APPLICATION OF STATISTICAL MODELS TO ABORTION CASES IN THE CENTRAL REGION OF GHANA

By

FRANCIS MAWULI ABUDE (BSc. Statistics, HND Statistics)

A Thesis submitted to the Department of Mathematics, Kwame Nkrumah University of Science and Technology in partial fulfillment of the requirements for the degree of

MASTER OF PHILOSOPHY (Industrial Mathematics)

Institute of Distance Learning

MAY 2013

CERTIFICATION

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

Francis Mawuli Abude	…………………	14/05/13
(PG 6326111)	Signature	Date

Certified by:

Mr. Nana Kena Frempong	…………………	31/5/2013
(Supervisor)	Signature	Date
Mr. F. K. Darkwah	…………………	31/5/2013
(HOD)	Signature	Date
Professor I. K. Dontwi	…………………	31/5/2013
(Dean, IDL)	Signature	Date

ABSTRACT

Abortion is a major health problem that contributes significantly to maternal mortality and mobility in Ghana especially due to its attending post-complications. This study therefore assessed and modelled the risk factors such as age groups of abortion cases, occupation and marital status with time in the Central Region. Data from 2008 to 2012 were obtained from the Regional Health Directorate. The Poisson and Negative Binomial regression models were fitted to the data. From the results of the analysis, it was found that all the risk factors were significantly associated with abortion. Also, the prevalence rates of abortion were 99.9% and 91.9% higher for women aged between 20-24 and 25-29 respectively than those aged at least 35 years, while traders and the unemployed women are more prone to abortion than other categories of professionals. Unmarried women were also found to be more likely to terminate their pregnancies than their married counterparts. We conclude that these risk factors must be considered in any intervention programmes aimed at reducing abortion and maternal mortalities in the region.

ACKNOWLEDGEMENTS

The successful completion of this work owes much to individuals who cannot go unnoticed. My heartfelt gratitude goes to Mr. Nana Kena Frempong for perusing through my work and most particularly his personal friendship with me. I sincerely appreciate the invaluable tuition and direction of all lecturers of the Department of Mathematics at the Kwame Nkrumah University of Science and Technology (KNUST), Kumasi especially Professor S. K. Amponsah, Dr. F. T. Oduro, Dr. Richard K. Avuglah and Dr. Richard Koufie. I also thank Messrs Alexander Boateng for his sage advice, and my colleagues, Cosmos Agbe Todoko and Japheth Arthur. My thanks go to Mr. Maxwell K. Abude, my father and Madam Juliana A. Kornu, my mother. I thank my brothers, sisters and friends particularly Peace, Solace, Victoria, George, Carl, Esther, Ivy, Anna, Noble, Eyram, Francis and Fuska for their morale support. I am also grateful to my past and present seniors, colleagues and subordinates at the Directorate of Academic Planning and Quality Assurance (DAPQA) of the University of Cape Coast especially Professor Kobina Yankson (former Director), and Professor Kofi Awusabo-Asare (current Director). I appreciate the efforts of officials at the Regional Health Directorate especially Mr. Gideon Agbeshie and Robo who freely assisted me in getting the data.

DEDICATION

To my father, Mr. Maxwell K. Abude.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC – Akaike's Information Criterion

BIC – Bayesian's Information Criterion

C/R – Central Region

D&C – Dilation and Curettage

GDHS – Ghana Demographic and Health Survey

GHS – Ghana Health Service

GLM – Generalised Linear Model

GSS – Ghana Statistical Service

IUD – Intrauterine device

MDG – Millennium Development Goal

MVA – Manual Vacuum Aspiration

Pdf – Probability Density Function

PHN – Public Health Nurse

PPAG – Planned Parenthood Association of Ghana

RHD – Regional Heath Directorate

SPSS – Statistical Product and Service Solutions

UN – United Nations

WHO – World Health Organisation

CHAPTER ONE

INTRODUCTION

1.1 Background of study

Abortion is the termination of pregnancy by the removal or expulsion from the uterus of a fetus or embryo prior to viability (Grimes et al., 2006). They indicated that an abortion can occur spontaneously, in which case it is usually called a miscarriage, or it can be purposely induced.

Abortion (unsafe) is a major health problem that contributes significantly to maternal mortality and morbidity (WHO, 1998). In 2007, the World Health Organisation estimated that globally, over 600,000 women and girls die of complications related to pregnancy and childbirth each year. Over 99% of those deaths occur in developing countries such as Ghana. Maternal deaths only tell part of the story because for every woman or girl who dies as a result of pregnancy-related causes, between 20 and 30 more will develop short and long-term disabilities, such as obstetric fistula, a ruptured uterus, or pelvic inflammatory disease. The death and the many complications that arise from pregnancy are related to several crude practices of pregnancy handling and management including abortion by girls and women. The commonest types of post-abortion complication recorded in the Central Region are bleeding/hemorrhage, sepsis, and perforations.

Maternal mortality of which abortion-related fatalities is a principal component, is one of the most tragic and serious health problems in the world and Ghana is no exception. The rate at which women are dying is so alarming that one of the Millennium Development Goals (MDGs) is targeted at reducing this phenomenon by three quarters by 2015 (UN, 2007). Current world rate of maternal mortality shows clearly that this goal can never be achieved unless something drastic is done to reverse the current trend. For instance, there are disturbing reports coming from England and Wales as national statistics show that more than a quarter of all deaths in the countries are caused by abortions especially unsafe abortions (Office for National Statistics of the United Kingdom, 2011).

Unsafe abortions are a major cause of injury and death among women worldwide. Although data are imprecise, it is estimated that approximately 20 million unsafe abortions are performed annually, with 97% taking place in developing countries (Grimes et al., 2006). They reported that unsafe abortion is believed to result in approximately 68,000 deaths and millions of injuries annually. Groups such as the WHO have advocated a public-health approach to addressing unsafe abortion, emphasising the legalisation of abortion, the training of medical personnel, and ensuring access to reproductive-health services (Berer, 2000).

Official statistics from the Ministry of Health indicated that abortion cases were on the increase where 16,182 girls and young women caused abortion in 2011 as against 10,785 in 2010 and 8,717 in 2009. According to the Adolescent Health and Development Programme of the Ministry of Health (2011), 216 cases of abortion involving girls aged 10 to 14 were recorded in 2009; 331 cases in 2010 and 582 in 2011. In the case of girls ~~15 to~~ 19 years of age, 5,525 abortion cases were recorded in 2009; 6,679 in 2010 and 7,800 in 2011. The report further indicated that for young women aged 20 to 24, 7,800 abortion cases were recorded in 2011; 6,679 in 2010 and 5,525 in 2009. The above figures excluded the numerous unreported abortion cases and deaths. According to Lassey and Wilson (1998), hospital-based studies at two leading hospitals in Ghana indicate that between 22% and 30% of maternal deaths are due to unsafe abortion. WHO

(1997) estimated that besides the health complications of unsafe abortion, the socioeconomic cost is so high that in some centres in developing countries, treating abortion-related complications may consume up to 50% of hospital budgets and resources, including medical staff time, medicines and supplies.

1.2 Study area profile

The Central Region has an estimated population of 1,961,994 (2011) and an annual population growth rate of 2.1%. With a population density of about 162 inhabitants per square kilometres, the region is the second most densely populated region after Greater Accra (GSS, 2011). Even though the Central Region has seen some reduction in Infant Mortality and under-five Mortality since 1998, the figures are still higher than the national average. The region recorded a steady decline in these indicators till 2003 before recording an increase in 2008 (GSS, 2008). The 2008 figures are quite high compared to the 2003 figures in the same survey of 50 and 90 per 1,000 live births respectively. In the southern sector of Ghana, the Central Region has the highest figures for both infant and maternal mortalities. The literacy rate among women age 15-49 years in the region has increased to 57% in 2008 from 46.7% in 2003. This improvement is very important considering the role women in that age group play in family health and family planning. The region has 224 health facilities and five Nurses Training Institutions. Abortion (unsafe) is always among the first three top causes of deaths in the region. According to the Population and Housing Census (GSS, 2010), unemployment is 8.0%; 2.4% lower than the national average of 10.4%. About 5 % of children under age 15 years are engaged in economic activities in most districts. The predominant industry in the districts, except Cape Coast Metropolis is agriculture (52.3%), followed by manufacturing (10.5%).

1.3 Problem statement

Abortion is a major health problem that contributes significantly to maternal mortality and mobility in Ghana, especially due to its attending post-complications (GHS, 2011). The World Health Organisation estimates that over 600,000 women and girls die annually from pregnancy and childbirth related causes mainly in the developing countries and that unsafe abortion is responsible for about 13 – 14% or in some cases 60% of these deaths (Grimes et al., 2006).

In the Central Region of Ghana, abortion cases increased from 2,739 in the year 2007 to 3,649 in the year 2012; making it the third cause of maternal deaths in the region. These alarming figures threaten the achievement of the Millennium Development Goal (MDG); that primarily aims at reducing maternal mortality by 75% by 2015. Despite several interventions (sexual and reproductive health, family life education) put in place, they have largely failed especially because many of such programmes did not take into accounts the various risk factors contributing to abortion. Although Turpin et al. (2002), and Yeboah and Kom (2003) identified some of these risk factors, their studies was basically descriptive and also did not develop any model to identify incidence rates among these risk factors. There is no such inferential analysis or model to be able to predict abortion cases in the region.

1.4 Objectives of the study

The main objective of the study is to model the abortion cases in the region. The study has the following specific objectives:

1. to use regression model to stud the model abortion cases based on the risk factors in order to identify the incidence rates among them.
2. to validate the Poisson model with the Negative Binomial model.

1.5 Methodology

Data was obtained from the Regional Health Directorate of the Central Region on monthly bases from 2008 to 2012. The R, SPSS and Microsoft Excel software were used to analyse the data. The data available are count and non-negative integers with the events being independent and average rate not changing over the period of interest hence the Poisson regression model could be most appropriate tool for the modeling. In statistics, Poisson regression is a form of regression analysis used to model count data of which malaria cases is no exception.

Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modelled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a log-linear model, especially when used to model contingency tables.

1.6 Justification

Abortion (unsafe) threatens the country's efforts in reducing maternal mortality by 75% by 2015 if the various risk factors associated with abortion are not known and invention programmes are well targeted. It is also believed that government will use the findings to strategies towards achieving MDG 5. The Ministry of Health and Ghana Health Service through the Central Regional Health Directorate will use the findings in their manpower, drugs and logistics planning in relation to reproductive health education programmes. The Regional Public Health Nurses (RHNs), PPAG and other health-concerned organisations in the region could also use these findings in advancing their campaigns for family planning and against unsafe abortions. Lastly, it will also serve as an additional source of reference for other researchers, organisations and students (nurse, medical doctors, and medical statisticians among others) whose studies relate to abortion and abortion-related deaths and Poisson regression modeling techniques.

1.7 Scope and Limitation

The thesis was restricted to the objectives and variables of the research. Research work was characterised by some constraints. Some of these setbacks included time constraint and untimely release of data by the various authorities since abortion issues are said to highly sensitive and confidential.

1.8 Thesis organisation

Chapter one is made up of introduction, which comprises the background of the study, study areas, problem statement and objective of the study. It also presents the justification and limitations of the study. Chapter two highlights related literature on the topic with ideas of different authors whose findings have been defined in relation to the topic under study. Chapter three focuses on methodological review in the light of mathematical and statistical tools that are relevant to the analyses of the data gathered. Basically, the study seeks to use time series model for the analyses. Chapter four deals with the analysis of data and the results, while chapter five presents of conclusion and recommendations.

1.9 Chapter summary

The chapter gave an introduction to the thesis report highlighting on issues relating to background of the study, problem statement and objectives guiding the study, methodology and justification of the study. In addition to these are limitations as well as thesis organisation. The chapter concludes with this summary.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter reviewed related literature on issues of abortion such as types of abortion, methods used in abortion, and some recent studies on the subject. It also examined the application of Poisson and Negative Binomial regression models in several fields of study especially epidemiology.

2.2 Types of Abortion

Three main groups of abortion are identified, namely, the spontaneous, induced and elective.

2.2.1 Spontaneous abortion

Spontaneous abortion or the early termination of pregnancy without outside interference may be caused by fetal, maternal or external factors (McBride, 1991). In many cases, a specific etiology may never be identified. A variety of clinical presentations are possible, ranging from imperceptible loss to profound life-threatening shock. Physicians should be able to diagnose and manage the six recognised types of spontaneous abortion: ~~threatened~~, inevitable, incomplete, complete, missed and septic. In all cases, uterine evacuation, avoidance of complications and psychological support of the family are important. The prognosis for a subsequent successful pregnancy is good, except in cases of habitual abortion.

2.2.2 Elective abortion

Elective abortion also known as the Therapeutic Abortion Obstetrics is a voluntary interruption of pregnancy before fetal viability, which is performed voluntarily at the request of the mother for reasons unrelated to concerns for maternal or fetal health or welfare; most abortions are elective. For instance, there is 1 EA per 3 live births in the US.

2.2.3 Induced Method

This is an intentional termination of pregnancy before the fetus has developed enough to live if born. From 20% to 50% of pregnancies are terminated deliberately at the request of the mother or for medical indications, during the first trimester by vacuum aspiration and/or curettage or during the second trimester by dilation and evacuation, induction of labour, or hysterotomy. Termination of pregnancy by a trained person under proper conditions is safe.

2.3 Methods of Performing Abortion

The study identified two (2) main ways of aborting pregnancies in the region. These are ~~the Man~~ual Vacuum Aspiration (MVA) and Dilation and Curettage (D&C).

2.3.1 Manual Vacuum Aspiration (MVA)

It uses aspiration to remove uterine contents through the cervix. It may be used as a method of induced abortion, a therapeutic procedure used after miscarriage, or a procedure to obtain a sample for endometrial biopsy. Manual Vacuum Aspiration also has the advantage of being quiet, without the noise of an electric vacuum pump. Manual Vacuum Aspiration is an alternative to electric suction curettage for first-trimester elective abortion. The rate of infection is lower than any other surgical abortion procedure at 0.5% (WebMD, 2006). This method may be used earlier in pregnancy than Dilation and Curettage (D&C). Manual Vacuum Aspiration is the only surgical abortion procedure available earlier than the 6th week of pregnancy. It also has lower rates of complications when compared to Dilation and Curettage (D&C). Although many studies have demonstrated that manual vacuum aspiration is safer than sharp curettage for abortion, only a few studies have directly compared it with electric suction curettage.

2.3.2 Dilation and Curettage (D&C)

Dilation and Curettage (D&C) is a surgical abortion procedure used to terminate a pregnancy up to 16 weeks gestation. It is also referred to as suction curettage or vacuum aspiration.

2.4 Post-abortion Complications

The Ghana Health Service has identified and classified post abortion complications into ~~three main~~ groups. These are bleeding/hemorrhage, sepsis/infections, and perforations.

