ANALYSIS OF THE EFFECTS OF OVERDISPERSION IN POPULATION DYNAMICS.

By

Seth Owusu

Registration Number: (MS300-0006/17)

A Research Thesis submitted to Pan African University Institute of Science, Technology and Innovation in partial fulfillment of the requirement for the award of the degree of Master of Science in (Mathematics) (Statistics option) of the Pan African University.

2018

Declaration

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

Seth Owusu:	....	....
	Signature	Date

This thesis has been submitted with our approval as University supervisors.

Dr. Jane Akinyi	....	....
(Supervisor)	Signature	Date
Prof. George Orwa	....	....
(Supervisor)	Signature	Date

Dedication

I dedicate this work to the Almighty God for his abundant grace, guidance and protection; to my Dad, Mr. Kwaku Duah, who gave up the ghost on the eve of submitting this thesis. May his soul rest in perfect peace.

Acknowledgment

I will like to show appreciation to my supervisor Dr. Jane Akinyi for her immense contribution towards this work and whose shoulder I stood.

Special appreciation also goes to my good friend Haruna Issaka for his great advises and support.

To Prof. George Orwa and the entire PAU staff I say a big thank you for the impact you have made in my life.

Lastly, my sincere thanks also goes to all who have one way or the other added value to my life to make this MSc. Program a success. I say Ayekoooo.

Abstract

Population size and growth rate have great impact on the society and economy of every country. Finding the determinants and growth rate of the population have become fundamental to policy makers in developing countries like Ghana, and its capital Accra in particular. Population growth models like the Exponential and Logistic Growth Models are mostly used in population dynamics. Most variables in this area are count variables. Count models are appropriate for modeling count variables. However, count models are hardly used in this field. Most count data have the problem of overdispersion, where data exhibits more variability than expected under an assumed model. There are possible advantages of including overdispersion in the modeling process. This study sought to investigate the effects of overdispersion in count data, and compare the Negative Binomial Model which is able to account for overdispersion to the Exponential and Logistic Growth Models which do not account for overdispersion. The results were then applied to a real dataset of the Greater Accra Region of Ghana.

The results showed that avoiding overdispersion when it do exist has dire consequences. Particularly, standard errors of the estimates are understated and the significance of some covariates are overstated. This was more evident when the data from the Greater Accra Region was modeled. Unless overdispersion is treated, inferences on such results are misleading. The study also revealed that the Negative Binomial Model actually performed better than the population growth models in modeling the population growth. The study recommends that analysts consider overdispersion when modeling count data and recommends the Negative Binomial Model to be used in modeling population data as compared to the Exponential and Logistic Growth Models.

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Background of Study

Population and census data have varied uses and benefits. In every country, information on the size, distribution and characteristics of the population is essential for describing and assessing its economic, social and demographic circumstances. Such information are also key in developing sound policies and programs (in fields such as education and literacy, employment and manpower, family planning, housing, maternal and child health, rural development, transportation and highway planning, urbanization and welfare) aimed at fostering the welfare of a country and its population (Fund, 1998).

The size of a population determines many factors such as its potential to supply skilled labor, its market demand, its ability to consume and cause an increase in aggregate demand and the growth of Gross Domestic Product (GDP), its geopolitical importance, its tax revenue potential among others. It also has the potential to generate competition, efficiency and innovation. However, irrespective of the potential gains that are associated with large population size, there is the need to ensure that the rate of growth matches the pace of economic growth otherwise it may create the problem of high youth unemployment rates, balance of payment problems as a result of increased net imports, corruption, inflation, food insecurity, deforestation, pollution, social vices and other environmental problems such as global warming and dwindling of natural resources, which hinder the well-being of the people.

Interestingly, one of the distinguishing features of low and high income countries is high population growth and size. This is evident in Africa where most of its member states are regarded as poor due to the fact that they cannot provide basic amenities like food, quality healthcare and water to their ever increasing high population growth and size. In as much as population growth provides a pool of labor force to be used in the production process with the potential to promote economic growth, it however poses some challenges which may equally impede the effort of countries in achieving desired economic growth and development. Increase in population size put financial constraint on families, which could prevent more people from getting access to quality food, education and good health care which retards human development for socio-economic growth.

The population of every country like Ghana is the nation's most valuable resource. The protection of its growth rate and enhancement of its welfare is the government's first responsibility and when that welfare is threatened government must act (May, 2005). Ghana, like most sub-Saharan countries, has been experiencing a rapid rise in its population since the early 1950's. From a population of slightly over ten million in 1980, Ghana's population rose to $25,904,598$ as at 2013 (Group, 2013), indicating a steep increase in the size. The growth rate of the population of Ghana stood at 2.7% in 2013 (Group, 2013). A slower rate of population growth ensures that more people will have better access to healthcare, education, social amenities and thus take advantage over the few job opportunities available to better their living conditions. This is, however, not the case of Ghana as its population growth is high.

Overall, the characteristics of the population of Ghana, particularly the Greater Accra Region (GAR) pose serious challenges for growth and development, and not until Ghana achieve a considerable decline in fertility rate and success in economic development, the nation's development efforts will be frustrated with high population growth rate which is an open challenge for management and policy makers. As the population continues to rise, there should be corresponding and appropriate methods and sources for calculating population size and growth, which would give a total human population figure at any given time. Also, there should be a mechanism to find the determinants of the population growth. Having such information about the population can aid the authorities to ensure that the increasing demand of services arising from the population growth is met.

Due to the ever increasing population growth naturally, it has become absolutely necessary to introduce the most common quantitative approach to population dynamics, taking note of the different theoretical foundations and assumptions to such population. Several authors, including Pearl and Reed (1920) and Meyer et al. (1999) have demonstrated that mathematical models can account for growth (increase or decline) in human population. The number of people that the environment can support, called the carrying capacity, gives an interesting background to the population research survey and evaluation of the data. Meyer and Ausubel (1999) hinted that, new technologies affect how resources are consumed and that models that employ dynamic carrying capacity are more reflective of the human condition.

At the base of population surveys and exercises, there are future projections and predictions, involving some types of mathematical models (Islam, 2009). Population values falls within the category of count variables. The Exponential and Logistic Growth Models are the most used in modeling population dynamics. To model a count variable, the count models are most appropriate for such course. However, these count models are hardly used in this area.

Count data mostly exhibit overdispersion under a specified model. Overdispersion refers to the scenario where data exhibits more variability than expected under an assumed model. This normally happens in models that exhibit mean - variance relationships. Since overdispersion is a major characteristic of count data, models that take into account overdispersion in the modeling process should be applied.

Population growth models do not consider overdispersion. There could be adverse outcomes and effects if overdispersion is ignored when it do exist. None of the current population growth models consider overdispersion in model structure. This study will reveal the effects of overdispersion and explore the importance of the count models in population dynamics.

1.2 Statement of the Problem

Most count data are faced with the issue of overdispersion. There are a number of causes of overdispersion in data. Handling this issue when modeling such data remains vital to the kind of information that can be retrieved from the data. None of the models popularly used in population dynamics is capable of handling overdispersion. Therefore, information derived from such models might be misleading.

There could be dire consequences of excluding overdispersion in modeling count data in population dynamics. Probably, most analysts do not involve overdispersion in analysing count data because they have no or little knowledge on the issues of overdispersion. This study seeks to bridge this gap by investigating the effects of overdispersion in count data modeling.

Hardly do most analysts apply count models in such a prominent area as population dynamics. Information in this area continuously have an effect on policy making. The current population growth models only depend on one independent variable which is time. Even though, humans live in such a complex environment that one would expect population values to be explained by more than one variable. Count models are capable of involving more than one independent variable in the model, having the tendency of handling overdispersion. However, such kind of models are hardly employed in population dynamics. This study therefore seeks to apply the count models that account for overdispersion in modeling population data.

1.3 Objectives of the Study

The broad objective of this research is to analyse the effects of overdispersion in count data modeling. The specific objectives are;

(i) To investigate the effects of overdispersion in count data modeling.

(ii) To compare the Exponential and Logistic Growth Models to the Negative Binomial Model in terms of predicting population values.

(iii) To determine factors that influence the population growth of the Greater Accra Region of Ghana.

1.4 Significance of the Study

Governments have viewed population data, especially, size and growth with serious concern, though they differ in the method of curtailment. At the center of each method is the projection and estimation. Islam (2009) justified the use of mathematical models for the estimation of population projection and estimation. Mathematical models are essentially an endeavour to find the relationship and their dynamic behaviour among the various elements in demography.

The study is aimed at recommending mathematical models, which the government could use for future population projection. Additionally, this study will make available information and effects of overdispersion in modeling count data. This study will make available the demographic variables that influence the population size of the GAR. This would help the government and authorities in managing the population issues of the region. Particularly, this study provides vital information to better understand population growth and the factors influencing population size, which would inform the policy instruments on the population of the region in order to achieve any set objectives. Policy makers of the region of study will have the knowledge on the problems associated with the population.

Organizations relying solely on population and census information data to provide social services can use the models herein to estimate. Population issues in the GAR and most parts of Ghana have not received much attention. Many governments have failed to address the population issues of the region. In order for Ghana to achieve its Millennium Development Goals, a critical look at the country's population, especially the GAR which has the largest population density is necessary and this work would contribute to this area of concern.

1.5 Organization of the Study

The study will be organized under five (5) chapters. Chapter 1 covers the introductory aspect of the study which highlights the statement of the research problem, objectives of the study and the significance of the study. Chapter 2 reviews the relevant literature on the subject area. Chapter 3 covers the methodology employed in this study while Chapter 4 provides the results and discussion of the study. Chapter 5 covers the summary, conclusion and recommendations for policy makers.

Chapter 2

LITERATURE REVIEW

2.1 Count Data Modeling

Count data regression modeling techniques have become important tools in empirical studies of statistics and their applicability continues to grow in other areas. In the literature like that of Winkelmann (2000), Cameron and Trivedi (2013) and Hilbe (2014), numerous examples utilizing count data methodologies are mentioned and treated. Although count data modeling techniques have a rather recent origin, the statistical analysis of counts has a long and rich history.

In count data regression, the main focus is the effect of covariates on the frequency of an event, measured by non-negative and integer-valued counts. The main motivation for using count data models to handle counts is that the standard Gaussian linear regression model ignores the restricted support (non-negativity and the integer-valued character) of the response count data variable (Hilbe, 2014). Using the standard linear regression in such cases may lead to significant deficiencies. Besides, most of the assumptions of the standard linear regression model are violated if used on count response variables.

The starting point for most cross-sectional data analyses is the Poisson Model. Due to common issues or problems like clustering in the data, repeated measurements of the response; the existence of overdispersion and the occurrence of excess zeros beyond what a Poisson Model can incorporate, the extensions of the Poisson Model is usually necessary and considered (Hilbe, 2014). Such issues can effectively be accounted for through the use of random subject-specific effects (Hens et al., 2005), through the use of an overdispersion model, such as the Negative Binomial Model for count data (Breslow, 1984; Lawless, 1987) and through the use of the so called Zero-Inflated Models (Lambert, 1992).

