MODELING THE OCCURRENCE AND INCIDENCE OF MALARIA CASES: A CASE STUDY AT OBUASI GOVERNMENT HOSPITAL

By

Boateng Alexander

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

MASTER OF PHILOSOPHY

COLLEGE OF SCIENCE

JUNE, 2012.

DECLARATION

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

Boateng Alexander .....

PG5070210 Signature Date

Certified by:

Nana Kenna Frempong ....

Supervisor's Name Signature Date

Certified by:

Mr. F.K. Darkwah ....

Head of Dept. Name Signature Date

ACKNOWLEDGMENT

My first thanks go to the Almighty God, without whose provisions and guidance, my participation in this program of study would have been futile. I would like to express my heartfelt gratitude to my supervisor, Nana Kenna Fempong, who read, criticized and provided necessary support and encouragement to accomplish this research. To Messrs Prof. M Louis, Dr.Kofi Z. Batse and Dr. N. Blay, Samuel Oduro, Kofi Agygarko Ababio, Eric Nimako Aidoo, Owusu Foster, Abdul Wahab and Bashiru Imoro Ibn Saeed, and my Head of Department, Mr. F. K. Darkwah, I extend special thanks for your immense contribution and support, particularly, in this research. I count myself blessed to have you all. To friends and classmates that contributed in diverse ways to my success in this program, I say a big thank you. Finally, my thanks go to my family, for the support, prayer, contributions and bearing with me throughout this program of study.

DEDICATION

I humbly dedicate this piece of work to the Almighty God, my mother Mad. Felicity Donkor and my supervisor, Nana Kenna Frempong.

ABSTRACT

Malaria has always been a major a major health problem and therefore timely and accurate information about its occurrence and incidence cannot be underestimated. The main objectives of this research is to model the occurrence of malaria cases given the age, gender and time in quarters ; to model the incidence of severe malaria cases given age , gender and time in years and lastly to validate the two models using negative binomial regression model. Poisson and negative binomial regression models were used in fitting the data obtained from Obuasi Government Hospital data based dated 2007 to 2010. Based on the results, the negative binomial regression model fitted the data better than the Poisson regression model. Both models indicated that malaria is independent of gender. With respect to time, more cases were recorded in quarters4 (October-December) in the first model and the incidence of severe malaria cases also increased with time in the second model. The prevalence of malaria and severe malaria cases were found to be prevalent among children with less than 1 year old, and those under 5 and 70+years old. More cases were recorded for those found 20-34 year groups with reference to occurrence of malaria and incidence of severe malaria cases. Consequently, we draw a conclusion that despite the various interventions such as the Internal Residual Mass Spraying (I R M S) exercise by the Malaria Control Programme of AngloGold Ashanti introduced in 2006 and other social programmes aimed at reducing the menace of malaria, it's still remains high particularly among children under 5 years and those found between 20-34 age groups.

KEY WORDS: Ghana, Malaria, Poisson, Negative binomial

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background of study

Malaria has been a long life-threatening parasitic disease transmitted by female anopheles mosquitoes. This has contributed to child morbidity in the world. It threatens 2.4 billion people, or about 40% of the world's population living in the world's poorest countries and more than one million deaths are attributable to the disease annually (WHO, 2000). Most of these deaths occur in children in high-transmission areas and malaria accounts for approximately one in five of all childhood deaths in Africa. However, the true burden of malaria is difficult to estimate as many people are treated at home and no proper post-mortem diagnosis is made in the case of death. As a result, many malaria cases go unreported. In areas of stable endemic malaria transmission in sub-Saharan western Africa, it has been estimated that in the year 1995 about 1 million deaths were directly attributable to malaria infection (Snow et al., 1999). Of these deaths, three quarters were recorded among children below the age of 5 years.

Accordingly, a World Bank report of 1993 noted that malaria accounts for an estimated 35 million disability-adjusted life years (DALYs) per year lost in Africa due to ill-health and premature death (World Bank, 1993). The discovery of an interactive effect between HIV infection and malaria morbidity (Whitworth et al., 2000; Chandramohan & Greenwood, 1998; Verhoef et al., 1999) exacerbates the potential for devastating health consequences in populations with large numbers of individuals who are co-infected. In resource-poor countries in Africa, malaria prevention and treatment consume a large

proportion of the health budget, and because it poses a threat to the in these countries. Malaria therefore not only affects the health status of Africa's population, but also has far-reaching economic consequences inhibiting economic development (Wernsdorfer & Wernsdorfer, 1988). The impact of malaria on the population and its significance on development in the region was recognized by the Abuja, Nigeria Summit in April 2000 as the first African summit of Heads of government on malaria control. The communiqué from the meeting calls, among other things, for more research on trends in incidence and prevalence of malaria, epidemic outbreaks and clinical epidemiology (Sachs, 2000). A better understanding of the distribution of malaria has been identified as an important tool in its control (Snow et al., 1996). More accurate maps make it possible for interventions to be mounted that are appropriate to the disease profile, which characterizes particular levels of endemicity. However, for clinical trials and evaluation, new approaches should be located correctly, and for planners of irrigation and other development schemes to take cognizance of the potential effects of these schemes on malaria transmission intensities.

In Ghana, the statistics as supplied by the National Malaria Control Programme (NMCP) are no less staggering. Established as the leading cause of illness, it causes about 8,200 cases daily and 3,000,000 illnesses every year with over 3000 deaths in 2010. As high mortality as this is, the NMCP is quick to point out that this represents a steady drop from the 40, 000 deaths reported ten years ago. The most vulnerable groups remain children under five years of age, pregnant women and non-immune.

Malaria is caused by the parasites of genus Plasmodium. The four species of Plasmodium are Plasmodium falciparium, Plasmodium malariae, Plasmodium ovale, and

Plasmodium vivax. In Africa, the predominant species of the disease-causing parasite is Plasmodium falciparum. Infection of the human host occurs when a person is bitten by a female Anopheles mosquito that has previously become infected. The parasite, called sporozoite at this stage of its cycle, enters the human body via the saliva of the mosquito that is injected into the blood. The parasites multiply in the liver and re-invade the blood via red blood cells as merozoites. These develop into a stage known as the trophozoite, which is the one visible in blood films, and subsequently divide by the process of schizogony to produce further merozoites, which invade non-infected blood cells. Some of the merozoites develop into new trophozoites, while others develop into malice micro or female macrogametocytes. Uninfected Anopheles mosquitoes become infected if they feed on a person with mature gametocytes in their peripheral blood. Within the mosquito, the micro gametocytes exflagellate into gametes before fertilizing the macrogametocytes, thereby forming zygotes. The zygote changes into an ookinete and then into an oocyst, which is found in the mid-gut wall of the mosquito. Large numbers of sporozoites are formed within the oocyst. The rate of development of sporozoites in the oocyst is temperature dependent. The sporozoites leave the oocyst to invade the mosquito's salivary glands, from where they can infect another human host when the mosquito takes a blood meal. The incubation period of the parasite in the vector takes 13 days to complete at 240 C. for P.falciparum. The vector will only become infective if it survives this sporogonic cycle (Gilles & Warrell, 1993). Malaria as a disease is therefore closely bound to conditions which favour the survival of the anopheles mosquito in the form of habitat and breeding sites and which favour the life cycle of the parasite in term of suitable temperatures. In the absence of any human intervention, these conditions are

predominantly determined by climatic and environmental factors. Clinically, malaria manifests itself in its mild form as an illness associated with other non-specific symptoms (Bruce-Chwatt, 1980). The first clinical sign will only appear after the incubation period, which varies between nine and fourteen days for falciparum malaria. Clinical diagnosis is usually confirmed by a blood test, involving microscopic evidence of parasites in the blood, or by a rapid diagnostic kit (Craig & Sharp, 1997).

1.2 Study area profile

The Obuasi Municipality is one of the 27 districts of the Ashanti Region and was created as part of the government's effort to further decentralized governance. It was carved out of the erstwhile Adansi West District Assembly on the strength of executive instruments (E. I.) 15 of December, 2003 and Legislative Instrument L.I 1795 of 17th March, 2007.

The Municipality is located at the southern part of Ashanti Region between latitude 5.35N and 5.65N and longitude 6.35N and 6.90N. It covers a land area of 162.4sqkm. There are 53 communities in the Municipality which share 30 electoral areas. It is bounded to the east by Adansi South, west by Amansie Central and to the north by Adansi North, to the south by Upper Denkyira District in the Eastern Region. It has Obuasi as its Administrative Capital where the famous and rich Obuasi Gold Mines, now Anglo Gold Ashansti is located.

The Municipality has a rather undulating topography and the climate is of the semi-equatorial type with a double rainfall regime. Mean annual rainfall ranges between 125mm and 175mm. Mean average annual temperature is 25.5OC and relative humidity is 75% - 80% in the wet season. The population of the Municipality is estimated at 205,000 using the 2000 Housing and Population Census as a base and applying a 4%

annual growth rate. The vegetation is predominantly a degraded and semi-deciduous forest. The forest consists of limited species of hard wood which are harvested as lumber. The Municipality has nice scenery due to the hilly nature of the environment. Obuasi Government hospital is located in one the beautiful suburbs in the municipality called Mensah Krom which about 5 minute drive from commercial area of the town and it has several clinics under its supervision.

1.3 Problem statement

Malaria has always been the most significant public health threat to the community. To deal with the spread of malaria, AngloGold Ashanti and Obuasi Municipal Assembly undertook to implement an Integrated Malaria Control Programme, focussing on Indoor Residual Spraying (IRS) in the Obuasi municipality and its surrounding villages.

The programme covered the entire Municipality. The total number of dwellings in the intervention area was 35000.

Malaria, still a major health concern in the Obuasi municipality recorded in 2005, an average of 12,000 cases monthly. Forty-eight percent (48%) of all Out Patient Attendants were due to malaria and the disease headed the top ten killers, being responsible for 22% of all deaths (GHS-District Annual Report, 2005).

There are consistent efforts to reduce malaria episodes which include chemical spraying, use of treated mosquito bed nets, clearing bushes, cleaning drains and subsidised treatments and yet prevalence rates and malaria incidence remain high. It is probable that the efforts to reduce malaria do not specifically take into account the risks factors likely to aggravate malaria disease. The high incidence of malaria cases among the age-structured population is unknown to the district, the season which recorded the highest

incidence despite the Integrated Malaria Control programme is also not known in the Obuasi municipality.

1.4 Objectives of the Thesis

The objectives of the thesis are as follows;

1. To model the occurrence of malaria cases given the age, gender and time in quarters.
2. To model the incidence of severe malaria cases given the age, gender and time in years.
3. To validate the two models using Negative Binomial model.