2.4.1 Bleeding/ Hemorrhage –

Heavy bleeding is an expected side effect, but it becomes a complication if hemorrhage necessitates transfusion or surgical evacuation (Kruse et al., 2000). They again posited that hemorrhage is best defined as the loss of large amounts of blood, as indicated by details of sanitary pad saturation, often associated with signs or symptoms of hypovolemia (state of decreased blood volume). The need for a blood transfusion for excessive bleeding in patients after medical abortion is estimated to be about 1 per 1,000 women. According to WHO (1994), the initial treatment of bleeding includes universal measures, oxygen, fluids, medicines and laboratories.

2.4.2 Sepsis/ Infection

Sepsis is a potentially deadly medical condition characterised by a whole-body inflammatory state (called a systemic inflammatory response syndrome or SIRS) caused by severe infection (Bone et al., 1992). Septicemia is a related medical term referring to the presence of pathogenic organisms in the bloodstream, leading to sepsis. The term has not been sharply defined. It has been inconsistently used in the past by medical professionals, for example as a synonym of bacteremia, causing some confusion. Sepsis is caused by the immune system's response to a serious infection, most commonly bacteria, but also fungi, viruses, and parasites in the blood, urinary tract, lungs, skin, or other tissues. Sepsis can be thought of as falling within a continuum from infection to multiple organ dysfunction syndrome (Annane et al., 2005).

Common symptoms of sepsis include those related to a specific infection, but usually accompanied by high fevers, hot, flushed skin, elevated heart rate, hyperventilation, altered mental status, swelling, and low blood pressure. In the very young and elderly, or in people with weakened immune systems, the pattern of symptoms may be atypical, with hypothermia and without an easily localisable infection. According to Munford and Pugin

(2001), in the United States, severe sepsis contributes to more than 200,000 deaths per year.

Sepsis is usually treated with intravenous fluids and antibiotics. If fluid replacement is not sufficient to maintain blood pressure, vasopressors can be used. Mechanical ventilation and dialysis may be needed to support the function of the lungs and kidneys, respectively. To guide therapy, a central venous catheter and an arterial catheter may be placed; measurement of other hemodynamic variables (such as cardiac output, mixed venous oxygen saturation or stroke volume variation) may also be used. Sepsis patients require preventive measures for deep vein thrombosis, stress ulcers and pressure ulcers, unless other conditions prevent this. Some might benefit from tight control of blood sugar levels with insulin. Patel and Balk (2012) indicated that the use of corticosteroids is controversial. Activated drotrecogin alfa (recombinant activated protein C), originally marketed for severe sepsis, has not been found to be helpful, and has recently been withdrawn from sale.

2.4.3 Perforations

This is an accidental puncture of the uterus. Perforation is usually caused by a surgical instrument shaped like a spoon or scoop used for scraping and removing material from an organ (curette) or by an intrauterine device (IUD). There have been some cases of uterine perforation following induced abortion. The instrument penetrates through the uterine wall, and rarely, may migrate into the abdominal cavity where the bowel or bladder may also be perforated.

According to Trupin (2004), a woman's uterus can become perforated during other intrauterine procedures, such as rotating an infant during delivery using forceps, dilation and curettage (D&C) procedures in which the lining of the uterus is scraped, or during a tubal ligation sterilisation procedure. During IUD insertion, perforations can occur when the uterus is abnormally positioned or unusually soft after a birth or abortion. An IUD can also become "lost" within the uterus, and the end of the device may pierce the muscular wall of the uterus. Breastfeeding (lactating) women are at higher risk for perforation of the uterus during insertion of an IUD or with D&C and should be carefully monitored. During dilation and curettage procedures, postmenopausal women are at a higher risk for uterine perforation because the cervix narrows and the wall of the uterus become thinner after menopause.

Individuals who have had past abortions, C-sections, or other surgeries on the cervix have an increased risk for uterine perforation. The internal cervical or may become injured during these procedures, causing scar tissue formation and weakness of cervical tissues that may lead to increased susceptibility to perforation. Uterine perforation occurs in about 1 out of every 250 (0.4%) abortion procedures; the perforation rate is higher when abortions are performed in the second trimester (Trupin, 2004).

2.5 Possible Risk Factors associated with Abortion

Buruh (2011) also conducted ~~a national~~ cross-sectional study in Ethiopia on women aged 15 to 49. This study sought to assess determinants of induced abortion among child bearing age women attending maternal and child health clinic in Mekelle Town, Tigray. A total of 260 women of reproductive age group were interviewed. The study revealed that the main determinants of induced abortion were health problem 56 (21.5%) and child spacing 39 (15%). The younger the women, the higher the risk of induced abortion. The mean age of health institution based induced abortion was 25.19 years. Contraceptive failure was also substantial determinant among the respondents. Marital status and occupation were some of the influential demographic factors. The study concluded that majority of women who experience induced abortion 120 (46.1%) were less than 25 years age. The most frequent determinant of induced abortion reported was health problems 56 (121.5%). When educational level and economy of the women increases, their interest to induce in health institution also increased. It was recommended that health professional, policy makers and leaders should create awareness in the community that the unwanted pregnancy would have cultural, economical and psychological impacts secondary to induced abortion.

Asamoah and Agardh (2012) reported that the fight against maternal deaths has gained attention as the target date for Millennium Development Goal 5 approaches. They revealed that induced-abortion is one of the leading causes of maternal deaths in developing countries which hamper this effort. The data they used was extracted from the Ghana Maternal Health Survey (GSS, 2007). This was a national survey conducted across the 10 administrative regions of Ghana. The survey identified 4,203 female deaths through verbal autopsy, among which 605 were maternal deaths in the 12 to 49 year-old age group. Asamoah and Agardh adopted a case control study design and used cross-tabulations and logistic regression models to investigate associations between the different variables. They found that alcohol consumption was significantly associated with abortion-related maternal deaths. Women who had ever consumed alcohol (OR adjusted 2.6, 95% CI 1.38-4.87), frequent consumers (OR adjusted 2.6, 95% CI 0.89-7.40) and occasional consumers (OR adjusted 2.7, 95% CI 1.29-5.46) were about three times as likely to die from abortion-related causes compared to those who abstained from alcohol. Maternal age, marital status and educational level were found to have a confounding effect on the observed association. They recommended that policy actions directed toward reducing abortion-related deaths should consider alcohol consumption, especially among younger women. Also, policy makers in Ghana should consider increasing the legal age for alcohol consumption, and information on the health risks posed by alcohol and abortion be disseminated to communities in the informal sector where vulnerable groups can best be reached.

Similarly, Yeboah and Kom (2003) examined the number of abortion cases attended to in the Chenard Ward of the Korle Bu Teaching Hospital, Accra during the years 2000 and 2001. A total of 1,935 abortion cases were handled in the year 2000 and 1,838 in 2001. Though they found that there was a 5% decrease in the number of cases in 2001, there was an increase in 'incomplete abortions', which happened to be the most frequent, 78% and 83% in 2000 and 2001 respectively. Majority of the abortions were found among women in the age bracket 21-30: 58% in 2000 and 55% in 2001. There were also 63 (3.3%) and 42 (2.3%) abortions in 2000 and 2001 respectively between the ages of 41 and 50 years. they concluded that the figures call for the intensification of the campaign for safer sex practices, family planning and the teaching that there is good care for those that call to the hospital early enough.

Again, Turpin et al. (2002) conducted a retrospective studies involving 1,301 out of 1,313 cases of abortion admitted to the Gynaecology Ward of the Komfo Anokye Teaching

Hospital (KATH) in Kumasi in Ghana. This was to determine their socio-demographic and clinical characteristics. In their finding, abortions were found to constitute 38.8% of the admissions to the Gynaecology Ward. Again, most (26.5%) patients were between 20-24 years and 25.9% aged between 25 and 29 years with 79.9% of them being married. Traders constituted 43.9% of the patients followed by unemployed with 19.9%.

In a case-control study South Africa by Jewkes et al. (1997), the authors concluded that even though the cases were relatively younger than the controls, there was no significant difference as far as the ages of the cases and controls are concerned. This is suggestive that age may not be a predictor for incomplete abortion. Marital status of the cases and controls, therefore, is presumed to influence decisions by women to desire to be pregnant in the first place, and secondly carry it to term. In this study, majority of women among both cases and controls were unmarried. Women of such category may be uncomfortable with having to being pregnancy yet unmarried (Adanu et al., 2005). The social stigma (Ahiadeke, 2001) and the stress associated with comments that may be made by close and distant associate affect such women negatively. It was worthy to note, that irrespective of the marital status of the cases and controls; this background character does not influence the experiencing of incomplete abortion. Similarly, Turpin et al. (2002) found that the occupational background did not predict the incidence of incomplete abortion among the women suggests, hence other factors could have accounted for it.

In a case-control study by Lokko (2009) in Bosomtwe District, Ashanti, Ghana. There was no difference in the age (p=0.61), marital status (p=0.11), educational status (p=0.71 and years of staying in the district (p=0.37) between cases and controls. Thus, there was no variation in the socio-demographic characteristics between cases and controls which could account for the incidence of incomplete abortion. On economic characteristics difference, Lokko (2009) concluded that there was no difference in the employment status (p=0.52), income earned (p=0.96), partner's occupation (p=0.40), partner's monthly income (p=0.95) and the rating of livelihood of the women (p=0.25). Thus, none of the economic characteristics of the women influence incomplete abortion status in the district.

2.6 Some Applications of Count Data Models in Health

2.6.1 Poisson Distribution

The Poisson distribution is often used to model information on counts of various kinds, particularly in situations where there is no natural "denominator", and thus, no upper bound or limit on how large an observed count can be. This is in contrast to the Binomial distribution which focuses on observed proportions. Possible areas of count data where a Poisson model is useful include: (i) the number of automobile fatalities in a given region over year intervals, (ii) the number of AIDS cases for a given risk group for a series of monthly intervals, (iii) the number of murders in Chicago by year, (iv) the number of server failures for a web-based company by ear, and (v) the number of earthquakes of a certain magnitude in a seismically active region by decade and modeling malaria prevalence or cases fits directly into this context (i.e.) three main factors of malaria prevalence are explored in this chapter, followed by a review of Poisson regression model. Poisson regression assumes that the data follows a Poisson distribution, a distribution frequently encountered when counting a number of events. The distribution was first used to characterise deaths by horse kicks in the Prussian army. Poisson distributions have three special problems that make traditional (i.e. least squares) regression problematic.

1. The Poisson distribution is skewed; traditional regression assumes a symmetric distribution of errors.
2. The Poisson distribution is non-negative; traditional regression might sometimes produce predicted values that are negative.
3. For the Poisson distribution, the variance increases as the mean increases; traditional regression assumes a constant variance.

2.6.2 Negative Binomial Distribution

The weakness of the Poisson distribution in accommodating heavy tails was recognised in the early twentieth century, when Greenwood and Yule (1920) postulated a heterogeneity model for the over-dispersion, in the context of disease and accident frequencies. This is the first appearance of the Negative Binomial as a compound Poisson distribution, as opposed to its derivation as the distribution of the number of failures till the r^{th} success. Newbold (1927), and Arbous and Kerrich (1951) illustrated compound Poisson distributions in the context of modeling industrial accidents. In the actuarial literature, Lundberg (1940) and Seal (1982) further considered the Negative Binomial as a compound Poisson distribution, as a result of heterogeneity of risk over either time or individuals, as a model for claim frequencies. The generalised inverse Gaussian is a three-parameter distribution which is highly flexible, but has the drawback that its computation is complex. Its two-parameter version, the inverse Gaussian, is computationally somewhat simpler. Poisson-inverse Gaussian distribution, which has greater skewness than the Negative Binomial, and so may be more suited to modeling heavy-tailed claim frequency distributions. Willmot (1987) compared their performance in fitting claim frequency distributions, and found that the Poisson-inverse Gaussian was more successful in accommodating the heavy tails than the Negative Binomial. However, this difference appears to be a marginal improvement only and the benefit of the Poisson-inverse Gaussian over the Negative Binomial was disputed by Lemaire (1991). In recent years, the Negative Binomial has gained popularity as the distribution of choice when modeling over-dispersed count data in many fields, possibly because of its simpler computational requirements and its availability in standard software.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

Statistical modeling is about finding general laws from observed data, which amounts to extracting information from the data. According to de Vries (2001), "there is no best model, only better models." However, White and Bennetts (1996) suggested that a good statistical model is the one that provides a good approximate mathematical representation of the data being modelled with particular emphasis being on structure or patterns in the data. This chapter presents theoretical analyses of Poisson regression and Negative Binomial models. It also looks at the data/variable description, coding and analysis plans.

3.2 Data and Variable Description

The data for the study were obtained from the Regional Health Directorate of the Central Region of Ghana after formally producing an introductory letter to the Regional Director of Health Services. The data spanned from 2008 to 2012, collated on monthly bases with variables such as age groups of women performing abortion, marital status, and occupation. The modeling was done by regressing abortion cases (response variable) on age, marital status, occupation and time using the Poisson and Negative Binomial distributions.

3.2.1 Coding Scheme

The age groups were coded as follows: 10 – 14 years (1); 15 – 19 years (2); 20 – 24 years (3); 25 – 29 years (4); 30 – 34 years (5); and 35 years and above (6), while marital status was coded as married (1) and not married (2). The variable occupation was coded as farmer (1), trader (2), public/civil servant (3), artisan (4), student (5) and unemployed (6). Various dynamics of time as in months, quarters and years were also considered.

