Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Research Article
Volume 9 Issue 5 - August 2019
DOI: 10.19080/BBOAJ.2019.09.555773

Biostat Biom Open Access J

Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros

Seyed Ehsan Saffari[1], John Carson Allen[1], Robiah Adnan[2], Seng Huat Ong[3], Shin Zhu Sim[3] and William Greene[4]*

[1]*Centre for Quantitative Medicine, Duke NUS Medical School, Singapore*

[2]*Department of Mathematical Sciences, Universiti Teknologi Malaysia, Malaysia*

[3]*Institute of Mathematical Sciences, University of Malaya, Malaysia*

[4]*Department of Economics, New York University, United States of America*

Submission: July 26, 2019; **Published:** August 26, 2019

***Corresponding author:** Professor William Greene, Stern School of Business, New York University, Department of Economics, 10012, New York, United States of America

Abstract

Count response variables are frequently encountered in medical data, which calls for the use of count regression models. In this study, we introduce the hurdle Conway-Maxwell Poisson (HCMP) regression model where the outcome variable is the number of doctor visits, complicated by excess zeros and over-dispersion from troublesome extreme values. A truncation approach is proposed to handle extreme values, leading to the definition of a truncated HCMP (THCMP) model. Parameter estimates are derived using maximum likelihood. Results of a case study on a RWM dataset investigated effects of response truncation at 6.65, 3.08 and 1.75% for the THCMP and truncated hurdle Poisson (THP) models. In a simulation study, responses were generated from a mixture of HCMP (50%) and HP (50%) probability models. THCMP and THP model performance was compared with respect to parameter estimation bias, goodness-of-fit and outcome estimates for truncation levels of 5 and 10%. As measured by AIC, the THCMP model exhibited better goodness-of-fit at all truncation levels compared to the THP model. Estimation bias increased with higher truncation levels for both models, but to a lesser degree for the THCMP model.

Keywords: Hurdle model; Conway-maxwell poisson; Over-dispersion, Parameter estimation; Model selection

Abbreviations: HCMP: Hurdle Conway-Maxwell Poisson; THP: Truncated Hurdle Poisson; CMP: Conway-Maxwell Poisson; GSOEP: German Socioeconomic Panel; LL: Log-Likelihood; AIC: Akaike's Information Criterion; BIC: Bayesian Information Criterion; TP: Truncated Poisson; TCMP: Right-Truncated CMP; HNB: Hurdle Negative Binomial; HGP: Hurdle Generalized Poisson

Introduction

Health care is one of the most important factors in human life. Good health care is a major contributor to quality of life, and ready access to a physician is an important component of a good health care system. The number of doctor visits for a household over a fixed interval is a useful metric for studying factors that affect physician accessibility. In this study, we introduce a new regression model for studying count data occurring in medical studies. The count variable in this study is number of doctor visits over a fixed time period.

There are numerous publications describing applications of count models to healthcare demand data, and applications of negative binomial models in particular. The negative binomial model has been applied to cross-sectional data, and in econometric models to analyze cross-sectional data with multiple outcomes per observation [1]. Modelling count data using a random effects negative binomial regression model is discussed in [2], and application of the negative binomial hurdle model to physician visit data is demonstrated in [3].

The Conway-Maxwell Poisson (CMP) distribution-a generalization of the Poisson-was introduced in [4] with applications to queues and service rates. The CMP distribution belongs to the exponential family and the two-parameter power series family of distributions. In the 50 plus years since its introduction, the CMP model has not been widely employed; however, a revival has arisen of late from a recognition of its utility in fitting discrete data [5]. The CMP distribution has two parameters and can handle both under- and over-dispersed data. This is in contrast to the commonly used negative binomial model which can only handle overdispersion.

We found several studies describing the properties of the CMP model with a variety of applications. The CMP distribution was applied to model timing of bid placement and the extent of multiple bidding in online auctions [6]. A Bayesian analysis of the CMP distribution is discussed in [7] and conjugate priors for the distribution parameters are derived. A flexible cure rate survival model is expanded to follow the CMP distribution in [8]. The joint generalized quasi-likelihood estimating equations are compared to the marginal equations in a CMP generalized linear model describing the number of car breakdowns in [9]. The structural properties of the CMP distribution, including moments and probability generating function are derived in [10]. Notwithstanding the many published applications, we have yet to find the CMP distribution used in a medical context-which is the motivation for this paper.

In many real world applications, the problem of excess zeros is encountered-the actual zero frequency outcome is higher than that predicted by the theoretical model. When this is the case, a hurdle model may be used that models the zero outcome separately from the non-zero outcomes [11]. In a hurdle model two densities are used, one that generates the zeroes, and another called the zero-truncated density that generates the positive values. The finite mixture distribution generated by combining two densities is discussed in [12]. The mechanisms by which excess zero frequencies occur for various types of count data, and how the zero-inflated Poisson model has application to such data in a medical context are described in [13]. Excess zeros in count data with application to public health employing a likelihood ratio test is addressed in [14].

Another aspect of our paper focuses on extreme values in the context of the CMP model and the adverse effect of troublesome 'outliers' on estimates of the CMP distribution mean and variance. The effect of outliers is to inflate the variance to make it larger than the mean, which is the definition of over-dispersion in the CMP model. One approach for reducing over-dispersion is right truncation, where values greater than a fixed constant are removed from the sample. A right-truncated Poisson regression model for handling over-dispersion is discussed in [15]. Applications of hurdle models with right censoring are the hurdle generalized Poisson regression model and the hurdle negative binomial regression model applied to the number of fish caught by fishermen at a state park, where the response was right censored [16,17].

The main focus of this study is on regression analysis based on CMP distribution. CMP has two parameters and this feature makes the distribution more flexible compared to Poisson model. Negative binomial (NB) model is a competitive model for CMP, however the dispersion parameter in NB model can only deal with over-dispersed data. CMP model is more flexible in that sense and can handle both over- and under-dispersion scenarios. The application part of this study (including read data example and simulation study) illustrates the performance of CMP model over alternative models under both under- and over-dispersed data.

