} y_j \quad \textbf{(3.56)}$$

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

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k} \quad \textbf{(3.57)}$$

is the average estimated probability. Using an extensive set of simulations, Hosmer and Lemeshow (1980) demonstrated that, when *J*=*n* and the fitted logistic regression model is the correct model, the distribution of the statistic $\hat{\boldsymbol{C}}$ is well approximated by the chi-square distribution with g – 2 degrees of freedom, χ^2(g- 2).

CHAPTER FOUR

DISCUSSION OF RESULTS

4.0 INTRODUCTION

The chapter describes in detail the results of our research. The results are in two parts. The preliminary results which involve our independent test of diabetes with some of our variables in the data and the test of association between diabetes and alcohol with gender as a confounder.

The second result involves the logistic regression model with our predictor variables.

The data used for the research consist of nine variables; age, gender, occupation, hypertension, diabetes, alcohol consumption, smoking, marital status, family history and cholesterol level. There were 119 people who were employed and 106 people who were unemployed. 187 people were hypertensive and 38 not hypertensive. 123 people consume alcohol and 102 did not, 15 were into smoking and 210 were not, 82 people had a family history of diabetes whiles 143 people did not, 89 were married and 136 were not. There were 134 males and 91 females, and 124 people with high cholesterol level and 101 with low cholesterol level.

4.1 PRELIMINARY RESULTS

Table 4.0: Independent test for diabetes verses alcohol

Alcohol	Diabetes: Yes	Diabetes: No	Total
Yes	49	74	123
No	24	78	102
Total	73	152	225

H_0: Outcome of diabetes is not associated with alcohol intake.

$H_{1:}$ Outcome of diabetes is associated with alcohol intake.

The computed expected frequencies are shown below.

$E_{1,1} = \frac{123 \times 73}{225} = 39.907$ $\quad$ $E_{2,1} = \frac{102 \times 73}{225} = 33.093$

$E_{1,2} = \frac{123 \times 152}{225} = 83.093$ $\quad$ $E_{2,2} = \frac{102 \times 152}{225} = 68.907$

The Pearson Chi-square test statistic

$$\chi^2 = \sum_{a=1}^{2} \sum_{d=1}^{2} \frac{(O_{a,d} - E_{a,d})^2}{E_{a,d}} \sim \chi^2_{(a-1)(d-1)}$$

$$\chi^2 = \frac{(49-39.907)^2}{39.907} + \frac{(74-83.093)^2}{83.093} + \frac{(24-33.093)^2}{33.093} + \frac{(78-68.907)^2}{68.907}$$

$$= 2.072 + 0.995 + 2.498 + 1.200$$

$$\chi^2 = 6.765$$

From the chi-square distribution table, we found P($\chi^2 > 6.765$) = 0.01. Here the P-value (0.01) is less than the significant level (0.05). Thus, we may conclude that there is statistical evidence between alcohol intake and the outcome of diabetes.

Table 4.1: Independent test for diabetes verses smoking

Smoking	Diabetes: Yes	Diabetes: No	Total
Yes	3	12	15
No	70	140	210
Total	73	152	225

H_0: Outcome of diabetes is not associated with smoking behavior

$H_{1:}$ Outcome of diabetes is associated with smoking behavior

The computed expected frequencies are shown below

$E_{1,1} = \frac{15 \times 73}{225} = 4.867$ $\quad E_{2,1} = \frac{210 \times 73}{225} = 68.133$

$E_{1,2} = \frac{15 \times 152}{225} = 10.133$ $\quad E_{2,2} = \frac{210 \times 152}{225} = 141.867$

The Pearson Chi-square statistic

$$\chi^2 = \sum_{s=1}^{2} \sum_{d=1}^{2} \frac{(O_{s,d} - E_{s,d})^2}{E_{s,d}} \sim \chi^2_{(s-1)(d-1)}$$

$$\chi^2 = \frac{(3-4.867)^2}{4.867} + \frac{(12-10.133)^2}{10.133} + \frac{(70-68.133)^2}{68.133} + \frac{(140-141.867)^2}{141.867}$$

$$= 0.716 + 0.344 + 0.051 + 0.025$$

$$\chi^2 = 1.136$$

From our chi-square distribution table, we will find $P(\chi^2 > 1.136) = 0.33$. Here the P-value (0.33) greater than the significant level (0.05). Thus, we may conclude that there is no statistical evidence between the behavior of smoking and the outcome of diabetes.

Table 4.2: Independent test for diabetes verses hypertension

Hypertension	Diabetes: Yes	Diabetes: No	Total
Yes	69	118	187
No	4	34	38
Total	73	152	225

H_0: Outcome of diabetes is not associated with hypertension.

$H_{1:}$ Outcome of diabetes is associated with hypertension

The computed expected frequencies are shown below:

$E_{1,1} = \frac{187 \times 73}{225} = 60.671$ $E_{2,1} = \frac{38 \times 73}{225} = 12.329$

$E_{1,2} = \frac{187 \times 152}{225} = 126.329$ $E_{2,2} = \frac{38 \times 152}{225} = 25.671$

The Pearson Chi-square statistic

$$\chi^2 = \sum_{h=1}^{2}\sum_{d=1}^{2}\frac{(O_{h,d}-E_{h,d})^2}{E_{h,d}} \sim \chi^2_{(h-1)(d-1)}$$

$$\chi^2 = \frac{(69-60.671)^2}{60.671} + \frac{(118-126.329)^2}{126.329} + \frac{(4-12.329)^2}{12.329} + \frac{(34-25.671)^2}{25.164}$$

$$= 1.143 + 0.549 + 5.627 + 2.757$$

$$\chi^2 = 10.076$$

From our chi-square distribution table, we will find $P(\chi^2 > 10.076) = 0.000$. Here the P-value (0.000) is less than the significant level (0.05). Therefore, we may conclude that there is statistical evidence between hypertension and the outcome diabetes.