3.3 Processes of Mathematical/Statistical Modeling

de Vries (2001) recommended the following processes involved in building any mathematical and statistical model:

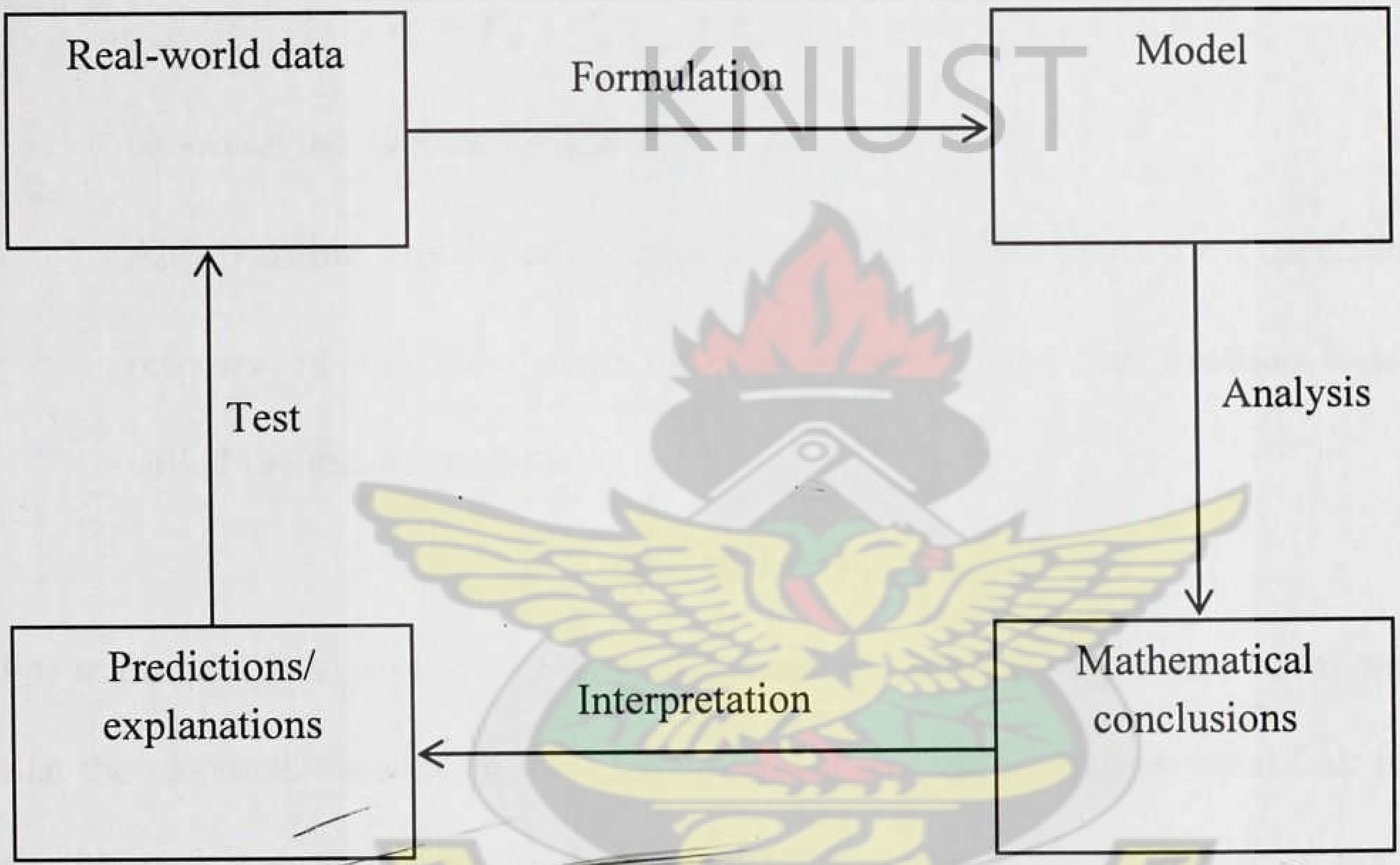

Figure 3.1: Processes of Building Mathematical/Statistical Model

3.3.1 Generalised Linear Models (GLMs)

Generalised Linear Model (GLM) was first introduced by Nelder and Wedderburn (1972). They provided a unified framework to study various regression models, rather than a separate study for each individual regression. Generalised linear models (GLM) are extensions of classical linear models. It includes linear regression models, analysis of variance models, logistic regression models, Poisson regression models, log-linear models, as well as many other models. The above models share a number of unique properties, such as linearity and a common method for parameter estimation. A generalised linear model consists of three components:

1. A random component, specifying the conditional distribution of the response variable, Y given the explanatory variables.
2. A linear function of the regressors, called the linear predictor,

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i_2} + \cdots + \beta_k x_{ik} = x_i \beta \tag{1}$$

   on which the expected value μ_i of Y_i depends.

3. An invertible link function, $g(\mu_i) = \eta_i$ which transforms the expectation of the response to the linear predictor. The inverse of the link function is sometimes called the mean function:

$$g^{-1}(\eta_i) = \mu_i \tag{2}$$

For traditional linear models in which the random component consists of the assumption that the response variable follows the Normal distribution, the canonical link function is the identity link. The identity link specifies that the expected mean of the response variable is identical to the linear predictor, rather than to a non-linear function of the linear predictor. The Generalised Linear Model is an extension of the General Linear Model to include response variables that follow any probability distribution in the exponential family of distributions. The exponential family includes such useful distributions as the Normal, Binomial, Poisson, Multinomial, Gamma, Negative Binomial and others.

3.3.2 The Poisson Distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

The Poisson regression model is a technique used to describe count data as a function of a set of predictor variables. In the last two decades it has been extensively used both in human and in veterinary epidemiology to investigate the incidence and mortality of chronic diseases. Among its numerous applications, Poisson regression has been mainly applied to compare exposed and unexposed cohorts and to evaluate the clinical course of ill subjects.

The distribution was first introduced by Simeon-Denis Poisson (1781–1840) and published together with his probability theory, in 1838 in his work "Recherchessur la probabilite des jugements en matierecriminelle et enmatierecivile (Research on the Probability of Judgements in Criminal and Civil Matters)." The work focused on certain random variables N that count, among other things, the number of discrete occurrences (sometimes called "arrivals") that take place during a time-interval of given length.

If the expected number of occurrences in this interval is λ, then the probability that there are exactly occurrences (k being ~~a~~ non-negative integer, $k = 0, 1, 2\ldots$) is equal to

$$f(k,\lambda) = \frac{\lambda^{k} e^{-\lambda}}{k!} \tag{3}$$

Where:

• e is the base of the natural logarithm (e = 2.71828...),

• k is the number of occurrences of an event - the probability of which is given by the function,

• $k!$ is the factorial of k,

• λ is a positive real number, equal to the expected number of occurrences that occur during the given interval.

For instance, if the events occur on average 4 times per minute, and one is interested in probability for k times of events occurring in a 10 minute interval, one would use as the model a Poisson distribution with $\lambda = 10 \times 4 = 40$.

The parameter λ is not only the mean number of occurrences, k but also its variance:

$$\sigma_k^2 = E(k^2) - [E(k)]^2 \tag{4}$$

Thus, the number of observed occurrences fluctuates about its mean λ with a standard deviation according equation (5):

$$\sigma_k = \sqrt{\lambda} \tag{5}$$

The variance as a function of k is the probability mass function. The Poisson distribution can be derived as a limiting case of the binomial distribution. The Poisson distribution can be applied to systems with a large number of possible events, each of which is rare. A classic example is the nuclear decay of atoms. The Poisson distribution is sometimes called a Poissonian, analogous to the term Gaussian for a Gauss or normal distribution.

Assumptions of Poisson distribution are:

- Observations are independent,

- Probability of occurrence in a short interval is proportional to the length of the interval,
- Probability of another occurrence in such a short interval is zero.

We verify that this Poisson distribution belongs to the exponential family as defined by Nelder and Wedderburn (1972). By taking logs of the Poisson distribution function, we find

$$log\ f_i(y_i) = y_i log(\mu_i) - \mu_i - log(y_i!) \tag{6}$$

Looking at the coefficient of y_i we see immediately from (6) that the canonical parameter is

$$\theta_i = log(\mu_i) \tag{7}$$

and therefore that the canonical link is the log. Solving for μ_i we obtain the inverse link

$$\mu_i = e^{\theta i} \tag{8}$$

and we see that we can write the second term in (8) the pdf as

$$b(\theta_i) = e^{\theta i} \tag{9}$$

The last remaining term in (9) is a function of y_i only, so we identify

$$c(y, \varphi_i) = log(y_i!) \tag{10}$$

Finally, note that we can take $a_i(\varphi)$ and $\varphi = 1$, just as it is in the binomial case. Let us verify the mean and variance. Differentiating the cumulant function $b(\theta_i)$ we have

$$\mu_i = b'(\theta_i) = e^{\theta i} \tag{11}$$

And differentiating again regarding equation (14) we have

$$v_i = a_i(\varphi)b''(\theta_i) = e^{\theta i} = \mu_i \tag{12}$$

Hence, the mean is equal to the variance.

3.3.3 The Exponential Family

GLMs may be used to model variables following distributions in the exponential family with probability density function:

$$f(y;\theta,\varphi) = exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y;\varphi)\right\}$$

or

$$log\, f(y;\theta,\varphi) = \frac{y\theta - b(\theta)}{a(\varphi)} + c(y;\varphi) \qquad (13)$$

Where φ is a dispersion parameter and $a(\varphi)$, $b(\varphi)$ and $c(y,\varphi_i)$ are known functions in equations (14). For distributions in the exponential families, the conditional variance of Y is a function of the mean, μ together with a dispersion parameter, φ.

That is, $E(Y_i) = \mu_i = b'(\theta)$, and

$$var(Y) = \sigma_i^2 = b''(\theta)a(\varphi) \qquad (14)$$

Where $b'(\theta)$ and $b''(\theta)$ are the first and second derivatives of $b(\theta)$. The dispersion parameter is usually fixed to one for some distributions.

Many commonly used distributions in the exponential family are the normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions. In addition, several other distributions are in the exponential family and they include the beta, multinomial, Dirichlet, and Pareto. Distributions that are not in the exponential family but are used for statistical modeling include the Student t and uniform distributions.

3.3.4 Poisson Regression Model

Poisson regression analysis is a technique which allows for modeling dependent variables that describe count data (Cameron & Trivedi, 1998). It is often applied to study the occurrence of small number of counts or events as a function of a set of predictor variables, in experimental and observational study in many disciplines including Economics, Demography, Psychology, Biology and Medicine (Gardener et al., 1995). The Poisson regression model may be used as an alternative to the Cox model for survival analysis, when hazard rates are approximately constant during the observation period and the risk of the event under study is small (e.g. incidence of rare diseases). For example, in ecological investigations, where data are available only in an aggregated form (typically as a count), Poisson regression model usually replaces Cox model, which cannot be easily applied to aggregated data. Furthermore, using rates from an external population selected as a referent, Poisson regression model has often been applied to estimate standardised mortality and incidence ratios in cohort studies and in ecological investigations (Breslow and Day, 1987). Finally, some variants of the Poisson regression model have been proposed to take into account the extra-variability (over-dispersion) observed in actual data, mainly due to the presence of spatial clusters or other sources of autocorrelation (Cameron and Trivedi, 1998). Besides medical studies, the Poisson regression model has been used in different fields of veterinary research, ranging from herd management assessment to animal health in domestic and wild animals and control of infectious diseases in different animal species. The Poisson model has been applied also to data analysis in a multidisciplinary study on cancer incidence in veterinary and other workers of veterinary industry compared to that of other part of active population in

Sweden (Travie et al., 2003). The most recent applications of the Poisson model and of its variations (e.g. Negative Binomial model, Poisson random effect model, Poisson model with autocorrelation terms, etc.) in veterinary medicine are aimed to evaluate: the effect of anthelmintic treatment with eprinomectin at calving on milk production in dairy herds with limited outdoor exposure (Sithole et al., 2006); the periparturient climatic, animal, and management factors influencing the incidence of milk fever in grazing systems in cows (Roche & Berry, 2006); the effects, both positive and negative, of widespread badger culling programs on Mycobacterium bovis tuberculosis in cattle in Britain (Donelly et al., 2006); the seasonality of equine gastrointestinal colic (Archer et al., 2006).

In spite of its recent wide application, Poisson regression model remains partly poor known, especially if compared with other regression techniques, like linear, logistic and Cox regression models.

The Poisson regression model assumes that the sample of n observations y_i are observations on independent Poisson variables Y_i with mean, μ_i.

Note that, if this model is correct, the equal variance assumption of classic linear regression is violated, since the Y_i have means equal to their variances.

So we fit the generalised linear model,

$$log(\mu_i) = x_i\beta \tag{15}$$

We say that the Poisson regression model is a generalised linear model with Poisson error and a log link so that (from 15):

$$\mu_i = \exp(x_i\beta) \tag{16}$$

This implies that one unit increases in an x_i are associated with a multiplication of

μ_i by $\exp(\beta_j)$

3.3.5 Exposure (offset)

Poisson regression model is appropriate for rate data, where rate is a count of events occurring to a particular unit of observations divided by some occurrence of that of exposure. It is given by:

$$\log(E(Y/x)) = \log(\exp osure) + \theta' x \tag{17}$$

Which implies:

$$\log(E(Y/x)) - \log(\exp osure) = \frac{\log(E(Y/x))}{\exp osure} = \theta' x \tag{19}$$

In Poisson regression, this is handled as an offset, where the exposure variable enters on the right hand side of equation (19), but with a parameter estimate constrained to 1.

3.4 Model specification

The primary equation of the model is

$$P(Y_i = y_i) = \frac{e^{\mu} \mu^{y_i}}{y_i!}, y_i = 0,1,2,... \tag{20}$$

The most common formulation of this model is the log-linear specification:

$$\log(\mu_i) = x_i'\beta \tag{21}$$

From (21) the expected number of events per period is given by

$$E\left(\frac{y_i}{x_i}\right) = \mu_i = e^{-x_i\theta} \tag{22}$$

Thus:

$$dE\left(\frac{y_i}{x_i}\right) = \beta e^{x_i\theta} = \beta_i\mu_i \quad (23)$$

The major assumption of Poisson model is

$$E\left(\frac{Y_i}{x_i}\right) = \mu_i = e^{x_i\beta} = var\left(\frac{Y_i}{x_i}\right) \quad (24)$$

This assumption would be tested later on. If $var(\frac{Y_i}{x_i}) > E(\frac{Y_i}{x_i})$ then there is over-dispersion. If, $var(\frac{Y_i}{x_i}) < E(\frac{Y_i}{x_i})$ then under-dispersion has occurred.

3.5 Estimation

Estimation involves estimating the regression parameters specifically using the maximum likelihood estimation.

3.5.1 Maximum Likelihood Estimation

The likelihood function for *n* independent Poisson observations is a product of probabilities given by

$$Pr(y_i) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda_i^{y_i}}{y_i!}, i = 0, 1, 2, \ldots \quad (25)$$

Taking logs and ignoring a constant involving $log(y_i)!$ we find that the log-likelihood function is

$$\log L(\beta) = \sum_{i=1}^{n}\left[-\lambda_i + y_i x_i \beta - \log y_i!\right] \quad (26)$$

$$= \sum_{i=1}^{n} \left[-e^{x_i'\beta} + y_i x_i'\beta - \log y_i! \right] \tag{27}$$

Where

$$y_i = \mu_i = e^{x_i'\beta} \tag{28}$$

The parameters of this equation can be estimated using maximum likelihood method

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \left(y_i - e^{x_i'\beta} \right) x_i = 0 \tag{29}$$

And

$$\frac{\partial^2 L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \left[e^{x_i'\beta} x_i x_i' \right] \tag{30}$$

this is the Hessian of the function and with typical element

$$\frac{\partial^2}{\partial \beta_j \partial \beta_i} = \sum_{i=1}^{n} \left[e^{x_i'\beta} x_{ij} x_{il} \right]; j, l = 1, 2, \ldots p. \tag{31}$$

As

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} \left[e^{x_i'\beta} x_{ij} x_{il} \right] \tag{32}$$

does not involve the y data

$$k_{jl} = E\left(\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} \right) = -\sum_{i=1}^{n} \left[e^{x_i'\beta} x_{ij} x_{il} \right]; j, l = 1, 2, \ldots p. \tag{33}$$

And the information matrix is

$$K = \sum_{i=1}^{n} \left[e^{x_i'\beta} x_i' x_i \right] \tag{34}$$

There is no closed form solution to, $\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n}(y_i - e^{x_i\beta})x_i = 0$ so the MLE for β must be obtained numerically. However, as the Hessian is negative definite for all x and β, the MLE $(\hat{\beta})$ is unique, if it exists. From

$$\frac{\partial^2 L}{\partial \beta_j \beta_l} = -\sum_{i=1}^{n}\left[e^{x_i\beta} x_{ij} x_{il}\right]$$

and:

$$k_{jl} = E\left(\frac{\partial^2 L}{\partial \beta_j \partial \beta_l}\right) = -\sum_{i=1}^{n}\left[e^{x_i\beta} x_{ij} x_{il}\right] \tag{35}$$

$$k_{jlr} = E\left(\frac{\partial^3 L}{\partial \beta_j \partial \beta_l \partial \beta_r}\right) = -\sum_{i=1}^{n}\left[e^{x_i\beta} x_{ij} x_{il} x_{ir}\right] \tag{36}$$

and:

$$k_{jl}^{(r)} = \left(\frac{\partial k_{jl}}{\partial \beta_r}\right) = -\sum_{i=1}^{n}\left[e^{x_i\beta} x_{ij} x_{il} x_{ir}\right], j,l,r = 1,2,\ldots,p. \tag{37}$$

To make matters more transparent, consider the case of a single covariate and an intercept. Then x_i is a scalar observation and

$$L = \sum_{i=1}^{n}\left[-\lambda_i + y_i(\beta_1 + \beta_2 x_i) - \log(y_i)\right] \tag{38}$$

Where

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i) \text{ for } i = 1, 2, 3, \ldots, n \tag{39}$$

The first order conditions, $\frac{\partial L}{\partial \beta} = 0$ yield a system of K equations (one for each β) of the form

$$\sum_{i=1}^{n}(y_i - e^{x_i\beta})x_i = 0 \tag{40}$$

Where $\hat{y}_i = e^{x_i\beta}$ is the fitted value of y_i. The predicted/fitted value has as usual been taken as the estimated value of $E\left(\frac{y_i}{x_i}\right)$. This first order condition tells us that the vector of residual is r orthogonal to the vectors of explicative variables.