Poisson Regression Model is traditionally known to model count data. Unlike the familiar Gaussian distribution which has two parameters $N(\mu, \sigma^2)$, the Poisson distribution is described by a single parameter (λ), as both the mean and the variance. Thus, a Poisson distribution is characterized by two parameters having the same value, the expected value (mean) and the variance (Hilbe, 2014). This can be a strong assumption. However, this does not always apply to count data. A common deviation from the basic Poisson Model is the failure of the equal conditional mean and variance restriction. When the observed variance exceeds that of the expected (predicted) variance, we have the situation called overdispersion, which simply means extra-variability. The most commonly used explanation for overdispersion is that unobserved heterogeneity is present and this could be as a result of omitted significant explanatory variables, clustering in the data, correlation, high relative frequency of zero observations (Lambert, 1992). Another common deviation from the Poisson Model is truncation and censoring. The observed counts are left truncated if small counts are excluded from the sample (zero truncation being the most frequently encountered) and are right censored if counts exceeding some value are aggregated (Hilbe, 2014).

When there is overdispersion, the data in question can be modeled using Negative Binomial (NB) Model (Hilbe, 2014). Negative Binomial regression, like Poisson Model, examines the relationship between predictors and count dependent variable through log link, which assumes a mixture distribution for count variable. For this model, it accommodates the variance been greater than the mean. A characteristic feature of this distribution is that it accounts only for overdispersion, not for underdispersion (i.e. the variance is smaller than the mean).

When the data are characterized by excess zero counts, the data can no longer be modeled using ordinary Poisson Regression Model. There exists substantial literature on the modeling of these Zero-Inflated count data using the Hurdel Model (Mullahy, 1986) and the Zero-Inflated Poisson, Negative Binomial Models (Lambert, 1992). The Zero-Inflated Models address the excess zero problem in count data. The high frequency of zeros is accounted for by allowing extra probability mass at zero and reducing the probability mass for other frequencies. The Zero-Inflated Poisson Model was first introduced by Lambert (1992) and it is assumed that there are two processes that can generate the zeros.

The Hurdle Model is seen as a two-part model for count data. The first part is the model for the binary variable which indicates whether the response outcome is zero or positive. Conditional on a positive outcome, the second part of the model uses a truncated Poisson distribution or a truncated Negative Binomial (NB) distribution to model the positive variable. The Hurdle Model was developed by Mullahy (1986) and builds on ideas from Cragg (1971).

For the truncated count models, the distribution is truncated if the whole distribution is not observable over the whole range of non-negative integers (Hilbe, 2014). Common among these distributions are the ones truncated at zero. The zero truncation is usually referred to as the left truncation or truncation from below. Right truncation or truncation from above do exist too. These models are very suitable for count variables that do not expect certain values to be predicted.

Most empirical work in population dynamics have neglected the fact that the dependent variable is a non-negative integer. Most authors use the standard linear regression to model count response variables. In the work done by Bloom et al. (2000), which they used multiple regression, they found fertility rate, immigration, population density and birth rate to have positive relationship with population size while death rate, emigration and literacy rate to have negative effect on population size.

A similar work done by Crenshaw et al. (1997), their results differed slightly with that of Bloom et al. (2000). They found population density, birth rate, literacy rate and poverty rate to have positive effect on population size while death rate and average household size to negatively affect population size. The work also done by Otani and Villanueva (1990) resulted in birth rate, life expectancy and household size to have a positive effect on population size whereas death rate and unemployment rate have negative effect on population size.

2.1.1 Overdispersion

Overdispersion is an issue that commonly arises in Generalised Linear Model (GLM) fitting. Key assumptions related to statistical models are that all relevant predictors are included, irrelevant ones excluded and that the predictors are not correlated. Wilson and Hardy (2002) defines overdispersion simply as the case where the residual deviance is greater than the residual degrees of freedom. In other words, one speaks of overdispersion when data have larger variance than expected under a fitted model.

Overdispersion is not uncommon in practice. In fact, some believe that overdispersion is the norm in practice and nominal dispersion is the exception. The incidence and degree of overdispersion encountered greatly depend upon the field of application. Lee et al. (2002) stated that overdispersion should be accounted for in a GLM analysis. Failure to do so in the presence of overdispersion results in type I error rates well above the nominal ones. When overdispersion is not present, the test for treatment effects is not negatively affected by considering overdispersion. McCullagh and Nelder (1989) advised that unless there are good external reasons for relying on model assumptions, it is wise to be cautious and to assume that overdispersion is present unless and until it is shown to be absent.

According to Cox (1983), overdispersion in general has two effects. One is that summary statistics have a larger variance than anticipated under the simple model. This has long been recognized and is commonly allowed for by an empirical inflation factor, either assumed from prior experience or estimated. The second effect is a possible loss of efficiency in using statistics appropriate for the assumed distribution.

There are several typical techniques employed when overdispersion is present to account for inflated deviance. It is important to note that these techniques do not remove the model of overdispersion, but rather embrace and account for it in the model structure. The first general technique is to take a quasi-likelihood approach (Hilbe, 2014). In this technique one uses the same error distribution (i.e. Poisson, binomial) but adjusts the standard errors and test statistics. Specifically, hypothesis testing with an F-test instead of Chi-squared is recommended. In other words, model overdispersion by letting the actual variance equal the assumed variance multiplied by an additional scale parameter that adjust for the discrepancy between assumed and actual (Hilbe, 2014). Another way of thinking of this is that using the F-test vs. Chi-squared (or any other) does not make overdispersion go away but simply takes it into account in the hypothesis testing.

In fact, studies of overdispersion in data have focused mainly on two regimes. One is a detailed representation of overdispersion by a specific model. The main models include a sampling density which is a mixture of the exponential families. Most of the time the log likelihood ratio is used to determine the goodness-of-fit of the model. The other is to introduce procedures which add a dispersion parameter to the exponential family (Albert and Pepple, 1989), and develop test statistics to detect the existence of the dispersion parameter. The test statistic is derived under the null hypothesis that the overdispersion parameter is zero.

In one-parameter exponential families such as the binomial and Poisson, the variance is a function of the mean. The existence of overdispersion, which is a common practical complication, leads to a failure of the variance-mean relation (Cox, 1983; Efron, 1989). Therefore, many efforts have been made to develop models to represent overdispersed exponential families.

A classical approach to the problem of the overdispersed Poisson is to treat the Poisson means associated with each observed count as latent variables that are sampled from a specific parametric distribution (Breslow, 1984). Most authors have considered a gamma mixing distribution, which leads to a negative binomial distribution for the observed data (Collings and Margolin, 1985). Paul and Plackett (1978) examined the effect of the Poisson mixtures, especially as represented by the negative binomial distribution, on statistical inference. They found that probabilities of rejecting the null hypothesis of no overdispersion are increased, sometimes considerably, confirmed by Collings and Margolin (1985).

Collings and Margolin (1985) proposed methods for detecting negative binomial departures from a Poisson model. They dealt with three kinds of problems namely; where the mean response is constant; where the mean response depends on a single covariate and the regression line is through the origin, or where it takes on a fixed number of values according to a one-way layout.

Dean and Lawless (1989) developed tests for detecting extra-Poisson variation in Poisson regression models. They obtained the tests as score tests against arbitrary mixed Poisson alternatives, and provided approximations for computing the significance level and power of the tests against the negative binomial alternatives. The mixed model they used was an extension of that of Collings and Margolin (1985). The model assumed that the random effects had finite first and second moments. If the random effects follow a gamma distribution, then the random variable has a negative binomial distribution. Dean and Lawless (1989) findings were that for detecting overdispersion, tests analogous to those based on their models were superior to tests based on the Pearson statistic for more general regression situations, at least for moderate amounts of overdispersion. But, as they pointed out, the test was designed to be particularly effective against one type of extra-Poisson variation.

2.2 Modeling Population Growth

Malthus (1888) in his research on human population stated that prior to the transformational stages of societies into well organized and developed ones, rapid population growth results in worsened standard of living because of low level of technological advancement in agriculture, and fixed supply of land. To him, rapid population growth may inevitably collide with the scarce resources and if not checked would lead to famine, war, and other natural disasters like diseases. Around the late 1960s, Neo-Malthusians analysts and commentators further expanded the view of Malthus (1888) and asserted that societal institutions cannot adjust economies to accommodate pressures from rapid population growth and high human density which may in turn cause havoc to the environment and social fabric.

The opposite view, that institutions can faultlessly handle these changes if allowed to occur without restriction, was shared by two groups of scholars, Neo-liberals and Distributionists. They opposed the Malthusian view where Simon (1972), a Neo-liberal argued that population growth is not a problem and that discussion of it only distracts attention from real problems since institutions can evolve, supported by technological progress with inventions which may provide all the necessary tools for adjustment. Distributionists were also of the view that concern about population growth is destruction from real problems and that state institutions may implement poverty alleviation and equity programs which may serve as adjustment mechanism. Both schools affirmed that the positive economic effects that may result from their own brand of institutional mediation may cause a decline in demand for children.

In the midst of this debate, emerged the Revisionist view in the 1980s. They opined that to adequately adjust for population growth, populous countries must have a broad array of functional modern institutions whose smooth performance may help to curb any possible negative effect of population growth. They cautioned that poor performance of these institutions would precipitate dire economic consequences in such economy. This includes low levels of output per worker and failure to meet societal goals for allocating goods and services. They concluded that irrespective of the size of national output, population growth can degrade renewable natural resources when property rights are nonexistent or inadequately assigned.

By the late 1980s, most of the population policy discussions went in favour of the Revisionist view that the effects of population growth on the economy were insignificant (Birdsall et al., 2001). This was revealed by the 1986 National Academy of Sciences report which stated that "On balance, we reach the qualitative conclusion that slower population growth would be beneficial to the economic development of most developing countries" (Landau et al., 1986). This is a very weak and qualified observation that led to a decline in the political and policy priority on reducing population growth in order to boost economic growth in developing counties. Cairo Population and Development Conference in 1994 emphasized the need of women's reproductive health and gender equality as the most important policy goals, not lowering world population growth (McIntosh and Finkle, 1995).

Thirlwall (1989) in his book "Growth and development with special reference to developing economies" said the relationship between population and economic development is a complex one, particularly concerning what is cause and what is effect. According to his book, many people consider rapid population growth in the third world to be a major obstacle to development, yet there are many ways in which population growth may be a stimulus to progress and there are many rational reasons why families in developing countries choose to have many children. The United Nations world population conference held in Bucharest in 1974 adopted a world population plan of Action that asserted "population and development are interrelated, population variables influence development and are also influenced by them" (Paul, 1994). The plan recommended that "population measures and programs should be integrated into comprehensive social and economic plans and programs and this integration should be reflected in the goals, instrumentality and organizations for planning within the countries.

Ausubel et al. (2001) in their work did indicate that decrease or growth comes from the interplay of death and birth (and locally migration). To justify the use of mathematical models to analyse population, Islam (2009) said, "Mathematical model is very important for the estimation of population projections. In fact, mathematical model is essentially an endeavour to find out structural relationships and their dynamic behaviour among the various elements in demography."

Ghana became the third country in Sub-Saharan Africa to adopt an explicit and a comprehensive population policy in 1969 after Mauritius (1958) and Kenya (1967) (NDPC, 2003). The policy was meant to affect the growth, structure and distribution of the country's population. Although the 1969 policy was retained by successive governments, very little progress was made during the next two decades in reducing the rate of population growth because political commitment was absent. In 1994, the policy was revised to ensure that Ghana achieves and maintains a level of population growth that is consistent with national development. Furthermore, the government's long term vision (Ghana-vision 2020) for the country is to attain a balanced economy that helps people attain a middle-income status and standard of living by the year 2020 (NDPC, 2003). Again, the vision sets a goal to reduce population growth from its present level of around 3% to 2% each year.

