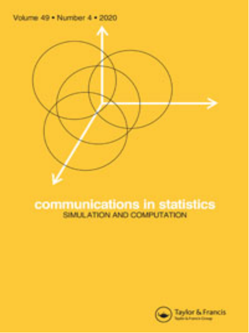
Volume 49 • Number 4 • 2020
communications in statistics
SIMULATION AND COMPUTATION
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: www.tandfonline.com/journals/lssp20

A simulation study for count data models under varying degrees of outliers and zeros

Fatih Tüzen, Semra Erbaş & Hülya Olmuş

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
2020, VOL. 49, NO. 4, 1078–1088
https://doi.org/10.1080/03610918.2018.1498886

Taylor & Francis
Taylor & Francis Group

A simulation study for count data models under varying degrees of outliers and zeros

Fatih Tüzen[a], Semra Erbaş[b], and Hülya Olmuş[b]

[a]TURKSTAT, Ankara, Turkey; [b]Department of Statistics, Faculty of Sciences, Gazi University, Ankara, Turkey

ABSTRACT

This study was aimed at examining the performance of count data models under various outliers and zero inflation situations with simulated data. Poisson, Negative Binomial, Zero-inflated Poisson, Zero-inflated Negative Binomial, Poisson Hurdle and Negative Binomial Hurdle models were considered to test how well each of the model fits the selected datasets having outliers and excess zeros. We found that Zero-inflated Negative Binomial and Negative Binomial Hurdle models were found to be more successful than other count data models. Also the results indicated that in some scenarios, the Negative Binomial model outperformed other models in the presence of outliers and/or excess zeros.

ARTICLE HISTORY
Received 5 October 2017
Accepted 5 July 2018

KEYWORDS
Count data; Hurdle models; Outliers; Zero-inflated models

1. Introduction

In many applications, count outcomes are quite common and often these count data have high proportion of zeros. They are not optimally modelled with a normal distribution. Because the assumptions of the ordinary least-squares regression are violated (homoscedasticity, normality, and linearity), the use of statistical techniques generally causes biased and inefficient results (Afifi et al. 2007). In these situations, the standard framework for explaining the relationship between the outcome variable and a set of explanatory variables includes the Poisson and Negative Binomial (NB) regression models. The drawback of basic Poisson regression model is that it assumes equality of the mean and variance. However, this is not a very practical assumption in real life. The Negative Binomial regression can be written as an extension of Poisson regression and it enables the model to have greater flexibility in modeling the relationship between the conditional variance and the conditional mean compared to the Poisson model (Cameron and Trivedi, 2013). Zero-inflated models and Hurdle models have been used when there exist excess zeros and overdispersion that the sample variance exceeds the sample mean.

Zero-inflated Poisson (ZIP) regression model is one of the zero-inflated regression models. The ZIP regression model was first introduced by Lambert (1992), and the author applied this model to the data collected from a quality control study, in which the response is the number of defective products in a sample unit. Also, Lambert (1992)

CONTACT Fatih Tüzen ✉ mfatihtuzen@gmail.com TURKSTAT, Necatibey 114, 06420 Çankaya, Ankara, Turkey.

found the ZIP model to be superior to the Negative Binomial, which was superior to the Poisson model.

Further applications for the ZIP regression model can be found in dental epidemiology (Bohning et al. 1999), occupational health (Lee, Wang and Yau 2001), psychology (Atkins and Gallop 2007), substance abuse (Xie et al. 2013) and computer science (Fagundes, Souza and Cysneiros 2016) and in different application fields (Tüzen and Erbaş 2017; Yang et al. 2017; Hassankiadeh et al. 2018; Wanjau et al. 2018). In practice, even after accounting for zero-inflation, the non-zero part of the count distribution is often over-dispersed. In this case, Greene (1994) described an extended version of the Negative Binomial model for excess zeros, the Zero-inflated Negative Binomial (ZINB), which may be more suitable than the ZIP. It has been established that the ZIP parameter estimates can be severely biased if the non-zero counts are over-dispersed in relation to the Poisson distribution. Slymen et al. (2006) found the ZIP and ZINB models to be equal. On the other hand, hurdle models first discussed by Mullahy (1986), are popular for modeling count data with many zeros. Welsh et al. (1996) found the hurdle and ZIP models to be equal while Pardoe and Durham (2003) found the ZINB model to be superior to both the Poisson and hurdle models. Both zero-inflated and hurdle models have been utilized in many studies (Rose et al. 2006, Buu et al. 2011; Hu, Pavlicova and Nunes 2011; Bandyopadhyay et al. 2011; Xie et al. 2013; Guo et al. 2016; Sakthivel and Rajitha 2017).