3.5.2 Likelihood Ratio Test (G^2) Statistics

A simple test on the overall fit of the model, as an analogue to the F-test in the classical regression model is a Likelihood Ratio test on the "slopes". The model with only the intercept is nothing but the mean of the counts, or

$$\lambda_i = \bar{\lambda}\,\forall \tag{41}$$

Where

$$\bar{y} = \sum_{i=1}^{n}\frac{y_i}{n} \tag{42}$$

The corresponding log-likelihood is:

$$L_R = n\bar{y} + \log(\bar{y})\left(\sum_{i=1}^{n} y_i\right) - \sum_{i=1}^{n}\log y_i! \tag{43}$$

where the R stands for the "restricted" model, as opposed to the "unrestricted" model with $K-1$ slope parameters. The last term in $\sum_{i=1}^{n} \log y_i!$ can be dropped, as long as it is also dropped in the calculation of the maximised likelihood

$$L = \sum_{i=1}^{n}\left[-e^{x_i\beta} + y_i x_i\beta - \log y_i!\right] \tag{44}$$

for the unrestricted model, L_u using $L = e^{x_i\hat{\beta}_i}$. The Likelihood Ratio test is then:

$$LR = 2(L_u\text{-}L_R) \tag{45}$$

and follows a χ^2 distribution with $K - 1$ degrees of freedom.

3.6 The Statistical Model

The canonical treatment of GLMs is McCullagh and Nelder (1989), and this review closely follows their notation and approach. Begin by considering the familiar linear regression model,

$$Y_i = x_i\beta + \varepsilon_i \tag{46}$$

Where, $i = 1, 2, 3, ..n$: Y_i is a dependent variable, x_i is a vector of k independent variables or predictors, β is a $k - by - 1$ vector of unknown parameters and the ε_i are zero-mean stochastic disturbances. Typically, the ε_i are assumed to be independent across observations with constant variance σ_i, and distributed normally. That is, the normal linear regression model is characterised by the following features:

1. **The random component**: identifies the response variable Y_i and assumes a distribution for it: $Y_i \sim P(\mu)$,
2. **Systematic component**: specifies the explanatory or the independent variables for the model:

$$\eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_p x_p \tag{47}$$

The covariates x_i combine linearly with the coefficients to form the linear predictor,

The covariates x_i combine linearly with the coefficients to form the linear predictor,

3. **Link function:** specifies a function of the expected value (mean) of Y_i, which the GLM relates to the explanatory or the independent variables through a prediction equation having a linear form

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{48}$$

The linear predictor $\eta_i = x_i\beta$ is a function of the mean parameter μ_i via a link function, $g(\mu_i)$. Note that for the normal linear model g is an identity.

3.7 Link Functions

In theory, link functions $\eta_i = g(\mu_i)$ can be any monotonic, differentiable function. In practice, only a small set of link functions are actually utilised. In particular, links are chosen such that the inverse link $\mu_i = g^{-1}(\eta_i)$ is easily computed, and so that g^{-1} maps from $X_i\beta = \eta_i \in \Theta$ into the set of admissible values for μ_i. A log link is usually used for the Poisson model, since while $\eta_i = g(\mu_i) \in \Theta$, because Y_i is a count, we have $\mu_i \in 0, 1, 2, \ldots$ For binomial data, the link function maps from $0 < \mu_i < 1$ to $\eta_i \in \Theta$. Examples of link functions that are used are the identity, log, inverse, logit, probit, log-log, complementary log-log, etc. Table 3.1 displays the various link functions that can be used in GLM frame work.

Table 3.1: Exponential Family and their Link Functions

Distribution	Link function	Canonical link	Dispersion	Expectation	Variance
$B(n,\pi)$	$\ln\frac{\pi}{1-\pi}$	$n\ln(1+e^{\theta})$	1	$n\pi$	$n\pi(1-\pi)$
$P(\mu)$	$\ln\mu$	e^{θ}	1	μ	μ
$N(\mu,\sigma^2)$	μ	$\frac{1}{2}\theta^2$	σ^2	μ	1
$G(\mu,v)$	$\frac{-1}{\mu}$	$-\ln(-\theta)$	σ^2	μ	μ^2
$IG(\mu,\sigma^2)$	$\frac{-1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2	μ	μ^3
$NB(\mu,k)$	$\ln\frac{k\mu}{1+k\mu}$	$-\frac{1}{k}\ln(1-ke$	1	μ	$\mu(1+k\mu)$

3.8 Log-linear Models

Suppose that we have a sample of n observations $y_1, y_2, \dots y_n$ which can be treated as realisations of independent Poisson random variables, with $Y_i \sim P(\mu)$ and suppose that we want to let the mean μ_i (and therefore the variance) depend on a vector of explanatory variables, x_i. We could entertain a simple linear model of the form

$$\mu_i = x_i\beta \tag{49}$$

But this model has the disadvantage that the linear predictor on the right hand side can assume any real value, whereas the Poisson mean on the left hand side, which represents an expected count, has to be non-negative.

A straightforward solution to this problem is to model instead the logarithm of the mean using a linear model. Thus, we take logs calculating

$$\eta_i = log(\mu_i) \tag{50}$$

and assume that the transformed mean follows a linear model

$$\eta_i = x_i\beta \tag{51}$$

Thus, we consider a generalised linear model with link log. Combining these two steps in one we can write the log-linear model as

$$log(\mu_i) = x_i\beta \tag{52}$$

In this model the regression coefficient, β_j represents the expected change in the log of the mean per unit change in the predictor, x_j. In other words, increasing x_j. by one unit is associated with an increase of β_j in the log of the mean.

Exponentiating Equation 52 we obtain a multiplicative model for the mean itself:

$$\mu_i = exp(x_i\beta) \tag{53}$$

In this model, an exponentiated regression coefficient $exp(\beta_j)$ represents a multiplicative effect of the $j-th$ predictor on the mean. Increasing x_j by one unit multiplies the mean by a factor, $exp(\beta_j)$. A further advantage of using the log link stems from the empirical observation that with count data the effects of predictors are often multiplicative rather than additive. That is, one typically observes small effects for small counts, and large effects for large counts. If the effect is in fact proportional to the count, working in the log scale leads to a much simpler model.

3.8.1 Fisher Scoring in Log-Linear Models

Fisher scoring algorithm is a form of Newton-Rapson method used in statistics to solve maximum likelihood equations numerically. Nelder and Wedderburn (1972) applied Fisher scoring algorithm to estimate $\hat{\beta}$ in generalised linear models. The Fisher scoring algorithm for Poisson regression models with canonical link would be considered, where it would be modelled as:

$$g(\eta_i) = log(\mu_i) \tag{54}$$

The derivative of the link is easily seen to be

$$g(\eta_i) = \frac{1}{\mu_i} \tag{55}$$

Specifically, given an initial estimate β, the algorithms update it to β^{new} by

$$\beta^{new} = \beta + \left\{ E\left(-\frac{\partial L}{\partial \beta \partial \beta^T} \right) \right\}^{-1} \frac{\partial L}{\partial \beta} \tag{56}$$

Where both derivatives are evaluated at β, and the expectation is evaluated as if β were the true parameter values. β is then replaced by β^{new} and the updating is repeated until convergence.

It can be shown that for a GLM, the updating equation (51) can be rewritten as

$$\beta^{new} = \beta + \left(X^T WX\right)^{-1} X^T W z \tag{57}$$

where z is the n-vector with i^{th} component

$$z_i = (Y_i - \mu_i) g'(\mu_i) \tag{58}$$

and W is the $n \times n$ diagonal matrix with

$$W_i = \left\{g'(\mu_i)^2 b''(\theta_i)\right\}^{1} \quad (59)$$

$$W_i = \left(\mu_i \cdot \frac{1}{\mu_i^2}\right)^{-1} \quad (60)$$

And this simplifies to

$$W_i = \mu_i \quad (61)$$

It is noted that the weight is inversely proportional to the variance of the working dependent variable.

3.9 Tests of Hypothesis

Likelihood ratio tests for log-linear models can easily be constructed in terms of deviances. In general, the difference in deviances between two nested models has approximately in large samples a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the models, under the assumption that the smaller model is correct. One can also construct Wald tests, based on the fact that the maximum likelihood estimator β has approximately in large samples a multivariate normal distribution with mean equal to the true parameter value β and variance-covariance matrix, $var(\hat{\beta})=X'WX$ where X is the model matrix and W is the diagonal matrix of estimation weights.

3.10 Goodness of Fit Test

In order to assess the adequacy of the Poisson regression model you should first look at the basic descriptive statistics for the event count data. If the count mean and variance are very different (equivalent in a Poisson distribution) then the model is likely to be over-dispersed. The model analysis option gives a scale parameter (sp) as a measure of over-dispersion; this is equal to the Pearson's Chi-square statistic divided by the number of observations minus the number of parameters (covariates and intercept).

The variances of the coefficients can be adjusted by multiplying by sp. The goodness of fit test statistics and residuals can be adjusted by dividing by sp. Using a quasi-likelihood approach sp could be integrated with the regression, but this would assume a known fixed value for sp, which is seldom the case. A better approach to over-dispersed Poisson models is to use a parametric alternative model, the Negative Binomial. The deviance (likelihood ratio) test statistic, D^2, is the most useful summary of the adequacy of the fitted model. It represents the change in deviance between the fitted model and the model with a constant term and no covariates; therefore, D^2 is not calculated if no constant is specified. If this test is significant then the covariates contribute significantly to the model. The deviance goodness of fit test reflects the fit of the data to a Poisson distribution in the regression. If this test is significant then a red asterisk is shown by the p-value, and you should consider other covariates and/or other error distributions such as Negative Binomial.

3.10.1 Technical validation

The deviance function is:

$$Deviance = 2\sum_{i=1}^{n} y_i \ln\left[\frac{y_i}{\hat{\mu}_i}\right] - (y_i - \hat{\mu}_i) \tag{62}$$

where y is the number of events, n is the number of observations and $\hat{\mu}_i$ is the fitted Poisson mean. The first term is identical to the binomial deviance, representing 'twice' a sum of observed times log of observed over-fitted. The second term, a sum of differences between observed and fitted values, is usually zero, because MLE's in Poisson models have the property of reproducing marginal totals, as noted above.

The log-likelihood function is:

$$L = \sum_{i=1}^{n} y_i \ln(\hat{\mu}_i) - \hat{\mu}_i - \ln(y_i!) \tag{63}$$

The maximum likelihood regression proceeds by iteratively re-weighted least squares, using singular value decomposition to solve the linear system at each iteration until the change in deviance is within the specified accuracy.

The Pearson's Chi-square residual is:

$$r_p = \frac{\left(y_i - \hat{\mu}_i\right)}{\hat{\mu}_i} \tag{64}$$

For large samples, the distribution of the deviance is approximately a Chi-square with $n - p$ degrees of freedom, where n is the number of observations and p the number of parameters. Thus, the deviance can be used directly to test the goodness of fit of the model. An alternative measure of goodness of fit is Pearson's Chi-square statistic, which is defined as

The Pearson's goodness of fit test statistic is:

$$\chi^2 = \sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \tag{65}$$

According to Cook and Weisberg (1982), the deviance residual is:

$$r_d = sign(y_i - \hat{\mu}_i)\sqrt{deviance(y_i, \hat{\mu}_i)} \tag{66}$$

The Freeman-Tukey, variance stabilised, residual is (Freeman & Turkey, 1950):

$$r_{ft} = \sqrt{y_i} + \sqrt{y_i + 1} - \sqrt{4\hat{\mu}_i + 1} \tag{67}$$

The standardised residual is:

$$r_s = \frac{y_i - \mu_i}{\sqrt{1 - h_i}} \tag{68}$$

where h is the leverage (diagonal of the Hat matrix).

3.11 Over-dispersion and the Negative Binomial Model

The major assumption of the Poisson model is

$$E[y_i / x_i] = \lambda_i = e^{x_i \beta} = \text{var}[y_i / x_i] \tag{69}$$

Implying that the conditional mean function equals the condition variance function.

This is very restrictive. If $E[{}^{y_i}/_{x_i}] < var[{}^{y_i}/_{x_i}]$ then we speak about over-dispersion, and when $E[{}^{y_i}/_{x_i}] > var[{}^{y_i}/_{x_i}]$, we say we have under-dispersion. The Poisson model does not allow for over or under-dispersion. A richer model is obtained by using the Negative Binomial distribution instead of the Poisson distribution. Instead of

$$P_r[Y_i = y_i] = \frac{e^{\lambda_i}\lambda_i^{y_i}}{y_i!} \tag{70}$$

we then use

$$P(Y_i = y_i/\beta, x_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)}\left(\frac{\lambda_i}{\lambda_i + \theta}\right)^{y_i}\left(1 - \frac{\lambda_i}{\lambda_i + \theta}\right)^{\theta} \tag{71}$$

This Negative Binomial distribution can be shown to have conditional mean, λ_i and conditional variance $\lambda_i(1 + \eta^2\lambda_i)$ with $\eta^2 := \frac{1}{\theta}$. Note that the parameter, η^2 is not allowed to vary over the observations. As before, the conditional mean function is modeled as

$$E[Y_i/x_i] = \lambda_i = e^{x_i'\beta} \tag{72}$$

The conditional variance function is then given by

$$\mathrm{var}[Y_i/x_i] = e^{x_i'\beta}\left(1 + \eta^2 e^{x_i'\beta}\right) \tag{73}$$

Using ~~maximum~~ likelihood, we can then estimate the regression parameter β, and also the extra parameter, η. The parameter η measures the degree of over (or under) dispersion. The limit case $\eta = 0$ corresponds to the Poisson model.

