KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI

COLLEGE OF SCIENCE

DEPARTMENT OF MATHEMATICS

MODELLING THE OCCURRENCE AND INCIDENCE OF SEROPOSITIVE HIV CASES

MODEL: A CASE STUDY AT KORLE - BU TEACHING HOSPITAL (2008 – 2010)

BY

FORSTER OWUSU

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF PHILOSOPHY IN MATHEMATICS.

SEPTEMBER, 2011

KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI

COLLEGE OF SCIENCE

FACULTY OF PHYSICAL SCIENCES

DEPARTMENT OF MATHEMATICS

MODELLING THE OCCURRENCE AND INCIDENCE OF SEROPOSITIVE HIV CASES

MODEL: A CASE STUDY AT KORLE - BU TEACHING HOSPITAL (2008 – 2010)

BY

FORSTER OWUSU

SEPTEMBER, 2011

DECLARATION

I hereby declare that this submission is my own work towards the award of the Mphil (Mathematics) Degree and that to the best of my knowledge, it contains no material previously published by another person or 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.

……. …….

FORSTER OWUSU

……. …….

NANA KENA FREMPONG

(SUPERVISOR)

……. …….

MR F. K. DARKWA

(HEAD OF DEPARTMENT)

ABSTRACT

Acquired Immune Deficiency Syndrome(AIDS) was identified in the early 1980s in the United States and since then a number of people have been affected by the pandemic. Though a lot of education has been made on the repercussions of AIDS on the socio economic development in our dear nation and a lot of research has been carried out in the fight against the menace, the disease continues to be endemic.

This study seeks to model the occurrence and incidence of Human Immune Virus(HIV) seropositive cases at Korle – Bu Teaching Hospital from January 2008 to August 2010. The data for this research is obtained from the chest clinic of the Korle – Bu Teaching Hospital. The data had a total of 220 observations and five variables under study. In the data consisting of gender, age and month of testing in a given year, the number of positive HIV cases was modeled using the Poisson regression. The Poisson regression was fitted to the data using the SAS statistical software (version 9.1). The negative binomial regression model was used to validate the Poisson regression model in the case of over dispersion. From the results, where the best model that fitted the data had an intercept estimate of -0.6424 which was the log of the expected number of occurrence of HIV positive cases when all other variables in the model are evaluated at zero, the number of occurrence of HIV positive cases for females was more than that of males. Also Adolescents between the ages of 35 -39 years have a higher incidence rate of HIV positive cases.

Key Words: AIDS, HIV, SERPOSITIVE, POISSON REGRESSION, OVRDISPERSSION, NEGATIVE BINOMIAL REGRESSION

ACKNOWLEDGEMENTS

My enormous thanks go to the Almighty God for giving me the knowledge, wisdom, strength, understanding, care and guidance in the writing of this thesis.

Special thanks also go to my Supervisor, Nana KenaFrempongwho has played a very good endless dedicated and supervisory role. His interest in the progress of this work led to the successful completion of this thesis. His brotherly love and patience will never be forgotten.

I would also want to thank all Lecturers of the Centre of Entrepreneurship and Enterprise Development, Kumasi Polytechnic especially the Head of Department, Mr. Solomon KwartengForkouh whose sense of direction have been tremendous. I appreciate every support they gave me during my studies. May the Almighty God richly bless them. I also say a big thank you to Mrs Dwomo-Fokuo Elizabeth for her encouragement and assistance in pursing this project.

To Mr U. U. P. Fleischer, head of department of Mathematics and Science Department, I say your wonderful encouragement and guidance will never be forgotten.

Last, but not least, I would like to thank my friends, my family members, especially my parents, my brother Abel, and my sisters, Selina and Rachel. Without them I would have never made it this far in life. They have been there for me every step of the way, have always loved me unconditionally, and have supported me through all of my tough decisions. Were it not for them, I would not be where I am today. I say thank you to them.

DEDICATION

This work is dedicated to my parents

Rev Oliver Owusu and Mrs Comfort Owusu

Who taught me that even the largest task can be accomplished if it is done one step at a time

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND OF STUDY

Acquired Immune Deficiency Syndrome (AIDS) is a medical condition that is clinically manifested when Human Immunodeficiency Virus (HIV) takes control over the immune system. As HIV progressively damages these cells, the body becomes more vulnerable to infections, which it will have difficulty in fighting off. It is at the point of very advanced HIV infection that a person is said to have AIDS. It can be years before HIV has damaged the immune system enough for AIDS to develop.

Since AIDS was first identified in the early 1980s, an unprecedented number of people have been affected by the global AIDS epidemic. According to estimates by WHO (2010) and UNAIDS (2010) reports, 33.4 million people were living with HIV at the end of 2008world wide. In that same year, some 2.7 million people became newly infected, and 2.0 million died of AIDS, including 280 000 children. Two thirds of HIV infections are in sub-Saharan Africa, according to the report.

The duration of infectiousness to the clinically manifested AIDS is quite long, so most often, people don't actually get AIDS. They might get infected with HIV and later develop AIDS. The blood, vaginal fluid, semen, and breast milk of people infected with HIV has enough of the virus in it to infect other people. Most people get the HIV virus by illicit sexual behavior i.e. having sex with an infected person, sharing a needle (shooting drugs) with someone who's infected, mother to child transmission ie. being born when their mother is infected, or drinking the breast milk of an infected woman and getting a

transfusion of infected blood used to be a way people got AIDS, but now the blood supply is screened very carefully and the risk is extremely low.

The impact on societies, economies and infrastructures is worrysome. In countries most severely affected, life expectancy has been reduced by as much as 20 years. Young adults in their productive years are the most at-risk population; so many countries have faced a slow-down in economic growth and an increase in household poverty. In Asia, HIV and AIDS causes a greater loss of productivity than any other disease. An adult's most productive years are also their most reproductive and so many of the age group who have died from AIDS have left children behind. In sub-Saharan Africa the AIDS epidemic has orphaned nearly 12 million children, UNAIDS, (2008)

When HIV/AIDS was first identified in Ghana in March 1986, the National rate of infection was 1.5%. Since 1986, more than 52, 961 HIV/AIDS cases have been reported in the health institutions in Ghana. This represents 30% of cases in the country as majority of the victims patronise the traditional health centres, prayer camps and others do not report their illnesses due to the fear of stigma and discrimination. Therefore, non-reported cases of HIV/AIDS are more than the reported cases in Ghana. This means the prevalent rate of 3.6% in the country is misleading because majority of carriers of HIV infections are difficult to be identified in the country. It has been reported that about 130 people in Ghana contract Aids daily and it is estimated that 125 people would die from the dreadful disease daily by the year 2009 if the rate of infection continues at 3.6%. Ghana Aids Commission, (2009).

With the previous prevalence rate of HIV/AIDS infections in the country standing at 3.6% as at 2004, from an initial rate of 2.9% some twenty four years ago, when the infection was first reported. Ghana had made remarkable success in combating the pandemic. Within two decades of strenuous combat against the menace, only 0.7% new infections had been recorded nationwide as at 2004 with 0.2% recorded in 2002. Statistics show that the prevalence rate had been very significant following constant improvement in the fight against the pandemic that had rocked much of the African work force over the past two decades. In 1986, Ghana recorded a prevalence rate of 2.9%. This rate dropped to 2.6% in 1999 and had an increment of 1.0% to settle at 3.6% in 2001. In the year 2004, the prevalence rate declined by 0.2% to make 3.4%, however, it shot up again in by 0.2%, increasing the prevalence rate to 3.6% currently, with the Northern region recording the lowest average rate of 2.1%, whiles the Eastern region recorded the highest average of 6.1%.

The Ghana 2003 HIV Sentinel Survey has shown a steady increase in HIV prevalence from 2.3% in 2000 to 3.6% in 2004, however, the prevalence rate in the 15-19 years age group experienced a decline. The youth, described as the high-risk age group, were targeted with seriousness and intensive programmes, leading to the declined of the mean prevalent rate to 1.9 per cent while the 45-49 years age group recorded a mean prevalence rate of 6.0 percent.. HIV/AIDS reported cases was increasing more among female than their male counterparts. Over 10,000 cases of female AIDS infection, between 20 to 24 years had been reported as compared to 6,000 male cases, between 30 to 39 years through 2003. Current AIDS estimated cases was about 200,000 with a daily record of 90 dead cases, 33,000 total dead cases and 132,000 orphans in 2004. It was also noted that almost

two-thirds of the AIDS-reported cases were female, an indication that the disproportionate rate of HIV infection among females had been existing for a while. Women tend to get infected at earlier ages than males for a variety of biological and sociocultural reasons.Ghana News Agency (2006)

These upward and downward adjustments in the HIV infectious rate do not allow us to determine the actual time and the number of people who get infected daily or annually. We may then tend to ask ourselves how come as certain age levels, the infection rate goes up and at certain age levels the infection rate declines.

Antiretroviral treatment can prolong the time between HIV infection and the onset of AIDS. Modern combination therapy is highly effective and someone with HIV who is taking treatment could live for the rest of their life without developing AIDS. An AIDS diagnosis does not necessarily equate to a death sentence. Many people can still benefit from starting antiretroviral therapy even once they have developed an AIDS defining illness. Treating some opportunistic infections is easier than others. Infections such as herpes zoster and candidiasis of the mouth, throat or vagina, can be managed effectively in most environments. On the other hand, more complex infections such as toxoplasmosis, need advanced medical equipment and infrastructure, which are lacking in many resource-poor areas. It is also important that treatment is provided for AIDS related pain, which is experienced by almost all people in the very advanced stages of HIV infection. Although there is no cure for AIDS, HIV infection can be prevented, and those living with HIV can take antiretroviral drugs to delay the onset of AIDS. However, in many countries across the world access to prevention and treatment services is limited.

Global leaders have pledged to work towards universal access to HIV prevention and care, so that millions of deaths can be averted.

End of life care becomes necessary when a person has reached the very final stages of AIDS. At this stage, preparing for death and open discussion about whether a person is going to die often helps in addressing concerns and ensuring final wishes are followed.

In recent years, the response to the epidemic has been intensified; in the past ten years in low- and middle-income countries there has been a 6-fold increase in spending for HIV and AIDS. The number of people on antiretroviral treatment has increased, the annual number of AIDS deaths has declined, and the global percentage of people infected with HIV has stabilised.

However, recent achievements should not lead to complacent attitudes. In all parts of the world, people living with HIV still face AIDS related stigma and discrimination, and many people still cannot access sufficient HIV treatment and care. In America and some countries of Western and Central and Eastern Europe, infection rates are rising, indicating that HIV prevention is just as important now as it ever has been. Prevention efforts that have proved to be effective need to be scaled-up and treatment targets reached. Commitments from national governments right down to the community level need to be intensified and subsequently met, so that one day the world might see an end to the global AIDS epidemic.

1.2 PROBLEM STATEMENT

Globally, the World Aids day was observed on the 1st December 2002 with route marches, concerts and seminars. However, the affected and infected individuals of

HIV/AIDS and their families marked the occasion in silence. Paradoxically, only a day is set aside yearly to provide the HIV/AIDS awareness crusades. Nonetheless, the virus/disease do not stop on 1st December each year because its ramifications stay with the victims, their families and the nations concerned in perpetuity because there is no cure for the disease at the moment.

Even though a lot of education has been made on the repercussions of HIV/AIDS on the socio-economic development of our region and a lot of research has gone on in the fight against the HIV menace, the disease continues to affect us.

In Ghana, antiretroviral drugs are given to HIV infected pregnant mothers so that the virus is not transmitted to the unborn child. Children are so given the antiretroviral drugs to prevent the HIV into developing into AIDS. Non-Governmental Organisations (NGO's) have and individuals have put on a lot of preventive strategies to curb this menace but yet the HIV/AIDS pandemic still prevails.

1.3 OBJECTIVES

Having found out that there is no cure for AIDS. There are drugs that can slow down the HIV virus, and slow down the damage to your immune system. There is no way to "clear" the HIV out of your body. The HIV virus does not discriminate among anyone and we are not always sure whether we are infected or not. The main objectives of this thesis is to

1. Model the number of occurrence of HIV(+) cases given age, time and gender.
2. Estimate the incidence rate of the HIV(+) cases with respect to age, time and gender.

3. Validate the models in (1) and (2).

1.4 JUSTIFICATION

Since its appearance in America in 1980, AIDS has shifted from a largely "unspeakable" and "untouchable" phenomenon--because of its first association with a gay male population already stigmatized in those terms--to a semi-acknowledged and still dismaying reality. Certain sectors of society now regularly "touch" and "speak" about AIDS, whereas the larger population remains generally ignorant of or indifferent to the disease and frustrated activists and practitioners see no cure in sight.

Epidemiological studies carried out in various settings to understand the role and complex relations of innumerable behavioral, social and demographic factors, which will help, interrupt and control the transmission of HIV/ AIDS have shown that the HIV virus is very difficult to curb. Therefore knowing the time and age of affected people will go a long way to help bring down the prevalence rate and also knowing the group of people who are more prone to the disease will help control it in a way before a cure is found.

1.5 SCOPE

The project will focus on the following:

- Research data and information will be gathered from Korle Bu Teaching Hospital.
- The study is based only on patients who voluntarily walk in to have the HIV test, those who were diagnosed and referred to be tested, those who were from the chest department, fever unit and other departments of the hospital who were also tested.

- Data collection and modelling of the data using the Poisson regression and the negative binomial which forms part of generalized linear models are used as methodology.