The novel contribution of our paper is the introduction of a hurdle model based on the CMP distribution, the truncated hurdle Conway-Maxwell Poisson (THCMP) distribution, which can handle excess zeros in right-truncated count data. We illustrate the THCMP model in an application involving an analysis of the number of doctor visits over a fixed interval and compare it with the truncated hurdle Poisson (THP) model—a less suitable, but possible choice among presently available alternatives. In section 2, we describe the health care data set and variables for a case study analysis. Inasmuch as we have found no published study applying the CMP distribution to an outcome in medicine or public health, our paper is unique in this regard. In section 3, the THCMP regression model for right truncated data is introduced and parameter estimates derived. A case study analysis of the THCMP model is presented in section 4 along with a description of the methodology for a simulation study in which the THCMP model is compared to the truncated hurdle Poisson (THP) model on both over- and under-dispersed data. In section 5, we discuss results of a simulation study and evaluate the performance of the THCMP regression model versus the THP model.

Description of RWM health care data

The RWM data set [18] used in this study is taken from the German Socioeconomic Panel (GSOEP). The GSOEP, conducted by the German Institute for Economic Research in Berlin, surveys a representative sample of East and West German households. Researchers have recently used this cross-sectional data set to evaluate performance of their proposed count regression models [18,19]. The RWM data set is an unbalanced panel survey of health care utilization of 27,326 German individuals. The frequency table for the number of visits to the doctor is given in Table 2.

Table 1: Descriptive statistics for RWM data variables[1], n = 27,326.

Variable	Mean	Variance	Minimum	Median	Maximum
Doctor visits	3.18	32.37	0	1	121
Sex (=1 if female;0 otherwise)	0.48	0.25	0	0	1
Age (years)	43.53	128.38	25	43	64
Children (=1 when children are present;0 otherwise)	0.40	0.24	0	0	1
Education (years)	11.32	5.41	7	10.5	18
Married (=1 when head of household is married; 0 otherwise)	0.76	0.18	0	1	1

[1]Data source: German Institute for Economic Research

How to cite this article: Seyed Ehsan Saffari, John Carson Allen, Robiah Adnan, Seng Huat Ong, Sim SZ, William Greene. Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros. Biostat Biom Open Access J. 2019; 9(5): 555773.
DOI: 10.19080/BBOAJ.2019.09.555773

Table 2: Observed frequency counts of RWM doctor visits, n=27,326.

DocVis	Count	DocVis	Count	DocVis	Count	DocVis	Count
0	10,135	10	524	20	113	30	46
1	3,692	11	174	21	37	31	16
2	3,412	12	354	22	36	32	7
3	2,711	13	129	23	20	33	10
4	1,584	14	143	24	47	34	4
5	1,169	15	176	25	44	35	14
6	979	16	91	26	16	36	16
7	539	17	63	27	10	37	7
8	489	18	65	28	14	38	7
9	275	19	31	29	12	>38	115

The dependent variable in our analyses, DocVis, is a count variable—the number of visits to a doctor (including dentists) during a fixed time interval. The RWM data set is viewed as a cross-sectional dataset in our analysis [18,19]-the outcome is not time varying, as contrasted with [18]-and we assume that counts among study subjects are independent. The particular irregularities/anomalies of this dataset make it a good candidate for illustrating the unique features of the THCMP model. In Table 1, the mean and variance of DocVis are 3.18 and 32.37, respectively, which indicates substantial overdispersion in the data; minimum and maximum values are 0 and 121, respectively. In addition, the frequency of the zero response in DocVis is higher than expected (median=1, mode=0) (Table 2).

The explanatory variables consist of socioeconomic characteristics and demographic variables. All count regression models on DocVis were fitted as functions of sex (1=female, 0=male), age (years), education (years of schooling), marital status (1=married, 0=single) and children in the household (1=children present). Independent variables are summarized in Table 1 which shows that 48% of visits are by females, average age is 43.5, children are present in 40% of households, average years of schooling is 11.3, and 24.1% of respondents are single. The base case count model used in the analysis included the following variables in addition to the constant term:

$$X_i = \left(sex_i, age_i, children_i, education_i, married_i\right)$$

The frequency distribution of DocVis is shown in Table 2. According to the percentage of zeros in the response variable (37.1%), there is an excess of zeros. In addition, the 95th percentile is 12 which means that there are some extreme values in the sample. It is apparent from the histogram in Figure 1 that the zero count frequency of DocVis exceeds that expected in a Poisson distribution.

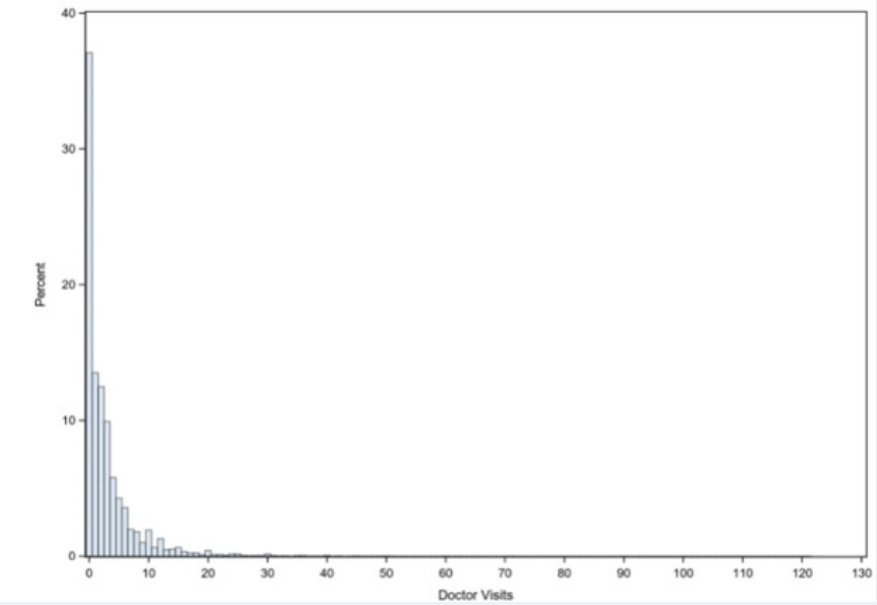

Figure 1: Histogram of number of doctor visits.

How to cite this article: Seyed Ehsan Saffari, John Carson Allen, Robiah Adnan, Seng Huat Ong, Sim SZ, William Greene. Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros. Biostat Biom Open Access J. 2019; 9(5): 555773.
DOI: 10.19080/BBOAJ.2019.09.555773

Methodology

In this section, the right-truncated hurdle Conway-Maxwell Poisson (THCMP) regression model is introduced for handling count data with excess zeros and right-tail data truncation. Parameter estimation and the goodness-of-fit statistics are discussed.