3.12 Model Selection Criteria

The procedure for choosing the best model relies on choosing the model with the minimum AIC, AICc, CAIC and BIC. The Akaike Information Criterion (AIC) (Akaike, 1974) is a way of selecting a model from a set of models. The chosen model is the one that minimises the Kullback-Leibler distance between the model and the truth. It is based on information theory, but a heuristic way to think about it is as a criterion that seeks a model that has a good fit to the truth but few parameters. It is defined as:

$$AIC = -2(\ln(\text{likelihood})) + 2K \tag{74}$$

where likelihood is the probability of the data given a model and K is the number of free parameters in the model. Again, the study relied on the Finite Sample Corrected AIC (AICC), Bayesian Information Criterion (BIC) (Schwarz, 1978) and Consistent AIC (CAIC) (Bozdogan, 1987) to select the most plausible model. The AICC is a modification of the AIC by Hurvich and Tsai (1989) and it is AIC with a second order correction for small sample sizes. Burnham and Anderson (1998) insist that since AICC converges to AIC as n gets large, AICC should be employed regardless of the sample size. Thus,

$$AIC = -2(I = \ln(likelihood)) + 2k * (\frac{n}{n-k-1}) \tag{75}$$

where n is the sample size. As n gets larger, AICC converges to $AIC(n-k-1) > n$ as n gets much bigger than k, and so $(n/(n-k-1))$ approaches 1, and so there is really no harm in always using AICC regardless of sample size. Similarly, the BIC is given as follows:

$$BIC = -2In(L) + kIn(n) \tag{76}$$

Where:

k: is the number of parameters in the statistical model, (p+q+P+Q+1).

L: is the maximised value of the likelihood function for the estimated model.

n: is the number of observation, or equivalently, the sample size.

In addition, the Likelihood ratio test (G^2) statistic was a criterion for adjudging the adequacy of models; by selecting the model with the largest G^2 value.

3.13 Data Analysis Plan

Data analysis is done using R-Consol software developed by the R Development Core Team (2008), SPSS and Microsoft Excel. In this software, GLMs are provided by the model setting functions glm (Chambers & Hastie, 1992) in the stats package and glm.nb in the MASS pack (Venables & Ripley, 2002) along with associated methods for diagnostics and inference.

3.14 Chapter summary

This chapter has described the methodological approach adopted in this study. Emphasis was laid on the data type and sources as well as the GLM approach for count data analysis. This approach has several features which make it particularly useful and popular. Aspects of GLM such as the random component, systematic component and link function are also reviewed. In this study, model selection criteria are the Akaike's Information Criterion (AIC) and Bayesian's Information Criterion (BIC) to select the most plausible model.

CHAPTER FOUR

DATA ANALYSIS AND RESULTS

4.1 Introduction

This chapter presents the data collected, techniques employed in the data analysis and the results that emerged from the analysis. It is divided into two sections, namely: preliminary, and modeling and validation. The preliminary analysis dealt with testing the relationship between the predictor and the response variables, while in the modeling and validation section, the Poisson and Negative Binomial regression models were used.

4.2 Preliminary Analysis

Data was collected from the Central Regional Health Directorate after a formal request to the Director was made. The data spanned from 2008 to 2012, collated on monthly bases with variables such as age groups of abortion patients, their marital status and occupations.

It can be seen from Figure 4.1 that most abortion cases were between 20-24 years age group, this was followed by those in the 25-29 years age group. The lowest cases were recorded among 10-14 years age group. Further details are summarised in Table A1 in the Appendix.

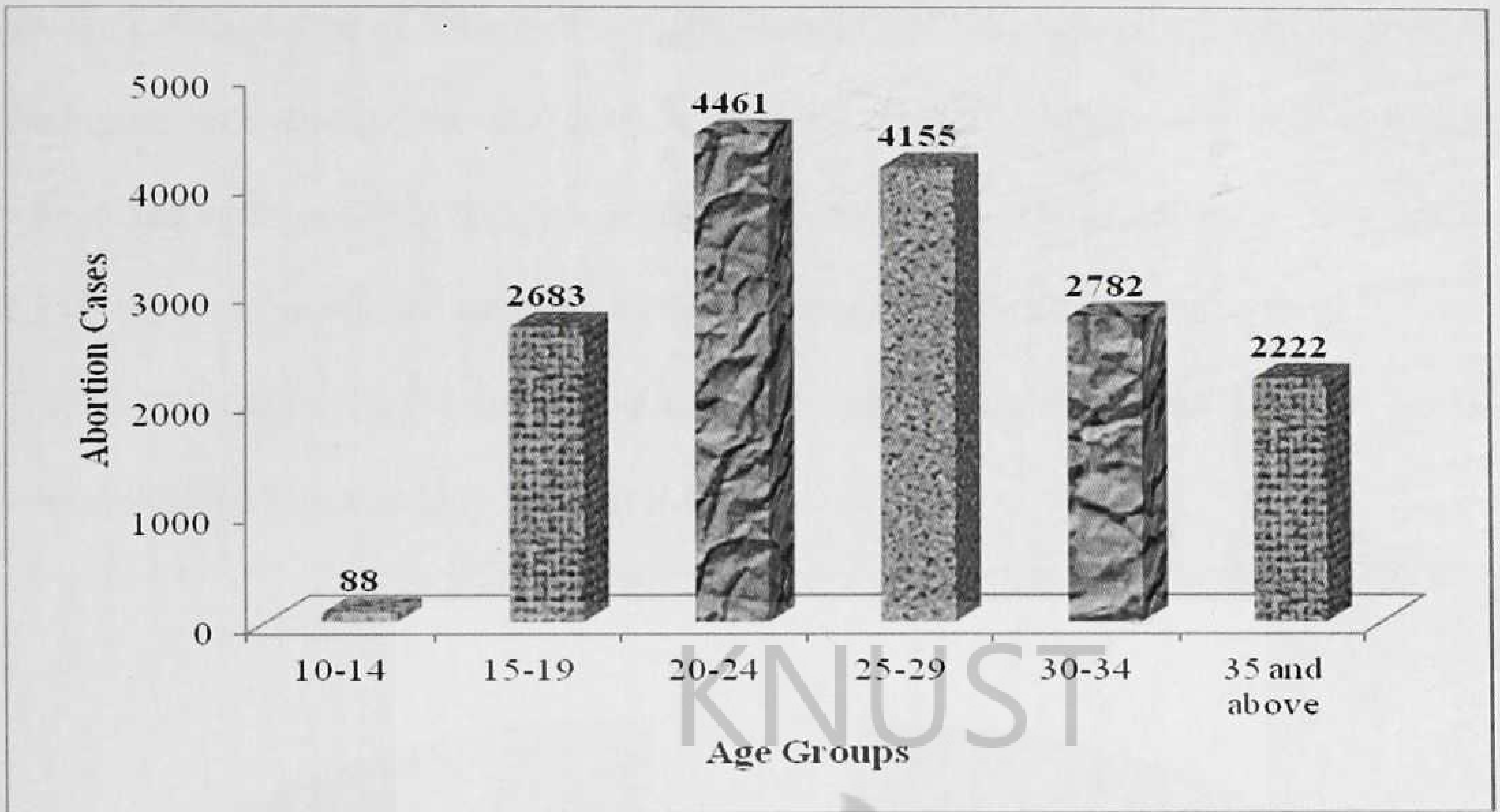

Figure 4.1: Bar chart of Age Group of Abortion Cases (2008-2012) in C/R of Ghana

Evident from Figure 4.2, traders are more involved in abortion than other categories of occupation, followed by the unemployed women. Artisans including seamstresses and bakers placed third position, while public/civil servants are few. Shockingly, students also recorded a sizeable amount of abortions.

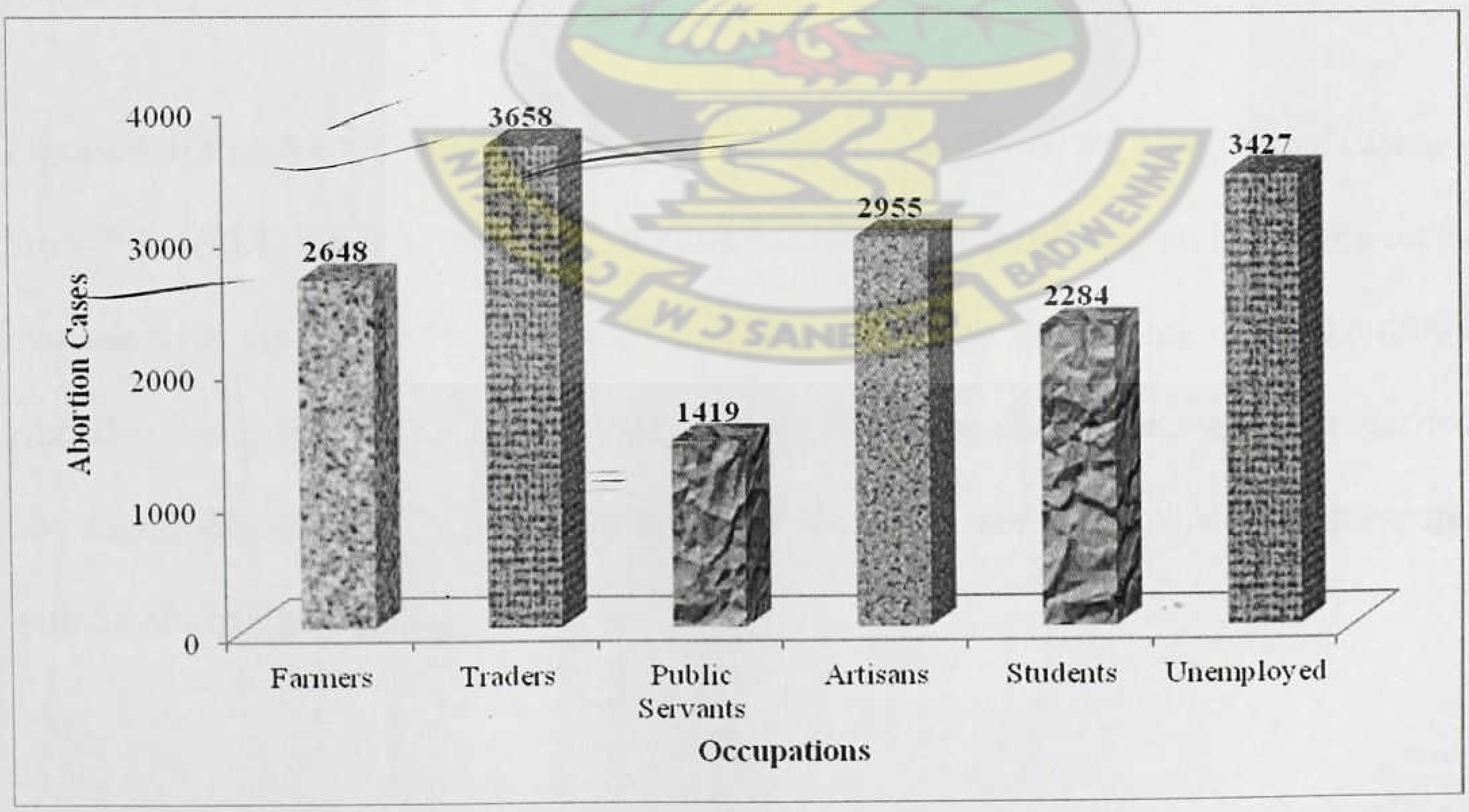

Figure 4.2: Bar chart of Occupations of Abortion Cases (2008-2012)

On the marital status of the abortion cases from Figure 4.3, unmarried women recorded the highest frequency of the total cases with 52.5%. This finding, however, is inconsistent with a finding from a retrospective studies by Turpin et al. (2002) involving 1,301 out of 1,313 cases of abortion admitted to the Gynaecology Ward of the Komfo Anokye Teaching Hospital (KATH) in Kumasi in Ghana that married dominated abortion women who visited the Gynaecology Ward at KATH.

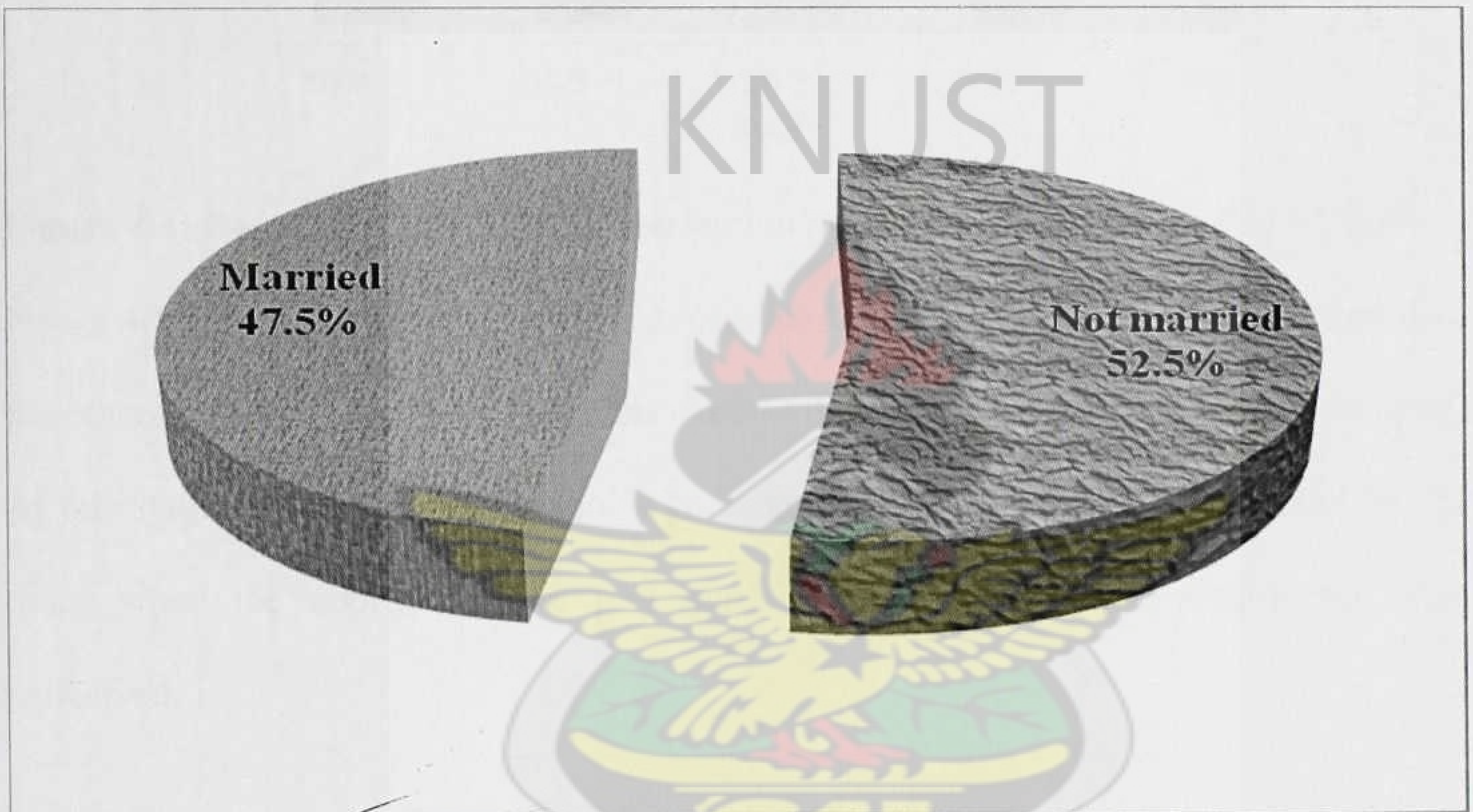

Figure 4.3: Pie chart of Marital Status of Abortion Cases (2008-2012) in C/R of Ghana

From Figure 4.4, it can be seen that abortion cases in the Central Region have been on the increase over the past from 2008 – 2012. This agrees with the finding of Lokko (2009) who also found in her case-control study that 60.5% of the participants were not married. She found that over sixty percent (68.9%) of the cases were not married as compared with 56.6% of the controls.