Population growth is the change in a population over time, and can be quantified as the change in the number of individuals of any species in a population using "per unit time" for measurement. In Biology, the term population growth is likely to refer to any known organism, but in this study it is applicable only to human populations (demography). According to Islam (2009), population growth is used informally for the more specific term population growth rate, which refers specifically to the growth of the human population of the world.

Islam (2009) noted four outstanding changes of world growth rate: first, the agricultural revolution which took place in Southwest Asia and China; second, the global agricultural revolution; third, the fall in death rate which was in the decades after 1950; and finally, the change in fertility rates since the 1980s. It is noteworthy that while the first three changes concern increases in the rate of population growth, the fourth is a decrease in population growth. He further maintained that, "There are enormous concerns about the consequences of human population growth for social, environmental and economic development. Intensifying these problems is population growth."

The United Nations collaborated the above observation, estimating world population to be 6.2 billion by 2000 (with the poorest areas of the world have the highest population growth rates. For example, out of the about 90 million babies born in the world in 1995, about 85 million were born in developing countries), which would increase to 9.8 billion by 2050 (Winkelmann, 2000). Globally, the growth rate of the human population has been declining since 1962 to 1963 at 2.20% per annum. In 2009 for example, the estimated annual growth rate was 1.1%. The CIA Factbook (Factbook, 2017) gives the world annual birth rate, mortality rate, and growth rate as 1.915%, 0.812%, and 1.092% respectively. The last one hundred years have seen a rapid increase in population due to medical advances and massive increase in agricultural productivity made possible by the Green Revolution. Some countries experience negative population growth, especially in Eastern Europe mainly due to low fertility rates, high death rates and emigration. In Southern Africa, growth is slowing due to the high number of HIV-related deaths.

Two mathematical models: Exponential Growth and Logistic Growth Models are popular in research of the population growth. The Exponential Growth Model was proposed by Malthus in 1798 (Malthus, 1888) , and it is sometimes called the Malthusian Growth Model. Verhulst proposed the logistic growth model in 1845 (Verhulst, 1838). Both originated from observations of biological reproduction process. However, when it comes to human population, the constant growth rate can be observed in doubt. Human reproduce sexually and have consciousness. We cannot be sure under what condition will an individual be reproduced. Therefore, it is in doubt whether models based on constant growth rate can explain human population growth. Taking carrying capacity into regards, Logistic Growth Model improves the preceding Exponential Growth Model.

The use of the Logistic Growth Model is widely established in many fields of modeling and forecasting (Banks, 2013). Through mathematical modeling approach, much work has been done to further develop these models so as to predict population growth accurately (Rasmussen et al., 2016; Kulkarni et al., 2014). Population growth model is a pattern of a mathematical model that is utilized to the study of population dynamics. Models allow the better perception of how processes and complex interactions work. Population growth modeling can provide a manageable way of understanding how numbers change over time or in relation to each other. Edwards and Weizi (2011) indicated that the Logistic Model with explicit carrying capacity is most convenient way to study population growth as the related equation contains few parameters. Ghana's population growth has been modeled by Ofori et al. (2013). In their work, they predicted Ghana's population to be 114.8207 million in 2050, with a growth rate of 3.15% per year using the Exponential Model. By the Logistic Growth Model, Ofori et al. (2013) predicted Ghana's population to be 341.2443 million in 2050, with a growth rate of 5.23% per year.

2.3 Summary of Chapter

Relevant literature on count models, and the issue of overdispersion have been reviewed. The Exponential and Logistic Growth Models are the two most used models in population dynamics. Even though count variables occur frequently in this field, count models are hardly used. In the next Chapter, the methodology employed in this study and the materials used will be looked at.

Chapter 3

METHODOLOGY

3.1 Effects of Overdispersion

Modeling count data, it is important to know what is meant by a count. In Statistics, a count refers to observations that have only non-negative integer values. Example of a count is the number of students that PAUSTI admits in a year. A count can be thought of as a verb which means to enumerate units, items, or events (Hilbe, 2014). The Poisson model is the basic count model. The nature of the count response variable determines which kind of the count models to use.

Overdispersion is mostly encountered when modeling count data. It occurs in models that have mean-variance relationship. It occurs when data exhibits more variability than expected under an assumed model. Thus, mathematically, it can be given as $Var(y) > Var(y|x)$. In the Poisson Model, there is the assumption of equidispersion, where the mean is expected to be equal to the variance, thus $Var(y|x) = E(y|x)$. When $Var(y|x) > E(y|x)$, then we have overdispersion under the Poisson Model. Several sources can be attributed to overdispersion. Common among these sources as given by Hilbe (2014), Mohammed and Laney (2006) and Iannario (2014) are variability of experimental material - individual level variability, correlation between individual responses, cluster sampling, aggregate level data, omitted significant variables, excess zero counts (structural and sampling zeros), misspecified link function, outliers.

There are consequences of not treating overdispersion if it does exist. Some authors (Hilbe, 2014; Iannario, 2014) have hinted that unless overdispersion is properly handled, it can lead to invalid inferences.

To investigate the effects of overdispersion, the study simulated data that exhibits overdispersion and fitted the Poisson Model, which does not account for overdispersion and the Negative Binomial Model, which takes into account the overdispersion in data. The simulations were done with the R statistical software. The simulated data was generated as follows:

The model used for the data generating process was

$$\ln(\mu) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6,$$

with parameter values $\beta_1 = 5$, $\beta_2 = 1$, $\beta_3 = 0$, $\beta_4 = -1$, $\beta_5 = 0$, $\beta_6 = -1$ and a sample size of $n = 200$. The covariates x_2, x_4 and x_6 were generated from the normal distribution whereas x_3 and x_4 were generated from the uniform distribution. The response variable y was generated from the Poisson distribution.

The models used on the simulated data are briefly discussed below:

3.1.1 The Poisson Model

In this model, the response variable Y follows the Poisson density given as

$$f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \ldots \tag{3.1}$$

where μ is the mean of Y. Figure 3.1 shows the distribution of the Poisson family with different expected values. Note that, small mean values makes the distribution skewed to the right and as the mean increases, the distribution approaches normal. λ is usually used to represent the expected value under the Poisson distribution but for purposes of modeling, we use μ.

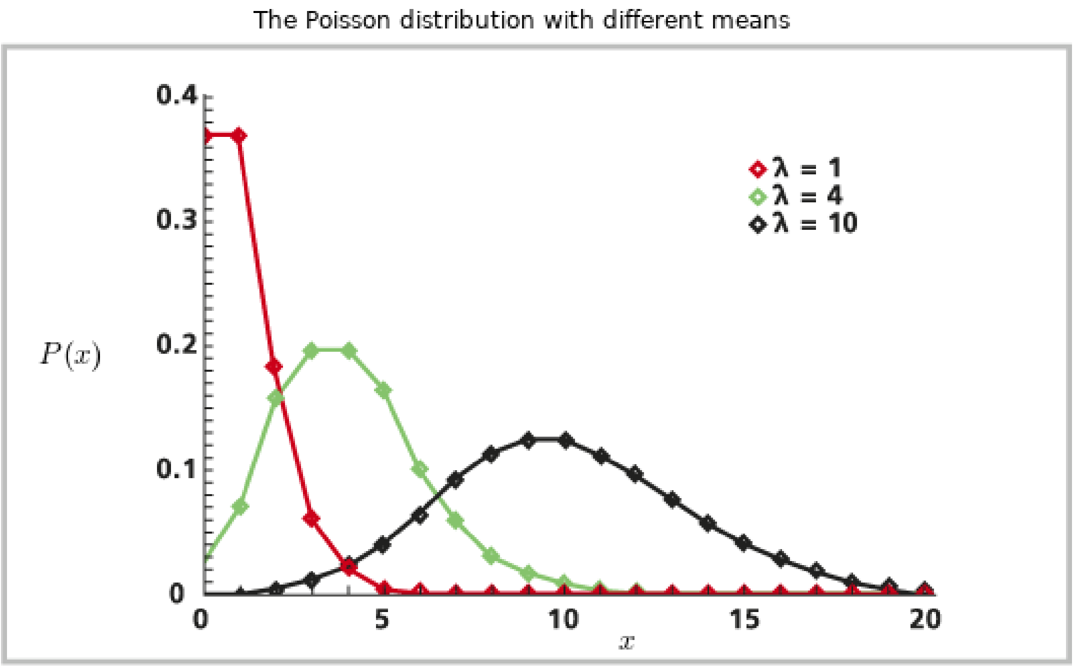


Figure 3.1: The Poisson distribution

The Poisson distribution belongs to the 1-Parameter Exponential Family of Distributions, whose density is generally given in the form:

$$f(y;\theta) = \exp\Big\{y_i\theta_i - b(\theta_i) + c(y_i)\Big\}, \tag{3.2}$$

where θ_i is the canonical parameter or link function, $b(\theta_i)$ is the cumulant and $c(y_i)$ is the normalization term. The first and second derivatives of $b(\theta_i)$ defines the mean and variance of Y respectively. Now, we write the Poisson density in the form of Equation (3.2) as

$$f(y;\theta) = \exp\Big\{y_i \ln(\mu) - \mu - \ln(y!)\Big\}. \tag{3.3}$$

So for the Poisson distribution, the link function is $\ln\mu$ and has both the mean and variance to be μ.

In the Poisson Model, μ_i is expressed as a function of some explanatory variables X through its link function in the form

$$\ln(\mu_i) = X_i'\beta, \tag{3.4}$$

where β are the regression parameters to be estimated. In expanded form, the Poisson model is given as

$$\ln(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \tag{3.5}$$

The regression parameters β are estimated via the maximum likelihood method. The Poisson log-likelihood is given as

$$\mathcal{L}(\mu; y) = \sum_{i=1}^{n} \Big[y_i \ln(\mu_i) - \mu_i - \ln(y_i!) \Big] \tag{3.6}$$

Recall from Equation (3.4) that $\mu_i = \exp(X_i'\beta)$. So the Poisson log-likelihood becomes

$$\mathcal{L}(\beta; y) = \sum_{i=1}^{n} \Big[y_i(X_i'\beta) - \exp(X_i'\beta) - \ln(y_i!) \Big] \tag{3.7}$$

The first derivative of the log-likelihood function with respect to β defines what we call the gradient or score function. Thus, the score function of the Poisson log-likelihood is

$$\frac{\partial(\mathcal{L}(\beta; y))}{\partial\beta} = \sum_{i=1}^{n} \Big[y_i - \exp(X_i'\beta)\Big] X_i' \tag{3.8}$$

Setting Equation (3.8) to zero and solving for β gives the estimates of β.

We find the second derivative of the log-likelihood function with respect to β as

$$\frac{\partial^2(\mathcal{L}(\beta; y))}{\partial\beta\,\partial\beta'} = -\sum_{i=1}^{n} \Big[\exp(X_i'\beta)\Big] X_i X_i' \tag{3.9}$$

The second derivative is a matrix called the Hessian matrix H. To find the maximum likelihood variance-covariance matrix, the Hessian H is inverted and negated. This is often represented as $\sum$.

$$\sum = -H^{-1} = \left\{\sum_{i=1}^{n} \Big[\exp(X_i'\beta)\Big] X_i X_i'\right\}^{-1} \tag{3.10}$$

The standard errors of the parameter estimates are found by finding the square root of the respective terms of the diagonal of $\sum$.