Table 4.3: Independent test for diabetes verses cholesterol level

Choles. level	Diabetes: Yes	Diabetes: No	Total
High	52	72	124
Low	21	80	101
Total	73	152	225

H_0: Outcome of diabetes is not associated with changes in cholesterol level

$H_{1:}$ Outcome of diabetes is associated with changes in cholesterol level

The Computed expected frequencies are shown below:

$E_{1,1} = \frac{124\times73}{225} = 40.231$ $E_{2,1} = \frac{101\times73}{225} = 32.769$

$E_{1,2} = \frac{124 \times 152}{225} = 83.769$ $E_{2,2} = \frac{101 \times 152}{225} = 68.231$

The Pearson Chi-square statistic

$$\chi^2 = \sum_{cl=1}^{2} \sum_{d=1}^{2} \frac{(O_{cl,d} - E_{cl,d})^2}{E_{cl,d}} \sim \chi^2_{(cl-1)(d-1)}$$

$$\chi^2 = \frac{(52-40.231)^2}{40.231} + \frac{(72-83.769)^2}{83.769} + \frac{(21-32.769)^2}{32.769} + \frac{(80-68.231)^2}{68.231}$$

$$= 3.443 + 1.653 + 4.227 + 2.030$$

$$\chi^2 = 11.353$$

From our chi-square distribution table, we will find $P(\chi^2 > 11.353) = 0.000$. Here the P-value (0.000) is less than the significant level (0.05). Thus, we conclude that there is statistical evidence between the changes in cholesterol level and the outcome diabetes.

Table 4.4: Independent test for diabetes verses occupational level

Occupation	Diabetes: Yes	Diabetes: No	Total
Employed	51	68	119
Unemployed	22	84	106
Total	73	152	225

H_0: Outcome of diabetes is not associated with the level of occupation

H_1: Outcome of diabetes is associated with the level of occupation

The Computed expected frequencies are shown below

$E_{1,1} = \frac{119 \times 73}{225} = 38.609$ $E_{2,1} = \frac{106 \times 73}{225} = 34.391$

$E_{1,2} = \frac{119 \times 152}{225} = 80.391$ $E_{2,2} = \frac{106 \times 152}{225} = 71.609$

The Pearson Chi-square statistic

$$\chi^2 = \sum_{occ=1}^{2} \sum_{d=1}^{2} \frac{(O_{occ,d} - E_{occ,d})^2}{E_{occ,d}} \sim \chi^2_{(OCC-1)(D-1)}$$

$$\chi^2 = \frac{(51-38.609)^2}{38.609} + \frac{(68-80.391)^2}{80.391} + \frac{(22-34.391)^2}{34.391} + \frac{(84-71.609)^2}{71.609}$$

$$= 3.977 + 1.910 + 4.464 + 2.144$$

$$\chi^2 = 12.495$$

From our chi-square distribution table, we will find $P(\chi^2 > 12.495) = 0.000$. Here, the P-value (0.000) is less than the significant level (0.05). Thus, we may conclude that there is statistical evidence between the outcome diabetes and occupational levels of individuals

Table 4.5: Independent test for diabetes verses all the predictor variables.

Variable	Chi-Square value	P-value
Alcohol	6.762	0.010
Gender	1.046	0.320
Smoking	1.136	0.332
Hypertension	10.076	0.000
Cholesterol level	11.353	0.000
Occupation	12.495	0.000
Family status	0.227	0.800
Marital status	0.065	0.400

4.1.1 Test of Association

A confounding variable (confounding factor, lurking variable, etc) is defined as an extraneous variable in a statistical model that correlates (positively or negatively) with both the dependent and the independent variable. ORs in our calculation represents odds ratio.

Table 4.6: The cross tabulation between diabetes and alcohol, while controlling the effect of a possible confounding variable, gender.

		Diabetes		
	Alcohol	Yes	No	Total
Male	Yes	32	40	72
	No	15	47	62
Female	Yes	17	34	51
	No	9	31	40
	Total	73	152	225

In table 4.6, 32 males consume alcohol and had diabetes, 40 of them consume alcohol and did not get diabetes, in all 72 males consume alcohol and 62 of them did not consume alcohol. Amongst the females, 17 of them consume alcohol and had diabetes, 34 consume alcohol and did not get diabetes, in all 51 females consume alcohol and 40 of them did not consume alcohol.

Table 4.7: The relationship between diabetes and alcohol for males

		Diabetes		
		Yes	No	Total
Alcohol	Yes	32	40	72
	No	15	47	62
	Total	47	87	134

Computing the odds ratio in table 4.7

$$\text{ORs} = \frac{32(47)}{40(15)} = \frac{1504}{600} = 2.507$$

From the results in table 4.7, the probability of a man getting diabetes whiles consuming alcohol verses the probability of another man getting diabetes whiles not consuming alcohol is 2.507

Table 4.8: The relationship between diabetes and alcohol for females

Alcohol	Diabetes: Yes	Diabetes: No	Total
Yes	17	34	51
No	9	31	40
Total	26	65	91

Computing the odds ratio in table 4.8

$$\text{ORs} = \frac{17(31)}{34(9)} = \frac{527}{306} = 1.722$$

From the results in table 4.8, the probability of a female getting diabetes whiles consuming alcohol verses the probability of female getting diabetes whiles not consuming alcohol is 1.722.

