KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI

MODELLING COUNT OUTCOMES FROM DENTAL CARIES IN ADULTS: A COMPARISON OF COMPETING STATISTICAL MODELS

By

ADETI, FELIX

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF M.PHIL MATHEMATICAL STATISTICS

October, 2016

Declaration

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

Felix Adeti ....

Student(20288680) Certified by: Signature Date

Dr. R.K.Avuglah ....

Supervisor Signature Date

Certified by:

Dr. R.K. Avuglah ....

Head of Department Signature Date

Dedication

To the memory of my father, Mr E.B.A. Adeti.

Abstract

The objectives of this research are to identify the key risk factors contributing to dental caries in adults and to determine an appropriate model suitable for the analysis of Decayed, Missing and Filled Teeth Data (DMFTDATA). The study was conducted between June, 2015 and January, 2016 in Sefwi Wiawso District Hospital, Sefwi Wiawso, Ghana, and it involved a systematic random samples of 1158. The average age of participants was about 40 years with 54% of them being females and the rest 46% being males. Count data models such as Poisson, negative binomial models, zero inflated Poisson (ZIP), zero inflated negative binomial (ZINB) and Poisson hurdle (PHURDLE) models were fitted to the DMFTDATA and compared using the Vuong test statistic (V) and Akaike information criterion (AIC). ZINB was found to be the appropriate model for analysing DMFTDATA characterised with overdispersion and contains 28.8% of zeros. It was found that age (in years), jaw accident, frequency in sweet consumption, rinsing habit after meal with water and frequency of brushing were significant risk factors of dental caries in adults. It is therefore recommended that adults should endeavour to brush their teeth at least twice daily and rinse their mouth with water after meal. Also proper dental care should be considered for adults with jaw accident and adults should reduce intake of sugar foods.

Acknowledgements

I would like to express my sincere appreciation and thanks to my supervisor Dr. R.K. Avuglah for his patience, motivation and knowledge during the time of this study. I am equally grateful to Mr. E. Harris for his immerse contribution to this research work.

Special gratitude goes to my wife Mrs. Juliana Adeti of Ghana Armed Forces for her love, time, patience, support and prayers. To my family, thanks for everything and to mathematics department of Sefwi Wiawso Senior High Technical School, your support is acknowledged.

To the entire staff of Sefwi Wiawso Dental Clinic, your immense support and contributions are appreciated.

My gratitude also goes to Miss Francisca Aba Acquah for her assistance.

Contents

List of Abbreviation

WHO	World Health Organisation
DMFT	Decayed, Missing and Filled Teeth
DMFTDATA	Decayed, Missing and Filled Teeth Data
NB	Negative Binomial
ZIP	Zero Inflated Poisson
ZINB	Zero Inflated Negative Binomial
PHURDLE	Poisson Hurdle
AIC	Akaike Information Criterion
D	Decayed due to caries
M	Missing due to caries
F	Filled due to caries
SPSS	Statistical Package for the Social Sciences
R	Statistical Computing Software
GLM(s)	Generalized Linear Model(s)
V	Vuong Test Statistic

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Introduction

This chapter comprises the background of the study, statement of problem, objectives of the study, Methodology, Significance of the the Research work, research questions and organisation of the thesis.

1.2 Background of the Study

Oral disease is one of the widespread diseases in modern society of which dental caries is one of such main oral diseases. Apart from common cold, dental caries is highly known among these common disorders. Generally, dental caries happens in children and adults, however, it can involve any individual and it is a known reason for tooth loss in individuals.

The branch of public health that concerns itself with the distribution and intensity of oral diseases like dental caries on individuals is the oral epidemiology. Gathering information with the purpose of detailing the amount of dental caries in an individual is of great significance in areas of funding, planning and delivery of such services so that adequate professionals of the right expertise are trained, adequate number of clinical centres are put up. New and advanced clinical techniques are also developed by means of sufficiently funded research and other programs.

A crucial problem for oral health is access to care. Therefore having knowledge of the need for care assists planners to make preventive programs such as water fluoridation, beverages and accessibility of low sugar foods to avoid this disease in the first place and to make sure an adequate amount of dentists are trained to provide care services and these care services are sufficiently financed.

Oral epidemiology on the other hand assesses disease-preventing interventions intended for controlling diseases and evaluate the efficacy and quality of interventions of instituted oral disease-preventing programs.

In spite of substantial advances in preventive dentistry, an immeasurable majority of the population will have incidences of dental decay in their lives. Advances in the prevention and treatment of dental caries imply that majority of children will preserve their teeth into adulthood, and these teeth will depend on how well they are cared for over their lifetime. The very important role oral health makes in our daily lives, in terms of speaking, eating, smiling and socialising, is regularly overlooked and consequently leads to dental caries and other oral diseases.

According to Arens (1999), dental caries arises as a consequence of demineralization of dentine and enamel by means of organic acids generated through dental plaque containing bacteria during sugar anaerobic metabolism obtained by eating food.

Bacteria usually reside in the mouth. These bacteria change foods particularly sugar and starch into acids. Food debris, acid, bacteria and saliva mix in the mouth to form an adhesive substance called plaque which sticks on the teeth. The plaque commences to build up on the teeth by 20 minutes immediately after eating which is the time that bacterial activities occur. Plaque is usually prominent on the molars, at the edges of fillings just and above the gum on all teeth. The plaque which is not gotten rid off from the teeth then mineralizes to form tartar. These tartar and plaque irritate the gums which later result into gingivitis and finally developing into periodontitis. Also, if the plaque is not taken off carefully and also regularly, cavities will not only start, but thrive.

Acids that are found in plaque soften the enamel of the tooth and make holes in the tooth that are usually called cavities. The cavities are typically painless until they develop big holes which have an effect on nerves and also cause tooth fissure. If cavities remain not treated, tooth abscess will be able to grow. Tooth decay that are left untreated break down the internal tooth structures and eventually resulting in tooth loss.

The risk of tooth decay is increased by carbohydrates such as starches and sugars. Adhesive foods are riskier than non-advesive foods because adhesive foods stay on the teeth’s surface. Also adding to the time that acids stay on the tooth’s surface is frequent snacking.

Signs and symptoms of tooth decay may differ, depending on the tooth decay’s degree and position. At the start of tooth decay, there may be no symptoms. With time, when the tooth decay develops bigger, there may be signs and symptoms like:

- tooth sensitivity
- painful feelings while biting down
- toothache
- observable pits or holes in the teeth
- feel mild to sharp ache while drinking or eating something hot, cold or sweet
- black, white or brown stains on any tooth’s surface
- foods are trapped regularly between teeth
- discomfort in your mouth
- bad breath

Potential complications of dental caries comprise the following

- tooth abscess
- pain
- tooth sensitivity
- breaking tooth

- incapability to bite down food on tooth

The dentist is competent to notify you of the three types of tooth decay you may have. Namely, root, pit and fissure and smooth surface. Generally, the dentist can detect tooth decay simply by:

- asking questions concerning tooth sensitivity and pain
- checking the teeth and mouth
- probing the teeth to check for soft areas with dental instruments
- observe dental X-rays to enable him give an idea about the extent of cavities and decay

Every person with teeth may be in danger of getting tooth decay but the danger can increase with the following factors:

- Location of tooth: Most tooth decay often occurs on the molars and premolars, that is the back teeth. These molars and premolars have plenty of pits, grooves and crannies which gather food substances. Consequently, these molars and premolars are difficult to maintain clean as compared to the front teeth which are easier to reach and smoother. When plaques assemble, bacteria are capable of thriving between the molars and premolars, producing acid which can tear down enamel of the tooth.
- Some type of foods and drinks can stick to the teeth over a long period time. These foods, for instance, cake, hard candy, honey, milk, dry cereal, soda, dried fruit, sugar, breath mints, cookies, ice cream, and chips are highly probable to cause tooth decay as compared to foods which are simply washed away.
- Regular sipping or snacking: Sip or snack of sodas increasingly provide the bacteria in the mouth more grounds to produce acids which will attack the teeth and subsequently wear the teeth down. Frequently sipping of soda and acidic drinks every day create a persistent acid bath over the teeth.

- Poor brushing : When the teeth is not cleaned almost immediately after drinking or eating, plaques are formed and the first phase of dental caries may start.

- Inadequate fluoride: A naturally occurring mineral that helps in the prevention of dental caries and may overturn the initial stages of damage of tooth is fluoride. Fluoride is added to many public water supplies since it benefits the teeth. Fluoride is as well a common ingredient in mouth rinses and also in toothpaste.

- Children and adults: Cavities are widespread among children and as well as teenagers. Also, adults are in danger of cavities since a lot of them maintain their teeth as they grow. With time, there may be wear down of teeth, and gums may recede. This makes the teeth susceptible to root decay. Usage of medications by adults which decrease flow of saliva also increase their risk of cavities.

- Dry mouth: Lack of saliva which aids in avoiding dental caries by means of washing away food particles as well as plaques from the teeth is the cause of dry mouth. Certain materials that are in the saliva can help restore early tooth decay and also help counter the acid produced by bacteria. Possibility of getting cavities by reducing saliva production in the mouth can increase by use of certain medications, certain chemotherapy drugs, some medical conditions and radiation to the neck and head .

- Damaged dental devices and fillings: Dental fillings may deteriorate and start to break down and have irregular edges over time and this permits plaque to develop easily making it difficult to remove. Also, dental devices can stop fitting well and this allows decay to start beneath these dental devices.

- Disorders in eating: Bulimia and anorexia can result into significant erosion of the tooth as well as cavities. Recurring vomiting washes over the teeth and commences softening the enamel as a result from stomach acids. Eating disorders also can hamper with saliva production.

- Heartburn: Gastroesophageal reflux disease or heartburn is capable of causing stomach acid to move into the mouth and then wear away the enamel of the teeth resulting into significant damage of tooth. The dentist may advise doctor consultation to know whether gastric reflux is the reason for the loss of tooth enamel.

Highly developed tooth decay is usually painful. This may result in tooth loss. With no treatment of the tooth decay, bacteria may pass through to the tooth may later develop into an abscess under the gums which is a severe infection. Such severe infection may move seriously to other parts of the body causing fatal consequences infrequently.

It is possible for the dentist to give education on ways to prevent tooth decay and even to repair early stages in tooth decay due to advances in science. These advances in science which includes use of pastes, filling materials and rinses which contain calcium, phosphates and fluoride are generally referred to as remineralization. These remineralization materials provide backbone to the enamel of the tooth and can help repair the tooth by itself.

Remineralization may not always be successful like any other treatments. Patients who go according to the recommendations of the dentist closely regarding care are usually the most successful. The dentist may get rid of the cavity and restore the tooth for more advanced cavity. The dentist may also place a filling on the tooth if the affected area is small. In other cases, the dentist may have to place a crown over the remaining tooth when the decay damages the structure of the tooth extensively. The tooth must be removed by the dentist when not healthy tooth is left, in more severe cases.

The first step in preventing tooth decay is practising good dental hygiene. These includes brushing the teeth with a fluoridated toothpaste two times a day.

Drinking fluoridated water strengthen the enamel of the teeth. Sipping or snacking drinks that are high in sugar an acids should be reduced. In cases of pits and grooves the dentist should place a protective coating on the molars and premolars to help reduce places for bacteria growth.

Below are some tips to help prevent cavities by detailed good oral and dental hygiene:

- Mouth rinse: use The dentist may recommend usage of mouth rinse with fluoride if there is a high risk of developing cavities.

- After eating or drinking, brush with toothpaste containing fluoride: The teeth should be brushed at least twice a day preferably after every meal with toothpaste containing fluoride. Floss or use an interdental cleaner to clean between the teeth. At least rinse the mouth with water where it is not possible to brush after eating.

- Use dental sealants A sealant is a shielding plastic covering which is put on the chewing surface of molars and premolars that closes up the crannies and grooves that collect food. The sealant shields enamel of the tooth from plaque and acid and it helps adults and children. Even though sealants need to be looked after regularly to ensure the sealants are still intact, it takes about 10 years for the sealants to be replaced.

- Visit dentist regularly: To help spot cavities early and prevent decay, visit dental care regularly for regular oral examinations from the dentist and get professional teeth cleanings.
- Drinking tap water: To help decrease cavities significantly, most public water supplies add fluoride. There is no fluoride benefits when drinking water which does not contain fluoride.

- Shun regular sipping and snacking Bacteria in the mouth create acids which can destroy the enamel of the tooth after drinking beverages and sodas. The teeth is under constant attack when there is snacking or sipping throughout the day.
- Fluoride treatments: When enough fluoride through fluoridated drinking water and other sources are not gotten, the dentist may suggest periodic fluoride treatments.
- Enquire on antibacterial treatments The dentist may suggest antibacterial mouth rinses and other treatments to help reduce harmful bacteria in the mouth when there is vulnerability to tooth decay because of some medical condition.
- Eat tooth-healthy foods: A number of foods and beverages are better for the teeth as compared to others. Foods and beverages which get trapped in grooves and pits of the teeth for a long time like candy or cookies, chips, etc should be avoided or brush the teeth almost immediately after taking such foods and beverages. Foods like fresh vegetables and fruits boost saliva flow.

Once dental caries occurs, it stays till lifetime. The early manifestation of the caries process is normally unseen from view in the fissures of the teeth and in between the teeth since the caries process is a small patch of enamel demineralisation on the surface of the tooth. Beneath the enamel, the damage spreads into the softer and most sensitive part of the tooth. There is a formation of cavity when the damaged enamel collapses and thereafter, the tooth is damaged gradually. Usually and more common in older adults is the recession of the gum which happens when dental caries attacks the roots of the teeth.

Dental caries is a major oral health problem, affecting a vast majority of adults and schoolchildren in most countries. Dental caries as a disease must be emphasized that it can only be controlled to a certain degree but not eradicated. James and Sheiham (n.d.) in their research observed that most policies, research programmes and surveys on dental caries have centred on children and regret that dental caries levels are rather very much advanced in adults than in children consuming even fluoridated water (Bernabe and Sheiham, 2014b) and fluoridated toothpaste (Broadbent et al., 2008; 2013). They therefore claimed that it is exceptionally significant that dental caries is emphasized in adults as adults caries accounts for about 80% of the dental care costs relating to caries compared with 20% for children.

Epidemiological studies involving dental caries are comparatively uncommon in adults according to Namal et al., (2008). During the last decades, however, studies have indicated that in the adult population of both developing and industrialized countries, major problem include dental caries (Luan et al., 2000; Treasure, et al., 2001). Diverse investigators in different parts of the world have assessed risk factors of dental caries among adults and the most regularly reported factor associated with a high number of dental caries in adults is age (Winn et al., 1996; Forshee & Storey, 2004). In some articles, women were reported to have higher incidence of dental caries than men (Splieth, et al., 2002; Lin, et al., 2001), while there were no significant differences in another study between men and women (Szoke & Petersen, 2004). Another reported factor that leads to increased dental caries is poor oral hygiene practices (Shah & Sundaram, 2004).

Several risk factors are associated with dental caries including injuries (trauma), sugar intake, gender and many more.

The need to assess dental caries status in an individual is essential. The DMFT (Decayed, Missing and Filled Teeth) index is a standard technique generally used for assessing dental caries incident in individuals. It consists of counts of decayed teeth due to caries, missing teeth due to caries and filled teeth due to caries in individuals. The tooth is recorded as D if there is the presence of a carious lesion(s) or both a restoration and carious lesion(s). When a recording of a tooth is M, then it has been removed due to caries. F is recorded for a tooth once a permanent or temporary tooth is filled due to caries. Whenever there is a presence of at least one permanent restoration and no persistent caries or primary caries in other area of the tooth, then the teeth are reported as filled without decay. Teeth are not recorded as F if they are restored for reasons other than dental caries as reported in Cappelli and Mobley (2007).

Dental caries affects the hard parts of the tooth of the human race and it is a progressive irreparable microbial disease. Even if the lesion of dental caries is treated, its symptoms persist throughout life once it occurs. Dental caries typically starts almost immediately the teeth erupted into the oral cavity hence it is referred to as a post eruptive disease. All genders, races, ages and socioeconomic groups are affected.

In 1982, the World Health Organization (WHO) at the World Health Assembly defined the suitable dental health level in areas of average decayed teeth due to caries (D), missing teeth due to caries (M) and filled teeth due to caries (FT) (DMFT) for adults in diverse age groups. They include: at age 18 years, 4 DMFT, for age 35–44 years, 6 DMFT and for age 65+ years, 12 DMFT. They recommended the use of DMFT index as a standard measure of dental caries status in individuals.

DMFTDATA (Decayed, Missing and Filled Data) as a count data is a major oral health problem in most countries with regards to this research. The modelling of DMFTDATA is of principal interest in several fields like epidemiology, public health, insurance, psychology and various research areas.

Statistical models to be considered for the modelling of DMFTDATA are the negative binomial (NB) model and Poisson model as one component models as well as Poisson Hurdle model, zero inflated negative binomial model and zero inflated Poisson model, as two component models to analyse the DMFTDATA. Over the years, the application of zero-inflated regression models by modellers has gained reputation. Literature has it that Zero-inflated Negative Binomial (ZINB), Poisson Hurdle (PHURDLE) and Zero-inflated Poisson (ZIP) models have been mainly used when outcome variable is characterized by a preponderance of zero counts. In other words Zero-inflated Poisson, Poisson Hurdle and Zero-inflated Negative Binomial models have been mostly applied when the data contain more zero counts than are accepted by Negative Binomial or Poisson distributions. The common reason for high percentage of zero counts is that, there is a second process that generates extra zeros in addition to the standard count data process.

