International Journal of Data Science and Analysis
2018; 4(1): 24-31
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20180401.15
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Assessment and Selection of Competing Models for Count Data: An Application to Early Childhood Caries

Agnes Njambi Wanjau, Samuel Musili Mwalili, Oscar Ngesa

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:
agiewanjau@gmail.com (A. N. Wanjau)

To cite this article:
Agnes Njambi Wanjau, Samuel Musili Mwalili, Oscar Ngesa. Assessment and Selection of Competing Models for Count Data: An Application to Early Childhood Caries. *Advances in Materials*. Vol. 4, No. 1, 2018, pp. 24-31. doi: 10.11648/j.ijdsa.20180401.15

Received: February 19, 2018; **Accepted**: March 19, 2018; **Published**: March 23, 2018

Abstract: Count data has been witnessed in a wide range of disciplines in real life. Poisson, negative binomial (NB), zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) are some of the regression models proposed to model data with count response. All the count models are potential candidates that can model count data, but there is no means to choose the one that would perform better than the others. This study aimed to assess the count models mentioned earlier at various degrees of zero inflation. Datasets were simulated with ZIP distribution with different conditions of zero inflation (0%, 2%, 5%, 10%, 15%, 20%, 30% and 40%). Poisson and NB were observed to predict regression coefficients well when the proportion of zero is below 15%. The two ZIM performed well at higher degrees of zero inflation; beyond 15% for ZIP and 20% for ZINB. Exploratory examination of the caries data revealed a zero inflation below 15%, that is, 3.23%. Analysis of early childhood caries (ECC) data among 3-6 year old children who visited Lady Northey Dental Clinic was then performed with Poisson and NB. Akaike information criterion (AIC) test was used to compare all the competing models both under simulation and with real data. Poisson yielded lower AIC values at lower zero inflation rates as compared to other three models. ZIP had the lowest AIC value at 10%, 15%, 20%, 30% and 40% levels of zero inflation. NB model had the lowest AIC value when real data was analyzed. Education level of the father- primary school completed, chewing gum several times a week, Feeding habit jam several times a day, Feeding habit juice every day, Feeding habit soda every day and Feeding habit sweets several times a week were found to be significant factors causing ECC.

Keywords: Simulation, RMSE, Competing Models

1. Introduction

Count regression models have been employed overtime to model count data and have found a wide application in real world [6]. The index dmft denoting number of decayed (d), missing (m) or filled (f) teeth (t) due to dental caries is used to denote presence of cavity among children with primary dentition. It is a count occurrence. Event count can be defined as the number of times an event occurs, for instance, the number of teeth affected by cavity for each subject. It takes on a random variable that is nonnegative and discrete. Count regression models include Poisson, NB, ZIP and ZINB models among others [1]. An assumption of the Poisson distribution is that the mean and variance are equal. Violation of this assumption leads to models such as the NB that allow modeling of Poisson heterogeneity [4, 7, 9].

Zero inflated data is as well common in dental caries research. In such situations, some subjects portray absence of caries due to chance while others may never experience dental caries [5]. In ZIM, which are two-part in nature, zeros result from two population groups, one involving subjects who never portray a study characteristic and therefore generate the structural zeros while others yield sampling zeros with a probability during the study [17, 19]. ZIM allow us to model both presence and abundance simultaneously. Logistic regression models the first part while count regression models such as the Poisson and NB model the second part [10, 13].

Artificial data has been produced on several occasions to mirror important features of data expected in real world [2]. Simulations were therefore done to inform practitioners on what level of excess zeros warrant use of sophisticated

models such as ZIP and ZINB. One merit of simulation study is the ability to generate several datasets within seconds inorder to evaluate stability of decisions [11]. This may not be feasible from true respondents. Medical modeling and simulation has been known to assist in several areas of medical profession such as disease modeling, training and treatment [12, 18]. Dental caries, being a health problem has posed challenges to dental practitioners and administration while trying to model real data and forecast future trends of caries.

The dmft counts are in most cases characterized by inflated zeros, making modeling of such data more complex. This property violates the classical Poisson distribution assumptions, which is the simplest and most popular count regression model. Although several count models capable of addressing count data are available, the advantages of one over competing models has not been absolutely discussed in existing literature [8, 14-16]. Assessment of these competing models and their traits still calls for research. The main objective of this article is to assess the performance of potential competing count models under various zero inflation levels. It focuses on comparing Poisson, NB, ZIP and ZINB models in modeling count data. Choice of suitable model(s) for the analysis of real data at hand is discussed. Validity of one or more models with simulation guides its application to caries data in order to determine the main causative agents of caries among 3-6 year old children attending Lady Northey Dental Clinic.