SUMMARY

This chapter introduces the project and discusses the problems to be solved in detail. The purpose of the chapter is to provide a deeper understanding of the project objectives and scope. The next chapter shall put forward literature review of HIV/AIDS.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Thirty years after AIDS was discovered in America, extensive research has been carried out in finding the cure to this menace. Activists and practitioners have also gone on to investigate the repercussions AIDS has on the society socially, economically and morally. A lot of research has also been carried out to find AIDS treatment and behavioral prevention. Soon after the emergence of the AIDS epidemic, it became evident that HIV was much more than just a disease. Unlike any other disease, HIV not only touches the lives of those infected, but it also impacts the lives of virtually everyone on earth. One would be hard pressed to find any group not affected by the HIV epidemic in some way. Simply put, it is clearly one of the most important public health issues.

Lewis et al. (1997) reviewed empirical studies dealing with the psychosocial correlates of HIV risk among heterosexual college students, including findings related to such theoretical variables as HIV/AIDS-related knowledge, personal and partner's attitudes toward condom use, perceived susceptibility, communication with sex partners, and sexual self-efficacy. Although college students are highly knowledgeable about basic HIV/AIDS facts, they retain some misperceptions about disease transmission. They hold neutral-to-negative hedonistic and practical attitudes about using condoms; those who have engaged in risky behavior accurately perceive their greater susceptibility to infection and experience anxiety regarding transmission of HIV infection. Heterosexual college students communicate infrequently with their partners about safer sex, but they often agree to a partner's suggestion that they use condoms. Higher levels of sexual self-efficacy among college students have been associated with a lower risk for HIV transmission. They discussed limitations and clinical implications of the findings and recommendations for future interventions.

Barnett et al. (1999) outlined accounts of some social and economic features of the HIV/AIDS epidemics in five countries: the United Kingdom, Botswana, Uganda, India

and Ukraine. It suggested that: (a) certain key features of society and economy are major determinants of the degree to which epidemics become generalised to whole populations; (b) these features can be conceptualised in ways that will assist in more effective targeting of preventive interventions and measures to confront the medium- and long-term impacts of raised morbidity and mortality associated with the occurrence of generalised HIV/AIDS epidemics.

Akhtet al. (2001) conducted a study on prison inmates in Sindh to determine whether HIV/AIDS related knowledge, attitudes and beliefs can predict their practices which risk HIV infection. They administered a questionnaire in this cross-sectional study to collect the data on HIV/AIDS related knowledge, attitudes, beliefs, practices and demographic variables in a systematic sample of 3,395 prison inmates during July 1994. The data on responses of inmates to HIV/AIDS related knowledge, attitudes, and beliefs were analyzed and a clear interpretable factor structure emerged for each set of questions labeled as knowledge, attitude and beliefs. Similarly based on responses of inmates to practice questions, three factors emerged and were labeled as heterosexuality, homosexuality and drugs. The standardized factor scores of inmates for each of these six factors were computed and used in further analyses. They used multiple linear regression analyses to identify if any of the independent variables (demographic variables, knowledge beliefs and attitude) predict these practice factors. The model for heterosexuality explained 23% of the variance and included HIV/AIDS related knowledge, beliefs, age, ethnicity and marital status and duration of imprisonment. The predictors in the model for homosexuality together explained 10% of the variance and included significant contribution by belief, marital status, ethnicity, education, age and duration of imprisonment. The model for drugs had significant contributions from HIV/AIDS related beliefs, marital status and ethnicity. Implications of prevention program based on these results were considered.

Whiteside et al. (2002) figured out that HIV/AIDS is the major threat to development, economic growth and poverty alleviation in much of Africa. And yet the full extent of the catastrophe facing the continent is only just being recognised, and still not by all. The international development targets set by the great and the good of the global community

or at least by those members of the community who attend the international summits that set these goals do not consider what HIV/AIDS means and are unachievable. He set the scene, describing the epidemic, explained why it is so important and what made HIV/AIDS different. He then explored how the poverty/epidemic cycle worked, whereby increased poverty increased the spread of HIV and AIDS increased poverty. He suggested we need to look beyond monetary poverty to understand these relationships. He also assessed what could and should be done to break the HIV/AIDS poverty cycle.

Gilbert et al. (2002) outlined aspects of the HIV/AIDS epidemic scenario and the complexities associated with it. They revealed the socio-epidemiological patterns of the epidemic and in doing so identified the populations with the greatest and fastest growing rates of infection. From the data they presented, it was evident that the pattern of HIV/AIDS in developing countries in sub-Saharan Africa in particular is unique. The pattern emerging in South Africa followed closely. The features of this pattern were as follows: the epidemic was mainly a heterosexual epidemic, the rates of infection in the general population were very high and the percentage of HIV-positive women was greater than men. An additional unique feature was the young age of onset of infection for women. Their data demonstrated the need to focus our attention on young African women and the factors underpinning their predicament. In order to understand their position they examined the long standing relationship between social inequalities and health in general and further invoked the concepts of vulnerability and social capital to shed light on the position of women in the epidemic. Within the constraints of limited and problematic statistical data, they argued that a mixture and complex interaction of material, social, cultural and behavioural factors shape the nature, process and outcome of the epidemic in South Africa.

Seidel et al. (2002) identified the competing discourses of HIV/AIDS circulating in sub-Saharan Africa. He said these were medical, medico-moral, developmental (distinguishing between 'women in development' and gender and development perspectives), legal, ethical, and the rights discourse of groups living with HIV/AIDS and of African pressure groups. The analytical framework was that of discourse analysis as exemplified by Michel Foucault. The medical and medico-moral are identified as

dominant. They shaped the perceptions of the pandemic, our responses to it, and to those living with HIV/AIDS. However, dissident activist voices are fracturing the dominant frameworks, and are mobilising a struggle for meaning around definitions of gender, rights, and development.

Parker et al (2002) reiterated that internationally, there had been a resurgence of interest in HIV and AIDS-related stigma and discrimination, triggered at least in part by growing recognition that negative social responses to the epidemic remain pervasive even in seriously affected communities. Yet, rarely were existing notions of stigma and discrimination interrogated for their conceptual adequacy and their usefulness in leading to the design of effective programmes and interventions. Taking as its starting point, the classic formulation of stigma as a 'significantly discrediting' attribute, but moving beyond this to conceptualize stigma and stigmatization as intimately linked to the reproduction of social difference, they offered a new framework to understand HIV and AIDS-related stigma and its effects. In so doing, they highlighted the manner in which stigma feeds upon, strengthens and reproduces existing inequalities of class, race, gender and sexuality. They highlighted the limitations of individualistic modes of stigma alleviation and calls instead for new programmatic approaches in which the resistance of stigmatized individuals and communities is utilized as a resource for social change.

Daborn et al. (2005) noted that the HIV/AIDS pandemic is associated with a number of opportunist mycobacterial infections, principally tuberculosis and disease due to the avian tubercle bacillus, Mycobacterium avium. Tuberculosis occurring early in the course of HIV infection is usually caused by M. tuberculosis. However some cases are due to the bovine tubercle bacillus, M. bovis, which, in turn, is transmissible from man to animals, principally by the aerogenous route although the majority of cases in man are non-pulmonary. These two mycobacterial species may be differentiated by means of a set of simple tests. The quality and quantity of information on the world-wide distribution and prevalence of bovine and human tuberculosis due to M. bovis is not uniform. There is a notable paucity of information from the tropics but available reports suggest that there are significant levels of bovine tuberculosis. If correct, this information has serious public health implications in the light of the current HIV/AIDS epidemic. Urgent investigation is

required so that appropriate control measures can be instituted where indicated and possible. The avian tubercle bacillus is a very common opportunistic pathogen in the late stage of AIDS but infection leading to disease is extremely rare in healthy, HIV-negative persons. Because of its widespread environmental distribution, infection by this pathogen cannot be prevented.

Peters et al. (2008) stated that Human immunodeficiency virus (HIV) was expanding rapidly to every region of the world. AIDS was the leading cause of death among people 15–59 years old and the world's most urgent public health challenge. Unprotected sexual contact was the predominant mode of HIV transmission throughout the world, although HIV could also be transmitted by exposure to infected blood and from mother to child. Sub-Saharan Africa has been disproportionately affected with 22.5 million people living with HIV (68% of the global infections) and 1.6 million AIDS deaths in 2007 (76%) of the AIDS deaths worldwide). In other regions, HIV infections had been concentrated in high-risk groups such as men who had sex with men, commercial sex workers and their clients, and injection-drug users. Many parts of the world had been slow to recognize the severity of the HIV pandemic but there had been notable prevention successes in Thailand and Uganda. In recent years, global efforts have increased substantially. The most encouraging improvements have been in sub-Saharan Africa where the number of people being treated with anti-retrovirals has increased tenfold from 2003 to 2006. A key challenge was to build on this success and increase access to evidence-based prevention programs, especially in low-income countries and marginalized high-risk groups.

Gould et al (2009) found out that although HIV was first known to science only in 1981, by 2006 HIV/AIDS had affected nearly 40 million people worldwide, 90% of whom live in developing countries and 63% of whom are in sub-Saharan Africa. The disease he was caused by the virus entering the body, most commonly during sexual intercourse, and destroying the immune system, and there is as yet no known cure. HIV/AIDS is an exceptional disease, in its demographic effects (largely affecting young sexually active adults, and not necessarily the poor), in its development impacts (causing major problems for rural and urban labor supply due to increased mortality and prolonged morbidity, as well as greatly increasing the direct costs of medical care), and in its management

(especially on the need for behavioral change to match the use of drug therapies and the search for a vaccine). The disease is 'beyond epidemiology', in that patterns of vulnerability to HIV/AIDS and its effects need to be explored more in cultural and behavioral terms, with explicit national political support, rather than in a narrowly biomedical realm.

Cluver et al. (2009) showed that AIDS-orphaned children are more likely to experience clinical-range psychological problems. Little is known about possible interactions between factors mediating these high distress levels. They assessed how food insecurity, bullying, and AIDS-related stigma interacted with each other and with likelihood of experiencing clinical-range disorder. In South Africa, 1025 adolescents completed standardised measures of depression, anxiety and post-traumatic stress. 52 potential mediators were measured, including AIDS-orphanhood status. Logistic regressions and hierarchical log-linear modelling were used to identify interactions among significant risk factors. Food insecurity, stigma and bullying all independently increased likelihood of disorder. Poverty and stigma were found to interact strongly, and with both present, likelihood of disorder rose from 19% to 83%. Similarly, bullying interacted with AIDS-orphanhood status, and with both present, likelihood of disorder rose from 12% to 76%. Approaches to alleviating psychological distress amongst AIDS-affected children addressed cumulative risk effects.

SUMMARY

This chapter provided intensively works people have done in the field of HIV/AIDS and its effect on the socio – economic development of our economy. Many practitioners and activists have dealt into a number of ways the HIV pandemic can be curbed. In the next chapter we shall put forward the relevant information concerning Generalised Linear Models (GLM's) and how it can be used to mode the Age - Time - Gender dependence on the number of occurrence of HIV positive cases.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

This chapter focuses on the detail and comprehensive understanding of the topic. The main contributions of this chapter are

i. The definition of Poisson distribution and the Poisson Regression model.
ii. The Seropositive HIV/AIDS cases.

In this chapter, examining and analysis on a few case studies involving several organizations from various researches are conducted and reported. Among the aspects that will come under scrutiny include the methodologies used in modelling the seropositive HIV/AIDS cases, the hardware and software specifications, and the features that are incorporated in the model.

3.2 DATA COLLECTION TECHNIQUE

This study is essentially on modelling the dependence on the occurrence of seropositive HIV cases using Korle Bu Teaching hospital as the case study.

Research data and information was gathered from Korle Bu Teaching Hospital.

The study is based only on patients who voluntarily walk in to have the HIV test, those who were diagnosed and referred to be tested, those who were from the chest department, fever unit and other departments of the hospital who were also tested. The data collected was done at the Korle Bu Teaching Hospital from January 2008 to August 2010.

3.3. DATA DESCRIPTION

The data has three main variables under study which are the year the data was collected, the gender that was tested (that is both males and females), and the ages of the genders that were tested. From the age groups we are able to determine that from those tested there were children, adolescents, adults and the aged. In all, there are 220 observations under the two genders for ten different age group ranges.

3.4. GENERALIZED LINEAR MODELS (GLM)

Generalized linear models (GLM) was first introduced by Nelder and Wedderburn (1972). They provided a unified framework to study various regression models, rather than a separate study for each individual regression. 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*,

$$\eta_i = \alpha + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} = x_i'\beta$$

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*: $g^{-1}(\eta_i) = \mu_i$.

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.

3.4.1. THE 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, $Y_i = X_i\beta + \varepsilon_i$ 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^2 , and

distributed normal. That is, the normal linear regression model is characterized by the following features:

1. **Stochastic component:** the Y_i are usually assumed to have independent normal distributions with $E(Y_i) = \mu_i$, with constant variance σ^2, or $Y_i \overset{iid}{\sim} N(\mu_i, \sigma^2)$

2. **Systematic component:** the covariates X_i combine linearly with the coefficients to form the linear predictor $\eta_i = X_i\beta$.

3. **Link between the random and systematic components:** the linear predictor $X_i\beta = \eta_i$ 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.

3.4.2. THE EXPONENTIAL FAMILY

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

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

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

where φ is a dispersion parameter and $a(\varphi)$, $b(\theta)$ and $c(y;\varphi)$ are known functions.

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,

$$E(Y_i) = \mu_i = b'(\theta)$$

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

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.

3.4.3. 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 $X_i\beta = \eta_i \in \mathbb{R}$ into the set of admissible values for μ_i. A log link is usually used for the Poisson model, since while $\eta_i = g(\mu_i) \in \mathbb{R}$, 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 $\eta_i \in \mathbb{R}$.