1.5 Methodology

Since the data are count of events (non-negative integers with no upper bound), with the events being independent and average rate not changing over the period of interest, Poisson regression model could be a necessary 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.

A characteristic of the Poisson distribution is that its mean is equal to its variance. In certain circumstances, it will be found that the observed variance is greater than the mean; this is known as over-dispersion indicates that the model is not appropriate. A common reason is the omission of relevant explanatory variables. Under some

circumstances, the problem of over-dispersion can be solved by using a negative binomial distribution instead.

The research will be restricted to the use of quantitative data. A routine time data will be taken from Obuasi Government Hospital for the analysis and modeling. The data is obtained from the Out Patients Department database and dates back from January 2007 to December 2010. Data analysis would be done using the R software.

1.6 Significance of the Thesis

Health is the level of functional and or metabolic efficiency of a living being. In humans, it is the general condition of a person in the mind, body and spirit, usually meaning to be free from illness, injury or pain (as in "good health" or "healthy"). The World Health Organisation (WHO) defined health in its broader sense in 1946 as "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity. Although this definition has been subject to controversy, in particular as having a lack of operational value and the problem created by use of the word "complete", it remains the most enduring. Classification systems such as the WHO Family of International Classifications, including the International Classification of Functioning, Disability and Health (ICF) and the International Classification of Diseases (ICD), are commonly used to define and measure the components of health.

The maintenance and promotion of health is achieved through different combination of physical, mental, and social well-being, together sometimes referred to as the "health triangle". The WHO's 1986 Ottawa Charter for Health Promotion furthered that health is not just a state, but also "a resource for everyday life, not the objective of living. Health is a positive concept emphasizing social and personal resources, as well as physical

capacities." Systematic activities to prevent or cure health problems and promote good health in humans are delivered by health care providers. Applications with regard to animal health are covered by the veterinary sciences. The term "healthy" is also widely used in the context of many types of non-living organizations and their impacts for the benefit of humans, such as in the sense of healthy communities' healthy cities or healthy environment. In addition to health care interventions and a person's surroundings, a number of other factors are known to influence the health status of individuals, including their background, lifestyle, and economic and social conditions; these are referred to as "determinants of health". Malaria is an integral part of the health and health practitioners are constantly making efforts in the area of research to reduce it menace hence the need for this research.

The result of the research will go a long way to help the municipal health directorate to make informed decisions and the various hospitals and the offices of the malaria control programme to understand and appreciate the underlying risk factors of malaria and modeling of malaria cases. This will enhance strategy formulation to assist in curtailing the prevalence and incidence of malaria.

1.7 Scope and Limitation of the Thesis

The thesis is restricted to the objectives of the research. Research work is often characterized by some constraints. Some of these setbacks include resource inadequacy since the project is solely self-sponsored, time constraints and the unavailability of relevant materials such as journals on the study.

1.8 Organisation of the Thesis

The thesis contains five chapters. Chapter 1 discuses the background of study, study area profile, problem statement, objectives, methodology, significance of the study, scope of study and the limitation of the study. Chapter 2 explores some previous research on malaria and other diseases and some applications. Count models such as a Poisson and Negative binomial regression models with applications in accidents, health, traffic and road crashes and criminology. Chapter 3 discusses the methodology and its process. Data Analysis and Results are explored in Chapter 4. Chapter 5 deals with the conclusions and recommendations.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The purpose of the review of related literature in a study is to discover facts, findings, concerning the area of study and how they can propel the researcher to explore the unknown Leedy (1989). Again this review also seeks to investigate the literature on the subject of modeling malaria incidence using a hospital based data. Kumekpor (2002), noted that the non-availability of relevant documentary sources poses serious challenge to would-be investigators and rightly so in the West Africa situation. The few that exist are normally poorly equipped and it difficult to get up-to-date information on many topics of interest to the investigator. This review will actually elaborate on the previous research on malaria and previous applications of count models such as Poisson regression model and negative binomial model.

2.2 Previous Research on Malaria

According to Claus et al. (2007) understanding local variability in malaria transmission risk is critically important when designing intervention or vaccine trials. Using a combination of field data, satellite image analysis, and GIS modeling, we developed a high-resolution map of malaria entomological inoculation rates (EIR) in The Gambia, West Africa. The analyses are based on the variation in exposure to malaria parasites experienced in 48 villages in 1996 and 21 villages in 1997. The entomological inoculation rate (EIR) varied from 0 to 166 infective bites per person per rainy season. Detailed field surveys identified the major *Anopheles gambiae* s.l. breeding habitats. These habitats were mapped by classification of a LANDSAT TM satellite image with

an overall accuracy of 85%. Village EIRs decreased as a power function based on the breeding areas size and proximity. We use this relationship and the breeding habitats to map the variation in EIR over the entire 2500-km^2 study area.

Greenwood et al. (1986) measured mortality and morbidity from malaria were measured among 3000 children under the age of 7 years in a rural area of The Gambia, West Africa. Using a post-mortem questionnaire technique, malaria was identified as the probable cause of 4% of infant deaths and of 25% of deaths in children aged 1 to 4 years. The malaria mortality rate was 6·3 per 1000 per year in infants and 10·7 per 1000 per year in children aged 1 to 4 years. Morbidity surveys suggested that children under the age of 7 years experienced about one clinical episode of malaria per year. Calculation of attributable fractions showed that malaria may be responsible for about 40% of episodes of fever in children. Although the overall level of parasitaemia showed little seasonal variation, the clinical impact of malaria was highly seasonal; all malaria deaths and a high proportion of febrile episodes were recorded during a limited period at the end of the rainy season.

Greenwood and Pickening, (2004) recognized malaria as an important cause of death among early European visitors to The Gambia, but the infection was first studied systematically in the local population only in the 1950s. Studies undertaken in the village of Keneba at that time showed that nearly all children under the age of 5 years had parasitaemia throughout the year. More recent surveys in rural areas of The Gambia have shown much lower levels of parasitaemia, probably as a result of a decline in rainfall in The Gambia during the past 30 years and because of an increase in the availability of anti-malarial drugs. Nevertheless, community surveys and reviews of

hospital statistics show that malaria is still one of the most important causes of death among Gambian children; about 1 in 25 rural Gambian children die from malaria before reaching the age of 5 years. Until recently, malaria control in The Gambia relied upon prompt treatment of clinical attacks, first with quinine and more recently with chloroquine, and upon some limited vector control in the capital, Banjul. However, during the past few years, it has been shown that mortality in rural children can be reduced substantially by means of chemoprophylaxis given by village health workers. Bed nets (mosquito nets) are used widely in The Gambia and epidemiological surveys have shown an association between the use of bed nets and protection against malaria. This observation led to a series of small scale intervention trials. These showed that conventional bed nets were not very effective at protecting against clinical attacks of malaria in children but that their protective effect was enhanced substantially when they were impregnated with the insecticide permethrin. The success of these pilot trials led to a much larger study of impregnated bed nets which had the objectives of determining whether the use of impregnated bed nets could reduce mortality in Gambian children and whether impregnation of bed nets could be accomplished successfully on a large scale through the national primary health care programme. Malaria has been a major cause of poverty and low productivity accounting for about 32.5 percent of all OPD attendances and 48.8 percent of under five years admissions in the country. (National Malaria Control Programme annual report, 2009). The attempt to control malaria in Ghana began in the 1950s. It was aimed at reducing the malaria disease burden till it's no longer of public health significance. It was also recognized that malaria cannot be controlled by the health sector alone therefore multiple strategies were being pursued with other health

related sectors. In view of this, interventions were put in place to help in the control of the deadly disease. Some of the interventions applied at the time included residual insecticide application against adult mosquitoes, mass chemoprophylaxis with Pyrimethamine medicated salt and improvement of drainage system. But malaria continued to be the leading cause of morbidity (illness) in the country.

Ghana then committed itself to the Roll Back Malaria (RBM) initiative in 1999 and developed a strategic framework to guide its implementation. Overall, the Ghana RBM emphasizes the strengthening of health services through multi and inter-sectoral partnerships and making treatment and prevention strategies more widely available. The goal was to reduce malaria specific morbidity and mortality by 50% by the year 2010 (National Malaria Control Program, Ghana Health Service, 2011).

McCombie (1999) review of literature on treatment seeking for malaria was undertaken to identify patterns of care seeking, and to assess what is known about the adequacy of the treatments used. There is considerable variation in treatment seeking patterns, with use of the official sector ranging from 10–99% and self-purchase of drugs ranging from 4–87%. The majority of malaria cases receive some type of treatment, and multiple treatments are common. The response to most episodes begins with self-treatment, and close to half of cases rely exclusively on self-treatment, usually with antimalarials. A little more than half use the official health sector or village health workers at some point, with delays averaging three or more days. Exclusive reliance on traditional methods is extremely rare, although traditional remedies are often combined with modern medicines. Although use of antimalarials is widespread, under dosing is extremely common.

Further research is needed to answer the question of what proportion of true malaria cases get appropriate treatment with effective antimalarial drugs, and to identify the best strategies to improve the situation. Interventions for the private and public sector need to be developed and evaluated. More information is needed on the specific drugs used, considering resistance patterns in a particular area. In order to guide future policy development, future studies should define the nature of self-treatment, record multiple treatments and attempt to identify the proportions of all cases that begin treatment with antimalarials at standardized time intervals. Hypothetical questions were found to be of limited usefulness in estimating rates of actual treatments. Whenever possible, studies should focus on actual episodes of illness and consider supplementing retrospective surveys with prospective diary-type methods. In addition, it is important to determine the specificity of local illness terms in identifying true malaria cases and the extent to which local perceptions of severity are consistent with clinical criteria for severity and symptoms of complicated malaria.

Olaleye et al. (1997) researched into diagnosis of malaria in children is difficult without laboratory support because the symptoms and signs of malaria overlap with those of other febrile illnesses such as pneumonia. Nevertheless, in many parts of Africa diagnosis of malaria must be made without laboratory investigation. Therefore, a scoring system has been developed to assist peripheral health care workers in making this diagnosis. Four hundred and seven Gambian children aged 6 months to 9 years who presented to a rural clinic with fever or a recent history of fever was investigated. A diagnosis of malaria was made in 159 children who had a fever of 38 °C or more and malaria parasitaemia of 5000 parasites/μL or more. Symptoms and signs in children with

malaria were compared with those in children with other febrile illnesses to identify features which predicted malaria. Symptoms and signs were incorporated into various logistic regression models to test which were best independent predictors of malaria and these regression models were used to construct simple scoring systems which predicted malaria. A nine terms model predicted clinical malaria with a sensitivity of 89% and a specificity of 61%, values comparable to those obtained by an experienced paediatrician without laboratory support. The ability of peripheral health care workers to diagnose malaria using this approach is now being investigated in a prospective study.