3.1.2 The Negative Binomial Model

This model is mostly used to correct or model an overdispersed Poisson Model. The Poisson Model has the equidispersion assumption implying that the mean and variance must be same. This equidispersion assumption is often not met when analyzing count data in practice. When a data exhibit more variablity than was expected under a model, then the scenario is termed overdispersion. The Negative Binomial Model is nearly always used to model an overdispersed Poisson data (Hilbe, 2014). This model helps to model a far wider range of variability than the Poisson.

The Negative Binomial is a two-parameter model; the mean (μ) and a second parameter called the dispersion parameter (α), which provides a wider distribution of counts than allowed under the Poisson distribution. The Negative Binomial Model has the same modeling form as the Poisson Model;

$$\ln(\mu_i) = X_i' \beta$$

or

$$\ln(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

The Negative Binomial Model has two forms: the NB1 model which has a variance of $\mu+\alpha\mu$ and the NB2 model which has a variance of $\mu + \alpha\mu^2$. The NB2 is the traditional parameterization of the Negative Binomial and it is the most widely used among the two forms. We used the NB2 model in study.

The Negative Binomial distribution can be obtained mathematically from the Binomial or Gamma or a mixture of the Poisson and Gamma distributions (Hilbe, 2014). For purposes of statistical modeling, the Poisson-Gamma distribution is predominant. Its probability distribution can be expressed in several ways, with the common parameterization of the form:

$$f(y;\mu,\alpha) = \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \tag{3.11}$$

The Negative Binomial Model uses the maximum likelihood method in estimating the regression parameters. The parameters can also be estimated using an Iteratively Re-weighted Least Squares (IRLS) algorithm within the domain of Generalized Linear Models (GLM). Both parameters μ and α are estimated when the Negative Binomial uses the full maximum likelihood algorithm. On the other hand, when the GLM algorithm is used, only μ is estimated; α is inserted into the algorithm as a constant.

The log-likelihood of the Negative Binomial is given as

$$\mathcal{L}(\mu;y,\alpha) = \sum_{i=1}^{n} y_i \log\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\log(1+\alpha\mu_i) + \log\Gamma\left(y_i + \frac{1}{\alpha}\right) - \log\Gamma(y_i+1) - \log\Gamma\left(\frac{1}{\alpha}\right) \tag{3.12}$$

Since $\mu_i = \exp(X_i'\beta)$, we can write the log-likelihood as

$$\begin{aligned}\mathcal{L}(\mu;y,\alpha) = \sum_{i=1}^{n} y_i \log\left(\frac{\alpha\exp(X_i'\beta)}{1+\alpha\exp(X_i'\beta)}\right) - \frac{1}{\alpha}\log\left(1+\alpha\exp(X_i'\beta)\right) + \log\Gamma\left(y_i + \frac{1}{\alpha}\right) \\ - \log\Gamma(y_i+1) - \log\Gamma\left(\frac{1}{\alpha}\right)\end{aligned} \tag{3.13}$$

The Poisson distribution has just one parameter (the mean). In the quest to treat overdispersion which almost always occur in count data, a second parameter, called the dispersion parameter (α) is introduced to solve the overdispersion issue. In the case of the Poisson Model, we expect this relation $Var(y) = \mu$. If α is introduced, then we would expect $Var(y) = \alpha\mu$.

This α is mostly assumed to be 1 in the modeling process so the equidispersion assumption can be met. When overdispersion is ignored when it does exist, results from the model would be biased. Overdispersion is supposed to be treated and the necessary adjustments made to the standard errors of the estimates. This is well catered for by the Negative Binomial Model.

Detecting Overdispersion and Model Assessment

Several methods exist on how to determine overdispersion after a model is fitted. Common among them is the Pearson dispersion statistic. This statistic is based on the Pearson Chi-squared(χ^2) statistic, which is the sum of squared residuals divided by the variance. Thus

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{Var(y_i)} \tag{3.14}$$

The Pearson dispersion statistic (d) is derived by dividing χ^2 by the residual degrees of freedom. The residual degree of freedom is simply the number of observations in the model less the number of predictors.

$$d = \frac{\chi^2}{n - p}, \tag{3.15}$$

where p is the number of predictors. The dispersion statistic should have a value of 1 if there is no unaccounted variability in the model other than what is expected based on the variance function. Values greater 1 indicate an overdispersed model.

The models were assessed based on their deviances and AIC values. The deviance (D) is defined as the difference between a saturated log-likelihood and full model log-likelihood.

$$D = 2\sum_{i=1}^{n}\Big\{\mathcal{L}(y_i; y_i) - \mathcal{L}(\mu_i; y_i)\Big\} \tag{3.16}$$

$$AIC = -2(\mathcal{L} - K), \tag{3.17}$$

where $\mathcal{L}$ is the model likelihood and K is the number of predictors.

3.2 Comparison of the Models

In determining population growth and predicting population values in demography, two models are mostly used. These two are the Exponential and Logistic Growth Models. These two, however, do not account for overdispersion if it exist. To this effect, we sought to compare these models with the Negative Binomial Model, which has the tendency of correcting overdispersion. These models were compared to see which one predicts population values correctly. The two population growth models are discussed below:

3.2.1 The Exponential Growth Model

The first growth model we considered was the Exponential Growth Model, also known as the Malthus Model or the Natural Growth Model. The model is based on the assumption that the population grows at a rate proportional to the size of the population. This assumption seems only realistic to a population with ideal conditions (eg. unlimited environment and space, adequate nutrition, absence of predators, immunity from disease and absence of natural disasters). In simple terms, this simple model is appropriate in the initial stages of growth when there are no restrictions or constraints on the population. But in more realistic situations, there are limits to growth such as finite space or food supply.

Let the population at a given time $t = t_0$ be P_0 and the population at a future time $t = t_1$ be P. Our interest was in finding a population function $P(t)$ for $t_0 \leq t \leq t_1$ with the initial condition $P(t_0) = P_0$. The Exponential Growth Model has the form:

$$\frac{dP}{dt} = rP(t), \quad t_0 \leq t \leq t_1, \quad P(t_0) = P_0, \tag{3.18}$$

where r is a constant called the *Malthus factor*, which represents the growth rate. If we assume that $P \neq 0$, then $P(t) > 0$ for all t. This shows that, for $r > 0$, then $\frac{dP}{dt} > 0$. This implies that the population is always growing and as $P(t)$ increases, $\frac{dP}{dt}$ increases.

Equation (3.18) is a first order differential equation. We find the solution to this Initial Value Problem (IVP) as follows:

Using separation of variables and integrating both sides of Equation (3.18), we have

$$\begin{aligned}
\frac{dP}{P} &= r\,dt \\
\int \frac{dP}{P} &= r \int dt \\
ln\,P &= rt + c_0 \\
P &= e^{(rt+c_0)} \\
P &= c\,e^{rt}, \quad \text{where } c = e^{c_0}
\end{aligned}$$

Now, applying the initial condition $P(t) = P_0$ at $t = 0$, we find c as

$$P_0 = c\, e^0$$

$$\Rightarrow c = P_0$$

Hence, the general solution to the IVP is

$$P(t) = P_0\, e^{rt} \tag{3.19}$$

3.2.2 The Logistic Growth Model

A slightly more realistic and largely used population growth model is the Logistic Growth Model proposed by Belgian scientist, Pierre Francois Verhulst in 1838 (Verhulst, 1838). In an environment that will support a limited population, it is assumed that the rate of growth of population decreases as the limiting population is approached. An appropriate model is the Logistic Model which has the form:

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right), \quad P(0) = P_0, \tag{3.20}$$

where $K > 0$ is the carrying capacity; the size of population that the environment can long term sustain and r is the constant growth rate. The right hand side of Equation (3.20) can be written as

$$rP - \frac{rP^2}{K} \tag{3.21}$$

From Equation (3.21) that the first term is responsible for growth of the population while the second term limits this growth due to lack of available resources or other reasons. The size of the population exerts negative feedback on its growth rate and as the population size increases, the rate of increase declines, leading eventually to an equilibrium population size known as the carrying capacity.

It is important to realize in Equation (3.20) that when the population size is very small (close to zero), the term in parentheses is approximately one and population growth is approximately exponential (similar to the Exponential Growth Model). When the population size is close to the carrying capacity ($P \approx K$), the term in parentheses approaches zero, and population growth ceases.

Equation (3.20) is also a first order differential equation. We solve this equation as follows: Separating the variables and integrating, we have

$$\begin{aligned} \frac{dP}{P\left(1-\frac{P}{K}\right)} &= r\,dt \\ \int \frac{dP}{P\left(1-\frac{P}{K}\right)} &= r \int dt \\ \int \frac{dP}{P\left(1-\frac{P}{K}\right)} &= rt + c \end{aligned} \tag{3.22}$$

Now, we resolve $\frac{1}{P\left(1-\frac{P}{K}\right)}$ into partial fractions so we can easily do the integration.

$$
\begin{aligned}
\frac{1}{P\left(1-\frac{P}{K}\right)} &= \frac{A}{P} + \frac{B}{\left(1-\frac{P}{K}\right)} \\
&= \frac{A\left(1-\frac{P}{K}\right) + BP}{P\left(1-\frac{P}{K}\right)} \\
\Rightarrow 1 &= A\left(1-\frac{P}{K}\right) + BP \\
&= A - \frac{PA}{K} + BP \\
&= A + P\left(B - \frac{A}{K}\right)
\end{aligned}
$$

Now, comparing coefficients, we have

$$A = 1$$

and

$$
\begin{aligned}
B - \frac{A}{K} &= 0 \\
\Rightarrow B &= \frac{1}{K}.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\frac{1}{P\left(1-\frac{P}{K}\right)} &= \frac{1}{P} + \frac{1}{K\left(1-\frac{P}{K}\right)} \\
&= \frac{1}{P} + \frac{1}{K-P}
\end{aligned}
$$

Now, putting the partial fractions into Equation (3.22), we get

$$
\begin{aligned}
\int \left(\frac{1}{P} + \frac{1}{K-P} \right) dP &= rt + c \\
\ln(P) - \ln(K-P) &= rt + c \\
\ln(K-P) - \ln(P) &= -rt - c \\
\ln\left(\frac{K-P}{P} \right) &= -rt - c \\
\frac{K-P}{P} &= e^{-rt-c} \\
\frac{K-P}{P} &= Q\, e^{-rt}, \quad \text{where } Q = e^{-c} \\
K - P &= PQ\, e^{-rt} \\
K &= P + PQ\, e^{-rt} \\
K &= P\left(1 + Q\, e^{-rt}\right) \\
P &= \frac{K}{1 + Q\, e^{-rt}}
\end{aligned}
$$

We now apply the initial condition $P(0) = P_0$ to find Q

$$
\begin{aligned}
P_0 &= \frac{K}{1 + Q\, e^{0}} \\
P_0 &= \frac{K}{1 + Q} \\
\Rightarrow Q &= \frac{K}{P_0} - 1
\end{aligned}
$$

So now, the solution to the differential equation is

$$
P(t) = \frac{K}{1 + \left(\frac{K}{P_0} - 1 \right)\, e^{-rt}} \tag{3.23}
$$

Now, taking limit as $t \to \infty$ of Equation (3.23) gives

$$\begin{aligned}
\lim_{t\to\infty} P(t) &= \lim_{t\to\infty} \left[\frac{K}{1 + \left(\frac{K}{P_0} - 1\right) e^{-rt}} \right] \\
&= \frac{K}{1 + \left(\frac{K}{P_0} - 1\right) e^{-\infty}} \\
&= K,
\end{aligned}$$

which clearly shows that the maximum value that $P(t)$ can take is K, the carrying capacity.