One striking characteristic of these articles is their differences in terms of the proportion of zeros and the distribution for the nonzeros. Some research (Bohning et al., 1999) analyzed data in which the percentage of zeros was as low as %21.6 while others (Zorn, 1996) used percentage as high as %95.8. Further, the nonzeros varied in terms of their distributions from highly positively skewed from normal to uniform. It is possible that different models yield different results depending on the proportion of zeros and the distribution for the nonzeros. This contradiction may resulted from not understanding the different underlying mechanism of zero-inflation, different degrees of zero-inflation and over-dispersion. It is necessary to undertake a comprehensive examination and comparison of these methods under different conditions to understand how to deal with data that include too many zeros. Thus, this study proposes to answer this question in terms of zero-inflation and outliers with a simulation study.

The aim of this paper is to compare the performance of count data models under various outliers and zero inflation situations. Poisson, Negative Binomial, Zero-inflated Poisson, Zero-inflated Negative Binomial, Poisson Hurdle (PH) and Negative Binomial Hurdle (NBH) are considered to test how well each of the model fits the selected data sets having outliers and excess zeros. The simulation design is explained in detail in chapter 3. Finally, the study focused on identifying models which can handle the impact of outliers and excess zeros in count data. Our study is aimed at examining the performance of count data models and how adequately did each model fit the data, base on AIC (Akaike Information Criterion) under varying degrees of outliers and zeros with simulated data.

2. Count data models

The basic count model is the Poisson (P) regression model which is based on the Poisson distribution with probability density function (pdf):

$$P(Y) = \frac{e^{-\mu}.\mu^{Y}}{Y!} \tag{1}$$

where $P(Y)$ is the probability that the number of events occurs during a time period, and μ is the parameter representing the expected value of Y; i.e., $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \mu$. The set of predictor variables X impacts the mean of the response variable Y via a link function such that $\mathrm{g}(\mu) = \ln(X\beta)$, and the inverse link function (mean function) is $\mu = \mathrm{e}^{X\beta}$ or $\ln(\mu) = X\beta$ where β is the model coefficient to be estimated from data. The restrictive condition that the mean must equal the variance is often violated by overdispersed. As a result of that Poisson model is generally considered inappropriate for count data, which is usually highly skewed and overdispersed (Cameron and Trivedi, 2013). The Negative Binomial (NB) regression distribution can be used for count data with overdispersion, i.e., when the sample variance exceeds the sample mean. The NB model addresses over-dispersion by containing a dispersion parameter (k) to accommodate unobserved heterogeneity in count data. The NB model used in this study has the following pdf:

$$P(Y) = \frac{\Gamma\left(Y + \frac{1}{k}\right)}{\Gamma(Y+1)\Gamma\left(\frac{1}{k}\right)} \left(\frac{1}{1+k\mu}\right)^{1/k} \left(\frac{k\mu}{1+k\mu}\right)^{Y} \tag{2}$$