Malaria in Thailand is endemic in forest regions and many cases occur along the national borders, particularly on the border with Myanmar to the east (Wattanavadee Scriwattanapongse, 2009). Although malaria cases and deaths had fallen substantially since 1999, the disease remained a considerable public health problem. Gomez-Elipe et al. (2007) developed a model to predict malaria incidence in an area of unstable transmission in Burundi by studying the association between environmental variables and disease dynamics. The model used time series of quarterly notifications of Malaria cases from local health facilities, rain and temperature records, and the normalized difference vegetation index. An autoregressive integrated moving average methodology was employed to obtain a model showing the relation between quarterly notifications of malaria cases and the environmental variables.

2.3 Count Models

2.3.1 Poisson distribution

The Poisson distribution is often used to model information on counts of various kinds particular 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 examples 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 characterize 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.

2.3.2 Negative Binomial Distribution

The weakness of the Poisson distribution in accommodating heavy tails was recognized 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 rth 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) 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; see also Seal (1982). There are alternative choices to the gamma for the mixing distribution $g(\lambda)$.

Two which have appeared in the actuarial literature are the generalized inverse Gaussian and inverse Gaussian distributions. The generalized 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.

2.4 Poisson and Negative Binomial Regression Models

Box (1979) wrote, "All models are wrong but some are useful." This statement is unquestionably true, but it raises the question: Useful for what? There are two ways in which a model can be useful, it can improve our understanding of the system generating the data or it can make accurate predictions of future observations. For example, linear models for designed factorial experiments are useful because the terms they contain may be interpreted as main and interaction effects. On the other hand, accurate weather prediction models are useful even if they are hard to interpret.

Strien et al (2000), loglinear Poisson regression method has been developed to analyze time series of count data. The method produces yearly indices and trend estimates. It is also capable of testing the effects of covariants on the changes so that the impact of human activities on changes can be investigated. The method can also deal with several difficulties inherent to monitoring data, especially missing values, over- and under sampling of particular strata, serial correlation and deviations from Poisson distribution.

Parodi and Bottarelli, (2006) described Poisson regression model as a technique used to describe count data which is 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 subject. This review provides a description of the Poisson regression in the framework of the prospective cohort study, which represents the conceptual ground of most epidemiological investigations. Furthermore, some strategies of modeling are illustrated, which allow to obtain estimates

of relative risk between exposed and unexposed individuals, adjusted for the effect of extraneous variables (confounding), or to assess the presence of an interaction between an exposure variable and another factor ("effect modifier"). Finally, a short description of some veterinary epidemiology studies, which have applied the Poisson regression model, is provided.

Applying linear regression to count data leads to inconsistent standard errors and may produce negative predictions for the dependent variable. Even with a logged dependent variable, the least squared estimates have these problems and are biased and inconsistent King (1989). Therefore count dependent variables require different modeling. The most common assumption of count data distribution is the Poisson distribution which restricts the data distribution to be equal-dispersion (the conditional variance equals the conditional mean). This stringent restriction cannot handle many empirical applications. Other modeling distributions have been developed. Mixed-Poisson distributions and negative binomial distributions have been widely used in situations where counts display over-dispersion (conditional variance exceeds the conditional mean). For under-dispersion (conditional variance is less than conditional mean) there are fewer modeling options. Since there is no model that handles only the underdispersed data, with underdispersed data we need to consider models that are flexible enough to cover both over- and under-dispersed data. Models that provide this flexibility include: the generalized event count (GEC (k)) model (Winkelmann (1991) and Zimmermann (1995), double Poisson, Efron (1986), Poisson polynomial expansion, hurdle models (Mullahy 1986), and the generalized Poisson models (Famoye 1993, Famoye and Singh. (2003).

Poisson regression is routinely used for analysis of epidemiological data from studies of large occupational cohorts. It is typically implemented as a grouped method of data analysis in which all exposure and covariate information is categorized and person-time and events are tabulated. Liu Sela (2008) describes an alternative approach to Poisson regression analysis using single units of person-time without grouping. Data for simulated and empirical cohorts were analyzed by Poisson regression. In analyses of simulated data, effect estimates derived via Poisson regression without grouping were compared to those obtained under proportional hazards regression. Analyses of empirical data for a cohort of 138 900 electrical workers were used to illustrate how the ungrouped approach may be applied in analyses of actual occupational cohorts .It was realized that using simulated data, Poisson regression analyses of ungrouped person-time data yield results equivalent to those obtained via proportional hazards regression: the results of both methods gave unbiased estimates of the ‘‘true’’ association specified for the simulation. Analyses of empirical data confirm that grouped and ungrouped analyses provide identical results when the same models are specified. However, bias may arise when exposure-response trends are estimated via Poisson regression analyses in which exposure scores, such as category means or midpoints, are assigned to grouped data. It was concluded that Poisson regression analysis of ungrouped person-time data is a useful tool that can avoid bias associated with categorizing exposure data and assigning exposure scores, and facilitate direct assessment of the consequences of exposure categorization and score assignment on regression results

Poisson regression has been widely used to model count data. However, it is often criticized for its restrictive assumption of equi-dispersion, meaning equality between the

variance and the mean. In real-life applications, count data often exhibits over-dispersion and excess zeroes. While Negative binomial regression is able to model count data with over-dispersion, both Hurdle (Mullahy, 1986) and Zero-inflated (Lambert, 1992) regressions address the issue of excess zeroes in their own rights. Different modeling strategies for count data and various statistical tests for model evaluation are illustrated through an example of healthcare utilization.

Researchers do not always evaluate the potential for bias in this method when the data are over-dispersed. This study used simulated data to evaluate sources of over-dispersion in public health surveillance data and compare alternative statistical models for analyzing such data. If count data are over-dispersed, Poisson regression will not correctly estimate the variance. A model called negative binomial 2 (NB2) can correct for over-dispersion, and may be preferred for analysis of count data. This paper compared the performance of Poisson and NB2 regression with simulated over-dispersed injury surveillance data. Methods: Monte Carlo simulation was used to assess the utility of the NB2 regression model as an alternative to Poisson regression for data which had several different sources of over-dispersion. Simulated injury surveillance datasets were created in which an important predictor variable was omitted, as well as with an incorrect offset (denominator). The simulations evaluated the ability of Poisson regression and NB2 to correctly estimate the true determinants of injury and their confidence intervals. Results: The NB2 model was effective in reducing over-dispersion, but it could not reduce bias in point estimates which resulted from omitting a covariate which was a confounder, nor could it reduce bias from using an incorrect offset. One advantage of NB2 over Poisson for over-dispersed data was that the confidence interval

for a covariate was considerably wider with the former, providing an indication that the Poisson model did not fit well. When over-dispersion is detected in a Poisson regression model, the NB2 model should be fit as an alternative. If there is no longer over-dispersion, then the NB2 results may be preferred. However, it is important to remember that NB2 cannot correct for bias from omitted covariates or from using an incorrect offset Kim and Kriebel (2009).

Count data regression models are used when the dependent variable takes on non-negative integer values. Cameron and Trivedi (1996) and Long (1997) provide good overviews of count regression models. Count data models are widely used in empirical studies. Some recent research used count models are as follows. Yang (2007) uses a Poisson distribution count model to explore factors affecting the potential entry into an industry. Hellström and Nordström (2008) using the count data modeling to analyze household's choice of total number of nights to spend on monthly recreational trips. Nelson and Young (2008) study the effects of various factors on alcohol advertising in magazines using the Poisson and negative binomial count regressions. Czado et al. (2007) proposed an extension of zero-inflated generalized Poisson regression models for count data. Guikema and Goffelt (2008) present a count model based on Conway-Maxwell Poisson (COM) distribution that is useful for both under-dispersed and over-dispersed count data.

Applying linear regression to count data leads to inconsistent standard errors and may produce negative predictions for the dependent variable. Even with a logged dependent variable, the least squared estimates have these problems and are biased and inconsistent King (1989). Therefore count dependent variables require different modeling. The most

common assumption of count data distribution is the Poisson distribution which restricts the data distribution to be equal-dispersion (the conditional variance equals the conditional mean). This stringent restriction cannot handle many empirical applications. Other modeling distributions have been developed. Mixed-Poisson distributions and negative binomial distributions have been widely used in situations where counts display over-dispersion (conditional variance exceeds the conditional mean). For under-dispersion (conditional variance is less than conditional mean) there are fewer modeling options. Since there is no model that handles only the underdispersed data, with underdispersed data we need to consider models that are flexible enough to cover both over- and under-dispersed data. Models that provide this flexibility include: the generalized event count (GEC (k)) model (Winkelmann and Zimmermann,1991 and 1995), double Poisson (Efron 1986), Poisson polynomial expansion, hurdle models (Mullahy 1986), and the generalized Poisson models (Famoye 1993, Famoye and Singh 2003) .

While Poisson regression is a popular tool for modeling count data, it is limited by its associated model assumptions. One assumption is that the response variable follows a Poisson distribution. However, over- or under-dispersion are common in practice and are not accommodated by Poisson regression. In addition, the dispersion is assumed fixed across observations, whereas in practice dispersion may vary across groups or according to some other factor. Recently, Sellers and Shmueli (2008) introduced the Conway-Maxwell-Poisson (CMP) regression, based on the CMP distribution. CMP regression generalizes both Poisson and logistic regression models and allows for over- or under-dispersed count data. The model structure introduced, however, assumes a fixed

dispersion level across all observations. In this paper, we extend the CMP regression model to account for observation-level dispersion. We discuss model estimation, inference, diagnostics, and interpretation, and present a variable selection technique. We then compare our model to several alternatives and illustrate its advantages and usefulness using datasets with varying types and levels of dispersion.

