Journal of Biostatistics and Biometric Applications
Volume 3 | Issue 2
ISSN: 2455-765X

Research article Open Access

Modelling the Occurrence of Dental Carries in Adult Population in Ghana; a Comparison of Competing Count Regression Models

Mensah IA[1], Alhassan EA[*2], Affi PO[3], Baah A[4], and Sarfo BKO[4]

[1]Faculty Science, Jiangsu University, Zhenjiang, Jiangsu, China
[2]Department of Mathematics University for Development Studies, Navrongo, Ghana
[3]Department of Statistics, University of Ghana, Legon, Ghana
[4]School of Finance and Economics, Jiangsu University, Zhenjiang, Jiangsu, China

[*]**Corresponding author:** Alhassan EA, Department of Mathematics University for Development Studies, Navrongo, Ghana, Tel: +233208726332, E-mail: aelvis@uds.edu.gh

Citation: Mensah IA, Alhassan EA, Affi PO, Baah A, Sarfo BKO (2018) Modelling the Occurrence of Dental Carries in Adult Population in Ghana; A Comparison of Competing Count Regression Models. J Biostat Biometric App 3(2): 201

Abstract

Dental carries is a disease with both high prevalence and severity in adults' worldwide population. Dental carries affect over half of the population in both developed and developing countries globally. Since the outcomes from dental carries are mostly discrete, they are often modelled using count regression models. It is therefore the purpose of this study to determine the appropriate count regression model that efficiently fits the Decayed Mining and Filled Teeth Data (DMFTDATA) and further identify the key risk factors contributing significantly to dental carries in adults in Ghana using the appropriate count regression model. The study was conducted between June 2015 through to June 2016 at the Sefwi Wiawso Government Hospital as a study case in Ghana. A cross sectional sample design was employed using clinical carries examination and questionnaire with interview. A systematic random sampling technique was then applied to the population of the hospital to obtain a sample of 1158 individuals. The average age of the participants was found to be 40 years with 54 percent being females and 46 percent being males. The count outcomes obtained were modelled using count regression models which included Poisson, Negative Binomial, Zero-Inflated Negative Binomial (ZINB), Zero-Inflated Poisson (ZIP) and Poisson Hurdle regression models. In order to compare the performance and the efficiency of the listed count regression models with respect to the DMFTDATA, the various model selection methods such as the Vuong statistic (V) and Akaikes Information Criterion (AIC) were used. The ZINB count regression model with reference to the values of the Vuong statistic and AIC was selected as the most appropriate and efficient count regression model for modelling dental carries using the DMFTDATA. Base on the ZINB model the variables age, jaw accident, frequency of sweet consumption, rinsing habit after meal with water and frequency of brushing were significant risk factors contributing to dental caries in adults.

Keywords: Dental Carries; Poisson; Negative Binomial; ZINB; ZIP; Poisson Hurdle

Introduction

Dental caries is one of the most common infectious multifactorial diseases globally, which is characterised by the progressive demineralization of the tooth, following the action of bacterial acid metabolism [1]. It is in essence a life style disease affecting 60 to 90 percent of school going children and vast majority of adults as well as gender, all races, all socioeconomic status and all age groups [2-4]. The impact of dental caries includes oral pain which may affect speech, eating, sleeping, swallowing as well as breathing. The causes of altered appearance of dental caries can result into low self-esteem and undermine social acceptance [5]. The prevalence and incidence of dental caries in a population is influenced by number risk factors such as age, sex, ethnic group, dietary patterns and oral hygiene [6]. Dental caries is a disease with both high prevalence and severity in adult worldwide populations [7]. According to the research conducted by [8], dental carries affects over half of the population of industrialized countries and since it is a cumulative process, the number of affected individuals increases with ageing. Although caries was not taken to be a major problem in Africa, there is now evidence of a rise in prevalence of dental caries in developing countries [8,9].

Dental caries is a major cause of tooth loss in Ghana yet there is evidently very little knowledge as to what causes dental caries or methods of its prevention among the public. This is due to the limited resources and self-care methods important in the prevention

of dental carries. Dental caries therefore imposes a considerable burden not only on the individual adult but also the economy of Ghana. Base on this gab, the research seeks to determine and appropriate count regression model that efficiently fits DMFTDATA and also use the appropriate count regression model to identify the key risk factor contributing to dental caries in adults in Ghana.

Materials and Methods

In order to achieve the objectives of the study, a cross sectional design was employed using a well-structured questionnaire with interview and clinical examination. The study was conducted between June 2015 and June, 2016 at Sefwi Wiawso District Hospital in Ghana. A systematic random sampling technique was employed to the population at the hospital to select a sample 1158 individuals who are within the age group 18-65 years. The protocol of this research along with the informed consent forms were approved by the Sefwi Wiawso District Hospital in Ghana. The research protocol was previously explained to all patients and informed consents were signed for each individual prior to entering the research.