One purpose of modeling count data is to enable prediction of effects changes have to a system. Inference will be made possible unlike when study is limited to exploratory analysis. Several studies have exhibited over dispersion and zero preponderance [17]. Real data considered for this study has a count response variable, dmft count, which requires choosing an appropriate count regression model to model it based on the degree of excess zeros as well as over-dispersion. Dental practitioners require more information about the best count regression model to employ for this and future case studies in order to plan for treatment.

Simulation modeling plays a key role in validating the models for prediction [14]. Comparison of model outputs under specified input conditions can only be achieved through simulation analysis. This counters the possibility of model's failure to meet specifications and eliminates over or under-utilization of regression models. Simulation tool will provide a better understanding of a system similar to the caries study at hand by developing mathematical models and observing how it operates under different inputs of zero inflation. In addition to simulation, goodness of fit statistic such as AIC is necessary. This is because it is not easy to determine the appropriateness of ZIP and ZINB as the zeros they account for cannot be observed directly but are latent.

Factors contributing to ECC should be recognized among infants in order to equip trainees and dental specialties with better skills for solving problems and making decisions. This study will be beneficial in developing new treatments and preventing progression of dental caries. Dental clinics can benefit from this study as information derived from it can be used to investigate the most optimal way to treat caries patients without compromising patient expectations.

2. Methods

2.1. Introduction

Significant developments in count models have taken place in demography, actuarial science and biostatistics. These models portray special features such as the features of generalized linear models. The main interest is to investigate the role of regressors which is achieved by regression modeling of count event. The response variable, dmft, is restricted to be a positive integer variable whose conditional mean is linked to a vector of regressors through the log link. In this chapter, both simulated and real data have been used for regression.

2.2. Simulation

Simulation has been defined as "the process of creating and experimenting with a computerized mathematical model of a physical system" [15]. Simulations enable researchers to check the performance of a statistical test on ideal data. Simulations were used to generate datasets with pre-specified properties and compared the parameter estimates resulting from regression to the specified parameters. A number of methods may exist for analyzing count data and suitability of such methods could be determined using simulations [16].

Two classical count regression models together with ZIP and ZINB have been discussed so far, as well as their estimation technique. Simulation of data was done under ZIP distribution. A count regression variable Y and two different types of covariates were simulated. The experiment was done 500 times on a sample of size n=1000 with two explanatory variables, age and sex, in the count component. The structural zero component assumed simple inflation, thus no regressors.

$$Y \sim age + sex|1 \tag{1}$$

Age is a continuous variable and was assumed to follow a normal distribution with mean=5 and standard deviation=0.7. The normal distribution is given as follows:

$age \sim N(5,0.7)$. Therefore:

$$f(age|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(age-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{0.98\pi}} e^{-\frac{(age-5)^2}{0.98}} \tag{2}$$

Age was simulated as age=rnorm(1000,5,0.7).

Sex is categorical binary and the function rbinom was used to generate random sample with n=1 and p=0.4.

$$sex = \begin{cases} 1\ if\ male \\ 0\ if\ female \end{cases} \tag{3}$$

$sex \sim Bernoulli\ (p)$. Therefore,

$$f(sex|p) = p^{sex}(1-p)^{1-sex} = 0.4^{sex}0.6^{(1-sex)} \tag{4}$$

The two regression coefficients were pre-specified as $\beta_1 = -0.1$ and $\beta_2 = 0.5$ a to give the regression function

$$log(\mu_i) = \beta_0 + \beta_1 age + \beta_2 sex = \beta_0 - 0.1age + 0.5sex \quad (5)$$

The underlying interest was to see the performance of the four count models discussed earlier for different proportions of zeros. Y was generated with a Poisson distribution with different zero percentages. These values include lower proportions of zeros to enable us assess the merit of the four models. It also helps us determine to what extent shall ZIM be employed. Proportions of zeros considered for this study are 0%, 2%, 5%, 10%, 15%, 20%, 30% and 40%. R software was used for analysis, where Poisson and NB models were fitted using glm() function in stats package and glm.nb() function in package MASS respectively.