Figure 3.2 shows the difference between the two growth models.

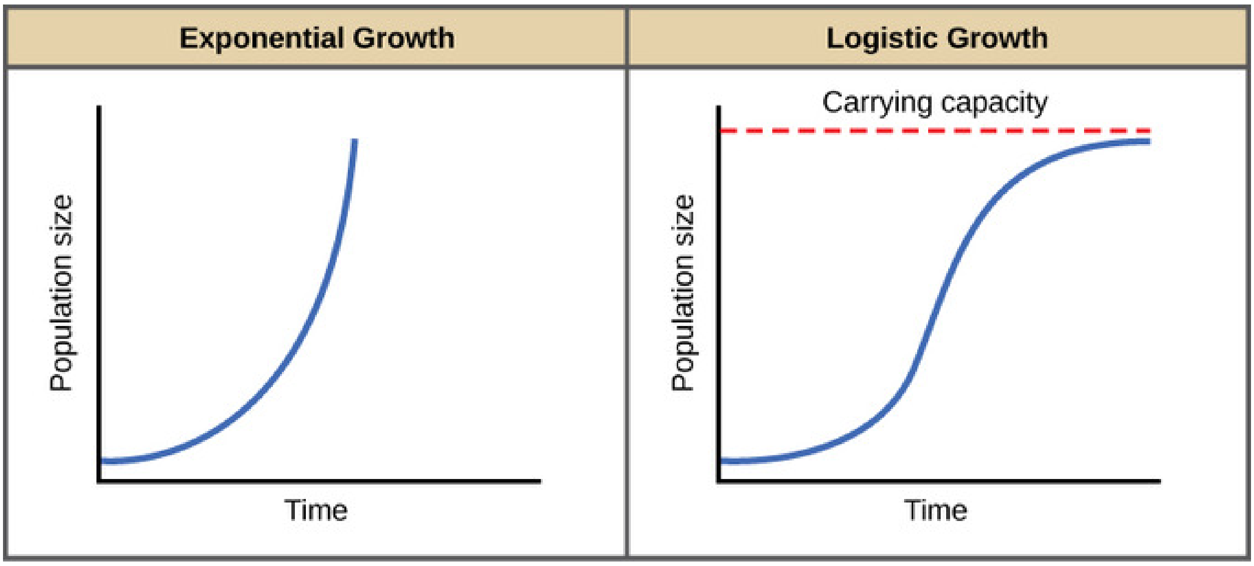


Figure 3.2: The difference between the Exponential and Logistic Growth Models

Mean Absolute Percentage Error (MAPE)

Both population growth models were assessed via the MAPE, to assess the performance of forecasting models. It is usualy used as a tool to assess the Goodness of Fit of forecasting models. It is expressed in percentages and lower values indicate better models. Lewis (1982) reports that MAPE values less than 10%, from 10% to 20%, from 21% to 50%, and above 50% represent a highly accurate, good, reasonable and inaccurate forecasting respectively. MAPE is given mathematically as

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i - \hat{Y}_i|}{Y_i} \times 100, \tag{3.24}$$

where Y and $\hat{Y}$ are the actual and predicted values respectively, and N is the number of observations.

Now, observe the differences between these three models. The Negative Binomial Model is capable of correcting overdispersion and accommodating more covariates. However, it cannot determine the growth rate of the population. On the other hand, the two growth models are able to determine growth rates. These two are, however, only time dependent. This implies they only depend on the time variable and not on any other covariates.

For the comparison, we again simulated a dataset for this purpose. The response variable, which is the Population, was generated using a combination of both growth models. In particular, we used Equation (3.25) in generating the response values.

$$P(t) = P_0 e^{rt}\left[\frac{K}{1 + \left(\frac{K}{P_0} - 1\right) e^{-rt}}\right] \tag{3.25}$$

We used these parameter values: $r = 0.01$, $P_0 = 500$ and $K = 20000$. The variable t was generated as a sequence of natural numbers. A sample size of 250 was used in the generation process. Note that the Negative Binomial Model was not involved in the generation process. This was to ascertain how well this model can predict the response values even though it was not used in generating them. All three models were fitted on the simulated data and their predictive prowess observed graphically.

3.3 Finding the determinants of population size

Our last objective in this study was to apply count models on a secondary data for the Greater Accra Region (GAR) of Ghana. This was to help determine variables that have influence on the population size. The data was obtained from the Ghana National Population Council and the Ghana Statistical Service. The data contained 13 variables in total and had 56 observations with no missing values.

These variables included: Year which spans from 1960 to 2015, Population which represents the population size of the GAR. The Population was the response variable in the models. The following variables were considered the covariates in the models: Fertility Rate which represents the average number of births per woman, Literacy Rate which indicates the percentage of the population that are educated (can read and write), Unemployed which is the percentage of the population that do not work, Population Density which is the number of persons per square kilometer, Poverty Level which is the percentage of the populates that fall below the poverty line, Birth Rate which is the number of births per 1000 individuals, Death Rate which is the number of deaths per 1000 individuals, Immigrants which represents the number of persons that move to settle in the region, Emigrants which represents the number of persons that move out of the region to settle elsewhere, Life Expectancy which represents the average number of years a person can live in the region and Household Size which represents the average number of persons per household.

Since the response variable (Population) is a count, we first fitted the Poisson Model. We observed overdispersion under this model and therefore went ahead to fit the Negative Binomial Model. We additionally fitted the Logistic Growth Model to this data.

3.4 Chapter Summary

In this Chapter, we have looked at the theoretical framework of the various models used in this study. Specifically, we presented the Exponential and Logistic Growth Models, the Poisson and the Negative Binomial Models. In the next Chapter, we present the results of applying these studied models on simulated and real dataset.

Chapter 4

RESULTS AND DISCUSSION

4.1 Effects of Overdispersion

We sought to find out the effects of overdispersion in count data. A dataset was simulated and two count models were fitted. In the simulation, two parameter values β_3 and β_5 were set to zero. This implied that covariates $x3$ and $x5$ were insignificant. This was allowed to check if both models could effectively produce those covariates insignificant. The results from the fitted models are presented in Tables 4.1 and 4.2. With a significance level of 5%, the Poisson Model fitted all the covariates to be significant whereas the Negative Binomial Model was able to reveal the insignificant covariates correctly. P-values less than 0.05 indicate statistical significance. This shows that when overdispersion is present, Poisson Models can lead to erroneous conclusions.

Table 4.1: Fitted Poisson and Negative Binomial Models

	Parameter	Poisson Model			NB Model		
		Estimate	Standard error	P - Value	Estimate	Standard error	P - Value
(Intercept)	5	6.542180	0.004963	0.0000	6.47987	0.17783	0.0000
x2	1	1.004240	0.001475	0.0000	1.05371	0.06800	0.0000
x3	0	-0.294190	0.005322	0.0000	-0.22038	0.22790	0.3340
x4	-1	-1.200497	0.002146	0.0000	-1.03900	0.07534	0.0000
x5	0	-0.156240	0.006510	0.0000	0.16158	0.23485	0.4910
x6	-1	-1.018769	0.001813	0.0000	-0.87627	0.06240	0.0000

From Table 4.2, we observe the overdispersion under the Poisson Model is far greater than 1, which indicates serious overdispersion. In Table 4.1, we can see that all the standard error values under the Poisson Model as smaller than that of the Negative Binomial. This confirms that, as a result of overdispersion, standard errors are underestimated, which in effect overstates the significance of some covariates. The overdispersion in the data was almost corrected by the Negative Binomial Model, as it can be observed in Table 4.2 that its dispersion statistic is almost 1.

In Table 4.2, we observe the word Deviance twice in the model output. Deviance is a measure of goodness of fit of a fitted model or rather, it is a measure of badness of fit; higher numbers indicate worse fit. The null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean). Both the null and residual deviance under the Negative Binomial Model are smaller than that of the Poisson Model, which indicates a better fit.

Table 4.2: The fitted Poisson and Negative Binomial Models assessment

	Poisson Model	NB Model
AIC	318570	2953.1
Null deviance	1545386 on 199 degrees of freedom	841.93 on 199 degrees of freedom
Residual deviance	317026 on 194 degrees of freedom	227.78 on 194 degrees of freedom
Dispersion statistic	1775.136	1.276

4.2 Comparison of the Models

To compare the growth models and the Negative Binomial Model, all three models were fitted to the simulated data. The results are shown in Figure 4.1. Observe that, even though the Negative Binomial was not involved in generating the data, it performed better than the two growth models. It almost fitted the data perfectly as manifested in the figure. The Logistic Growth Model, in turn, performed better than the Exponential Growth Model.

Table 4.3: MAPE values of the growth models

Model	MAPE
Exponential Growth	18.06%
Logistic Growth	14.01%

The Exponential Growth Model predicted a constant growth rate of 0.035, while that of the Logistic Growth Model was 0.048, and the MAPE values for both models were 18.06% and 14.01% respectively, which both exhibit a fairly good fit.

To ensure a fair comparison, only one covariate, which is the time variable was fitted in the Negative Binomial since the other two models depend only on time. After the fit, the Negative Binomial reported a dispersion statistic of 1, which meant that all the overdispersion in the data was perfectly corrected by this model.

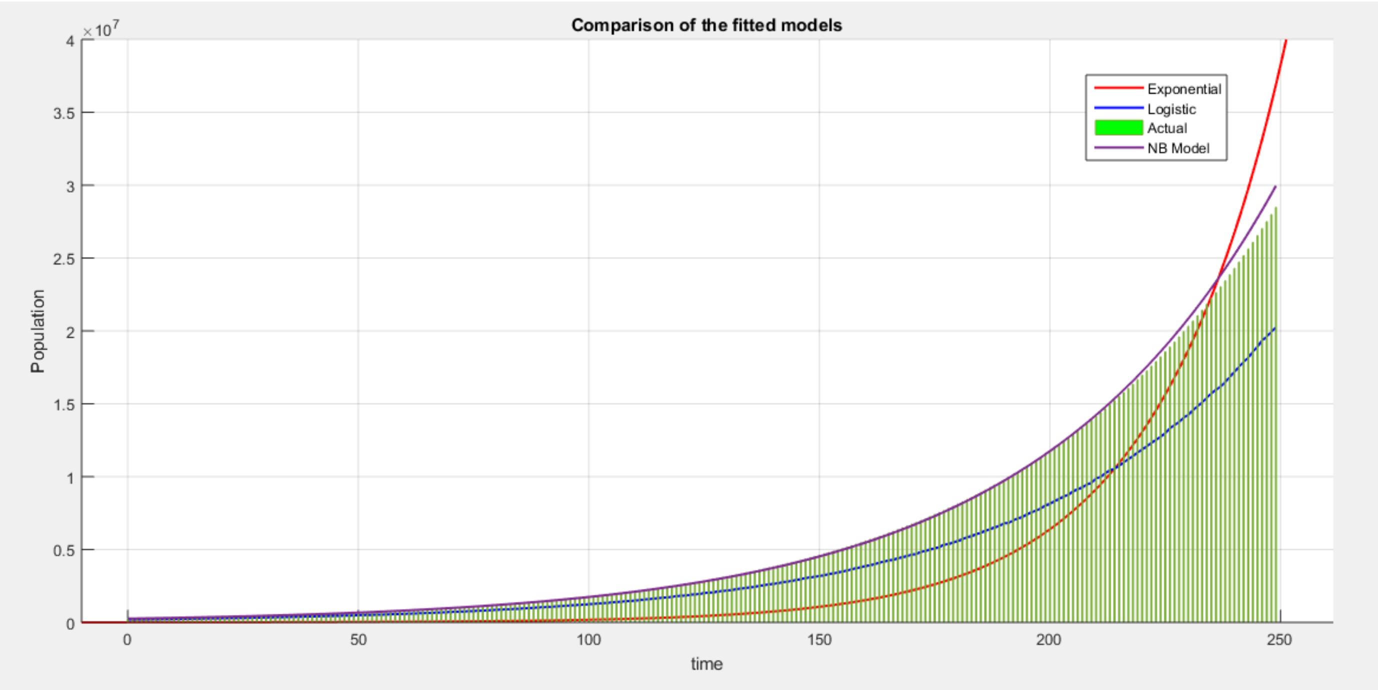


Figure 4.1: Comparison of the fitted models

4.3 Finding the determinants of population size

4.3.1 Poisson Model

The response variable **Population** was regressed on the other variables to ascertain those that have influence on this variable. The study first observed how the response variable is distributed. Figure 4.2 shows its distribution. It can be observed from the figure that the distribution is skewed to the right (positive skewness) as it is generally the case of most count variables.