Example of link functions that are used are the identity, log, inverse, logit, probit, log - log, complementary log – log,etc. Note that binary data are handled in the GLM framework as special cases of binomial data.

3.5 THE POISSON DISTRIBUTION

The Poisson distribution (pronounced [pwasɔ̃]) (or Poisson law of small numbers) 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

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

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

$$\sigma_k^2 = E\left(k^2\right) - E\left(k\right)^2 .$$

Thus, the number of observed occurrences fluctuates about its mean λ with a standard deviation

$$\sigma_k = \sqrt{\lambda} .$$

As a function of *k*, this 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

$$\log f_i(y_i) = y_i \log(\mu_i) - \mu_i - \log(y_i!)$$

Looking at the coefficient of y_i we see immediately that the canonical parameter is

$$\theta_i = \log(\mu_i)$$

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

$$\mu_i = e^{\theta_i}$$

and we see that we can write the second term in the p.d.f. as

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

The last remaining term is a function of y_i only, so we identify

$$c(y_i, \varphi) = \log(y_i!)$$

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

$$\mu_i = b'(\theta_i) = e^{\theta_i} = \mu_i$$

And differentiating again we have

$$\mu_i = a_i(\varphi) b''(\theta_i) = e^{\theta_i} = \mu_i$$

Hence the mean is equal to the variance.

3.6 LOG – LINEAR MODELS

Suppose that we havea sample of n observations $y_1, y_2, ..., y_n$ 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

$$\mu_i = x_i'\beta$$

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

$$\eta_i = \log(\mu_i)$$

and assume that the transformed mean follows a linear model

$$\eta_i = x_i'\beta$$

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

$$\log(\mu_i) = x_i'\beta \qquad (4.2)$$

In this model in equation 4.2, the regression coefficients 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, when x_j is continuous.

Exponentiating Equation 4.2 we obtain a multiplicative model for the mean itself:

$$\mu_i = \exp(x_i'\beta)$$

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.

3.7 POISSON REGRESSION MODEL

Poisson regression analysis is a technique which allows to model 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,

$$\log(\mu_i) = x_i'\beta$$

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

$$\mu_i = \exp(x_i'\beta)$$

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

3.8 SPECIFICATION OF THE MODEL

The primary equation of the model is $\Pr(Y_i = y_i) = \frac{e^{\mu}\mu_i^{y_i}}{y_i}, y_i = 0,1,2,3...$

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

$$\log(\mu_i) = x_i'\beta$$

The expected number of events per period is given by

$$E(y_i/x_i) = \mu_i = e^{x_i'\beta}$$

Thus:

$$\frac{dE(y_i/x_i)}{dx_i} = \beta e^{x_i'\beta} = \beta_i\mu_i$$

The major assumption of Poisson model is

$$E(y_i/x_i) = \mu_i = e^{x_i'\beta} = V\text{ar}(y_i/x_i)$$

This assumption would be tested later on. If $Var(y_i/x_i) > E(y_i/x_i)$ then there is over-dispersion. If $Var(y_i/x_i) < E(y_i/x_i)$, 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

$$prob(y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}.$$

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

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

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

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

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

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

and

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

which is the Hessian of the function and with typical element

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

As $\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n}\left[e^{x_i'\beta} x_{ij} x_{il} \right]$ does not involve the y data

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

And the information matrix is

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

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 theHessian is negative definite for all x and β, the MLE ($\hat{\beta}$) is unique, if it exists. From $\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n}\left[e^{x_i'\beta} x_{ij} x_{il} \right]$ and $k_{jl} = E\left(\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} \right) = -\sum_{i=1}^{n}\left[e^{x_i'\beta} x_{ij} x_{il} \right]$:

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

And

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

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

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

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

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

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

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

3.10 FISHER SCORING IN LOG - LINEAR MODELS

Fisher scoring algorithm is a form of Newton-Raphson 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:

$$g(\eta_i) = \log(\mu_i)$$

The derivative of the link is easily seen to be

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

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

$$\beta^{new} = \beta + \left\{ E\left(-\frac{\partial L}{\partial \beta \partial \beta^T} \right) \right\}^{-1} \frac{\partial L}{\partial \beta} \qquad (3.1)$$

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 (3.1) can be rewritten as

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

where z is the n-vector with ith component

$$z_i = \left(Y_i - \mu_i \right) g'\left(\mu_i \right)$$

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

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

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

And this simplifies to

$$W_i = \mu_i$$

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

3.11 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 theassumption 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 $\text{var}\left(\hat{\beta}\right) = X'WX$, where X is the model matrix and W is the diagonal matrix of estimation weights.

3.11.1 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

$$\lambda_i = \forall \bar{y}$$

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

The corresponding log-likelihood is:

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

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 thecalculation of the maximized likelihood $L = \sum_{i=1}^{n}\left[-e^{x_i'\beta} + y_i x_i'\beta - \log y_i !\right]$ for the unrestricted model $\left(L_U\right)$, using $L = e^{x_i'\hat{\beta}_t}$. TheLikelihood Ratio test is then:

$$LR = 2\left(L_U - L_R\right)$$

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

3.11.2GOODNESS 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, G^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 G^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:

$$Deviance = 2\sum_{i=1}^{n} y_i \ln\left[\frac{y_i}{\hat{\mu}_i}\right] - \left(y_i - \hat{\mu}_i\right) \quad \text{.....(3.1)}$$

where in equation 3.1, y_i 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' the 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:

$$L = \sum_{i=1}^{n} y_i \ln\left(\hat{\mu}_i\right) - \hat{\mu}_i - \ln\left(y_i!\right) \quad \text{.....(3.2)}$$

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:

$$r_p = \frac{\left(y_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i}$$

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:

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

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

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

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

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

The standardized residual is:

$$r_s = \frac{y_i - \mu_i}{\sqrt{1-h_i}} \quad \text{.....(3.3)}$$

In equation 3.3, h is the leverage (diagonal of the Hat matrix).

3.12 OVERDISPERSION AND THE NEGATIVE BINOMIAL MODEL

The major assumption of the Poisson model is

$$E[y_i \mid x_i] = \lambda_i = e^{x_i'\beta} = Var[y_i \mid x_i]$$

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

This is very restrictive. If $E[y_i \mid x_i] < Var[y_i \mid x_i]$ then we speak about overdispersion, and when $E[y_i \mid x_i] > Var[y_i \mid x_i]$ we say we have underdispersion. The Poisson model does not allow for over or underdispersion. A richer model is obtained by using the negative binomial distribution instead of the Poisson distribution. Instead of

$$\Pr[Y_i = y_i] = \frac{e^{\lambda_i}\lambda_i^{y_i}}{y_i!},$$

we then use

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

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 ismodeled as

$$E[y_i \mid x_i] = \lambda_i = e^{x_i'\beta}$$

The conditional variance function is then given by

$$Var[y_i \mid x_i] = e^{x_i'\beta}\left(1 + \eta^2 e^{x_i'\beta}\right)$$

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.

3.13 AKAIKE'S INFORMATION CRITERION (AIC)

The Akaike's 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……(3.4)

where likelihood is the probability of the data given a model and K is the number of free parameters in the mode in equation 3.4. 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:

$$AICc = -2 (\ln (likelihood)) + 2 K * (n / (n - K - 1))\ldots\ldots\ldots\ldots\ldots\ldots(3.5)$$

In equation 3.5, 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)

3.14 SEROPOSITIVE

Serostatus is a word used to describe whether particular antibodies are present in the body. Seropositive refers to blood that shows traces of HIV antibodies (i.e., HIV-infected persons, but without symptoms)seronegative means they do not. If someone is seropositive for HIV (the Human Immunodeficiency Virus), it means their body has been producing antibodies for HIV, which can be detected with an HIV antibody test, the most common type of HIV test used.A seropositive test result means your blood tested positive for the human immunodeficiency virus (HIV) in a blood serum test. Seroconversion is the incubation period before you test positive for HIV antibodies in your blood. It can take between 4 to 12 weeks after you are infected with HIV. During this period, you may experience flu-like symptoms that include fever, rashes, night sweats and chills. When seroconversion is finished, there are HIV antibodies in your blood and you are considered seropositive.

3.15 STATISTICAL ANALYSIS SOFTWARE (SAS)

The Statistical Analysis Software (SAS) would be used for the analysis and interpretation of the data.

In SAS, several procedures in both STAT and ETS modules can be used to estimate Poisson regression. The procGENMOD was used for the analysis.GENMOD is easy to use with standard MODEL statement. Similar to Poisson regression, Negative binomial regression can be modeled by SAS directly with GENMOD.

CHAPTER 4

ANALYSIS

4.1INTRODUCTION

This chapter presents the analysis of the study which involves exploratory data analysis as well as further analysis using Poisson regression.

4.2. DATA COLLECTION

The data for this research was a secondary data which was originally obtained from the chest unit of the Korle – Bu Teaching Hospital from January 2008 to August 2010. In this research, the number of HIV positive (hiv_pos) cases recorded was used as the dependent variable and the other variables used as the explanatory variables were demographic data (Age_cate, Age_avge and Gender) and period of data collection (time, time quarterly and time_cont) were recorded.

4.3 NUMBER OF HIV POSITIVE CASES

Table 4.3 HIV positive cases recorded for various age categories for the period

AGE CATEGORIES	HIV POSITIVE CASES
0 – 9	18
10 – 14	9
15 – 19	10
20 – 24	60
25 – 29	181
30 – 34	221
35 – 39	228
40 – 44	204
45 – 49	120
50+	163
Total	**1214**

Source: Korle – Bu Teaching Hospital

The overall number of HIV positive cases recorded in the chest unit of the Korle – Bu Teaching Hospital from 2008 t0 2010 was 1,214. It can be seen in table 4.3 that, children between the ages of 10 and 14 years had the least number of HIV positive cases with adults between the ages of 35 and 39 years having the highest number of HIV positive cases. Only 9 children between the ages of 10 and 14 were infected with the HIV and that represent about 0.74% of the total number of people who were infected with the HIV. About 228 adults between the ages of 35 and 39 were HIV infected and that represented about 18.78% of the total HIV infected people. It can also be seen in table 4.3 that majority of the people who were HIV positive were between 25 years and over 50 years. Out of the total 1,214 HIV positive cases recoreded, 1,117 were people who were between that age(25 – 50+ years) and that represented 92% of the total HIV infected people. The result in the table 4.3 above is clearly shown figure 4.1.

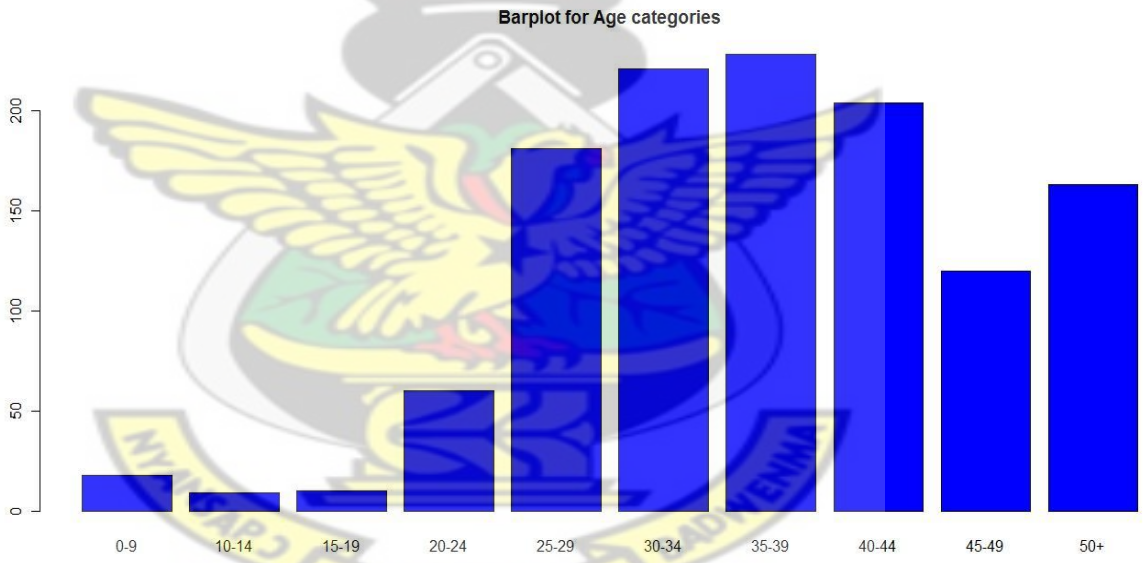

Figure 4.1, Bar chart showing the number of HIV cases recorded for various age categories

It can clearly be seen in the figure 4.1 above that, the age category that recorded the hishest HIV positive cases was 35 – 39 years and the age category that recorded the least

number of HIV positive cases was children between 10 – 14 years. The count for the distribution is negatively skewed since majority of the people who were infected with HIV from figure 4.1 above were between 25 years and over 50 years.

Table 4.4 HIV positive cases recorded from January 2008 to August 2010

QUARTERLY TIME	HIV POSITIVE CASES
TC 1	122
TC 2	171
TC 3	103
TC 4	120
TC 5	115
TC 6	104
TC 7	90
TC 8	111
T C 9	131
TC 10	85
TC 11	62
Total	**1214**

Source: Korle – Bu Teaching Hospital

In table 4.4 above, the HIV positive cases recorded in quarterly time in years was shown. TC 1, TC 2, TC 3 and TC 4 represent the first, second, third and fourth quarters of 2008 respectively. TC 5, TC 6, TC 7 and TC 8 represent the first, second, third and fourth quarters of 2009 respectively and TC 9, TC 10 and TC 11 also represent the first, second and third quarters of 2010. It can be seen in table 4.3 that the highest number of HIV positive cases was recorded in the second quarter of 2008 with a number of 171 and that represented about 14% of the total number of HIV positive cases recorded over the period. The lowest case of HIV positive case was recorded in the third quarter of 2010

where 62 HIV positive cases were recorded. The result in the table 4.5 above is clearly shown in figure 4.2.