To validate the simulation process, performance measures such as bias and MSE (mean square error) were used to make comparisons between simulation results and prespecified parameters as described by Beaujean [3].

2.3. Application to Real Data Set

2.3.1. Data

Collection point for data used in this study was Lady Northey Dental Clinic, situated in Westlands constituency, Nairobi City County along State house Avenue. The sampling frame consisted of patients between the age of three and six years whose parents or guardians accompanied them and agreed to be interviewed. This data was collected from September to November 2014. Only 83 observations with all values for every variable were used.

2.3.2. Study Variables

The outcome variable is the number of decayed (d), missing (m) or filled (f) teeth (t) due to dental caries. Predictor variables included age, gender, highest education of the father, highest education of the mother, employment state of the father, feeding habit biscuits, feeding habit gum, feeding habit jam, feeding habit juice, feeding habit soda, feeding habit sweets, feeding habit tea with sugar, brushing frequency and use of fluoridated toothpaste.

3. Results and Discussion

Regression coefficients estimates from Poisson, NB, ZIP and ZINB resulting from simulation were recorded alongside their respective bias and root mean square error (RMSE) as shown in table 1. AIC values from all the resulting models were also recorded in table 2. Table 3 presents estimates of proportions of zeros with corresponding bias and RMSE values. Poisson and NB regression coefficients estimates with real data are shown in table 4. In addition, table 4 shows the covariates statistically significant at the various significant levels. The AIC values of Poisson and NB as well as the levels of significance as given by the R software are outlined at the bottom of table 4. Significance levels included 0.001, 0.01, 0.05 and 0.1.

Table 1. *Regression coefficients estimates for Poisson, NB, ZIP and ZINB with synthetic data.*

	True value	Poisson				NB			
		parameter estimate	Std.Error	Bias	RMSE	parameter estimate	Std.Error	Bias	RMSE
0%									
b1	-0.1	-0.1092	0.05197	1.27E-07	0.050955	-0.12737	0.05058	1.29E-07	0.050956
b2	0.5	0.48081	0.07291	3.01E-07	0.07227	0.54057	0.0735	3.04E-07	0.072283
2%									
b1	-0.1	-0.08121	0.05056	1.37E-06	0.053844	-0.08125	0.05063	1.54E-06	0.053875
b2	0.5	0.52177	0.07241	6.71E-06	0.074852	0.52178	0.07252	6.88E-06	0.074856
5%									
b1	-0.1	-0.19425	0.04987	2.18E-06	0.055465	-0.19542	0.05084	2.28E-06	0.055452
b2	0.5	0.58154	0.07153	4.01E-07	0.073561	0.582	0.07287	4.12E-07	0.073548
10%									
b1	-0.1	-0.16763	0.05389	1.26E-06	0.057508	-0.12073	0.05652	1.29E-06	0.05745
b2	0.5	0.43875	0.07656	6.94E-06	0.077188	0.45289	0.07902	6.84E-06	0.077172
15%									
b1	-0.1	-0.104	0.05304	6.50E-06	0.061055	-0.1041	0.05388	7.01E-06	0.06112
b2	0.5	0.61098	0.07883	3.04E-06	0.081542	0.61096	0.07999	2.95E-06	0.081569
20%									
b1	-0.1	-0.1248	0.05813	6.59E-07	0.060166	-0.12575	0.06322	4.97E-07	0.060153
b2	0.5	0.54914	0.08195	8.67E-06	0.086828	0.54917	0.08871	8.48E-06	0.086829
30%									
b1	-0.1	-0.01454	0.06457	1.31E-06	0.068234	-0.01547	0.07308	1.76E-06	0.068284
b2	0.5	0.60847	0.08765	1.12E-05	0.092464	0.60852	0.09874	1.17E-05	0.092544
40%									
b1	-0.1	-0.14767	0.06474	1.12E-07	0.070264	-0.15298	0.07642	9.26E-08	0.069892
b2	0.5	0.7062	0.09689	4.87E-06	0.097421	0.70858	0.11372	4.46E-06	0.097206