Consider a patient visiting dental clinic for dental care. The patient might have zero dental caries because (i) he visited the dental clinic for a dental assessment though he has healthy teeth or (ii) he has no caries despite heightened symptoms of dental caries. The zeros product of (i) called 'structural' zero comes from a binary process. The zeros product of (ii) called 'incidental' zeros result from a count process which is subjected to the individual at risk. Zero-inflated count models are models known to allow for these two separate types of zeros. The most notable being the zero-inflated negative binomial, Poisson hurdle and zero-inflated Poisson models (Mullahy, 1986; Lambert, 1992).

In practice, overdispersion is common with count data having lots of zeros. The standard Poisson model offers a baseline for count data analysis. This model which is widely assumed for modelling the distribution of the observed count data is limited in applications because of its restrictive nature of equidispersion assumption. In practice, however, count data are usually with regards to the Poisson distribution, over-dispersed. One frequent manifestation of overdispersion is the incidence of zero counts especially when the zero counts are greater than expected under the Poisson distribution and these zero counts are of interest because such zero counts have special status. Though Negative Binomial comes in handy to model the data due to its ability to handle overdispersion, it fell short in its ability to handle two processes data. However, both models, Poisson and Negative Binomial, when extended can contain extra dispersion in the data which will end the violation of distributional properties of each respective distribution. The improved Poisson and Negative Binomial models can therefore be considered as a solution to a violation of distributional assumptions of the Poisson and Negative Binomial models. And to get a better fit, an over-dispersed model which can incorporate excess zeros can provide an alternative to Negative Binomial model and Poisson model according to Chipeta, et al., (2014).

One model capable of handling excess zero counts is the zero-inflated models like the Zero-inflated Negative Binomial model and the Zero-inflated Poisson model (Mullahy, 1986; Lambert, 1992). These models are two-component models combining “structural” zeros with a count distribution like the negative binomial or Poisson. That is, the zero counts can arise from both the count component and the “structural” component.

Many count data show more zero realizations or counts than Poisson model would allow, in addition to over-dispersion. The Poisson Hurdle model is one other type of model that is also competent in capturing both properties (overdispersion and excess zeroes), which was initially proposed by Mullahy (1968). The Poisson Hurdle is a two state model that has a zero-truncated component to predict positive or non-zero counts and binary model to predict zeros.

Application areas of zero-inflated regression models and Poisson Hurdle model by researchers vary and include recreational facilities usage (Shonkwiler & Shaw, 1996; Gurmu & Trivedi, 1996), abundance of species (Welsh et al., 1996), manufacturing defects (Lambert, 1992), road safety (Miaou, 1994), patent applications (Crepon & Duguet, 1997), medical consultations (Gurmu, 1997), political science (Zorn, 1996) and sexual behaviour (Heilbron, 1994).

Various model selection methods such AIC and the Voung statistic are used to determine the model that performs efficiently with respect to the DMFTDATA. The Vuong non-nested test is used as a comparison of the predicted probabilities of two models which are non-nested. Typical is the comparison of zero-inflated models such as zero-inflated negative-binomial and zero-inflated Poisson with their traditional models like negative-binomial and Poisson respectively. Indication of a large positive test statistic is the superiority of the first model over the second model whilst indication of a large negative test statistic is the superiority of the second model over the first model. The test statistic is asymptotically distributed standard normal under the null hypothesis that two models are indistinguishable.

Similarly, Akaike Information Criterion (AIC) is used to measure the comparative superiority of statistical models for a specified data set especially when comparing statistical models fitted by maximum likelihood to the same data usually for non-nested models. The statistic considers model parsimony and penalizes for the number of predictors used in the model hence $AIC = -2L+2k$ Basically, the first term represents the deviance and the second term is a penalty for number of parameters. A statistical model with the smallest AIC value is the best of the models involved.

The following steps can help to compare and select the best model amongst the Poisson Hurdle, ZINB, ZIP, NB and Poisson. A single study population is inferred by the Vuong test if the ZINB model is rejected in support of the NB model, then the null of no "structural" zero is not rejected hence the heterogeneity parameter α in the NB model is estimated. This implies the dispersion parameter accounts for unobservable heterogeneity responsible for overdispersion if this parameter is significant. When the Vuong test confirms that the NB is rejected in support of the ZINB model, test whether the parameter α is significant in the ZINB model. There is both extra Poisson variation and structural zero counts in the data if the estimate of α is again significant.

1.3 Statement of Problem

In several countries, dental caries still remains a dental health problem and numerous studies have been carried out in different age-groups. Prevalence of dental caries was reported in preschool children, group of 12-year-old children, group of 15-year-old adolescents and adolescents that are older than 15. Mainly, studies on dental caries have therefore been carried out in groups of children, adolescents and adults. James and Sheiham (n.d.) posited that most policies, research programmes and surveys based on dental caries have centred on children. Regrettably, dental caries levels are very much advanced in adults than in children consuming even fluoridated water (Bernabe and Sheiham, 2014b) and fluoridated toothpaste (Broadbent et al., 2008; 2013). It is exceptionally significant that this is emphasized as adult caries accounts for about 80% of the dental care costs relating to caries as compared to just 20% for children (James & Sheiham, n.d.). Dental caries therefore imposes a considerable burden not only on the individual adult but also on the economy of Ghana. Once dental caries occurs, it stays for the lifetime of the affected person and this imposes a considerable burden not only on the individual but also on the economy. Hence the need to know risk factors contributing to dental caries in adults using count data models such as the negative binomial (NB) model and Poisson model as one component models as well as Poisson Hurdle model, zero inflated negative binomial model and zero inflated Poisson model as two component models to analyse the DMFTDATA.

1.4 Objectives

Dental caries as a disease can only be controlled to a certain measure but cannot be eradicated hence this research will achieve the following objectives with respect to DMFTDATA (Decayed, Missing and Filled Teeth Data):

1. to identify the key risk factors contributing to dental caries in adults
2. to determine an appropriate model suitable for the analysis of DMFTDATA

1.5 Methodology

A cross sectional design was employed using clinical caries examination and questionnaire with interview. A systematic random sampling technique will be used to select the samples. The dental caries assessment will be carried out using DMFT (Decayed, Missing and Filled Teeth) index thus the addition of Decayed teeth due to caries (D), Missing teeth due to caries (M) and Filled teeth due to caries (F). Information will be collected about oral hygiene practices by faceto-face interviews. Information will generally be recorded on DMFT calculated, gender, age, duration of change of toothbrush, frequency of sweet consumption per day, frequency of brushing, use of fluoridated toothpaste, rinsing habit after every meal with water, dental visit experience and jaw accident.

The data will be put together using Microsoft Access 2007 and then transferred to SPSS version 16.0 and R for analysis.

Generalized Linear Regression models specifically Poisson Hurdle (PHURDLE), zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), Negative Binomial (NB) and Poisson models will be used as statistical tools in the analysis. Dispersion test would be performed to determine overdispersion or under-dispersion in the DMFTDATA using variance and mean ratio. Finally, model selection methods such (AIC) and the Voung test statistic will be used to determine the model that performs efficiently with respect to the DMFTDATA.

1.6 Significance of the Research Work

This thesis will serve the following purposes.

First, it will create awareness of risk factors of dental caries relating to adults.

Also, this work will help in planning for dental care for adults in areas of policies, research programs and surveys by institutions and individuals since adults form about 80% of dental caries susceptibility.

This thesis will again serve as reference for other research works.

1.7 Research Questions

This thesis will seek to answer the following questions

a. What are the key risk factors that contribute to dental caries in adults?

b. What statistical model is appropriate for modelling DMFTDATA?

1.8 Organisation of the Thesis

This thesis comprises of five main chapters. Chapter one contains the Introduction, the problem statement, objectives of the research work, significance of the research work and the organization of the research work. Chapter two of this research work contains the literature review. Chapter three discusses the various methods used in this research work. Chapter four on the other hand features the analysis and modelling of the data. Findings are also discussed in this chapter. The final chapter, chapter five, talks about summary of findings, conclusions and recommendations. References and the appendixes then follow this chapter.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents relevant literature for this thesis.

2.2 Theoretical and Impirical Literature Review

Worldwide, approximately 2.43 billion people have tooth decay in their permanent teeth according to SANTINI and in Ghana, 485 mixed adults have 28 percent dental caries rate with 12 percent severe (MacGregor, 1964). Kidd and Fejerskov (2013) opined that in spite of all the efforts of the dental profession, tooth decay remains one of the commonest chronic diseases. They emphasised that dental caries is ubiquitous, present in all populations and is as old as mankind and added that the caries incidence rate varies extensively between and within populations. According to them as age increases, signs and symptoms of dental caries accumulate, and in most adult populations the caries prevalence approaches 100%. Prevention and operative treatment of caries lesions, and their clinical consequences, occupy the majority of the dental profession life-long around the world and the cost of dental healthcare is a major societal burden, they stated.

According to Lee (2005), dental caries is one the most prevailing chronic diseases globally which is also an expensive burden for health care services. Lee (2005) averred that treatment of dental caries is costly in industrialized countries with a budget ranging from 5% to 10% of the total health care expenditures, and put forward that prevalence rate of dental caries is severe in most developing low-income countries, adding that with an estimated 5 billion people globally suffering from dental caries, more than 90% of this dental caries is left mostly untreated.

Dental caries is termed as a post-eruptive, localized, pathological process of external source including softening of the hard tooth tissue and continue to the development of cavity as said by World Health Organization (1962). The organization simplified dental caries in various forms and purposes. According to the organization, clinical caries is for the purposes of recording, a cavity diagnosed by mouth mirror and probe examination. It added that clinical caries is a phase in the course of dental caries and indicated that dental caries advances from a microscopic lession that cannot be positively detected by clinical methods to a cavity that can be detected by clinical examination. Life caries is explained by the orgainisation as past and existing clinical caries put together and may be expressed as the sum of the number of decayed teet due to caries, missing teeth due to caries and filled teeth due to caries.

In throwing more light on dental caries, Fujiwara (2005) recognized dental caries as a multifactorial disease caused by actions of three main factors, thus, cariogenic bacteria, teeth, and fermentable sugars. Dental caries as an infectious and transmissible disease is caused by mutans streptococci mainly Streptococcus sobrinus, Streptococcus mutans and lactobacilli. S. mutans and S. sobrinus produce three and four glucosyltransferases, respectively, whose cooperative actions is essential for adhesive glucan synthesis (Fujiwara, 2005). Adhesive glucans mediate attachment of bacteria to the tooth surface as well as to each other and mutans streptococci require the tooth surface as a habitat according to (Fujiwara, 2005).

Fujiwara (2005) asserted that the detection rate of mutans streptococci increases with age.

Arens (1999) on the occurrence of dental caries said, dental caries occur as a result of enamel and dentine breakdown by bacteria in dental plaque via sugar anaerobic metabolism from diet. Fujiwara (2005) added that initiation of dental caries went through dental plaque which is a biofilm consisting of microorganisms and their products such as adhesive glucan. Fujiwara (2005) again put forward that the plaque bacteria produces a large quantity of acids such as lactic acid from fermentable carbohydrates which are entrapped between the tooth surface and plaque biofilm, and when the pH of the enamel falls below 5.6, then loss of mineral from enamel is induced.

It must be accentuated that dental caries can only be controlled to some extent as a disease, but cannot be eradicated according to Petersen, et al. (2005).

According to Petersen, et al. (2005) the inconvenience with eating, experience of pain, chewing, communication and smiling because of the discoloured, missing and damaged teeth have a serious effect on the life of individuals and comfort. They said as a consequence that dental caries and other oral diseases impede the individuals' performance at home, at school and at work causing especially schooling and working hours to be lost annually worldwide.

Epidemiological studies of dental caries in adults are relatively rare according to Namal et al. (2008). During the last decades, however, studies have showed that in both developing and industrialized countries dental caries comprise a major problem in the adult population (Luan et al. 2000; Treasure, et al. 2001).

Various investigators in different parts of the world have evaluated risk factors of dental caries among adults and found age as the most commonly reported risk factor associated with dental caries (Winn et al. 1996; Forshee & Storey, 2004). In some articles, women have been reported to have more dental caries than men (Splieth et al. 2002; Lin et al. 2001), while there have been no significant differences between men and women in another study (Szoke & Petersen, 2004). Also, poor oral hygiene practices showed significant increase in dental caries (Shah & Sundaram, 2004).

When the habit of sweet eating is correlated with the caries incidence, it is found that 55 percent of adults ate sweets, and it is of interest that if adults are considered, the correlation between dental caries and sweet eating is even more striking MacGregor (1964) asserted.

Javali and Pandit (2010) in their research put forward that duration of change of toothbrush, frequency of brushing, systematic toothpaste, rinsing habit, frequency of sweet consumption and smoking habit are found to have significant influences on dental caries. Epidemiological studies also revealed the relationship between caries prevalence and sugar consumption, Fujiwara (2005) claimed.

James and Sheiham (n.d.) opined that most policies, research programmes and surveys on dental caries have centred on children and regret that dental caries levels are rather very much advanced in adults than in children consuming even fluoridated water (Bernabe & Sheiham, 2014b) and fluoridated toothpaste (Broadbent et al, 2008; 2013). They therefore claimed that it is exceptionally significant that dental caries is emphasized in adults as adults caries accounts for about 80% of the dental care costs relating to caries compared with 20% for children.

The basic assessment to confirm the presence and absence of dental caries is a cautious visual examination and early and consistent means of identifying caries involves visual and radiographic methods (Cappelli & Mobley, 2008)

To measure, per person, the intensity or extent of dental caries in relation to the number of teeth affected, DMF (DMFT) index is recommended according to World Health Organization (1962). The DMFT index has been widely utilized in epidemiological surveys of oral health (Cypriano, et al, 2005) and the DMFT index is used for more than seven decades. It is a standard measure of dental caries incident (Larmas, 2010). The World Health Organization (1962) said the DMFT index, per person, is the average number of permanent teeth which are decayed (D), missing due to caries (M), or filled due to caries (F). It added that DMFT index is a life-time quantitative expression of dental caries incident of the permanent teeth and that in the calculation of the DMFT index the numerator is the total number of DMFT for individuals while the denominator consist of the total number of individuals examined. The counts per individual DMFT index is between 0 and 28 or 32, and dependent respectively on whether the third molars are involved in the counting as indicated by Mahyudin and Hermawan (2016).

Javali and Pandit (2010) in their study considered age, duration of change of toothbrush, frequency of sweet consumption per day, frequency of brushing, use of systemic toothpaste, rinsing habit after every meal with water and many more as risk factors associated with dental caries. Gender, jaw accident or trauma and dental visit experience are also considered by other researchers as risk factors associated with dental caries (García-Cortés, et al. (2009); Fontana and Zero (2006); Petersen, et al. (2005).).

James and Sheiham (n.d.) recommended that researches should not only be based on response and risk factors of dental caries but also assessing the risk factors and their magnitude.

Modelling count data usually is a regular task in statistics, econometrics, epidemiology and other related fields (Zeileis, et al. (2007); Liu & Cela (2008, March)). Count data is a data that arises from counting as explained by Miller (2007). They are the realization of an integer-valued random variables that are non-negative, Cameron and Travedi (2013) added. The data are constrained by lower bound of zero and no upper bound as asserted by Miller (2007).

Various researchers modelled count data in various fields with different models. Gurmu (1997) assessed the impact of managed care program on healthcare utilization using hurdle model. Deb and Trivedi (1997) modelled the demand for healthcare utilization by the elderly using a finite mixture negative binomial regression. Winkelmann (2004) researched on the effect of healthcare reform on the number of doctor visits in Germany using a number of modified count data models.

Empirical count data characteristically shows an excess number of zeros and/or over-dispersion therefore the use of Poisson model for such count data is usually inadequate in these disciplines. The issue of over-dispersion may be solved by expanding the Poisson model in a variety of directions to estimate an additional dispersion parameter. The use of a negative binomial (NB) regression is more formal way to solve the issue of over-dispersion.

All count models are from the family of generalized linear models (GLMs) (McCullagh & Nelder, 1989). Even though these models normally may be capable of dealing with over-dispersion relatively well, however, these models are insufficient for modelling excess zeros in many applications. There is increased interest in zero-augmented models since Mullahy (1986) and Lambert (1992), both in the statistics and econometrics literature which address the issue of excess zero counts by a second model part dealing with the excess zero counts. To some extent, Zero-inflated models (Lambert, 1992) take a different approach, which is; the zero-inflated models are mixture models that combine a count component and a point mass at zero. Hurdle models (Mullahy, 1986) combine a rightcensored hurdle component with a left-truncated count component. An overview of hurdle and zero-inflated models and other count data models in econometrics are discussed in Cameron and Trivedi (2005).

Count data models can be understood and represented by the GLM framework which appeared in the early 1970s (McCullagh & Nelder, 1989). GLMs express the response variable $Y_i(i = 1,2,\ldots,n)$ on a vector of covariates X_i. The conditional distribution of $y|X_i$ is a linear exponential family that has a probability density function as

$$f(y,\lambda,\phi) = exp(\frac{y.\lambda - b(\lambda)}{\phi}) + c(y,\phi) \tag{2.1}$$

Where λ, the canonical parameter, relates the covariates by a linear predictor and φ is usually known as the dispersion parameter. Usually, $b(.)$ and $c(.)$

are known and they also show which member of the family is used, for example, Poisson distribution. Conditional variance and mean of Y_i are $V\,ar(y|X_i) = \varphi.b^{00}(\lambda_i)$ and $E(y|X_i) = \mu_i = b^0(\lambda_i)$ respectively. That is, the distribution of Y_i up to the dispersion parameter φ, is determined by its variance also called variance function is proportional to $V\,(\mu) = b^{00}(\lambda(\mu))$.

