ISSN No:-2456-2165

Are Zero Inflated Distributions Compulsory in the Presence of Zero Inflation?

B.P. Tlhaloganyang and K. Thaga
Department of Statistics
University of Botswana
Private Bag UB00705
Gaborone
Botswana

Abstract:- If overdispersion occur as a result of excessive zero counts "i.e, zero inflation", Zero Inflated Poisson/Negative Binomial distributions are preferred over the standard Negative Binomial distribution as they have a parameter that handles excessive zeros. Without doubting their outstanding performance in modeling overdispersed and zero inflated datasets, there are concerns as to whether zero inflated distributions should always substitute standard distributions in these types of datasets. For this reason, this paper intended to use different real datasets to show that zero inflated models are not always necessary even if the data is characterized by overdispersion and zero inflation. This was achieved through comparing Negative Binomial distribution with Zero Inflated Poisson/Negative Binomial distributions in datasets that went through the test of overdispersion and zero inflation. With respect to goodness of fit of these distributions, zero inflated distributions scored higher AIC scores in all datasets when compared to Negative Binomial distribution. Negative Binomial was marked as the outstanding distribution in all datasets suggesting that zero inflated models are not always necessary in datasets caharacterized by overdispersion and zero inflation.

I. INTRODUCTION

Poisson distribution remains the classic and natural distribution for discrete data when the data is equidispersed, that is, when the variance is equal to the mean. However, when the count outcome is overdispersed, practitioners routinely recommends the use of Negative Binomial distribution as it allows for extra variability that cannot be accommodated by Poisson distribution, Ismail and Zamani (2013). Overdispersion occurs when the conditional variance is greater than the conditional mean.

In real life, many count datasets are found to be highly skewed to the right, Gupta et al. (2013); Javali et al. (2010); Khan et al. (2011). Datasets that are highly skewed to the right usually consists a lot of observed zero counts. These datasets are usually termed as Zero Inflated datasets as they contain excessive zero counts. Excessive zero counts increases the sample size while the total sum of counts remain unchanged. Thus, excessive zero counts can lead to smaller conditional mean in relation to the conditional variance.

Often, if a count outcome variable is overdispersed and zero inflated, Negative-Binomial distribution usually underestimate its observed zero count frequency. In research, use of inappropriate statistical distributions can result in making incorrect conclusions that may bring uncertainty in research. Thus, use of inappropriate statistical distributions can yield biased standard errors of the regression coefficients will obviously produce wrong statistical tests that can overstate the significance of the predictors.

For this reason, several studies prefer the use of Zero Inflated distributions to model overdispersed and zero inflated count datasets, see (Perumean-Chaney et al. (2013); Ismail and Zamani (2013); Sellers and Raim (2016); Gupta et al. (2013)). From statistical point of view, zero inflated distributions are often appropriate for count datasets that are characterized by both overdispersion and excessive zero counts, (Lambert (1992) ; Edwin (2014); Perumean-Chaney et al. (2013)). Among available range of zero inflated distributions seen from Saengthong et al. (2015), Yamrubboon et al. (2017) and Sim et al. (2018), the commonly used distributions are Zero Inflated Poisson (ZIP) and Zero Inflated, Negative-Binomial (ZINB), see (Chipeta et al. (2014); Edwin (2014); Yusuf et al. (2017) ; Gupta et al. (2013)).

ZIP and ZINB distributions have an advantage over standard distributions as they have both the distributional and the zero inflation parameters. This made ZIP and ZINB distributions to become fairly popular in literature as most of real life datasets are usually characterized by overdispersion and excessive zero counts. Without doubt, these distributions often provide a reasonable fit when applied to positively skewed datasets when compared to standard distributions.

Consequently, standard distributions are usually neglected when it comes to modelling overdispersed and zero inflated datasets, Yusuf et al. (2017) and Muniswamy et al. (2015).However, care should be taken when deciding which distribution to apply as different datasets adhere to different distributions. Numerous studies proposed decision flow charts that can be used when deciding on a count data distribution to apply in the presence of overdispersion and zero inflation, see Perumean-Chaney et al. (2013), Elhai et al. (2008) and Walters (2007). They recommended that if overdispersion and zero inflation are noted, the data should be modelled using ZINB distribution.

Nevertheless, concerns have been raised as to whether zero inflated distributions should always substitute standard distributions in datasets that are characterized by overdispersion and excessive zero counts, Allison (2012). Xie et al. (2013) analyzed longitudinal count data using zero inflated models and standard models. In their results, standard regression models were marked as the best fitting models over zero inflated regression models. The results were the same as those of Schmidt et al. (2008) who studied time series count data from a Bayesian point of view. From these studies, we are not sure if the datasets were actually zero inflated as larger frequency of zeros does not always correspond to zero inflation.