In particular, Poisson regression implicitly uses a log transformation which adjusts for the skewness and prevents the model from producing negative predicted values. Poisson regression also models the variance as a function of the mean (Professor Mean, 2007).

care hospitals. Their study indicated that race, income and education are significant factors in differential hospital utilization rates. They also found that admission rates for medical reasons declined with increasing community income levels and were elevated in blacks. Kudur and Demlo (1985) found average income levels in areas of high admission rates were significantly lower than those areas of low admission. Wilson and Tedeschyi (1984) found income levels positively associated with surgical discharges rates and a positive association between the percent of Medicaid population and medical discharge rates. Wennberg and Freeman (1987) found in a population-based study that the relationship between hospitalization rates for avoidable hospital conditions (AHC) cases and median household income, revealed consistent correlation between low income and high rate of hospitalization. Also, Codman Research Group (1991), a group of investigators in California found a negative correlation between income and the rate of hospitalization for these AHC and suggested that a reduction in these AHC's offers a considerable cost savings to the community. DeShazo (1997) admits that the random nature of the discharge count data of his research suggests a fit of the pure Poisson

model because the data indicates the average number of discharges per person per time interval. It indicated that the goodness-of-fit of pure Poisson model for the count data was poor and observed that the count data indicated extra-Poisson variation; and thus, decided to fit a log-linear model. DeShazo argued that the log-linear model was observed to be appropriate for this purpose since it can be extended to include measured community characteristics in a regression model, while still yielding an estimate of the amount of systematic variability beyond that predicted by the factors included in the model. The results of DeShazo research support the hospital discharge findings of Wilson and Tedeschyi (1984).

Again, in another research it was revealed that Poisson regression model is another natural choice for fitting a log-linear model, since it estimates incidence rate ratio and since most medical applications of the Poisson distribution arise via the Poisson approximation to the binomial distribution. This approach has been proposed by Traissac et al (1999), McNutt et al (2003), Zou (2004), and Carter et al (2005). The estimating equations were those for a generalized linear model with log link and variance proportional to mean.

Malaria in Thailand is endemic in forest regions and many cases occur along the national borders, particularly on the border with Myanmar to the east (Wattanavadee Scriwattanapongse, 2009). Although malaria cases and deaths had fallen substantially since 1999, the disease remained a considerable public health problem. Gomez-Elipe et al(2007). developed a model to predict malaria incidence in an area of unstable transmission in Burundi by studying the association between environmental variables and disease dynamics. The model used time series of quarterly notifications of Malaria

cases from local health facilities, rain and temperature records, and the normalized difference vegetation index. An autoregressive integrated moving average methodology was employed to obtain a model showing the relation between quarterly notifications of malaria cases and the environmental variables . Devi and Jauhari (2006) investigated the relationship between climate variables and malaria transmission in India. Earlier, Bi et al (2008), also explored the impact of climate on the transmission of malaria in China and suggested that climatic variables should be considered as possible predictors for regions with similar geographic and socio-economic conditions. Hoshen and Morse (2004) described a mathematical-biological model of the parasite dynamics in Africa, comprising the weather-dependent stages, both within vectors and within hosts. Gagnon et a (2001), found a statistically significant relationship between El Ni o and malaria epidemics in South America and thus postulated that global warming will be an important factor in the spatial distribution of infectious disease.

Kleinschmidt et al,(2002) investigated malaria incidence in children under 10 in South Africa by using logistic regression modeling. The model used climatic, population and topographic variables as potential predictors and described a simple two-stage procedure for producing maps of predicted risk, including environmental factors such as land use. Built up areas were found to have the highest incidence rates. Studies like these aim to identify risk factors for the disease, which could provide a basis for health organizations in countries affected by malaria to establish effective prevention programs. However, when resources are limited it is also important to know in which area prevention should be targeted for treatment and control patterns and trends, and this is the focus of the present study. The objective of their study was to identify the spatial patterns and trends

of hospital-diagnosed malaria incidences based on case data aggregated by quarterly periods in 65 districts of the North-western region of Thailand. The provinces in our study comprise Lamphun, Phrae, Nan, and Chiang. They made use data are available in computer files with a record for each case and fields comprising characteristics of the subject and the disease, including dates of sickness and disease diagnosis, the subject's age, gender, address, severity of the illness, and date of death for mortality cases. Counts for malaria cases(incidence) were created for each combination of quarter (24 periods from January, March 1999 to October-December 2004), age group (0-4, 5-14, 15-39 and 40+ years), and district. Incidence rates were computed as the number of cases per 1,000 residents in the district according to the 2,000 Population and Housing Census of Thailand. Since there was little evidence of a gender effect, the data for the two sexes were combined. They considered two alternative statistical models for describing the relation between malaria incidence and the age group, district and period factors, namely a negative binomial and log-normal distribution, respectively. In each case the mean function for the selected distribution was a specified combination of demographic factors. The negative binomial model was found to be an extension of the Poisson model for incidence rates that allows for the over-dispersion that commonly occurs for disease counts in re

CHAPTER 3

METHODOLOGY

3.1 Introduction

A good statistical model is the one that provides a good approximate mathematical representation of the data being modeled with particular emphasis being on structure or patterns in the data (White and Bennetts, 1996). Statistical analysis and modelling of data have become increasingly important in scientific research and study inquiries and the process involves application of appropriate statistical procedure, testing hypotheses, interpreting data results, and coming up with valid conclusions (Clinical Science Research, 2009).

In this chapter, we shall define and give a detailed description of some count models mainly Poisson regression and negative binomial models in the analysis and modeling of malaria occurrence and incidence. We shall begin by describing the data and the coding scheme used. The main emphasis of this chapter is to make available a detailed and inclusive understanding of the topic. A comprehensive elaboration of the analysis plan will be dealt with. Issues related to quantities, estimations and inferences will also be discussed.

3.2. Data Description

This thesis utilized a hospital based data from Obuasi Government Hospital database and it is a routine time data (secondary administrative data). Modelling the incidence of malaria using count models such as Poisson regression and negative regression models will be used with malaria cases being the response variable and gender, age (age group), and time in years and quarters with respect to the stipulated year of study between 2007

to 2010 being the predictors .This study is basically on modeling the occurrence of malaria cases and the incidence of malaria and severe malaria cases using Obuasi Government Hospital as the case study. The study is based on the records of the Out Patients Department (OPD) who were diagnosed with severe malaria (laboratory confirmed). The data was put together at the Obuasi Government Hospital and dates back from January 2007 to December 2010.

3.3 Coding Scheme

The data is categorized into four main groups. The response variable being the malaria cases and these are patients who were diagnosed with severe malaria (laboratory confirmed) and other malaria cases with or without laboratory confirmation given the independent variables gender, age (age group) and time (years and quarters). The age groups provided by the hospital were used, gender (with the males coded as 0 and the females coded as 1) and the time in quarters numbering quarters 1 to quarters 4 between 2007 to 2010 and the years also coded as 1,2,3,and 4 from 2007 to 2010.

3.4 Generalized Linear Models (GLM)

Generalized linear models (GLM) was first introduced by Nelder and Wedderburn (1972, JRSSA). They provided a unified framework to study various regression models, rather than a separate study for each individual regression. Generalized 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 generalized linear model consists of three components:

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

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

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 Generalized 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.

3.5 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 Judgments 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 k occurrences (k being a non-negative integer, k = 0, 1, 2…) 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 λ = 10×4 = 40.

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

$$\sigma_k^2 = E(k^2) - \left[E(k)\right]^2 = \lambda \quad (4)$$

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

$$\sigma_k = \sqrt{\lambda} \quad (5)$$

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 (7) 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 (14) the p.d.f. as

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

The last remaining term in (14) 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

3.6 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

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

Where φ is a dispersion parameter and $a(\varphi)$, $b(\theta)$ and $c(y;\varphi)$ 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

$$\text{var}(Y) = \sigma_i^2 = b''(\theta)a(\varphi) \tag{14}$$

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 modelling include the student's t and uniform distributions.

3.7 Poisson Regression

Poisson regression analysis is a technique which allows for modeling dependent variables that describe count data (Cameron et al, 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 Economy, 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 standardized mortality and incidence ratios in cohort studies and in ecological investigations (Breslow et al. 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 (Trivedi et al. 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 (Travier 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 et al. 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 generalized linear model,

$$\log(\mu_i) = \chi_i^{'}\beta \tag{15}$$

We say that the Poisson regression model is a generalized linear model with Poisson error and a log link so that (from 17)

$$\mu_i = \exp(\chi_i^{'}\beta) \tag{16}$$

This implies that one unit increases in an x_i are associated with a multiplication of μ_i by $\exp(\beta_j)$.

3.7.1 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

$$\log(E(Y/x)) = \log(\exp osure) + \theta^{'}x \tag{18}$$

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 (18), but with a parameter estimate constrained to 1.

3.8 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:

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

From (20) the expected number of events per period is given by

$$E(y_i/x_i) = \mu_i = e^{x_1^{'}\theta} \tag{22}$$

Thus:

$$dE(y_i/x_i) = \beta e^{x_i\beta} = \beta_i \mu_i \tag{23}$$

The major assumption of Poisson model is

$$E(Y_{iI}/x_i) = \mu_i = e^{x_i\beta} = \text{var}(Y_i/x_i) \tag{24}$$

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

3.9 Estimation

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

3.9.1 Maximum Likelihood Estimation

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

$$\Pr(y_i) = \frac{e^{-\lambda} \lambda_i^{y_i}}{y_i!}, i = 0,1,2,... \tag{25}$$

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

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

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

Where $y_i = \mu_i = e^{x_i'\beta}$ (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

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

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

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

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, $\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n}\left(y_i - e^{x_i'\beta}\right)x_i = 0$ so the MLE for β must be obtained numerically. However, as the Hessian is negative definite for all x and β, the MLE $\left(\hat{\beta}\right)$ 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

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

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 $\lambda_i = \exp\left(\beta_1 + \beta_2 x_i\right)$, for $i = 1,2,...,n.$.

(39)

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

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

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.

3.9.2 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,

$$Y_i = x_i\beta + \varepsilon_i \,, \tag{41}$$

Where $i = 1,2,...,n$: 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 characterized 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:

$$\eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \quad (42)$$

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

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \quad (43)$$

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.

3.10 The Link Function

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 utilized. 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 $\mu_i \in 0,1,...$. 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. The table 4.1 below display the various link functions that can be used in GLM frame work.

4.1: Exponential Family and their link functions

Distribution	Link function	Canonical link	Dispersion	Expectation	Variance
	θ	$a(\theta)$	ϕ	$E(y)$	$Var(\mu) = \frac{\text{var}(y)}{\phi}$
$B(n,\pi)$	$\ln\frac{\pi}{1-\pi}$	$n\ln(1+e^{\theta})$	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
$NB(\mu,k)$	$\ln\frac{k\mu}{1+k\mu}$	$-\frac{1}{k}\ln(1-ke^{\theta})$	1	μ	$\mu(1+k\mu)$

3.11 Log-linear Models

Suppose that we have a sample of n observations $y_1, y_2, \cdots y_n$ which can be treated as realizations of independent Poisson random variables, with, $Y_i \sim P(\mu_i)$ 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

$$\mu_i = x_i' \beta \tag{44}$$

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

$$\eta_i = \log(\mu_i) \tag{45}$$

and assume that the transformed mean follows a linear model

$$\eta_i = x_i' \beta \tag{46}$$

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

$$\log(\mu_i) = x_i' \beta \tag{47}$$

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 47 we obtain a multiplicative model for the mean itself:

$$\mu_i = \exp(x_i^{'}\beta) \tag{48}$$

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.