The dependence of the conditional mean $E(y|X_i) = \mu_i$ on the covariates X_i is expressed by $g(\mu_i) = X_i^T\beta$, and β is the vector of regression coefficients that are normally estimated by maximum likelihood and $g(.)$ is the link function. (Zeileis, et al. 2007)

With regards to count data such as DMFTDATA, counts are discrete in nature and non-negative, and it makes a lot of sense to model such count data using Poisson, Negative Binomial, Hurdle and Zero Inflated models (Kibria, 2006). According to Kibria (2006) it will be assumed that, the occurrences of outcomes will be independent of each other regardless of whether the assumed model is a Poisson, Negative Binomial, Hurdle Zero or Inflated model. But all these models can be understood and presented using the Generalized Linear model that appeared in the early 1970's (McCullagh & Nelder, 1989). GLM may also be seen as a regression model for the mean only where the estimating functions applied for fitting the model are from a particular family instead of presenting GLM as models for the full likelihood (Zeileis, et al. 2007).

The most basic and common regression model is the Poisson regression model for count data according to Xia, et al. (2012). They said with the Poisson regression model, each observed count, Y_i, is sampled from a Poisson distribution with its mean μ_i given as a vector of covariates X_i for each ith subject. The Poisson distribution, according to them is obtained by modelling the number of independent events through a memory-less Poisson process with fixed rate. Generally, the Poisson distribution has its probability density function as:

$$P(Y_i = y|X_i) = \frac{exp(-\mu_i)\mu_i^y}{y!}, \ \ y = 0, 1, 2, ... \quad (2.2)$$

where $\mu_i = exp(X_i^T\beta)$

A distinguishing assumption of the Poisson distribution is the variance and mean equality, thus, $V\,ar(y|X_i) = u_i$ that regrettably becomes the main limitation in application of this model, and the estimates obtained by the Poisson regression model will therefore be inefficient when overdispersion is an issue (Cameron & Trivedi, 2013).

The negative binomial (NB) model, the mostly known option to Poisson regression, according to Xia, et al. (2012), deals with overdispersion by modelling explicitly the correlated events through a latent variable. They added that NB extends the Poisson regression model by adding that the mean μ_i of Y_i is not determined only by X_i but also by adding a heterogenity component $_i$ independent of X_i. If $\exp(\epsilon_i)$ is assumed to be distributed with a gamma (α,β), according to Xia, et al. (2012) then the NB model has the following density function

$$P(Y_i = y|X_i) = \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}(\frac{\theta}{\theta+\mu_i})^{\theta}(\frac{\mu_i}{\theta+\mu_i})^{y}, \;\; y = 0, 1, 2, ... \quad (2.3)$$

where

$\theta = \frac{1}{\alpha}$ which is the dispersion parameteer

$$\mu_i = \exp(X^T\beta + \epsilon_i) = \exp(X_i^T\beta)\exp(\epsilon_i)$$

and that $E(\exp(\epsilon_i)) = 1$ hence $E(\exp(X_i^T\beta + \epsilon_i)) = E(exp(X_i^T\beta))$, thus, the expected value, μ_i does not change even if we assume a negative binomial distribution or a Poisson distribution. They also asserted that since the dispersion parameter, $\theta > 0$, under the negative binomial distribution, $V\,ar(Y_i = y|X_i) = \mu_i(1 + {}^{\mu}{}_{\theta}{}^{i}) > \mu_i$, implying that. $V\,ar(Y_i = y|X_i)/E(Y_i = y|X_i) = 1 + {}^{\mu}{}_{\theta}{}^{i}$. This implies the variance of the negative binomial is much greater than its mean, hence addressing the issue of overdispersion. Xia, et al. (2012) also noted that as θ approaches 0 then Poisson and negative binomial models can be seen as nested for the reason that NB approaches the Poisson.

Negative binomial model is inappropriate for modelling a high percentage of zero counts in a data even though it is capable of addressing overdispersion. Such excess zero counts may be modelled using zero-inflated Poisson (ZIP) regression (Lambert, (1992); Long, (1997); Greene, (1994); Cameron & Trivedi, (2013)). The uses of ZIP regression models are more popular especially from the time of the publication of Lambert in 1992. Origin of zero inflated Poisson regression model is in the econometrics literature (Mullahy, 1986). When the zeros in a dataset are said to be caused by both chance and systematic factors, ZIP model is a model which can be used (Min & Agresti, 2005). This model is a combination of two statistical processes, where one is producing both nonzero and zero counts, the other always generating zero counts according to Xia, et al. (2012). They said that in ZIP, each observation originates from one of two probable distributions where one of the distribution is producing constant zeros while the other following Poisson distribution. The Poisson regression models the count data while the probability of the constant zero in a ZIP model is modelled normally using a logit model, they asserted, and added that the ZIP models two kinds of zeros: the structural zero counts beyond and above the expected zero incidence following Poisson and the sampling zeros due to sampling inconsistency following Poisson distribution. To them, either the Poisson process or the logistic process generates the observed zeros.

According to Xia et al. (2012), if $\omega_i = Pr(i \in (structural\ zero)|Z_i)$ and $1 - \omega_i = Pr(i \in sampling\ zero|Z_i)$ then the density function of zero inflated

Poisson is:

$$P(Y_i = y|X_i, Z_i) = \omega_i + (1 - \omega_i)exp(-\mu_i)\ for\ y = 0$$

$$P(Y_i = y|X_i, Z_i) = (1 - \omega_i)\frac{exp(-\mu_i)\mu_i^y}{y!}\ for\ y > 0 \qquad (2.4)$$

When X_i and Z_i are two sets of covariates relating to the count data and logit models by $\log(\omega_i/(1-\omega))$ = $Z_i\gamma$ and the probability of being in the nonzero state is determined by $\log(\mu_i) = X^T\beta.\omega_i(1 < \omega_i < 1)$ (Lambert 1992, Long 1997). It is clearly seen that the observed zeros are coming from the two sources: sampling and structural sources. The variance and mean of Y_i according to Xia, et al.

(2012) are

$$Var(Y_i|X_i,Z_i) = \mu_i(1 - \omega_i)(1 + \mu_i\omega_i) \quad (2.5)$$

And

$$E(Y_i|X_i,Z_i) = \omega_i 0 + \mu_i(1 - \omega_i) = \mu_i(1 - \omega_i) \quad (2.6)$$

To Xia, et al. (2012), $Var(Y_i|X_i,Z_i)/E(Y_i|X_i,Z_i) = 1 + \mu_i\omega_i$ indicating that $Var(Y_i|X_i,Z_i) > E(Y_i|X_i,Z_i)$. They stated that ZIP reduces to Poisson,if ω_i approaches zero, thus the quantity of structural zeros reduces to zero.

By replacing the Poisson component in ZIP with that of the negative binomial, according to Xia, et al. (2012) and Chipeta, et al. (2014), results in zero-inflated negative binomial (ZINB). According to Xia, et al. (2012) and Chipeta, et al. (2014), ZINB has the following density:

$$P(Y_i|X_i, Z_i) = \omega_i + (1 - \omega_i)(\frac{\theta}{\mu_i + \theta})^{\theta}\ if\ y = 0$$

$$P(Y_i|X_i, Z_i) = (1 - \omega_i)\frac{\Gamma(y + \theta)}{y!\Gamma(\theta)}(\frac{\theta}{\theta + \mu_i})^{\theta}(\frac{\mu_i}{\theta + \mu_i})^{y}\ if\ y > 0 \quad (2.7)$$

They asserted that either the probit, logit or other models for binary outcomes can be modelled using the binary process.

The ZINB regression model according to Chipeta et al. (2014) relates ω_i and μ_i to covariates by

$$\log(\mu_i) = X_i^T\beta \quad (2.8)$$

And

$$\text{log}it(\omega_i) = Z_i\gamma \qquad (2.9)$$

where i = 1,2,...,n and X_i and Z_i have $j-$ and $k-$ dimensional vectors of covariates regarding the ith subject respectively, with β and γ their corresponding regression coefficients' vectors.

The mean and variance of the zero inflated negative binomial model as posited by Xia, et al. (2012) are;

$$E(Y_i = y|X_i, Z_i) = (1 - \omega_i)\mu_i \qquad (2.10)$$

$$Var(Y_i = y|X_i, Z_i) = \mu_i(1 - \omega_i)(1 + \mu_i(\omega_i + \theta)) \qquad (2.11)$$

$$\frac{Var(Y_i = y|X_i, Z_i)}{E(Y_i = y|X_i, Z_i)} = 1 + \mu_i(\omega_i + \theta) \qquad (2.12)$$

It follows that for ZINB, $V\ ar(Y_i = y|X_i,Z_i) > E(Y_i = y|X_i,Z_i)$ indicating clearly that ZINB has the potential of addressing overdispersion since $(\omega_i + \theta)/(1-\omega_i)$ is a function of both the dispersion parameter θ and the zero=inflated parameter ω_1 as posited by Xia et al.(2012), Hence ZINB accounts for both overdispersion and population heterogenity in the distribution of the negative binomial component of the ZINB according to Xia et al. (2012).

Also capable of addressing excess zeros is the Hurdle model according to Mullahy (1986). The Hurdle model, developed by Mullahy in 1986 in economics though the term itself was largely likely coined by Cragg (1971). Welsh, et al.(1996) called it a 'conditional Poisson model'. Miller (2007) said a binomial probability model directs the binary outcome irrespective of if a count variate has a zero or a positive outcome. This gives the idea fundamental to the hurdle formulations. The 'hurdle' is crossed if the outcome is positive, with the conditional distribution of the positive outcomes being determined by a truncated-at-zero count data model (Mullahy, 1986) such as a truncated Poisson (Min & Agresti, 2005). A hurdle model is a modified two processes count data model in which one process producing the positive outcomes while the other generating the zeros according to Cameron and Trivedi (2013). They stated that the two generating processes are the same. If the outcome of the generating process is positive, then the "hurdle" is crossed, hence the concept fundamental in the hurdle model is that a binomial probability model determines the binary outcome of whether a count variable has a zero outcome or a positive outcome and that a zero-truncated count model determines the conditional distribution of the positive outcomes (Mullahy, 1986). Gurmu (1998) introduced generalized hurdle model for the analysis of over-dispersed or under-dispersed count data.

In diverse applications and comparisons of hurdle model, some researchers, such as Arulampalam and Booth (1997), Pohlmeier and Ulrich (1995) and others discussed the applications of hurdle models. The comparison between hurdle and zero-inflated models as two state models was discussed by Greene (2005). Saffari, et al. (2011) used a right truncated Poisson regression model to argue the overdispersion problem on count data. Gurmu and Trivedi (1996) applied a hurdle model to the annual number of recreational boating trips by a family. Generalized heterogeneous, hurdle, zero-inflated and compound frequency models were compared by Boucher, et al. (2007) for the annual number of claims reported to the insurer. Dalrymple, et al. (2003) discussed the incidence of sudden infant death syndrome and applied three mixture models including a hurdle model.

Logit-Poisson model is the most common Hurdle regression formulation, which is the mixture of a truncated Poisson regression that models positive counts conditional on nonzero outcomes and a logit regression that models nonzero verses zero outcomes as said by Liu and Cela (2008). They said its probability density function is given as

$$P(Y_i = y|X_i,Z_i) = \omega_i \, for \; y = 0$$

$$P(Y_i = y|X_i, Z_i) = \frac{(1-\omega_i)exp(-\mu)\mu_i^y}{(1-exp(-\mu_i))y!} \; for \; y > 0 \qquad (2.13)$$

where $\mu_i = exp(X_i^T\beta)$

According to Chipeta, et al. (2014) using a logistic regression model is the most normal preference to model probability of zero counts in a Hurdle model. Thus $\mathrm{log}it(\omega_i) = Z_i\gamma$.

And the estimates of covariates of X_i on only positive count outcomes are modelled via Poisson regression, thus

$$\log(\mu_i) = X^T\beta \tag{2.14}$$

Given that $\pi = P(y > 0)$, the probability of a nonzero outcome with η, the expected value, then the variance and mean are given according to Chipeta, et al. (2014) as:

$$Var(Y_i = y|X_i, Z_i) = \eta(\mu_i - \eta) + \frac{\pi\sigma^2}{1 - P(0;\alpha)} \tag{2.15}$$

And

$$E(Y_i = y|X_i, Z_i) = \eta = \frac{\pi\mu_i}{1 - P(0;\alpha)} \tag{2.16}$$

Clearly, $V\,ar(Y_i = y|X_i,Z_i) > E(Y_i = y|X_i,Z_i)$ The results from both simulated and actual data sets in the zero-inflation literature are in much disagreement according to Miller (2007). The ZIP model is found to do better than the NB model, and the NB model was also found to perform better than the Poisson model (Lambert, 1992). Greene (1994) found the NB model to perform better than the ZIP model, which was better than the Poisson model. Slymen, et al. (2006) found the ZINB and ZIP models to be the same. Welsh, et al. (1996) found the Hurdle model to be equal to ZIP models while Pardoe and Durham (2003) found the ZINB model to be better than both the Poisson and Hurdle models.

The difference with regards to the distribution of nonzero counts and the proportion of zero counts is one important characteristic of these and other models for count outcomes according to Miller (2007). In some research, the percentage of zero counts was as low as 21.6% in the count outcome analyzed (Bhning, et al. 1999) while others (Zorn, 1996) used a percentage of zeros as high as 95.8%. Miller (2007) explained that the nonzero counts differ with regards to their distributions ranging from positively skewed, to normal, then to uniform and that depending on the percentage of zeros and the distribution for the nonzero counts, different models produce varied results.

DMFTDATA, according to Javali and Pandit (2010), consists of counts of decayed due to caries, missing due to caries and filled teeth due to caries and that such data has high percentage of excess zero counts with respect to probability distributions for example the Poisson distribution. They asserted that this model often underestimate the observed DMFT calculated since the data is usually overdispersed and a single Poisson parameter is not enough to explain the population. The negative binomial model can deal with such outcome by modelling overdispersion due to unobserved heterogeneity. The incidence of structural zero counts that is a violation of the Poisson model by the population heterogeneity can be modelled using ZIP while ZINB tackles both overdispersion and heterogeneity (Xia, et al, 2012). When count outcomes emanate from two states, Lee, et al. (2007) asserted that the ZIP and ZINB are known to provide robust statistics especially when zero counts are present.

An essential part of statistical analysis is model selection and it is certainly vital to the pursuit of science according to Kadane and Lazar (2004). In most empirical work model selection is a significant step and accordingly, there is an immeasurable literature dedicated to this issue. The Kullback-Leibler (KL) information criterion, since Akaike (1973,1974), has become a known measure for chosing among models taking the form of parametric likelihoods. Some literature (Inoue & Kilian(2006) and Vuong(1989)) employ this criterion in addition to previous knowledge on embedding the model selection problem into a classical hypothesis testing framework (Hotelling, 1940). This approach as a goodness-offit measure uses the maximum of the likelihood function. Model 1 is preferred to model 2 if model 1 is found to have a statistically and significantly larger maximum likelihood than model 2.

There is no consensus on if the Poisson model is nested with ZIP model and also if NB model is also nested with ZINB model. The Poisson model and the NB model are one-state models for single population whereas the zero inflated negative binomial model and the zero inflated Poisson model are twostate mixture models for population comprising of two subpopulations. That is, the two classes of one-state and two-state mixture models can therefore not be used to explain the same study population. The Poisson versus ZIP as well are as NB versus ZINB models are considered by several researchers as nonnested for this reason. As a result, the score test and log-likelihood ratio test therefore cannot be used to evaluate these models. See (Atkins and Gallop (2007), Minami, et al. (2007), Sheu, et al. (2004) and Tin (2008)).

Whether competing models are correctly specified models or whether competing models are non-nested, nested, or overlapping, Vuong proposed a general approach to select between competing models (Vuong, 1989) and that Vuong's statistic is the average log-likelihood ratio correctly normalized in other that it can be compared to a standard normal. According to Xia, et al. (2012) the Vuong test statistic (V) is defined by

$$V = \frac{\bar{m}}{S_m/\sqrt{n}} \tag{2.17}$$

Where $\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ and $S_m = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2}$ represent the mean and standard deviation of the measurement of m_1. m_1 is defined as:

$$m_i = \log\left[\frac{\hat{f}_1(Y_i = y|X_i)}{\hat{f}_2(Y_i = y|X_i)}\right] \tag{2.18}$$

Where $\hat{f}_1(Y_i = y|X_i)$ and $\hat{f}_2(Y_i = y|X_i)$ are the predicted probabilities of the corresponding midels $\hat{f}_1(Y_i = y|X_i)$ and $\hat{f}_2(Y_i = y|X_i)$ respectively.

The test is directional, with a large positive value favoring f_1 and a large negative value favoring f_2,and that a value asymptotic to zero signifying neither model fits the data well given that the test statistic has an asymptotically standard normal distribution (Long (1997) and Vuong (1989)).

The Akaike information criterion (AIC) is used to determine the comparative superiority of statistical models for a given set of data. The AIC was developed and published in 1973 by Hirotugu Akaike (Akaike, 1973). Mazerolle (2004) posits that the AIC gives an objective way to decide which model amongst competing models is parsimonious. The AIC value for a given data has no meaning but when it is compared to the AIC values of competing models, it becomes meaningful, and the model with smallest AIC value is the best model amongst competing models for the specified data set (Mazerolle, 2004). Mazerolle (2004) said the AIC is used to select a model that fits the data well but has a minimum number of parameters thus the AIC penalizes for the addition of parameters. The greatest strength of the AIC as posited by Mazerolle (2004) is its potential in model selection. This is because the AIC is independent of the order by which the models are calculated. We can include model uncertainty to get precise and robust estimates, and confidence intervals in situations where there are a lot of models ranked highly based on the AIC. Information theory is the basis for AIC. Mazerolle (2004) put forward a comparative estimate of the information lost when a specified model is used to signify the process which produces the data. In that regard, the AIC addresses the trade-off between model complexity and the goodness of fit of the model. In the case of testing a null hypothesis the AIC does not provide a test of the model, thus, in an absolute sense the AIC cannot tell about the quality of the model used. For various applications in conservation biology, physiology, ecology and behavioural ecology, the information-theoretic approach revolving about the AIC showed great promise (Mazerolle, 2004). Burnham and Anderson (2002) ascertained that AIC has no model restrictions.