1.1. The model

Let the response variable $Y_i^*, i=1,\ldots,n$ be the number of visits to a doctor over a fixed time period. The HCMP regression model $f\left(\lambda_i, v, y_i^*\right)$, is given by

$$P\left(Y_i^*=y_i^*\left|x_i,z_i\right.\right)=\begin{cases} w_0 & , \quad y_i^*=0 \\ \left(1-w_0\right)\dfrac{\lambda_i^{y_i^*}}{\left(Z\left(\lambda_i,v\right)-1\right)\left(y_i^*!\right)^v} & , \quad y_i^*=1,2,\ldots \end{cases} \quad (1)$$

where

$$Z\left(\lambda_i,v\right)=\sum_{s=0}^{\infty}\frac{y_i^s}{\left(s!\right)v} \quad (2)$$

Where $\lambda_i=E\left(Y_i^*\right)$ of a Poisson distribution associated with observation , and $v\geq 0$ is the dispersion parameter. The CMP regression model can handle both over-dispersion $(v<1)$ and under-dispersion $(v>1)$, and when $v=1$, the probability function (1) reduces to a Poisson model. A geometric distribution is obtained from (1) when $v=0$ and $\lambda_i<1$. When $v\to\infty$ in (1) with probability $\frac{\lambda_i}{1+\lambda_i}$, the result is a Bernoulli distribution.

In many practical applications, it is common to assume that the parameter λ_i depends on a vector of explanatory variables $x_i=\left(x_{i1},x_{i2},\ldots,x_{im}\right)$. When there is interest in capturing possible systematic variation in λ_i as a function of x_i, explanatory variables are commonly incorporated in the context of a log-linear model, where log indicates the base e or natural logarithm,

$$\log\lambda_i=\sum_{j=1}^{m}x_{ij}\beta_j$$

The β_j's are coefficients of the explanatory variables in the regression model and m is the number of explanatory variables.

Furthermore, $0<w_0<1$ and $w_0=w_0\left(z_i\right)$ satisfy

$$\operatorname{logit}\left(w_0\right)=\log\left(\frac{w_0}{1-w_0}\right)=\sum_{j=1}^{m}z_{ij}\delta_j$$

where $z_i=\left(z_{i1}=1,z_{i2},\ldots,z_{im}\right)$ is the i-th row of a covariate matrix z, and $\delta=\left(\delta_1,\delta_2,\ldots,\delta_m\right)$ is an m-dimensional column vector of unknown parameters. In this model set up, the non-negative function w_0 is modeled using a logit link function.

The moments of the HCMP distribution are obtained as

$$E\left[Y_i^{r+1}\right]=\left(1-w_0\right)\frac{Z\left(\lambda_i,v\right)}{Z\left(\lambda_i,v\right)-1}E\left[Y_i^{*r+1}\right]$$

Where $E\left[Y_i^{*r+1}\right]$ is derived using [5], as

$$E\left[Y_i^{*r+1}\right]=\begin{cases} \lambda_i E\left(Y_i^*+1\right) & r=0 \\ \lambda_i\dfrac{d}{d\lambda_i}E\left[Y_i^{*r}\right]+E\left[Y_i^*\right]E\left[Y_i^{*r}\right] & r>0 \end{cases} \quad (3)$$

Now, we can define the right truncated hurdle Conway-Maxwell Poisson regression model as

$$P\left(Y=y_i\left|x_i,z_i\right.\right)=\begin{cases} \dfrac{w_0}{1-B}, & y_i=0 \\ \dfrac{1}{1-B}\dfrac{\left(1-w_0\right)\lambda_i^{yi}}{\left(Z\left(\lambda_i,v\right)-1\right)\left(y_i!\right)^v}, & 1\leq y_i\leq t \end{cases}$$

Where t is the truncation point for y_i. This means that we truncate the response variable when $y_i>t$, leading to the definition of B as

$$B=B\left(\lambda_i,v,w_0\right)=\frac{1-w_0}{Z\left(\lambda_i,v\right)-1}\sum_{y_i^*=t+1}^{\infty}\frac{\lambda_i^{y_i^*}}{\left(y_i^*!\right)^v}$$

Thus, the log-likelihood function for the HCMP model with right truncation can be written as

$$LL=\sum_{i=2}^{k}\left\{I_{\{y_i=0\}}\left[\log w_0-\log\left(1-B\right)\right]+I_{\{1\leq y_i\leq t\}}\left[\log\left(1-w_0\right)+y_i\log\lambda_i-\log\left(Z\left(\lambda_i,v\right)-1\right)-v\log\left(y_i!\right)-\log\left(1-B\right)\right]\right\} \quad (4)$$

Where k is the number of observations after truncation.

Parameter estimation

In this section we obtain parameters estimates using maximum likelihood. The likelihood equations for estimating β_r,δ_t and v are obtained by taking the partial derivatives of (4) and setting them equal to zero yielding

$$\frac{\partial LL}{\partial\beta_r}=\sum_{i=1}^{k}\left[\frac{\lambda_i}{1-B}\frac{\partial B}{\partial\lambda_i}+I_{\{1\leq y_i\leq t\}}\left(y_i-\frac{\lambda_i}{Z\left(\lambda_i,v\right)-1}\cdot\frac{\partial Z\left(\lambda_i,v\right)}{\partial\lambda_i}\right)\right]x_{ir}=0$$

$$\frac{\partial LL}{\partial\delta_t}=\sum_{i=1}^{k}\left[I_{\{y_i=0\}}-w_0+\frac{w_0\left(1-w_0\right)}{1-B}\frac{\partial B}{\partial w_0}\right]z_{it}=0$$

$$\frac{\partial LL}{\partial\delta_t}=\sum_{i=1}^{k}\left[\frac{1}{1-B}\frac{\partial B}{\partial v}-I_{\{1\leq y_i\leq t\}}\left(\frac{1}{Z\left(\lambda_i,v\right)-1}\cdot\frac{\partial Z\left(\lambda_i,v\right)}{\partial v}+\log\left(y_i!\right)\right)\right]=0$$

Where $\frac{\partial Z\left(\lambda_i,v\right)}{\partial\lambda_i}=\frac{1}{\lambda_i}\sum_{s=0}^{\infty}\frac{\lambda_i^s}{\left(s!\right)v}s$

$$\frac{\partial Z\left(\lambda_i,v\right)}{\partial\lambda_i}=-\sum_{s=0}^{\infty}\frac{\lambda_i^s}{\left(s!\right)v}\log\left(s!\right)$$