3.12 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 generalized linear models. The Fisher scoring algorithm for Poisson regression models with canonical link would be considered, where it would be modelled as:

$$g(\eta_i) = \log(\mu_i) \tag{49}$$

The derivative of the link is easily seen to be

$$g'(\eta_i) = \frac{1}{\mu_i} \tag{50}$$

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

$$\beta^{new} = \beta + \left\{ E\left(-\frac{\partial \mathbf{L}}{\partial \beta \partial \beta^T} \right) \right\}^{-1} \frac{\partial \mathbf{L}}{\partial \beta} \tag{51}$$

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

$$\beta^{new} = \beta + \left(X^T W X\right)^{-1} X^T W_z \tag{52}$$

where z is the n-vector with ith component

$$z_i = (Y_i - \mu_i) g'(\mu_i) \tag{53}$$

and W is the n × n diagonal matrix with

$$W_i = \left\{ g'(\mu_i)^2 b''(\theta_i) \right\}^{-1} \tag{54}$$

$$W_i = \left(\mu_i \cdot \frac{1}{\mu_i^2} \right)^{-1} \tag{55}$$

And this simplifies to

$$W_i = \mu_i \tag{56}$$

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

3.13 Tests of Hypotheses

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 $\hat{\beta}$ has approximately in large samples a multivariate normal distribution with mean equal to the true parameter value β and variance-covariance matrix, $\text{var}(\hat{\beta}) = X^{'}WX$ where X is the model matrix and W is the diagonal matrix of estimation weights.

3.14 Likelihood Ratio Test

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

$$\lambda_i = \bar{\lambda} \forall \tag{57}$$

Where $$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n} \tag{58}$$

The corresponding log-likelihood is:

$$L_R = n\bar{y} + \log(\bar{y})\left(\sum_{i=1}^{n} y_i\right) - \sum_{i=1}^{n} \log y_i! \tag{59}$$

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 maximized likelihood

$$L = \sum_{i=1}^{n}\left[-e^{x_i'\beta} + y_i x_i'\beta - \log y_i!\right] \tag{60}$$

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

$$LR = 2(L_U - L_R) \tag{61}$$

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

3.15 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 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.

Technical validation:

The deviance function is:

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

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 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- squared 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-squared statistic, which is defined as

The Pearson goodness of fit test statistic is:

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

The deviance residual is (Cook and Weisberg, 1982):

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

The Freeman-Tukey, variance stabilized, residual is (Freeman and Turkey, 1950):

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

The standardized 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)

3.16 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 $E[y_i / x_i] < \text{var}[y_i / x_i]$ then we speak about over- dispersion, and when $E[y_i / x_i] > \text{var}[y_i / x_i]$ 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 $\lambda_i\left(1+\eta^2\lambda_i\right)$ 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

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

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.

3.17 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a way of selecting a model from a set of models. The chosen model is the one that minimizes the Kullback-Leibler distance between the model and the truth. It's 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 (likelihood)) + 2 K (74)

where likelihood is the probability of the data given a model and K is the number of free parameters in the model. AIC scores are often shown as ΔAIC scores, or difference

between the best model (smallest AIC) and each model (so the best model has a ΔAIC of zero).

The second order information criterion, often called AICc, takes into account sample size by, essentially, increasing the relative penalty for model complexity with small data sets. It is defined as:

$$AIC = -2\left(\ln\left(likelihood\right)\right) + 2k * \left(n/\left(n-k-1\right)\right) \quad (75)$$

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's really no harm in always using AICc regardless of sample size. In model selection in comparative methods, sample size often refers to the number of taxa (Butler and King, 2004; O'Meara et al., 2006).

3.18 Software (R)

R: In R (R Development Core Team 2008), GLMs are provided by the model setting functions glm (Chambers and Hastie 1992) in the stats package and glm.nb in the MASS pack (Venables and Ripley 2002) along with associated methods for diagnostics and inference.

3.19 Analysis plan

Modelling the incidence of malaria will be analyzed using R software with malaria cases being the dependent variable and gender, age (age group) and time in quarter and years being the independent variables. Total percentage of variation in the dependent variable

explained by these factors will be analyzed and discussed. All outputs shall be discussed and analyze.

CHAPTER 4

DATA ANALYSIS AND RESULTS

4.1 Introduction

This chapter introduces the analysis of the various models and discussion of findings. Preliminary analysis, summary of results and snapshot of the data will also be presented and discussed. Poisson and Negative Binomial regression models shall be used in the modeling.

4.2 Source of Data

A routine time data was obtained from Obuasi Government Hospital from 2007 to 2010. Laboratory confirmed cases of malaria (severe malaria and simple malaria) and simple malaria cases (non-laboratory confirmed) recorded at the Out Patients Department (OPD) section of the hospital is the response (dependent) variable and the age , gender , time (years and quarters) are independent variables.

4.2The Occurrence of Malaria cases for the various age groups.

Age Category	Cases
<1	2849
1-4	6268
5-9	3768
10-14	3456
15-17	2737
18-19	2450
20-34	8567
35-49	6179
50-59	2759
70+	1686
Total	**40719**

Source: Obuasi Government Hospital

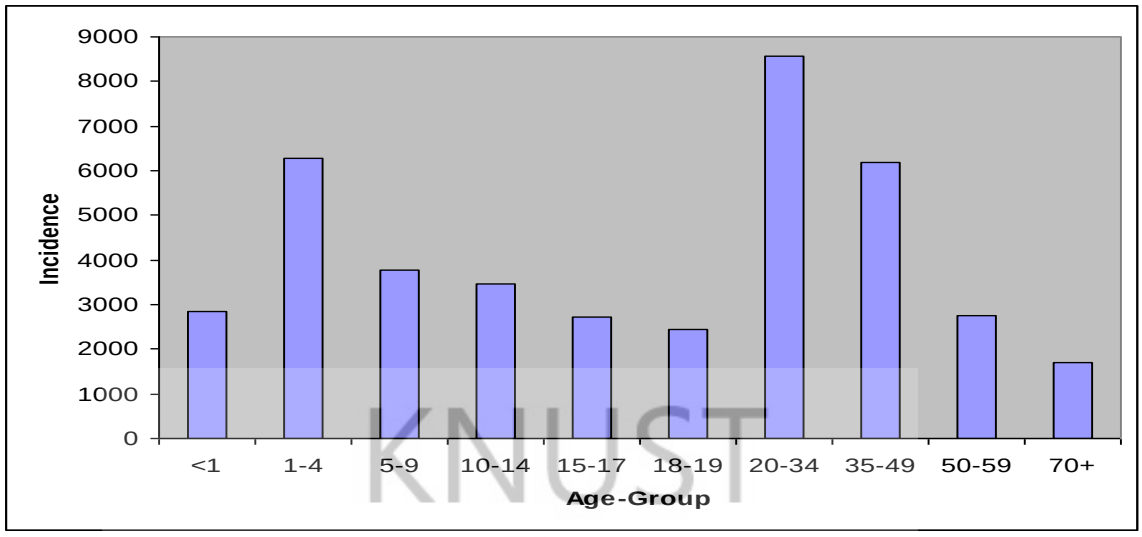

Figure 4.1: A bar chart depicting the number of cases of malaria for the various age groups.

From Table 4.2, it can be seen that the highest malaria cases (incidence) occurred among those found between 20-34 age group and it directly followed by 1-4 and 35-49 age groups. The lowest cases or incidence were recorded for 70+ age group (adults).

A bar chart in Figure 4.1 confirms it with the highest bar depicting more cases or incidence being those in 20-34 age group followed by 1-4 and 35-49 age groups.

Table 4.3: The Occurrence of Malaria Cases for Gender for the period.

Gender	**Cases**
Female	23460
male	17259
Total	**40719**

Source: Obuasi Government Hospital.

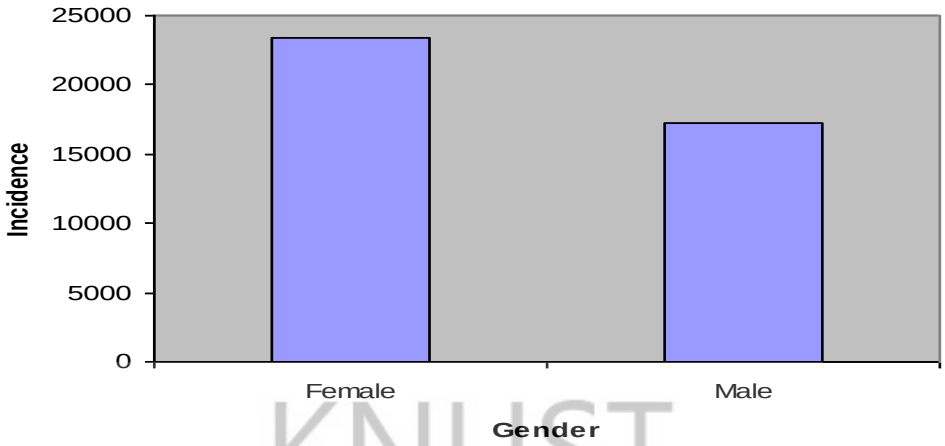

Figure 4.2: A bar chart depicting cases of malaria with respect to gender.

Out of a total of 40719 cases or incidence, the female recorded the highest followed by males. Both T able 4.3 and Figure 4.2 confirm it.

Table 4.4: The Occurrence of Malaria cases for time in quarters for 2007 to2010.