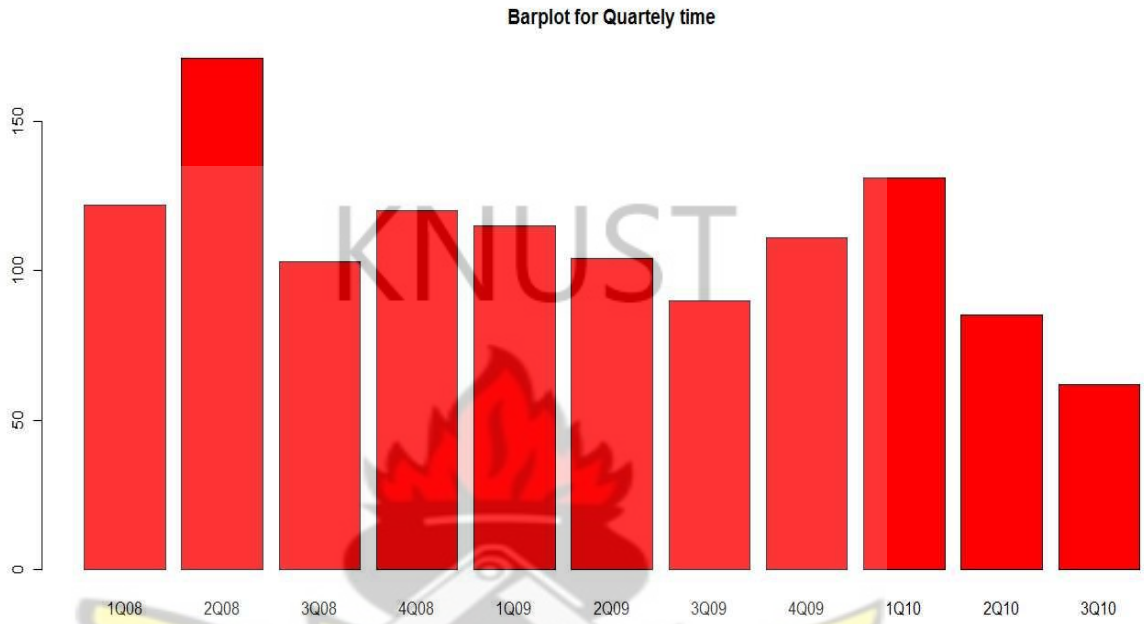

Figure 4.2, Bar Chart showing the number of HIV positive cases recorded for the period of January 2008 to August 2010

In the figure 4.2 above, it can be seen that in the fourth quarter of 2008, HIV positive cases were high but it started reducing in the first quarter of 2009 and continued to reduce to the third quarter of 2009 after which it started increasing again in the fourth quarter of that same year. Also when 2010 began, it could be seen in the figure 4.2 that a lot of HIV positive cases were recorded but it started to decline in the subsequent months of the same year. It can be seen that the third quarter of 2010 recorded the least number of HIV positive cases.

In table 4.5 in appendix A, it can be observed that the number of females who were infected with the HIV were more than that of males as per the years under study. Females in the ages between 30 years and 34 years were the most affected female group and males between the ages of 35 years and 39 years were the most affected male group. The result in the table 4.5 in appendix A is clearly shown in the chart below.

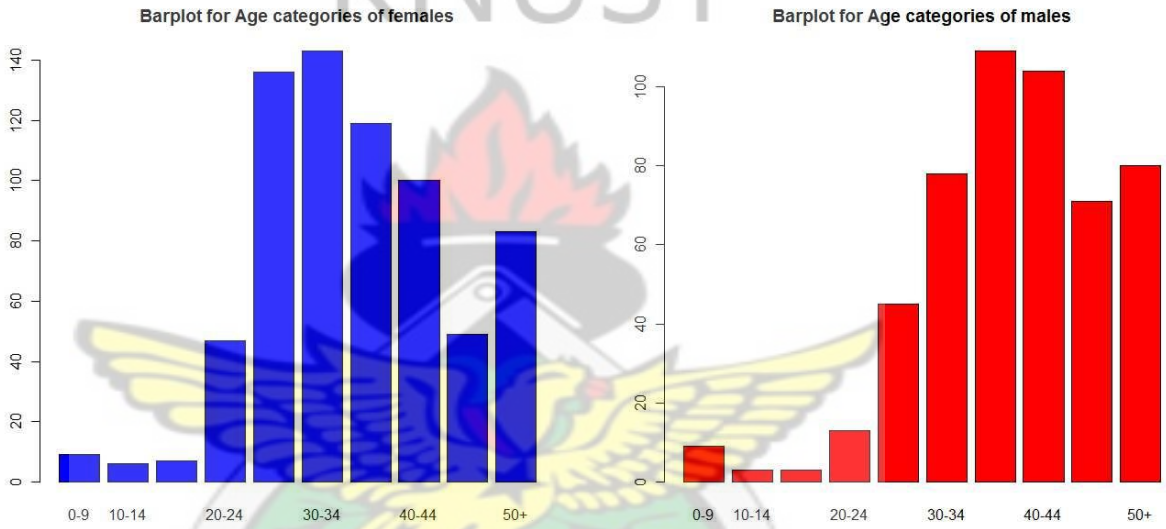

Figure 4.3, Bar charts showing HIV positive cases recorded for various age categories for both females and males

In figure 4.3, it can be seen that for all the age categories females recorded the highest number of HIV positive cases. The count of the distribution for both age categories of both males and females are negatively skewed.

4.4. MODELLING THE OCCURRENCE OF HIV POSITIVE CASES

In modelling the occurrence of the number of HIV positive cases, SAS statistical package version 9.1 was used. Using the PROC GENMOD procedure, the distribution specified

was Poisson with a Log link function. The total number of observations used was 220 and dependent variable used in the model was HIV positive. The following models were obtained in the table 4.5.

Table 4.6 Candidate models of the occurrence of HIV positive cases and their respective AICs

Model No.		AICs
1.	$\log(mean\_HIV^+) = -0.6362 - 0.3055G + 0.6769TC_1 + 1.0145TC_2 + 0.5076TC_3 + 0.3155TC_{10} - 0.6931AG_{12} - 0.5878AG_{17} + 1.2040AG_{22} + 2.3081AG_{27} + \ldots.. + 1.8971AG_{47} + 2.2034AG_{55}$	-2523.669
2.	$\log(mean\_HIV^+) = 1.8279 + 0.3308T_1 + 0.1250T_2 - 2.2034AC_1 - 2.8965AC_2 - 2.7912AC_3 - 0.9994AC_4 + 0.1047AC_5 + 0.3044AC_6 + 0.3356AC_7 + 0.2244AC_8 - 0.3063AC_9$	-2462.9068
3.	$\log(mean\_HIV^+) = -0.3650 + 0.3055G - 0.6931AG_{12} - 0.5878AG_{17} + 1.2040AG_{22} + 2.3081AG_{27} + 2.5078AG_{32} + 2.5390AG_{37} + 2.4277AG_{42} + 1.8971AG_{47} + 2.2034AG_{55}$	-2476.9002
4.	$\log(mean\_HIV^+) = 1.9691 - 0.3055G + 0.3308T_1 0.1250T_2 - 2.2034AC_1 - 2.8965AC_2 - 2.7912AC_3 - 0.9994AC_4 + 0.1047AC_5 + 0.3044AC_6 + 0.3356AC_7 + 0.2244AC_8 - 0.3063AC_9$	-2490.9026
5.	$\log(mean\_HIV^+) = 0.9761 - 0.3055G - 0.0499TC + 0.0346AG$	-2088.7402
6.	$\log(mean\_HIV^+) = 1.4261 + 0.6769TC_1 + 1.0145TC_2 + 0.5076TC_3 + \ldots.. + 0.3155TC_{10} - 2.8965AG_{12} - 2.7912AG_{17} - 0.9994AG_{22} + 0.1047AG_{27} + \ldots.. + 0.2244AG_{42} - 0.3063AG_{47}$	-2497.6734

NB: T_1, T_2 and T_3 means years 2008, 2009 and 2010 respectively TC_1, TC_2, TC_3 and TC_4 represent the first, second, third and fourth quarters of 2008 respectively. TC_5, TC_6, TC_7 and TC_8 represent the first, second, third and fourth quarters of 2009 respectively and TC_9, TC_{10} and TC_{11} also represent the first, second and third quarters of 2010. AC_1,AC_2, AC_3,……..AC_9means age categories 0 – 9, 10 – 14, 15 – 19,……..50+ respectively. $AG_{4.5}$, AG_{12}, AG_{17}……AG_{55}means average ages 4.5, 12,17 ……55 respectively and G means gender.

Table 4.6 shows the different models that were fitted as a result of the occurrence of HIV positive cases. In model one (1), after treating the continuous time and average ages as class variables , the log of the occurrence of HIV positive cases given gender, continuous time and average ages were fitted. The continuous time for the third quarter of 2010 (time_cont 11) was the reference time _cont and also the average age of 4.5 was also the reference average age by default. In model two (2), time (in years) and age categories were treated as class variables and the log of HIV positive occurrence given time (in years) and age categories were fitted. The reference variables were time (year 2010) and age category of 50+ .

Model three (3) showed a model of the log of the occurrence of HIV positive cases given gender and average age. This model was fitted after treating gender and average age as class variables. By default males were used as a reference variable and average age of 55 was also used as a reference variable. In model four (4) which fitted the log of the occurrence of HIV positive cases given gender, time (in years) and age categories, time (in years) and age categories were treated as class variables. By default, time (year 2010) and age category 50+ were taken as reference variables.

The variables, gender, continuous time (quarterly year) and average age were assumed to be continuous and the log of HIV positive occurrence was fitted in model five (5). The last model six (6) was fitted treating continuous time (quarterly year) and average age as class variables. By default, continuous time for the third quarter of the third year was a reference variable and average age 55 was also set as a reference variable.

The Akaike's information criterions (AICs) of the respective models were reported in table 4.5. The highest AIC was computed from model number 5 which had -2088.7402 where as the lowest AIC was computed from model number 1 which also had -2523.669

4.4.1. MODEL SELECTION

In selecting the best model in table 4.5 that provides an adequate fit to the data, the Akaike's information criterions (AICs) of each of the models in the table were computed. The model with the minimum AIC was selected as the best fit. From table 4.6, model number one (1) had the lowest AIC compared to the other candidate models and that it was selected as best model that fit the data. Hence the fitted model is given as

$$\begin{aligned}\log\left(mean\_HIV^{+}\right) = &-0.6362 - 0.3055G + 0.6769TC_1 + 1.0145TC_2 + 0.5076TC_3 + ...\\ &+ 0.3155TC_{10} - 0.6931AG_{12} - 0.5878AG_{17} + 1.2040AG_{22}\\ &+ 2.3081AG_{27} + ... + 1.8971AG_{47} + 2.2034AG_{55}\end{aligned} \qquad \textbf{(M. 4)}$$

Analysing the data in SAS (version 9.1) software showed that, the model (M.4) that fitted the data had deviance and Pearson chi – square (χ^2) distribution of 295.5438 and 286.5650 respectively. The deviance had an approximate chi – square distribution with 199 degrees of freedom. The ratio of the deviance and the degrees of freedom yielded a value of 1.4851 and the ratio of the Pearson chi – square and its degree of freedom yielded 1.4400. The scaled deviance and the scaled Pearson chi – square (χ^2) had the same approximate values as the deviance and the chi – square values of 295.5438 and 286.5650 respectively and they had the same degrees of freedom of 199. The log likelihood of the model was 1284.8345. A detailed analysis of the Criteria for Assessing Goodness of Fit was reported in table 4.4.1 in appendix A.

Table 4.7 Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr>ChiSq
Intercept	1	-0.6362	0.2673	-1.1602	-0.1122	5.66	0.0173
Gender	1	-0.3055	0.0581	-0.4193	-0.1917	27.67	<.0001
Time_cont 1	1	0.6769	0.1560	0.3712	0.9826	18.83	<.0001
Time_cont 2	1	1.0145	0.1482	0.7240	1.3051	46.83	<.0001
Time_cont 3	1	0.5076	0.1607	0.1925	0.8226	9.97	0.0016
Time_cont 4	1	0.6604	0.1564	0.3538	0.9669	17.83	<.0001
Time_cont 5	1	0.6178	0.1576	0.3090	0.9266	15.37	<.0001
Time_cont 6	1	0.5173	0.1605	0.2028	0.8317	10.39	0.0013
Time_cont 7	1	0.3727	0.1650	0.0492	0.6962	5.10	0.0239
Time_cont 8	1	0.5824	0.1585	0.2716	0.8931	13.49	0.0002
Time_cont 9	1	0.7481	0.1542	0.4459	1.0502	23.55	<.0001
Time_cont 10	1	0.3155	0.1670	-0.0118	0.6429	3.57	0.0589
Time_cont 11	0	0.0000	0.0000	0.0000	0.0000	..	
Age_avge 12	1	-0.6931	0.4082	-1.4933	0.1070	2.88	0.0895
Age_avge 17	1	-0.5878	0.3944	-1.3608	0.1852	2.22	0.1361
Age_avge 22	1	1.2040	0.2687	0.6772	1.7307	20.07	<.0001
Age_avge 27	1	2.3081	0.2471	1.8237	2.7925	87.22	<.0001
Age_avge 32	1	2.5078	0.2451	2.0274	2.9882	104.68	<.0001
Age_avge 37	1	2.5390	0.2448	2.0591	3.0188	107.54	<.0001
Age_avge 42	1	2.4277	0.2459	1.9458	2.9097	97.49	<.0001
Age_avge 47	1	1.8971	0.2528	1.4017	2.3925	56.33	<.0001
Age_avge 55	1	2.2034	0.2484	1.7166	2.6902	78.70	<.0001
Age_avge 4.5	0	0.0000	0.0000	0.0000	0.0000	. .	
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

A detailed analysis of the parameter estimates were reported in table 4.7. The first variable, gender was treated as a continuous variable with females forming part of the intercept. The second variable continuous time (cont_time) was treated as a class variable. SAS orders categories alphabetically or, if numeric, from the smallest to largest and takes the last one as a reference category by default. The value of the intercept estimate was -0.6362. Positive parameters for continuous times (cont_time 1, 2, 3, 4, ….., 10) indicated that the mean number of occurrence of HIV positive for those times were

greater than that of the mean of the continuous time (cont_time 11), the reference category.