How to cite this article: Seyed Ehsan Saffari, John Carson Allen, Robiah Adnan, Seng Huat Ong, Sim SZ, William Greene. Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros. Biostat Biom Open Access J. 2019; 9(5): 555773.
DOI: 10.19080/BBOAJ.2019.09.555773

$$\frac{\partial B}{\partial \lambda_i}=\sum_{y_i^*=t+1}^{\infty}\frac{1-w_0}{\left(y_i^*!\right)v}\frac{y_i^*\lambda_i^{y_i^*-1}}{Z\left(\lambda_i,v\right)-1}-\frac{B}{Z\left(\lambda_i,v\right)-1}\frac{\partial Z\left(\lambda_i,v\right)}{\partial \lambda_i}$$

$$\frac{\partial B}{\partial w_0}=\frac{B}{1-w_0}$$

$$\frac{\partial B}{\partial v}=-B^*-\frac{B}{Z\left(\lambda_i,v\right)-1}\frac{\partial Z\left(\lambda_i,v\right)}{\partial v}$$

and where

$$B^*=B^*\left(\lambda_i,v,w_0\right)=\frac{1-w_0}{Z\left(\lambda_i,v\right)-1}\sum_{y_i^*=t+1}^{\infty}\frac{\lambda_i^{y_i^*}}{\left(y_i^*!\right)v}\log y_i^*$$

These partial derivative equations cannot be further simplified. Calculating the Hessian matrix directly is computationally laborious, and so the Conjugate Gradient Optimization method implemented in SAS was used to numerically obtain the Hessian variance-covariance matrix. In approximating standard errors of the parameter estimates, the Hessian matrix must be computed at least once, regardless of the optimization technique.

The Fisher information matrix for the THCMP regression model is obtained as

The elements of the Fisher information matrix are available in the Appendix.

$$I\left(\beta,v,\delta\right)=\begin{bmatrix}\frac{\partial^2 LL}{\partial\beta_r\partial\beta_u} & \frac{\partial^2 LL}{\partial\beta_r\partial v} & \frac{\partial^2 LL}{\partial\beta_r\partial\delta_t}\\ \frac{\partial^2 LL}{\partial v\partial\beta_u} & \frac{\partial^2 LL}{\partial v^2} & \frac{\partial^2 LL}{\partial v\partial\delta_t}\\ \frac{\partial^2 LL}{\partial\delta_t\partial\beta_r} & \frac{\partial^2 LL}{\partial\delta_t\partial v} & \frac{\partial^2 LL}{\partial\delta_t\partial\delta_v}\end{bmatrix}$$

Model selection and test for dispersion

Goodness-of-fit statistics for the THCMP model are based on the deviance statistic, defined as

$$D=-2\left[\log L\left(\hat{\theta};\hat{\mu}\right)-\log L\left(\hat{\theta};y\right)\right]$$

Where $\log L\left(\hat{\theta};\hat{\mu}\right)$ and $\log L\left(\hat{\theta};y\right)$ are obtained from the model likelihood function evaluated at μ and y, respectively. The log-likelihood function is defined in equation (4).

The deviance statistic can be approximated by a chi-square distribution when $\mu's$ is large. In the application section, we use $-2LL, AIC$ and BIC to compare the different regression models in terms of goodness-of-fit. For all of these statistics, a smaller value indicates a better fit.

From section 3.1, it is obvious that the THCMP model reduces to THP model when $v=1.$ To assess the adequacy of the THCMP model over the truncated hurdle Poisson model, we test the hypothesis

$$H_0: v=1 \quad against \quad H_a: v\neq 1 \quad (5)$$

The purpose of (5) is to evaluate the significance of the dispersion parameter. It follows that the THCMP model should be used instead of the THP model whenever H_0 is rejected. To test the null hypothesis H_0 in (5), the likelihood ratio statistic can be used. An alternative statistic for the parameter v is the asymptotic Wald statistic where the dispersion parameter is calculated after fitting the THCMP regression model.

Results

Case study

In this case study using the RWM data set (n=27,326), the THCMP regression model is used to model the number of doctor visits (DocVis) per patient over a period of three months as a function of the independent variables sex, age, children, education and married. The THP model is also considered as an alternative model in the analysis of the RWM. The THCMP model will be compared to the THP model relative to parameter estimates, standard errors, goodness-of-fit statistics and accuracy in modelling the response.

Five independent variables are used in the model and all are incorporated into both the logit and non-logit parts of the model. Therefore, the link functions can be written as

$$\log\lambda=\beta_0+\beta_1 sex+\beta_2 age+\beta_3 children+\beta_4 education+\beta_5 married$$

$$\text{logit}\, w_0=a_0+a_1 sex+a_3 age+a_3 children+a_4 education+a_5 married$$

Thus, the parameters $\left(a_0,a_1,a_3,a_3,a_4,a_5,\beta_0+\beta_1+\beta_2+\beta_3+\beta_4+\beta_5\right)$ plus the dispersion parameter (when used) will be estimated using the ML method.

Three truncation points, t=10, t=15, and t=20 are employed in comparing the effects of truncation, and correspond to truncation percentages of 6.65, 3.08 and 1.75, respectively.

Parameter estimates for the THP and the THCMP model were obtained for the specified truncation points and are summarized in Table 3. Link functions are obtainable from the estimates shown in Table 3. For example, the respective log and logit link functions for the THCMP regression model for truncation point $t_1=10$ is

$$\log\lambda=-0.1863+0.0471 sex+0.0036 age-0.0307 children-0.0081 education-0.0297 married$$

$$\text{logit}\, w_0=0.0766-0.6756 sex-0.0217 age+0.3609 children+0.0655 education-0.0211 married$$

Using the log link function coefficients in Table 3, one can see the estimated change in DocVis per unit change in each independent variable, all others being held constant. To illustrate, the positive β_1 coefficients corresponding the variable sex in the HP model in Table 3 for the three truncation times are 0.0997, 0.11 and

How to cite this article: Seyed Ehsan Saffari, John Carson Allen, Robiah Adnan, Seng Huat Ong, Sim SZ, William Greene. Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros. Biostat Biom Open Access J. 2019; 9(5): 555773.
DOI: 10.19080/BBOAJ.2019.09.555773

0.0935, which indicate higher values of $\log(DocVis)$ for females compared to males for all truncation points. For the THCMP model, decreases in $\log(DocVis)$ of 0.0158 and 0.003 for females versus males are predicted for $t_2 = 15$ and $t_3 = 20,$ respectively. For a one-unit increase in age, expected $\log(DocVis)$ is estimated to increase by ~0.01 on average using the THP model and ~0.003 using THCMP model for all truncation points. Thus, older patients are predicted to have a higher number of doctor visits per unit time. $\log(DocVis)$ is estimated to be lower for households with children relative to those with no children for both THP and THCMP models at all truncation points—with the exception of THCMP when $t_2 = 15.$ For a one-unit increase in the number of years of schooling, the estimated change in the number of doctor visits decreased for all truncation points. $\log(DocVis)$ showed more frequent doctor visits for married versus single individuals for all truncation points.

