Appropriate statistical model for zero-inflated count data: simulation based study

Ayan CHOWDHURY[1*], Soma Chowdhury BISWAS[2]
[1]Department of Statistics, University of Chittagong, ayan.statcu09@yahoo.com, Chittagong-4331, Bangladesh
[2]Department of Statistics, University of Chittagong, soma.stat@cu.ac.bd, Chittagong-4331, Bangladesh

Abstract

Zero-inflated count data is easily reached in real field. Over-dispersion is the consequence of zero inflation in count data. For modeling this kind of count data, several zero adjusted models such as Zero Inflated Poisson, Zero Inflated Negative Binomial, Hurdle Poisson and Hurdle Negative Binomial models are more suitable than basic statistical models. The best zero adjusted model selection is the key aim of this research. In this study, R code has been used to simulate datasets as well as to compare these models based on Akaike information criterion, Bayesian information criterion and Vuong test. The result of this study suggests that Hurdle Negative Binomial model has been preferred as the best fitted model for count data with excess of zero.

Keywords: ***Hurdle model, Simulation, Vuong test, Zero-inflated model.***

1. Introduction

Count data means discrete number of occurrences of an event in a fixed period of time. A count variable can takes positive integer values or zero because an event cannot occur a negative number of times. There are numerous example of the use of count variables in psychology, public health, insurance, epidemiology, behavioral sciences and many other research areas. Poisson regression analysis is a technique used for modeling the count data [2]. It is a non-linear regression analysis of the Poisson distribution, where the analysis is highly suitable for use in analyzing count data. Poisson model is a part of class of models in generalized linear models (GLM). It uses natural log as the link function and models the expected value of response variable. The natural log in the model ensures that the predicted values of response variable will never be negative. This model is used under two principal assumptions: one is that events occur independently over given time and the other is that the conditional mean and variance are equal. However, this restriction is violated in many applications because data often exhibits over-dispersion. Over-dispersion occurs when the variance is significantly larger than the mean. Generally, two sources of over-dispersion are determined which are heterogeneity of the data and excess of zeros.

In case of over-dispersion problem due to heterogeneity of the data, Negative Binomial (NB) model may be used instead of Poisson model [4], [10]. In real field, it is possible that count data is heterogeneous with excess of zero. As a result, over-dispersion problem is occurred due to both causes heterogeneity of the data and excess of zero. Zero-inflated count data cannot be modeled accurately with Negative Binomial model.

In such situation, zero-inflated models (i.e. Zero Inflated Poisson and Zero Inflated Negative Binomial) and hurdle models (i.e. Hurdle Poisson and Hurdle Negative Binomial) are more appropriate for modeling this kind of count data. The main motivation for zero-inflated count models is that real life data frequently display over-dispersion and excess of zero [5]. In zero-inflated model it is assumed that the zero observation have two different sources i.e. “structural” and “sampling”. The structural zero observation happened naturally but the sampling zero observation happened by chance.

Hurdle model is another model which also provides a way of modeling the excess zeros in addition to allowing for over-dispersion, which is proposed by Mullahy (1986) [8]. But hurdle model assumes that all zeros of data are from only “structural” source. Moreover, both zero-inflated and hurdle models have statistical advantage to standard Poisson and Negative Binomial models in such a way that these models the preponderance of zeros as well as the distribution of positive counts simultaneously. This study has been conducted to compare zero inflated models (ZIP, ZINB) and hurdle models (HP, HNB) using simulated data. Furthermore, this study

* Corresponding author: ayan.statcu09@yahoo.com

will provide valuable information about several zero adjusted models as well as will present the best model for modeling zero-inflated count data which is selected based on AIC, BIC and Vuong test.

2. Materials and Methods

2.1. Models

2.1.1. Zero-inflated Model

Zero-inflated model is a mixture of two statistical processes, one always generating zero counts and the other generating both zero and non-zero counts, which is introduced by Lambert (1992) [5]. This is a two parts model which provides a way of count data modeling with excess zeros additionally to allowing for over-dispersion. In case of zero inflated model, a Logit model with Binomial assumption is used to determine if an individual count outcome is from the always zero or the not always zero group and then Poisson or Negative Binomial model is used to model outcomes in the not always zero group [3], [6]. For case *i*, with probability π_i the only possible response of the first process is zero counts, and with probability of $(1-\pi_i)$ the response of the second process is governed by Poisson or Negative Binomial distribution with mean μ_i .