It can also be seen in table 4.7 that the there are positive values of the parameter estimates for the average ages (age_avge 22, 27, 32, 37, 42, 47, 55) and that is an indication that the mean of the reference category average age (age_avge 4.5) was smaller than the mean of ages that gave positive parameter estimates. The negative values of the parameter estimates for the average ages (age_avge 12 and 17) indicated that the mean of the reference category (age_avge 4.5) is greater than the mean of the average ages that gave negative parameter estimates.

The table also presented values of the standard errors, confidence limits, the Wald chi-squares as well as the P –values of the individual parameter estimates. The results of the Wald chi – square test showed that there are statistically significant differences between continuous time (cont_time 11) and the continuous times 1, 2, 4, 4, 5, 6, 7, 8 and 9 but the statistical differences between continuous time (time_cont 11) and continuous times (time_cont 10) was not significant.

Also, the results of the Wald chi – square test showed that there were statistically significant differences between average age (age_avge 4.5) and the average ages 22, 27, 32, 37, 42, 47 and 45 but the differences between average age (age_avge 4.5) and the average ages (age_avge 12 and 17) was not statistically significant.

The Wald chi – square also showed that gender was statistically significant in the model with chi – square of 27.67 and a p – value less than 0.0001.

The negative parameter estimate value for the variable gender in model M.4 indicates that as the number of females increases, it will lead to an average decrease in the mean number of the occurrence of HIV positive cases for men.

4.5. MODELLING THE INCIDENCE OF HIV POSITIVE CASES.

In modelling the incidence of HIV positive cases, SAS statistical package version 9.1 was used. PROC GENMOD procedure in SAS System was used, where the distribution specified was Poisson with the Log as link function. The total number of observations used was 220 with HIV positive as dependent variable in the model. The offset variable used was the Log of the total number of people tested for the virus (i.e. total in the data set). The following models were obtained in the table 4.8.

Table 4.8 Candidates models of incidence of the number of HIV positive cases with their respective AICs

Model No.		AICs
1	$\log\left(\frac{mean\_HIV^{+}}{total}\right) = -1.3229 - 0.5190G - 0.0280TC - 0.2521AC_1 - 0.4133AC_2 - 1.2097AC_3 - 0.7199AC_4 + 0.2847AC_5 + 0.5528AC_6 + 0.6825AC_7 + 0.6422AC_8 + 0.5163AC_9$	-2606.0812
2	$\log\left(\frac{mean\_HIV^{+}}{total}\right) = -1.8572 + 0.1798T_1 + 0.0111T_2 - 0.1706AC_1 - 0.2686AC_2 - 1.1129AC_3 - 0.6093AC_4 + 0.3860AC_5 + 0.6031AC_6 + 0.7153AC_7 + 0.6808AC_8 + 0.5258AC_9$	-2523.5382
3	$\log\left(\frac{mean\_HIV^{+}}{total}\right) = -1.4191 - 0.4968G - 0.0333TC + 0.0110AG$	-2420.8032
4	$\log\left(\frac{mean\_HIV^{+}}{total}\right) = -2.0578 + 0.2138TC_1 + 0.3882TC_2 + 0.0007TC_3 + ... - 0.1507TC_{10} - 0.0918AG_{12} - 0.9116AG_{17} + ... + 0.7134AG_{47} + 0.1918AG_{55}$	-2529.973
5	$\log\left(\frac{mean\_HIV^{+}}{total}\right) = -2.2585 + 0.5220G - 0.1341AG_{12} - 0.9533AG_{17} - 0.4703AG_{22} + 0.5397AG_{27} + 0.8135AG_{32} + 0.9401AG_{37} + 0.9038AG_{42} + 0.7720AG_{47} + 0.2501AG_{55}$	-2597.239
6	$\log\left(\frac{mean\_HIV^{+}}{total}\right) = -1.8416 - 0.5163G + 0.2002TC_1 + 0.3673TC_2 + 0.0161TC_3 + ... - 0.1447TC_{10} - 0.1487AG_{12} - 0.9383AG_{17} - 0.4448AG_{22} + ... + 0.7916AG_{47} + 0.2741AG_{55}$	-2606.5093
7	$\log\left(\frac{mean\_HIV^{+}}{total}\right) = -1.5561 - 0.5199G + 0.1729T_1 + 0.0108T_2 - 0.2498AC_1 - 0.4024AC_2 - 1.2152AC_3 - 0.7233AC_4 + 0.2869AC_5 + 0.5537AC_6 + 0.6871AC_7 + 0.6420AC_8 + 0.5219AC_9$	-2601.3694

NB: T_1, T_2 and T_3 means years 2008, 2009 and 2010 respectively TC_1, TC_2, TC_3 and TC_4 represent the first, second, third and fourth quarters of 2008 respectively. TC_5, TC_6, TC_7 and TC_8 represent the first, second, third and fourth quarters of 2009 respectively and TC_9, TC_{10} and TC_{11} also represent the first, second and third quarters of 2010. AC_1,AC_2, AC_3,…….AC_9means age categories 0 – 9, 10 – 14, 15 – 19,……..50+ respectively. $AG_{4.5}$, AG_{12}, AG_{17}……AG_{55}means average ages 4.5, 12,17 ……55 respectively and G means gender.

The table 4.8 exhibited seven different models that were fitted as a result of modelling the log of incidence of HIV positive cases using the SAS (v 9.1) software.

In model one (1), a model of the log of incidence of HIV positive cases given gender, continuous time and age categories was fitted. Age categories were treated as a class variable and age_ cate 50+ was set as the reference age category by default. In model two (2), a model of the log of incidence of HIV positive cases given time and age categories was fitted. In this model, both time (in years) and age categories were treated as class variables. The reference variables were time (year 2010) and age category of 50+.

Model three (3) showed a model of the log of incidence of HIV positive cases given gender,continuous time (quarterly year) and average age. In this model all the variables were treated as continuous variables. Model four (4) also fitted the log of incidence of HIV positive cases treating continuous time (quarterly year) and average age as class variables. By default, continuous time for the third quarter of the third year was a reference variable and average age 55 was also set as a reference variable.

In model five (5), the log of incidence of HIV positive cases given gender and average ages was fitted. This model was fitted after treating gender and average age as class variables. By default males were used as a reference variable and average age of 55 was also used as a reference variable. The log of incidence of HIV positive cases given gender, continuous time and average ages was fitted in model six (6). The continuous time and average ages were treated as class variables.The continuous time for the third quarter of 2010 (time_cont11) was the reference continuous time and also the average age of 4.5 was also the reference average age by default.

The last model in table 4.8 showed the log of incidence of HIV positive cases gender, time (in years) and age categories, time (in years) and age categories were treated as class variables. By default, time (year 2010) and age category 50+ were set as reference variables.

The Akaike's Information Criterions (AICs) of the respective models were reported in the table 4.8. The model with the highest AIC was model three (3) with an AIC value of -2420.8032 and the model with the minimum AIC was model six (6) with an AIC value of -2606.5093.

4.5.1 MODEL SELECTION

In selecting the best model in table 4.8 that provides an adequate fit to the data, the Akaike's information criterions (AICs) of each of the models in the table were computed. The model with the minimum AIC was selected as the best fit. From table 4.8, model number six (6) had the lowest AIC compared to the other candidate models and that it was selected as best model that fit the data. Hence the fitted model is given as

$$\log\left(\frac{mean\_HIV^{+}}{total}\right) = -1.8416 - 0.5163G + 0.2002TC_1 + 0.3673TC_2 + 0.0161TC_3 + ... - 0.1447TC_{10} - 0.1487AG_{12} - 0.9383AG_{17} - 0.4448AG_{22} + ... + 0.7916AG_{47} + 0.2741AG_{55} \quad \textbf{(M. 5)}$$

Analysing the data in SAS (version 9.1) software showed the model (M.5) that fitted the data had deviance and Pearson chi – square (χ^2) distribution of 212.7031 and 209.3270 respectively. The deviance had an approximate chi – square distribution with 197 degrees of freedom. The ratio of the deviance and the degrees of freedom yielded a value of 1.0797 and the ratio of the Pearson chi – square and its degree of freedom yielded 1.0626.

The scaled deviance and the scaled Pearson chi – square (χ^2) had the same approximate values as the deviance and the chi – square values of 212.7031 and 209.3270 respectively and they had the same degrees of freedom of 197. The log likelihood of the model was 1326.2549. A detailed analysis of the Criteria for Assessing Goodness of Fit was reported in table 4.5.1 in appendix A.

Table 4.9 Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr>ChiSq
Intercept		1	-1.8416	0.2683	-2.3675	-1.3157	47.11	<.0001
Gender		1	-0.5163	0.0587	-0.6314	-0.4013	77.41	<.0001
Time_cont	1	1	0.2002	0.1561	-0.1058	0.5062	1.64	0.1998
Time_cont	2	1	0.3673	0.1485	0.0763	0.6583	6.12	0.0134
Time_cont	3	1	0.0161	0.1608	-0.2991	0.3314	0.01	0.9201
Time_cont	4	1	0.0909	0.1565	-0.2159	0.3976	0.34	0.5616
Time_cont	5	1	0.0802	0.1577	-0.2289	0.3893	0.26	0.6110
Time_cont	6	1	-0.0014	0.1606	-0.3163	0.3134	0.00	0.9929
Time_cont	7	1	-0.1877	0.1652	-0.5114	0.1360	1.29	0.2557
Time_cont	8	1	0.1870	0.1587	-0.1241	0.4980	1.39	0.2388
Time_cont	9	1	0.1324	0.1543	-0.1700	0.4348	0.74	0.3908
Time_cont	10	1	-0.1447	0.1671	-0.4723	0.1828	0.75	0.3865
Time_cont	11	0	0.0000	0.0000	0.0000	0.0000	. .	
Age_avge	12	1	-0.1487	0.4087	-0.9498	0.6524	0.13	0.7160
Age_avge	17	1	-0.9383	0.3946	-1.7118	-0.1649	5.65	0.0174
Age_avge	22	1	-0.4448	0.2691	-0.9722	0.0827	2.73	0.0984
Age_avge	27	1	0.5662	0.2476	0.0809	1.0515	5.23	0.0222
Age_avge	32	1	0.8235	0.2456	0.3420	1.3049	11.24	0.0008
Age_avge	37	1	0.9621	0.2454	0.4812	1.4431	15.38	<.0001
Age_avge	42	1	0.9122	0.2465	0.4292	1.3953	13.70	0.0002
Age_avge	47	1	0.7916	0.2534	0.2950	1.2882	9.76	0.0018
Age_avge	55	1	0.2741	0.2490	-0.2138	0.7621	1.21	0.2709
Age_avge	4.5	0	0.0000	0.0000	0.0000	0.0000	..	
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

A detailed analysis of the parameter estimates were reported in table 4.9. The first variable, gender was treated as a continuous variable with females forming part of the intercept. The second variable continuous time (cont_time) and the third variable average age (age_avge) were treated as class variables with the third quarter of the third year (cont_time 11) and average age 4.5 being their respective reference categories.. The value of the intercept estimate was -1.8416. Positive parameters for continuous times (cont_time 1, 2, 3, 4, 5, 8, 9) indicated that the mean number of incidence of HIV positive for those times were greater than that of the mean of the continuous time (cont_time 11), the reference category. The negative values of the parameters for continuous times (cont_times 6, 7, 10) indicated that mean of the reference category continuous time 11 was greater than the means of those times.

It can also be seen in table 4.9 that the there are positive values of the parameter estimates for the average ages (age_avge 27, 32, 37, 42, 47, 55) and that is an indication that the mean of the reference category average age 4.5 was less than the mean of ages that gave positive parameter estimates. The negative values of the parameter estimates for the average ages (age_avge 12, 17, 22) indicated that the mean of the reference category (age_avge 4.5) is greater than the mean of the average ages that gave negative parameter estimates.

The table also presented values of the standard errors, confidence limits, the Wald chi-squares as well as the P –values of the individual parameter estimates. The results of the Wald chi – square test showed that there was statistically significant difference between continuous time (cont_time 11) and the continuous time (time_cont 2). The chi-square value (χ^2) was given as 6.12 and the p – value was also less than 0.0001.

The statistical differences between continuous time (time_cont 11) and continuous times (time_cont 1, 3, 4, 5, 6, 7, 8 and 10) were not significant.

Also, the results of the Wald chi – square test showed that there were statistically significant differences between average age (age_avge 4.5) and the average ages 17, 27, 32, 37, 42 and 47 but the differences between average age (age_avge 4.5) and the average ages 12, 22, and 55 were not statistically significant.

The Wald chi – square also showed that gender was statistically significant in the model with chi – square of 77.41 and a p – value less than 0.0001.