While ZIP, ZINB and PHURDLE are known to address extra zero counts, it is however tricky to tell the suitable choice for the data at hand, because such excess zero counts are indirectly observed and latent. It is essential to apply goodness of fit statistics to guide the choice of models suitable and best for the data according to Xia, et al. (2012). With regards to DMFTDATA as a count data in this research, all models, tests, finding etc. would applied with respect to the relevant literature reviews in this research.

Chapter 3

METHODOLOGY

3.1 Introduction

This chapter presents the methods used in the thesis.

3.2 Research Design

A cross sectional design was employed using questionnaire through interview and clinical examination.

3.3 Sample

The study was conducted between June, 2015 and January, 2016 in Sefwi Wiawso District Hospital, Sefwi Wiawso, Ghana, and involved systematic random samples of 1158 participants of age 18-65 years.

Ethical approval for the thesis was obtained from the Sefwi Wiawso District Hospital before the start of the study and individual consent was obtained from the participants. Confidentiality of information is also assured in conjunction with the hospital.

3.4 Data Collection

The data collection was in two parts: a questionnaire survey and a clinical examination. Individuals were clinically examined by a dentist and information on individual participant’s oral health practices such as DMFT calculated (DMFT index), gender, age, duration of change of toothbrush, frequency of sweet consumption per day, frequency of brushing, use of fluoridated toothpaste, rinsing habit after every meal with water, dental visit experience and jaw accident were collected. Dental caries statuses of the participants were recorded, according to criteria proposed by World Health Organization, by counting the number of teeth that were decayed due to caries, missing due to caries, and filled due to caries for calculation of the DMFT index (the DMFT index is a standard method for measuring dental caries incident in individuals). Dental caries was identified visually at the cavitation level and early caries was not recorded. The dental caries examination was carried out by a trained examiner using DMFT index. Each tooth of each participant is examined using artificial light, plane mouth mirrors and probes.

Before the clinical examination, participants were assisted to fill out or asked to self-fill-out a structured questionnaire. The questionnaire was used to collect information on participant's oral health and others.

The questionnaires were collated and transferred by recording into Microsoft Access 2007. For the analysis of the data, the data was again transferred to SPSS 16.0 and R as statistical tools.

3.5 DMFT Index Calculation Guidelines

The teeth not counted were supernumerary teeth, congenitally missing teeth, unerupted teeth, primary teeth reserved in the permanent dentition and avulsion teeth due to trauma or accident teeth removed for causes other than dental caries. The third molars were optional in counting. The tooth was recorded as a D when both carious lesion(s) and a restoration or a carious lesion(s) is seen. A tooth is recorded as M when it has been removed due to caries. When a temporary or permanent tooth is filled due to caries, this is counted as F. When there was no other area of the tooth with primary caries or there was no recurrent caries and one or more permanent restorations were present, the teeth were considered filled without decay. Teeth are not counted as an F when restored for reasons other than caries

3.6 Principles and Rules in Recording DMFT

- A tooth having a number of restorations was recorded as F and as it is counted as one tooth
- A tooth was counted as decayed, D, if the tooth has restoration on one surface and caries on the other
- For decayed tooth due to caries, missing tooth due to caries and filled tooth due to caries, no tooth was counted more than once

3.7 Calculation of DMFT

For individual

$$DMFT\ index = D + M + F \tag{3.1}$$

The DMFT index has a minimum score of zero and a maximum score of 32 for adults.

3.8 Count Data Regression

With regards to count data, for example DMFTDATA, they are discrete in nature and non-negative and using Zero Inflated Negative Binomial, Poisson Hurdle, Zero Inflated Poison, Negative Binomial and Poisson makes more sense to model this DMFTDATA. It will be assumed that the occurrences of outcomes will be independent of each other regardless of whether the assumed model is a Zero Inflated Negative Binomial, Poisson Hurdle, Zero Inflated Poison, Negative Binomial or Poisson. The above models being used in this research work are described in the following subsections. All these models can be understood and presented using the Generalized Linear models that appeared in the 1970's (McCullagh, 1992).

3.8.1 Generalized Linear Models(GLMs)

GLMs were first introduced in 1972. Instead of a different study for different regression models, GLMs provide a joined framework to study these different regression models. GLMs are an expansion of the classical linear models. The Generalized Linear Models include Poisson regression models, linear regression models, Zero-inflated regression models, logistic regression models, Negative Binomial regression models, Poisson Hurdle model and many more models. These models have a number of distinctive properties like the use of a common method in estimating parameters.

3.8.2 Components of Generalized Linear Models

Lawal (2003) explains that GLMs are a subset of the traditional linear models that permit other possibilities than modelling the mean as a linear function of the independent variables. All GLMs have random component, systematic component and link function.

With the random component of a GLM, the dependent variable follows a certain distribution Y_i given the explanatory variables X_i. As explained by Agresti (1996), the indication of the distribution for the response variable is a requirement for the random component. One could specify this distribution to be normal; hence, classical models such as ordinary least squares regression and analysis of variance models are included within this broader class of generalized linear models. Other possible random components that could be specified include the binomial distribution, negative-binomial distribution, gamma distribution and Poisson distribution. Specifying the random component depends on the expected population distribution of the outcome variable. A requirement for GLM is that the response variable has a distribution in the exponential family of the models according to McCullagh and Nelder (1989). These distributions are defined primarily by a scale parameter φ and a vector of natural parameters λ.

The exponential family is given by

$$f(y;\lambda,\phi) = exp\Big(\frac{y.\lambda - b(\lambda)}{\phi} + c(y,\phi)\Big). \tag{3.2}$$

The model for the predictors is simply the systematic component established as a linear combination and is denoted η_i given as

$$\eta_i = \alpha + \beta_1 X_{i1} + \dots + \beta_k X_{ik} = X_i^T\beta \tag{3.3}$$

The link function $g(.)$ brings together the random component and the systematic component hence linking the function for the mean, μ_i, and the function for the systematic component, η_i, as $\eta_i = g(\mu_i)$. In other words, it specifies how the population mean of the outcome variable with a specific distribution is associated with the predictors in the model. If $g(\mu_i)$ redundantly equals μ_i, then the population mean itself is related to the predictors. This is termed the identity link and is exactly the function used to link the mean of the normal distribution to its covariates.

$$g(E(y)) = X_i^T\beta = \eta_i \tag{3.4}$$

The inverse of the link function is sometimes called the mean function given by:

$$E(y) = g^{-1}(\eta_i) \tag{3.5}$$

3.8.3 Exponential Family(Canonical Form)

Generalized Linear Models can be used to model variables following distributions in the exponential family in their canonical forms with its density function as

$$f(y;\lambda,\phi) = exp\Big(\frac{y.\lambda - b(\lambda)}{\phi} + c(y,\phi)\Big). \tag{3.6}$$

Where λ, the the canonical parameter, depends on the regressors by the use of a linear predictor and the dispersion parameter, φ, is often known. The functions $b(.)$ and $c(.)$ are also known and they establish the member of the family used, for example the Poisson or the normal distribution.

The variance and conditional mean of Y_i for distributions in the exponential families in their canonical form are respectively given by $V\,ar(y|X_i)$ = $\varphi.b^{00}(\lambda_i)$ and $E(y|X_i) = \mu_i = b^0(\lambda_i)$.

Where $b^0(\lambda_i)$ and $b^{00}(\lambda_i)$ denote the first and second derivatives of $b(\lambda)$. The dispersion parameter is normally fixed to one for some distributions.

The key advantage of GLM is they are not limited to a specific link function. As one can indicate the logit link as $g(\mu) = \log \frac{\mu}{1-\mu}$ and the log link as $g(u) = \log(u)$.

3.8.4 Poisson Regression Model

The Poisson regression model the most common regression model for count data. With the Poisson regression model, each observed count, Y_i, is sampled form a Poisson distribution with its mean μ_i given as a vector of predictors X_i for each ith subject. The Poisson distribution is obtained by modelling the number of independent events through a memory-less Poisson process with a fixed rate. The Poisson distribution has its probability density function as:

$$P(Y_i = y|X_i) = \frac{exp(-\mu_i)\mu_i^y}{y!},\ y = 0, 1, 2, ... \qquad (3.7)$$

Where Y_i represents the number of DMFT calculated for a specified period i, and μ_i represents the expected number of DMFT calculated per given period that can be expressed as

$$\mu_i = exp(X_i^T\beta) \qquad (3.8)$$

where β is the vector of unknown regression parameters to be estimated and X_i^T is the vector of independent variables. The equality of the mean and variance, thus, V $ar(y|X_i)$ = μ_i is a characteristic assumption of the Poisson regression model which regrettably is a main limitation in applications of this model, and the estimates based on the Poisson regression model when overdispersion is an issue will be inefficient (Cameron & Trivedi, 2013). As a result, the Poisson regression does not work well when there is an over-dispersion or a heterogeneity hence it is important to fit a model that is more dispersed than the Poisson model.

3.8.5 Derivation of the Poisson Regression Model

If Y_i is a random variable, such that Y_i follows a Binomial distribution. It therefore follows that

$$P(Y_i = y|X_i) = \binom{n}{y} p^y (1-p)^{n-y} \tag{3.9}$$

$$\begin{aligned}
But\ E(Y_i = y|X_i) &= np = \mu_i \\
&\Rightarrow p = \frac{\mu_i}{n} \\
P(Y_i = y|X_i) &= \binom{n}{y} p^y (1-p)^{n-y} \\
&= \binom{n}{y}\left(\frac{\mu_i}{n}\right)^y \left(1-\frac{\mu_i}{n}\right)^{n-y} \\
But \left(1-\frac{\mu_i}{n}\right)^{n-y} &= \left(1-\frac{\mu_i}{n}\right)^{n}\left(1-\frac{\mu_i}{n}\right)^{-y} \\
&As\ n \to \infty \\
lim_{n\to\infty}\left(1-\frac{\mu_i}{n}\right)^{n} &\to 1 \\
lim_{n\to\infty}\left(1-\frac{\mu_i}{n}\right)^{-y} &\to e^{-\mu} \\
also,\ \binom{n}{y}\left(\frac{\mu_i}{n}\right)^y &= \frac{n!}{(n-y)!y!}.\frac{\mu_i^y}{n^y} \\
but\ \frac{n!}{(n-y)!} &= n(n-1)(n-2)...(n-y+1) \\
&\approx n^y \\
lim_{n\to\infty}\binom{n}{y}\left(\frac{\mu_i}{n}\right)^y &= \lim_{n\to\infty}\frac{n^y}{n^y}.\frac{\mu^y}{y!} = \frac{\mu^y}{y!}
\end{aligned}$$

$$\therefore P(Y_i = y|X_i) = \frac{e^{-\mu_i}\mu_i^y}{y!} \quad (3.10)$$

3.8.6 Derivation of the Mean of the Poisson Regression Model

$$Let\ P(Y_i = y|X_i) = \frac{e^{-\mu_i}\mu_i^y}{y!};\ y = 0, 1, 2, ... \quad (3.11)$$

$$\begin{aligned}
then\ E(Y_i = y|X_i) &= \sum_{y=0}^{\infty} yP(Y_i = y|X_i) \\
&= \sum_{y=0}^{\infty} y\frac{e^{-\mu_i}\mu_i^y}{y!} \\
&= e^{-\mu_i}\sum_{y=1}^{\infty} y\frac{\mu_i^y}{y!} \\
&= e^{-\mu_i}\sum_{y=1}^{\infty} \frac{\mu_i^y}{(y-1)!} \\
&= \mu_i e^{-\mu_i}\sum_{y=1}^{\infty} \frac{\mu_i^{y-1}}{(y-1)!}
\end{aligned}$$

$$Let\ k = y - 1$$

$$then\ E(Y_i = y|X_i) = \mu_i e^{-\mu_i}\sum_{k=0}^{\infty} \frac{\mu_i^k}{(k)!}$$

$$But\ e^{\mu_i} = \sum_{k=0}^{\infty} \frac{\mu_i^k}{(k)!}$$

$$\Rightarrow E(Y_i = y|X_i) = \mu_i e^{-\mu_i} e^{\mu_i}$$

$$= \mu_i e^{-\mu_i+\mu_i}$$

$$= \mu_i \quad (3.12)$$

3.8.7 Derivation of the Variance of the Poisson Regression model

Let

$$V ar(Y_i = y|X_i) = E[(y - \mu_i)^2] \quad (3.13)$$

$$= \sum_y (y - \mu_i)^2 P(Y_i = y|X_i)$$

But

$$E[(y - \mu_i)^2] = E(y^2) - [E(y)]^2 \qquad (3.14)$$

And

$$(y^2) = \sum_y y^2 P(Y_i = y|X_i) \qquad (3.15)$$

$$= \sum_{y=0}^{\infty} y^2 \frac{e^{-\mu_i}\mu_i^y}{y!}$$

Also

$$E[y(y-1)] = \sum_{y=0}^{\infty} y(y-1)\frac{e^{-\mu_i}\mu_i^y}{y!} \qquad (3.16)$$

and

$$E(y^2 - y) = \sum_{y=2}^{\infty} y(y-1)\frac{e^{-\mu_i}\mu_i^y}{y!} \qquad (3.17)$$

$$= e^{-\mu_i}\sum_{y=2}^{\infty}\frac{\mu_i^y}{(y-2)!}$$

$$= \mu_i^2 e^{-\mu_i}\sum_{y=2}^{\infty}\frac{\mu_i^{y-2}}{(y-2)!}$$

Let $k = y - 2$, then

$$E(y^2 - y) = \mu_i^2 e^{-\mu_i}\sum_{k=0}^{\infty}\frac{\mu_i^k}{(k)!} \qquad (3.18)$$

But $e^{\mu_i} = \sum_{k=0}^{\infty}\frac{\mu_i^k}{(k)!}$
giving

$$E(y_2 - y) = \mu_{2i}\, e_{-\mu_i} e_{\mu_i} = \mu_{2i}\, e_{-\mu_i+\mu_i} = \mu_{2i}$$

$$\Rightarrow E(y^2 - y) = \mu^2{}_i \qquad (3.19)$$

Since $E(y^2 - y) = \mu^2{}_i$

$$\Rightarrow E(y^2) - E(y) = \mu^2{}_i$$

But $E(y) = \mu_i$

$$\Rightarrow E(y^2) - \mu_i = \mu^2{}_i$$

$$\therefore E(y^2) = \mu^2{}_i + \mu_i \tag{3.20}$$

But

$$Var(Y_i = y|X_i) = E[(y - \mu_i)^2] = E(y^2) - [E(y)]^2 \tag{3.21}$$

$$Var(Y_i = y|X_i) = \mu^2{}_i + \mu_i - \mu^2{}_i = \mu_i \tag{3.22}$$

3.8.8 Poisson as a Member of Exponential Family

$$P(Y_i = y|X_i) = \frac{e^{-\mu_i}\mu_i^y}{y!}, \qquad \mu_i > 0 \quad and \quad y = 0, 1, 2, \ldots \tag{3.23}$$

The exponential family in canonical form is given by the function

$$P(Y_i = y|X_i) = exp\left(\frac{y.\lambda - b(\lambda)}{\phi} + c(y, \phi)\right)$$

$$\ln(P(Y_i = y|X_i)) = \ln\left(\frac{e^{-\mu_i}\mu_i^y}{y!}\right)$$

$$= y \ln \mu_i - \mu_i - \ln y! \tag{3.24}$$

$$P(Y_i = y|X_i) = exp(y \ln\mu_i - \mu_i - \ln y!) \tag{3.25}$$

Implying that

$$\lambda = \ln\mu_i \qquad \Rightarrow \mu_i = e^\lambda \tag{3.26}$$

$$b(\lambda) = \mu_i = e^\lambda \tag{3.27}$$

$$\varphi = 1 \tag{3.28}$$

$$c(y,\varphi) = -\ln y! \tag{3.29}$$

For the mean and variance

$$E(Y_i = y|X_i) = \frac{\partial b(\lambda)}{\partial \lambda} = \frac{\partial e^\lambda}{\partial \lambda} = e^\lambda = \mu_i \tag{3.30}$$

$$Var(Y_i = y|X_i) = \phi\frac{\partial^2 b(\lambda)}{\partial \lambda^2} = \frac{\partial e^\lambda}{\partial \lambda} = e^\lambda = \mu_i \tag{3.31}$$

For Canonical link

$$g(\mu_i) = \lambda \tag{3.32}$$

$\ln\mu_i = \lambda$ implies that ln (natural log) is the canonical link for Poisson regression model

3.8.9 Estimation of the Poisson Regression model

μ_i must be parameterized in the estimation of Poisson regression model, and μ_i must be positive because it is a count, thus, $E(Y_i = y|X_i) > 0$. For this reason, to estimate the Poisson regression model we assume that

$$E(Y_i = y|X_i) = \mu_i = e^{X_i\beta} \tag{3.33}$$

This yields the following probability model:

$$(P(Y_i = y|X_i)) = \frac{e^{-\mu_i}\mu_i^y}{y!} \tag{3.34}$$

And the likelihood function therefore is:

$$L = \prod_{i=1}^{N}\left(\frac{e^{-\mu_i}\mu_i^y}{y!}\right) \tag{3.35}$$

The log likelihood function is:

$$\ell = \ln\prod_{i=1}^{N}\left(\frac{e^{-\mu_i}\mu_i^y}{y!}\right)$$

$$= \sum_{i=1}^{N}\ln\left(\frac{e^{-\mu_i}\mu_i^y}{y!}\right)$$

$$= \sum_{i=1}^{N}(-\mu_i + y\ln\mu_i - \ln y!)$$

$$= -n\mu_i + \left(\sum_{i=1}^{N} y\right)\ln\mu_i - \sum_{i=1}^{N}\ln y \quad ! \tag{3.36}$$

In maximizing the log-likelihood function, we take the derivative of l with respect to μ_i and equating it to zero to give:

$$\frac{\partial\ell}{\partial\mu_i} = -n + \left(\sum_{i=1}^{N} y\right)\frac{1}{\mu_i} = 0 \tag{3.37}$$

!

$$\mu_i = \frac{1}{N}\left(\sum_{i=1}^{N} y\right. \tag{3.38}$$

The following characteristics are evident in the Poisson distribution •

Occurrences of each outcome are independent of other outcomes.