As a result, this paper concentrated on using different real datasets from different fields to show that zero inflated models are not always necessary even if the data is characterized by both overdispersion and zero inflation. The datasets initially went through statistical tests to confirm overdispersion and zero inflation rather than relying on descriptive statistics and observed zero frequencies. Moreover, the performance of Negative Binomial distribution against ZIP and ZINB distributions was assessed based on Akaike Information Criteria (AIC) goodness of fit measure.

The article is outlined as follows; Section 2 gives the distributions that were considered in the study being Poisson, Negative Binomial, ZIP and ZINB distributions. This section ends with provision of methods used to test for overdispersion and zero inflation together with a goodness of fit measure. Section 3 gives datasets that were used to illustrate that zero inflated models are not always necessary even if the data is characterized by both overdispersion and zero inflation. Section 4 provides conclusions and recommendations guided by the study findings.

II. METHODS AND MATERIALS OVERVIEW

- *Poisson*

Poisson distribution is a well known count data model with a probability mass function (pmf) expressed as

$$P(Z=z)=\frac{e^{-\theta}\theta^{z}}{z!},\quad z\in\ \{0\}\cup\mathbb{Z}^{+} \qquad (2.1)$$

in which $\theta\in(0,1)$ and $\theta=E(Z)=Var(Z)$. Poisson distribution is suitable only when equidispersion, thus, when $E(Z)=Var(Z)$.

- *Negative Binomial distribution*

Negative Binomial distribution is usually considered when Poisson distribution cannot fit well as a result of overdispersion, Cameron and Trivedi (1986). It is suitable in presence of overdispersion as it has an additional parameter, *r*, that allows for extra variability. Its pmf is expressed as

$$P(Z=z)=\binom{r+z-1}{z}\theta^{z}(1-\theta)^{r},\quad z\in\ \{0\}\cup\mathbb{Z}^{+}$$
(2.2)

where $r\geq 0$ and $\theta\in(0,1)$. Its mean and variance are given as $E(Z)=\frac{r\theta}{1-\theta}$ and $Var(Z)=\frac{r\theta}{(1-\theta)^2}$.

- *Zero Inflated distributions*

Zero Inflated distributions are extensions of the assumed standard distributions with an additional parameter, δ, that handles excessive zero counts underpredicted by the assumed standard distributions. This additional parameter is termed as the zero inflation parameter. The probability function of this two component distribution concentrated at zero is expressed as

$$P(Z=z)=\begin{cases}\delta+(1-\delta)P(K=0), & z=0\\ (1-\delta)P(K=z), & z=1,2,3,\ldots\end{cases} \qquad (2.3)$$

where $0\leq\delta<1$. The scenario $\delta=0$ implies that there is no inflation at zero while $\delta>0$corresponds to zero inflation.A random variable K has either a Poisson or Negative Binomial distribution for our case, Edwin (2014); Gupta et al. (1995); Lambert (1992). From equation (2.3), when $P(K=z)$has a Poisson distribution given in equation (2.1),equation (2.3) is termed as Zero Inflated Poisson (ZIP) distribution with mean and variance stated as as $E(Z)=\theta(1-\delta)$ and $Var(Z)=\theta(1-\delta)(1+\delta\theta)$. Moreover, when $P(K=z)$ has a Negative Binomial distribution given in equation (2.2), equation (2.3) is now termed as Zero Inflated Negative Binomial (ZINB) distribution characterized by the mean and the variance defined as

$$E(Z)=\frac{r\theta(1-\delta)}{1-\theta}\quad and\quad Var(Z)=\frac{r\theta(1-\delta)(1+r\delta\theta)}{(1-\theta)^2}$$

- *Overdispersion test*

For all the datasets, it was checked if the random variable Z is characterized by equidispersion or not. As stated before, if the random variable Z is equidispersed, then it follows Poisson distribution with mean and variance stated as $E(Z)=\theta$and $Var(Z)=\theta$respectively. If overdispersion is present, the random variable Z follows mixed Poisson distribution with mean and variance$E(Z)=\theta$ and $Var(Z)=\theta+\rho\theta$respectively, Dean and Lawless (1989). An extra parameter in mixed Poisson, ρ, measures the level of extra variation. As a result, it was investigated if $\rho=0$ or $\rho>0$ using hypotheses $H_0;\rho=0$ vs $H_1;\rho>0$. Based on the test statistic developed by Cameron and Trivedi (1986), the above hypotheses can be evaluated using AER-package in R. Rejection ofH_0corresponds to the presence of overdispersion.