Data Collection

The data collection was in two parts; a quantitative survey and clinical examination. Individuals were clinically examined by a dentist and information on an individuals' oral health practices such as DMFT calculated (index), gender, age, duration of change of tooth brush, frequency of sweet consumption per day, frequency of brushing, use of fluoridated toothpaste, rinsing habit after every meal with water, dental visit experience (frequency of dental visit) and jaw accident were recorded. Dental caries statuses of the individuals involved in the research were recorded based on the criteria proposed by World Health Organization (WHO), by counting the number of teeth which were decayed due to caries, missing due to caries, and filled due to carries for the calculation of the DMFT index (the DMFT index is a standard method for measuring the dental caries incident in individuals) [10]. Dental caries was identified visually at the cavitation level; hence early caries was not recorded. The dental caries examination was carried out by a trained examiner using the DMFT index. Each tooth of each individual was examined using artificial light, plane mouth mirrors and probes. The questionnaires were collated and transferred by recording them into Microsoft excel 2013. The data was then again transferred into SPSS 22.0 and R statistical software for further statistical analysis. (2.1)

DMFT Index Calculation Guidelines: The teeth counted were supernumerary teeth, congenitally missing teeth, unerupted teeth, primary teeth reserved in permanent dentition and accident teeth removed due to causes other than dental carries. A tooth was recorded as D when both carious lesion(s) is seen. A tooth was also recorded as M when it has been removed due to carries. When a temporary or permanent tooth is filled due to carries, this is recorded as F. When there is no other area of the tooth with primary caries or there is 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 carries.

DMFT Index Calculation: For an individual the DMFT index is calculated by using the relation;

$$DMFT_{index} = D + M + F \qquad (1)$$

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

Data Analysis

The objective of this research is to determine an appropriate count regression model suitable for the analysis of DFMTDATA and to identify the risk factors contributing to dental caries in adults using the appropriate count regression model. Information on DMFT calculated, age (in years), gender (0-male, 1-female), dental visit (0-less than or equal to 3 months, 1-more than 3 months), frequency of sweet consumption per day (0-not frequent, 1-frequent), frequency of brushing, 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 a structured questionnaire and analysed using R and SPSS 22.0. The DMFT calculated was treated as the response variable and the rest as explanatory variables. The average age of participants was about 40 years with 54 percent being females and the rest being males (46 percent).

Regression Models

Count data such as the DMFTDATA are better modelled using Zero Inflated Negative Binomial, Poisson Hurdle, Zero Inflated Poison, Negative Binomial and Poisson regression models since it assumes non-negative values, discrete in nature and are often zero-inflated. These regression models employed to model the outcomes of dental caries in adults are briefly explained as follows;

Poisson Regression Model: The Poisson regression model is the most basic model for count data. If the variance of the counts approximately equals to the mean counts, then the Poisson regression model is expressed as;

$$P(Y_i = y / X_i) = \frac{\exp(-\mu_i)\mu_i^y}{y!} \quad for\ y = 0,1,2,3\ldots \qquad (2)$$

where Y_i represents the number of DMFT calculated for a specific period *i* and μ_i represents the expected number of DMFT calculated per given period which can be expressed as;

$$\mu_i = \exp\left(X_i^T \beta\right) \tag{3}$$

where β is the vector of unknown regression parameters to be estimated and X_i^T is the vector of explanatory variables. The equation (3) gives the indication that a unit increase in an explanatory variable increase the expected value μ_i by a multiplicative factor of exp(β).

The main constraint of the Poisson regression model is that, the mean and the variance are approximately equal that is;

$$E(Y_i = y / X_i) = Var(Y_i = y / X_i) = \mu \tag{4}$$

Due to this, in the presence of heterogeneity or over-dispersion (when the variance increases faster than what the Poisson regression allows), the Poisson regression does not work well hence there is the need to fit a parametric model that is more dispersed than the Poisson model and a natural choice is the Negative Binomial, Poisson Hurdle and the Zero-inflated regression models.

The log-likelihood function of the Poisson regression model is expressed as;

$$\ell = \sum_{i=1}^{n} \left(-\mu_i + y ln\mu_i - lny!\right) \tag{5}$$

By substituting equation (3) into equation (5), we further obtain the log-likelihood function as;

$$\ell = \sum_{i=1}^{n} \left(-e^{X\beta} + yX_i\beta - lny!\right) \tag{6}$$

In order to estimate the regression coefficients by the Maximum Likelihood Estimation (MLE) procedure, the derivative of the log-likelihood with respect to β_s must be set to zero as;

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=0}^{1} \left(y_i - -e^{X_i\beta}\right) X_i = 0 \tag{7}$$

Estimating the regression coefficient in the Poisson regression model is not obtained from a direct equation but rather the Newton Raphson method used for estimating the unknown parameters in the model.