Table 4.9: Association of diabetes and alcohol ignoring gender as the confounder.

		Diabetes		
		Yes	No	Total
Alcohol	Yes	49	74	123
	No	24	78	102
	Total	73	152	225

Computing the odds ratio in table 4.9

$$\text{ORs} = \frac{49(78)}{74(24)} = \frac{3822}{1776} = 2.152$$

From the results in table 4.8, the probability of one getting diabetes whiles taking alcohol verses the probability of another person getting diabetes whiles not taking alcohol is 2.152.

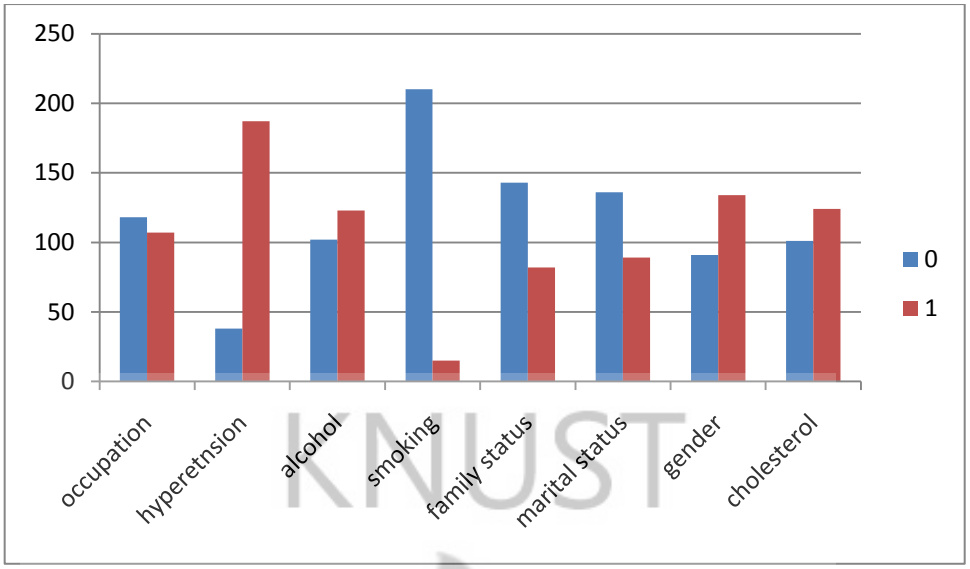

Fig4.0. A stack bar chart of 8 categorical variables in the data.

The fig 4.0 shows stack bar chart of the frequencies of the 8 categorical variables. The 1 or brown bars represent 'yes' and male and 0 or the blue bars represent 'no' and females. There were 119 people who were employed (1) and 106 people who were unemployed (0). 187 people were hypertensive (1) and 38 not hypertensive (0). 123 people consume alcohol (1) and 102 did not (0), 15 were into smoking (1) and 210 were not (0), 82 people had a family history of diabetes (1) whiles 143 people did not (0), 89 were married (1) and 136 were not (0). There were 134 males (1) and 91 females (0), and 124 people with high cholesterol level (1) and 101 with low cholesterol level (0).

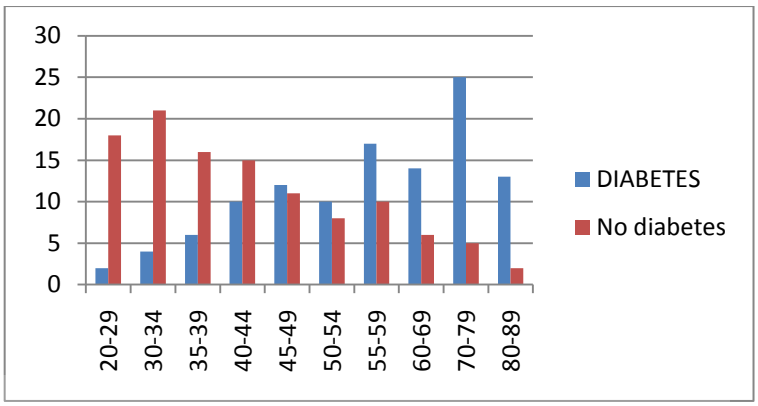

Fig. 4.1: A stack bar chart of the various ages and diabetes status of all the 225 subjects from our data.

From the figure, age category from 20 to 44 years, shows that patients without diabetes dominates in terms of frequency counts whiles from the ages of 45 to 89 years, patients with diabetes increase with frequency counts. From figure 4.1, we may say that increase in the level of age may result in an increase risk of patients having diabetes.

4.2 Logistic Modeling with Categorical Predictors.

The data set contains eight predictor variables: Gender, Occupation, Hypertension, Alcohol Consumption, Smoking, Marital Status, Family history, Cholesterol level. The response variable was the diabetes status.

Table 4.10: Testing Global Null Hypothesis: BETA=0

Test	Chi-square	DF	Pr>ChiSq
Likelihood Ratio	146.7632	9	<0.000
Score	106.1841	9	<0.000

The table , displays the Likelihood Ratio, and the Score test which shows that at least one of the predictors' regressions co-efficient is not equal to zero. There is also the Chi-square test statistic, degrees of freedom (DF) and associated p-value (Pr > ChiSq). We are testing the probability (Pr > ChiSq) of observing a Chi-square statistic under the null hypothesis; the null hypothesis is that all of the regression coefficients in the model is equal to zero. Typically, Pr > ChiSq is compared to a specific alpha level, which we have set at 0.05. The small p-values (0.000 and 0.000) from the two tests which are less than 0.05 would lead us to reject our null hypothesis and conclude that at least one of the regression coefficients in the model is not equal to zero.