- The Poisson distribution is a discrete distribution.
- The mean number of outcomes throughout the experiment must be fixed.
- The Poisson distribution describes discrete outcomes over an interval.
- Outcomes in each interval can range from zero to infinity. μ_i, the mean number of occurrences in the interval, is characterized by the Poisson distribution. The μ_i is the long-run average of the process if a Poisson-distributed occurrence is studied over a long period of time i. The Poisson distribution formula is used to calculate the probability of outcomes over an interval for a fixed μ_i value.

3.8.10 Negative Binomial Model

As the mostly known option to Poisson regression, the negative binomial (NB) regression model, according to Xia, et al. (2012), deals with overdispersion by modelling explicitly the correlated events through a latent variable. They added that the NB extends the Poisson regression model by adding that the mean μ_i of Y_i is not determined only by X_i but also in adding to a heterogeneity component $_i$ independent of X_i.

If $exp(\epsilon_i)$ is assumed to be distributed with a gamma (θ,θ), then the NB model has its density function as

$$P(Y_i = y|X_i) = \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu_i}\right)^{\theta} \left(\frac{\mu_i}{\theta+\mu_i}\right)^{y} \quad y = 0,1,2,... \quad (3.39)$$

Where $\theta = \frac{1}{\alpha}$ which is the dispersion parameter

$$E(Y_i = y|X_i) = \mu_i = E(exp(X_i^T\beta + \epsilon_i)) = E(exp(X_i^T\beta)exp(\epsilon_i)) = exp(X_i^T\beta)$$

and that $E(exp(\epsilon_i)) = 1$ hence $E(exp(X_i^T\beta + \epsilon_i)) = E(exp(X_i^T\beta))$ thus, the expected value, μ_i does not change even if we assume that a negative binomial distribution or a Poisson distribution. Since the dispersion parameter, $\theta > 0$, under the negative binomial distribution, $Var(Y_i = y|X_i) = \mu_i\left(1 + \frac{\mu_i}{\theta}\right) > \mu_i$, implying that $Var(Y_i = y|X_i)/E(Y_i = y|X_i) = 1 + \frac{\mu_i}{\theta}$. This implies that, the variance of the negative binomial is greater than its mean, hence addressing the issue of overdispersion.

θ represents the dispersion parameter in the Negative Binomial distribution, which indicates the degree of over dispersion. The Negative binomial regression model also turns to become a Poisson regression model if $\theta = 0$. Some researchers have considered the use of both Negative Binomial and Poisson models in their studies in several different fields such as Miaou (1994), Xia, et al. (2012), Kibria (2006) and Chipeta, et al.(2014).

3.8.11 Derivation of the Negative Binomial Distribution

Suppose Γ is a gamma (θ,θ) random variable and $Y_i = y|X_i,\Gamma$ is a Poisson random variable. Beginning with a Poisson model, we create a new kind of random variable but making the Poisson model more variable by permitting parameter of the mean to itself be random. Then

$$P(Y_i = y|X_i) = \int_0^\infty P(y,\mu_i)d\mu_i = \int_0^\infty P(y|\mu_i)P(\mu_i)d\mu_i \qquad (3.40)$$

but

$$P(\mu_i|y) = \frac{P(y|\mu_i) \times P(\mu_i)}{P(Y_i = y|X_i)}$$

implying that

$$P(Y_i = y|X_i) = \frac{P(y|\mu_i) \times P(\mu_i)}{P(\mu_i|y)}$$

now

$$P(Y_i = y|X_i) = \frac{P(y|\mu_i) \times P(\mu_i)}{P(\mu_i|y)}$$

$$= \frac{\frac{e^{-\mu_i}\mu_i^y}{y!} \times \frac{\theta^\theta \mu_i^{\theta-1} e^{-\theta\mu_i}}{\Gamma(\theta)}}{\frac{(1+\theta)^{y+\theta}\mu_i^{y+\theta-1}e^{-(1+\theta)\mu_i}}{\Gamma(y+\theta)}}$$

But

$$\theta = \frac{1}{\alpha}$$

$$\mu_i^y \mu_i^{\theta-1} = \mu_i^{(y+\theta-1)}$$

$$e^{-\mu_i}(e^{-\beta\mu_i}) = e^{-(1+\theta)\mu_i}$$

$$P(Y_i = y|X_i) = \frac{\theta^\theta}{(1+\theta)^{y+\theta}} \times \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}$$

$$= \left(\frac{\theta^\theta}{(1+\theta)^\theta}\right)\left(\frac{1}{(1+\theta)^y}\right) \times \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}$$

$$= \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}\left(\frac{\theta}{1+\theta}\right)^\theta \left(\frac{1}{1+\theta}\right)^y$$

But for gamma (θ, θ), $\mu_i = \frac{\theta}{\theta} = 1$.

Substituting it gives

$$P(Y_i = y|X_i) = \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}\left(\frac{\theta}{\mu_i+\theta}\right)^\theta \left(\frac{\mu_i}{\mu_i+\theta}\right)^y \quad (3.41)$$

$$= \binom{y+r-1}{r-1} p^r (1-p)^y \quad (3.42)$$

Where Y_i is the count of the number of failures with known r before the rth success and $0 \leq y < \infty$ and $0 \leq p \leq 1$.

3.8.12 Derivation of the mean of the negative Binomial

$$P(Y_i = y|X_i) = \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}\left(\frac{\theta}{\mu_i+\theta}\right)^\theta \left(\frac{\mu_i}{\mu_i+\theta}\right)^y = \binom{y+r-1}{r-1} p^r (1-p)^y \quad (3.43)$$

And by comparing above,

$$p = \frac{\theta}{\mu_i+\theta} \quad and \quad r = \theta \quad (3.44)$$

$$E(Y_i = y|X_i) = \sum_{y=0}^{\infty} y \binom{y+r-1}{r-1} p^r (1-p)^y \qquad (3.45)$$

$$y\binom{y+r-1}{r-1}p^r = \frac{y(y+r-1)!}{y!(r-1)!}$$

$$= \frac{r(y+r-1)!}{(y-1)!r!}$$

$$= r\binom{y+r-1}{r}$$

$$E(Y_i = y|X_i) = \sum_{y=0}^{\infty} r \binom{y+r-1}{r} p^r (1-p)^y \qquad (3.46)$$

$$= rp^r \sum_{y=0}^{\infty} \binom{y+r-1}{r} (1-p)^y \qquad (3.47)$$

let $k = y - 1$ and $r = s - 1$ then

$$E(Y_i = y|X_i) = rp^r \sum_{k=0}^{\infty} \binom{k+s-1}{s-1} (1-p)^{k+1} \qquad (3.48)$$

$$= rp^r(1-p) \sum_{k=0}^{\infty} \binom{k+s-1}{s-1} (1-p)^k$$

but

$$\sum_{k=0}^{\infty} \binom{k+s-1}{s-1} (1-p)^k = (1-(1-p))^{-s} = p^{-s} = p^{-r-1} \qquad (3.49)$$

implying that

$$E(Y_i = y|X_i) = rp^r(1-p)p^{-r-1} = \frac{r(1-p)}{p} \qquad (3.50)$$

but $p = \frac{\theta}{\mu_i+\theta}$ and $r = \theta$

Implying that

$$E(Y_i = y|X_i) = \frac{\theta\left(1 - \frac{\theta}{\mu_i+\theta}\right)}{\frac{\theta}{\mu_i+\theta}} = \mu_i \qquad (3.51)$$

3.8.13 Derivation of the variance of the negative Binomial

$$P(Y_i = y|X_i) = \binom{y+r-1}{r-1} p^r (1-p)^y \qquad (3.52)$$

Let

$$Var(Y_i = y|X_i) = E[(Y_i = y|X_i)^2] - [E(Y_i = y|X_i)]^2 \tag{3.53}$$

then

$$E(Y_i = y|X_i) = \frac{r(1-p)}{p} = \mu_i \tag{3.54}$$

and

$$E[(Y_i = y|X_i)^2] = \sum_{y=0}^{\infty} y^2 \binom{y+r-1}{r-1} p^r (1-p)^y \tag{3.55}$$

let

$$y\binom{y+r-1}{r-1} = r\binom{y+r-1}{r} \tag{3.56}$$

then

$$E[(Y_i = y|X_i)^2] = r\sum_{y=1}^{\infty} y\binom{y+r-1}{r} p^r (1-p)^y \tag{3.57}$$

Let also $k = y - 1$

$$E[(Y_i = y|X_i)^2] = r\sum_{k=0}^{\infty}(k+1)\binom{k+r}{r} p^r (1-p)^{k+1}$$

$$= r(1-p)\sum_{k=0}^{\infty}(k+1)\binom{k+r}{r} p^r (1-p)^k$$

$$= r(1-p)\left(\sum_{k=0}^{\infty} k\binom{k+r}{r} p^r (1-p)^k + \sum_{k=0}^{\infty}\binom{k+r}{r} p^r (1-p)^k\right.!$$

Let $r = s - 1$, then

$$E[(Y_i = y|X_i)^2] = r(1-p)\left(\sum_{k=0}^{\infty} k\binom{k+s-1}{s-1} p^{s-1}(1-p)^k + \sum_{k=0}^{\infty}\binom{k+s-1}{s-1} p^{s-1}(1-p)^k\right.!$$

$$= \frac{r(1-p)}{p}\left(\sum_{k=0}^{\infty} k\binom{k+s-1}{s-1} p^s (1-p)^k + \sum_{k=0}^{\infty}\binom{k+s-1}{s-1} p^s (1-p)^k\right.!$$

But $\sum_{k=0}^{\infty} k\binom{k+s-1}{s-1} p^s (1-p)^k = 1$ giving

$$E[(Y_i = y|X_i)^2] = \frac{r(1-p)}{p}\left(\frac{s(1-p)}{p} + 1\right) \tag{3.58}$$

But $s = r + 1$

$$E[(Y_i = y|X_i)^2] = \frac{r(1-p)}{p}\left(\frac{(r+1)(1-p)}{p} + 1\right)$$

Now

$$\begin{aligned}
Var(Y_i = y|X_i) &= E[(Y_i = y|X_i)^2] - [E(Y_i = y|X_i)]^2 \\
&= \frac{r(1-p)}{p}\left(\frac{(r+1)(1-p)}{p} + 1\right) - \left(\frac{r(1-p)}{p}\right)^2 \\
&= \frac{r(r+1)(1-p)^2}{p^2} + \frac{pr(1-p) - r^2(1-p^2)}{p^2} \qquad (3.59) \\
&= \frac{r(1-p)(r+1)(1-p) + p - r(1-p)}{p^2} \\
&= \frac{r(1-p)}{p^2}
\end{aligned}$$

But $p = \frac{\theta}{\mu_i + \theta}$ and $r = \theta$

giving

$$Var(Y_i = y|X_i) = \frac{r(1-p)}{p^2} = \frac{\theta\left(1 - \frac{\theta}{\mu_i + \theta}\right)}{\left(\frac{\theta}{\mu_i + \theta}\right)^2} = \mu_i\left(1 + \frac{\mu_i}{\theta}\right) \tag{3.60}$$

3.8.14 Negative Binomial As Exponential Family

Let

$$P(Y_i = y|X_i) = \binom{y + r - 1}{r - 1} p^r (1-p)^y \tag{3.61}$$

The exponential family in canonical form is given by

$$P(Y_i = y|X_i) = exp\left(\frac{y.\lambda - b(\lambda)}{\phi} + c(y, \phi)\right) \tag{3.62}$$

$$P(Y_i = y|X_i) = exp\left(ylog(1-p) + rlogp + log\binom{y + r - 1}{r - 1}\right) \tag{3.63}$$

Let

$$\lambda = log(1 - p)$$

$$e^{\lambda} = 1 - p$$

$$p = 1 - e^{\lambda}$$

$$\phi = 1$$

$$c(y, \phi) = log\binom{y + r - 1}{r - 1}$$

$$E(Y_i = y|X_i) = \frac{\partial b(\lambda)}{\partial \lambda} = \frac{\partial(-rlog(1 - e^{\lambda}))}{\partial \lambda}$$

$$= \frac{-r(-e^{\lambda})}{1 - e^{\lambda}} = \frac{re^{\lambda}}{1 - e^{\lambda}} = \frac{r(1 - p)}{p} = \mu_i$$

$$Var(Y_i = y|X_i) = \frac{\partial^2 b(\lambda)}{\partial \lambda^2} = \frac{\partial}{\partial \lambda}\left(\frac{re^{\lambda}}{1 - e^{\lambda}}\right)$$

$$= \frac{re^{\lambda}}{(1 - e^{\lambda})^2} = \frac{r(1 - p)}{p^2} = \mu_i(1 + \theta\mu_i)$$

$$logp = log(1 - e^{\lambda})$$

$$P(Y_i = y|X_i) = exp\left(\lambda y - (-rlog(1 - e^{\lambda})) + log\binom{y + r - 1}{r - 1}\right) \quad (3.64)$$

Implying that

$$b(\lambda) = -rlog(1 - e^{\lambda})$$

(3.65)

(3.66)

(3.67)

(3.68)

Canonical link

Let

$$\eta = E(Y_i = y|X_i) = \frac{re^{\lambda}}{1 - e^{\lambda}} = \frac{r}{e^{-\lambda} - 1} \quad (3.69)$$ Inversing it

$$\eta^{-1} = \frac{e^{-\lambda} - 1}{r}$$

$$r\eta_{-1} = e_{-\lambda} - 1$$

$$r\eta_{-1} + 1 = e_{-\lambda}$$

$$log(r\eta^{-1}+1) = -\lambda$$

$$log\left(\frac{1}{r\eta^{-1}+1}\right) = \lambda$$

$$log\left(\frac{1}{\frac{r}{\eta}+1}\right) = \lambda$$

$$\frac{r}{\eta}+1 = \frac{r+\eta}{\eta} = \frac{r+E(Y_i=y|X_i)}{E(Y_i=y|X_i)}$$

$$\lambda = log\left(\frac{E(Y_i=y|X_i)}{r+E(Y_i=y|X_i)}\right) = log\left(\frac{\mu_i}{\theta+\mu_i}\right) \qquad (3.70)$$

3.8.15 Negative Binomial Regression Model Estimation

The probability density function is defined previously as

$$P(Y_i=y|X_i) = \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta}\left(\frac{\mu_i}{\mu_i+\theta}\right)^{y} = \frac{\Gamma(y+r)p^r(1-p)^y}{\Gamma(y+1)\Gamma((r))} \qquad (3.71)$$

The likelihood function for the negative binomial model is

$$L = \prod_{i=1}^{N} P(Y_i=y|X_i) = \prod_{i=1}^{N}\left(\frac{\Gamma(y+r)p^r(1-p)^y}{\Gamma(y+1)\Gamma(r)}\right) \qquad (3.72)$$

And the log-likelihood is

$$\ell = \sum_{i=1}^{N} \ln\left(\frac{\Gamma(y+r)p^r(1-p)^y}{\Gamma(y+1)\Gamma((r))}\right) \qquad (3.73)$$

$$= \sum_{i=1}^{N}\ln(\Gamma(y+r)) + \sum_{i=1}^{N}\ln(p^r) + \sum_{i=1}^{N}\ln((1-p)^y) - \sum_{i=1}^{N}\ln(\Gamma(y+1)) - \sum_{i=1}^{N}\ln(\Gamma(r))$$

$$= \sum_{i=1}^{N}\ln(\Gamma(y+r)) + Nr\ln(p) + y\ln(1-p) - \ln(\Gamma(y+1)) - N\ln(\Gamma(r)) \qquad (3.74)$$

To find the maximum likelihood estimates of r and p , we take derivatives of ` with respect to r and p and equate them to zero.

$$\frac{d\ell}{dp} = \frac{Nr}{p} + \sum_{i=1}^{N}\left(\frac{-y}{1-p}\right) = 0 \qquad (3.75)$$

To get the maximum likelihood estimate of p ,

$$p = \frac{Nr}{Nr + \sum_{i=1}^{N} y} \tag{3.76}$$

Differentiating `in relation to r and substituting the maximum likelihood estimate of p to eliminate p from the equation,

$$\frac{d\ell}{dr} = N\ln(p) - N\psi(r) + \sum_{i=1}^{N} \psi(y + r) \tag{3.77}$$

$$\frac{d\ell}{dr} = N\ln\left(\frac{Nr}{Nr + \sum_{i=1}^{N} y}\right) - N\psi(r) + \sum_{i=1}^{N} \psi(y + r) = 0 \tag{3.78}$$

The equation $\frac{d\ell}{dr}$ has no closed-form solution so the root has to be found with numerical methods.

When excess zero counts occur in the outcome variable, the Poisson regression and the Negative Binomial models are not appropriate to fit such data. In this case, the Hurdle models and Zero Inflated models are the appropriate alternatives.

3.8.16 Poisson Hurdle Models

Many count data exhibit more zero observations or counts in addition to over-dispersion than may be permitted by the Poisson model. One type of model that can deal with both over-dispersion and excess zero counts is the Poisson Hurdle, which was proposed by Mullahy (1968).The Hurdle model is a two-state model; a binary model to predict zeros and a zero-truncated component like Poisson to predict non-zero counts.