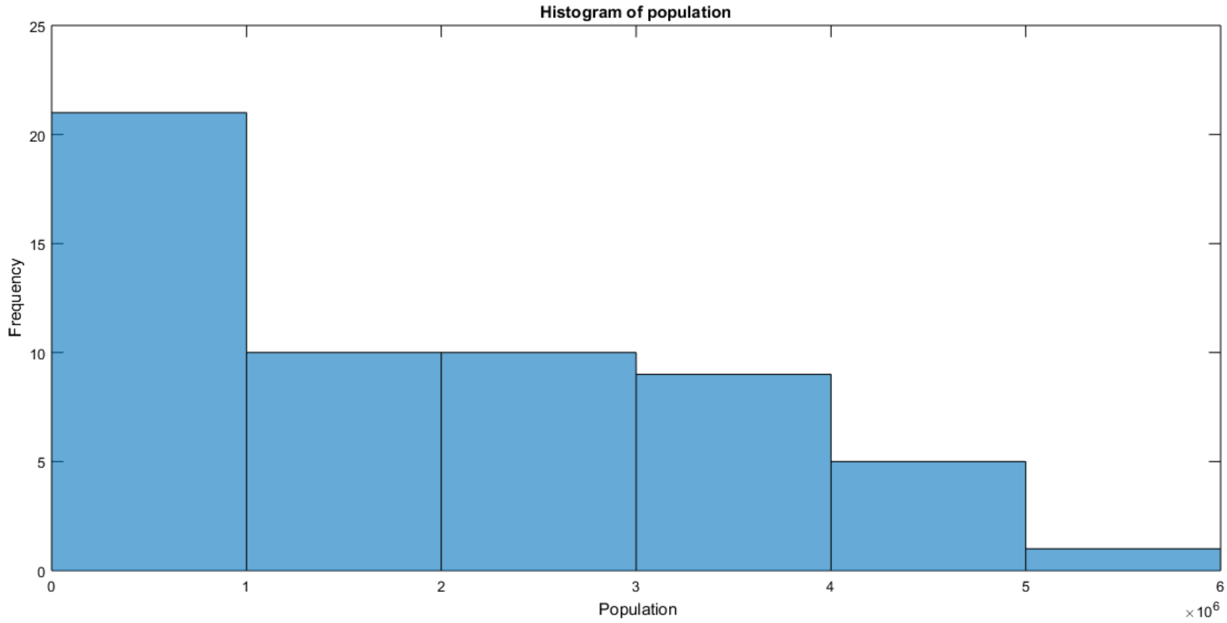


Figure 4.2: Histogram of the Population

The Poisson Model was fitted to the data in R and the results are shown in Table 4.4 . All the covariates we considered were highly significant in this model, as portrayed by their respective P-values. Out of the 11 covariates, 5 of them have negative influence on the Population. Those variables are the Fertility Rate, Literacy Rate, Unemployed, Poverty Level and Household Size. The rest of the covariates, however have positive influence on the response variable (Population). This indicates that all the covariates have an effect on the population size of the GAR under this model.

Table 4.4: Table of results for the fitted Poisson Model

	Estimate	**Standard error**	**P-value**
(Intercept)	9.414×10^{00}	3.806×10^{-02}	2.000×10^{-16}
FertilityRate	-1.165×10^{-02}	3.301×10^{-04}	2.000×10^{-16}
LiteracyRate	-3.182×10^{-02}	3.558×10^{-04}	2.000×10^{-16}
Unemployed	-3.452×10^{-02}	4.991×10^{-04}	2.000×10^{-16}
PopDensity	1.574×10^{-03}	8.073×10^{-06}	2.000×10^{-16}
PovertyLevel	-5.623×10^{-03}	2.203×10^{-04}	2.000×10^{-16}
BirthRate	8.833×10^{-02}	3.115×10^{-04}	2.000×10^{-16}
DeathRate	4.393×10^{-03}	1.473×10^{-04}	2.000×10^{-16}
Immigrants	3.398×10^{-07}	1.673×10^{-09}	2.000×10^{-16}
Emigrants	4.470×10^{-07}	4.241×10^{-09}	2.000×10^{-16}
LifeExpectancy	4.959×10^{-02}	3.390×10^{-04}	2.000×10^{-16}
HouseholdSize	-5.330×10^{-03}	7.587×10^{-04}	2.130×10^{-12}

Table 4.5: The fitted Poisson Model assessment

AIC	167401
Null deviance	52148620 on 55 degrees of freedom
Residual deviance	166477 on 44 degrees of freedom
Dispersion statistic	3742.795

4.3.2 Negative Binomial Model

Before this model was fitted, the study had to be sure of overdispersion based on the Poisson Model fitted. The Pearson dispersion statistic obtained for the Poisson Model was 3742.795, which is far greater than 1. This indicated that the Poisson Model we fitted was highly overdispersed and hence, care must be taking in using the results for inferences. The dispersion statistic was expected to be 1.0 or very close to it. Looking at how surprisingly all the covariates considered under the Poisson Model came out significant, was just a confirmation from the results from our first objective on the effects of overdispersion. This is obvious that the overdispersion in the data under the Poisson Model overstated the significance of the covariates.

We therefore moved ahead to fit the Negative Binomial Model to the data and the results are presented in Table 4.6. This model saw a drastic improvement over the Poisson Model. Among the covariates considered, only 4 of them were significant in this model. Among the 4 variables, only the variable Unemployed has a negative relationship with the response variable; the rest have a positive influence. Therefore, based on this model, only 4 out of the 11 covariates are most important and require attention.

In Table 4.7, the model assessment values are presented. We observe a great reduction in the null deviance by just the addition of 4 covariates. The dispersion statistic for this model is 1.082208, which is almost 1.0 and indicates that the overdispersion in the data is adequately catered for. Mean while, in correcting the overdispersion in the data, this model saw only 4 of the covariates significant. This is an indication that, the Poisson Model which failed to correct the overdispersion in the data, underestimated the standard errors and overstated the significance of 7 covariates. Adverse consequences can arise in using such results for inferences.

Table 4.6: Table of results for the fitted Negative Binomial Model

	Estimate	**Standard error**	**P-value**
(Intercept)	1.202×10^{01}	1.989×10^{-01}	2.00×10^{-16}
Unemployed	-1.469×10^{-01}	2.543×10^{-02}	7.58×10^{-09}
PopDensity	1.626×10^{-03}	4.073×10^{-04}	6.57×10^{-05}
BirthRate	5.978×10^{-02}	1.452×10^{-02}	3.84×10^{-05}
Immigrants	5.502×10^{-07}	7.881×10^{-08}	2.94×10^{-12}

Table 4.7: The fitted Negative Binomial Model assessment

AIC	1439.4
Null deviance	10259.397 on 55 degrees of freedom
Residual deviance	56.025 on 51 degrees of freedom
Dispersion statistic	1.082208

The Logistic Growth Model was additionally fitted on the GAR data and compared its fit to the two count models already fitted. Figure 4.3 shows the comparison of the three models on the data. The Negative Binomial Model fitted the data better than the rest as can be visualised in the figure.

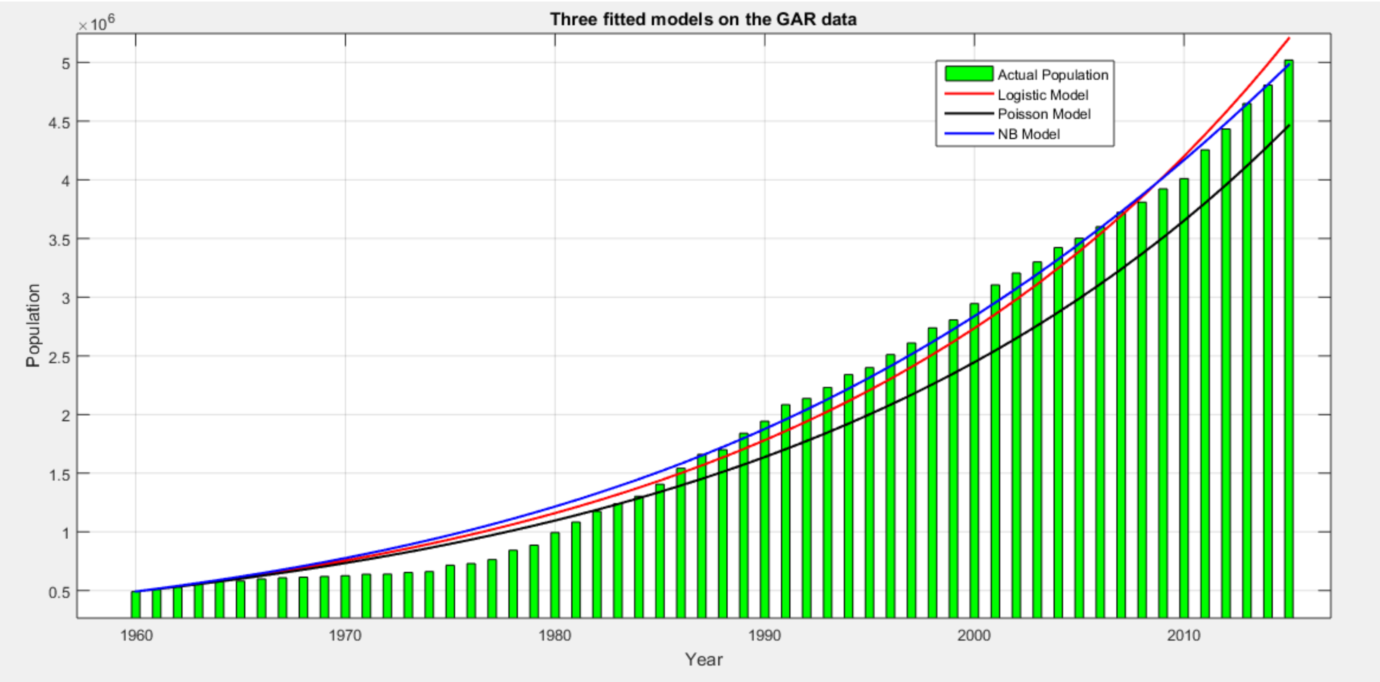


Figure 4.3: Comparison of the Logistic Growth, Poisson the and Negative Binomial on the GAR data

Table 4.8 presents the Mean Absolute Percentage Error (MAPE) values of the three models. Lower MAPE values indicate a better fit. The Negative Binomial Model has the least value among the three models.