Table 4.11: Analysis of Maximum Likelihood Estimates for model 1.

Parameter	Estimate	Standard. Error	Wald Chi-sq	Pr>ChiSq	Exp(Est)	95%Con Inte. for Esp(Est)	
						lower	upper
Intercept	-0.757	0.883	0.735	0.3912	0.469	0.0831	2.648
Age	0.021	0.010	4.410	0.036	1.021	1.001	1.041
Gender	0.161	0.321	0.252	0.616	1.175	0.626	2.204
Occupation	-0.696	0.346	4.046	0.044	0.499	0.237	1.048
Hypertension	0.241	0.428	0.317	0.573	1.272	0.550	2.944
Alco. Cons	0.667	0.310	4.629	0.031	1.948	1.061	3.578
Smoke	-0.891	0.621	2.059	0.151	0.410	0.121	8.280
Marr. Status	-0.098	0.318	0.095	0.758	1.103	0.486	1.691
Family Hist.	- 0.047	0.392	0.120	0.729	0.954	0.442	2.057
Cholesterol lev.	0.938	0.364	6.641	0.010	2.555	1.252	5.214

Parameter estimates are displayed in table 4.12. The Exp(Est) column contains the exponentiated parameter estimates. These values may, but do not necessarily, represent odds ratios for the corresponding predictor variables. For CLASS variables using the effect coding, the Exp(Est) values have no direct interpretation as a comparison of levels. Following the parameter estimates in the above table 4.12, PROC LOGISTIC displays the odds ratio estimates for all the predictor variables. The odds ratio estimates for all the predictors are precisely the values given in the Exp(Est) column in the table 4.12. The 95% Wald Confidence Limit shows the confidence interval (CI) for the odds ratio given the other predictors in the model. For a given predictor with a level of 95%

confidence, we say that we are 95% confidence that the true population odds ratio lies between the lower and upper limit of the interval.

Here, the significant predictors were age, occupation, alcohol consumption and cholesterol level.

4.2.1 Testing for the significance of the full model and the model with the significant predictors.

In other to know if the full model or the model with only the significant predictors is the best model to be used, we will use the likelihood ratio and the Hosmer and Lemeshow test to test for the significant one.

H_0: $\boldsymbol{\beta}_j = 0$

H_1: $\boldsymbol{\beta}_j \neq 0$

Table 4.12: Model fit statistics

-2Log likelihood(G)	Intercept and Covariates
-2Log L_0	265.030
-2Log L_1	263.045

$G = -2(L_0 - L_1)$

Where L_0 = Log likelihood of overall model = 265.030

L_1 = Log likelihood of model with significant variables = 263.045

$G = -2(265.030 - 263.045) = 3.97$ on 5 degrees of freedom

p-value = 0.632 > 0.05

The value of p indicates that we will not reject the null hypothesis. We will conclude that the $\boldsymbol{\beta_i}$ is equal to zero so we cannot use the full model but the reduced model.

The degrees of freedom is equal to the number of parameters that were removed from the original model.

Table 4.13: Homers and Lemeshow test

Step	Chi-square	DF	Sig
1	12.676	8	0.123
2	15.949	8	0.043

The table 4.13 shows the Homers and Lemeshow test. This table displays the significance of our two models. The step 1 shows the chi-square value for our full model with a p-value of 0.123. The second step shows the p-value for our reduced model. The reduced model fits better than our full model because it's p-value is less than our significant level of 0.05.

Table 4.14: Analysis of the Maximum Likelihood Estimate of the reduced model 2.

Parameter	Estimate	Standard Error	Wald Chi-sq	Pr>ChiSq	Exp(Est)	95%Confi. Int for Exp(Est)	
						Lower	Upper
Intercept	-0.533	0.490	1.183	0.278	0.587	0.225	1.533
Age	0.021	0.010	4.41	0.036	1.021	1.001	1.041
Occupation	-0.700	0.351	3.977	0.046	0.497	0.239	1.032
Alco. Cons	0.672	0.308	4.760	0.029	1.958	1.071	3.581
Cholesterol lev.	0.815	0.310	6.912	0.009	2.259	1.230	4.148

From table 4.15, we dropped the insignificant predictors and used the reduced model since it fits better than the full model.

4.2.2 Multiple Logistic Regression

The model was formulated as follows. For the diabetes, Let Y_i be the binary outcome diabetes , (Yes/No) for individual i. $Y_i \sim$Bernoulli (π_i)

Logit $(P(Y=1 \mid x)) = \beta_0 + \beta_1$age $+ \beta_2$gender $+ \beta_3$occupation $+ \beta_4$hypertensive $+ \beta_5$alcohol $+ \beta_6$smoking $+ \beta_7$cholesterol level $+ \beta_8$marital status$+ \beta_9$family history.

4.2.3 The fitted multiple logistic regression model with all parameters

So we used the estimates from the multiple logistic regression for making our interpretation of the model from table 4.10.

Logit $(P(Y=1 \mid x)) = -0.757 + 0.021 * \text{age} + 0.161 * \text{gender} - 0.696 * \text{occupation} + 0.241 * \text{hypertensive} + 0.667 * \text{alcohol} - 0.891 * \text{smoking} - 0.0980 * \text{marital status} - 0.047 * \text{fam. history} + 0.8208 * \text{Cholesterol level}$

4.2.4 The fitted multiple logistic regression model with only significant parameters

Logit$(P(Y=1 \mid x)) = -0.533 + 0.021 * \text{age} - 0.700 * \text{occupation} + 0.667 * \text{alcohol} + 0.815 * \text{cholesterol level}$

Example 4.0

Calculating the Log odds of some individuals with our model.