The probability mass function of Zero Inflated Poisson (ZIP) distribution is given by:

$$\Pr(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i} & ;y_l = 0 \\ (1-\pi_i)\dfrac{e^{-\mu_i}\mu_i y_i}{y_i!} & ;y_l > 0 \end{cases} \quad (1)$$

And the probability mass function [14] of Zero Inflated Negative Binomial (ZINB) distribution is given by:

$$\Pr(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)(1+\alpha\mu_i)^{-\alpha^{-1}} & ;y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma(y_i+\alpha^{-1})(\alpha\mu_i)^{y_i}}{\Gamma(y_i+1)\Gamma(\alpha^{-1})(1+\alpha\mu_i)^{y_i+\alpha^{-1}}} & ;y_i > 0 \end{cases} \quad (2)$$

where $0 \le \pi_i \le 1$, $\mu_i \ge 0$ and α is the dispersion parameter. The parameter μ_i is expressed as: $\mu_i = \exp(\theta' x_i)$ where, θ is the (p+1)×1 vector of unknown parameters associated with the known covariates vector x_i and p is the number of covariates.

The parameter π_i is often referred as the zero-inflation factor, which is the probability of zero counts from the binary process. According to Lambert, we can model π_i using a Logit model given by: $\log it(\pi_i) = \delta' z_i$, where, δ is the (q+1)×1 vector of zero-inflated coefficients to be estimated which is associated with the known zero-inflation covariates vector z_i and q is the number of the covariates in the model. In the terminology of generalized linear models (GLMs), $\log(\mu_i)$ and $\text{logit}(\pi_i)$ are the log link for the Poisson or Negative Binomial mean and logit link for Bernoulli probability of success respectively [5].

Zero-inflated models (ZIP and ZINB) are expressed as:

$$\begin{aligned} \log(\mu_i) &= \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \theta_3 x_{i3} + \ldots\ldots\ldots\ldots + \theta_p x_{ip} \\ \log it(\pi_i) &= \delta_0 + \delta_1 z_{i1} + \delta_2 z_{i2} + \delta_3 z_{i3} + \ldots\ldots\ldots\ldots + \delta_q z_{iq} \end{aligned} \quad (3)$$

The log-likelihood function of ZIP model is:

$$\log L = \sum_{y_i=0} \log\left[\pi_i + (1-\pi_i)e^{-\mu_i}\right] + \sum_{y_i\neq 0}\left[\log(1-\pi_i) - \mu_i + y_i \log\mu_i - \log(y_i!)\right] \quad (4)$$

And the log-likelihood function of ZINB model is:

$$\log L = \sum_{y_i=0} \log\left[\pi_i + (1-\pi_i)(1+\alpha\mu_i)^{-1/\alpha}\right] + \sum_{y_i\neq 0}\left[\log(1-\pi_i) + \log\frac{\Gamma(y_i+1/\alpha\alpha}{\Gamma(y_i+1)\Gamma 1/\alpha\alpha} + y_i\log(\alpha\alpha_i) - (y_i+1/\alpha\alpha\log(1+\alpha\mu_i)\right] \quad (5)$$

The parameters of this model can be estimated using maximum likelihood estimation.

2.1.2. Hurdle Model

Hurdle model is another model for modeling over-dispersed count data with excess zeros which is introduced by Mullahy (1986) [8]. This model assumes that two different processes drive the zero and non-zero counts respectively. The hurdle component of the model corresponds to the probability that the count is non-zero, while the count component corresponds to the distribution of positive counts. This model contains also two parts: the first part is a binary (presence/absence) outcome model (e.g. a logistic regression) and the second part is a count model which is governed by truncated distribution. In case of Hurdle Poisson (HP) model and Hurdle Negative Binomial (HNB) model, second part is governed by truncated Poisson distribution and truncated Negative Binomial distribution respectively.