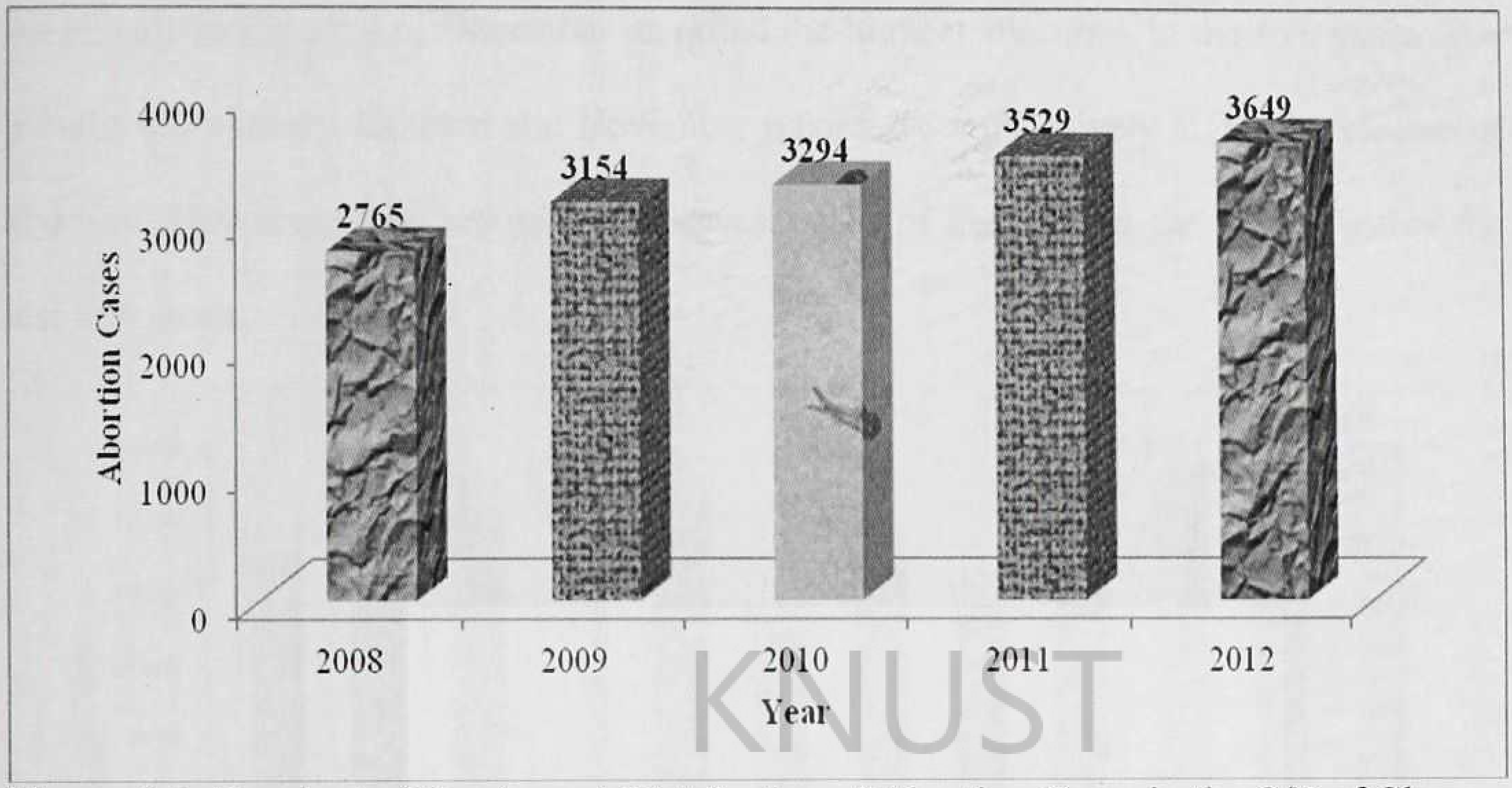

Figure 4.4: Bar chart of the Annual Distribution of Abortion Cases in the C/R of Ghana

Figure 4.5 shows that the fourth quarter recorded the highest prevalence of abortion than the others, followed by the first quarter. Quarter three recorded the least number of cases. At this stage, it is difficult to hazard a guess since these periods only accounted for the times when the abortions were performed and not when these pregnancies were conceived.

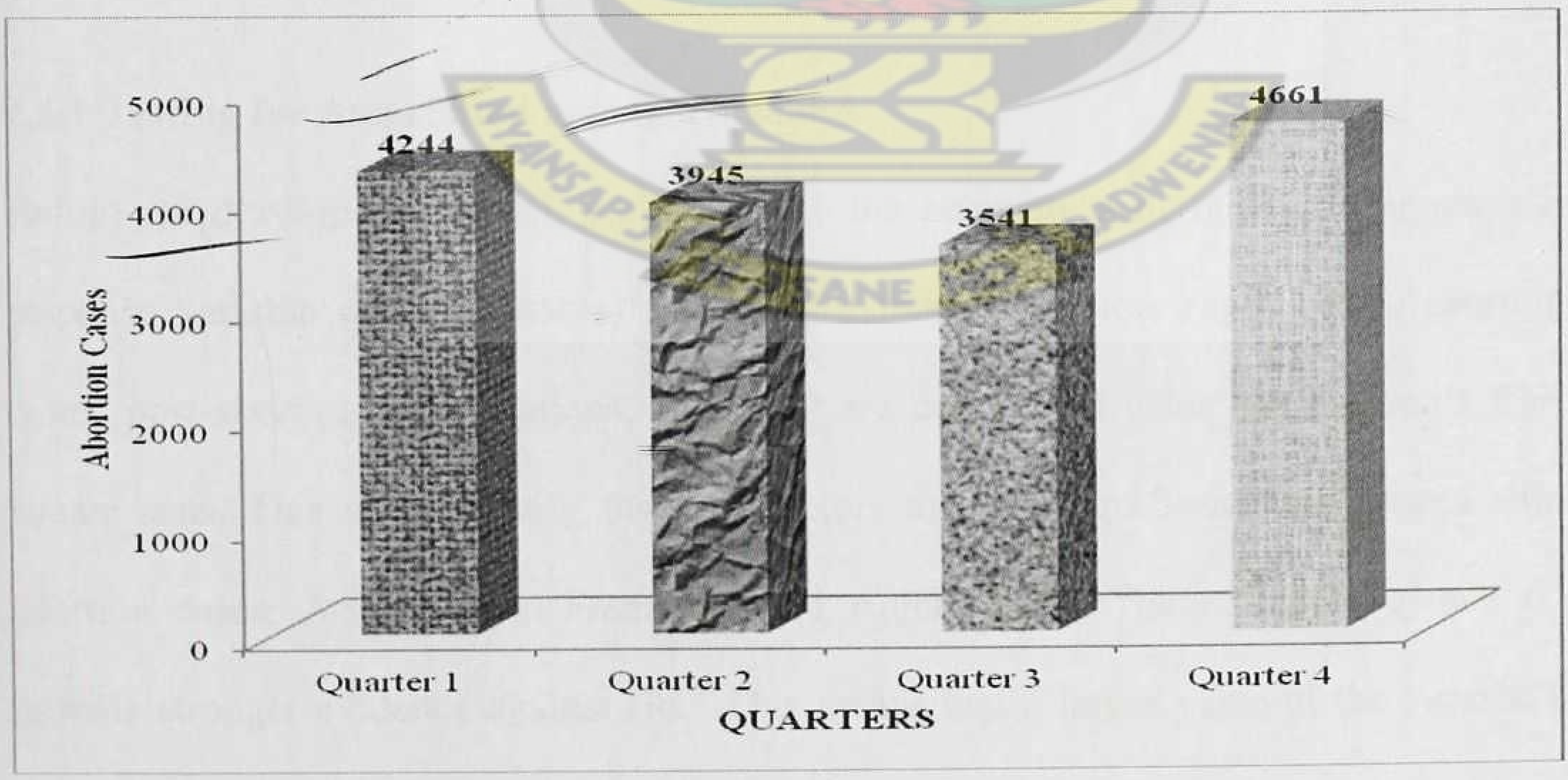

Figure 4.5: Bar chart of Quarterly Distribution of Abortion Cases in C/R of Ghana

As reveals in Figure 4.6, December recorded the highest abortions in the five years. The months of February, October and November recorded comparatively higher incidences of abortion. However, June recorded the lowest cases of abortion in the region within the last five years.

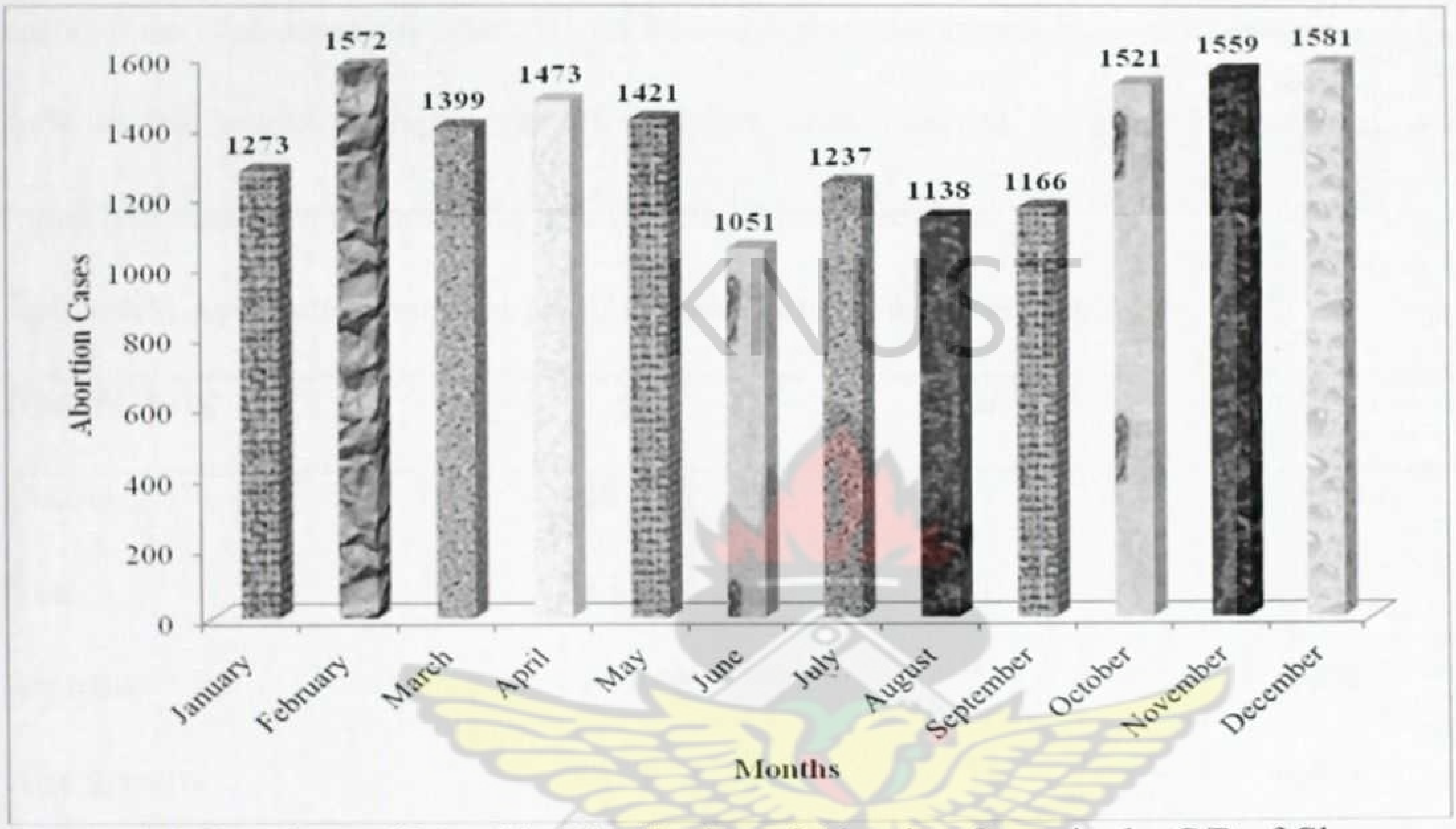

Figure 4.6: Bar chart of Monthly Distribution of Abortion Cases in the C/R of Ghana

4.2.1 Testing for Association among Variables

Before we develop the model, we determine the association or otherwise among the response variable (abortion cases) and the possible risk factors (ages of the abortion cases, post-abortion complications, and time) are determined using the Pearson's Chi-square tests. This is to identify those predictors that are significantly associated with abortion cases. According to Frempong and Adjei (2011), "larger values χ^2 and G^2 provide stronger evidence against Ho." This means that a larger value of the Pearson's Chi-square value signals a high level of association among the variables.

Table 4.1 summarised results from Tables A2, A3, A4, A5, A6 and A7 in the Appendix with the relationships between abortions and predictors including time (months, quarters and years). In selecting time, the Chi-square values are assessed and month is selected since it has the largest values than quarter and year; implying it has the strongest association with abortion cases. It can be concluded that month should be considered for time in the model. Again, ages of abortion cases, marital status and occupation are important variables in modeling abortion cases in the region.

Table 4.1: Association between possible Risk Factors and Abortion Cases

Risk factors	χ^2	df	p
Quarters	38.326	6	0.000
Year	43.050	8	0.000
Months	**66.358**	**22**	**0.000**
Age groups	**547.674**	**10**	**0.000**
Marital status	**6.843**	**2**	**0.033**
Occupation	**236.030**	**10**	**0.000**

4.3 Modeling, Selection and Validation

4.3.1 Model Building

The goal here is to find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome and a set of predictor variables (Frempong & Adjei, 2011). Considering the number of predictors (4) under consideration that are statistically significantly associated with the number of abortion cases recorded between 2008 and 2012 in the Central Region, 14 possible Poisson's regression models can be obtained. This is regressing abortion cases on the following independent variables:

Model 1: Time (months),

Model 2: Age groups,

Model 3: Occupation,

Model 4: Marital status,

Model 5: Age and occupation,

Model 6: Age and marital status,

Model 7: Age and time (months),

Model 8: Occupation and marital status,

Model 9: Occupation and time (months),

Model 10: Marital status and time (months),

Model 11: Age, marital status and occupation,

Model 12: Age, occupation and time (months),

Model 13: Age, marital status and time (months),

Model 14: Occupation, marital status and time (months),

Model 15: Age, occupation, marital status and time (months).