A 60 year old person who works, takes alcohol and have high cholesterol level.

Logit $(P(Y=1 \mid \mathbf{x})) = -0.533 + 0.021(60) - 0.700(1) + 0.667(1) + 0.815(1)$

$= 1.509.$

This person has 1.509 risk of getting diabetes.

A 50 year old person who is unemployed, takes alcohol and have high cholesterol level

Logit $(P(Y=1 \mid \mathbf{x})) = -0.533 + 0.021(50) - 0.700(0) + 0.667(1) + 0.815(1)$

$= 1.999.$

This person has 1.999 risk of getting diabetes.

A 30 year old person who is employed, takes no alcohol and have low cholesterol level.

$$\text{Logit}\,(P(Y=1 \mid \mathbf{x})) = -0.533 + 0.021(30) - 0.700(1) + 0.667(0) + 0.815(0)$$

$$= 0.097$$

This person has a very low risk of getting diabetes as compared to the others.

CHAPTER FIVE

FINDINGS, CONCLUSION AND RECOMMENDATIONS

5.0 INTRODUCTION

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

5.1 Findings From the study

The data used for the research consist of nine variables; age, gender, occupation, hypertension, diabetes, alcohol consumption, smoking, marital status, family history and cholesterol level. There were 119 people who were employed and 106 people who were unemployed. 187 people were hypertensive and 38 not hypertensive. 123 people consume alcohol and 102 did not, 15 were into smoking and 210 were not, 82 people had a family history of diabetes whiles 143 people did not, 89 were married and 136 were not. There were 134 males and 91 females, and 124 people with high cholesterol level and 101 with low cholesterol level.

The preliminary results showed the results from our independent test of diabetes with all the predictor variables (alcohol, smoking, hypertension, cholesterol level, family history, occupation, and marital status). It was reported that alcohol consumption, cholesterol level, occupation and hypertension were associated with the outcome of having diabetes. The predictors, smoking, family history, gender and marital status had no association with having diabetes.

From the test of association, we considered the association between diabetes and alcohol with gender as a confounder. That is, we wanted to find out if gender had any influence on one getting diabetes when one consumes alcohol. The odds ratio test showed the following results; the probability of male getting diabetes whiles taking alcohol to the probability of a male getting diabetes whiles not taking alcohol was 2.507. This means that, men who consume alcohol were more likely in getting diabetes as compared to those who consume no alcohol. Females who consume no alcohol also had a lesser chance of getting diabetes than those who consume alcohol. (Odds ratio of 1.722). The marginal association between diabetes and alcohol ignoring the confounder gave an odds ratio of 2.152. This proves that alcohol consumption increases one's risk of getting diabetes and this can be influenced by the gender of the patient.

Considering age which was the only continues variable in our data, a pictorial representation showed increase in age with increase in getting diabetes.

The likelihood ratio and Score test with p-values of 0.000 and 0.000 respectively proved the significance of the regression coefficients in the full model.

From the model with all the predictors, the fitted model showed a positive association between age and diabetics. This is because the P-value of age (0.0357) is far below the chosen significant level of 0.05. Results also in fig 1.1 also showed that an increase in age, increases one's risk of getting diabetes. The odds ratio also showed that, the multiplicative effect of a year increase in age on the odds of having diabetes is 1.021.

Study on ≥20 years of age subjects, prevalence of diabetes was found 7.6%, and the prevalence of diabetes increased with age and peaked in the oldest age group in Africa. Kim et al, (2001). Similarly, Choi and Shu, (2001) also documented that increased age was 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.

Again, the study also demonstrated a positive association between diabetes and cholesterol level (p=0.009). Again, the odds ratio ($e^{0.815}$ = 2.259) showed that, those with high cholesterol levels have more than twice the likelihood of getting diabetes than those with low cholesterol levels. Our reports is also in agreement with results from other countries such as Olinto et al. (2003) in Brazil and Chu. (2005) in Taiwan that suggested that high cholesterol level was an independent determinant for diabetes and obesity as a wide spread and growing problem in the world with significant medical, psychological and economical consequences. 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.

Our findings also establish a relationship between diabetes and alcohol consumption. The Odds ratio ($e^{0.672}$ = 1.958) also showed that individuals who took alcohol have almost two times the likelihood of getting diabetes than those who took no alcohol. A study by American Diabetes Association (2009) that is already stated in our literature review showed that the consumption of alcohol is an influencing factor of diabetes.

Motala et al, (2008) in rural communities of South Africa and Wandell & Gafvels (2007) in Sweden educated that subjects consuming alcohol were at higher risk of getting diabetes. In addition, different studies such as Burchfeil et al. (1995) and Ajani et al. (2000) observed an association of intake of alcohol with diabetes. Lapidus et al. (2005) also saw an association between alcohol intake and diabetes

The findings from the study showed that subjects who were involved in various social activities when it comes to working were less likely to develop diabetes. The odds ratio ($e^{-0.700} = 0.497$) also showed that those who were employed were at a lesser risk of getting diabetes than the unemployed. Sundiquist et al. (2004) also observed a positive association between occupation and diabetes. They related their findings with these two reasons; First, social participation which is involved with work could be associated with exercise. Secondly, mental satisfaction produced by social activities and self esteem as well.

Finally, though gender was not significant to one been diagnose of diabetes, our test of association with gender as a confounder showed that there was a higher risk of men getting diabetes than women. This may be due to the fact that alcohol intake was high in men than in women from our data.