Table 1. *Continued.*

	True value	ZIP				ZINB			
		parameter estimate	Std.Error	Bias	RMSE	parameter estimate	Std.Error	Bias	RMSE
0%									
b1	-0.1	-0.1786	0.05144	1.81E-07	0.051032	-0.07694	0.05201	4.05E-08	0.051129
b2	0.5	0.55362	0.07352	2.97E-07	0.074091	0.52503	0.07233	8.30E-08	0.072415
2%									
b1	-0.1	-0.08101	0.05086	1.28E-06	0.054013	-0.08111	0.05086	1.58E-06	0.05405
b2	0.5	0.52082	0.07291	6.86E-06	0.07494	0.52096	0.07291	7.10E-06	0.074965
5%									
b1	-0.1	-0.19497	0.05059	2.42E-06	0.055566	-0.19542	0.05096	2.58E-06	0.05562
b2	0.5	0.58249	0.07242	6.83E-07	0.073702	0.58201	0.07288	7.14E-07	0.07366
10%									
b1	-0.1	-0.11966	0.05609	9.55E-07	0.057265	-0.11974	0.05618	1.17E-06	0.057258
b2	0.5	0.45281	0.07852	5.28E-06	0.076842	0.45286	0.07863	6.19E-06	0.076967
15%									
b1	-0.1	-0.09631	0.05491	5.48E-06	0.061276	-0.09631	0.05492	5.66E-06	0.061321
b2	0.5	0.61223	0.08104	2.74E-06	0.081693	0.61223	0.08104	3.71E-06	0.081661
20%									
b1	-0.1	-0.12252	0.06229	1.22E-06	0.059607	-0.123	0.06258	5.28E-07	0.059841
b2	0.5	0.54616	0.0876	1.01E-05	0.087369	0.54668	0.0879	7.33E-06	0.087252
30%									
b1	-0.1	-0.009262	0.0719	3.07E-06	0.06761	-0.01033	0.0726	3.59E-06	0.067835
b2	0.5	0.616471	0.096291	8.98E-06	0.092805	0.61577	0.09705	1.33E-05	0.093045
40%									
b1	-0.1	-0.14234	0.0721	2.08E-08	0.070755	-0.14526	0.07362	1.38E-07	0.070498
b2	0.5	0.727	0.10918	9.18E-06	0.098772	0.72537	0.11056	1.32E-05	0.098451

Table 2. *AIC values for Poisson, NB, ZIP and ZINB models under simulation.*

percentage of zeros	Poisson	NB	ZIP	ZINB
0%	2289.9	2291.7	2249.125	2302.816
2%	2283.2	2285.2	2285.124	2287.125
5%	2310.8	2312.1	2312.387	2314.13
10%	2159	2244.8	2243.608	2245.6
15%	2108.2	2109.8	2107.996	2109.997
20%	2085.8	2076.2	2074.315	2076.275
30%	1959	1935.4	1932.152	1933.944
40%	1761.2	1721.1	1716.917	1718.6

Table 3. *Estimates and RMSE values of various zero percentages under simulation.*

	ZIP			ZINB		
percentage of zeros	Estimate	Bias	RMSE	Estimate	Bias	RMSE
0%	3.50E-05	0.000202	0.027794	5.03E-05	9.22E-05	0.022638
2%	0.091042	6.82E-05	0.032932	0.091037	1.62E-06	0.028789
5%	0.08863	2.10E-10	0.039739	0.060649	0.000192	0.041331
10%	0.089948	1.04E-05	0.044692	0.089921	0.000757	0.060344
15%	0.11424	9.66E-06	0.046319	0.114177	0.000966	0.070507
20%	0.168981	4.33E-06	0.044325	0.098086	0.001094	0.073924
30%	0.322839	6.19E-06	0.042015	0.233671	0.000923	0.073247
40%	0.409784	3.55E-05	0.037477	0.352896	0.001202	0.077425

Table 4. *Poisson and NB regression coefficients estimates with caries data.*

	Parameter	Poisson			
		Estimate	Std.Error	Z value	Pr(>|z|)
	Intercept	1.329011	0.80459	1.652	0.09858
	Age	-0.02129	0.065773	-0.324	0.74614
Gender	female	0.270115	0.126057	2.143	0.03213
Highest education level of father					
	Primary school completed	-0.99628	0.298362	-3.339	0.00084
	Secondary school completed	-0.60864	0.309075	-1.969	0.04893
	College/University	-0.58212	0.337086	-1.727	0.08418
	No male adult	-0.73175	0.465784	-1.571	0.11618
Highest education level of mother					
	Primary school completed	1.021829	0.499212	2.047	0.04067 *