4.3.2 Model Evaluation and Selection

In comparing and selecting 'best' models from the 15 models developed, the study relies on the Likelihood ratio test (G^2), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) and the Likelihood ratio test (G^2). Tables 4.2, 4.3, 4.4 and 4.5 present parameter estimates of the four 'best' Poisson regression models respectively.

Model 5: Abortion cases dependent on Age and Occupation

Figures from Model 5 indicate that all two predictors are significant at 5% significance level with an expected number of incidence of abortion cases in the among those between 20-26 to be 0.87 times higher relatively to those aged at least 35 years *if other variable (occupation) is unchanged.* This is similar to women aged between 25-29 years. This means that abortions are common among age groups 20-24 and 25-29 years than the other age groups. With respect to their occupations, traders are much engaged in abortion as compared to the unemployed if their ages are constant.

Table 4.2: Parameters Estimates from Model 5

Parameter	Estimate	df	S.E	Wald χ^2	p
intercept	1.358	1	0.0261	2711.348	.000
age 10-14	-1.234	1	0.1104	125.062	.000
age 15-19	0.189	1	0.0287	43.199	.000
age 20-24	0.697	1	0.0260	720.489	.000
age 25-29	0.626	1	0.0263	567.172	.000
age 30-34	0.225	1	0.0285	62.407	.000
Farmers	-0.257	1	0.0259	98.831	.000
Traders	0.066	1	0.0238	7.682	.006
Public/civil servants	-0.881	1	0.0316	778.993	.000
Artisans	-0.148	1	0.0251	34.534	.000
Students	-0.443	1	0.0273	263.609	.000

AIC = 2253; BIC = 2260; $G^2 = 3100.098$*; df = 10; p < .05*

Model 11: Abortion cases dependent on Age, Marital status and Occupation

Similar to Model 5, age group, occupation and marital status are significantly associated with abortion cases in the region. Here again, age groups 20-24 and 25-29 years remain more prone to abortion compared to those in age group 35 and above provided that marital status and occupation are intact. With regards to their marital status, Table 4.3 reveals that the difference in the logs of expected counts is expected to be 0.906 abortion cases lower for the marital women compared to the unmarried (all things being equal).

Table 4.3: Parameters Estimates from Model 11

Parameter	Estimate	df	S.E	Wald χ^2	*p*
intercept	1.406	1	0.0271	2688.352	.000
age 10-14	-1.234	1	0.1104	125.226	.000
age 15-19	0.189	1	0.0287	43.199	.000
age 20-24	0.697	1	0.0260	720.489	.000
age 25-29	0.626	1	0.0263	567.172	.000
age 30-34	0.225	1	0.0285	62.407	.000
Married	-0.099	1	0.0156	39.754	.000
Farmers	-.257	1	0.0259	98.831	.000
Traders	.0.066	1	0.0238	7.682	.006
Public/civil servants	-0.881	1	0.0316	778.994	.000
Artisans	-0.148	1	0.0251	34.534	.000
Students	-0.443	1	0.0273	263.610	.000

AIC = 2249; BIC = 2257; $G^2 = 3139.900$; *df = 11; p < .05*

Model 12: Abortion cases dependent on Age, Occupation and Time (months)

Like Models 5 and 11, Model 12 has all levels of the predictors significant except time (months) whose April, October and November are not with respect to December abortion figure at 5% significance level. In adding time (months), abortion remains higher among 20-24 and 25-29 age group women in the region.

Table 4.4: Parameters Estimates from Model 12

Parameter	Estimate	df	S.E	Wald χ^2	p
intercept	1.505	1	0.0354	1809.012	.000
age 10-14	-1.234	1	0.1104	125.056	.000
age 15-19	0.188	1	0.0287	43.064	.000
age 20-24	0.696	1	0.0260	715.969	.000
age 25-29	0.626	1	0.0263	566.608	.000
age 30-34	0.224	1	0.0285	62.242	.000
Farmers	-0.257	1	0.0259	98.517	.000
Traders	0.066	1	0.0238	7.768	.005
Public/civil servants	-0.881	1	0.0316	778.201	.000
Artisans	-0.148	1	0.0251	34.580	.000
Students	-0.444	1	0.0273	263.057	.000
January	-0.123	1	0.0377	10.685	.001
February	-0.083	1	0.0356	5.440	.020
March	-0.132	1	0.0367	12.856	.000
April	-0.071	1	0.0362	3.818	.051

May	-0.107	1	0.0366	8.520	.004
June	-0.408	1	0.0398	105.254	.000
July	-0.245	1	0.0380	41.783	.000
August	-0.329	1	0.0389	71.530	.000
September	-0.304	1	0.0386	62.214	.000
October	-0.039	1	0.0359	1.160	.281
November	-0.014	1	0.0357	0.154	.695

AIC = 2230; BIC = 2246; $G^2 = 3355.500$*; df = 21;* $p < .05$

Model 15: Abortion cases dependent on Age, Occupation, Marital status and Time (months)

This model has in it all four independent variables. Again, ages of the abortion cases, marital status, occupation and some levels of time are statistically significant at 5% significance level.

Table 4.5: Parameters Estimates from Model 15

Parameter	Estimate	df	S.E	Wald χ^2	*p*
intercept	1.553	1	0.0362	1845.000	.000
age 10-14	-1.235	1	0.1104	125.230	.000
age 15-19	0.188	1	0.0287	43.064	.000
age 20-24	0.696	1	0.0260	715.958	.000
age 25-29	0.626	1	0.0263	566.606	.000
age 30-34	0.224	1	0.0285	62.242	.000

Farmers	-0.257	1	0.0259	98.535	.000
Traders	.066	1	0.0238	7.764	.005
Public/civil servants	-0.881	1	0.0316	778.250	.000
Artisans	-0.148	1	0.0251	34.602	.000
Students	-0.444	1	0.0273	264.084	.000
Married	-0.099	1	0.0156	39.767	.000
January	-0.123	1	0.0377	10.721	.001
February	-0.083	1	0.0356	5.423	.020
March	-0.132	1	0.0367	12.856	.000
April	-0.071	1	0.0362	3.818	.051
May	-0.107	1	0.0366	8.520	.004
June	-0.408	1	0.0398	105.254	.000
July	-0.245	1	0.0380	41.783	.000
August	-0.329	1	0.0389	71.530	.000
September	-0.304	1	0.0386	62.214	.000
October	-0.039	1	0.0359	1.160	.281
November	-0.014	1	0.0357	0.154	.695

AIC = 2226; BIC = 2240; $G^2 = 3395.315$; $df = 22$; $p < .05$

Using the AIC and BIC criteria which state, "the smaller the AIC and BIC for a model, the better the fit," Table 4.6 summarises the various AICs and BICs of the four 'good" Poisson regression models. Also are their respective Likelihood ratio (G^2) statistics, whose condition is that the best model must have the largest G^2 value.

While Model 4 has AIC and BIC values of 2253 and 2260 respectively with a relatively larger Likelihood ratio statistic of 3100.098, Model 11 comes with comparatively minimum AIC and BIC values of 2249 and 2257 respectively. Also, its Likelihood ratio statistic of 3139.900 is larger than that of Model 4. Similarly from Model 12, AIC of 2230, BIC of 2246 and G^2 are obtained. These values indicate that Model 12 appears to be better than the previous two models. This presupposes that Model 12 is the best model for statistically understanding abortion dynamics in the Central Region of Ghana although it has failed the parsimonious test.

Table 4.6: Assessment of Models

Model	df	*p*	AIC	BIC	G^2
5	10	000	2253	2260	3100.098
11	11	.000	2249	2257	3139.900
12	21	.000	2230	2246	3355.500
15	**22**	**.000**	**2226**	**2240**	**3395.315**

Therefore, the recommended Poisson regression model from Table 4.5 is given as:

$$\begin{aligned}
In(mean\ abortion\ cases) &= 1.553 - 1.235age(10-14) + 0.188age(15-19) \\
&+ 0.696age(20-24) + 0.626age(25-29) \\
&+ 0.224age(30-34) - 0.257(farmer) + 0.066(trader) \\
&- 0.881(public\ or\ civil\ servant) - 0.148(artisan) \\
&- 0.444(student) - 0.099(married) - 0.123(january) \\
&- 0.083(february) - 0.132(march) - 0.107(may) \\
&- 0.408(june) - 0.245(july) - 0.329(august) \\
&- 0.304(september)
\end{aligned}$$

4.3.3 Model Validation using the Negative Binomial Model

After the identification of Model 15 as most adequate Poisson regression model, the shortcoming especially over-dispersion of the model has not been overlooked. Therefore, from Table 4.5, the Poisson model is validated using the Negative Binomial regression model. The criteria here is that the best model must have the minimum AIC, BIC, deviance and scaled deviance. Clearly from Table 4.7, the Negative Binomial model produced the least values of AIC, BIC, deviance and scaled deviance. This is an indication that the Negative Binomial model should rather be adopted in explaining incidence of abortion cases in the region.

Table 4.7: Validation of Model

Assessment criteria	Poisson Model	Negative Binomial Model
AIC	2226	1858
BIC	2240	1873
Deviance	1231	322.7
Scaled deviance	1231	322.7

Hence, the parameter estimates from the Negative Binomial model involving all possible risk factors are presented in Table 4.8. The expected number of incidence of abortion cases for women aged between 20-24 years is 99.9% since $e(0.693) = 1.9997$ higher than for those aged between 35 and above year if all other things are constant in the model. A similar figure of 91.9% was recorded for those aged 25-29 years. This again confirms the earlier finding in this study. Again, the incidence abortion rate for married is comparatively lower than those not married by 0.837 cases since $e(-0.178) = 0.8367$ if other variables do not changed. The expected abortions recorded in for traders are about 5% higher than for the unemployed if all other things remain unchanged. However, the difference in the logs of expected counts is expected to be 75.0%, 39.9%, 86.5% and 56.2% cases lower for farmers, public/civil servants, artisans and students respectively relative to those who are unemployed (if all other predictors are the same). In terms of time, the incidence rate ratio (IRR) indicates that abortion cases are relatively lower in January, February, March, May, June, July, August and September by 0.896, 0.832, 0.986, 0.898, 0.664, 0.778, 0.729 and 0.740 cases respectively as compared to December, holding other variables unchanged. Statistically, April, October and November are insignificant at 5% significance level, hence their exclusions from the final model.

Table 4.8: Parameters Estimates from Negative Binomial Model

Parameter	Estimate	df	S.E	Wald χ^2	*p*
intercept	1.624	1	0.0877	342.820	.000
age 10-14	-1.187	1	0.1552	58.500	.000
age 15-19	0.134	1	0.0623	4.622	.032
age 20-24	0.693	1	0.0604	131.652	.000
age 25-29	0.652	1	0.0600	118.267	.000
age 30-34	0.236	1	0.0607	15.098	.000
Farmers	-0.288	1	0.0638	20.376	.000
Traders	0.049	1	0.0629	0.607	.436
Public/civil servants	-0.920	1	0.0668	189.823	.000
Artisans	-0.145	1	0.0635	5.191	.023

Students	-0.576	1	0.0661	75.837	.000
Married	-0.178	1	0.0385	21.344	.000
January	-0.110	1	0.0919	1.440	.230
February	-0.084	1	0.0887	0.890	.345
March	-0.140	1	0.0897	2.448	.118
April	-0.023	1	0.0899	0.064	.801
May	-0.108	1	0.0901	1.430	.232
June	-0.410	1	0.0914	20.116	.000
July	-0.251	1	0.0906	7.675	.006
August	-0.316	1	0.0910	12.088	.001
September	-0.301	1	0.0909	10.969	.001
October	-0.031	1	0.0897	0.116	.734
November	-0.035	1	0.0896	0.152	.697

AIC = 2226; BIC = 2240; $G^2 = 3395.315$*; df = 22;* $p < .05$

From Table 4.8, the most adequate model for abortion cases in the region is given by the Negative Binomial distribution as:

$$\begin{aligned}
In(\textit{mean abortion cases}) &= 1.624 - 1.187age(10-14) + 0.134age(15-19) \\
&+ 0.693age(20-24) + 0.652age(25-29) \\
&+ 0.236age(30-34) - 0.288(farmer) + 0.049(trader) \\
&- 0.920(\textit{public or civil servant}) - 0.145(artisan) \\
&- 0.576(student) - 0.178(married) - 0.110(january) \\
&- 0.084(february) - 0.140(march) - 0.108(may) \\
&- 0.410(june) - 0.251(july) - 0.316(august) \\
&- 0.301(september)
\end{aligned}$$

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATONS

5.1 Introduction

This chapter presents the conclusion drawn from based on the findings and the recommendations made for appropriate intervention programmes to reduce abortion cases in particular and maternal mortalities in general in the region.

5.2 Conclusions

The study attempted to model the incidence of abortion cases in the Central Region of Ghana based on predictors like age groups of abortion cases, marital status, occupation, and time (months, quarters and years) using the Poisson and Negative Binomial regression models. Data for the study were obtained from the Central Regional Health Directorate from 2008 to 2012 organised on monthly bases. From the data exploratory analysis, it was revealed that abortion was significantly associated with age of cases, marital status, occupation and time (quarters only).

In selecting the best Poisson regression model that fitted abortion cases, statistical model selection criteria like AIC, BIC and G^2 were used. The Poisson model selected involved all four predictors, namely, age group, marital status, occupation and time (quarters). However, comparing the Poisson regression model with the Negative Binomial model, the model produced by the Negative Binomial regression was adopted, which also involved all four predictors

From the Negative Binomial regression model, the expected number of incidence of abortion cases for women aged between 20-24 years was 99.9% higher with reference to those aged at least 35 year if all other things are constant in the model. A similar figure of 91.9% was recorded for those aged 25-29 years *(ceteris paribus).* This again confirms the earlier finding by this study and Turpin et a. (2002) whose study on abortion at Komfo Anokye Teaching Hospital revealed that abortion was common among females aged between 20-24 years (26.5%), followed by those within 25-29 years (25.9%).

Abortion among married women was less as compared to those not married by 83.7% if other variables do not changed. The expected abortion recorded in for traders was found to be 5% higher for the unemployed women. However, the incidence ratio rates for farmers, public or civil servants, artisans and students were 75.0%, 39.9%, 86.5% and 56.2% lower respectively with regards to those who were unemployed (if all other risk factors are the same). December recorded the highest cases of abortion over the five-year period.

Based on the findings of the study, the following conclusions are drawn:

- The main risk factors in abortion are ages of cases, occupation, marital status and time (months) *(in order of importance).*
- The model can be used to forecast abortion cases in the region.
- Incidence rates of abortion were 99.9% and 91.9% higher for women aged between 20-24 and 25-29 respectively than those aged at least 35 years.
- Traders and the unemployed women are more prone to abortion than other categories of professional.
- Unmarried women are more likely to terminate their pregnancies than their married counterparts.

- Higher cases of abortion are recorded in the first and last quarters in the years under study.