Table 4.8: MAPE values for the fitted models

Model	MAPE
Poisson	18.02%
Logistic Growth	13.07%
Negative Binomial	12.25%

4.4 Chapter Summary

The results from the models been applied to both simulated and real dataset have been presented in this Chapter and the results discussed. We have seen the importance of treating overdispersion in modeling. In the next Chapter, we summarize the results and findings of this study and make recommendations.

Chapter 5

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Summary of the Results and Findings

Vital and accurate information about the human population go a long way in assisting policy makers in coming up with sound policies to make the lives of humans comfortable. In the absence of such information about the population, conditions necessary to meet the ever increasing demand associated with population growth might not be met. The Exponential and Logistic Growth Models are almost always used in demographic analysis to determine growth rate and predict population vales. However, these models do not account for overdispersion and they depend on just the time variable. In a complex ecosystem like the human environment, one would expect several variables to have influence on the population. To this effect, we investigated the effects of overdispersion in count data, compared the growth models to the Negative Binomial Model and determined variables that have influence on the population values.

To investigate the effects of overdispersion, the study simulated a dataset that exhibits overdispersion and fitted the Poisson and Negative Binomial Models to the data. The simulation results revealed that the Poisson Model, which does not account for overdispersion understated the standard errors and overstated the significance of some covariates. Two variables were made insignificant in the simulation process. However, the Poisson Model fitted these variables to be significance. Meanwhile, the Negative Binomial Model, which takes into account overdispersion, fitted correctly the variables to be statistically insignificant. This showed that when overdispersion is present and overlooked, inferences made from such results might be misleading.

Since the Negative Binomial Model can correct overdispersion in a data and have the tendency of incorporating more than one covariate in the model, we compared this model to the two population growth models. We simulated dataset using both growth models without the Negative Binomial Model. Then all three models were fitted to the simulated dataset. Interestingly, even though the Negative Binomial Model was not involved in the data generating process, it fitted the data better than the growth models. This implies that the Negative Binomial Model can be used in predicting population values better compared to the Exponential Growth Model performed worst in fitting the data.

In our quest to find the factors that affect the population size, we first regressed the population size on the covariates of the Greater Accra Region (GAR) real data using the Poisson Model. 11 covariates were considered and each one of them was statistically significant. Thus, based on this model, all the covariates have an influence on the population size. However, as with most count data, the data exhibited overdispersion, where the variability in the data was higher than was expected under the Poisson Model. The Pearson dispersion statistic was used to assess the extra variability under this model. In the absence of overdispersion, we expect the dispersion statistic to be 1 or close to 1. This statistic for the Poisson Model was 3742.795, which was extremely greater than 1.

To account for the overdispersion in the data, this study employed the Negative Binomial Model, which is almost always used to model an overdispersed Poisson Model. Under the second model, which reported a dispersion statistic of 1.082208, the overdispersion was well catered for. Meanwhile, only 4 out of the 11 covariates were statistically significant under this model. This meant that the Poisson Model overstated the significance of 7 variables, which is a confirmation that when overdispersion is overlooked, then erroneous conclusion from such results might be inevitable.

5.2 Conclusion

The study has revealed that overdispersion is very key when modeling count data. Most of the variables involved in demographic analysis are count variables. There is almost always the issue of overdispersion when modeling count data. In fact, it would be wise to assume the presence of overdispersion than its absence when modeling count data. This study has revealed that when overdispersion does exist, some variables might appear statistically significant, which would have appeared not significant in the absence of overdispersion. Therefore, it is very important to consider and treat overdispersion when modeling count data.

This study has also revealed the importance of the Negative Binomial Model in predicting population values. The model must be considered in modeling human populations. It has the capability of involving any possible relevant variable in the modeling process, aside correcting overdispersion in data. Applying this model on the GAR dataset revealed the variables Unemployed, Population Density, Birth Rate and Immigrants to have an effect on the population size of the region. The other variables were not significant under this model. This model performed better on the GAR than the rest.

5.3 Recommendation

Based on the study's results and findings, this study recommend the following;

It is important to consider overdispersion in demography and therefore models that are capable of addressing overdispersion must be employed. All information about human population, in one way or the other influence policy making. It is therefore paramount to ensure that the right tools and models are used to extract such information.

The study recommends the Negative Binomial Model to all stakeholders to be used when modeling population growth. As seen in this study, the Poisson Model is weak when dealing with overdispersion. Extra care must be taking when using such a model in analysis.

Important variables have been revealed in this study to have an effect on the population values of GAR. Authorities of the region can pay more attention to those variables so as to be able to have more control of the population of the region.

References

Albert, J. H. and Pepple, P. A. (1989). A bayesian approach to some overdispersion models. *Canadian Journal of Statistics*, 17(3):333–344.

Ausubel, J. H., Meyer, P. S., and Wernick, I. K. (2001). Death and the human environment: the united states in the 20th century. *Technology in Society*, 23(2):131–146.

Banks, R. B. (2013). *Growth and diffusion phenomena: Mathematical frameworks and applications*, volume 14. Springer Science & Business Media.

Birdsall, N., Kelley, A. C., Sinding, S. W., Sinding, S., et al. (2001). *Population matters: demographic change, economic growth, and poverty in the developing world*. Oxford University Press.

Bloom, D. E., Canning, D., and Malaney, P. N. (2000). Population dynamics and economic growth in asia. *Population and development review*, 26:257–290.

Breslow, N. E. (1984). Extra-poisson variation in log-linear models. *Applied statistics*, pages 38–44.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

Collings, B. J. and Margolin, B. H. (1985). Testing goodness of fit for the poisson assumption when observations are not identically distributed. *Journal of the American Statistical Association*, 80(390):411–418.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika*, 70(1):269–274.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, pages 829–844.

Crenshaw, E. M., Ameen, A. Z., and Christenson, M. (1997). Population dynamics and economic development: Age-specific population growth rates and economic growth in developing countries, 1965 to 1990. *American Sociological Review*, pages 974–984.

Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in poisson regression models. *Journal of the American Statistical Association*, 84(406):467–472.

Edwards, W. and Weizi, C. (2011). Population limiting factors. *Nature Education Knowledge*, 3(10):1.

Efron, B. (1989). *Overdispersion Estimates Based on the Method of Asymmetric Maximum Likelihood*. Stanford University. Division of Biostatistics.

Factbook, C. (2017). The world factbook; 2010. *See also: http://www cia gov/library/publications/the-world-factbook, accessed October*, 30.

Fund, U. N. P. (1998). *The State of World Population 1998: The New Generations*. United Nations Publications.

Group, W. B. (2013). *World Development Indicators 2013*. World Bank Publications.

GSS (2014). *2010 Population & Housing Census: District Analytical Report: Agotime Ziope District*. Ghana statistical service.

Hens, N., Aerts, M., Molenberghs, G., Thijs, H., and Verbeke, G. (2005). Kernel weighted influence measures. *Computational statistics & data analysis*, 48(3):467–487.

Hilbe, J. M. (2014). *Modeling Count Data*. Cambridge University Press.

Hirschman, C. (2004). The origins and demise of the concept of race. *Population and development review*, 30(3):385–415.

Iannario, M. (2014). Modelling uncertainty and overdispersion in ordinal data. *Communications in Statistics-Theory and Methods*, 43(4):771–786.

Islam, R. (2009). Mathematical modeling of age specific marital fertility rates of bangladesh. *Research Journal of Mathematics and Statictics*, 1(1):19–22.

Kulkarni, S., Kulkarni, S., and Patil, S. (2014). Analysis of population growth of india and estimation for future. *International Journal of Innovative Research in Science, Engineering and Technology*, 3(9):843–850.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Landau, R., Rosenberg, N., and of Engineering, N. A. (1986). *The positive sum strategy: Harnessing technology for economic growth*. National Academies Press.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, 15(3):209–225.

Ledder, G. (2005). *Differential equations: a modeling approach*. McGraw-Hill Higher Education.

Lee, A. H., Stevenson, M. R., Wang, K., and Yau, K. K. (2002). Modeling young driver motor vehicle crashes: data with extra zeros. *Accident Analysis & Prevention*, 34(4):515–521.

Lewis, C. D. (1982). *Industrial and business forecasting methods: A practical guide to exponential smoothing and curve fitting*. Butterworth-Heinemann.

Malthus, T. R. (1888). *An essay on the principle of population: or, A view of its past and present effects on human happiness*. Reeves & Turner.

May, J. F. (2005). Population policy. In *Handbook of population*, pages 827–852. Springer.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

McIntosh, C. A. and Finkle, J. L. (1995). The cairo conference on population and development: A new paradigm? *Population and development review*, pages 223–260.

Meyer, P. S. and Ausubel, J. H. (1999). Carrying capacity: a model with logistically varying limits. *Technological forecasting and social change*, 61(3):209–214.

Meyer, P. S., Yung, J. W., and Ausubel, J. H. (1999). A primer on logistic growth and substitution: the mathematics of the loglet lab software. *Technological Forecasting and Social Change*, 61(3):247–271.

Mohammed, M. and Laney, D. (2006). Overdispersion in health care performance data: Laney's approach. *BMJ Quality & Safety*, 15(5):383–384.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365.

NDPC, G. (2003). *Ghana poverty reduction strategy, 2003-2005: an agenda for growth and prosperity*, volume 1. Government of Ghana, National Development Planning Commission.

Ofori, T., Ephraim, L., and Nyarko, F. (2013). Mathematical model of ghana's population growth. *International Journal of Modern Management Sciences*, 2(2):57–66.

Otani, I. and Villanueva, D. (1990). Long-term growth in developing countries and its determinants: An empirical analysis. *World Development*, 18(6):769–783.

Paul, S. and Plackett, R. (1978). Inference sensitivity for poisson mixtures. *Biometrika*, 65(3):591–602.

Paul, S. S. (1994). Report on the international conference on population and development. *Ageing International*, 21(4):49–51.

Pearl, R. and Reed, L. J. (1920). On the rate of growth of the population of the united states since 1790 and its mathematical representation. *Proceedings of the national academy of sciences*, 6(6):275–288.

Rasmussen, R. O., Ensign, P., and Huskey, L. (2016). The forecasting of remote area populations: Numbers aren't everything. In *Demography at the Edge*, pages 65–78. Routledge.

Simon, H. A. (1972). Theories of bounded rationality. *JSPMAD*, 1(1):161–176.

Thirlwall, A. P. (1989). *Growth and development: with special reference to developing economies*. Springer.

Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. *Corresp. Math. Phys.*, 10:113–126.