The negative parameter estimate value for gender in model M.5 indicates that the mean number of incidence of HIV positive cases for females increases as that of males also decreases when all other variables in the model are held constant.

4.6. MODEL VALIDATION

If the model fits the data well, then the Poisson regression assumes that, the ratio of the deviance to the degree of freedom should be about one.

In model M.4 in section 4. 3. 1 and model M.5 in section 4. 3. 2, the deviance for model M.4 divided by its degrees of freedom was greater than one but the deviance divided by the degrees of freedom for model M.5 was approximately equal to one. The ratio of the deviance to the degree of freedom for model M.4 gave a value of 1. 4851 and the ratio of the deviance to the degree of freedom in model M.5 gave a value of 1. 0797. This was an indication that the assumption of equality of the variance to the mean has been violated in model M.4 but was satisfied in model M.5. This shows that the model M.4 in section 4. 3.

1 has been misspecified and therefore over dispersion existed in the model. We therefore run the negative binomial regression to validate the Poisson regression model as a corrective measure in validating the model.

4.6.1 VALIDATING THE OCCURRENCE OF HIV POSITIVE CASES USING THE NEGATIVE BINOMIAL REGRESSION.

Since the model, M.4 in section 4.3.1 does not fit the data well since there is over dispersion, the negative binomial regression is used to validate the model. We assume the model is the same as the model M.4 described in section 4.3.1, that is

$$\begin{aligned}\log\left(mean\_HIV^{+}\right) = &-0.6362 - 0.3055G + 0.6769TC_1 + 1.0145TC_2 + 0.5076TC_3 + ...\\ &+ 0.3155TC_{10} - 0.6931AG_{12} - 0.5878AG_{17} + 1.2040AG_{22}\\ &+ 2.3081AG_{27} + ... + 1.8971AG_{47} - 0.3063AG_{55}\end{aligned}$$

Instead of assuming as before that the distribution of the number of occurrence of HIV positive cases is Poisson, we will now assume that the number of occurrence of HIV positive cases has a negative binomial distribution. That means the assumption about the equality of the mean and the variance (Poisson distribution property) is relaxed, since the variance of the negative binomial is equal to $u + ku^2$, where $k > 0$ is a dispersion parameter. The maximum likelihood method is used to estimate k as well as the parameter of the regression model for $\log(u)$.

The SAS syntax for running the negative binomial regression is almost the same as that for the Poisson regression. The only change is that, the *dist* option in the model statement. Instead of dist = Poisson, dist = nb is used for the negative binomial regression. However

the Log of the total number of casualties (total) in the data set was used as an offset variable as specified in the model statement.

In analyzing the data in SAS (version 9.1), the table below contains information on assessing the Goodness of fit.

Table 4.10 Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	199	232.1318	1.1665
Scaled Deviance	199	232.1318	1.1665
Pearson Chi-Square	199	221.9631	1.1154
Scaled Pearson X2	199	221.9631	1.1154
Log Likelihood		1289.6700	

From table 4.10, the deviance was 232.1318 with 199 degrees of freedom. The deviance divided by the degree of freedom yielded a value of 1.1665 which is approximately equal to one. The Pearson chi - square value is 221.9631 with 199 degrees of freedom. The scaled deviance and the scaled Pearson chi – square had the same degrees of freedom values of 199 and had approximate values of the deviance and the Pearson chi – square with values 232.1318 and 22`.9631 respectively. The log likelihood was 1289.6700.

Since the deviance divided by the degrees of freedom gave a value of 1.1665 which is approximately equal to one, we say over dispersion has been taken care of and therefore the fitted model for the occurrence of HIV positive case was negative binomial regression model.Hence the fitted model is given as

$$\log\left(mean\_HIV^{+}\right) = -0.6424 - 0.3105G + 0.6640TC_1 + 1.0310TC_2 + 0.5132TC_3 + 0.6647TC_4 + ... + 0.3126TC_{10} - 0.6955AG_{12} - 0.5876AG_{17} + 1.1923AG_{22} + ... + 1.9134AG_{47} + 2.2121AG_{55}$$

(M.6)

In model M.6, the intercept estimate was -0.6424 which is the log of the expected number of occurrence of HIV positive cases when all other variables in the model are evaluated at zero.

The parameter estimates for quarterly time were all positive and this is an implication that, a unit increase in each of the quarterly times given, will lead to an average increase in the log of expected number of the occurrence of HIV positive cases when all other variables are held constant.

The positive estimates 1.1923, 2.3056, 2.4999, 2.5434, 2.4359, 1.9134 and 2.2121 for the average ages 22, 27, 32, 37, 42, 47 and 55 respectively, is an indication that there will be an average increase in the log of the expected number of occurrence of HIV positive cases when there is a unit increase in each of the average ages above when all other variables are held constant in the model.

The negative estimates -0.6955 and -0.5876 for the average ages 12 and 17 respectively, also shows that a unit increase in each of the average ages above when all other variables are held constant will lead to an average decrease in the log of the expected number of the occurrence of HIV positive cases.

The negative value for the estimate for gender which is -0.3105 is an indication that a unit increase in the number of females will lead to an average decrease in the log of the

expected number of the occurrence of HIV positive cases for male when all other variables are held constant.

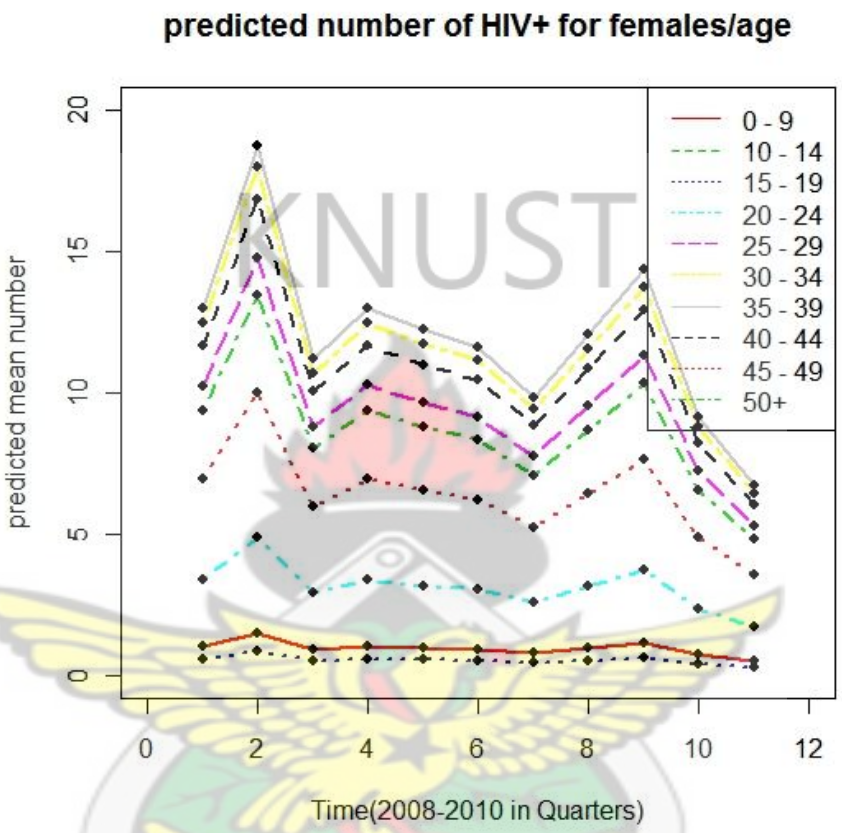

Figure 4.4 A graph showing the predicted mean number of HIV positive cases for females with their various age categories

From the figure 4.4 above, it can be observed that the predicted mean number of HIV positive cases for females was high in the second quarter of 2008 than the first quarter of 2008 for the various age categories. We also observe that, in the third quarter of 2008 there was a decrease in the predicted number of HIV positive cases for females in almost all the age categories. There was a slight increase in the predicted mean number of HIV positive cases of females for almost all the age categories in the fourth quarter of 2008.

Between the fourth quarter of 2008 and the third quarter of 2009 the predicted mean values of HIV positive cases for almost all the age categories decreased but mean number of HIV positive cases for females for almost all the age categories also decreased. Also but between the third quarter of 2009 and the first quarter of 2010 three was an increase in predicted mean values of HIV positive cases for females. It was also observed that between the first quarter of 2010 and the third quarter of 2010 there was a decline in the predicted mean values of HIV positive cases for all age categories of females. It was seen from the figure 4.4 that, females between the ages of 15 – 19 years recorded the least values of HIV positive cases while females between the ages of 35 – 39 years recorded the highest values of HIV positive cases.

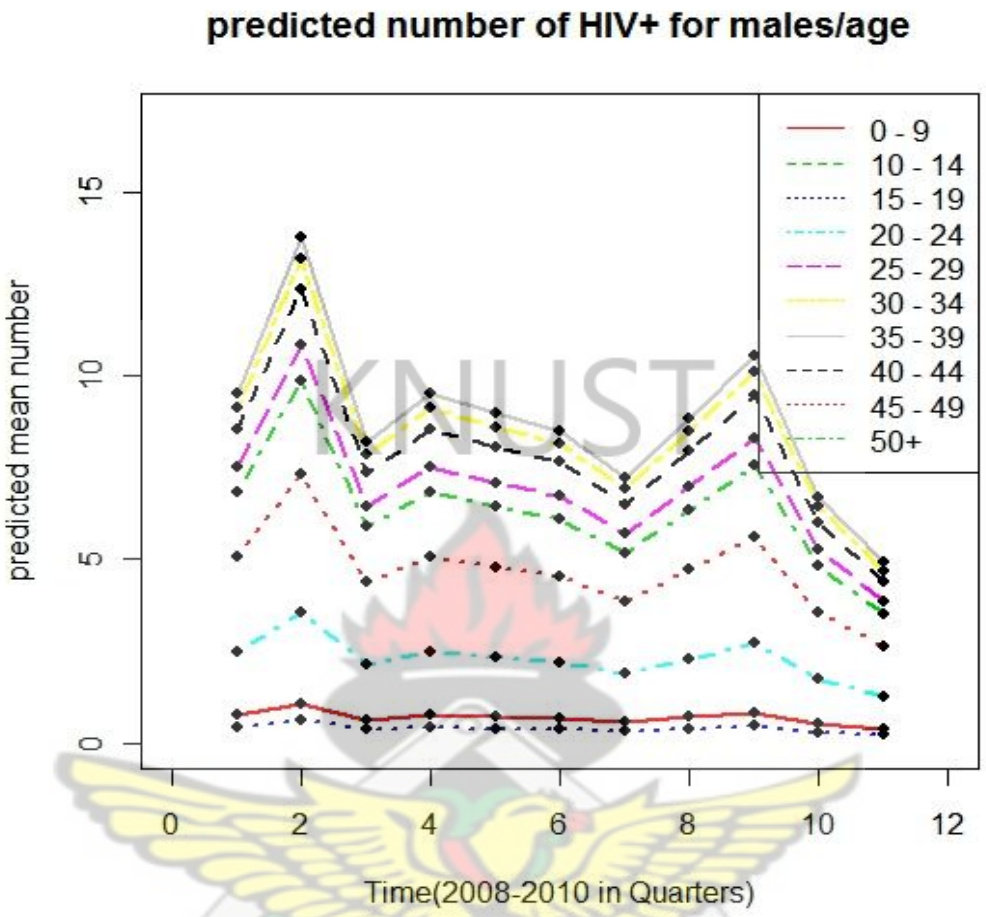

Figure 4.5 A graph showing the predicted mean number of HIV positive cases for males with their various age categories

In figure 4.5 above,we observed that there was an increase in the predicted mean number of HIV positive in males for the various age categories in the second quarter of 2008 as compared to that of the first quarter of 2008.In the third quarter of 2008, the predicted mean number of HIV positive cases in male decreased for all the age categories. Between the third quarter of 2008 and the third quarter of 2009, the predicted mean number of HIV positive cases decreased in men for all the age categories but increased in the fourth quarter of 2009 and the first quarter of 2010. The predicted mean number of HIV positive cases in male, decreased in the second and third quarter of 2010 for all the age categories.

We also observe from the figure 4.5 that, males between the ages of 15 – 20 years have the least predicted mean number of HIV positive cases while their counterparts between the ages of 35 – 39 years recorded the highest predicted mean number of HIV positive cases.

CHAPTER 5

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1INTRODUCTION

This chapter presents the summary of findings, conclusions drawn from the study, recommendation for future and discussion plan for future dissemination.

5.2 SUMMARY OF FINDINGS

In modelling the number of occurrence of HIV positive cases, the study revealed that

The intercept estimate -0.6424 represents the log of the occurrence of HIV positive cases for the third quarter of 2010 and children whose average ages are 4.5years which are the reference quarterly time and the reference average ages respectively with no other variable. Since {exp (-0.6424) = 0.53}, it can be deduced that in the third quarter of 2010, the average occurrence of HIV positive cases for children with average age of 4.5years is 53 out of every 100 children.

The negative parameter for gender indicated that, an average decrease in the mean number of the occurrence of HIV positive cases will increase the number of females. At any given quarterly time, given any number of people and for any particular age, we can say that the average number of occurrence of HIV positive cases for females would be 73 more than the number of males since exp(-0.3105) = 0.73.

The effects of quarterly time showed that the third quarter of 2010 recorded the least occurrence of HIV positive cases. Given any number of people and for a particular

average age given, it was seen that the average number of occurrence of HIV positive cases for the second quarter of 2008 reported {exp (1.0310) = 2.80} 2.80 times more than the number of occurrence of HIV positive cases in the third quarter of 2010. Also, given any number of people and for a particular average age given, the average number of occurrence of HIV positive cases for the second quarter of 2010 reported {exp (0.3126) = 1.37 times more than the occurrence of HIV positive cases in the third quarter of 2010.