Negative Binomial Regression Model: If a Poisson regression model doesn't fit the data and it appears that the variance of is increasing faster than the Poisson model allows, then a simple scale-factor adjustment is not appropriate. One way to handle this situation is to a fit a parametric regression model that is more dispersed than the Poisson. A natural choice is the negative binomial.

Suppose; $y/\lambda \sim poisson\ (\mu)$ and $\lambda \sim Gamma(\alpha,\beta)$, where $Gamma(\alpha,\beta)$ is the gamma distribution with mean $\alpha\beta$ and $\alpha\beta^2$ variance , whose density is given by;

$$P(\lambda) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \lambda^{\alpha-1} \exp\left(-\lambda/\beta\right) \tag{8}$$

for $\lambda>0$ and zero otherwise. Then it is easy to show that the unconditional distribution of *y* is negative binomial;

$$P(Y_i = y / X_i) = \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)y!}\left(\frac{\beta}{1+\beta}\right)^{y_i}\left(\frac{1}{1+\beta}\right)^{\alpha} \quad y = 0,1,2,.. \tag{9}$$

This distribution has the mean and variance as; $E(y)= \alpha\beta$ and $var(y) == \alpha\beta+ \alpha\beta^2$ respectively.

In order to build the regression model, it is natural to express the negative binomial in terms of the parameters $\mu = \alpha\beta$ and $\varpi=1/\alpha$ so that $E(y) = \mu$ and $var(y) = \mu + \varpi\mu^2$ where the variances function is quadratic. The distribution of then is formulated as;

$$P(Y_i = y / X_i) = \frac{\Gamma\left(\omega^{-1}+y\right)}{\Gamma\left(\omega^{-1}\right)y!}\left(\frac{\omega\mu}{1+\omega\mu}\right)^{y_i}\left(\frac{1}{1+\omega\mu}\right)^{\omega^{-1}} \tag{10}$$

which approaches Poisson (μ) as $\varpi \to 0$. The negative binomial can accommodate over-dispersion but not under-dispersion with respect to the Poisson model. For regression purposes, it is typically assumed; $y_i Negbin(\mu_i, \varpi)$ and apply a log-link, so that

$$log\,\mu_i = \eta = X_i^T \beta \tag{11}$$

The log-likelihood function of the negative binomial regression model is obtained from the following equation;

$$\ell = \sum_{i=1}^{n} \left\{ \left[\sum_{i=1}^{y-1} \ln\left(y_i + \alpha^{-1}\right) - lny_i ! + \alpha^{-1}\left[\ln\left(\alpha^{-1}\right) - \ln\left(\alpha^{-1} + e^{x\beta}\right)\right] + y_i\left[x_i\beta - ln\left(\alpha^{-1} + e^{x\beta}\right)\right]\right]\right\} \tag{12}$$

In order to estimate β and α as in the Poisson regression model, the iteration procedure or the method of Newton Raphson is applied (Lee and Mannering, 2002).

Poisson Hurdle Regression Model: Many count data exhibit more zero counts and are in addition over-dispersed. One type of count regression model that is capable of dealing with both excess zeros and over-dispersion is the Poisson hurdle regression model, which was proposed [11]. The Poisson Hurdle count regression model is a two state model; a binary component to predict zeros and a zero-truncated component such as the Poisson to predict the non-zero counts.

The probability density function of the Poisson hurdle model is given by;

$$P(Y_i / X_i, Z_i) = \frac{(1-\omega_i)\exp(-\mu_i)\mu_i^y}{(1-\exp(-\mu_i))y!} \; for\; y > 0 \tag{13}$$

where $\mu i = exp(X_i^1 \beta)$

The variance and mean of a Poisson hurdle model according to Chipeta *et al.* (2014) are given as;

$$Var(Y_i / X_i, Z_i) = \eta(\mu_i - \eta) + \frac{\pi\sigma^2}{1-P(0;\alpha)} \text{ and } E((Y_i / X_i, Z_i) = \eta - \frac{\pi\sigma^2}{1-P(0;\alpha)}$$

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

$$f_{hurdle}(y, X_i, Z_i, \beta, \gamma) = \begin{cases} f_{zero}(0, Z_i, \gamma) if\; y = 0 \\ (1 - f_{zero}(0, Z_i, \gamma).f_{count}(y, X_i, \beta) / f_{count}(0, X_i, \beta) if\; y > 0 \end{cases} \tag{14}$$