Table 3: Parameter Estimates of THP and THCMP models for three truncation scenarios, RWM data

Parameter	Truncated at t=10		Truncated at $t = 15$		Truncated at $t = 20$	
	HP[1]	HCMP[2]	HP	HCMP	HP	HCMP
Log link parameters						
β_0	0.955	-0.1863	1.0489	-0.1328	1.2042	-0.126
	(-0.0378)	(-0.0284)	(-0.0334)	(-0.02)	(-0.0318)	(-0.0168)
β_1	0.0997	0.0471	0.11	-0.0158	0.0935	-0.002
	(-0.0097)	(-0.0077)	(-0.0085)	(-0.0047)	(-0.008)	(-0.0047)
β_2	0.0072	0.0036	0.0091	0.003	0.0096	0.0026
	(-0.0005)	(-0.0004)	(-0.0004)	(-0.0003)	(-0.0004)	(-0.0002)
β_3	-0.0418	-0.0307	-0.0359	0.0074	-0.035	-0.0119
	(-0.0115)	(-0.0086)	(-0.0101)	(-0.0051)	(-0.0095)	(-0.0046)
β_4	-0.0091	-0.0081	-0.0116	-0.0126	-0.0186	-0.0103
	(-0.0022)	(-0.0018)	(-0.002)	(-0.001)	(-0.0019)	(-0.0009)
β_5	-0.0437	-0.0297	-0.0387	-0.0112	-0.0504	-0.0209
	(-0.012)	(-0.0087)	(-0.0104)	(-0.0053)	(-0.0098)	(-0.0051)
ν		0.108		0.069		0.0621
		(-0.0078)		(-0.0023)		(-0.0019)
Logit link parameters						
α_0	0.1506	0.0766	-0.0186	-0.0757	-0.0172	0.0365
	(-0.0954)*	(-0.1304)	(-0.0944)	(-0.1896)	(-0.0941)	(-0.1821)
α_1	-0.5548	-0.6756	-0.566	-0.3004	-0.5755	-0.4729
	(-0.0267)	(-0.0369)	(-0.0264)	(-0.0466)	(-0.0263)	(-0.0469)
α_2	-0.0155	-0.0217	-0.0163	-0.0345	-0.0165	-0.034
	(-0.0013)	(-0.0021)	(-0.0013)	(-0.0024)	(-0.0013)	(-0.0025)
α_3	0.247	0.3609	0.2391	0.0698	0.2452	0.3268
	(-0.0304)	(-0.0443)	(-0.0301)	(-0.0472)	(-0.0299)	(-0.0491)
α_4	0.0287	0.0655	0.0397	0.1606	0.039	0.1463
	(-0.0056)	(-0.0102)	(-0.0056)	(-0.011)	(-0.0056)	(-0.0107)
α_5	-0.1123	-0.0211	-0.0781	-0.0294	-0.0818	0.0726
	(-0.0338)	(-0.0435)	(-0.0335)	(-0.0497)	(-0.0334)	(-0.0503)

*Standard errors in parentheses.
[1]Hurdle Poisson
[2]Hurdle Conway-Maxwell Poisson

The THCMP regression model indicated overdispersion in the RWM data. Dispersion parameter estimates corresponding to truncation percentages of 6.65,3.08 and 1.75 were $\nu = 0.108,$ $\nu = 0.069$ and $\nu = 0.0621$ respectively, where $\nu < 1$ indicates overdispersion.

THP and THCMP regression model coefficients for the RWM analysis based on the logit link function are also given in Table 3. These coefficients correspond to the excess zeros component of the models. Both models show a negative effect of sex for all truncation percentages, indicating a higher rate of zero doctor visits in males than in females. The log odds of excess zeros increases

How to cite this article: Seyed Ehsan Saffari, John Carson Allen, Robiah Adnan, Seng Huat Ong, Sim SZ, William Greene. Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros. Biostat Biom Open Access J. 2019; 9(5): 555773.
DOI: 10.19080/BBOAJ.2019.09.555773

for each unit decrease in age in both the THP and THCMP models at all truncation points. This means that zero doctor visits were increasingly more likely with aging. A positive coefficient for children in both THP and THCMP models indicated that households with children exhibited a higher rate of zero visits to a doctor compared to households without children for all truncation percentages. The log odds of excess zeros increased for each unit increase in the number of years of schooling for all truncation points for both regression models. This indicates that fewer years of schooling were associated with higher odds of zero visits to a doctor. Both THP and THCMP models resulted in negative coefficients for the married variable for all truncation points with the exception of the $t_3 = 20$ truncation point for the THCMP model. So, generally, single individuals exhibited higher rates of zero doctor visits than married individuals.

Table 4 compares the THP and THCMP regression models on three goodness-of-fit measures: log-likelihood (LL), Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The right-truncated Poisson (TP) and right-truncated CMP (TCMP) regression models are included in Table 4 as special cases to investigate whether a zero frequency of 37.1% should be considered an excess zero scenario. We were also able to investigate whether the TP and TCMP regression models (without excess zero scenario) were able to fit the data as well as the hurdle regression models. The THP and THCMP models demonstrated better goodness-of-fit than the TP and TCMP models on all measures (smaller is better). The THCMP regression model exhibited superior goodness-of-fit compared to the THP regression model for all truncation levels in the RWM data analysis.

Table 4: Goodness-of-fit measures for THP and THCMP regression models, RWM data.