The effect of average age showed that the average age 4.5 had the lowest occurrence of HIV positive cases. Given any number of people and at any fixed quarterly time, the average number of the occurrence of HIV positive cases for average age 37 {exp (2.5434 = 12.72} was 12.72 times more than that of the occurrence of HIV positive cases for average age 4.5. Also at any fixed quarterly time and given a specific number of people, the average number of occurrence of HIV positive cases for average age 12 {exp (-0.6955) = 0.50} was 0.5 times fewer than that of the occurrence of HIV positive cases on average age 4.5. The multiplicative effect on the mean number of occurrence of HIV positive cases for a unit increase in Average age 37 was exp (2.5434) = 12.72 and that of average age 12 was exp (-0.6955) = 0.5.

In modelling the incidence of HIV positive cases, the study revealed that,

The intercept estimate -1.8416 represents the log of the mean incidence of HIV positive cases for the third quarter of 2010 (reference quarterly time) and for average age 4.5years (reference average age) when no fixed number of people is given, no particular quarterly time is given and no average age is also given. Since exp (-1.8416) =0.16, it can be

deduced that the average incidence of HIV positive cases for children whose average age is 4.5 in the third quarter of 2010 is 16 out of every 100 children.

The effect of gender indicated that given any fixed quarterly time and any particular average age, a unit increase in the number of females will reduce the average incidence of HIV positive cases. That means the mean incidence of HIV positive cases for males will be 0.60 times fewer than that of females since exp (-0.5165) = 0.60.

The effect of quarterly time showed that the third quarter of 2010 had the lowest incidence of HIV positive cases. Given any fixed number of people and a particular average age, the second quarter of 2008 reported {exp (0.3673) = 1.44}, 1.44 times more incidence cases than that of the third quarter of 2010. Also, given any fixed number of people and a particular average age, the third quarter of 2009 {exp (-0.1877) = 0.83} had 17% fewer incidence of HIV positive cases than that of the third quarter of 2010. The multiplicative effect of the mean incidence of HIV positive cases for a unit increase in the second quarter of 2008 is exp (0.3673) = 1.44 and that of the third quarter of 2009 is exp (-0.1877) = 0.83.

The effect of average age showed that the average age 4.5years recorded the least incidence of HIV positive cases. Given any number of people and any fixed quarterly time, the mean incidence of HIV positive cases for average age 37years reported{exp (0.9621) = 2.62}, 2.62 times more incidence than children with average age 4.5years. Again given any number of people and at any fixed quarterly time children with average age 17years {exp (-0.9383)=0.39} had 61% fewer incidence of HIV positive cases than that of children with average age 4.5years. The multiplicative effect on the mean

incidence of HIV positive cases for a unit increase in adults with average age of 37 is exp(0.9621) = 2.62 and that of children with average age of 17years is exp (-0.9383) = 0.39.

5.3 CONCLUSIONS

Upon the findings of the research it can be concluded that

i) The number of occurrence of HIV positive cases for females is more than that of males.

ii) Adolescents between the ages of 15 – 19 years recorded the least occurrence of HIV positive cases.

iii) Adolescents and adults between the ages of 20 -44 years have a higher incidence rate of HIV positive cases.

5.4 RECOMMENDATIONS

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

i) Since most of the affected group (risk group) is people between the age groups 20 – 44 years, which are the working force of the country, government must intensify its campaign on the effect of HIV/AIDS. The economic gains of government aretend to increase of these people live and work for the betterment of our country.

ii) There should be interventions that focused on high – risk individuals, such as female sex workers by ensuring an increase in the use of condoms and encouraging the adoption of other safer sex practices.

iii) Mother – to – child transmission of HIV should be prevented by providing antiretroviral drugs to HIV infected pregnant mothers and children and if possible avoid infected mothers from breastfeeding their children.

iv) Government should strategize with industry to provide incentive and support for the role of business development to alleviate poverty.

v) Effective public education and sensitization should be made in schools, on the radio in the print media and all available platforms on the repercussions of HIV.

vi) Government should develop community involvement through training and appointment of community development workers who will engage with communities and determine health needs regarding HIV/AIDS.

REFERENCES

1. Akhtar S, Luby SP, Rahbar MH, Azam I. (2001) HIV/AIDS knowledge, attitudes and beliefs based prediction models for practices in prison inmates, Sindh, Pakistan. The Southeast Asian Journal of Tropical Medicine and Public Health. ; 32(2): 351-61
2. Annete J. D. (2002) An introduction to generalized linear models. Chapman & Hall/CRC, Boca Raton (Florida).
3. Barnett T. and Whiteside A. (1999) HIV/AIDS and Development: Case Studies and a Conceptual Framework, *The European Journal of Development Research* 11, 200–234;
4. Bozdogan H. (2007), Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions, PsychometrikaVolume 52, Number 3, 345-370,
5. Cameron A.C. and Trivedi P.K. (1998) Regression analysis of count data. Econometrics Society Monographs No. 30. Cambridge University Press. Cambridge (UK). Chapman and Hall.
6. CluverL. and Orkin M. (2009),Cumulative risk and AIDS-orphanhood: Interactions of stigma, bullying and poverty on child mental health in South Africa, *Social Science & Medicine,*Volume 69, Issue 8, Pages 1186-1193
7. Daborn C. J. and Grange J. M. (1993), HIV/AIDS and its implications for the control of animal tuberculosis, *British Veterinary Journal*, Volume 149, Issue 5, Pages 405-417
8. Famoye F, Wulu J. T. Jr.and Singh K. P. (2004), On the Generalized Poisson Regression Model with an Application to Accident Data, Journal of Data Science 2, 287-295
9. Giles E. D .and Feng H. (2009), Almost Unbiased Estimation of the Poisson Regression Model, Econometrics Working Paper EWP0909, ISSN 1485-6441
10. Ismail N and Jemain A. A., (2007)Handling Overdispersion with Negative Binomial andGeneralized Poisson Regression Models
11. Kim H. and Kriebel D. (2009), Regression models for public health surveillance data: a simulation study, *Occup Environ Med*2009 66: 733-739
12. Leah G. and Walker L. (2002), Treading the path of least resistance: HIV/AIDS and social inequalities—a South African case study, *Social Science & Medicine*, Volume 54, Issue 7, April, Pages 1093-1110
13. Lewis, J.E., Malow, R.M.& Ireland, S.J. (1997). HIV/AIDS risk in heterosexual college students. *J Am Coll Health*, 45(4):147-58.

14. Liu W. and Cela J. (2008), Count Data Models in SAS, Choice Point Precision Marketing, Alpharetta, GA
15. Loomis D., Richardson D. B. and Elliott L. (2005), Poisson regression analysis of ungrouped data, *Occup Environ* Med 62: 325-329.
16. McCullagh, P. and Nelder J. A. (1989), Generalized Linear Models. Second ed. London:
17. Nelder, J.A. and Wedderburn R.W.M. (1972), ‘‘Generalized linear models.’’ *Journal of the Royal Statistical Society*, Series A135:370--84.
18. Parker R. and Aggleton P. (2003), HIV and AIDS-related stigma and discrimination: a conceptual framework and implications for action, *Social Science & Medicine*,Volume 57, Issue 1, Pages 13-24
19. Ronald J. E, Boersma J., Blachman N. M. and Jagers A. A. (1988). "The Entropy of a Poisson Distribution: Problem 87-6". SIAM Review, Vol. XXVI, 30 (2):
20. Seidel G. (1993), The competing discourses of HIV/AIDS in sub-Saharan Africa: Discourses of rights and empowerment vs discourses of control and exclusion, *Social Science & Medicine*, Volume 36, Issue 3, Pages 175-194
21. Sithole F, Dohoo I, Leslie K, DesCôteaux L, Godden S, Campbell J, Stryhn H, Sanchez J.(2005), Effect of eprinomectin treatment at calving on milk production in dairy herds with limited outdoor exposure. *J Dairy Sci* ;88(3):929-37
22. Steenhard D., Nexis L. and Dayton, O.H. (2009), A Flexible Count Data Regression Model Using SAS PROC NLMIXED, Nan-Ting Chou, University of Louisville, Louisville, KY
23. Stefano P. and Ezio B. (2006), Poisson regression model in ePidemiology - an introduction, *Ann. Fac. Medic. Vet. di Parma* page 25 - page 44
24. Travier N., Gridley G., Blair A., Dosemeci M. and Boffetta P. (2003), Cancer Incidence among Male Swedish Veterinarians and Other Workers of the Veterinary Industry: *Cancer Causes & Control*, Vol. 14, No. 6, pp. 587-593
25. Whiteside A. (2002), Poverty and HIV/AIDS in Africa, Third World Quarterly, Vol. 23, No. 2, *Global Health and Governance*: HIV/AIDS, pp. 313-332

APPENDIX A

Table 4.4Cross tabulation of age categories and gender for HIV positive cases

AGE CATEGORIES	GENDER	
	Females	Males
0 – 9	9	9
10 – 14	6	3
15 – 19	7	3
20 – 24	47	13
25 – 29	136	45
30 – 34	143	78
35 – 39	119	109
40 – 44	100	104
45 – 49	49	71
50+	83	80
Total	**699**	**515**

Source – Korle – Bu Teaching Hospital

Table 4.5.1Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	199	295.5438	1.4851
Scaled Deviance	199	295.5438	1.4851
Pearson Chi-Square	199	286.5650	1.4400
Scaled Pearson X^2	199	286.5650	1.4400
Log Likelihood		1284.8345	

Table 4.7.1Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	197	212.7031	1.0797
Scaled Deviance	197	212.7031	1.0797
Pearson Chi-Square	197	209.3270	1.0626
Scaled Pearson X2	197	209.3270	1.0626
Log Likelihood		1326.2549	

Table 4.9.1 Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr>ChiSq
Intercept		1	-1.8416	0.2683	-2.3675	-1.3157	47.11	<.0001
Gender		1	-0.5163	0.0587	-0.6314	-0.4013	77.41	<.0001
Time_cont	1	1	0.2002	0.1561	-0.1058	0.5062	1.64	0.1998
Time_cont	2	1	0.3673	0.1485	0.0763	0.6583	6.12	0.0134
Time_cont	3	1	0.0161	0.1608	-0.2991	0.3314	0.01	0.9201
Time_cont	4	1	0.0909	0.1565	-0.2159	0.3976	0.34	0.5616
Time_cont	5	1	0.0802	0.1577	-0.2289	0.3893	0.26	0.6110
Time_cont	6	1	-0.0014	0.1606	-0.3163	0.3134	0.00	0.9929
Time_cont	7	1	-0.1877	0.1652	-0.5114	0.1360	1.29	0.2557
Time_cont	8	1	0.1870	0.1587	-0.1241	0.4980	1.39	0.2388
Time_cont	9	1	0.1324	0.1543	-0.1700	0.4348	0.74	0.3908
Time_cont	10	1	-0.1447	0.1671	-0.4723	0.1828	0.75	0.3865
Time_cont	11	0	0.0000	0.0000	0.0000	0.0000	.	.
Age_avge	12	1	-0.1487	0.4087	-0.9498	0.6524	0.13	0.7160
Age_avge	17	1	-0.9383	0.3946	-1.7118	-0.1649	5.65	0.0174
Age_avge	22	1	-0.4448	0.2691	-0.9722	0.0827	2.73	0.0984
Age_avge	27	1	0.5662	0.2476	0.0809	1.0515	5.23	0.0222
Age_avge	32	1	0.8235	0.2456	0.3420	1.3049	11.24	0.0008
Age_avge	37	1	0.9621	0.2454	0.4812	1.4431	15.38	<.0001
Age_avge	42	1	0.9122	0.2465	0.4292	1.3953	13.70	0.0002
Age_avge	47	1	0.7916	0.2534	0.2950	1.2882	9.76	0.0018
Age_avge	55	1	0.2741	0.2490	-0.2138	0.7621	1.21	0.2709
Age_avge	4.5	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

APPENDIX B

Secondary data obtained from the chest unit of the Korle – Bu Teaching Hospital with the predicted mean number of HIV positive cases

Gender	Age_cate	Times_quarterly	Time	Time_cont	Age_avge	HIV_pos	HIV_neg	Wexpos	Pred
1	0 - 9	2008	1	1	4.5	0	7	7	0.7491
0	0 - 9	2008	1	1	4.5	0	2	2	1.0218
1	10 – 14	2008	1	1	12	0	4	4	0.3736
0	10 – 14	2008	1	1	12	2	5	7	0.5097
1	15 - 19	2008	1	1	17	0	5	5	0.4162
0	15 - 19	2008	1	1	17	1	6	7	0.5678
1	20 - 24	2008	1	1	22	0	26	26	2.4679
0	20 - 24	2008	1	1	22	2	20	22	3.3665
1	25 - 29	2008	1	1	27	3	29	32	7.5132
0	25 - 29	2008	1	1	27	15	26	41	10.2491
1	30 - 34	2008	1	1	32	6	21	27	9.1247
0	30 - 34	2008	1	1	32	18	19	37	12.4474
1	35 - 39	2008	1	1	37	9	24	33	9.5304
0	35 - 39	2008	1	1	37	13	15	28	13.0008
1	40 - 44	2008	1	1	42	11	21	32	8.5591
0	40 - 44	2008	1	1	42	13	13	26	11.6758
1	45 - 49	2008	1	1	47	6	15	21	5.0757
0	45 - 49	2008	1	1	47	8	4	12	6.924
1	50+	2008	1	1	55	7	38	45	6.8423