The added dispersion parameter allows the variance to exceed the mean and thus cause the NB distribution to account for overdispersion. The theoretical distribution of NB is more skewed (with more values stacked at zero on the left and a longer tail on the right) than the Poisson distribution. When dispersion parameter becomes zero, the NB model is reduced to the Poisson model. For this reason, the Poisson model is a special type of NB model. In addition to that the NB model is an overdispersed Poisson model (Xie et al. 2013). In recent years, zero-inflated and hurdle models have gained popularity for modeling count data with excess zeros. According to Cameron and Trivedi (2013), zero-inflated and hurdle models can be viewed as finite mixture models with a degenerate distribution whose mass is condensing at zero. Excess zeros arise when the event of interest is not experienced by many of the subjects. For sparse events or event counts with low means, both the Poisson and NB models can virtually predict a high proportion of zeros, with NB being able to predict more zeros. However, when data have excess zeros neither the Poisson nor the NB models are adequate. Zero-inflated Poisson (ZIP) model has been considered by Lambert (1992) as a mixture of a zero point mass and a Poisson, while Heilborn (1989) similarly considers the Negative Binomial model case. The ZIP model assumes that data are generated from two different processes: one process generates only the zero count (structural zeros, e.g., non-smokers) from the always-zero group (non-risk group), and another process generates both the zero (sampling zeros, e.g., smokers who did not smoke on a particular day) and non-zero counts from the not-always-zero group (at-risk group). Hence, ZIP combines the two modeling processes; the logistic regression part predicts the presence or absence of the outcome owing to structural zeros, and the Poisson regression part predicts the extent of the count outcome, which may include zero counts (sampling zeros), too. Similar to the connection between ZIP and Poisson model, the zero-inflated Negative Binomial model (ZINB) is an extension of the NB model. Like ZIP, ZINB is a two-component mixture model, but it combines the logistic regression model (the same as ZIP) and the NB model (different from ZIP). The logistic regression model predicts the excessive or structural zero-only count, and the NB model predicts both positive counts and the sampling zeros defined by the Negative

Binomial distributions (Xie et al. 2013). Zero-inflated Poisson model and zero-inflated Negative Binomial model can be expressed as respectively:

$$P(Y) = \begin{cases} \omega + (1-\omega)e^{-\mu} & Y = 0 \\ (1-\omega)\dfrac{e^{-\mu}\mu^{Y}}{Y!} & Y \geq 1 \quad 0 \leq \omega \leq 1 \end{cases} \tag{3}$$

$$P(Y) = \begin{cases} \omega + (1-\omega)\left(\frac{1}{1+k\mu}\right)^{\frac{1}{k}} & if \;\; Y = 0 \\ (1-\omega)\dfrac{\Gamma\left(Y+\dfrac{1}{k}\right)}{\Gamma(Y+1)\Gamma\left(\dfrac{1}{k}\right)}\left(\dfrac{1}{1+k\mu}\right)^{\frac{1}{k}}\left(\dfrac{k\mu}{1+k\mu}\right)^{Y} & if \;\; Y \geq 1 \end{cases} \tag{4}$$

where ω denotes the probability of being an individual having zero count and μ denotes the mean of the underlying distribution. Eq. 3 shows that the marginal distribution of Y exhibits over-dispersion if $\omega<0$, and it reduces to the standard Poisson model when $\omega = 0$. The alternative is that Y has the ZINB distribution where $k \geq 0$ is a dispersion parameter that is assumed independent of covariates. The ZINB model reduces to the ZIP model in the limit $k \rightarrow 0$. Hurdle models, first discussed by Mullahy (1986), are popular for modeling count data with many zeros. The hurdle model uses a two-stage modeling process. The first stage determines whether event occurrence falls below or above the hurdle. Once the hurdle has been crossed, the second stage determines the number of subsequent event occurrences above the hurdle (Min and Agresti, 2005). The first stage is a binary response model like in the ZIP and ZINB model, and the second stage is usually a truncated-at-zero count model (either the truncated Poisson or Negative Binomial model) to clarify observations above the hurdle. Like zero-inflation models (ZIP and ZINB), the hurdle model is a two-component model. There are, however, crucial conceptual differences between them: (1) The hurdle model supposes that zeros are generated from one process, rather than two processes, as supposed in the ZIP and ZINB models. That is, all zeros are considered as sampling zeros. Thereby, the first part (logit model) in the ZIP and ZINB models is predicting excess zeros, whereas the first part of the hurdle model includes all zeros. (2) The second part of the ZIP and ZINB models may include zero counts, but the second part of the hurdle model has only positive counts. For this reason, the hurdle model is generally called truncated-at-zero model, or a two-part model (Coxe, West and Aiken 2009). Poisson Hurdle model and Negative Binomial Hurdle model can be expressed as:

$$P(Y) = \begin{cases} \omega & if \;\; Y = 0 \;\; ise \\ (1-\omega)\dfrac{e^{-\mu}\mu^{Y}}{(1-e^{-\mu}).\Gamma(Y+1)} & if \;\; Y \geq 1 \;\; ise \end{cases} \tag{5}$$

$$P(Y) = \begin{cases} \omega & if \;\; Y = 0 \;\; ise \\ (1-\omega)\dfrac{\Gamma\left(y+\frac{1}{k}\right)^{\omega}}{\left[1-\left(\frac{1}{1+k\mu}\right)^{\frac{1}{k}}\right]\Gamma(Y+1)\Gamma\left(\dfrac{1}{k}\right)}\left(\dfrac{1}{1+k\mu}\right)^{\frac{1}{k}}\left(\dfrac{k\mu}{1+k\mu}\right)^{Y} & if \;\; Y \geq 1 \;\; ise \end{cases} \tag{6}$$

where ω is the probability of a zero count, $1-\omega$ is the probability of overcoming the hurdle and also k is overdispersion parameter.

Table 1. Simulation design factors.

Factor A Degree of zero-inflation (%)	Factor B Degree of outlier ratio (%)	Factor C Degree of outlier magnitude
0.25	0	Low (20-39)
0.50	0.01	Medium (40-59)
0.75	0.05	High (60-79)
	0.10	

3. Simulation Design

Effect of outliers and excess zeros on count data were both studied by creating simulated data sets under different conditions in order to compare Poisson (P), Negative Binomial (NB), Zero-inflated Poisson (ZIP), Zero-inflated Negative Binomial (ZINB), Poisson Hurdle (PH) and Negative Binomial Hurdle (NBH) models. Three conditions with varying probability of zeros for the response variable were tested in the current study. In order to be able to evaluate count data models in a different way, the dependent variable was also designed according to whether it contains outliers or not. In this direction, 4 different outlier ratio conditions were designed for 0%, 1%, 5% and 10%. Low (20–39), medium (40–59) and high (60–79) levels were also included in the simulation as a factor in the outlier presence (when the outlier ratio is 1%, 5% or 10%). For example, low outlier values are randomly selected according to the desired outlier ratio (1%, 5% and 10%) from the numbers 20–39. The response variable was generated from Negative Binomial distribution and these generated numbers are transformed with uniform distribution to create zeros in the desired range. The simulation study was 3*3*3 factorial design that was examined for situations involving outliers. In addition, 3 scenarios that involves only three zero-inflation (25%, 50% and 75%) conditions were added for cases without outliers (outlier ratio is 0%), and 30 scenarios were analyzed in total. The factors specified in the simulation design are shown in Table 1.

To provide a reasonable prediction model to explore in this study, two different kinds of covariates, X_1 and X_2, were simulated. Both of the covariates were assumed to be categorical variables. X_1 is a binary variable whose values are 1 and 2 with $P(X_1 = 1) = P(X_1 = 2) = 0.5$. X_2 has three categories and values are 1, 2, and 3 with $P(X_2 = 1) = 0.40$, $P(X_2 = 2) = 0.35$ and $P(X_2 = 3) = 0.25$. Regression coefficients β_0, β_1 and β_2 were set to be 1, 0.5 and -0.7 respectively. To ensure accurate results, 1000 replications (simulation size), each with sample size $n = 500$ were generated. The decisions on the number of simulations and sample size were made by referring to previous simulation studies on zero-inflated data (Lambert, 1992; Min and Agresti, 2005; Williamson et al. 2007).

The distribution of the numbers produced for the dependent variable was examined during the determination of the low, medium and high values of the outliers. The range of numbers generated for the dependent variable is from 0 to 18. From this point of view, the lowest outlier value was determined as 20 and the highest outlier value is set to 79. Thus, the values in the range of 20–79 are divided into 3 parts and the levels of the outliers are set as low (20–39), medium (40–59) and high (60–79). The outliers from the relevant levels were randomly selected for designs made for each level. For better understanding of the simulation design, the algorithm steps related to number

Table 2. AIC significance levels.