	Parameter	Poisson			
		Estimate	Std.Error	Z value	Pr(>|z|)
	Secondary school completed	0.768398	0.484204	1.587	0.11253
	College/University	0.687741	0.469466	1.465	0.14294
Employment Father					
	Formal employment	0.137963	0.152093	0.907	0.36436
Feeding habit biscuits					
	Several times a month	0.257256	0.214833	1.197	0.23112
	once a week	-0.20042	0.188844	-1.061	0.28856
	Several times a week	-0.27657	0.194309	-1.423	0.15464
	Everyday	-0.21809	0.176594	-1.235	0.21684
	Several times a day	-0.07485	0.379255	-0.197	0.84354
Feeding habit gum					
	Several times a month	0.341327	0.216093	1.58	0.11421
	once a week	-0.12499	0.163831	-0.763	0.44552
	Several times a week	-0.62884	0.188325	-3.339	0.00084 ***
	Everyday	-0.54279	0.249505	-2.175	0.02959 *
	Several times a day	-0.02998	0.290174	-0.103	0.91771
Feeding habit jam					
	Several times a month	0.349425	0.21098	1.656	0.09768.
	once a week	0.342109	0.279251	1.225	0.22054
	Several times a week	0.027382	0.241389	0.113	0.90969
	Everyday	0.220301	0.1527	1.443	0.1491
	Several times a day	-0.64984	0.261054	-2.489	0.01280 *
Feeding habit juice					
	Several times a month	0.004304	0.183572	0.023	0.98129
	once a week	0.02906	0.181663	0.16	0.87291
	Several times a week	-0.35748	0.182567	-1.958	0.05022.
	Everyday	0.765603	0.237764	3.22	0.00128 **
	Several times a day	-0.30498	0.449318	-0.679	0.49729
Feeding habit soda	Several times a month	0.084501	0.191285	0.442	0.65867
	once a week	0.147669	0.188471	0.784	0.43333
	Several times a week	-0.25665	0.180157	-1.425	0.15428
	Everyday	0.952474	0.365081	2.609	0.00908 **
Feeding habit sweets					
	Several times a month	0.379049	0.235252	1.611	0.10713
	once a week	0.074773	0.157531	0.475	0.63503
	Several times a week	0.485374	0.233509	2.079	0.03765 *
	Everyday	0.741052	0.252853	2.931	0.00338 **
	Several times a day	0.094973	0.267982	0.354	0.72304
Feeding habit teasugar					
	Several times a week	0.768187	0.507404	1.514	0.13004
	Everyday	-0.09114	0.381086	-0.239	0.81098
	Several times a day	0.622179	0.345326	1.802	0.07159.
Brushing Frequency					
	Once a day	0.187313	0.258003	0.726	0.46783
	Two or more times a day	-0.03528	0.304523	-0.116	0.90776
Use of flouridated toothpaste					
	Non-flouridated toothpaste	0.668804	0.347475	1.925	0.05426.
Signif. codes:		0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			
AIC		571.34			
Theta					

Table 4. *Continued.*

	Parameter	NB			
		Estimate	Std.Error	Z value	Pr(>|z|)
	Intercept	1.21408	1.113357	1.09	0.2755
	Age	0.004997	0.087674	0.057	0.9546
Gender	female	0.245977	0.17158	1.434	0.1517
Highest education level of father					
	Primary school completed	-1.00146	0.437782	-2.288	0.0222 *
	Secondary school completed	-0.52352	0.444881	-1.177	0.2393
	College/University	-0.57001	0.48718	-1.17	0.242
	No male adult	-0.63161	0.647432	-0.976	0.3293
Highest education level of mother					
	Primary school completed	1.000256	0.701178	1.427	0.1537
	Secondary school completed	0.657924	0.6805	0.967	0.3336