Its probability density function is given as

$$P(Y_i|X_i,Z_i) = \omega_i \, for \, y = 0$$

$$P(Y_i|X_i,Z_i) = \frac{(1-\omega_i)exp(-\mu_i)\mu_i^y}{(1-exp(-\mu_i))y!} \qquad for \quad y > 0 \tag{3.79}$$

where $\mu_i = exp(X_i^T\beta)$

The variance and mean of a Poisson hurdle model according to Chipeta, et al. (2014) are:

$$Var(Y_i|X_i,Z_i) = \eta(\mu_i - \eta) + \frac{\pi\sigma^2}{1-P(0;\alpha)} \tag{3.80}$$

and

$$E(Y_i|X_i,Z_i) = \eta = \frac{\pi\mu_i}{1-P(0;\alpha)} \tag{3.81}$$

The hurdle model combines a zero-truncated component which is specified more formally $f_{count}(y,X_i,\beta)$ and a hurdle component, that models zero counts, which is also specified as $f_{zero}(y,Z_i,\gamma)$ and which as a result is given by the relation:

$$f_{hurdle}(y,X_i,Z_i,\beta,\gamma) = \begin{cases} f_{zero}(0,Z_i,\gamma) & if \quad y = 0 \\ (1 - f_{zero}(0,Z_i,\gamma)).f_{count}(y,X_i,\beta)/f_{count}(0,X_i,\beta) & if \quad y > 0 \end{cases} \tag{3.82}$$

The parameters γ and β of the Poisson hurdle model can be estimated using the Maximum Likelihood estimation, the advantage is that the zero-truncated component and the hurdle component can be maximized separately by likelihood specification. The likelihood function of the hurdle model has the general form as

$$L = \underset{i=\Omega_0}{Y}\{f_{zero}(0;Z_i,\gamma)\} \underset{i=\Omega_1}{Y}\{(1-f_{zero}(0;Z_i,\gamma)).f_{count}(y,X_i,\beta)/1-f_{count}(0,X_i,\beta)\} \tag{3.83}$$

Where $\Omega_0 = \{i/y = 0\}, \Omega_1 = \{i/y \; 6= 0\}$ and $\Omega_0 \cup \Omega_0 = \{1,2,...,N\}$

The log likelihood, by taking the log of L and rearranging the terms, can be written as

$$` = \sum_{i=\Omega_0} \ln\{f_{zero}(0,Z_i,\gamma)\} + \sum_{i=\Omega_1} \ln\{1 - f_{zero}(0,Z_i,\gamma)\} + \sum_{i=\Omega_1} \ln\{f_{count}(y,X_i,\beta)\} - \ln\{1 - f_{count}(0,X_i,\beta)\}$$

(3.84)

The log-likelihood can at all times be expressed as the sum of the log likelihoods from the two different models because the likelihood function is separable with regards to the parameter vectors β and γ . The log-likelihood of the hurdle model as can always be maximized by maximizing the two components without loss of information. Hence mean regression relationship can be expressed as

$log(\lambda) = X_i^T\beta + log(f_{zero}(0,Z_i,\gamma)) - log(1 - f_{count}(0,X_i,\beta))$ (3.85) by canonical log

link

3.8.17 Zero-inflated Models

Zero inflated Poisson Zero and inflated Negative Binomial are zero-inflated models capable of addressing issues of excess zero counts and overdispersion (Mullay1968; Lambert 1992). The Zero-inflated models are two-state models that have a count distribution following negative binomial or Poisson and a point mass at zero. The zero counts may come from both the count component and the point mass, indicating the two sources of zero counts.

According to Xia, et al. (2012), if $\omega_i = Pr(i \in (structuralzero)|Z_i)$ and $1-\omega_i = Pr(i \in (samplingzero)|Z_i)$, then, Zero Inflated Poisson has its distribution as:

$$P(Y_i|X_i,Z_i) = \omega_i + (1 - \omega_i)exp(-\mu_i) \qquad for \quad y = 0$$

$$P(Y_i|X_i,Z_i) = (1-\omega_i)\frac{exp(-\mu_i)\mu_i^y}{y!} \quad for \quad y > 0 \qquad (3.86)$$

With the variance and the mean of Y_i are given according to Xia, et al. (2012) as

$$Var(Y_i|X_i,Z_i) = \mu_i(1-\omega_i)(1+\mu_i\omega_i) \quad (3.87)$$

and

$$E(Y_i|X_i,Z_i) = \omega_i 0 + \mu_i(1-\omega_i) = \mu_i(1-\omega_i) \quad (3.88)$$

On the other hand, the Zero inflated negative binomial has the form:

$$P(Y_i = y|X_i, Z_i) = \omega_i + (1-\omega_i)\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta} \quad if \quad y = 0$$

$$P(Y_i = y|X_i, Z_i) = (1-\omega_i)\frac{\Gamma(y+\theta)}{y!\Gamma(\theta)}\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta}\left(\frac{\mu_i}{\mu_i+\theta}\right)^{y} \quad if \quad y > 0 \quad (3.89)$$

The variance and mean of the ZINB as posited by to Xia, et al. (2012) are;

$$Var(Y_i|X_i,Z_i) = \mu_i(1-\omega_i)(1+\mu_i(\omega_i+\theta)) \quad (3.90)$$

and

$$E(Y_i|X_i,Z_i) = (1-\omega_i)\mu_i \quad (3.91)$$

More formally, the zero-inflated density function is a combination of a count distribution $f_{count}(y,X_i,\beta)$ and a point mass at zero $I_{(0)}(y)$. The probability of observing a zero count is inflated with a probability

$$\pi = f_{zero}(0;Z_i,\gamma) :$$

$$f_{zeroinfl}(y,X_i,Z_i,\beta,\gamma) = f_{zero}(0,Z_i,\gamma)*I_{(0)}(y)+(1-f_{zero}(0,Z_i,\gamma))*f_{count}(y,X_i,\beta) \quad (3.92)$$

Where the unobserved probability π belong to the point mass component and $I(.)$ is the indicator function. The related mean regression equation is

$$\lambda = \pi_i.0 + (1-\pi_i)exp(X_i^T\beta) \quad (3.93)$$

by the canonical log link.

3.9 Selecting Appropriate Models

3.9.1 Selecting Inflated Models over Traditional

The score, the likelihood ratio test and the Wald test tests and other tests are available for testing the zero inflation in the model (see Vanden Broek 1995 and Lee et. al. 2004). For ease in the selection of zero inflated models over their traditional counterpart, the Vuong statistic will be considered. In defining the Vuong statistic (V), assume $f_1(Y_i = y|X_i)$ and $f_2(Y_i = y|X_i)$ are the probability density functions of Hurdle model or Zero-inflated models and their traditional models (Poisson regression model or Negative Binomial model) respectively and

$F_1(Y_i = y|X_i)$ and $F_2(Y_i = y|X_i)$ represent their corresponding cumulative distribution functions.

The Vuong test statistic (V) is therefore defined as

$$V = \frac{\bar{m}}{S_m/\sqrt{n}} \tag{3.94}$$

Where $\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ and $S_m = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2}$ represent the mean and standard deviation of the measurement of m_i. m_i is defined as;

$m_i = log\left[\frac{\hat{f}_1(Y_i=y|X_i)}{\hat{f}_2(Y_i=y|X_i)}\right]$

Where $\hat{f}_1(Y_i = y|X_i)$ and $\hat{f}_2(Y_i = y|X_i)$ indicate the predicted probabilities of their corresponding models $f_1(Y_i = y|X_i)$ and $f_2(Y_i = y|X_i)$ respectively.

The test is directional, with a value close to zero signifying that neither model fits the data well and a large positive value favours $\hat{f}_1(Y_i = y|X_i)$ and a large negative value favours $\hat{f}_2(Y_i = y|X_i)$ The V statistic will test the following hypotheses:

H_o: Two distribution functions are equivalent

H_a: $F_1(Y_i = y|X_i)$ is worse than $F_2(Y_i = y|X_i)$

H_a : $F_1(Y_i = y|X_i)$ is better than $F(Y_i = y|X_i)$

H_a: Two distribution functions are different

3.9.2 Akaike's Information Criterion (AIC) For Selecting Best Model

Akaike's information criterion (Akaike, 1973) is used to compare models in order to select the best one. The model that maximizes the Kullback-Leibler distance between the model and the truth is the chosen model. The AIC is as a result
defined as;

$$AIC = -2L + 2k \qquad (3.95)$$

Where k indicates the number of parameters in the model and L indicates the log-likelihood. The lowest AIC value among the models is the best fitted model. AIC is used when comparing nonnested models fitted by maximum likelihood to the same data set. The statistic considers model parsimony and penalizing for the number of predictors used in the model, thus, $AIC = -2\log L + 2$ number of parameters. The first term is basically the deviance and the second term is a penalty for the number of parameters. The better the model fit, the smaller the AIC value.

3.9.3 Parameter Estimation

In estimating the parameters used in the models, the maximum likelihood estimation method has been considered. It is therefore important to check the significance of the variables included in the models in other to evaluate the models involved. The regression coefficients estimated have to be statistically significant for a better model. To determine the significance of the regression coefficients, the t-test is usually used.

Chapter 4

ANALYSIS AND RESULTS

4.1 Introduction

In this chapter, the results from the analysis and its findings found from the research are discussed.

4.2 Data

The study was conducted between June, 2015 and January, 2016 in Sefwi Wiawso District Hospital, Sefwi Wiawso, Ghana, and it involved systematic random samples of 1158 participants of age 18-65 years.

Ethical approval for the thesis was obtained from the Sefwi Wiawso District Hospital before the start of the study and individual consent was obtained from the participants. Confidentiality of information was also assured in conjunction with the hospital.

The objectives of this research are to identify the key risk factors contributing to dental caries in adults and to determine an appropriate model suitable for the analysis of DMFTDATA. Information on DMFT calculated, age (in years), gender (0-male, 1-female), dental visit or treatment experience (0-no, 1-yes), frequency of change of toothbrush (0-less than or equal to 3 months, 1more than 3 months), frequency of sweet consumption per day (0-not frequent, 1-frequent), frequency of brushing (0-less or equal to 1, 1-more than 1), use of fluoridated toothpaste (0-no, 1-yes), rinsing habit after every meal with water (0-no, 1-yes), jaw accident (0-no, 1-yes) were recorded using the questionnaire in Appendix 1 and analysed using R and SPSS 16.0. The DMFT calculated was treated as the response variable and the rest as predictor variables. The average age of participants was about 40 years with 54% of them being females and the rest 46% being males. Also with regards to the participants, 26.5% have dental visit/treatment experience, 68.1% consume sweet frequently in a day,

45.8% rinse their month with water after meal, 78.6% use fluoridated toothpaste, 20.9% experience blows to their jaws. 65.6% brush their teeth at most once a day and 34.4% brush their teeth at least twice a day. 27.7% at most change their toothbrush in 3 months and 72.3 at least change their toothbrush in 3 months.

Bar chart of DMFT calculated in Figure 4.1 clearly showed that the DMFT calculated is characterized by many zero-valued observations and also positively skewed. Further screening of the DMFT calculated showed that the variance (5.074) is much greater than the mean (2.25) indicating overdispersion.

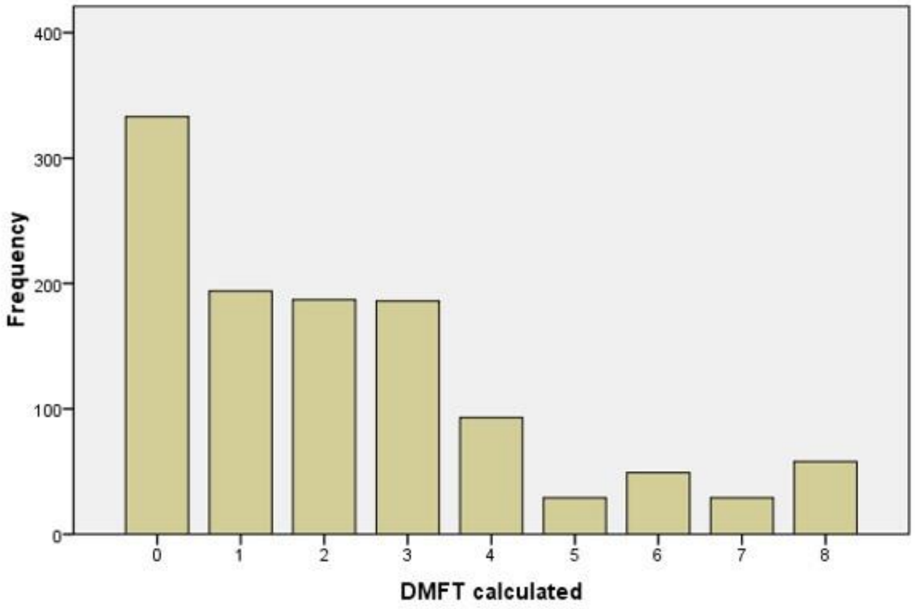


Figure 4.1: Bar chart showing the distribution of DMFT calculated

Also, the distribution of DMFT calculated in Table 4.1 showed that the observed zero counts in the data was 28.8% indicating excess zeros and the observed DMFT calculated has non-negative integer values. From Figure 4.1 and Table 4.1, clearly, the best starting point in analysing DMFTDATA is the use of Poisson regression model since the observed counts are non-negative integers and also the Poisson regression model has some extensions that are useful for count data. The outcome variable occurred in a given period of time and did not assume normality hence the use of Poisson model. Ordinary Least Square (OLS) regression can be used for modelling DMFTDATA when the count outcomes are log transformed but would lead to loss of data because of log transformation of response variable especially when there are zero counts and therefore resulting into biased estimates. And also, normality is assumed for OLS regression.

Table 4.1: Distribution of DMFT Calculated

DMFT Calculated	Frequency	Percentage
0	333	28.8
1	194	16.8
2	187	16.1
3	186	16.1
4	93	8.0
5	29	2.5
6	49	4.2
7	29	2.5
8	58	5.0

The use of Poisson model with the mean estimated from the data is a good starting point of count data modelling. Due to the restrictive assumption of equidispersion regarding the Poisson model, other competing models were fitted to the data.

4.3 The Poisson Regression Model

The Poisson regression model was applied to the DMFTDATA at a significant level of $\alpha = 0.05$. Table 4.2 shows the parameters, estimates, standard errors, z values and p-values from Poisson regression model. From Table 2, all the factors, thus, age, gender, dental visit or treatment experience, frequency of change of toothbrush, frequency of sweet consumption per day, frequency of brushing, use of fluoridated toothpaste, rinsing habit after every meal with water and jaw accident were all found to be significant as contributing factors to dental caries when their p-values were compared to the $\alpha = 0.05$.

Table 4.2: Poisson model showing Parameter, Estimates, Std. Error, z value and $Pr(>|z|)$ at $\alpha = 0.05$

Parameter	Estimate	Std. Error	z-value	Pr(> $\|z\|$)
Intercept	-0.064507	0.117120	-0.551	0.5818
Age	0.012344	0.001925	6.414	1.4×10^{-10}*
Gender	0.092011	0.039934	2.304	0.0212*
Dental Treatment	-0.591670	0.057785	-10.239	$< 2 \times 10^{-16}$*
Sweet Consumption	0.351273	0.050834	6.910	4.8×10^{-12}*
Rinsing habit	0.236086	0.040081	5.890	3.86×10^{-9}*
Fluoridated toothpaste use	-0.191712	0.047967	-3.997	6.42×10^{-5}*
Jaw accident	0.113489	0.045757	2.480	0.0131*
Brushing times	-0.354106	0.063486	-5.578	2.44×10^{-8}*
Toothbrush change	0.332932	0.077337	4.305	1.67×10^{-5}*

* means significant at $\alpha = 0.05$

Due to overdispersion and excess zero counts in the data, as seen in Figure 4.1 and Table 4.1, models such as the negative binomial model which accounts for over-dispersion in the data, Poisson hurdle model which accounts for excess zero counts assumed to come from structural source, zero inflated negative binomial model which accounts for both overdispersion due to heterogeneity and excess zero counts when the zero counts are assumed to come from both structural and chance sources, and zero inflated Poisson models which accounts for overdispersion due to excess zero counts when the zero counts are believed to come from both structural and chance sources were all applied as alternative models either/both for overdispersion and/or excess zero counts. These are to investigate the one state or two state processes of the DMFTDATA and thereafter apply comparison and selection tests to select the appropriate model for the DMFTDATA

4.4 The Negative Binomial Regression Model

The negative binomial regression model was applied to the DMFTDATA at a significant level of $\alpha = 0.05$ to account for overdispersion due to heterogeneity.

Table 4.3: Negative binomial showing parameter Coefficients, Estimates, Std. Error, z value and $Pr(|z|)$ at $\alpha = 0.05$

Parameter	Estimate	Std. Error	z-value	Pr(> $\|z\|$)

Intercept	-0.185677	0.156132	-1.189	0.234350
Age	0.015274	0.002699	5.659	1.53×10^{-8}*
Gender	0.094704	0.055240	1.714	0.086455
Dental treatment	-0.597701	0.072616	-8.231	$< 2 \times 10_{-16}$*
Sweet consumption	0.376873	0.066807	5.641	1.69×10^{-8}*
Rinsing habit	0.223354	0.055306	4.039	5.38×10^{-5}*
Fluoridated toothpaste use	-0.225416	0.066748	-3.377	0.000732*
Jaw accident	0.100375	0.065473	1.533	0.125258
Brushing times	-0.326793	0.082891	-3.942	$8.0710 - 5$*
Toothbrush change	0.344019	0.097318	3.535	0.000408*
θ	2.860			

$*$ means significant at $\alpha = 0.05$

Table 4.3 shows the parameters, estimates, standard errors, z values and p-values from the negative binomial regression model. From Table 4.3, age, dental visit or treatment experience, frequency of change of toothbrush, frequency of sweet consumption per day, frequency of brushing, use of fluoridated toothpaste and rinsing habit after every meal with water were all found to be significant as contributing factors to dental caries when their p-values were compared to the $\alpha = 0.05$. The dispersion parameter $\theta = 2.860$, confirmed that there was overdispersion in the DMFTDATA hence Poisson regression model would be insufficient in modelling DMFTDATA since Poisson dispersion parameter $\theta = 0$.