Time in quarters	**Cases**
Quarter1	8671
Quarter 2	9642
Quarter 3	10486
Quarter 4	11920
Total	**40719**

Source: Obuasi Government Hospital

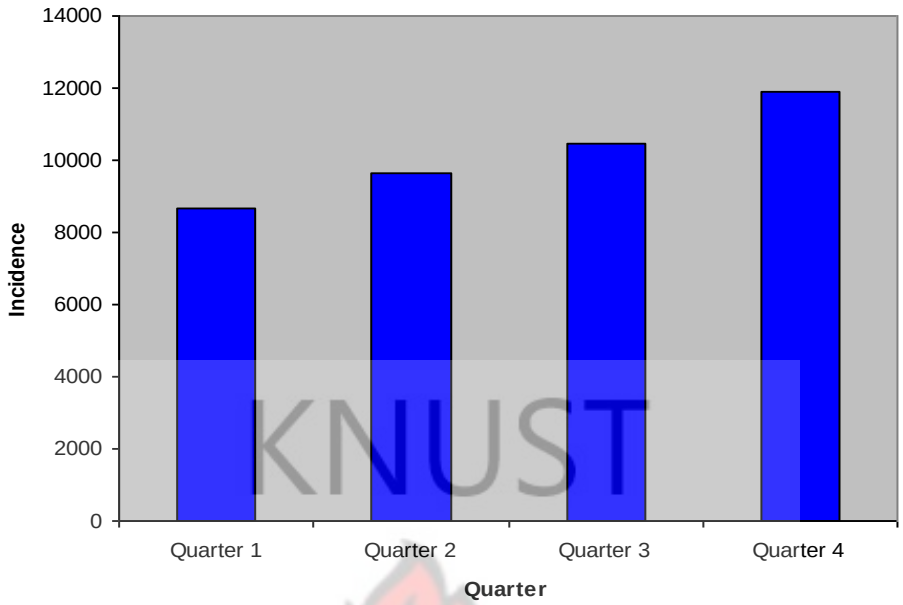

Figure 4.3: A bar chart of malaria cases in quarterly time.

With respect to the quarterly time, quarter 4 saw the highest cases or incidence of malaria and it directly followed by quarter 3, quarter 2 and quarter 1 in that order. This is consistent with Table 4.4 and Figure 4.3.

4.3 Modeling and Criteria for assessing Model Goodness of Fit

The model goodness of fit assessment criteria results as outputted from the R software using the glm procedure for Poisson regression models are shown in Table 4.5 and its coefficients estimate are also depicted in Table 4.6 with malaria cases being the response variable given the independent variables age, gender and time in quarters. Again let α_i, be the respective intercept estimates where i=1, 2… 10 represent the various intercept for the 10 models and $\beta_i, \beta_j \text{ and } \beta_l$ denote the estimates of the independent variables for i= 1,2,3 and 4;j=0 and 1 and l=1,2,3,…,10 representing time in quarters, gender and age.

4.3.1 Modeling the Occurrence of malaria Cases

Table 4.5: Poisson Regression Models with their AIC's

Models	AIC's
1. $\ln(mean\_cases) = \alpha_1 + \beta_k quarters_i, k = 1,2,..,3$	19393
2. $\ln(mean\_cases) = \alpha_2 + \beta_k gender_j, k = 1$	19733
3. $\ln(mean\_cases) = \alpha_3 + \beta_k age_l, k = 1,...,9$	17030
4. $\ln(mean\_cases) = \alpha_4 + \beta_k^* quarters_i + \beta_m^{**} gender_j, k = 1,2,..,3 and m = 1$	19367
5. $\ln(mean\_cases) = \alpha_5 + \beta_1 (quarters_i * gender_j)$	19366
6. $\ln(mean\_cases) = \alpha_6 + \beta_k^* quarters_i + \beta_n^{***} age_l, k = 1,2,...,3 and n = 1,2,...9$	16664
7. $\ln(mean\_cases) = \alpha_6 + \beta_2 (quarters_i * age_l)$	11737
8. $\ln(mean\_cases) = \alpha_7 + \beta_m^{**} gender_j + \beta_n^{***} age_l, m = 1 and n = 1,2,...9$	17004
9. $\ln(mean\_cases) = \alpha_8 + \beta_3 (gender_j * age_l)$	16638
10. $\ln(mean\_cases) = \alpha_{10} + \beta_k^* quarters_i + \beta_m^{**} gender_j + \beta_n^{***} age_l, k = 1,2,..3, m = 1 and, n = 1,2,3,...,9$	16638

From Table 4.5 model 10 was chosen to assess the goodness of fit test because it satisfied all the assumptions with an AIC value of 16638, a deviance of 14505 on 306 degrees of freedom following the chi-square distribution $\chi^2_{(1)}$ .The corresponding p-value associated with this model is < 0.00 and this indicates over-dispersion. Table 4.6 shows the parameter estimates of the selected Poisson regression model for model 10.

Table 4.6: Parameter Estimates of the Selected Poisson Model

Coefficients	Estimates	Standard errors	z values	pr(>|z|)
intercept	5.246723	0.015673	337.962	<0.001
quarters2	-0.148015	0.012591	-11.755	<0.001
quarters3	-0.083806	0.02379	-6.77	<0.001
quarters4	-0.239595	0.012912	-18.557	<0.001
female	0.047856	0.009074	5.224	<0.001
age1-4	-0.644854	0.022304	-28.912	<0.001
age5-9	-0.113363	0.019051	-5.95	<0.001
age10-14	-0.0607	0.018792	-3.23	0.001237
age15-17	0.041582	0.018315	2.27	0.023184
age18-19	-0.062156	0.018799	-3.306	0.000945
age20-34	-0.444213	0.020932	-21.222	<0.001
age35-49	-0.698124	0.022701	-30.754	<0.001
age50-59	-0.039282	0.018689	-2.102	0.03559
age70+	-0.126099	0.019116	-6.597	<0.001

A dispersion parameter of 47.40196 shows that the data is over-dispersed i.e. a situation where the variance of the response variable exceeds the mean. In the nut shell, Poisson regression model cannot fit the data. A negative binomial regression model is considered to be convenient and practical; they handle over-dispersion; they allow the likelihood ratio and other standard maximum likelihood tests to be implemented. Table 4.7 depicts the parameter estimates after validating the Poisson regression model using negative binomial regression model and the parameter estimates are depicted in Table 4.7.

Table 4.7: Negative Binomial Regression Model Parameter Estimates

Coefficients	Estimates	Standard Error	z value	pr(>|z|)
intercept	5.33495	0.13895	38.396	<0.001
quarters2	-0.15686	0.10516	-1.35305	0.1358
quarters3	-0.10387	0.10514	-0.988	0.32317
quarters4	-0.2734	0.07439	-2.599	0.00936
female	0.05084	0.16651	0.683	0.49435
age1-4	-0.67796	0.16611	-4.072	<0.001
age5-9	-0.15571	0.16609	-0.937	0.34856
age10-14	-0.1255	0.16601	-0.756	0.44988
age15-17	0.02526	0.16605	0.152	0.87904
age18-19	-0.06559	0.16632	-0.395	0.69283
age20-34	-0.46634	0.16632	-2.804	0.00505
age35-49	-0.72386	0.16656	-4.346	<0.001
age50-59	-0.05718	0.16605	-0.344	0.7306
age70+	-0.14917	0.1661	-0.898	0.36916

The AIC of this model is 3767.9; a deviance of 347.06 on 306 degrees of freedom also following the chi-square distribution $\chi^2_{(1)}$ with one degree of freedom .The dispersion parameter was found to be 1.227679 and a p-value of 0.0528855 indicating the significance of the model.

$$\ln(mean\_cases) = 5.33495 - 0.15686quarters2 - 0.10387quarters3 - 0.2734quarters4 + 0.05084\,female - 0.67796age(1-4) - 0.15571age(5-9) - 0.1255age(10-14) + 0.02526age(15-17) - 0.06559age(18-19) - 0.46634age(20-34) - 0.72386age(35-49) - 0.05718age(50-59) - 0.14917age(70+)$$

(76)

The goodness of fit results shown in Table 4.8 below clearly shows that the negative binomial regression model fits better the occurrence of malaria cases data better than the Poisson model. First, the ratios of deviance and Pearson chi-square to degree of freedom $\frac{D}{DF} \sim \chi^2_{(1)}$ for the Poisson are much larger than 1, which indicate an over-dispersion in the data and hence Poisson does not do a good job of modeling such kinds of data. Second, ($\frac{D}{DF} \sim \chi^2_{(1)}$) ratios for negative binomial model are both close to one which shows a good fit to the data. Third, the lower deviance, Pearson chi-square and the larger log likelihood values of negative binomial as against those of Poisson, all of them together share the same conclusion of favouring the negative binomial model.

Table 4.8 Assessment Criteria for Poisson and Negative Binomial Regression Models

Assessment Parameter	Poisson Regression model	Negative Binomial Regression Model
AIC	16638	3767.9
Residual Deviance	14505	347.06
Degrees of Freedom	306	306
Dispersion parameter	47.40196	1.13418

From the Table 4.7, the expected number of occurrences of malaria cases is 0.15685, 0.15685 and 0.10387 times lower in quarters2, quarters3 and quarters4 respectively compared to the base level (quarter1). Meanwhile quarters4 is the most significant quarter at 5% α-level, with the expected cases of $e^{-0.2734}(0.7607)$ approximately 76% of all cases. For gender, the expected occurrence of malaria cases is 0.05084 times

higher in females than in their male (base level) counterparts. It is not significant because malaria is independent of gender, but relevant and significant in a non – technical sense and therefore gender cannot be ignored. Gender norms and values that influence the division of labour, leisure patterns, and sleeping arrangements may lead to different patterns of exposure to mosquitoes for men and women. There are also gender dimensions in the accessing of treatment and care for malaria, and in the use of preventative measures such as mosquito nets. A thorough understanding of the gender related dynamics of treatment-seeking behaviour, as well as of decision-making, resource allocation and financial authority within households is key to ensuring effective malaria control programmes. Therefore, gender and malaria issues are increasingly being incorporated into malaria control strategies in order to improve their coverage and effectiveness in different contexts.

Regarding the age categories, it can be seen that, the expected difference in log count of malaria cases between the all the ages and the bases level. The expected occurrence of malaria cases for 1-4, 20-34, and 35-49 age groups are 0.67796, 0.46634 and 0.72386 times lower compared to the base level (<1 year old). They were all significant at 5% α-level with those found in 20-34 age category being the most significant accounting for $e^{-046634}$ (0.6272), which is about 62% percent of all cases in malaria compared to cases recorded for those in 1-4 and 35-49 with $e^{-0.67796}$ (0.5070) approximately 50% and $e^{-0.7238}$ (0.4848) which is approximately 48% respectively. Similarly, the expected occurrence of malaria cases is 0.15571, 0.06559, 0.05718 and 0.14917 times lower for

5-9, 18-19, 50-59 and 70+age categories respectively compared to the base level (<1year group) but it is 0.02526 times higher in 15-17 age group compared to the base level year group but none was significant at 5% α - level.