Difference between models A and B	Result if A < B
> 0.0 and ≤ 2.5	No difference in models
> 2.5 and ≤ 6.0	Prefer A if $n > 256$
> 6.0 and ≤ 9.0	Prefer A if $n > 64$
> 9.0	Prefer A

generation of scenario that have 25% zero-inflation, 1% outlier ratio and low outlier magnitude is described in the following example:

Step 1: $n = 500$ is set to sample size.
Step 2: The zero-inflation is defined as %25.
Step 3: The outlier ratio is defined as %1.
Step 4: The integers between 20 and 39 are defined as the set of outliers.
Step 5: The number of observations without outliers is determined by the formula $n1 = n$ – (outlier ratio × n). That is, $n1 = 500 – (0.01 \times 500) = 495$ is obtained.
Step 6: $n1 = 495$ numbers are generated from Negative Binomial distribution and Uniform
distribution to obtain count data at the desired zero-inflation ratio (%25).
Step 7: The remaining 5 observations are randomly selected from the set of outliers specified in step 4.
Step 8: By combining the observations obtained in steps 6 and 7, the number generation for the dependent variable (Y) is completed with $n = 500$ numbers.
Step 9: According to the assumptions expressed above, $n = 500$ observations are randomly
generated for two categorical independent variables.
Step 10: AIC values of the models are obtained with generated dependent and independent variables by setting the number of simulations to 1000.

AIC was used to evaluate the goodness of fit of the six models and is defined as follows:

$$\text{AIC} = -2\text{logL} + 2p \tag{7}$$

where logL is the maximum of the likelihood function for a fitted model and p is the number of parameters in the fitted model. The preferred model is the one with the minimum AIC value (Burnham and Anderson, 2004). Hilbe's AIC rule-of-thumb criterion (Hilbe, 2011) was adopted for this study to determine whether there was a statistically significant difference between two values of the AIC. Table 2 shows the significance levels for AIC based on the number of observations. In the case of our study ($n = 500$), if the difference in the AIC value was more than 2.5, then the model with the lower AIC was significantly different from another. The statistical software R (R Development Core Team, 2016) was used for data simulation.

4. Results

The results of AIC values of the 3 scenarios that involve three zero-inflation cases without outliers are presented in Table 3.

Table 3. AIC values for count data models (outlier ratio = % 0).

No	Zeroinflation (%)	P	NB	ZIP	ZINB	PH	NBH
1	25	2891.3482	2442.1799	2324.4086	**2297.4314**	2324.4088	**2297.4665**
2	50	2936.9700	2012.2259	1859.9364	**1842.6384**	1859.9369	**1842.6390**
3	75	2328.6235	1328.1406	1142.4274	**1135.2064**	1142.4287	**1135.2080**

Table 4. AIC values for count data models (outlier ratio = % 1).

No	Zero inflation (%)	Magnitude of outliers	P	NB	ZIP	ZINB	PH	NBH
4	25	Low	3238.2513	2504.2091	2596.0255	**2412.8440**	2596.0252	**2412.8435**
5		Medium	3682.8489	2546.8857	2980.1896	**2492.7073**	2980.1891	**2492.7062**
6		High	4198.7024	2583.3385	3435.4105	**2551.5758**	3435.4097	**2551.5742**
7	50	Low	3374.7194	2053.8314	2126.3526	**1948.2593**	2126.3533	**1948.2604**
8		Medium	3889.453	2079.1227	2499.7577	**2012.9099**	2499.7589	**2012.9121**
9		High	4470.5068	2100.1965	2939.2800	**2057.7328**	2939.2817	**2057.7360**
10	75	Low	2917.2044	1278.2065	1395.5711	**1223.2060**	1395.5728	**1223.2088**
11		Medium	3541.5095	1295.2751	1740.5473	**1265.2837**	1740.5503	**1265.2897**
12		High	4220.6888	1308.4230	2140.2415	**1291.6211**	2140.2454	**1291.6303**

ZINB and NBH models were found to be more successful than other models under different degree of zero-inflation and at 0% ratio of outliers. As the results showed that there is no significant difference between these two models since the difference between the AIC values of these two models is less than 2.5. Also, we can say that the higher the zero-inflation, the lower the AIC values. Tables 4–6 shows the AIC results of scenarios for outlier ratio at 1%, 5% and 10% respectively.