5.2 CONCLUSION

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

Firstly, the independent test of diabetes versus all the predictor variables showed that occupational status, alcohol consumption, hypertension and cholesterol level were associated with one's risk of being diagnosed of diabetes.

Secondly, our statistical analysis showed that age, alcohol consumption, occupational status and cholesterol level are significant to the diagnosis of diabetes. The model fitted shows that getting diabetes does not depend on the gender of a person, a medical history of one been hypertensive or not, smoking, marital status, and having a family history of diabetes but rather increase in age increases your risk of getting the disease. Also, been unemployed, taking alcohol, and having a high cholesterol level increases ones risk of been diagnosed as diabetic.

Lastly, though, gender is not a significant predictor in the model, men and women who consume alcohol have higher risk of getting diabetes than those who consume no alcohol.

5.3 RECOMMENDATION

Based on the findings from this study, the following recommendations are made to reduce the burden of most disabling disease like diabetes in people.

- The data used for the study was a secondary data from the Komfo Anokye Teaching Hospital in Kumasi in the Ashanti Region of Ghana. We therefore recommend that researchers who will want to carry on with a project like this should seek for primary data from patients. Because of the disadvantages associated with secondary data.
- Women and men of all ages should avoid becoming overweight, by maintaining their body mass index below 25 kg/m^2 and 27 kg/m^2, respectively. They should maintain a moderate level of physical activity.
- An adequate fruit and vegetable intake is linked to several health benefits, due to an assumed complex interaction of containing biological active compounds.

  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)

- Access to and use of health services need to be increased in our communities particularly for diabetes and other cardiovascular diseases. This will help individuals to visit the place early for treatment than to wait for the disease to reach it climax before they visit hospitals in the cities. We also recommend awareness and regular screening programs which will promote early diagnosis of undetected disease in our communities including diabetes.

REFERENCE

Addo, J., Smeeth, L., &Leon, D.L.(2007). Hypertension in Sub Saharan Africa a systematic review. Hypertension. 50(6),1012-18.

Addo, J.,Amoah, A.G., & Koram, K.A (2006). The changing pattern of hypertension in Ghana. A study of four rural communities in the GA district. Ethn Dis. 16(4), 89-99.

Agyeman, C. (2006). Rural and urban differences in Blood pressure and Hypertension in Ghana, West Africa. The Journal of Public Health. 120(6), 525-33

Agreti, A. (1990). *Categorical Data Analysis*, Wiley, Inc., New York

Ajani, U.A., Gaziano, J.M., Lotufo, P.A., Liu, S., Hennekens, C.H., Buring, J.E., et al. (2000). Alcohol consumption and risk of coronary heart disease by diabetes status. *Circulation.* 102(5), 489-90.

Ali H. Mokdad, PhD; Earl S. Ford, MD, MPH; Barbara A. Bowman, PhD; William H. Dietz, MD, PhD; Frank Vinicor, MD, MPH; Virginia S. Bales, MPH; James S. Marks, MD, MPH, Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors, 2001

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, Second Edition, Wiley, Inc., New York.

Bachand, A.M., and Hosmer, D.W. (1999). Defining confounding when the outcome is dichotomous, Submitted for Publication.

Baltazar, J.C., Ancheta, C.A., Aban, I.B., Fernando, R.E., & Baquila, M..M. (2004). Prevalence and correlation of diabetes mellitus and impaired glucose tolerance among adults in Luzon, Phillipines. *Diabetes Research and Clinical Medicine.* 64(2), 107-552.

Basnet, S., & Sharma, S. (2000). *Ageing Problem Two Sides of Coin.* Kathmandu: Nepal Participatory Action Network.

Bedrick, E.J., and Hill, J.R. (1996). Assessing the fit of logistic regression models to individuals matched sets of case-control data. *Biometrics,*52, 1-9.

Beaty T. H. (Johns Hopkins U. School of Hygiene and Public Health, Baltimore, MD 21205), J. V. Neel and S. S. Fajans. Identifying risk factors for diabetes in first degree relatives of non-Insulin dependent diabetic patients. *Am J Epidemiol* 1982;115:380-97.

Bendel, R. B., and Afifi, A. A. (1977). Comparism of stopping rules in forward regression. *Journal of the American Statistical Association,*72, 46-53.

Burchfiel, C.M., Curb, J.D., Rodriguez, B.L., Yano, K., Hwang, L.J., Fong, K.O., et al. (1995). Incidence and predictors of diabetes in Japanese-American men. The Honolulu Heart Program. *An Epidermiology.* 5(1), 33-43.

Campbell, N. R., Ashley, M. J., Carruthers, S.G., Lacourciere, Y., & Mckay, D .W. (1999). Lifestyle modification to prevent and control hypertension. 3 recommendations on alcohol consumption. Canadian hypertension society, Canadian coalitation for high blood pressure prevention and control, laboratory center for Disease control and health, Canada, heart and stroke foundation of Canada. *Canadian Medical Association Journal.* 160(9) , 513-520.

Carballo, M. & Friedrik, S. (2006). Migration and Diabetes: the emerging challenge.*Diabetes Voice.* 51(2), 31-33.

Chao, W-H., Palta, M., and Young, T. (1997). Effect of omitted confounders in the analysis of correlated binary data. *Biometrics,* 53, 678-689

Collett, D. (1991) *Modelling Binary Data.* Chapman & Hall, London.

Cox, D. R. (1970). *The Analysis of Binary Data.* Methuen, London.