Based on the discussion of model selection coupled with established model outputs, the negative binomial was selected on the grounds of producing relatively low ratio deviance (347.06) and chi-square distribution to the degree of freedom (306) values (depicting better model fit to the data), thus aiding the estimation of the parameters shown in Table 4.7

4.3.2 Modeling the incidence of severe malaria cases.

The total number of observations used was 40717with severe malaria cases as the response variable in the model. The offset variable used as the log of the total number of malaria cases. The following models were obtained with δ_s, s=1,2,…10 being the respective intercepts and $\gamma_k^*, \gamma_m^{**} and, \gamma_n^{***}$ denoting the parameter estimates of gender, time and the age categories.

Table 4.9 Poisson Regression Models for the Incidence of Severe Malaria Cases

Models	AIC's
1. $\ln\left(\frac{severe\_cases}{total}\right) = \delta_1 + \gamma_k gender_i, k = 0$	33341
2. $\ln\left(\frac{severe\_cases}{total}\right) = \delta_2 + \gamma_k time_j,$	22937
3. $\ln\left(\frac{severe\_cases}{total}\right) = \delta_3 + \gamma_k age_l, k = 1,2,...,9$	25623
4. $\ln\left(\frac{severe\_cases}{total}\right) = \delta_4 + \gamma^*_k gender_i + \gamma^{**}_m time_j, k = 0$	22080
5. $\ln\left(\frac{severe\_cases}{total}\right) = \delta_5 + \gamma_1 (gender_i * time_j)$	21858
6. $\ln\left(\frac{severe\_cases}{total}\right) = \delta_6 + \gamma^*_k gender_i + \gamma^{***}_n age_l, k = 0, n = 1,2,...9$	24736
7. $\ln\left(\frac{severe\_cases}{total}\right) = \delta_7 + \gamma_2 (gender_i * age_l)$	24177
8. $\ln\left(\frac{severe\_cases}{total}\right) = \delta_8 + \gamma^{**}_m time_j + \gamma^{***}_n age_l$	14483
9. $\ln\left(\frac{severe\_cases}{total}\right) = \delta_9 + \gamma_3 (time_j * age_l)$	13587
10. $\ln\left(\frac{severe\_cases}{total}\right) = \delta_{10} + \gamma^*_k gender_i + \gamma^{**}_m time_j + \gamma^{***}_n age_l$	13613

Table 4.9 shows the various models for the incidence of severe malaria cases and the respective AICs. Assessing the goodness of fit, model 10 below, was chosen based on an AIC of 13613, a deviance of 12955 on 75 degrees of freedom following a chi-square distribution and a p-value of < 0.00, thus statistically significant and the parameter estimates are displayed in Table 4.10

Table 4.10 Parameter Estimates for the Incidence of Severe Malaria Cases using the Poisson regression model

Coefficient	Estimate	Standard error	z-value	Pr(<|z|)
Intercept	4.086346	0.028027	145.798	<2e-16
male	-0.329643	0.011236	-29.337	<2e-16
time	0.552520	0.005598	98.706	<2e-16
age10-14	0.255481	0.028308	9.025	<2e-16
age15-17	-0.029768	0.030300	-0.982	0.326
age18-19	-0.164388	0.031296	-5.257	<2e-16
Age1-4	0.814577	0.025531	31.905	<2e-16
age20-34	1.074733	0.024636	43.625	<2e-16
age35-49	0.753653	0.0225855	29.149	<2e-16
age5-9	0.321030	0.027975	11.476	<2e-16
age50-59	-0.027483	0.032402	-0.904	0.366
age70+	-0.602318	0.036181	-16.647	<2e-16

However, it was over-dispersed meaning the variance far exceeded the expected mean. It therefore became necessary to validate the model using negative binomial regression model. Validating the model 10 yielded model 10.1 whose parameter estimates are displayed in Table 4.11 and the equation is displayed below.

Table 4.11 Parameter Estimates for the Incidence of Severe Malaria Cases using the Negative Binomial Regression Model

Coefficients	Estimate	Standard error	z-value	Pr(>|z|)
Intercept	4.7376	0.3180	14.896	<2e-16
male	-0.2508	0.1543	-1.626	0.1040
time	0.3470	0.0690	5.029	4.94e-07
age10-14	0.1236	0.3616	0.342	0.7326
age15-17	-0.1047	0.3618	-0.289	0.7723
age18-19	-0.1983	0.3618	-0.548	0.5836
age1-4	0.6443	0.3614	1.783	0.0746
age20-34	0.8535	0.313	2.363	0.0182
age35-49	0.5838	0.3614	1.616	0.1062
age5-9	0.1200	0.3616	0.332	0.7400
age50-59	-0.1295	0.3618	-0.358	0.7205
age70+	-0.7706	0.3626	-2.125	0.0336

$$\ln\left(\frac{severe\_cases}{total}\right) = 4.7376 - 0.2508male + 0.3470time + 0.1236age(10-14) - 0.1047age(15-17) - 0.1983age(18-19) + 0.6443age(1-4) + 0.8535age(20-34) + 0.5838age(35-49) + 0.1200age(5-9) - 0.1295age(50-59) - 0.7706age70+ \quad (77)$$

Model 10.1 above gave a good fit to the data because it was statistically significant with reference to the AIC of 1176.9, a deviance of 95.563 on 75 degrees of freedom following a chi-square distribution, a dispersion parameter of 1.271507 and a p-value of <0.00. Table 4.12 summarizes the assessment criteria for selecting negative binomial regression model.

Table 4.12 Assessment Criteria for Poisson and the Negative Binomial Regression Models for the Incidence of Severe Malaria Cases

Assessment parameter	Poisson regression model	Negative binomial regression model
AIC	13613	1176.9
Deviance	12955	95.563
Degrees of freedom	75	75
Dispersion parameter	172.7333	1.271507

From the results in Table 4.11, it can be seen that the expected incidence of severe malaria cases is 0.2508 lower in the base level (female) compared to their male counterparts. It is not significant in the sense that malaria is not dependent on gender.

The variable time (years) has a coefficient of 0.3470 which is statistically significant at 5% α level. This means that for each one-unit increase in time (years), the expected log count of the incidence of severe malaria cases will increase by 0.35262 (i.e.) $e^{0.35262}$ (1.4227). It also follows that as time goes on there will an increase in severe malaria cases.

Now, for the age categories, the expected incidence of severe malaria cases for 10-14, 1-4, 20-34, 35-49 and 5-9 are respectively 0.1236, 0.6443, 0.8535, 0.5838 and 0.1200 higher compared to the base level (< 1 year old). Among these age groups, 20-34 age category was significant at 5% α level with an associated p-value of < 0.0182. Again the expected log count of the incidence of severe malaria cases is 0.1047, 0.1983, 0.1295 and 0.7706 lower in 15-17, 18-19,50-59 and 70+ age categories respectively compared

to the base level (<1 year old) with 70+age category being the most significant at 5% α level with an associated p-value of 0.0336.

CHAPTER 5

CONCLUSION AND RECOMMENDATONS

5.1 Introduction

This chapter presents conclusions from the research and some recommendations

5.2 Conclusions

The objective of this research was to model the occurrence of malaria cases given the age , gender and time (quarters); to model the incidence of severe malaria cases given the age, gender and time in years and lastly to validate the two models using negative binomial regression model. Data from Obuasi Government Hospital were utilized in this study. Severe malaria cases confirmed by the laboratory, simple malaria (laboratory confirmed) and simple malaria (non-laboratory confirmed) were used in the analysis and modeling. Both Poisson and negative binomial regression models were used and well known statistical goodness of fit model assessment criteria were used in selecting which model will fit the malaria cases better.

Based on the results, the negative binomial regression model was found to fit the data better than the Poisson regression model. In modeling the occurrence of malaria cases, the analysis produced a reasonable AIC values, (16638) deviance (14505); p-value <0.00 for the Poisson model and a dispersion parameter of 47.40196 showing an extra-Poisson variation or over-dispersion in the data; leading to overestimated standard errors, thus inaccurate parameter estimates apparently due to a violation of one of its main assumption of the equality of mean and variance parameters. Because of over-dispersion, it became a necessary tool to validate the Poisson regression models using

negative binomial. The deviance for the negative binomial regression model displayed in Table 4.8 is 347.06 on 306 degrees of freedom.

The occurrence of malaria cases in quarter4 (October- December) was found to be $e^{-0.2734}$ (0.7607) accounting for 76% of all laboratory confirmed cases recorded between 2007 to 2010. Regarding the various age groups, the occurrence of malaria found to be statistically significant at 5% α-level for 1-4, 20-34 and 35-49 age groups with the expected cases being $e^{-0.67796}$ (0.5070), $e^{-0.67796}$ (0.5070) and $e^{-0.7238}$ (0.4848). 50% laboratory confirmed cases were children below 5 years with more cases recorded for 20-34 age groups accounting for (48%) and those found in 20-34 age groups recording the highest number of cases accounting for 62% of laboratory confirmed cases.

Again the negative binomial regression model fitted the data. The variable time (in years) had a coefficient of 0.3470 which is statistically significant at 5% α level. This means that for each one-unit increase in time (years), the expected log count of the incidence of severe malaria cases will increase by 0.35262 (i.e.) $e^{0.35262}$ (1.4227). It also follows that as time goes on there will be an increase in severe malaria cases. Similarly, for the age categories, the expected incidence of severe malaria cases is 0.1236, 0.6443, 0.8535,0.5838 and 0.1200 higher in 10-14, 1-4,20-34, 35-49 and 5-9 age categories respectively compared the base level. Among these age groups, 20-34 age category was significant at 5% α level with an associated p-value of < 0.0182. Again the expected log count of the incidence of severe malaria cases are 0.1047, 0.1983, 0.1295 and 0.7706 lower in 15-17, 18-19,50-59 and 70+ age categories respectively compared the base

level with 70+age category being the most significant at 5% α level with an associated p-value of 0.0336. Both model showed the independence of malaria with respect to gender.

With reference to the findings of the research, it can be concluded that:

- the occurrence of malaria cases and the incidence of severe malaria cases were independent of gender.
- the occurrence of malaria cases was found to be very high in the last quarter (October-December) between 2007 to 2010.
- the occurrence of malaria was significantly high among infants , children under 5 years old and adults aged between 20-49 years.
- between 2007 to 2010, the incidence of severe malaria cases increased.
- the incidence of severe malaria cases was found to be very high among infants and adults aged 20-34 and 70+ years.