- *Zero inflation test*

For a count dataset, it was examined if the zero inflation parameter, δ, is equal to zero or not. This was done using the hypotheses $H_0;\delta=0$ vs $H_1;\delta>0$where rejection of H_0 corresponds to the presence of zero inflation. Van den Broek (1995) developed a simple test statistic distributed as chi-square with 1 degrees of freedom under the null hypothesis that can be used to test for zero inflation presence. The test statistic is expressed as

$$T = \frac{(n_0 - n\bar{P}_0)^2}{n\bar{P}_0(1-\bar{P}_0) - n\bar{z}\bar{P}_0^2}$$

where$\bar{P}_0$ = exp(−$\bar{z}$), n :Total number of counts and n_0 : Number of zeros in the data.

- *Best Model selection criteria*

Best model selection was based on Akaike Information Criteria (AIC) which is a goodness of fit measure that marks the best model as the one with lowest AIC score among competing models, Cameron and Trivedi (2013). To define AIC of a given model, let log$L$be the logarithm of the maximized likelihood value, then AIC score is expressed as

$$AIC = -2\, logL + 2K$$

Where $K$is the number of parameters estimated in a given model.

III. RESULTS AND FINDINGS

In this section, Negative Binomial, ZIP and ZINB distributions are considered to see how they fit two real datasets that are overdispersed and zero inflated. Table 1 gives the two datasets that were used for illustration purpose. Dataset 1 shown in Table 1 was obtained from Muniswamy et al. (2015) who was studying the frequency of deaths of women aged 80 years and above appearing in the “London Times”. Muniswamy et al. (2015) analyzed this dataset using ZIP and ZINB distributions without considering standard distributions and found that ZINB distribution provided a better fit when compared to ZIP distribution.

Table 1: Observed frequencies for Dataset 1 & 2

Count	0	1	2	3	4	5+
Dataset 1 Frequencies	162	267	271	185	111	100
Dataset 2 Frequencies	3719	232	38	7	3	1

Moreover, Table 1 also contains Dataset 2 which gives Zaire insurance company policy claim counts data obtained from Zhang et al. (2018). These claim counts were analyzed by Zhang et al. (2018) using COM-type of distributions and found that COM-Negative Binomial distribution provided a reasonable fit. Table 2 gives the calculated p-values for overdispersion and zero inflation tests for each dataset contained in Table 1 together with decision guided by the calculated p-values evaluated at 5% level of significance.

Table 2: Overdispersion and Zero inflation p-values and decision for Dataset 1 & 2

	Dispersion test	Inflation test	Dataset characteristic
Dataset 1	1.4E-05	8.7E-06	Overdispersed& Zero inflated
Dataset 2	5.2E-06	1.2E-44	Overdispersed& Zero inflated

With respect to overdispersion and zero inflation tests in Table 2 & 2, p-values calculated generated statistically significant results that suggests overdispersion and zero inflation presence in both datasets. Table 3 gives AIC scores of distributions under investigation for each dataset contained in Table 1.

Table 3: AIC scores for competing distributions inDataset 1 & 2

	Poisson	Negative-Binomial	ZIP	ZINB
Dataset 1	4004.8	3985.7	3992.1	3986.8
Dataset 2	2494.2	2371.1	2379.6	2373.1

AIC scores are examined to identify a model that provides a reasonable fit within each dataset in the presence of overdispersion and zero inflation. Table 3 shows decreasing AIC scores from Poisson distribution and ZIP distribution to ZINB distribution and Negative Binomial distribution. On basis of lower is better criteria of AIC goodness of fit measure, Negative Binomial distribution stands as the best model.

IV. CONCLUSIONS

Zero inflated distributions are known to be good in modeling datasets with ultrahigh zero counts when compared to standard distributions. As a result, standard distributions receives less attention in datasets that are overdispersed and zero inflated as it is believed they can provide a poor fit. For this reason, this paper used real life datasets to show that zero inflated models are not always necessary even if the data is characterized by both overdispersion and zero inflation. Before model comparisons, datasets initially went through statistical test of overdispersion for confirmation of overdispersion and zero inflation. With respect to model performances, Negative Binomial provided a reasonable fit in all datasets when compared to ZIP and ZINB distributions in overdispersed and zero inflated datasets. To conclude, we recommend that in the analysis of count datasets characterized by overdispersion and zero inflation, one should fit both zero inflated and standard distributions.

REFERENCES

[1]. Allison, P. (2012). Do we really need zero-inflated models? https:// statisticalhorizons.com/zero-inflated-models. Blog post accessed on 12April-2019.

[2]. Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data. comparisons and applications of some estimators and tests. *Journal of applied econometrics*, 1(1):29–53.