The unconditional probability mass function of Hurdle Poison (HP) distribution is:

$$\Pr(Y_i = y_i) = \begin{cases} \pi_i & ;y_i = 0 \\ (1-\pi_i)\dfrac{e^{-\mu_i}\mu_i y_i}{(1-e^{-\mu_i})y_i!} & ;y_i > 0 \end{cases} \quad (6)$$

And the unconditional probability mass function of Hurdle Negative Binomial (HNB) distribution is:

$$\Pr(Y_i = y_i) = \begin{cases} \pi_i & ;y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\dfrac{(1+\alpha\mu_i)^{-\alpha^{-1}-y_i}\alpha^{y_i}\mu_i^{y_i}}{1-(1+\alpha\mu_i)^{-\alpha^{-1}}} & ;y_i > 0 \end{cases} \quad (7)$$

where, $0 \le \pi_i \le 1$, $\mu_i \ge 0$ and α is the dispersion parameter. The conditional mean μ_i of the Poisson or Negative Binomial distribution is expressed as, $\mu_i = \exp(\theta' x_i)$, where x_i is a $(p+1)\times 1$ vector of covariates, θ is a $(p+1)\times 1$ vector of parameters to be estimated and p is the number of covariates in the model.

The parameter π_i is the probability of observing a zero count and $(1-\pi_i)$ is the probability of observing a positive count. For the hurdle model, the zero hurdle component describes the probability of observing a positive count whereas, for the zero-inflated model, the zero-inflation component predicts the probability of observing a zero count from the point mass component [15]. We can model $(1-\pi_i)$ using a Logit model given by: $\log it(1-\pi_i) = \delta' z_i'$, where z_i is a $(q+1)\times 1$ vector of covariates, δ is a $(q+1)\times 1$ vector of parameters to be estimated and q is the number of the covariates in the model.

Hurdle models (HP and HNB) are expressed as:

$$\begin{aligned} \log(\mu_i) &= \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \theta_3 x_{i3} + \ldots\ldots\ldots\ldots + \theta_p x_{ip} \\ \log it(1-\pi_i) &= \delta_0 + \delta_1 z_{i1} + \delta_2 z_{i2} + \delta_3 z_{i3} + \ldots\ldots\ldots + \delta_q z_{iq} \end{aligned} \quad (8)$$

The log-likelihood function of HP is:

$$\log L = \sum_{y_i=0} \log \pi_i + \sum_{y_i \neq 0} \left[\log(1-\pi_i) - \mu_i + y_i \log \mu_i - \log\left(1-e^{-\mu_i}\right) - \log(y_i!) \right] \quad (9)$$

And the log-likelihood function of HNB is:

$$\log L = \sum_{y_i=0} \log \pi_i + \sum_{y_i \neq 0} \left[\log(1-\pi_i) + \log \gamma - \log\left\{1-(1+\alpha\mu_i)^{-\alpha^{-1}}\right\} \right] \quad (10)$$

where, $\gamma = \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma\Gamma(\alpha^{-1})}(1+\alpha\mu_i)^{-\alpha^{-1}-y_i}(\alpha\mu_i)^{y_i}$ (11)

The parameters of this model can be estimated using maximum likelihood estimation.

2.2. Model Selection

In this study, to select the best model among ZIP, ZINB, HP and HNB models Akaike information criterion (AIC), Bayesian information criterion (BIC) and Vuong test have been used.

2.2.1. AIC and BIC

Information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are extensively used to compare and select the best model among a set of models. Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data which is introduced by Akaike (1973) [1]. Bayesian information criterion (BIC) is another important approach to compare and select the consistent model from a set of candidate models which is introduced by Schwarz (1978) [11]. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC and BIC value.

Akaike information criterion (AIC) is defined as:

$$\text{AIC} = -2\log(\text{L}) + 2\text{k} \quad (12)$$

Bayesian information criterion (BIC) is defined as:

$$\text{BIC} = -2\log(\text{L}) + \text{k}\log(\text{n}) \quad (13)$$

where, L is maximum value of the likelihood function for the model, n is sample size and k is the number of parameters to be estimated.

2.2.2. Vuong Test

Vuong test is used to compare two statistical models for count data [6], [7] which is introduced by Vuong (1989) [12]. It is a test that is based on a comparison of the predicted probabilities of two models.