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

$$L = \prod_{i=\hat{U}_o} \left\{f_{zero}(0; Z_i, \gamma)\right\} \prod_{i=\hat{U}_1} \left\{\left(1 - f_{zero}(0; Z_i, \gamma)\right).f_{count}(y, X_i, \beta) / 1 - f_{count}(0, X_i, \beta)\right\} \tag{15}$$

where $\Omega_o = (i / y = 0), \Omega_1 = (i / y \neq 0)$ and $\Omega_o \cup \Omega_1 = \{1, 2, .., N\}$

The log-likelihood, by taking log of the likelihood function and rearranging the terms, gives the following relation;

$$\ell = \sum_{i=\hat{U}_o} ln\left\{f_{zero}(0; Z_i, \gamma)\right\} + \sum_{i=\hat{U}_1} ln\left\{1 - f_{zero}(0, Z_i, \gamma)\right\} + \sum_{i=\hat{U}_1} ln\left\{f_{count}(y, X_i, \beta)\right\} - ln\left\{1 - f_{count}(0, X_i, \beta)\right\} \tag{16}$$

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. Hence the mean regression relationship can be expressed;

$$\log(\mu) = X_i\ \beta + \log\left(f_{zero}(0; Z_i, \gamma)\right) - log\left(1 - f_{count}(0, X_i, \beta)\right) \tag{17}$$

by a canonical link.

Zero-Inflated Models: Zero-inflated Poisson and Zero-inflated Negative Binomial are zero-inflated models capable of addressing issues of excess zero counts and over-dispersion [11,12]. The Zero-inflated models as compared to the Hurdle models also 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 $\omega_i = P(i \in (structural\ zero)/Z_i)$ and $1-\omega_i = P(i \in (sampling\ zero)/Z_i)$, then the Zero-inflated Poisson has the distribution;

$$P(Y_i / X_i, Z_i) = \begin{cases} \omega_i + (1-\omega_i)\left(\dfrac{\theta}{\mu_i+\theta}\right)^{\theta} \ if\ y=0 \\ (1-\omega_i)\dfrac{\exp(-\mu_i)\mu_i^{y}}{y!} \ for\ y>0 \end{cases} \tag{18}$$

The variance and the mean of Y_i according to Xia *et al.* (2012) are given respectively as $Var(Y_i / X_i, Z_i) = \mu_i(1-\omega_i)(1+\mu_i\omega_i)$ and $E(Y_i / X_i, Z_i) = (1-\omega_i)\mu_i$

On the other hand, the Zero-inflated Negative Binomial has the form;

$$P(Y_i / X_i, Z_i) = \begin{cases} \omega_i + (1-\omega_i)\left(\dfrac{\theta}{\mu_i+\theta}\right)^{\theta} \ if\ y=0 \\ (1-\omega_i)\dfrac{\tilde{A}(y+\grave{e})}{y\tilde{A}(\theta)}\left(\dfrac{\theta}{\mu_i+\theta}\right)^{\theta}\left(\dfrac{\mu_i}{\mu_i+\theta}\right)^{y} \end{cases} \tag{19}$$

The variance and mean of the ZINB also are also given respectively by the relations; $Var(Y_i / X_i, Z_i) = \mu_i(1-\omega_i)(1+\mu_i(\omega_i+\theta))$ and $E(Y_i / X_i, Z_i) = (1-\omega_i)\mu_i$

Generally the zero-inflated density function is a combination of the 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;

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

$$f_{zero}(y, X_i, Z_i, \beta, \gamma) = f_{zero}(0; Z_i, \gamma) * I_{(0)}(y) + (1 - f_{zero}(0; Z_i, \gamma)) * f_{count}(y, X_i, \beta) \tag{19}$$

Where the unobserved probability belong to the point mass component and *I(.)* is the indicator function. The related mean regression equation is formulated as;

$$\mu_i = \pi_i.0 + (1-\pi_i)\exp\left(X_i^T\beta\right) \tag{20}$$

By a canonical link

Parameter Estimation: In estimating the parameters used in the models, the maximum likelihood estimation (MLE) has been considered. It is therefore very necessary to check the significance of the variables included in the models in order to evaluate the models involved in the study. The regression coefficients estimated have to be statistically significant for a better model.