5.2 Recommendations

On the basis of the findings of the research, the following recommendations were made:

- since more cases were recorded in the last quarter (October-December) of the years considered apparently due to some seasonal changes like rainfall, it is imperative that programmes and campaigns meant to reduce the menace of malaria should be carried out before, during and after the seasonal changes.
- because most of the affected group were people between 0-4years (infants and children) and 20-34 and 35-49 age groups, which are working force of the country and that matter the municipality and therefore government under

auspices of the municipal health directorate must intensify its campaign on the effects of severe malaria. The economic gains of the municipality will increase if these people are free from malaria.

- the education on the use of treated mosquito nets and environmental cleanliness such clearing of bushes, desilting of choked gutters and getting rid of stagnant waters which serves as breeding ground for plasmodium larvae, must be enforced in order to bring the disease under control.
- the Internal Residual Mass Spraying exercise introduced by the Anglo Gold Ashanti as part of its corporate social responsibility in the municipality must intensify its campaign in order to bring the disease under control.

It is therefore suggested that more studies and research be carried out in highly utilized hospital in other to check the incidence of malaria so that appropriate measures and strategies could be adopted to curtail its spread.

REFERENCES

Alan, Agresti (2007), An introduction to Categorical data analysis, 2nd edition, John Wiley and sons; Hoboken, New Jesey.

AngloGold Ashanti Annual Report (2006), Report to Society, Campaign at Obuasi halves malaria incidence.

AngloGold Ashanti Annual Report (2006), Report to Society, Campaign at Obuasi halves malarial.

Hoshen M B and Morse, A P (2004), A weather driven model of malaria transmission, Malaria journal 3(32).

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

Olaleye,B.O., Williams,A, D'Alessandro U, Weber M.M.,Mulholland K, Okorie C, Lenqerock P, Bennett S, Greenwood B.M., (1997) Clinical predictors of malaria in Gambia children with fever or history of fever, Royal society of tropical medicine and hygiene.

Greenwood B.M., Pickering H, (2004), A malaria control trial using insecticide-treated bed nets and targeted chemoprophylaxis in a rural area of The Gambia, West Africa: 1. A review of the epidemiology and control of malaria in The Gambia, West Africa, Journal, science direct magazine.

Box, G. E. P. (1979) Robustness in the strategy of scientific model building, in R. L. Launer and G. N. Wilkinson (eds), Robustness in Statistics, Academic Press, New York, pp. 201–236.

Bruce-Chwatt LJ & de Zuleta J (1980) The Rise and Fall of Malaria in Europe. Oxford University Press, Oxford.

Cameron, A. C. and Trivedi, P. K. (1996), Count Data Models for Financial Data, and book of Statistics, Vol. 14, Statistical Methods in Finance, 363-392, Amsterdam, North-Holland.

Cameron, A. C. and Trivedi, Pravin K. (1998), Regression Analysis of Count Data, econometric Society Monographs, No. 30, Cambridge University Press.

Carter R E, Lipsitz SR, Tilley BC (2005), Quasi-likelihood estimation for relative risk regression models.

Chandramohan, D. & Greenwood, B. (1998) Is there an interaction between human immunodeficiency virus and plasmodium falciparum. International Journal of Epidemiology 27:296-301.

Craig, M. & Sharp, B. (1997) Comparative evaluation of four techniques for the diagnosis of Plasmodium Falciparam infections.Pub Med. Variables, Advanced Quantitative Techniques in the Social Sciences Series 7, SAGE Publications.

Czado C, Vinzenz E, Aleksey M, Stefan W. (2007). Zero-inflated Generalized Poisson Models with Regression Effects on the Mean, Dispersion and Zero-inflation Level Applied to Patent Outsourcing rates. Statistical Modelling, 7(2), 125-153.

Devi N.P. and Jauhari R.K (2006)., Climatic variables and malaria incidence in Dehradun, Uttaranchal, India, Journal Vector Borne Disease, 43(1): 21–28.

Efron, B. (1986), Double Exponential Families and Their Use in Generalized Linear Regressions, Journal of the American Statistical Association, 81, 709-721.

Famoye F, and Karan P. Singh (2006) Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data 1Central Michigan University and Health Science Centre Journal of Data Science, 117-130.

Gagnon AS, Smoyer-Tomic KE, Bush ABG: The El Niño Southern Oscillation and malaria epidemics in South America. International Journal of Biometeorology 2002, 46:81-89. Pub Med, Malaria journal.

Gardner, Howard (1995). Reflections on multiple intelligences. Phi Delta Kappan, 77 200-Generalized Linear Models (with J.A. Nelder),(1983), Chapman and Hall, London.

Ghana Malaria Alert, 2007, Volume 1, Issue 1 • March 2007 Volume.

Gilles, H. & Warrell, D. (1993). Bruce-Chwatt's Essential Malariology. Edward Arnold. London.

Gomez-Elipe A., Otero A., Herp van M. and Aguirre-Jaime A. (2007), Forecasting malaria incidence based on quarterly case reports and environmental factors in Karuzi, Burundi, 1997–2003, Malaria Journal; 6: 129.

Greenwood BM , Bradley A.K., Byass P, Grenwood A.M., Snow R.W., Hayes R.J., Abhn J, (1988) Comparison of two strategies for the control of malaria within a primary health care programme in the Gambia. Lancet, 1988, 1: 1121±1127.

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, pp 228-279.

Guikema, S.D., Coffelt, J.P., (2007). A flexible count data regression model for risk analysis. Risk Analysis 28(1), 213-223.

Jörgen N .and Jonas N, (2008), A Count Data Model with Endogenous Household Specific Censoring: the Number of Nights to Stay, Empirical Economics, Vol. 35, No. 1, 179-193. 54.

Kleinschmidt I & Sharp B (2002). Patterns in age-specific malaria incidence inapopulation exposed to low levels of malaria transmission intensity. Tropical Medicine and International Health, 6: 986-991.

Kudur, J. M., Demlo, L. K. (1985), Small area variation in Iowa hospital utilization. Iowa Medicine 75, 213-217.

Kumekpor Tom K.B. (2002) Research Methods and Techniques of Social Research. Accra Ghana: Son Life Press and Services.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Techno metrics 34, 1-14.

Leedy, P. (1989), Practical Research: planning and design, 4th edition (New York, Macmillan).

Lemaire L, (1991), Negative binomial or Poisson inverse Gaussian, Astin Bulleting, 21, 167-168.

Liu, W and Cela J, (2008), Count Data Models in SAS, Proceedings SAS Global Forum 2008, paper 371-2008.

Lundberg, O.(1940), "On Random Processes and their Application to Sickness and Accident Statistics". Almquist and Wiksell, Uppsala.

Mc Nutt L A, Wu C, Xue X, Hafner JP (2003), Estimating relative risk in cohort studies and clinical trials of common events. American Journal of Epidemiology. 157: 940-943.

Mullahy, J. (1986). Specification and testing of some modified count models. Journal of Econometrics 33, 341-365.

Nelder, J .A. and Wedderburn R.W.M., (1972), Generalized linear models, Journal of the statistical society Cameron, A.C., Trivedi, P.K., (1998), Regression Analysis of Count Data, Cambridge University Press, Cambridge, U.K.

Claus B, Steven W. L., Sian E. C., Andy D, Musa J, Margaret P, and Christopher J.T., (2007), High spatial resolution mapping of malaria transmission risk in the Gambia, West Africa using landsat satellite imagery, The American Journal of tropical magazine and hygiene.

Winkelmann R and Zimmernann K.F.,(1991), A new approach for modelling economic count data. Theory and applications, Journal of economic surveys, vol. 9. 1-24.

Nelson, Jon P. and Douglas Young (2008), Effects of youth, price, and audience size on alcohol advertising in magazines, Health Economics, 17(4), 551-556.

Newbold, E. M. (1926). "A Contribution to the Study of the Human Factor in the Causation of Accidents". Med. Res. Count. industry. Fatigue Res. Bd, Report No. 34. H.M.S.O., London.

Parodi S and Bottarelli E. (2005) Control for confounding in case-control studies. Annali della Facoltà di Medicina Veterinaria di Parma, XXV, 14-46.

Piet de Jong and Gillian Z. Heller (2008), Generalized Linear Models for Insurance Data, Cambridge University Press, New York.

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

Roll Back Malaria,(1999) National Malaria Control Programme, Ghana Health Service.

McCombie S.C., (1999), Treatment seeking for malaria: A review of recent research, Journal, science direct magazine.

Sachs, J. (2000) Economic analyses indicate the burden of malaria is great. In: The African Summit on Roll Back Malaria, Abuja WHO/CDS/RBM/2000.17, WHO, Geneva, pp. 27±35.

Seal, H L (1982), Mixed Poisson- an ideal distribution of claim number mvsv,293-295.

Snow R W, Craig M H, Deichmann U, Le Sueur D (1999), A continental risk map for malaria mortality among African children. Parasitology today, 15: 99±104.

Snow RW, Korenromp, G, (1996), Pediatric mortality in Africa: plasmodium falciparum malaria as a cause or risk? Kenya Medical Research Institute.

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.

UN (2003), United Nations Technical Health and Housing Report.

Verhoeff FH, Brabin BJ, Hart CA, Chimsuku L, Kazembe P, Broadhead RL (1999). Increased prevalence of malaria in HIV-infected pregnant women and its implications for malaria control. Tropical Medicine and International Health, 4: 5-12.

Wattanavadee Sriwattanapongse and Metta Kuning(2008) Modeling Malaria Incidence in North-Western Thailand Department of Statistics, Faculty of Science, Chiang Mai University, Chiang .

Wennberg, J. and Freeman, J. (1987), Are hospital services rationed in New Haven or over utilized in Boston? The Lancet 1, 1185- 1188.

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

Whitworth J, Morgan D, Quigley M, Smith A, Mayanja B, Eotu H, Omoding N, Okongo M, Malamba S, Ojwiya A (2000). Effect of HIV-1 and increasing immunosuppression on malaria parasitaemia and clinical episodes in adults in rural Uganda: a cohort study.Lancet, 356: 1051-1056.

WHO (2003), Expect Committee on Malaria-Twentieth Report. World Health Organisation, Geneva.

Willmot, G E (1987), The Poisson inverse Gaussian distribution as an alternative to Negative binomial, Sandinavian actuarial journal.

Wilson, P. and Tedeschyi, P.,(1984), Community correlates of hospital use. Health Services Research 19, 133-146.

World Bank Report (1993). World Development Report: Investing in Health. Oxford University Press: New York.

Yang, Chih-Hai (2007), What factors inspire the high entry flow in Taiwan's manufacturing industries-A count entry model approach, Applied Economics. Vol. 39, 1817-1831.

Zou G (2004), A modified Poisson regression approach to prospective studies with binary data, American Journal of Epidemiology,154: 702-706.