Goodness of fit*	Truncated at $t = 10$				Truncated at $t = 15$				Truncated at $t = 20$			
	HP[1]	HCMP[2]	P[3]	CMP[4]	HP	HCMP	P	CMP	HP	HCMP	P	CMP
-2LL	99,625	95,034	117,595	96,259	118,293	106,771	143,350	109,348	129,498	111,829	158,667	114,751
AIC	99,649	95,060	117,607	96,273	118,317	106,797	143,362	109,362	129,522	111,855	158,679	114,765
BIC	99,747	95,166	117,656	96,330	118,416	106,903	143,411	109,419	129,621	111,961	158,729	114,822

*LL=Log-Likelihood, AIC=Akaike Information Criterion, BIC= Bayesian Information Criterion
[1]Hurdle Poisson
[2]Hurdle Conway-Maxwell Poisson
[3]Poisson
[4]Conway-Maxwell Poisson

Table 5 shows that the THCMP regression model exhibited a better fit to the RWM data versus the THP model based on the predicted versus observed frequency count criterion for all truncation points. The RWM dataset observed zero frequency was 10135. Predicted zero frequencies for THCMP/THP at the truncation points were $t_1 : 10118/10012$; $t_2 : 10319/9971$; and $t_3 : 9626/9008$.

Table 5: Observed versus predicted doctor visit counts for THP and THCMP models at three truncation points.

Doctor visits	Observed count	Truncated at $t = 10$		Truncated at $t = 15$		Truncated at $t = 20$	
		HP[1]	HCMP[2]	HP	HCMP	HP	HCMP
0	10,135	10,011.87	10,118.24	9,970.72	10,319.23	9,008.43	9,626.32
1	3,692	2,095.52	3,896.52	1,455.30	3,668.79	1,230.78	3,689.69
2	3,412	3,361.66	3,157.25	2,752.27	2,992.24	2,524.27	3,043.88
3	2,711	3,595.20	2,448.63	3,470.07	2,373.15	3,451.44	2,448.63
4	1,584	2,883.73	1,840.95	3,281.30	1,845.17	3,539.36	1,934.88
5	1,169	1,850.44	1,351.13	2,482.24	1,412.75	2,903.62	1,507.87
6	979	989.5	972.3	1,564.81	1,068.15	1,985.06	1,161.86
7	539	453.53	688.13	845.53	799.06	1,163.21	886.71
8	489	181.89	480.04	399.77	592.28	596.42	671.13
9	275	64.84	330.65	168.01	435.46	271.83	504.26
10	524	20.8	225.17	63.55	317.85	111.5	376.41
11	174			21.85	230.48	41.58	279.31
12	354			6.89	166.12	14.21	206.15
13	129			2	119.08	4.48	151.39
14	143			0.54	84.92	1.31	110.67

How to cite this article: Seyed Ehsan Saffari, John Carson Allen, Robiah Adnan, Seng Huat Ong, Sim SZ, William Greene. Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros. Biostat Biom Open Access J. 2019; 9(5): 555773.
DOI: 10.19080/BBOAJ.2019.09.555773

15	176			0.14	60.28	0.36	80.55
16	91					0.09	58.4
17	63					0.02	42.18
18	65					0.01	30.35
19	31					0	21.77
20	113					0	15.57

[1]Hurdle Poisson

[2]Hurdle Conway-Maxwell Poisson

Table 6: Simulation scenarios based on selected parameterizations of HCMP and HP regression models.

Scenario*	50% of data simulated from HCMP[1] model					50% of data simulated from HP[2] model			
	β_0	β_1	β_2	a_0	ν	β_0	β_1	β_2	a_0
Model (a)	1	0.8	0.8	0.3	2	1	0.8	0.8	0.3
Model (b)	1	1	1	2	2	1	1	1	2
Model (c)	1	1.5	0.9	1	2	1	1.5	0.9	1
Model (d)	1	0.8	0.8	0.3	0.85	1	0.8	0.8	0.3
Model (e)	1	1	1	2	0.85	1	1	1	2
Model (f)	1	1.5	0.9	1	0.85	1	1.5	0.9	1

* a_1 is solved by using the equation $logit(W_0) = a_0 + a_1 z_1$.
[1]Hurdle Conway-Maxwell Poisson

[2]Hurdle Poisson

Table 7: Mean AIC and log-likelihood (in brackets) comparisons for fitted truncated hurdle regression models.

Scenario	Proportion of zeros=0.4		Proportion of zeros=0.3		Proportion of zeros=0.2	
	THP[1]	THCMP[2]	THP	THCMP	THP	THCMP
(i) 5% truncation at the tail						
Model (a)	703.06	691.92	774.2	758.69	826.7	806.29
	(-346.53)	(-339.96)	(-382.10)	(-373.35)	(-408.35)	(-397.15)
Model (b)	770.55	742.74	854.73	817.39	921.06	873.48
	(-380.28)	(-365.37)	(-422.36)	(-402.69)	(-455.53)	(-430.74)
Model (c)	854.48	793.68	956	876.16	1041.2	941.91
	(-422.24)	(-390.84)	(-473.00)	(-432.46)	(-515.60)	(-464.95)
Model (d)	761.15	760.85	839.51	838.76	898.03	896.9
	(-375.57)	(-374.42)	(-414.76)	(-413.38)	(-444.01)	(-442.45)
Model (e)	806.48	804.2	892.68	889.28	958.91	954.52
	(-398.24)	(-396.10)	(-441.34)	(-438.64)	(-474.45)	(-471.26)
Model (f)	855.12	848.16	950.65	941.06	1025.7	1013.68
	(-422.56)	(-418.08)	(-470.33)	(-464.53)	(-507.85)	(-500.84)
(ii) 10% truncation at the tail						
Scenario	Proportion of zeros=0.4		Proportion of zeros=0.3		Proportion of zeros=0.2	
	THP	THCMP	THP	THCMP	THP	THCMP
Model (a)	594.13	592.08	664.17	659.8	714.2	707.19
	(-292.07)	(-290.04)	(-327.09)	-323.90)	(-352.10)	(-347.60)
Model (b)	647.77	639.55	731.68	716.49	796.49	774.54
	(-318.27)	(-313.78)	(-360.84)	-352.25)	(-393.24)	(-381.27)

How to cite this article: Seyed Ehsan Saffari, John Carson Allen, Robiah Adnan, Seng Huat Ong, Sim SZ, William Greene. Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros. Biostat Biom Open Access J. 2019; 9(5): 555773.
DOI: 10.19080/BBOAJ.2019.09.555773