	Parameter	NB			
		Estimate	Std.Error	Z value	Pr(>|z|)
	College/University	0.646837	0.660124	0.98	0.3271
Employment Father					
	Formal employment	0.21843	0.212826	1.026	0.3047
Feeding habit biscuits					
	Several times a month	0.293339	0.29686	0.988	0.3231
	once a week	-0.19385	0.257445	-0.753	0.4515
	Several times a week	-0.22438	0.260388	-0.862	0.3889
	Everyday	-0.21036	-0.21036	-0.851	0.395
	Several times a day	-0.05726	0.488018	-0.117	0.9066
Feeding habit gum					
	Several times a month	0.277405	0.301831	0.919	0.3581
	once a week	-0.14589	0.227399	-0.642	0.5212
	Several times a week	-0.64685	0.255583	-2.531	0.0114 *
	Everyday	-0.4717	0.334951	-1.408	0.1591
	Several times a day	0.048025	0.385663	0.125	0.9009
Feeding habit jam					
	Several times a month	0.269538	0.29568	0.912	0.362
	once a week	0.307992	0.381756	0.807	0.4198
	Several times a week	0.012144	0.323766	0.038	0.9701
	Everyday	0.189707	0.208245	0.911	0.3623
	Several times a day	-0.71163	0.341768	-2.082	0.0373 *
Feeding habit juice					
	Several times a month	0.03674	0.260745	0.141	0.8879
	once a week	0.030208	0.252628	0.12	0.9048
	Several times a week	-0.33918	0.245294	-1.383	0.1667
	Everyday	0.769093	0.331255	2.322	0.0202 *
	Several times a day	-0.22528	0.632436	-0.356	0.7217
Feeding habit soda	Several times a month	0.067732	0.261608	0.259	0.7957
	once a week	0.143818	0.267179	0.538	0.5904
	Several times a week	-0.28196	0.248864	-1.133	0.2572
	Everyday	1.000423	0.497039	2.013	0.0441 *
Feeding habit sweets					
	Several times a month	0.420812	0.331297	1.27	0.204
	once a week	0.101585	0.211888	0.479	0.6316
	Several times a week	0.571144	0.319391	1.788	0.0737.
	Everyday	0.677801	0.337135	2.01	0.0444 *
	Several times a day	0.023235	0.3655	0.064	0.9493
Feeding habit teasugar					
	Several times a week	0.721712	0.678945	1.063	0.2878
	Everyday	-0.05159	0.491724	-0.105	0.9164
	Several times a day	0.60409	0.447279	1.351	0.1768
Brushing Frequency					
	Once a day	0.198809	0.373428	0.532	0.5945
	Two or more times a day	-0.12754	0.434526	-0.294	0.7691
Use of flouridated toothpaste					
	Non-flouridated toothpaste	0.592004	0.485037	1.221	0.2223
Signif. codes:		0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			
AIC		557.28			
Theta		9.75			

This study aimed at evaluating four count regression models after making changes to data, by varying the proportions of zero. Results from table1 show that Poisson estimated the two regression coefficients under simulation experiment with 0%, 2% and 20% proportions of zeros relatively well and the coefficient of age at 15% and 40% zero proportions. This model underestimated β_1 and β_2 when p was fixed at 0.1 and 0.3. Poisson model was also observed to overestimate β_1 at 5% and 10% and β_2 at 5%, 15%, 30% and 40% percentages. NB model under-predicted the value of β_1 at 0.3 fraction of zero while over-predicting the same regression coefficient at 5% and 40% and β_2 at 5%, 15%, 30% and 40%. However, the NB model performed well in estimating β_1 and β_2 at 0%, 2%, 10% and 20% levels of zero proportion as well as β_1 at 15%.

The ZIP regression model estimated β_1 at 2%, 10%, 15%, 20% and 40% and β_2 at 10% and 20% proportions of zeros approximately well while over-predicting β_1 at 0% and 5% and β_2 at 0%, 5%, 15%, 30% and 40% percentages of zeros. Only β_1 was under-predicted by the ZIP model when the value of p was set at 0.3. ZINB approximated β_1 well when the proportion of zero was set at 0%, 2%, 10%, 15% and 20% and β_2 when p was specified as 0%, 2%, 10% and 20%. The regression coefficient of x1 was underestimated at 30% and overestimated at 5% zero proportions. This model overestimated β_2 at 5%, 15%, 30% and 40% zero

percentages. Poisson model yielded the lowest AIC value under 0%, 2%, 5% and 10% of zeros, as can be seen in table 2. ZIP proved to outperform all the four models at 15%, 20%, 30% and 40% percentages of zeros with the lowest AIC values followed by ZINB.