Model Selection Methods

Selection of Zero-Inflated Models over Traditional Models: The score, likelihood ratio test, Wald test just to mention few are available for testing the zero-inflation in the model [13]. For easiness in selecting zero-inflated models over their traditional counterparts, the Vuong statistic will be considered. In defining the Vuong statistic, we assume $f_1(Y_i = y / X_i)$ and $f_2(Y_i = y / X_i)$ are both the probability density functions of Hurdle or Zero-inflated models and their traditional models (Poisson regression model and Negative binomial model) respectively whilst $f_1(Y_i = y / X_i)$ and $F_2(Y_i = y / X_i)$ as their corresponding cumulative distribution functions. Then the Vuong statistic (V) is therefore defined as;

$$V = \frac{\bar{m}}{S_m / \sqrt{n}} \tag{21}$$

where $\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ and $S_m = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2}$ represents the mean and the standard deviation of the measurement of *mi* . *mi* on the other hand, is defined as;

$$m_i = log\left(\frac{\hat{f}_1(Y_i = y / X_i)}{\hat{f}_2(Y_i = y / X_i)}\right) \quad (22)$$

where $\hat{f}_1(Y_i = y / X_i)$ and $\hat{f}_2(Y_i = y / X_i)$ indicate the predicted probabilities of the corresponding probability distribution functions $f_1(Y_i = y / X_i)$ and $f_2(Y_i = y / X_i)$ respectively.

Akaike's Information Criterion (AIC): Akaike's information criterion (AIC), is a measure of the relative quality of a statistical model for a given data [14]. That is, given a collection of models for a data, AIC estimates the quality of each model, relative to other models. Hence, AIC provides a means of model selection. For any statistical model, the AIC value is computed using the relation;

$$AIC = -2L + 2K \quad (23)$$

Where L is the maximized value of the likelihood function and k is the number of parameters in the model.

The model with the lowest AIC value among the models being compared is said to be the best fitted model. In other words, the better the model fit, the smaller the AIC. AIC is used when comparing non-nested models fitted by maximum likelihood estimation.

Results

Bar chart of the DMFT calculated from the Figure 1 below clearly reveals that, the DMFT calculated is characterised by many zero-valued observations and additionally positively skewed. Furthermore, the DMFT calculated (response variable) descriptively gave a variance of 5.074 which is much greater than the mean (2.25) indicating the presence of over-dispersion.

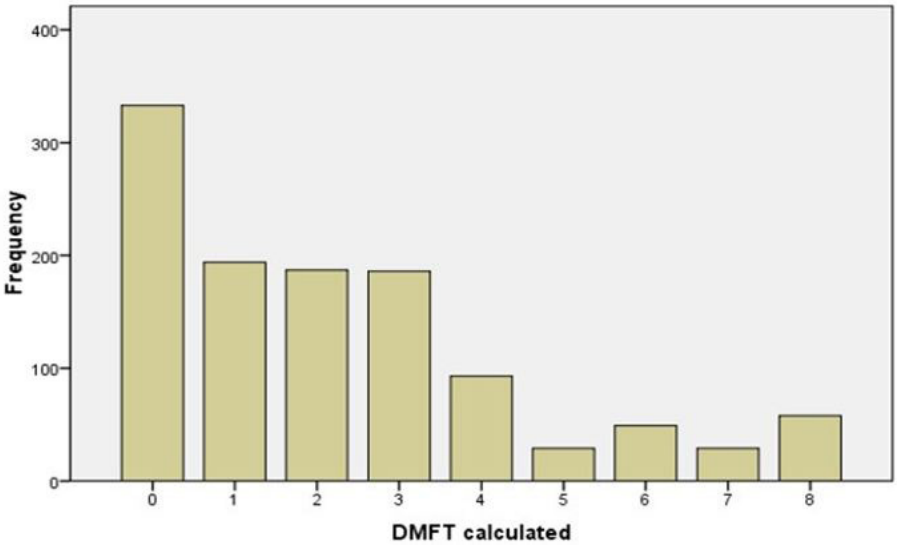

Figure 1: Bar chart indicating the distribution of DMFT calculated

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

Table 1: Distribution of DMFT calculated

Also the distribution of DMFT calculated in Table 1, showed that the observed number of zero counts in data is 28.8 percent indicating the presence of excess zeros. Additionally, from Table 1, it can be affirmed that the DMFT calculated contains non-negative integer values. Analysed results from Table 1 and Figure 1 clearly indicates that the best starting point in analysing DFMTDATA is to use the Poisson regression model since the observed counts are non-negative and also the Poisson regression model has some extensions that are useful for count data. The outcome variable (DFMT) occurred within a given period of time and did not assume normality hence the use of Poisson count regression model. The Poisson regression model was fitted to the DMFTDATA at a level of significance of 0.05. Below shows the parameter estimates, standard errors, standard normal (z) values and its respective p-values from the Poisson regression analysis (Table 2). All the factors from the Table 2 which includes age, gender, dental visit or treatment experience, frequency of brushing, use of fluoridated toothpaste, rinsing habit after every meal with water and jaw accident with the exception of the intercept are all found to be significant as contributing factors to dental carries since their respective p-values are less than the level of significance $\alpha = 0.05$.