Cox, D. R., and Snell, E. J. (1989). *Analysis of Binary Data,* Second Edition. Chapman & Hall, London.

Djousse, L., Biggs, M.L., Mukamal, K.J., & Sisicovic, D.S. (2007). Alcohol consumption and type 2 diabetes among older adults; The cardiovascular health study. *Obesity* 15(7), 1758-1765.

Dobson, A. (1990). *An Introduction to Generalised Linear Models.* Chapman & Hall, London.

Hosmer, D. W., Jovanovic, B., and Lemeshow, S. (1989). Best subsets logistic regression. *Biometrics,* 45, 1265-1270.

Hosmer, D.W., and Lemershow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics,* A10, 1043-1069.

Hosmer, T., Hosmer, D. W., and Klar .J. (1988). Goodness-of-fit testing for mulitiple logistic regression analysis when the estimated probabilities are small. *Biometrical Journal,* 30, 911-924.

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997). Acomparism of goodness-of-fit tests for the logistic regression model, *Statistic in Medicine,*8, 795-802.

Hosmer, D, W., and Lemeshow, S.(1999). *Applied Survival Analysis: Regression Modelling of Time to Event Data.* Wiley, Inc., New York.

Keil U, Liese AD, Hense HW, et al. Classical risk factors and their impact on incident non-fatal and fatal myocardial infarction and all-cause mortality in southern Germany. Results from the MONICA Augsburg cohort study 1984–1992. Eur Heart J 1998; 19: 1197–207.

Kim, S.M., Lee, J.S., Lee, J., Na, J.K., Han, J.H., & Yoon, D.K. (2006).Prevalence of diabetes and impaired fasting glucose in Korea. *Diabetes Care.* 29(2), 226-32.

Lemeshow, S., and Hosmer, D .W.,(1982). The use of goodness-of-fit statistics in the development of logistic regression models. *American Journal of Epidemiology,* 115, 92-106.

Lemesho, S., and Hosmer, D. W., (1983). Estimation of odds ratios with categorical scaled covariates in multiple logistic regression analysis. *American Journal of Epidemiology*, 119, 147-151.

Lemeshow, S., Hosmer, D, W., Klar, J., and Lwanga, S.K. (1990). *Adequacy of Sample Size in Health Studies.* Wiley, Chichester.

Lemeshow, S., Teres, D., Avrunin, J. S., and Pastides, H. (1988). Predicting the outcome of intensive care uniy patients. *Journal of the American Statistical Association,* 83, 348-356.

McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *Journal of the American Statistical Association,* 81, 104-107.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models,* Second Edition. Chapman & Hall, London.

Meisinger C, Thorand B, Schneider A, Stieber J, Doering A, Loewel H. Sex differences in risk factors for incident type 2 diabetes mellitus. The MONICA Augsburg Cohort Study. Arch Intern Med 2002; 162:

Motala, A., Esterhuizen, T., Gouws, E., Pirie, F.J., & Omar, A.K.M. (2008). Diabetes and other disorders of glcemia in a rural South African community. *Diabetes Care* 31(9), 1783-85.

Mufunda, J., Mebrahutu,G., Uman, A., Nayarango, P., Kosia, A., Ghebart, Y., et al. (2006). The prevalence of hypertension and it relationship with obesity; results from a national blood pressure survey in Eritera. *J Hum Hypertension.* 20(1), 59-65

Muni, C. (2008). Prevalence and determination of hypertension and diabetes among elderly population in the Kathmandu valley of Nepal. *Journal of Public Health.* 145-50

Nakanishi, N., Suzuki, K., & Tatara, K. (2003). Alcohol consumption and risk of development of impaired fasting glucose or type 2 diabetes in middle aged Japanese men. *Diabetes Care* 26(1), 48-54.

Olinto, M.T., Nacul, L.C., Gigante, D.P., Costa, J.S.D., Menezes, A.M.B. & Macedo, S. (2003). Waist circumference as a determinant of hypertension and diabetes in Brizalian women: a population based study. *Public Health Nutrition* 7(5), 629-635.

Okosun, I.K.E., Richard, S., Cooper, S., Rotimt, C.N., Ostomehin, B., & Forrester, T. (1998). Associated waist circumference with risk of hypertension and type 2 diabetes in Nigerians, Jamaicans, and African-American. *Diabetes Care.* 21(11), 1836-1842.

SAS Institute Inc. (1999). *SAS Guide for Personal Computers, Version 6.12.* SAS Institue Inc., Cary, NC.

SAS Institute Inc. (2000). *SAS Guide for personal Computers, Version 8.0.* SAS Institute Inc., Cary, NC.

Scott, A. J., and Wild, C. J. (1986). Fitting models under case control or choice based sampling. *Journal of the Royal Statistical Association, Series, B,* 48, 170-182.

Siddiqui, H., Qudsia, A., Oman, A., Usman, J., Rizvi, R., & Ashfaq, T. (2005). Risk factor assessment for hypertension in a squatter settlement of Karachi. *Journal of Pakistan Medical Association.* 55(9), 390-92

Wang Y, Rimm EB, Stampfer MJ, Willett WC, Hu FB. Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men. Am J Clin Nutr 2005; 81: 555–63.

World Health Organisation [WHO]. (1994). *Prevention of diabetes mellitus: Report of a WHO study group.* WHO technical report series 844; Geneva: Geneva.

Zozuliniska, D. & Wysoca, B.W. (2006). Type 2 diabetes mellitus as inflammatory disease. *Diabete Research and Clinical Practice.* 6(7)

APPENDIX A

INDEPENDENT TEST OF DIABETES WITH FAMILY HISTORY, MARITAL STATUS AND GENDER.