Let's define,

$$u_i = \log\left\{\frac{P_1(Y_i|X_i)}{P_2(Y_i|X_i)}\right\} \quad (14)$$

where, $P_1(Y_i|X_i)$ and $P_2(Y_i|X_i)$ are the predicted probability of observed count for case *i* from model 1 and model 2 respectively.

Null hypothesis of Vuong test is considered as:

H_0: Both models are equally appropriate.

Against the hypothesis:

H_A: Model 1 is better than model 2,

or model 2 is better than model 1.

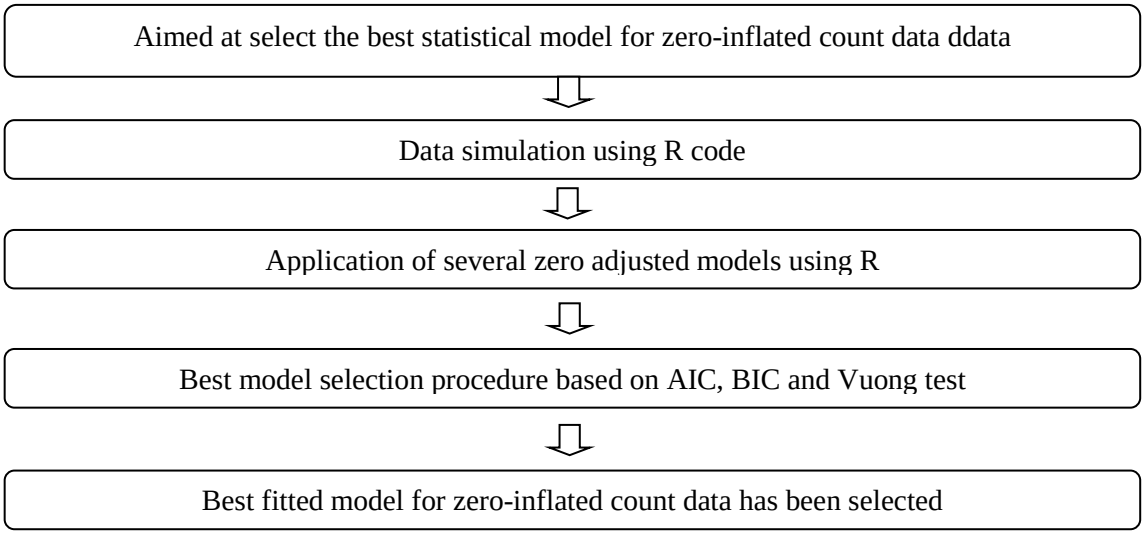
Under the null hypothesis, the Vuong test statistic is given by:

$$V = \frac{\bar{u}\sqrt{n}}{\sqrt{\mathrm{var}(u)}} \sim N(0,1) \tag{15}$$

where, $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$, $Var(u) = \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2$ and n is sample size.

Mathematically, if V is greater than Z_α then model 1 is better than model 2 at α level of significance. Conversely, if V is less than $-Z_\alpha$ then model 2 is better than model 1 at α level of significance. Otherwise, model 1 and model 2 both models are equally appropriate at α level of significance.

The flowchart of this study has been shown in **Fig. 1**.



[1]
[2] **Fig. 1.** Flowchart of this study

2.3. Simulation Study

Simulation studies allow researchers to answer specific questions about data analysis, statistical power and best practices for obtaining accurate results in empirical research. This study has been conducted based on simulated datasets. To complete this study all programs have been written in **R** (**version 3.2.3; packages: MASS, pscl)** codes. The parameter vector $(\beta_0, \beta_1, \beta_2)$ has been used in the simulation study. The parameters values have been fixed as $\beta_0 = 0.5, \beta_1 = 1, \beta_2 = 1$ to simulate count data with excess zeros. In this study, 3 (three) sets zero-inflated count data consisting of 100, 500 and 1000 observations respectively have been simulated. The following equation(s) has been used to simulate zero-inflated count datasets using R.

$$Y \sim Poisson(\mu)$$

$$\log \mu = 0.5 + X_1 + X_2 \tag{16}$$

where, X_1 and X_2 have been generated from Normal distribution. For simplicity of zero-inflated count data modeling, 2 (two) binary covariates have been generated by random sampling method and then ZIP, ZINB, HP and HNB models have been applied to each dataset. AIC, BIC and Vuong test values have been computed for

ZIP, ZINB, HP and HNB models using each simulated dataset to compare these models as well as to select the best model.