[3]. Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

[4]. Chipeta, M. G., Ngwira, B. M., Simoonga, C., and Kazembe, L. N. (2014). Zero adjusted models with applications to analysinghelminths count data. *BMC research notes*, 7(1):856.

[5]. Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in poisson regression models. *Journal of the American Statistical Association*, 84(406):467 – 472.

[6]. Edwin, T. (2014). *Power series distributions and zero-inflated models*. PhD thesis,

[7]. Doctoral dissertation, University of Nairobi.

[8]. Elhai, J. D., Calhoun, P. S., and Ford, J. D. (2008). Statistical procedures for analyzing mental health services data. *Psychiatry research*, 160(2):129–136.

[9]. Gupta, P., Gupta, R., and Tripathi, R. (1995). Inflated modified power series distributions with applications. *Communications in Statistics-Theory and Methods*, 24(9):2355–2374.

[10]. Gupta, R., Marino, B. S., Cnota, J. F., and Ittenbach, R. F. (2013). Finding

[11]. the right distribution for highly skewed zero-inflated clinical data. *Epidemiology, Biostatistics and Public Health*, 10(1).

[12]. Ismail, N. and Zamani, H. (2013). Estimation of claim count data using nega-

[13]. tive binomial, generalized poisson, zero-inflated negative binomial and zero-inflated generalized poisson regression models. In *Casualty Actuarial Society E-Forum*, volume 41, pages 1–18.

[14]. Javali, S. B., Pandit, P. V., et al. (2010). Using zero inflated models to analyze dental caries with many zeroes. *Indian Journal of Dental Research*, 21(4):480.

[15]. Khan, A., Ullah, S., and Nitz, J. (2011). Statistical modelling of falls count data with excess zeros. *Injury prevention*, 17(4):266–270.

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

[17]. Muniswamy, B., Molla, D. T., and Reddy, N. K. (2015). Comparison of test

[18]. statistic for zero-inflated negative binomial against zero-inflated poisson model. *Indian Journal of Science and Technology*, 8(4):349.

[19]. Perumean-Chaney, S. E., Morgan, C., McDowall, D., and Aban, I. (2013). Zeroinflated and overdispersed: what's one to do? *Journal of Statistical Computation and Simulation*, 83(9):1671–1683.

[20]. Saengthong, P., Bodhisuwan, W., and Thongteeraparp, A. (2015). The zero inflated negative binomial–crack distribution: some properties and parameter estimation. *Songklanakarin J. Sci. Technol*, 37(6):701–711.

[21]. Schmidt, A., Pereira, B., and Vieira, P. (2008). Do we always need a zero-inflated model to capture an apparent excess of zeros? http://www.dme.im.ufrj.br/ arquivos/publicacoes/arquivo211.pdf. [Online; accessed 28- April -2019].

[22]. Sellers, K. and Raim, A. (2016). A flexible zero-inflated model to address data dispersion. *Computational Statistics & Data Analysis*, 99:68–80.

[23]. Sim, S. Z., Gupta, R. C., and Ong, S. H. (2018). Zero-inflated conway-maxwellpoisson distribution to analyze discrete data. *The international journal of bio-*

[24]. *statistics*.

[25]. Van den Broek, J. (1995). A score test for zero inflation in a poisson distribution.

[26]. *Biometrics*, pages 738–743.

[27]. Walters, G. D. (2007). Using poisson class regression to analyze count data in correctional and forensic psychology: A relatively old solution to a relatively new problem. *Criminal Justice and Behavior*, 34(12):1659–1674.

[28]. Xie, H., Tao, J., McHugo, G. J., and Drake, R. E. (2013). Comparing statistical methods for analyzing skewed longitudinal count data with many zeros: An example of smoking cessation. *Journal of substance abuse treatment*, 45(1):99–108.

[29]. Yamrubboon, D., Thongteeraparp, A., Bodhisuwan, W., and Jampachaisri, K.

[30]. (2017). Zero inflated negative binomial-sushila distribution and its application. In *AIP Conference Proceedings*, volume 1905, page 050044. AIP Publishing.

[31]. Yusuf, O., Bello, T., and Gureje, O. (2017). Zero Inflated Poisson and Zero Inflated Negative Binomial Models with Application to Number of Falls in the Elderly. http://juniperpublishers.com/bboaj/pdf/BBOAJ.MS.ID. 555566.pdf. [Online; accessed 15- April -2019].

[32]. Zhang, H., Tan, K., and Li, B. (2018). Com-negative binomial distribution: modeling overdispersion and ultrahigh zero-inflated count data. *Frontiers of Mathematics in China*, 13(4):967–998.