It can be seen clearly at 1% ratio of outliers, ZINB and NBH models outperform the other models having the smallest AIC values and are considered the best. For all scenarios, there is no significant difference between these two models at 1% ratio of outliers. But Tables 5 and 6 have shown that the results for 5% and 10% ratio of outliers are different. In scenario 15 and 23 the NB model outperforms other models. It is a striking result that we have obtained from this study. Besides, the NB model have similar results in scenario 14, 18, 21, 24 and 27. According to these findings we can say that, as the ratio and magnitude of the outliers increase, the NB model performs the same or better than the ZINB and NBH models. The predominance of ZINB and NBH models is particularly prevalent in the case of excess zeros. But it is seen that the NB model achieved as successful results as ZINB and NBH, especially when the magnitude of outliers are high even at every zero-inflation level.

Overall, Negative binomial models were found to be successful than Poisson models in situations where they were outliers. Because they have lower AIC values at different level of zero-inflation and at different outlier ratios and magnitudes. As the outlier ratio and magnitude increase, model performances decrease; conversely, as zero-inflation increase, model performances increase.

5. Discussion

In this paper, we discussed the problem of outliers and excess zeros together for count data models. Effect of outliers and excess zeros on count data were both studied by creating simulated data sets under different conditions in order to compare P, NB, ZIP, ZINB, PH and NBH models.

Table 5. AIC values for count data models (outlier ratio = % 5).

No	Zero inflation (%)	Magnitude of outliers	P	NB	ZIP	ZINB	PH	NBH
13	25	Low	4475.6312	2688.2399	3548.9859	**2659.7296**	3548.9866	**2659.7392**
14		Medium	6425.1712	**2809.3281**	5195.8584	**2808.7546**	5195.8592	**2808.9845**
15		High	8619.6295	**2894.9338**	7087.1590	2898.7320	7087.1596	2897.7685
16	50	Low	4931.6511	2197.9986	3004.8723	**2154.4708**	3004.8728	**2154.4747**
17		Medium	7136.1857	2275.6682	4503.6171	**2266.2515**	4503.6175	**2266.2660**
18		High	9541.0029	**2329.7374**	6202.4103	**2330.5963**	6202.4106	**2330.8088**
19	75	Low	4959.3284	1397.4790	2077.0441	**1369.4474**	2077.0444	**1369.4530**
20		Medium	7510.3772	1448.7393	3232.2021	**1440.6735**	3232.2022	**1440.6895**
21		High	10184.545	**1482.5453**	4510.4433	**1481.6141**	4510.4433	**1481.6980**

Table 6. AIC values for count data models (outlier ratio = % 10).

No	Zero inflation (%)	Magnitude of outliers	P	NB	ZIP	ZINB	PH	NBH
22	25	Low	5782.5495	2860.0187	4486.6370	**2846.7641**	4486.6370	**2846.7751**
23		Medium	9152.3544	**3025.9371**	7251.7925	3029.3055	7251.7925	3029.2446
24		High	12860.248	**3135.1424**	10354.872	3140.0524	10354.872	**3133.8123**
25	50	Low	6553.3420	2339.3033	3756.5482	**2305.6629**	3756.5482	**2305.6654**
26		Medium	10303.673	2452.1061	6095.0736	**2446.9090**	6095.0736	**2446.9201**
27		High	14302.225	**2525.9487**	8681.7909	**2528.0222**	8681.7909	**2528.3377**
28	75	Low	7000.0879	2522.0242	2415.9125	**1478.7387**	2415.9125	**1478.7405**
29		Medium	11223.022	2762.6059	3857.3859	**1578.6087**	3857.3859	**1578.5991**
30		High	15562.831	3492.6322	5415.7288	**1638.4398**	5415.7288	**1638.3108**

In literature, there has been a few research on outliers, which is an important issue in analyzing these types of models. Yang, Xie and Goh (2009) studied some outlier identification methods and their applications. Two robust parameter estimates based on the trimmed mean and the Winsorized mean are proposed to eliminate the effect of outliers. Simulation results showed the robustness of proposed parameter estimates. Usman and Oyejola (2013) studied the impact of outliers and excess zeros on Poisson, Negative Binomial, ZIP and ZINB regression models by creating simulated data set for 20, 50 and 100 sample sizes. Outliers were introduced into the generated data adding 5 to 5%, 10% and 15%. According to the results, ZINB outperformed other models.