3. Results

Each dataset contains with excess of zero (i.e. frequency of zero is maximum). The percentage of zero in case of each dataset has been shown in the Table 1.

Table 1. Percentage of zero in several datasets

Dataset (sample size)	Dataset 1 (n = 100)	Dataset 2 (n = 500)	Dataset 3 (n = 1000)
Percentage (%) of zero	28.0	33.4	29.5

The result of AIC, BIC and Vuong test values of ZIP, ZINB, HP and HNB in case of several datasets have been shown in the following Table(s).

Table 2. AIC and BIC value of ZIP, ZINB, HP and HNB models

Model	Data set 1 (n = 100)		Data set 2 (n = 500)		Data set 3 (n = 1000)	
	AIC	BIC	AIC	BIC	AIC	BIC
ZIP	859.60	875.23	4350.00	4375.29	9820.00	9849.45
ZINB	498.40	516.64	2402.00	2431.50	5042.00	5076.35
HP	859.60	875.23	4350.00	4375.29	9820.00	9849.45
HNB	487.20	505.44	2384.00	2413.50	4992.00	5026.35

Table 3. Vuong test result to compare ZIP, ZINB, HP and HNB models

[3]

Data set	Model	ZIP	ZINB	HP	HNB
Data set 1 (n = 100)	ZIP	----------			
	ZINB	V = -2.8501 P = 0.002 ZINB is better	----------		
	HP	V = -0.8479 P = 0.198 ZIP = HP	V = 2.8468 P = 0.002 ZINB is better	----------	
	HNB	V = -2.8730 P = 0.002 HNB is better	V = -1.7754 P = 0.038 HNB is better	V = -2.8700 P = 0.002 HNB is better	----------
Data set 2 (n = 500)	ZIP	----------			
	ZINB	V = -3.4546 P = 0.002 ZINB is better	----------		
	HP	V = -0.03641 P = 0.485 ZIP = HP	V = 3.4546 P = 0.002 ZINB is better	----------	
	HNB	V = -3.4394 P = 0.002 HNB is better	V = -1.6826 P = 0.046 HNB is better	V = -3.4394 P = 0.002 HNB is better	----------
Data set 3 (n = 1000)	ZIP	----------			
	ZINB	V = -6.2726 P = 0.000	----------		

[4]

		ZINB is better			
	HP	V = -0.2548 P = 0.399 ZIP = HP	V = 6.2726 P = 0.000 ZINB is better	----------	
	HNB	V = -6.2487 P = 0.000 HNB is better	V = -3.113 P = 0.001 HNB is better	V = -6.2487 P = 0.000 HNB is better	----------

V=Vuong statistic, P = P-value; V > 1.64 indicates column model had significantly better fit than the column model at < -1.64 indicates row model had significantly better fit than the significance. 5% level of

From Table 2 it is seen that, the AIC and BIC value of Hurdle Negative Binomial (HNB) model is lowest in case of each dataset, which indicate that HNB model is the best model for modeling zero-inflated count data. The result of Vuong test has been shown in the Table 3 for 3 (three) datasets of sample size 100, 500 and 1000 respectively, which also indicates that HNB model is the best model for modeling this kind of count data.

4. Discussion

For modeling over-dispersed count data with excess of zero, zero-inflated models and hurdle models are more suitable than standard Poisson and Negative Binomial (NB) models [5], [8]. Although the choice between the hurdle and zero-inflated models should be based on the aim and endpoints of the study, but it is noted that hurdle model allows for over-dispersion and also accommodates presence of excess zeros, is more appropriate than zero-inflated model [3]. In some real fields, Zero-Inflated Negative Binomial (ZINB) and Hurdle Negative Binomial (HNB) models both lead to the same qualitative results and very similar model fits. But the Hurdle Negative Binomial (HNB) model is slightly preferable because it has the nicer interpretation [15]. Under the conditions of zero-inflation and over-dispersion, the hurdle model is more suitable because it performed well consistently and relatively easy to interpret and implement [9]. This study has been conducted using simulated data to compare several zero adjusted count data models such as ZIP, ZINB, HP and HNB models. According to AIC, BIC and Vuong test values, this study has been focused on Hurdle Negative Binomial (HNB) model as the best fitted model for modeling zero-inflated count data.