Variable	Estimates	Standard error	z-value	pr(>|z|)
Intercept	-0.064507	0.117120	-0.551	0.5818
Age	0.012344	0.001925	6.414	$1.4\times10^{-10*}$
Gender	0.092011	0.039934	2.304	0.0212*
Dental Treatment	-0.591670	0.057785	-10.239	$2\times10^{-16*}$
Sweet consumption	0.351273	0.050834	6.910	$4.8\times10^{-12*}$
Rising habit	0.236086	0.040081	5.890	$3.86\times10^{-9*}$
Use of fluoridated toothpaste	-0.191712	0.047967	-3.997	$6.42\times10^{-5*}$
Jaw accident	0.113489	0.045757	2.480	0.0131*
Brushing times	-0.354106	0.063486	-5.578	$2.44\times10^{-8*}$
Toothbrush change	0.332932	0.077337	4.305	$1.67\times10^{-5*}$
*means significant at $\alpha = 0.05$				

Table 2: Parameter estimates of the Poisson Regression Model

Due to over-dispersion and excess zero counts in the data, as observed in Figure 1 and Table 1, count regression 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 over-dispersion 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 model which accounts for over-dispersion 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 for over-dispersion and(or) excess zero counts. These alternative models are to investigate the one state or two state processes of the DFMTDATA and thereafter apply comparison and selection tests to select the appropriate model for the DMFTDATA.

The Negative binomial regression model was estimated using DFMTDATA at a significant level of $\alpha = 0.05$ to account for over-dispersion due heterogeneity.

Variable	Estimates	Standard error	z-value	pr(>|z|)
Intercept	-0.185677	0.156132	-1.189	0.234350
Age	0.015274	0.002699	5.659	$1.53\times10^{-8*}$
Gender	0.094704	0.055240	1.714	0.086455
Dental Treatment	-0.597701	0.072616	-8.231	$2\times10^{-16*}$
Sweet consumption	0.376873	0.066807	5.641	$1.69\times10^{-8*}$
Rising habit	0.223354	0.055306	4.039	$5.38\times10^{-5*}$
Use of fluoridated toothpaste	-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*
ϖ	2.860			
*means significant at $\alpha = 0.05$				

Table 3: Parameter estimates of the Negative Binomial Regression Model

Table 3, therefore shows the parameter estimates, standard errors, standard normal values (z-values) and p-values form the negative binomial regression analysis. From the table, age, dental treatment, 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 found to contribute statistically significant to dental caries when their respective p-values were compared to the level of significance $\alpha = 0.05$. The dispersion parameter from the Table 3 was obtained as 2.860 which as a result greater than 1 indicating the presence of over-dispersion in the DMFTDATA, hence the Poisson regression model will be insufficient in modelling the DMFTDATA. The 28.8% of zero counts in the DMFTDATA (Table 1) implies the use of Zero-inflated models to account for the excess zeros.

Furthermore, since the zero-inflated Poisson regression model is capable of dealing with over-dispersion in which the Poisson model cannot handle especially dealing with over-dispersion due to excess zero counts, the zero-inflated Poisson model was applied to DFMDATA at a significant level of $\alpha = 0.05$. Table 4 as a result shows the parameters estimates, standard errors, z-values and p-values from the zero-inflated Poisson regression analysis. From the Table 4, the variables which include 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 contributing factors to dental caries with respective p-values being less than the level of significance 0.05. Also worth noting is that, when all the variables in this model were evaluated at zero, the model is significant at $\alpha = 0.05$.

Count model coefficients(pois with log link)				
Variable	**Parameter Estimate**	**Std. Error**	**z-value**	**Pr(> |z|)**
Intercept	0.709730	0.148216	4.788	$1.68\times10^{-6*}$
Age	0.010753	0.002116	5.083	$3.72\times10^{-7*}$
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	$5.88\times10^{-16*}$
Toothbrush change	-0.263445	0.096013	-2.744	0.00607*
Zero-inflated model coefficients (binomial with logit link)				
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	$2.33\times10^{-5*}$
Dental treatment	3.55248	0.37388	9.502	$2\times10^{-16*}$
Sweet consumption	-1.21280	0.27956	-4.338	$1.44\times10^{-5*}$
Rinsing habit	-1.58816	0.30198	-5.259	$5.21\times10^{-7*}$
Fluoridated toothpaste	2.28966	0.39204	5.840	$1.4\times10^{-9*}$
Jaw accident	0.36588	0.34355	1.065	0.2869
Brushing times	-3.25643	0.83048	-3.921	$8.81\times10^{-5*}$
Toothbrush change	-4.88095	0.84501	-5.776	$7.64\times10^{-9*}$
*means significant at $\alpha = 0.05$				