The 28.8% of zero counts in the DMFTDATA implied the use of zero inflated models to account for excess zero counts.

4.4.1 Zero Inflated Poisson Model

Zero inflated Poisson model is capable of dealing with extra zero counts in which Poisson model cannot handle especially dealing with overdispersion due to excess zero counts hence zero inflated Poisson model is applied to DMFTDATA at a significant level of $\alpha = 0.05$ Table 4.4 shows the parameters, estimates, standard errors, z values and p-values from the Zero inflated Poisson model. From Table 4.4, age, frequency of change of toothbrush, frequency of sweet consumption per day, frequency of brushing and rinsing habit after every meal with water were all found to be significant as contributing factors to dental caries when their p-values were compared to the $\alpha = 0.05$. Also worth noting is when all variables in this model were evaluated at zero, the model is significant at $\alpha = 0.05$.

Table 4.4: Zero inflated Poisson showing parameter Coefficients, Estimates, Std. Error, z value and Pr(>|z|) for both count model and zero inflated model at $\alpha = 0.05$

Parameter	Estimate	Std. Error	z-value	Pr(> $\|z\|$)
Intercept	0.709730	0.148216	4.788	$1.68 \times 10^{-6*}$
Age	0.010753	0.002116	5.083	$3.72 \times 10^{-7*}$
Gender	0.002321	0.041577	0.056	0.95547
Dental treatment	0.056079	0.066422	0.844	0.39851
Sweet consumption	0.165018	0.054615	3.021	0.00252^{*}
Rinsing habit	0.148342	0.041494	3.575	0.00035^{*}
Fluoridated toothpaste use	0.031125	0.051214	0.608	0.54336
Jaw accident	0.123997	0.047534	2.609	0.00909^{*}
Brushing times	-0.599701	0.074112	-8.092	$5.88 \times 10_{-16*}$
Toothbrush change	-0.263445	0.096013	-2.744	0.00607^{*}
Zero-inflation model coefficients (binomial)				
Parameter	Estimate	Std. Error	z-value	Pr(> $\|z\|$)
Intercept	2.07670	0.74478	2.788	0.0053^{*}
Age	-0.01543	0.01383	-1.116	0.2646
Gender	-1.23396	0.29169	-4.230	$2.33 \times 10^{-5*}$
Dental treatment	3.55248	0.37388	9.502	$< 2 \times 10 - 16^{*}$
Sweet consumption	-1.21280	0.27956	-4.338	$1.44 \times 10^{-5*}$
Rinsing habit	-1.58816	0.30198	-5.259	$5.21 \times 10^{-7*}$
Fluoridated toothpaste	2.28966	0.39204	5.840	$1.4 \times 10^{-9*}$
Jaw accident	0.36588	0.34355	1.065	0.2869
Brushing times	-3.25643	0.83048	-3.921	$8.81 \times 10^{-5*}$
Toothbrush change	-4.88095	0.84501	-5.776	$7.64 \times 10^{-9*}$

$*$ means significant at $\alpha = 0.05$.

4.4.2 Zero Inflated Negative Binomial Model

Zero inflated negative binomial model is capable of dealing with extra zero counts in which negative binomial model cannot handle especially dealing with overdispersion due to excess zero counts and heterogeneity especially when the zero counts are believed to come from both structural and chance sources hence zero inflated negative binomial model is applied to DMFTDATA at a significant level of $\alpha = 0.05$.

Table 4.5 shows the parameters, estimates, standard errors, z values and p-values from the Zero inflated negative binomial model. From Table 4.5, age, frequency of sweet consumption per day, rinsing habit after every meal with water, jaw accident and brushing times were all found to be significant as contributing factors to dental caries when their p-values were compared to the $\alpha = 0.05$. When all variables in the zero inflated negative binomial model were evaluated at zero, the model is significant at $\alpha = 0.05$. Assessment of the dispersion parameter θ, showed that there was overdispersion due to excess zero since θ is significant at $\alpha = 0.05$.

Table 4.5: Zero Inflated Negative Binomial Showing Parameter Coefficients, Estimates, Std. Error, z value and $Pr(> z)$ for both Count Model and Zero Inflated Model at $\alpha = 0.05$

Count model Coefficients (negbin with log link)				
Parameter	Estimate	Std. Error	z-value	$Pr(>\lvert z\rvert)$
Intercept	0.494766	0.202835	2.439	0.0147^{*}
Age	0.012833	0.002744	4.676	$2.92 \times 10^{-6*}$
Gender	-0.006316	0.050420	-0.125	0.9003
Dental Treatment	0.046556	0.081162	0.574	0.5662
Sweet consumption	0.215668	0.066069	3.264	0.0011^{*}
Rinsing habit	0.147140	0.051261	2.870	0.0041^{*}
Fluoridated toothpaste use	0.020552	0.062294	0.330	0.7415
Jaw accident	0.118588	0.058808	2.017	0.0467^{*}
Brushing times	-0.608800	0.084933	-7.168	$7.61 \times 10_{-13*}$
Toothbrush change	-0.171539	0.122214	-1.404	0.1604
$\log(\theta)$	1.854763	0.178263	10.405	$< 2 \times 10_{-16*}$
Zero-Inflation Model Coefficients (Binomial with Logit Link)				
Intercept	2.0096369	0.8483111	2.369	0.01784^{*}
Age	-0.000444	0.0200270	-0.022	0.98230
Gender	-1.598473	0.3989885	-4.006	$6.17 \times 10^{-5*}$
Dental treatment	4.2478853	0.6366233	6.673	$2.51 \times 10_{-11*}$
Sweet Consumption	-1.096493	0.3423760	-3.203	0.00136^{*}
Rinsing habit	-1.841739	0.3584712	-5.138	$2.78 \times 10^{-7*}$
Fluoridated toothpaste use	2.4995587	0.4213329	5.933	$2.98 \times 10^{-9*}$
Jaw accident	0.4215990	0.4310619	0.978	0.32805
Brushing times	-4.656025	1.1365804	-4.097	$4.19 \times 10^{-5*}$
Toothbrush change	-6.074075	1.0261714	-5.919	$3.24 \times 10^{-9*}$
* means significant at $\alpha = 0.05$				

4.4.3 Poisson Hurdle Model

Poisson hurdle model is capable of dealing with extra zero counts in which Poisson model cannot handle especially dealing with overdispersion due to excess zero counts which are believed to come from structural sources hence Poisson hurdle model was applied to DMFTDATA at a significant level of $\alpha = 0.05$. Table 4.6 showed the parameters, estimates, standard errors, z-values and p-values from the Zero inflated Poisson model. From Table 4.6, age, gender, dental visit or treatment experience, frequency of change of toothbrush, frequency of sweet consumption per day, frequency of brushing, use of fluoridated toothpaste and rinsing habit after every meal with water were all found to be significant as contributing factors to dental caries when their p-values were compared to the $\alpha = 0.05$. Also worth noting is that when all variables in this model were evaluated at zero, the model was significant at $\alpha = 0.05$.

Table 4.6: Poisson Hurdle showing parameter Coefficients, Estimates, Std. Error, z value and $Pr(>|z|)$ for both count model and zero inflated model at $\alpha = 0.05$ Count Model Coefficients (truncated poisson with log link)

Parameter	Estimate	Std. Error	z-value	Pr(> $\|z\|$)
Intercept	0.725658	0.147604	4.916	8.82×10^{-7}*
Age	0.011092	0.002173	5.105	3.32×10^{-10}*
Gender	-0.007831	0.043792	-0.179	0.8581
Dental treatment	0.037822	0.067462	0.561	0.5750
Sweet Consumption	0.024328	0.054704	0.445	0.6565
Rinsing habit	0.122532	0.043847	2.795	0.0052*
Fluoridated toothpaste use	0.008041	0.054190	0.148	0.8820
Jaw accident	0.100041	0.049805	2.009	0.0446*
Brushing times	-0.373023	0.085936	-4.341	1.42×10^{-5}*
Tootbrush change	-0.162238	0.102320	-1.586	0.1128
Zero hurdle model coefficients (binomial with logit link)				
Parameter	Estimate	Std. Error	z-value	Pr(> $\|z\|$)
Intercept	-0.588521	0.469452	-1.254	0.20997
Age	0.025407	0.009106	2.790	0.00527*
Gender	0.741820	0.181476	4.088	4.36×10^{-5}*
Dental treatment	-2.358197	0.201381	-11.710	$< 2 \times 10^{-16}$*
Sweet Consumption	1.632692	0.188961	8.640	$< 2 \times 10^{-16}$*
Rinsing habit	1.080031	0.190177	5.679	1.35×10^{-8}*
Fluoridated toothpaste use	-1.319496	0.236760	-5.573	2.50×10^{-8}*
Jaw accident	0.088776	0.241051	0.368	0.71266
Brushing times	-0.607081	0.239426	-2.536	0.01123*
Toothbrush change	1.173696	0.243159	4.827	1.39×10^{-6}*
* means significant at $\alpha = 0.05$				

4.5 Models Comparison and Selection

The Vuong test statistic was used to compare the models used in this research since the models are non-nested models and were fit to the same data. Since many zero valued models were compared to their traditional counterparts, excess zeros were also tested. AIC was used to compare non-nested models fitted by maximum likelihood to the same data set. $AIC = -2\log L + 2\ number\ of\ parameters.$ Table 4.7 showed the test statistics for Poisson, NB, ZIP, ZINB and PHURDLE.

Table 4.7: Test Statistics for Poisson, NB, ZIP, ZINB and PHURDLE

Test Statistics					
Parameter	POISSON	NB	ZIP	ZINB	PHURDLE
AIC	4500.309	4307.981	4123.256	4066.447	4124.996
Theta		2.860		6.3902*	

Vuong test			8.34683^*	9.124266^*	7.750094^*

* means significant at $\alpha = 0.05$

From Table 4.7, comparing the Poisson model to the ZIP model, the Poisson model to the PHURDLE model and NB model to the ZINB model using the Vuoung test, the results from the test statistics, V=7.750094 for PHURDLE versus Poisson $V = 9.124233$ for ZINB versus NB and $V = 8.346832$ for ZIP versus Poisson, showed that PHURDLE ZINB and ZIP offered a better fit as compared to their traditional counterpart models with one-component data. There was evidence also of overdispersion in the DMFTDATA due to excess zero counts as confirmed by the dispersion parameter $\theta = 2.86$ from NB and $\theta = 6.39$ from ZINB (Note: The dispersion parameter α is equal to the inverse of the dispersion parameter θ in R).

Though PHURDLE, ZIP and ZINB were statistically significant models for analysing DMFTDATA, we need the best model amongst these models for analysing DMFTDATA. The AIC was therefore used to select the best model amongst the models.

The AIC obtained from the models were in the following ascending order:

ZINB >> ZIP >> PHURDLE >> NB >> POISSON with AIC values as

4066.447, 4123.256, 4124.996, 4307.981 and 4500.309 respectively.

Thus, though PHURDLE, ZIP and ZINB were statistically significant models for analysing DMFTDATA, we need the best model amongst these models for analysing DMFTDATA. The AIC was therefore used to select the best model amongst the models. Under the AIC, ZINB was the appropriate model amongst the models considered.

4.6 Findings

The response variable, DMFT calculated, in the DMFTDATA was characterized by preponderance of zeros of 28.8% which is evident by the Vuong test that favoured two component models as against one component models. The dispersion parameter θ = 6.39 from ZINB also indicated that there was overdispersion in the DMFTDATA due to excess zeros in the data. Modelling the DMFTDATA with the various regression models showed agreement with Lee, et al., (2007) who asserted that the ZIP and ZINB were known to provide robust statistics especially when zero counts are present and in addition to Poisson Hurdle models. Likewise, there was an agreement with Lambert (1992) who also found the ZIP model to be better than the NB model, and that the NB model was also better than the Poisson model. The AIC value of the ZINB of 4066.447 put the ZINB as the preferred model for the analysis of DMFTDATA.

Though Javali and Pandit (2010) put forward that duration of change of toothbrush, frequency of brushing, systematic toothpaste, rinsing habit, frequency of sweet consumption and smoking habit were found to have significant influences on dental caries, there was a disagreement on duration of change of toothbrush and use of systemic toothpaste. But largely, we have the same opinion on sweet consumption, rinsing habit and frequency of brushing. And to add that age and jaw accident are also significant risk factors of dental caries.

Chapter 5

CONCLUSION and RECOMMENDATIONS

5.1 Summary of Findings

The DMFTDATA contained 28.8% of excess zeros and overdispersed due to excess number of the zeros which was confirmed by θ = 6.39 in ZINB model. Poisson Hurdle, zero-inflated Poisson, zero-inflated negative binomial, Negative Binomial and Poisson models were statistical tools used in analysing the DMFTDATA to identify key risk factors of dental caries and to determine an appropriate model suitable for the analysis of DMFTDATA. The Vuong test statistic showed that Poisson Hurdle, zero-inflated Poisson and zero-inflated negative binomial models which are two-component models were preferred to one-component models like the Negative Binomial and Poisson models for the analysis of DMFTDATA which implied that the DMFTDATA contained excess zeros. To select the best model among Poisson Hurdle, zero-inflated Poisson and zero-inflated negative binomial models, the AIC was used. The AIC values showed that the ZINB model is the most appropriate model suitable for the analysis of DMFTDATA since the ZINB model has the least AIC value. Analysis of DMFTDATA with ZINB model showed that age (in years), jaw accident, frequency in sweet consumption, rinsing habit after meal with water and frequency of brushing were significant risk factors of dental caries.

5.2 Conclusions

Collecting data on dental caries incidence is an essential field in oral epidemiology since dental caries is a severe issue in public health. In this thesis, an appropriate model suitable for the analysis of DMFTDATA was determined. The Poisson model for count data modelling was a good starting point but has a restrictive equidispersion assumption. The Negative Binomial regression model with a more relaxed assumption on variance provided a better solution with the evidence of over-dispersion in the DMFTDATA since the variance was greater than the mean. Advanced composite models such as ZINB, Poisson Hurdle and ZIP regression models gave a more suitable fit to the data with over-dispersion as a result of a high frequency of zero counts. Based on the AIC and Vuong test values, the ZINB was therefore the appropriate model suitable for the analysis of DMFTDATA characterised with overdispersion and many zero-values. It was found that age (in years), jaw accident, frequency in sweet consumption, rinsing habit after meal with water and frequency of brushing were significant risk factors of dental caries with 28.8% of excess zeros in the DMFTDATA.

5.3 Recommendation

This thesis is of great contribution to oral epidemiology especially, dental caries. Count models were discussed and comparison tests like the Vuong test and the AIC were also performed. It is therefore recommended that:

1. Adults should reduce intake of sugar foods and consider consumption of free sugar foods since the intake of sugar foods is a significant cause of dental caries in adults.

2. Adults should endeavour to brush their teeth with a fluoridated toothpaste at least twice daily.

3. Mouth rinsing with water after meal should be encouraged in adults.

4. Proper dental care should be considered for adults with jaw accident.

5. Age should be considered in planning dental care involving adults.
6. Adequate amount of dentists should be tranined to provide dental care services and these care services should be sufficiently financed since adult caries accounts for about 80% of the dental care costs relating to caries.

7. Preventive programs such as water fluoridation and accessibility of low sugar foods to avoid this disease should be put in place.

8. This thesis will create awareness of risk factors of dental caries in adults which will help in planning for dental care.

9. This thesis can be an important reference material for further research.

REFERENCES

Agresti, A. (1996). Categorical data analysis (Vol. 996). New York: John Wiley & Sons.

Akaike information criterion.Retrieved from Wikipedia

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika, 60(2), 255-265.

Akaike, H. (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19(6), 716-723.

Arens, U. (1999). Oral health: diet and other factors. Nutrition Bulletin, 24(1), 41-44

Arulampalam, W., & Booth, A. L. (1997). Who gets over the training hurdle? A study of the training experiences of young men and women in Britain. Journal of Population Economics, 10(2), 197-217.

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. Journal of Family Psychology, 21(4), 726.

Beltra´n-Valladares, P., Cocom-Tum, H., Casanova-Rosado, J. F., VallejosSa´nchez, A. A., Medina-Solís, C. E., & Maupome, G. (2006). Caries prevalence and some associated factors in 6-9-year-old schoolchildren in Campeche, Mexico. Rev Biome´d, 17, 25-33.

Ben, L. (2014, September 8). 43 - Prior predictive distribution (a negative binomial) for gamma prior to poisson likelihood 2 [Video file] Retrieved from www.youtube.com

Bernabe´, E., & Sheiham, A. (2014a). Age, period and cohort trends in caries of permanent teeth in four developed countries. American journal of public health, 104(7), e115-e121.

Bernabe´, E., & Sheiham, A. (2014b). Extent of differences in dental caries in permanent teeth between childhood and adulthood in 26 countries. International dental journal, 64(5), 241-245.

Boucher, J. P., Denuit, M., & Guille´n, M. (2007). Risk Classification for Claim Counts: A Comparative Analysis of Various Zeroinflated Mixed Poisson and Hurdle Models. North American Actuarial Journal, 11(4), 110-131.

Broadbent, J. M., & Thomson, W. M. (2005). For debate: problems with the DMF index pertinent to dental caries data analysis. Community dentistry and oral epidemiology, 33(6), 400-409. Broadbent, J. M., Page, L. F., Thomson, W. M., & Poulton, R. (2013). Permanent dentition caries through the first half of life. British dental journal, 215(7), E12-E12.

Broadbent, J. M., Thomson, W. M., & Poulton, R. (2008). Trajectory patterns of dental caries experience in the permanent dentition to the fourth decade of life. Journal of Dental Research, 87(1), 69-72.