5.3 Recommendations

On the basis of the above findings, the following recommendations are made:

- The Central Regional Health Directorate could used the model to predict abortion cases.
- The risk factors (such as age, occupation and marital status) identified in this study should guide all intervention programmes aimed at reducing abortion and maternal mortalities in the region.
- Women especially aged between 20-29 years should be given special sex and reproductive health education in order to curb the high abortion incidence with that cohort.
- Traders and the unemployed in the region must be advised to adopt safe sexual practices to avoid unwanted pregnancies.
- Direct policies targeted at the unmarried women should be formulated to help them avoid unprotected sex, while married women must also work towards avoiding unwanted pregnancies.
- The Regional Health Directorate should have effective post-abortion care facilities to handle any post-abortion complications as a result in the region to avoid abortion-related deaths.
- The study could not unearth reasons for the increasing abortion cases, and the high incidence of abortions in the first and fourth quarters of every year in the

region. It is, therefore, suggested that further studies be carried out into the reasons accounting for this phenomenon.

REFERENCES

Adanu R. M. K., Ntumy, M. N. E., and Tweneboah E. (2005). Profile of women with abortion complications in Ghana. *Trop Doct, 35,* 139-142.

Ahiadeke, C. (2001). Incidence of induced abortion in Southern Ghana. *International Family Planning Perspectives, 27*(2), 96-108.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control, AC-19*, 716-723.

Annane, D., Bellissant, E., and Cavaillon, J. M. (2005). Septic shock. *Lancet, 365*(9453), 63-78.

Arbous, A. G., and Kerrich, J. E. (1951). Accident statistics and the concept of accident proneness. *Biometrics, 7,* 340.

Archer, D. C., Pinchbeck, G. L., Proudman, C. J., and Clough, H. E. (2006). Is equine colic seasonal? Novel application of a model based approach. *BMC Veterinary Research, 2,* 27.

Asamoah, B. O., and Agardh, A. (2012). Alcohol consumption in relation to maternal deaths from induced-abortions in Ghana. *Reproductive Health, 9*(10).

Berer, M. (2000). Making abortions safe: a matter of good public health policy and practice. *Bull. World Health Organ., 78*(5), 580-592.

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M., and Sibbald, W. J. (1992). *Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Chest., 101*(6), 1644-1655.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345-370.

Breslow, N., and Day, N. (1987). Statistical methods in cancer research: The design and analysis of cohort studies. Lyon, France.

Burnham, K. P., and Anderson, D. R. (1998). *Model selection and inference: A practical information-theoretic approach.* New York: Springer-Verlag.

Buruh, G. (2011). *Assessment of determinants of induced abortion among child bearing age women attending maternal and child health clinic in Mekelle Town, Tigray, Ethiopia.* A thesis submitted to the School of Graduate Studies, Addis Ababa University.

Cameron, A. C., and Trivedi, P. K. (1998). Regression analysis of count data. *Econometric Society Monographs, 30,* Cambridge University Press.

Cameron, A. C., and Trivedi, P. K. (1998). *Regression analysis of count data.* New York: Cambridge University Press.

Chambers, J. M., and Hastie, T. J. (1992). Statistical Models in S. London: Chapman and Hall.

Churchill Livingstone (2008). *Churchill Livingstone medical dictionary.* Edinburgh New York: Churchill Livingstone Elsevier.

Cook, R. D., and Weisberg, S. (1982). *Residuals and influence in regression.* London: Chapman and Hall.

de Vries, G. (2001). *What is mathematical modeling?* Department of Mathematical Sciences, University of Alberta, Canada.

Donnelly, C. A., Woodroffe, R., Cox, D. R., Bourne, F. J., and Cheeseman, C. L. (2006). Positive and negative effects of widespread badger culling on cattle tuberculosis. *Nature, 439*, 843-846.

Frempong, N. K., and Adjei, I. A. (2011). *MATH 518: Statistical model 2.* M.Sc. (Industrial Mathematics), Institute of Distance Learning (IDL), KNUST, Kumasi, Ghana.

Gardner, W., Mulvey, E. P., and E.C. Shaw, E. C. (1995). Regression Analysis of Count and Rates: Poisson, and Negative Binomial Models. *Psychological Bulletin, 118,* 392-404.

Greenwood, M., and Yule, G. U. (1920). An inquiry into the nature and frequency distribution of multiple happenings with particular reference to the occurrence of multiple attacks o disease or reported accidents. *Journal of Royal Statistical Society,* 228-279.

Grimes, D. A., Benson, J., Singh, S., Romero, M., Ganatra, B., Okonofua, F. E., and Shah, I. H. (2006). Unsafe abortion: The preventable pandemic (PDF). *The Lancet, 368*(9550), 1908-1919.

GHS (2011). *Central Region Annual Report.* Cape Coast: RHD.

GSS (2011). *2010 Population and Housing Census (provisional results).* Accra: GSS.

GSS (2008). *Ghana Demographic and Health Survey*. Accra: GSS.

GSS (2007). *Ghana Maternal Health Survey.* Accra: GSS.

Hurvich, C. M., and Tsai, C. – L. (1989), Regression and time series model selection in small samples. *Biometrika, 76,* 297–307.

Kruse, B., Poppema, S., Creinin, M. D., and Paul, M. (2000). Management of side effects and complications in medical abortion. *Am. J. Obstet Gynecol., 183*(2).

Lassey, A., and Wilson, J. B. (1998). Trends in maternal mortality in the Korle-Bu Teaching Hospital, in 1984–1994. *Ghana Medical Journal, 32a,* 910-916.

Lemaire, J, (1991). Negative binomial or Poisson inverse Gaussian? *ASTIN Bulleting,* 21, 167-168.

Lokko, C. N. M. (2009). Abortion in Bosomtwe District, Ashanti, Ghana: A case control study at St. Michael's Hospital, Jachie-Pramso. An MPH thesis, KNUST, Kumasi, Ghana.

Lundberg, O. (1940). On random processes and their application to sickness and accident statistics. Uppsala. Mc: Almquist and Wiksell.

Jewkes, R., Wood, K., and Maforah, F. (1997). Backstreet abortion: Women's experiences. *Curationis, 20*(2), 70-82.

McCullagh, P., and Nelder J. A. (1989). Generalised Linear Models (2nd ed.). New York: Chapman and Hall.

Munford, R. S., and Pugin, J. (2001). Normal responses to injury prevent systemic inflammation and can be immunosuppressive. *Am J Respir Crit Care Med, 163,* 316-321.

McBride, W. Z. (1991). Spontaneous abortion. *Am Fam Physician, 43*(1), 175-182.

Ministry of Health (2011). *More girls cause abortion.* Accra: Ministry of Health.

Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalised linear models. *Journal of the Statistical Society, 135*(3), 370.

Newbold, E. M. (1927). Practical applications of the statistics of repeated events particularly to industrial accidents. *Journal of the Royal Statistical Society, 90,* 487-547.

Office of National Statistics of UK (2011). *Abortion statistics, England and Wales, 2010.* Wales: Department of Health.

Patel, G. P., and Balk, R. A. (2012). Systemic steroids in severe sepsis and septic shock. *Am J Respir Crit Care Med., 185*(2), 133-139.

R Development Core Team (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Roche, J. R., and Berry, D. P. (2006). Periparturient climatic, animal, and management factors influencing the incidence of milk Fever in grazing systems. *Journal of Dairy Science, 89*(7), 2775-2783.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Seal, H. L. (1982). Mixed Poisson – an ideal distribution of claim? MVSV, 293-295.

Sithole, F., Dohoo, I., Leslie, K., DesCôteaux, L., Godden, S., Campbell, J., Keefe, G., and Sanchez, J. (2006). Effect of eprinomectin pour-on treatment around calving on reproduction parameters in adult dairy cows with limited outdoor exposure. *Preventive Veterinary Medicine, 75*(3-4), 267-279.

Travier, N., Gridley, G., Blair, A., Dosemenci, M., Boffetta, P. S. (2003). *Cancer incidence among male Swedish veterinarians and other workers of the veterinary industry: A record-linkage study.* Unit of Environmental Cancer Epidemiology, International Agency for Research on Cancer, Lyon, Pub med.

Trupin, S. R. (2004). Abortion. *eMedicine. Retrieved on June 15, 2012 from* http://emedicine.com/med/topic5.htm.

Turpin, C. A., Danso, K. A., and Odoi, A. T. (2002). Abortion at Komfo Anokye Teaching Hospital. *Ghana Medical Journal, 36*(2), 60-64.

UN (2007). *The Millennium Development Goals Report*. New York: United Nations.

Venables, W. N., and Repley, B. D. (2002). *Statistics and computing: Modern applied statistics* (4th ed.). New York: Springer.

WebMD. (2006). *Manual and vacuum aspiration for abortion: A-Z health guide.* Retrieved on May 18, 2012.

White, G. C., and Bennetts, R. E. (1996). Analysis of frequency count data using the negative binomial distribution. *Ecology, 77,* 2549-2557.

WHO (1997). *Abortion: A tabulation of available information.* Geneva: WHO.

WHO (1994). Clinical management of abortion complications: A practical guide. Geneva: WHO.

Willmot, G. E. (1987). The Poisson inverse Gaussian distribution as an alternative to Negative Binomial. *Sandinavian Actuarial Journal,* 113-127.

Yeboah, R. W. N., and Kom, M. C. (2003). Abortion: The case of Chenard Ward, Korle-Bu from 2000 to 2001. *Research Review, 19*(1), 57-66.

APPENDIX

Table A1: Age Distribution of Abortion Cases

Age group	Frequency	Percentage
10 – 14	88	0.5
15 – 19	2683	16.4
20 – 24	4461	27.2
25 – 29	4155	25.3
30 – 34	2782	17.0
35 and above	2222	13.6
Total	**16391**	**100.0**

Table A2: Cross-classification of Abortion Cases by Age Groups

	Abortion Cases			
Age groups	0 – 5	6 – 10	11 and above	Total
10 – 14	706	14	0	720
15 – 19	517	143	60	720
20 – 24	385	217	118	720
25 – 29	382	239	99	720
30 – 34	541	136	43	720
35 and above	589	101	30	720
Total	3120	850	350	4320

$\chi^2 = 547.674$; $G^2 = 307.741$; df = 10; $p = .05$

Table A3: Cross-classification of Abortion Cases by Marital Status

Married status	Abortion Cases 0 – 5	6 – 10	11 and above	Total
Married	1604	401	161	2115
Not married	1516	444	189	2205
Total	3120	850	350	4320

$\chi^2 = 6.843$; $G^2 = 6.984$; df = 2; $p = .033$

Table A4: Cross-classification of Abortion Cases by Occupation

Occupation	Abortion Cases 0 – 5	6 – 10	11 and above	Total
Farmer	528	154	38	720
Trader	440	180	100	720
Public/civil servant	634	80	6	720
Artisan	496	170	54	720
Student	569	77	74	720
Unemployed	453	189	78	720
Total	3120	850	350	4320

$\chi^2 = 236.030$; df = 10; $p < .05$

Table A5: Cross-classification of Abortion Cases by Quarter

Quarters	Abortion Cases 0 – 5	6 – 10	11 and above	Total
Quarter 1	754	238	88	1080
Quarter 2	806	192	82	1080
Quarter 3	833	185	62	1080
Quarter 4	727	235	118	1080
Total	3120	850	350	4320

$\chi^2 = 38.326$; df = 6; $p < .05$

Table A6: Cross-classification of Abortion Cases by Year

Years	Abortion Cases 0 – 5	6 – 10	11 and above	Total
2008	632	177	55	864
2009	630	162	72	864
2010	593	195	76	864
2011	587	177	100	864
2012	678	139	47	864
Total	3120	850	350	4320

$\chi^2 = 43.050$; df = 8; $p < .05$

Model 1: Abortions ~ Time (months)

Table A8: Parameters Estimates from Model 1 (Months)

Parameter	Estimate	df	S.E	Wald χ^2	p
intercept	1.629	1	0.0251	4196.646	.000
January	-0.150	1	0.0377	15.865	.000
February	-0.056	1	0.0356	2.475	.116
March	-0.135	1	0.0367	13.551	.000
April	-0.071	1	0.0362	3.818	.051
May	-0.107	1	0.0366	8.520	.004
June	-0.408	1	0.0398	105.254	.000
July	-0.245	1	0.0380	41.783	.000
August	-0.329	1	0.0389	71.530	.000
September	-0.304	1	0.0386	62.214	.000
October	-0.039	1	0.0359	1.160	.281
November	-0.014	1	0.0357	0.154	.695

AIC = 2537; BIC = 2545; G^2 = 260.913; df = 11; $p < .05$

Model 2: Abortion ~ Age groups of Abortion cases

Table A9: Parameters Estimates from Model 2 (age groups)

Parameter	Estimate	df	S.E	Wald χ^2	p
intercept	1.127	1	0.0212	2821.785	.000
age 10-14	-1.437	1	0.1087	174.811	.000
age 15-19	0.189	1	0.0287	43.199	.000
age 20-24	0.697	1	0.0260	720.489	.000
age 25-29	0.626	1	0.0263	567.172	.000
age 30-34	0.225	1	0.0285	62.407	.000

AIC = 2387; BIC = 2390; G^2 = 1755.319; df = 5; $p < .05$

Model 3: Abortion ~ Occupation

Table A10: Parameters Estimates from Model 3 (occupation)

Parameter	Estimate	df	S.E	Wald χ^2	p
intercept	1.739	1	.0171	10365.845	.000
Farmers	-.255	1	.0259	96.792	.000
Traders	.069	1	.0238	8.317	.004
Public/ civil servants	-.878	1	.0316	774.288	.000
Artisans	-.145	1	.0251	33.297	.000
Students	-.582	1	.0270	464.192	.000

AIC = 2408; BIC = 2412; G^2 = 1543.910; df = 5; $p < .05$

Model 4: Abortion ~ Marital status

Table A11: Parameters Estimates from Model 4 (marital status)

Parameter	Estimate	df	S.E	Wald χ^2	p
intercept	1.739	1	.0171	10365.845	.000
Farmers	-.255	1	.0259	96.792	.000
Traders	.069	1	.0238	8.317	.004
Public/ civil servants	-.878	1	.0316	774.288	.000
Artisans	-.145	1	.0251	33.297	.000
Students	-.582	1	.0270	464.192	.000

AIC = 2408; BIC = 2412; G^2 = 1543.910; df = 5; $p < .05$

Table A12: Selection Criteria for All Models

Model	df	*p*	AIC	BIC	G^2
1	11	.000	2537	2545	260.913
2	5	.000	2387	2390	1755.319
3	3	.000	2408	2412	1543.910
4	1	.000	2557	2559	39.080
5	**10**	**.000**	**2253**	**2260**	**3100.098**
6	6	.000	2383	2387	1795.117
7	16	.000	2363	2374	2010.628
8	6	.000	2404	2409	1583.230
9	16	.000	2384	2394	1805.573
10	12	.000	2534	2542	299.912
11	**11**	**.000**	**2249**	**2257**	**3139.900**
12	**21**	**.000**	**2230**	**2246**	**3355.500**
13	17	.000	2360	2371	2050.420
14	17	.000	2380	2391	1844.873
15	**22**	**.000**	**2226**	**2240**	**3395.315**