Table 4: Parameter estimates of the Zero-inflated Poisson (ZIP) regression model

The Zero-inflated negative binomial regression (ZINB) model also noted to be capable of dealing with extra zero counts especially when dealing with over-dispersion due to excess zero counts and heterogeneity, that is assumed to come from both the structural and chance sources, was estimated using DMFTDATA at significant level of 0. 05. Results from the Table 5 reveals age, frequency of sweet consumption per day, rinsing habit after every meal with water, jaw accident and brushing times as the significant contributing factors to dental carries in adults in Ghana. When all the variables in the ZINB regression model were evaluated at zero, the model was significant at $\alpha = 0.05$. Assessment of the dispersion parameter ϖ showed that there is over-dispersion due to excess zero counts since $\log(\varpi)$ from Table 5 is significant at $\alpha = 0.05$

Count model coefficients(negbin with log link)				
Variable	Parameter Estimate	Std. Error	z-value	Pr(> |z|)
Intercept	0.494766	0.202835	2.439	0.0147*
Age	0.012833	0.002744	4.676	2.92×10^{-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	7.61×10^{-13}*
Toothbrush change	-0.171539	0.122214	-1.404	0.1604
log(ϖ)	1.854763	0.178263	10.405	2×10^{-16}*
Zero-inflated model coefficients (binomial with logit link)				
Parameter	Estimate	Std. Error	z-value	Pr(> |z|)
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	6.17×10^{-5}*
Dental treatment	4.2478853	0.6366233	6.673	2.51×10^{-11}*
Sweet consumption	-1.096493	0.3423760	-3.203	0.00136*
Rinsing habit	-1.841739	0.3584712	-5.138	2.78×10^{-7}*
Fluoridated toothpaste	2.4995587	0.4213329	5.933	2.98×10^{-9}*
Jaw accident	0.4215990	0.4310619	0.978	0.32805
Brushing times	-4.656025	1.1365804	-4.097	4.19×10^{-5}*
Toothbrush change	-6.074075	1.0261714	-5.919	3.24×10^{-9}*
*means significant at $\alpha = 0.05$				

Table 5: Parameter estimates of the Zero-inflated Negative Binomial (ZINB) regression model

Finally, the Poisson hurdle regression model which has the capability to deal with extra zero counts in which the Poisson regression model cannot handle especially dealing with over-dispersion due to excess zero counts which was fitted at significant level of $\alpha = 0.05$. Fitting the data to the Poisson Hurdle regression model revealed the variables; age, gender, dental treatment, 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 as significant contributing factors to dental caries in adults (Table 6). Also worth noting, the estimated Poisson Hurdle model was statistically significant at $\alpha = 0.05$ when all the variables were evaluated at zero.

Count model coefficients(truncated poisson with log link)				
Variable	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}*
Toothbrush 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	2×10^{-16}*
Sweet consumption	1.632692	0.188961	8.640	2×10^{-16}*
Rinsing habit	1.080031	0.190177	5.679	1.35×10^{-8}*
Fluoridated toothpaste	-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 $\alpha = 0.05$				

Table 6: Parameter estimates of the Poisson Hurdle regression model

Model Evaluation and Comparison

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. Table 7 below shows the Vuong and AIC test statistics of the various count regression models employed in the study.

Characteristic	Poisson	NB	ZIP	ZINB	PHURDLE
AIC	4500.309	4307.981	4123.256	4066.447	4124.996
Dispersion parameter		2.860		6.3902*	
Vuong test			8.34683*	9.124266*	7.750094*

Table 7: Model Comparison and Evaluation

By comparing the Poisson regression model and the NB regression model to the ZIP and ZINB regression models respectively using the Vuong test statistic, the results from Table 7 above that is V=9.124266 for ZINB versus NB and V=8.34683 for ZIP versus Poisson, shows that both ZINB followed by the ZIP regression models offered a better fit to the DFMTDATA compared to their traditional counterpart regression models with one-component data. Results from Table 7, also shows evidence of over-dispersion in the DFMTDATA due to excess zero counts as confirmed by the values of the dispersion parameters for NB and ZINB as 2.860 and 6.3902 respectively. Also the results from Table 7 gave the AIC values for all the count regression models involved in the study of which the ZINB had the smallest AIC value indicating that, ZINB count regression model fits best to the DFMTDATA compared to the other count regression models. Thus, with the respect to the results from the Vuong test and the AIC we can best conclude that ZINB as compared to the other regression models employed in the study is the appropriate and efficient model for fitting DFMTDATA in order to determine key factors that contributes to dental caries in adults in Ghana. Based on the ZINB regression model the key factors that contributed significantly to dental carries in adults were age, frequency of sweet consumption per day, rinsing habit after every meal with water, jaw accident and brushing times.