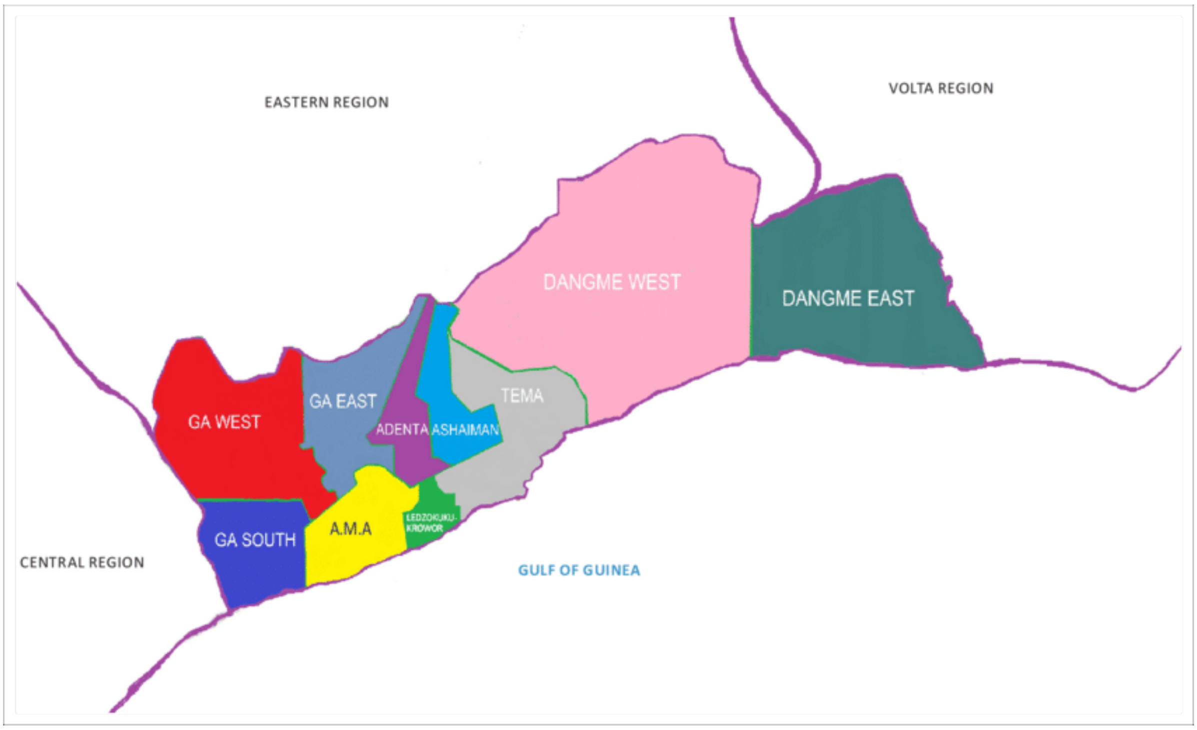
Wilmoth, J. R. (2000). Demography of longevity: past, present, and future trends. *Experimental gerontology*, 35(9-10):1111–1129.

Wilson, K. and Hardy, I. C. (2002). Statistical analysis of sex ratios: an introduction. *Sex ratios: concepts and research methods*, 1:48–92.

Winkelmann, R. (2000). Seemingly unrelated negative binomial regression. *Oxford Bulletin of Economics and Statistics*, 62(4):553–560.

Appendix I: Map of Ghana

Appendix II: Map of the Greater Accra Region



Appendix III: R code for simulating data and fitting the Poisson and NB models

```
library(COUNT)
library(lmtest)
library(rcompanion)

# Generating data

set.seed(444)
n=200

#The covariates
x2 <- rnorm(n)
x3 <- runif(n)
x4 <- rnorm(n)
x5 <- runif(n)
x6 <- rnorm(n)

# Setting Parameter values

b1=5
b2=1
b3=0
b4=-1
b5=0
b6=-1

#Getting mu
mu <- exp(b1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6)

#Generating the response values y
y <- rpois(n, mu)

#Putting all together in data form
Cdata <- as.data.frame(cbind(y, x2, x3, x4, x5, x6))
names(Cdata) <- c("y", "x2", "x3", "x4", "x5", "x6")

#Fitting the Poisson model
fm_pois <- glm(y ~ x2 + x3 + x4 + x5 + x6, data = Cdata, family = poisson)
summary(fm_pois)
P__disp(fm_pois)

#Fitting the NB model
nb_glm <- glm.nb(y ~ x2 + x3 + x4 + x5 + x6, data = Cdata)
summary(nb_glm)
P__disp
```

Appendix IV: The GAR data

Year	population	Population	FertilityRate	LiteracyRate	Unemployed	PopDensity	PovertyLevel	BirthRate	DeathRate	Immigrants	Emigrants	LifeExpectancy	HouseholdSize
1960	0.4918	491800	7.25	55.0	1.3	215.5	1.6	15.5	3	14945	9321	77.60	2.3
1961	0.5119	511900	7.23	55.2	1.4	221.2	1.7	15.8	3	15045	10341	77.50	2.3
1962	0.5276	527600	7.20	55.2	1.4	229.2	1.8	16.2	4	16034	11031	77.00	2.4
1963	0.5523	552300	7.10	55.3	1.5	236.6	1.7	16.5	3	17800	12821	76.50	2.4
1964	0.5789	578900	7.03	55.3	1.5	241.8	1.8	16.9	3	18341	12989	76.40	2.8
1965	0.5808	580800	7.05	55.5	1.6	256.7	1.9	17.3	3	19450	13200	76.40	2.5
1966	0.5988	598800	7.00	55.6	1.7	269.1	2.0	17.6	4	22340	13567	76.20	2.7
1967	0.6089	608900	6.30	55.7	1.7	277.2	2.2	18.0	4	25782	13721	76.10	2.5
1968	0.6134	613400	6.62	55.7	1.8	281.7	2.1	18.2	5	28502	13928	75.90	2.6
1969	0.6199	619900	6.00	55.8	1.8	293.4	2.3	18.4	4	29003	14328	75.87	2.9
1970	0.6267	626700	6.70	55.8	1.9	303.8	2.6	18.7	4	34001	15761	75.15	2.7
1971	0.6389	638900	6.69	55.8	2.0	315.5	2.2	19.2	3	44098	18987	75.10	2.8
1972	0.6401	640100	6.60	55.9	2.1	328.9	2.4	19.5	3	53004	21109	74.95	2.9
1973	0.6543	654300	6.55	55.0	2.0	342.2	2.7	19.7	4	66210	22002	74.93	2.8
1974	0.6623	662300	6.53	55.9	2.2	367.5	2.5	19.9	5	76000	23412	74.72	2.9
1975	0.7156	715600	6.52	56.0	2.3	384.0	2.6	20.1	4	81200	24211	74.69	2.8
1976	0.7298	729800	6.57	56.0	2.3	391.6	2.9	20.3	3	85678	25002	74.55	2.8
1977	0.7645	764500	6.54	56.0	2.4	408.8	3.6	20.5	3	89700	28111	74.49	2.7
1978	0.8434	843400	6.45	56.2	2.5	428.1	3.3	20.8	3	90342	30456	74.00	2.7
1979	0.8867	886700	6.44	56.3	2.5	442.8	3.0	21.0	3	120343	53090	73.70	2.9
1980	0.9943	994300	6.42	56.5	2.5	467.8	3.1	21.3	4	150089	64377	73.40	2.8
1981	1.0834	1083400	6.37	56.7	2.5	489.4	3.5	21.6	5	190334	75211	73.20	2.9
1982	1.1745	1174500	6.35	56.6	2.6	509.4	3.2	21.8	5	280321	76089	73.10	2.8
1983	1.2408	1240800	6.40	56.8	2.7	531.4	3.3	22.1	5	307009	83511	73.00	2.9
1984	1.3034	1303400	6.32	56.9	2.8	558.2	3.5	22.3	5	389112	99661	72.90	2.9
1985	1.4056	1405600	6.30	56.8	2.8	577.9	3.7	22.5	3	418004	112701	72.80	3.2
1986	1.5423	1542300	6.46	57.0	2.9	598.5	3.9	22.7	3	500459	117303	72.70	3.2
1987	1.6608	1660800	6.42	57.0	3.0	614.0	4.1	23.2	3	600002	119621	72.60	3.3
1988	1.7004	1700400	6.40	57.2	3.2	638.6	4.3	23.4	4	723000	120403	72.50	3.5
1989	1.8409	1840900	6.10	57.4	3.3	666.9	4.0	23.6	4	821456	124980	72.40	3.4
1990	1.9432	1943200	6.07	57.6	3.4	681.7	4.2	23.9	4	921590	127711	72.30	3.3
1991	2.0843	2084300	5.80	57.8	3.6	696.5	4.6	24.1	4	930004	129451	72.20	3.0
1992	2.1367	2136700	5.82	57.8	3.7	712.2	4.8	24.3	3	938056	132189	72.10	3.0
1993	2.2312	2231200	5.70	58.0	3.9	738.8	5.0	24.6	3	943213	138376	72.00	3.1
1994	2.3409	2340900	5.67	58.3	4.0	762.0	5.3	24.8	3	956007	238098	71.70	3.5
1995	2.4003	2400300	5.62	59.6	4.2	781.9	6.0	25.0	3	967336	241700	71.50	3.4
1996	2.5111	2511100	5.50	58.9	4.3	804.2	6.1	25.4	4	980206	245672	71.00	3.3
1997	2.6098	2609800	5.10	59.1	4.6	828.7	6.3	25.7	3	990458	251056	70.50	3.2
1998	2.7389	2738900	4.40	59.5	5.0	852.6	5.1	26.0	4	1000500	253098	70.00	3.1
1999	2.8065	2806500	3.91	59.7	5.2	871.2	5.2	26.3	3	1010040	256002	69.70	3.2
2000	2.9457	2945700	3.93	59.8	5.6	895.5	5.9	26.5	4	1017004	258074	69.50	3.3
2001	3.1045	3104500	3.95	60.2	5.5	930.6	6.2	26.8	4	1025600	262456	69.30	3.6
2002	3.2052	3205200	3.98	60.5	5.7	960.3	7.4	27.0	4	1037605	267908	68.50	3.5
2003	3.3005	3300500	3.95	60.9	6.1	996.0	8.5	27.5	4	1054000	269321	67.00	3.4
2004	3.4223	3422300	3.92	61.4	6.4	1025.2	9.6	27.8	6	1090345	311098	66.40	3.5
2005	3.5022	3502200	3.97	62.5	6.8	1060.0	10.4	28.1	4	1145900	375008	65.70	3.6
2006	3.6009	3600900	3.98	63.1	7.0	1078.7	11.8	28.3	5	1200006	377432	65.20	3.6
2007	3.7234	3723400	3.99	63.6	7.8	1128.3	12.1	28.5	5	1258790	381032	64.60	3.9
2008	3.8098	3809800	3.40	64.4	8.6	1170.0	12.5	28.7	5	1278000	384008	64.00	3.6
2009	3.9234	3923400	3.50	64.7	8.4	1200.0	12.7	29.1	5	1323000	389245	63.40	3.7
2010	4.0101	4010100	3.57	64.9	8.2	1235.8	12.9	29.3	4	1350067	392761	63.10	3.8
2011	4.2555	4255500	3.56	65.1	8.5	1270.1	13.4	29.5	6	1390120	395900	62.50	3.8
2012	4.4333	4433300	3.58	65.3	8.7	1293.8	13.7	29.9	6	1420300	484067	62.00	3.7
2013	4.6502	4650200	3.57	67.0	9.2	1320.4	14.1	30.0	5	1450000	496000	61.29	3.8
2014	4.8072	4807200	3.59	67.4	9.5	1346.9	14.5	30.3	6	1490150	512304	61.35	3.9
2015	5.0202	5020200	3.60	68.0	10.7	1378.5	15.2	30.5	7	1500200	526030	61.49	4.0