Burnham, K. P., & Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife research, 28(2), 111-119.

Burnham, K. P.; Anderson, D. R. (2002), Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.), SpringerVerlag, ISBN 0-387-95364-7.

Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: methods and applications. Cambridge university press.

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

Cappelli, D. P., & Mobley, C. C. (2007). Prevention in clinical oral health care. Elsevier Health Sciences.

Casanova-Rosado, A. J., Medina-Solís, C. E., Casanova-Rosado, J. F., VallejosSánchez, A. A., Maupomé, G., & Ávila-Burgos, L. (2005). Dental caries and associated factors in Mexican schoolchildren aged 6–13 years. Acta odontologica Scandinavica, 63(4), 245-251.

Chipeta, M. G., Ngwira, B. M., Simoonga, C., & Kazembe, L. N. (2014). Zero adjusted models with applications to analysing helminths count data. BMC research notes, 7(1), 856.

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, 829-844.

Crépon, B., & Duguet, E. (1997). Estimating the innovation function from patent numbers: GMM on count panel data. Journal of Applied Econometrics, 12(3), 243-263.

Cypriano, S., Sousa, M. D. L. R. D., & Wada, R. S. (2005). Evaluation of simplified DMFT indices in epidemiological surveys of dental caries. Revista de Saúde Pública, 39(2), 285-292.

Dalrymple, M. L., Hudson, I. L., & Ford, R. P. K. (2003). Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. Computational Statistics & Data Analysis, 41(3), 491-504.

Deb, P. and Trivedi, P. (1997), Demand for Medical Care by the Elderly: A Finite Mixture Approach, Journal of Applied Econometrics, Vol. 12, No. 3, 313-336.

Deirdre, T. (2013, February 2). Poisson distribution: a member of the exp family. [Video file]. Retrieved from www.youtube.com

Deirdre, T.(2015, December 7). Negative Binomial - a member of the Natural Exponential Family.[Video file]. Retrieved from www.youtube.com

Dental caries. Retrieved from https://en.wikipedia.org/wiki/Dental caries

DTUbroadcast. (2013, December 15). EXTRA MATH 6C: Maximum likelihood estimation for the poisson model.[Video file]. Retrieved from https://www.youtube.com/watch?v=lHxDy2Ddwnk

Engineer Clearly. (2012, March 14). Poisson Distribution Derivation. [Video file]. Retrieved from https://www.youtube.com/watch?v=yR0hX7C1pYU

Fejerskov, O., & Kidd, E. (Eds.). (2009). Dental caries: the disease and its clinical management. John Wiley & Sons.

Fontana, M., & Zero, D. T. (2006). Assessing patients' caries risk. The Journal of the American Dental Association, 137(9), 1231-1239.

For the dental patient. Tackling tooth decay. (2013)

Forshee, R. A., & Storey, M. L. (2004). Evaluation of the association of demographics and beverage consumption with dental caries. Food and chemical toxicology, 42(11), 1805-1816.

Fujiwara, T. (2005). Etiology and clinical symptoms of dental caries. FOODS AND FOOD INGREDIENTS JOURNAL OF JAPAN, 210(4), 305.

García-Cortés, J. O., Medina-Solis, C. E., Loyola-Rodriguez, J. P., Mejía-Cruz, J. A., Medina-Cerda, E., Patiño-Marín, N., & Pontigo-Loyola, A. P. (2009). Dental caries' experience, prevalence and severity in Mexican adolescents and young adults. Revista de Salud Publica, 11(1), 82-91.

Greene, W. (2005). Functional form and heterogeneity in models for count data. Foundations and Trends in Econometrics, 1(2), 113–218.

Greene, W. (2007). Functional form and heterogeneity in models for count data. Now Publishers Inc.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models.

Gurmu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. Journal of applied econometrics, 12(3), 225-242.

Gurmu, S. (1998). Generalized hurdle count data regression models. Economics Letters, 58(3), 263-268.

Gurmu, S., & Trivedi, P. K. (1996). Excess zeros in count models for recreational trips. Journal of Business & Economic Statistics, 14(4), 469-477.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. The Annals of Mathematical Statistics, 11(3), 271-283.

Inoue, A., & Kilian, L. (2006). On the selection of forecasting models. Journal of Econometrics, 130(2), 273-306.

Iqbal, S. (15 Jan 2016, January 16). Negative Binomial Distribution: Expected Value.[Video file]. Retrieved from https://www.youtube.com/watch?v=yEZtLnulyg4

Iqbal, S. (16 Jan 2016, January 16). Negative Binomial Distribution: Variance.[Video file]. Retrieved from https://www.youtube.com/watch?v=U73kPsJZwVM

James, W. P. T., & Sheiham, A. Comments on the WHO Draft Guideline:

Sugars intake for adults and children.

Javali, S. B., & Pandit, P. V. (2010). Using zero inflated models to analyze dental caries with many zeroes. Indian Journal of Dental Research, 21(4), 480.

Jeremy, B. (2013, July 27). The Poisson Distribution: Mathematically Deriving the Mean and Variance.[Video file].

Kadane, J. B., & Lazar, N. A. (2004). Methods and criteria for model selection. Journal of the American statistical Association, 99(465), 279-290.

Kibria, B. G. (2006). Applications of some discrete regression models for count data. Pakistan Journal of Statistics and Operation Research, 2(1), 1-16.

Kidd, E., & Fejerskov, O. (2013). Caries control in health service practice. Primary dental journal.

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

Larmas, M. (2010). Has dental caries prevalence some connection with caries index values in adults?. Caries research, 44(1), 81-84.

Lawal, B. (2003). Categorical data analysis with SAS and SPSS applications.

Lee, E. (2005). World Health Organization's Global Strategy on Diet, Physical Activity, and Health: Turning Strategy into Action, The. Food & Drug LJ, 60, 569.

Lee, Y. G., Lee, J. D., Song, Y. I., & Lee, S. J. (2007). An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST. Scientometrics, 70(1), 27-39.

Lesaffre, E., Mwalili, S. M., & Declerck, D. (2004). Analysis of caries experience taking inter-observer bias and variability into account. Journal of dental research, 83(12), 951-955.

Lin, H. C., Wong, M. C. M., Zhang, H. G., Lo, E. C. M., & Schwarz, E. (2001). Coronal and root caries in Southern Chinese adults. Journal of dental research, 80(5), 1475-1479.

Liu, W., & Cela, J. (2008, March). Count data models in SAS. In SAS Global Forum (Vol. 317, pp. 1-12).

Long, J. S. (1997). Regression models for limited and categorical dependent variables.

Luan, W. M., Baelum, V., Fejerskov, O., & Chen, X. (2000). Ten–Year Incidence of Dental Caries in Adult and Elderly Chinese. Caries research, 34(3), 205213.

MacGregor, A. B. (1964). Diet and Dental Disease in Ghana: Hunterian Lecture delivered at the Royal College of Surgeons of England on 24th May 1963. Annals of the Royal College of Surgeons of England, 34(3), 179.

Mahyudin, F., & Hermawan, H. (Eds.). (2016). Biomaterials and Medical Devices: A Perspective from an Emerging Country (Vol. 58). Springer.

Mazerolle, M. J. (2004). Making sense out of Akaike's Information Criterion (AIC): its use and interpretation in model selection and inference from ecological data. World Wide Web.< http://www. theses. ulaval. ca/2004/21842/apa. html.

McCullagh, P. (1992). Introduction to Nelder and Wedderburn (1972) Generalized Linear Models. In Breakthroughs in Statistics (pp. 543-546).
Springer New York.

McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (Vol. 37). CRC press.

Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions.
Accident Analysis & Prevention, 26(4), 471-482.

Miller, J. M. (2007). Comparing Poisson, Hurdle, and ZIP model fit under varying degrees of skew and zero-inflation (Doctoral dissertation, University of Florida).

Min, Y., & Agresti, A. (2002). Modeling nonnegative data with clumping at zero: A survey. Journal of Iranian Statistical Society, 1(1), 7-33.

Min, Y., & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. Statistical Modelling, 5(1), 1-19.

Minami, M., Lennert-Cody, C. E., Gao, W., & Roman-Verdesoto, M. (2007). Modeling shark bycatch: the zero-inflated negative binomial regression model with smoothing. Fisheries Research, 84(2), 210-221.

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

Namal, N., Can, G., Vehid, S., Koksal, S., & Kaypmaz, A. (2008). Dental health status and risk factors for dental caries in adults in Istanbul, Turkey. Nowjack-Raymer, R. E., & Sheiham, A. (2007). Numbers of natural teeth, diet, and nutritional status in US adults. Journal of dental research, 86(12), 1171-1175.

Pardoe, I., & Durham, C. A. (2003). Model choice applied to consumer preferences. In Proceedings of the 2003 Joint Statistical Meetings. Citeseer.

Petersen, P. E. (2003). The World Oral Health Report 2003: continuous improvement of oral health in the 21st century–the approach of the WHO Global Oral Health Programme. Community Dentistry and oral epidemiology, 31(s1), 3-24.

Petersen, P. E., Bourgeois, D., Ogawa, H., Estupinan-Day, S., & Ndiaye, C. (2005). The global burden of oral diseases and risks to oral health. Bulletin of the World Health Organization, 83(9), 661-669.

Pohlmeier, W., & Ulrich, V. (1995). An econometric model of the two-part decisionmaking process in the demand for health care. Journal of Human Resources, 339-361.

Saffari, S. E., Adnan, R., & Greene, W. (2011). Handling of over-dispersion of count data via truncation using Poisson regression model. Journal of Computer Science and Computational Mathematics, 1(1), 1-4.

Saffari, S. E., Adnan, R., & Greene, W. (2012). Hurdle negative binomial regression model with right censored count data. SORT: statistics and operations research transactions, 36(2), 181-194.

Santini, A. (2013). The effective management of caries. Primary dental journal, 2(3), 5.

Segovia-Villanueva, A. Estrella-Rodriguez, Medina-Solis, C.E., & Maupome, G. (2006). Dental caries experience and factors among preshoolers in Southeastern Mexico: A brief communication. Journal of public health dentistry, 66(2), 88-91

Shah, N., & N., Sundaram K.R. (2004).Impact of socio-demographic variables, oral hygiene practices, oral habits and diet on dental caries experience of Indian elderly: a community-based study. Gerodontology, 21(1),43-50

Sheu, M.L.,Hu, T.W., Keeler, T.E., Ong, M & Sung, H.Y. (2004). The effect of a major cigarette price change on smoking behaviour in California: a zero-inflated negative binomial model. Health Economics, 13(8),781-791

Shonkwiler, J. S., & Shaw, W. D. (1996). Hurdle count-data models in recreation demand analysis. Journal of Agricultural and Resource Economics, 210-219.

Silva, R. P. D., Assaf, A. V., Ambrosano, G. M. B., Mialhe, F. L., Meneghim, M. D. C., & Pereira, A. C. (2015). Different methods of dental caries diagnosis in an epidemiological setting. Brazilian Journal of Oral Sciences, 14(1), 78-83.

Slymen, D. J., Ayala, G. X., Arredondo, E. M., & Elder, J. P. (2006). A demonstration of modeling count data with an application to physical activity. Epidemiologic Perspectives & Innovations, 3(1), 3.

Splieth, C. H., Schwahn, C. H., Bernhardt, O., Kocher, T., Born, G., John, U., & Hensel, E. (2002). Caries prevalence in an adult population: results of the Study of Health in Pomerania, Germany (SHIP). Oral health & preventive dentistry, 1(2), 149-155. Standards for clinical dental hygiene practice. The Association, 2008.

Szoke, J., & Petersen, P. E. (2004). [State of oral health of adults and the elderly in Hungary]. Fogorvosi szemle, 97(6), 219-229.

Team, R. C. (2015). R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. Document freely available on the internet at: http://www. r-project. org.

Tin, A. (2008). Modeling zero-inflated count data with underdispersion and overdispersion. In SAS Global Forum Proceedings.

Treasure, E., Kelly, M., Nuttall, N., Nunn, J., Bradnock, G., & White, D. (2001). Factors associated with oral health: a multivariate analysis of results from the 1998 Adult Dental Health survey. British dental journal, 190(2), 60-68.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. Biometrics, 738-743.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. Biometrics, 738-743.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica: Journal of the Econometric Society, 307-333.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., & Lindenmayer, D. B. (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecological Modelling, 88(1), 297-308.

Winkelmann, R. (2004). Health care reform and the number of doctor visits—an econometric analysis. Journal of Applied Econometrics, 19(4), 455-472.

Winn, D. M., Brunelle, J. A., Selwitz, R. H., Kaste, L. M., Oldakowski, R. J., Kingman, A., & Brown, L. J. (1996). Coronal and root caries in the dentition of adults in the United States, 1988-1991. Journal of dental research, 75, 642-651.

World Health Organization. (1962). Standardization of reporting of dental diseases and conditions: report of an Expert Committee on Dental Health [meeting held in Geneva from 14 to 20 November 1961]. World Health Organization. (1997). Oral health surveys-basic methods 4th edition.

Geneva: WHO.

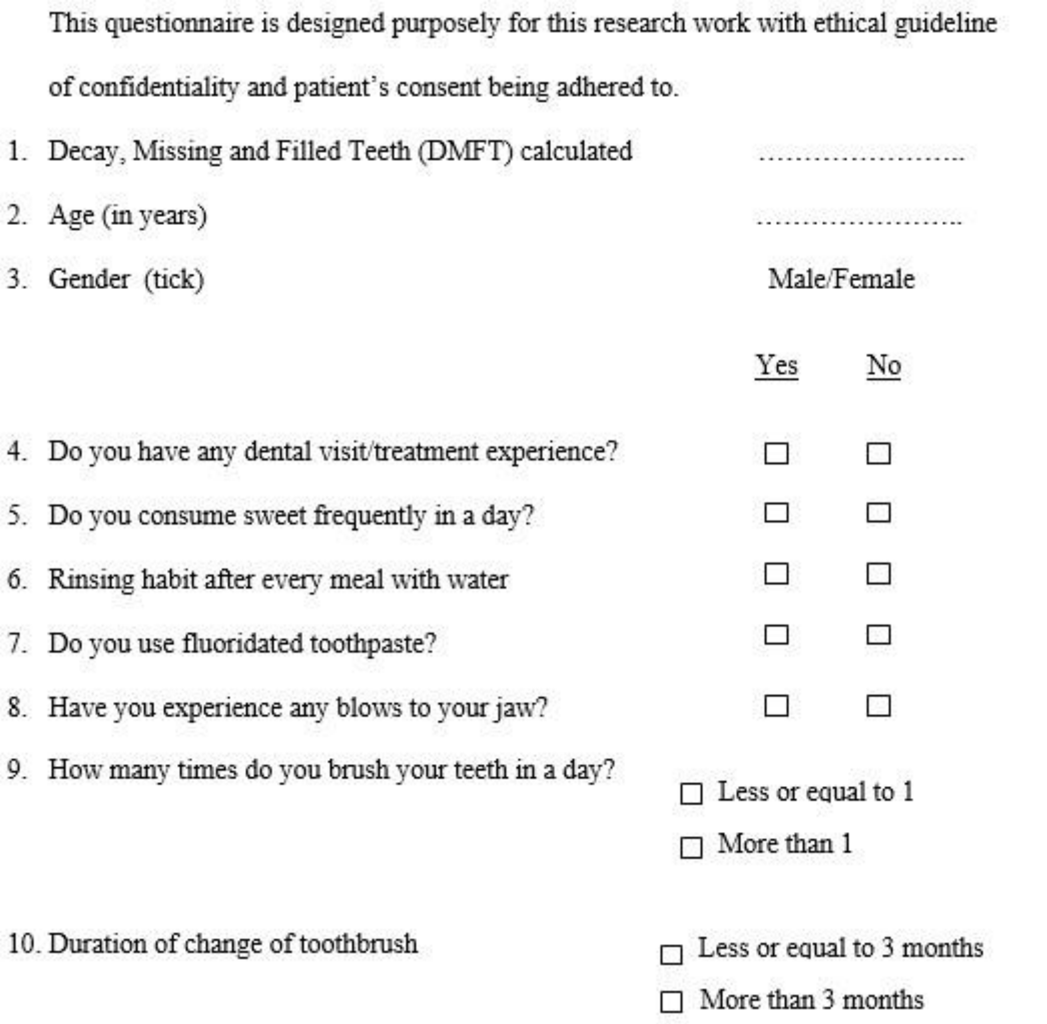
Xia, Y., Morrison-Beedy, D., Ma, J., Feng, C., Cross, W., & Tu, X. (2012). Modeling count outcomes from HIV risk reduction interventions: a comparison of competing statistical models for count responses. AIDS research and treatment, 2012.

Xu, L., Paterson, A. D., Turpin, W., & Xu, W. (2015). Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. PloS one, 10(7), e0129606.

Zeileis, A., Kleiber, C., & Jackman, S. (2007). Regression models for count data in R. Zorn, C. J. (1996). Evaluating zero-inflated and hurdle Poisson specifications. Midwest Political Science Association, 18(20), 1-16.

Appendix 1 DENTAL CARIES

QUESTIONNAIRE



This questionnaire is designed purposely for this research work with ethical guideline of confidentiality and patient's consent being adhered to.

1. Decay, Missing and Filled Teeth (DMFT) calculated ……………………..
2. Age (in years) ……………………..
3. Gender (tick) Male/Female

	Yes	No
4. Do you have any dental visit/treatment experience?	□	□
5. Do you consume sweet frequently in a day?	□	□
6. Rinsing habit after every meal with water	□	□
7. Do you use fluoridated toothpaste?	□	□
8. Have you experience any blows to your jaw?	□	□

9. How many times do you brush your teeth in a day?
   □ Less or equal to 1
   □ More than 1

10. Duration of change of toothbrush
    □ Less or equal to 3 months
    □ More than 3 months