5. Conclusions

Zero-inflated count data have been implemented in real life and zero adjusted count models are being usually used in various disciplines such as public health, insurance, epidemiology, behavioral sciences, econometrics etc. Over-dispersion is the result zero-inflation which leads to serious underestimation of standard errors and ambiguous implication for the estimated parameters [13]. As a result, several estimation methods for several models have been anticipated to handle over-dispersed count data. Appropriate statistical model is indispensable for estimating parameters correctly which play significant role on interpretations of any study. The estimated parameters of the best fitted statistical model lead the accurate result of the analysis. According to this study, we suggest to apply Hurdle Negative Binomial (HNB) model as the best fitted statistical model in case of zero-inflated count data modeling which overcomes the over-dispersion problem.

An appropriate statistical model for zero-inflated over-dispersed count data has been suggested in this research. Hurdle model may be also applied for under-dispersed count data with zero-inflation. In rare practical field, under-dispersion may be occurred in case of count data without excess of zero which is also a crisis to estimate the parameters accurately. Further study may be conducted to choice a suitable statistical model for under-dispersed count data without zero-inflation.

References

[1] Akaike, H., "Information theory and an extension of the maximum likelihood Principle", In B. N. Petrov and F. Csaki (Eds.), Second international symposium on information theory, Budapest: Academiai Kiado, pp. 267-281, 1973.

[2] Cameron, A. C., Trivedi, P. k., "Regression Analysis of Count Data", Cambridge: Cambridge University Press, 1998.

[3] Chipeta, M. G., Ngwira, B. M., Simoonga, C., Kazembe, L. N., "Zero adjusted models with applications to analysing helminths count data", BMC Research Notes, 7:856, pp. 1-11, 2014, Available: http://www.biomedcentral.com/1756-0500/7/856.

[4] Gardner, W., Mulvey, E. P. and Shaw, E. C., "Regression analysis for counts and rates: Poisson, Over-dispersed Poisson and Negative Binomial

models",Psychological Bulletin, 118, pp. 392-404, 1995.
[5] Lambert, D., "Zero-inflated Poisson regression, with an application to defects in manufacturing", Technometrics, 34, pp. 1-14, 1992.
[6] Liu, W. S., Cela, J., "Count Data Models in SAS®", Statistics and Data Analysis, SAS Global Forum, Paper 371, 2008.
[7] Moineddin, R., Meaney, C., Agha, M., Zagorski, B., Glazier, R. H., "Modelling factors influencing the demand for emergency department services in Ontario: A comparison of methods", BMC Emergency Medicine, 11:13, 2011.
[8] Mullahy, J., "Specification and Testing of Some Modified Count Data Models", Journal of Econometrics, 33, pp. 341-365, 1986.
[9] Potts, J. M., Elith, J., "Comparing species abundance models", ecological Modelling, Elsevier, 199, pp. 153–163, 2006.
[10] Saffari, S. E., Adnan, R., Greene W., "Handling of Over-Dispersion of Count Data via Truncation using Poisson Regression Model", Journal of Computer Science & Computational Mathematics, 1(1), August 2011.
[11] Schwarz, G., "Estimating the dimension of a model", Annals of Statistics, 6, pp. 461-464, 1978.
[12] Vuong, Q. H., "Likelihood ratio tests for model selection and non-nested Hypotheses", Econometrica, 57(2), pp. 307–333, 1989.
[13] Yang, Z., Hardin, J.W., Addy, C.L., "Testing over-dispersion in the zero-inflated Poisson model" Journal of Statistical Planning and Inference, 139(9), pp. 3340-3353, 2009.
[14] Yau, K., Wang, K., Lee, A., "Zero-inflated negative binomial mixed regression modelling of over dispersed count data with extra zeros", Biometrical Journal, 45, pp. 437-452, 2003.
[15] Zeileis, A., Kleiber, C., Jackman, S., "Regression Models for Count Data in R", Journal of Statistical Software, 27(8), 2008.