According to obtained simulation results, it has been seen that different results are obtained in scenarios containing outliers. It has been found that as the zero-inflation increases, the AIC values decrease and as outlier ratio and magnitude increase, AIC increase. It is understood that NB models are more successful than Poisson models in both outlier and non-outlier situations. Especially when the outlier ratio and magnitude are increased in the scenario where the zero-inflation is 50% or less, it is seen that the NB model produces more successful results than the two superior models such as ZINB and NBH. In this type of scenarios, we can say that simpler models such as NB can be used instead of using complex models like ZINB and NBH. The dominant sides of the ZINB and NBH models appear more prominently when the zero-inflation goes up to 50% and above.

Results from this study support using special zero inflated models for zero-inflated data. Both zero-inflated and hurdle models have the ability to identify the factors that have significant effects on the probability that the participant is from the nonsusceptible group by means of a binary regression model. On the other hand, these models provide important information on the magnitude of the counts given that the participant is from the susceptible group by means of a Poisson regression or Negative Binomial regression. Factors or explanatory variables do not need to be the same for the Binomial model and the count model. Although the NB model can also effectively offer accurate estimation under some degrees of zero-inflation and outliers, it cannot provide information about possible mechanisms underlying the zero-inflation. Statistically, zero inflated models provide more accurate estimates.

Our fitted models brought out that zero-inflated and hurdle models were identical with AIC results with simulated data. However, choosing between the zero-inflated and hurdle models, assuming the Poisson and NB were insufficient because of excess zeros, should generally be based on study design and objectives. The biggest difference between them is that zero-inflated models distinguish between structural zeros and sampling zeros, hurdle models do not. In many studies, zero-inflated models may be conceptualized as allowing zeros to arise from at-risk and not-at-risk populations. In contrast, we may conceptualize hurdle models as having zeros only from an at-risk population (Rose et al., 2006). It is more appropriate to use zero-inflated models in these kind of situations when the study design has a greater chance of having sampling zeros. On the other hand, however, zero-inflated and hurdle modeling framework should be equivalent when the primary purpose of the study is to make predictions as both of them tend to yield similar model fit. There is a need for more different scenarios to study the performance of these models. For example the study can be extended further on very large sample to investigate the performance of these models in the presence of outliers and/or excess zeros. We think that this study is enlightening for researchers on how proportions of outliers and excess zeros affect the count data models. This study will assist researchers to understand what proportion of outliers or excess zeros may cause serious effects to the these count data models. This may give them a kind of an overview of their study.

References

Afifi, A. A., J. B. Kotlerman, S. L. Ettner, and M. Cowan. 2007. Methods for improving regression analysis for skewed continuous or counted responses. *Annual Review of Public Health* 28:95–111.

Atkins, D., and R. Gallop. 2007. Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology* 21 (4):726–35.

Bandyopadhyay, D., S. DeSantis, J. Korte, and K. Brady. 2011. Some considerations for excess zeroes in substance abuse research. *American Journal of Drug and Alcohol Abuse* 37 (5):376–82.

Bohning, D., E. Dietz, P. Schlattmann, L. Mendonca, and U. Kirchner. 1999. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society (Series A, Statistics in Society)*, 162 (2):195–209.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* 33:261–304.

Buu, A., N. Johnson, R. Li, and X. Tan. 2011. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine* 30:2326–40.

Cameron, A. C., and P. K. Trivedi. 2013. *Regression Analysis of Count Data*, 2nd ed. NewYork: Cambridge University Press.

Coxe, S., S. G. West, and L. S. Aiken. 2009. The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment* 91:121–36.

Fagundes, R. A. A., R. M. C. R. Souza, and F. J. A. Cysneiros. 2016. Zero-inflated prediction model in software-fault data. *The Institution of Engineering and Technology*, 10 (1):1–9.