0, 0.1 and 0.4 fractions of zeros were approximately well estimated by the ZIP and ZINB models. These two models overestimated the value of p at 2% and 5% while underestimating p=15%. 20% and 30% proportions were observed to be predicted more accurately by ZIP in relation to ZINB. This can be observed from table 3. The root mean square error (RMSE) was observed to increase with increase in the value of p up to when p=15% for ZIP. On the other hand, RMSE increased with increase in p up to when p=20% for ZINB.

The proportion of zeros in the caries data was only 3.23\%. In other words, only four children did not display any sign of dmft. This number of zeros is below the threshold of 15\% recommended for application of ZIM from simulation results. The NB model had the lowest AIC value, hence fitted the caries data well as compared to Poisson. The following covariates' levels were observed to be significant at 5% level under NB regression: education level of the father- primary school completed, chewing gum several times a week, Feeding habit jam several times a day, Feeding habit juice every day, Feeding habit soda every day and Feeding habit sweets several times a week. These explanatory variables are marked with * against their p-values as indicated in table 4.

4. Conclusion

The ZIM can be employed when the proportion of zero p exceeds 15%, otherwise, the two classical count regression models apply. NB model fitted the caries data well.

The following covariates are the main risk factors associated with caries among children attending Lady Northey dental clinic: education level of the father- primary school completed, chewing gum several times a week, Feeding habit jam several times a day, Feeding habit juice every day, Feeding habit soda every day and Feeding habit sweets several times a week.

The simulation study dismissed the use of complicated ZIM, while favoring use of Poisson and NB models with real data. Classical count regression models should therefore not be overlooked as datasets have distinct properties. The ZIM should be employed with high percentages of zero in data, above 15% zero inflation. This is evidenced in the simulation study, where ZIP and ZINB models gave lower AIC values at 15% level of zero preponderance and beyond. NB should be used to fit the caries data at hand as it portrayed lower AIC value as compared to Poisson model. The main risk factors observed from regression with caries data should be considered in planning for prevention and treatment of caries among the children attending Lady Northey clinic. Further study should be done to determine the effect of each category of respective factors responsible for dental caries.

References

[1] Agresti, A. (2003). *Categorical data analysis* (Vol. 482). John Wiley & Sons.

[2] Benson, N. F. (2018). Introduction to a Special Issue on Simulation Studies as a Means of Informing Psychoeducational Testing and Assessment. *Journal of Psychoeducational Assessment*, *36*(1), 3-6.

[3] Beaujean, A. A. (2018). Simulating data for clinical research: A tutorial. *Journal of Psychoeducational Assessment*, 0734282917690302.

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

[5] Çolak, H., Dülgergil, Ç. T., Dalli, M., Hamidi, M. M., et al. (2013). Early child- hood caries update: A review of causes, diagnoses, and treatments. *Journal of Natural Science, Biology and Medicine*, *4* (1), 29.

[6] Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of personality assessment, 91* (2), 121–136.

[7] Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letters*, *99* (3), 585–590.

[8] Hallgren, K. A. (2013). Conducting simulation studies in the r programming environment. *Tutorials in quantitative methods for psychology, 9* (2), 43.

[9] Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

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

[11] Morgan, G. B., Moore, C. A., & Floyd, H. S. (2018). On using simulations to inform decision making during instrument development. *Journal of Psychoeducational Assessment*, *36*(1), 82-94.

[12] Morris, T. P., White, I. R., & Crowther, M. J. (2017). Using simulation studies to evaluate statistical methods. *arXiv preprint arXiv:1712.03198*.

[13] Mwalili, S. M., Lesaffre, E., & Declerck, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical Methods in Medical Research*, *17* (2), 123–139.

[14] Padhi, S. S., & Mohapatra, P. K. (2007). A discrete event simulation model for awarding of works contract in the government–a case study. In *5th international conference on e-governence-2007*.

[15] Sainani, K. L. (2015). What is computer simulation? *PM&R*, *7* (12), 1290–1293.

[16] Sokolowski, J. A., & Banks, C. M. (2011). *Principles of modeling and simulation: a multidisciplinary approach*. John Wiley & Sons.

[17] Wenger, S. J., & Freeman, M. C. (2008). Estimating species occurrence, abun- dance, and detection probability using zero-inflated distributions. *Ecology, 89* (10), 2953–2959.

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

[19] Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). Zero-truncated and zero-inflated models for count data. In *Mixed effects models and extensions in ecology with r* (pp. 261–293). Springer.