Model (c)	707.16	685.25	806.9	771.56	889.89	838.79
	(-348.58)	(-336.62)	(-398.45)	-379.78)	(-439.95)	(-413.40)
Model (d)	665.24	665.61	744.51	744.68	802.65	802.59
	(-327.62)	(-326.80)	(-367.25)	-366.34	(-396.32)	(-395.30)
Model (e)	707.56	706.98	795.05	793.79	862.12	860.3
	(-348.78)	(-347.49)	(-392.53)	-390.89	(-426.06)	(-424.15)
Model (f)	750.31	747.29	846.13	841.69	922.78	916.63
	(-370.07)	(-367.65)	(-418.07)	-414.84	(-456.39)	(-452.32)

[1]Truncated Hurdle Poisson
[2]Truncated Hurdle Conway-Maxwell Poisson

A subset of RWM data where under-dispersion is present was considered. We have focused on the individuals with low health satisfaction score (<4 out of 10) and split up the data set by marriage status to come up with two under-dispersed scenarios. The married covariate was excluded from the list of independent variables and other covariates were kept in. TP and THCMP models with a right truncation point of 4 applied to the data. The goodness-of-fit statistics (-2LL and AIC) for THCMP/THP were 653.2/686 and 675.2/706 for unmarried individuals, and 1847.5/1908.9 and 1869.5/1928.9 for married individuals. The dispersion parameter of THCMP model was significant in both scenarios (4.12 (3.02, 5.33) and 3.42 (2.77, 4.07), p-value<0.001).

A Simulation study

We conducted a simulation study to assess and compare THP and THCMP regression model performance. Goodness-of-fit was measured using the Akaike information criterion (AIC). Data were simulated with 50% of responses generated by the HCMP regression model and 50% by the HP model. Simulation models were structured as $\log it(W_0) = a_0 + a_1 z_1$ and linear function $\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$, where z_1, x_1 and x_2 were generated from a uniform distribution on $[0,1]$. A sample size of $n = 200$ was used in conjunction with varying proportions of zero outcomes and w_0 set to values of 0.2, 0.3 and 0.4. Truncation percentages were set at 5% and 10% of the simulated distribution tail, although actual simulated truncation percentages were not always exactly equal to 5% and 10%.

The six working simulation models, differing in their coefficient parameters, are shown in Table 6. Three are mixtures of under-dispersed (υ>1) HCMP regression models with the HP regression model. These models generate count data outcomes of 0, 1, 2, and 3 for the most part, resulting in short-tailed count frequency distributions. Conversely, the remaining three models are mixtures of over-dispersed (υ<1) HCMP regression models with the HP regression model, which generate more extended right tails in the response distribution. Using the three combinations of coefficient parameters, the model with $\beta_0 = 1$, $\beta_1 = 1.5$, $\beta_2 = 0.9$, $a_0 = 1$ generates long right-tailed count distributions with high count outcomes; the model with $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 1$, $a_0 = 2$ generates distributions with tails of intermediate length; and the model with $\beta_0 = 1$, $\beta_1 = 0.8$, $\beta_2 = 0.8$, $a_0 = 0.3$ distributions with relatively short tails.

The number of replications was set at 1,000, which was sufficient for our purposes. Additional replications would have unnecessarily increased the computational burden. The simulation was programmed in FORTRAN. Maximum likelihood estimates were obtained via numerical maximization using the simulated annealing algorithm [20].

Simulation results are summarized in Table 7(i) and (ii) where the values given are the average values of 1,000 replications. From Table 7(i), the average AIC for the THCMP regression model was significantly lower than that for the THP regression model for simulation models (a) to (c). When the truncation percentage was increased to 10% as in Table 7(ii), the average AIC difference was smaller as compared to Table 7(i). Nevertheless, the THCMP regression model performed slightly better for model (a) and definitely outperformed models (b) and (c). For model (d), the average AIC difference between the THP and THCMP regression models was less than 2, regardless of the proportion of zeros and the percentage of truncation. For models (e) and (f), the THCMP regression model was slightly better than the THP model with 5% truncation in the tail, where the AIC average difference was greater than 2.

Discussion

In this paper, we introduce the THCMP regression model and illustrate its application in an analysis of the RWM data in which the outcome variable of interest is the number of doctor visits occurring in a fixed interval. We show how the THCMP model can be used to handle dual data anomalies of excess zeros and extreme values in a count response variable. Parameter estimates and standard errors for the TCMP model—and the alternative THP model-were obtained for selected data truncation levels using ML estimation. Covariate effects were interpreted in the context of model link functions.

A comparison of goodness-of-fit of the THCMP and THP models to the RWM data, as assessed by -2LL, AIC and BIC, showed better performance for the THCMP model at the three truncation levels studied (6.65%, 3.08% and 1.75%). The percentage of zeros in the response variable of the RWM case study was 37.1%—which is cited in the literature as a threshold for excess zeros [18,19].

How to cite this article: Seyed Ehsan Saffari, John Carson Allen, Robiah Adnan, Seng Huat Ong, Sim SZ, William Greene. Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros. Biostat Biom Open Access J. 2019; 9(5): 555773.
DOI: 10.19080/BBOAJ.2019.09.555773

The results showed that a right truncated hurdle model can fit these data with inflation at zero better than a simple right-truncated model where excess zero part of the model is not taken into account. We also examined goodness-of-fit for the right truncated Poisson and CMP models as well as the excess zero models in the RWM case study analysis.

In the RWM analysis, the THCMP model exhibited better agreement between observed and predicted zero frequency counts than the THP model for all three truncation points. However, greater truncation levels resulted in greater bias in estimating the zero frequency for both models, although to a lesser degree for the THCMP model. In addition, the THCMP model generally exhibited better goodness-of-fit in modelling counts greater than zero. Exceptions favoring the THP model were a few cases involving frequency estimates for 2 and 7 at 6.65% truncation, and 8 and 9 at 3.08% and 1.75% truncation. The case study analysis results suggest advantages of the THCMP regression model compared to the THP model for analyzing 'real life' count data when there are both an excess of zeros and extreme values in the observed response.

In the under-dispersed subset of RWM data sets, we have tried hurdle negative binomial (HNB) and hurdle generalized Poisson (HGP) models with right truncation approach to investigate the performance of HCMP over HNB and HGP. HNB model was not converged because the final Hessian matrix, though full rank, had at least one negative eigenvalue, and therefore the second-order optimality condition violated. This is expected as NB model can handle over-dispersion scenario not under-dispersion. HGP model also was not converged as the final Hessian matrix was not positive definite and therefore the estimated covariance matrix was not full rank and may not be reliable. Hence, the HCMP model outperform HNB and HGP when the under-dispersed right truncated outcome has excess zeros.