0	50+	2008	1	1	55	8	18	26	9.3339
1	0 - 9	2008	1	2	4.5	0	6	6	1.0812
0	0 - 9	2008	1	2	4.5	2	5	7	1.4749
1	10 – 14	2008	1	2	12	0	2	2	0.5393
0	10 - 14	2008	1	2	12	2	3	5	0.7357
1	15 - 19	2008	1	2	17	1	5	6	0.6008
0	15 - 19	2008	1	2	17	1	12	13	0.8195
1	20 - 24	2008	1	2	22	2	27	29	3.5621
0	20 - 24	2008	1	2	22	9	23	32	4.8593
1	25 - 29	2008	1	2	27	2	21	23	10.8446
0	25 - 29	2008	1	2	27	12	27	39	14.7936
1	30 - 34	2008	1	2	32	13	31	44	13.1707
0	30 - 34	2008	1	2	32	20	22	42	17.9667
1	35 - 39	2008	1	2	37	17	16	33	13.7562
0	35 - 39	2008	1	2	37	12	16	28	18.7654
1	40 - 44	2008	1	2	42	17	23	40	12.3542
0	40 - 44	2008	1	2	42	18	19	37	16.8529
1	45 - 49	2008	1	2	47	6	19	25	7.3263
0	45 - 49	2008	1	2	47	11	10	21	9.9941
1	50+	2008	1	2	55	15	43	58	9.8763
0	50+	2008	1	2	55	11	22	33	13.4726
1	0 - 9	2008	1	3	4.5	0	8	8	0.6442
0	0 - 9	2008	1	3	4.5	1	6	7	0.8788
1	10 – 14	2008	1	3	12	0	2	2	0.3213

0	10 – 14	2008	1	3	12	0	2	2	0.4383
1	15 - 19	2008	1	3	17	1	16	17	0.3579
0	15 - 19	2008	1	3	17	2	9	11	0.4883
1	20 - 24	2008	1	3	22	1	23	24	2.1223
0	20 - 24	2008	1	3	22	2	42	44	2.8952
1	25 - 29	2008	1	3	27	6	32	38	6.4613
0	25 - 29	2008	1	3	27	7	33	40	8.8141
1	30 - 34	2008	1	3	32	13	33	46	7.8472
0	30 - 34	2008	1	3	32	10	16	26	10.7047
1	35 - 39	2008	1	3	37	7	22	29	8.196
0	35 - 39	2008	1	3	37	12	11	23	11.1805
1	40 - 44	2008	1	3	42	10	23	33	7.3607
0	40 - 44	2008	1	3	42	7	15	22	10.0411
1	45 - 49	2008	1	3	47	6	15	21	4.3651
0	45 - 49	2008	1	3	47	4	8	12	5.9546
1	50+	2008	1	3	55	3	53	56	5.8843
0	50+	2008	1	3	55	11	19	30	8.0271
1	0 - 9	2008	1	4	4.5	0	5	5	0.7496
0	0 - 9	2008	1	4	4.5	0	5	5	1.0226
1	10 – 14	2008	1	4	12	0	2	2	0.3739
0	10 – 14	2008	1	4	12	0	7	7	0.5101
1	15 - 19	2008	1	4	17	1	7	8	0.4165
0	15 - 19	2008	1	4	17	0	11	11	0.5682
1	20 - 24	2008	1	4	22	3	30	33	2.4696

0	20 - 24	2008	1	4	22	8	33	41	3.3689
1	25 - 29	2008	1	4	27	4	16	20	7.5186
0	25 - 29	2008	1	4	27	9	27	36	10.2564
1	30 - 34	2008	1	4	32	4	35	39	9.1312
0	30 - 34	2008	1	4	32	15	17	32	12.4563
1	35 - 39	2008	1	4	37	15	33	48	9.5371
0	35 - 39	2008	1	4	37	13	11	24	13.01
1	40 - 44	2008	1	4	42	11	21	32	8.5651
0	40 - 44	2008	1	4	42	10	23	33	11.6841
1	45 - 49	2008	1	4	47	6	21	27	5.0793
0	45 - 49	2008	1	4	47	3	7	10	6.9289
1	50+	2008	1	4	55	8	55	63	6.8472
0	50+	2008	1	4	55	10	25	35	9.3406
1	0 - 9	2009	2	5	4.5	2	3	5	0.7049
0	0 - 9	2009	2	5	4.5	1	3	4	0.9615
1	10 – 14	2009	2	5	12	0	2	2	0.3516
0	10 – 14	2009	2	5	12	1	5	6	0.4796
1	15 - 19	2009	2	5	17	0	8	8	0.3917
0	15 - 19	2009	2	5	17	0	5	5	0.5343
1	20 - 24	2009	2	5	22	2	28	30	2.3222
0	20 - 24	2009	2	5	22	1	29	30	3.1679
1	25 - 29	2009	2	5	27	3	24	27	7.0699
0	25 - 29	2009	2	5	27	8	29	37	9.6443
1	30 - 34	2009	2	5	32	7	27	34	8.5863

0	30 - 34	2009	2	5	32	17	8	25	11.7129
1	35 - 39	2009	2	5	37	18	27	45	8.968
0	35 - 39	2009	2	5	37	17	18	35	12.2336
1	40 - 44	2009	2	5	42	7	22	29	8.054
0	40 - 44	2009	2	5	42	12	12	24	10.9868
1	45 - 49	2009	2	5	47	5	25	30	4.7762
0	45 - 49	2009	2	5	47	2	10	12	6.5154
1	50+	2009	2	5	55	6	47	53	6.4386
0	50+	2009	2	5	55	6	34	40	8.7831
1	0 - 9	2009	2	6	4.5	2	3	5	0.6691
0	0 - 9	2009	2	6	4.5	2	10	12	0.9127
1	10 - 14	2009	2	6	12	1	5	6	0.3338
0	10 – 14	2009	2	6	12	0	4	4	0.4553
1	15 - 19	2009	2	6	17	0	7	7	0.3718
0	15 - 19	2009	2	6	17	1	11	12	0.5072
1	20 - 24	2009	2	6	22	2	20	22	2.2044
0	20 - 24	2009	2	6	22	4	32	36	3.0072
1	25 - 29	2009	2	6	27	3	33	36	6.7112
0	25 - 29	2009	2	6	27	10	23	33	9.155
1	30 - 34	2009	2	6	32	7	28	35	8.1507
0	30 - 34	2009	2	6	32	12	14	26	11.1187
1	35 - 39	2009	2	6	37	9	32	41	8.513
0	35 - 39	2009	2	6	37	9	18	27	11.6129
1	40 - 44	2009	2	6	42	11	27	38	7.6454

0	40 - 44	2009	2	6	42	3	14	17	10.4294
1	45 - 49	2009	2	6	47	13	10	23	4.5339
0	45 - 49	2009	2	6	47	4	21	25	6.1849
1	50+	2009	2	6	55	5	53	58	6.1119
0	50+	2009	2	6	55	6	17	23	8.3375
1	0 - 9	2009	2	7	4.5	1	4	5	0.5675
0	0 - 9	2009	2	7	4.5	0	3	3	0.7742
1	10 - 14	2009	2	7	12	0	2	2	0.2831
0	10 -14	2009	2	7	12	0	2	2	0.3862
1	15 - 19	2009	2	7	17	0	5	5	0.3153
0	15 - 19	2009	2	7	17	0	4	4	0.4302
1	20 - 24	2009	2	7	22	1	18	19	1.8698
0	20 - 24	2009	2	7	22	6	22	28	2.5506
1	25 - 29	2009	2	7	27	0	36	36	5.6924
0	25 - 29	2009	2	7	27	15	32	47	7.7652
1	30 - 34	2009	2	7	32	4	30	34	6.9133
0	30 - 34	2009	2	7	32	6	26	32	9.4308
1	35 - 39	2009	2	7	37	4	32	36	7.2207
0	35 - 39	2009	2	7	37	6	21	27	9.85
1	40 - 44	2009	2	7	42	10	22	32	6.4848
0	40 - 44	2009	2	7	42	8	13	21	8.8461
1	45 - 49	2009	2	7	47	6	20	26	3.8456
0	45 - 49	2009	2	7	47	3	18	21	5.2459
1	50+	2009	2	7	55	10	40	50	5.1841

0	50+	2009	2	7	55	10	37	47	7.0718
1	0 - 9	2009	2	8	4.5	0	5	5	0.6963
0	0 - 9	2009	2	8	4.5	1	4	5	0.9499
1	10 - 14	2009	2	8	12	1	0	1	0.3473
0	10 - 14	2009	2	8	12	0	5	5	0.4738
1	15 - 19	2009	2	8	17	0	3	3	0.3869
0	15 - 19	2009	2	8	17	0	7	7	0.5278
1	20 - 24	2009	2	8	22	0	25	25	2.2942
0	20 - 24	2009	2	8	22	4	23	27	3.1295
1	25 - 29	2009	2	8	27	14	22	36	6.9843
0	25 - 29	2009	2	8	27	20	11	31	9.5276
1	30 - 34	2009	2	8	32	4	33	37	8.4824
0	30 - 34	2009	2	8	32	13	18	31	11.5712
1	35 - 39	2009	2	8	37	5	21	26	8.8595
0	35 - 39	2009	2	8	37	8	9	17	12.0856
1	40 - 44	2009	2	8	42	4	17	21	7.9566
0	40 - 44	2009	2	8	42	8	15	23	10.8539
1	45 - 49	2009	2	8	47	11	21	32	4.7184
0	45 - 49	2009	2	8	47	4	7	11	6.4366
1	50+	2009	2	8	55	6	51	57	6.3607
0	50+	2009	2	8	55	8	22	30	8.6769
1	0 - 9	2010	3	9	4.5	4	8	12	0.8282
0	0 - 9	2010	3	9	4.5	2	14	16	1.1298
1	14-Oct	2010	3	9	12	1	1	2	0.4131

0	14-Oct	2010	3	9	12	1	3	4	0.5636
1	15 - 19	2010	3	9	17	0	8	8	0.4602
0	15 - 19	2010	3	9	17	1	12	13	0.6278
1	20 - 24	2010	3	9	22	2	21	23	2.7287
0	20 - 24	2010	3	9	22	6	49	55	3.7224
1	25 - 29	2010	3	9	27	6	30	36	8.3073
0	25 - 29	2010	3	9	27	17	23	40	11.3324
1	30 - 34	2010	3	9	32	7	27	34	10.0892
0	30 - 34	2010	3	9	32	15	18	33	13.7631
1	35 - 39	2010	3	9	37	14	28	42	10.5377
0	35 - 39	2010	3	9	37	16	12	28	14.3749
1	40 - 44	2010	3	9	42	9	22	31	9.4637
0	40 - 44	2010	3	9	42	10	13	23	12.9098
1	45 - 49	2010	3	9	47	4	20	24	5.6122
0	45 - 49	2010	3	9	47	3	11	14	7.6558
1	50+	2010	3	9	55	6	73	79	7.5655
0	50+	2010	3	9	55	7	30	37	10.3205
1	0 - 9	2010	3	10	4.5	0	1	1	0.5271
0	0 - 9	2010	3	10	4.5	0	2	2	0.7191
1	10 - 14	2010	3	10	12	0	3	3	0.2629
0	10 – 14	2010	3	10	12	0	0	0	0.3587
1	15 - 19	2010	3	10	17	0	6	6	0.2929
0	15 - 19	2010	3	10	17	0	9	9	0.3995
1	20 - 24	2010	3	10	22	0	25	25	1.7366

0	20 - 24	2010	3	10	22	3	43	46	2.369
1	25 - 29	2010	3	10	27	2	22	24	5.287
0	25 - 29	2010	3	10	27	14	20	34	7.2122
1	30 - 34	2010	3	10	32	8	31	39	6.421
0	30 - 34	2010	3	10	32	12	18	30	8.7592
1	35 - 39	2010	3	10	37	6	29	35	6.7064
0	35 - 39	2010	3	10	37	6	17	23	9.1485
1	40 - 44	2010	3	10	42	8	30	38	6.0229
0	40 - 44	2010	3	10	42	4	17	21	8.2161
1	45 - 49	2010	3	10	47	4	17	21	3.5717
0	45 - 49	2010	3	10	47	5	8	13	4.8724
1	50+	2010	3	10	55	9	44	53	4.8149
0	50+	2010	3	10	55	4	29	33	6.5682
1	0 - 9	2010	3	11	4.5	0	2	2	0.3856
0	0 - 9	2010	3	11	4.5	0	3	3	0.526
1	10 - 14	2010	3	11	12	0	0	0	0.1923
0	10 - 14	2010	3	11	12	0	1	1	0.2624
1	15 - 19	2010	3	11	17	0	5	5	0.2143
0	15 - 19	2010	3	11	17	1	8	9	0.2923
1	20 - 24	2010	3	11	22	0	14	14	1.2704
0	20 - 24	2010	3	11	22	2	22	24	1.733
1	25 - 29	2010	3	11	27	2	17	19	3.8677
0	25 - 29	2010	3	11	27	9	18	27	5.276
1	30 - 34	2010	3	11	32	5	14	19	4.6972

0	30 - 34	2010	3	11	32	5	7	12	6.4077
1	35 - 39	2010	3	11	37	5	14	19	4.906
0	35 - 39	2010	3	11	37	7	9	16	6.6925
1	40 - 44	2010	3	11	42	6	12	18	4.406
0	40 - 44	2010	3	11	42	7	11	18	6.0105
1	45 - 49	2010	3	11	47	4	12	16	2.6129
0	45 - 49	2010	3	11	47	2	4	6	3.5643
1	50+	2010	3	11	55	5	41	46	3.5223
0	50+	2010	3	11	55	2	18	20	4.8049