Greene, W. H. 1994. Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models (March). NYU Working Paper No. EC-94-10. Available at: https://ssrn.com/abstract =1293115.

Guo, F., G. Wang, J. L. Innes, Z. Ma, and A. Liu. 2016. Comparison of six generalized linear models for occurrence of lightning-induced fires in northern Daxing'an Mountains, China. *Journal of Forestry Research* 27 (2):379–88.

Hassankiadeh, R. F., A. Kazemnejad, M. G. Fesharaki, and S. K. Jahromi. 2018. Efficiency of zero-inflated generalized poisson regression model on hospital length of stay using real data and simulation study. *Caspian Journal of Health Research* 3 (1):5–9.

Heilborn, D. 1989. Generalized Linear Models for Altered Zero Probability in Count Data, Technical Report, Department of Epidemiology and Biostatistics, University of California, San Francisco.

Hilbe, J. M. 2011. *Negative Binomial Regression*. NewYork: Cambridge University Press.

Hu, M., M. Pavlicova, and E. Nunes. 2011. Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. *American Journal of Drug and Alcohol Abuse* 37 (5):367–75.

Lambert, D. 1992. Zero-inflated Poisson regression with an application to detects in manufacturing. *Technometrics* 34:1–14.

Lee, A. H., K. Wang, and K. K. W. Yau. 2001. Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical Journal* 43:963–75.

Min, Y., and A. Agresti. 2005. Random effect models for repeated measures of zero-inflated count data. *Statistical Modeling* 5:1–19.

Mullahy, J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics* 33:341–65.

Pardoe, I., and C. A. Durham. 2003. Model choice applied to consumer preferences. In Proceedings of the 2003 Joint Statistical Meetings, Alexandria, VA, American Statistical Association.

R Development Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R foundation for Statistical Computing. Available at: http://www.R-project.org/

Rose, C., S. Martin, K. Wannemuehler, and B. Plikaytis. 2006. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics* 16 (4):463–81.

Sakthivel, K. M., and C. S. Rajitha. 2017. A comparative study of zero-inflated, hurdle odels with artificial neural network in claim count modeling. *International Journal of Statistics and Systems* 12 (2):265–76.

Slymen, D. J., G. X. Ayala, E. M. Arredondo, and J. P. Elder. 2006. A demonstration of modeling count data with an application to physical activity. *Epidemiologic Perspectives & Innovations* 3:1–9.

Tüzen, M. F., and S. Erbaş. 2017. A comparison of count data models with an application to daily cigarette consumption of young persons. *Communications in Statistics – Theory and Methods* 47 (23):5825–44. doi:10.1080/03610926.2017.140205010.

Usman, M., and B. A. Oyejola. 2013. Models for Count Data in the Presence of Outliers and/or Excess Zero. *Mathematical Theory and Modeling* 3 (7):94–103.

Wanjau, A. N., S. M. Mwalili, and O. Ngesa. 2018. Assessment and selection of competing models for count data: an application to early childhood caries. *International Journal of Data Science and Analysis* 4 (1):24–31.

Welsh, A. H., R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmayer. 1996. Modelling abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling* 88:297–308.

Williamson, J. M., H. Lin, R. H. Lyles, and A. W. Hightower. 2007. Power calculations for ZIP and ZINB models. *Journal of Data Science* 5 (4):519–34.

Xie, H., J. Tao, G. J. McHugo, and R. E. Drake. 2013. Comparing statistical methods for analyzing skewed longitudinal count data with many zeros: An example of smoking cessation. *Journal of Substance Abuse Treatment* 45:99–108.

Yang, J., M. Xie, and T. N. Goh. 2009. Outlier identification and robust parameter estimation in a zero-inflated Poisson model. *Journal of Applied Statistics* 38 (2):421–30.

Yang, S., G. Puggioni, L. L. Harlow, C. A. Redding. 2017. A comparison of different methods of zero-inflated data analysis and an application in health surveys. *Journal of Modern Applied Statistical Methods* 16 (1):518–43.

Zorn, C. J. W. 1996. Evaluating zero-inflated and hurdle Poisson specifications. Midwest Political Science Association, 1–16.