Table 1: Independent test of diabetes versus family history

Fam. history	Diabetes: Yes	Diabetes: No	Total
Yes	25	57	82
No	48	95	143
Total	73	152	225

H_0: Outcome of diabetes is not associated with one's family history

H_1: Outcome of diabetes is associated with one's family history

The Computed expected frequencies are calculated below:

$E_{1,1} = \frac{82 \times 73}{225} = 26.604$ $\quad$ $E_{2,1} = \frac{143 \times 73}{225} = 46.395$

$E_{1,2} = \frac{82 \times 152}{225} = 55.395$ $\quad$ $E_{2,2} = \frac{143 \times 152}{225} = 96.604$

The Pearson Chi-square statistic

$$\chi^2 = \sum_{f=1}^{2} \sum_{d=1}^{2} \frac{(O_{f,d} - E_{f,d})^2}{E_{f,d}} \sim \chi^2_{(f-1)(d-1)}$$

$$\chi^2 = \frac{(25-26.604)^2}{26.604} + \frac{(48-46.395)^2}{46.395} + \frac{(57-55.395)^2}{55.395} + \frac{(95-96.604)^2}{96.604}$$

$$= 0.097 + 0.056 + 0.047 + 0.027$$

$$\chi^2 = 0.227$$

From our chi-square distribution table, we will find P($\chi^2 > 0.227$) = 0.800. Here, the P-value (0.800) is greater than the significant level (0.05). Thus, we conclude that there is no statistical evidence between diabetes and family history. That is, one cannot get diabetes because a family member had it.

Table 2: Independent test of diabetes verse marital status

		Diabetes		
		Yes	No	Total
Marital status	Yes	28	61	89
	No	45	91	136
	Total	73	152	225

H_0: Outcome of diabetes is not associated with one's marital status

H_1: Outcome of diabetes is associated with one's marital status

The Computed expected frequencies are calculated below:

$E_{1,1} = \frac{89 \times 73}{225} = 28.876$ $\quad$ $E_{2,1} = \frac{136 \times 73}{225} = 44.124$

$E_{1,2} = \frac{89 \times 152}{225} = 60.124$ $\quad$ $E_{2,2} = \frac{136 \times 152}{225} = 91.876$

The Pearson Chi-square statistic

$$\chi^2 = \sum_{f=1}^{2} \sum_{d=1}^{2} \frac{(O_{f,d} - E_{f,d})^2}{E_{f,d}} \sim \chi^2_{(f-1)(d-1)}$$

$$\chi^2 = \frac{(28-28.876)^2}{28.876} + \frac{(61-60.124)^2}{60.124} + \frac{(45-44.124)^2}{44.124} + \frac{(91-91.876)^2}{91.876}$$

$$= 0.0265 + 0.0128 + 0.0174 + 0.0084$$

$$\chi^2 = 0.0650$$

From our chi-square distribution table, we will find $P(\chi^2 < 0.0650) = 0.400$. Here, the P-value (0.400) is greater than the significant level (0.05). Thus, we conclude that there is no statistical evidence between diabetes and marital status.

Table 3: Independent test for diabetes verses Gender

Gender	Diabetes: Yes	No	Total
Male	47	87	134
Female	26	65	91
Total	73	152	225

H_0: Outcome of diabetes is not associated with the gender of a person

$H_{1:}$ Outcome of diabetes is associated with the gender of a person

The Computed expected frequencies are shown below

$E_{1,1} = \frac{134 \times 73}{225} = 43.476$ $E_{2,1} = \frac{91 \times 73}{225} = 29.524$

$E_{1,2} = \frac{134 \times 152}{225} = 90.524$ $E_{2,2} = \frac{91 \times 152}{225} = 61.476$

The Pearson Chi-square statistic

$$\chi^2 = \sum_{gen=1}^{2} \sum_{d=1}^{2} \frac{(O_{gen,d} - E_{gen,d})^2}{E_{gen,d}} \sim \chi^2_{(gen-1)(D-1)}$$

$$\chi^2 = \frac{(47-43.476)^2}{43.476} + \frac{(87-90.524)^2}{90.524} + \frac{(26-29.524)^2}{29.524} + \frac{(65-61.476)^2}{61.476}$$

$$= 0.286 + 0.137 + 0.421 + 0.202$$

$$\chi^2 = 1.046$$

From our chi-square distribution table, we will find $P(\chi^2 < 1.046) = 0.300$. Here, the P-value (0.300) is greater than the significant level (0.05). Thus, we may conclude that there is no statistical evidence between the outcome diabetes and the gender of a patient.

Table 4: Association between Gender and diabetes

Gender	Diabetes: Yes	No	Total
Male	47	87	134
Female	26	65	91
Total	73	152	225

$$\text{ORs} = \frac{47(65)}{87(26)} = \frac{3055}{2262} = 1.351$$

From the calculation above, the probability of one getting diabetes as a man verses the probability of female getting diabetes is 1.351

Table 5: The description of the 8 variables in the data with their percentages.

Variables	Categories	Number in each group	Percentage (%)
Occupation	employed	119	52.89
	unemployed	106	47.11
Hypertension	yes	187	83.11
	no	38	16.89
Alcohol cons.	yes	123	54.67
	no	102	45.33
Smoking	yes	15	6.67
	no	210	93.33
Family history	yes	82	36.44
	no	143	63.56
Marital status	married	89	39.56
	single	136	60.44
Gender	male	34	59.56
	Females	91	40.44
Cholesterol level	high	124	55.11
	low	101	44.89