In the simulation study, mean AIC for the THCMP model was significantly lower than mean AIC for the THP model (AIC mean difference >11) in under-dispersed scenarios with 5% tail truncation, indicating substantial improvement in fitting the data [21]. At the 5% truncation level, a clear lack-of-fit of THP is exhibited for simulation models (a) to (c), indicating under-dispersed scenarios, compared to THCMP. However, at 10% truncation, the mean AIC difference between models was smaller in under-dispersed scenarios. In the simulation study, the THCMP regression model performed slightly better than the THP model in the short-tailed data scenario (model (a)) and clearly outperformed THP in the medium- and long-tailed data scenarios (models (b) and (c)). In an over-dispersed scenario with short tail (model (d)), the average AIC difference for the THP and THCMP models was less than 2 regardless of the proportion of zeros or the truncation percentage, implying similar performance in this case. For over-dispersed scenarios involving medium and long tails (model (e) and (f)), and less than 5% truncation, the THCMP regression model performed slightly better than the THP model (AIC mean difference >2). In short, the comparison of average AIC for the THCMP and THP regression models indicated superiority of the THCMP to the THP model for outcome data with lower than expected proportions of zeros, lower percentages of tail truncation, and consists of mostly low values of the response variable.

A strong point of the simulation study is that the CMP model, unlike more frequently used models such as the negative binomial model, is more flexible and able to handle under-dispersion as well as over-dispersion. The negative binomial model was not entertained as an alternative model in this study because the NB model cannot accommodate under-dispersed data. The THCMP model clearly outperformed THP model for under-dispersed scenarios. Therefore, the THCMP regression model would be expected to provide better outcomes in terms of the parameter estimates and goodness-of-fit statistics when data are under-dispersed—even when compared to some alternative models such as hurdle negative binomial model.

In summary, we introduced the truncated hurdle Conway-Maxwell Poisson regression model. We carried out a simulation study, based on data generated from a mixture of HCMP (50%) and HP (50%) probability models, which showed the THCMP regression model accommodated various degrees of truncation and anomalous zero frequencies better than the THP regression model. We recommend the THCMP regression model as a flexible distribution for analyzing count data exhibiting the dual anomalies involving zero frequencies and a low level of tail truncation in handling over- and under-dispersion.

References

1. Cameron AC, Trivedi PK (1986) Econometrics Models Based on Count Data: Comparison and Applications of Some Estimators and Tests. Journal of Applied Econometrics 1: 29-54.
2. Geil P, Million A, Rotte R, Zimmermann KF (1997) Economic Incentives and Hospitalization in Germany. Journal of Applied Econometrics 12(3): 295-311.
3. Pohlmeier W, Ulrich V (1995) An Econometric Model of the Two-Part Decision Making Process in the Demand for Health Care. Journal of Human Resources 30: 339-361.
4. Conway RW, Maxwell WL (1962) A queuing model with state dependent service rates. Journal of Industrial Engineering 12: 132-136.
5. Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P (2005) A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. Applied Statistics 54: 127-142.
6. Borle S, Boatwright P, Kadane JB (2006) The timing of bid placement and extent of multiple bidding: An empirical investigation using eBay online auctions. Statistical Science 21(2): 194–205.
7. Kadane J, Shmueli G, Minka G, Borle T (2006) Boatwright P. Conjugate analysis of the Conway Maxwell Poisson distribution. Bayesian Analysis 1: 363-374.
8. Rodrigues J, de Castro M, Cancho VG, Balakrishnan N (2009) COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. Journal of Statistical Planning and Inference 139: 3605-3611.
9. Khan NM, Jowaheer V (2010) A comparison of marginal and joint generalized quasi-likelihood estimating equations based on the COM-

How to cite this article: Seyed Ehsan Saffari, John Carson Allen, Robiah Adnan, Seng Huat Ong, Sim SZ, William Greene. Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros. Biostat Biom Open Access J. 2019; 9(5): 555773.
DOI: 10.19080/BBOAJ.2019.09.555773

Poisson GLM: Application to car breakdowns data. International Journal of Mathematical and Statistical Sciences 67: 543-546.

10. Gupta RC, Sim SZ, Ong SH (2014) Analysis of discrete data by Conway-Maxwell Poisson distribution. Advances in Statistical Analysis 98(4): 327-343.

11. King G (1989) Event Count Models for International Relations: Generalizations and Applications. International Studies Quarterly 33: 123-147.

12. Cameron AC, Trivedi PK (1998) Regression Analysis of Count Data. Cambridge University Press: New York.

13. Bohning D, Dietz E, Schlattmann P, Mendona L, Kirchner P (1999) The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Association 162(A): 195-209.

14. Zhou X, Tu W (1999) Comparison of Several Independent Population Means when Their Samples Contain Log-Normal and Possibly Zero Observations. Biometrics 55: 645-651.

15. Saffari SE, 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.

16. Saffari SE, Adnan R, Greene W (2012) Investigating the impact of excess zeros on hurdle-generalized Poisson regression model with right censored count data. Statistica Neerlandica 67: 67-80.

17. Saffari SE, Adnan R, Greene W(2012) Hurdle negative binomial regression model with right censored count data. SORT-Statistics and Operations Research Transactions 36:181-193.

18. Riphahn RT, Wambach A, Million A (2003) Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation. Journal of Applied Econometrics 18(4): 387-405.

19. Greene W (2005) Functional Form and Heterogeneity in Models for Count Data. Foundations and Trends in Econnometrics 1(2): 113-218.

20. Metropolis N, Rosenbluth AW, Rosenbluth MN, Telle AH, Teller E (1953) Equation of State Calculations by Fast Computing Machines. Journal of Chemical Physics 21: 1087-1092.

21. Burnham KP (2002) Anderson Model Selection and Multimodel Inference: a practical information-theoretic approach, 2nd edition: Springer-Verlag, New York, USA.

DOI: 10.19080/BBOAJ.2019.09.555773

How to cite this article: Seyed Ehsan Saffari, John Carson Allen, Robiah Adnan, Seng Huat Ong, Sim SZ, William Greene. Frequency of Visiting a Doctor: A right Truncated Count Regression Model with Excess Zeros. Biostat Biom Open Access J. 2019; 9(5): 555773.
DOI: 10.19080/BBOAJ.2019.09.555773