Discussion of Findings

The response variable, DMFT calculated, in the DMFTDATA was characterized by excess zero counts of about 28.8% which is evident by the Vuong test that favoured two component models as against one component models. The value of the dispersion parameter (6.3902) from ZINB also indicated that there was over dispersion in the DMFTDATA due to excess zeros in the data. Modelling the DMFTDATA with the various regression models (Poisson, Negative Binomial, ZINB, ZIP and Poisson hurdle) showed an agreement with (15) who asserted that the ZIP and ZINB are known to provide robust statistics especially when zero counts are present and in addition to Poisson Hurdle models. Likewise, there was an agreement with (16) 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 (4066.447) placed the ZINB as the preferred model for fitting DMFTDATA efficiently. Though (17) 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. Largely, we had the same opinion on sweet consumption, rinsing habit and frequency of brushing. Age and jaw accident were also in addition significant risk factors contributing to the occurrence of dental caries in adults.

Conclusions and Recommendations

Collecting data on dental carries incidence is an essential field in oral epidemiology since dental carries is a severe issue in public health. In this research, an appropriate model suitable for fitting DFMTDATA was determined. The Poisson count regression model for count data modelling was a good starting point but has a restrictive equidispersion assumption. Due to this, the Negative binomial count regression model with more relaxed assumption on variance provided a better solution with evidence of over-dispersion in the DMFTDATA since the variance was greater than the mean. Advanced composite count regression models such as ZINB, ZIP and the Poisson Hurdle count regression models gave a more suitable fit to the data with over-dispersion as a result of high frequency of zero counts. Based on the AIC and Vuong test statistics, the ZINB was selected as the appropriate and significantly efficient model suitable for fitting DFMTDATA which is characterised with over-dispersion and many zero counts. It was finally found that age in years, jaw accidents, frequency of sweet consumption, rinsing habit after meal with water and frequency of brushing were the significant risk factors contributing to dental caries in adults in Ghana with 28.8% of excess zero counts in the DMFTDATA. With respect to these conclusions, the research therefore recommends adults to reduce the intake of sugar foods and consider consumption of more sugar-free foods since the intake of more sugar foods is a significant cause of dental caries. Also adults should endeavour to brush their teeth with fluoridated toothpaste at least twice daily. Additionally, mouth rinsing with water after meal should be encouraged in adults. Age should also be considered when planning dental care involving adults and finally adequate amount of dentists should be trained 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.

References

1. Kidd EA, Gierdrys-Leeper E, Simons D (2000) Take two dentists: a tale of root cares. Dent Update 27: 222-30.

2. Elderton RJ (1990) The definition and dental care. Oxford: Heinemann Medical Books, USA.

3. Poul Erik Petersen (2003) World Oral Health Report, Geneva, Switzerland.

4. Prakash H, Sidhu SS, Sundaram KR (1999) Prevalence of dental caries among Delhi school children. J Ind Dent Assoc 70: 12-4.

5. Weir E (2002) Dental caries: a nation divided. CMAJ 167: 1035.

6. Shingare P, Jogani V, Sevekar S, Patil S, Jha P (2012) Dental Caries Prevalence among 3 to 14 years old school children, Uran, Raigad District, Maharashtra. J Contemporary Dentistry 2: 11-4.

7. Urzua I, Mendoza C, Arteaga O, Rodríguez G, Cabello R, et al. (2012) Dental caries prevalence ad tooth loss in Chilean adult population: first National Dental Examination Survey. Int Dent J 2012: 810170.

8. Petersen P, Razanamihaja N, Poulsen VJ (2004) Surveillance of Oral Health among Children and Adults in Madagascar. WHO, Geneva, Switzerland.

9. Adegbembo AO, El-Naddeef MAI, Adeyinke A (1995) National survey of dental caries status and treatment needs of Nigeria. Int Dent J 45: 35-44.

10. World Health Organization (2000) Global Data on Dental Caries Prevalence (DMFT) in Children Aged 12 years. Global Oral Data Bank, Oral health country/area profile programme, Management of noncommunicable diseases. Geneva, Switzerland.

11. Mullahy J (1986) Specification and Testing of Some Modified Count Data Models. J Econometrics 33: 341-65.

12. Lambert D (1992) Zero-inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics 34: 1-14.

13. Lee AA, Xiang L, Fung WK (2004) Sensitivity of score tests for zero inflation in count data. Stat Med. 23: 2757-69.

14. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. Proc. 2nd Inter Symposium Information Theory 1973: 267-81.