Research Journal of Mathematical and Statistical Sciences _____ISSN 2320-6047
Vol. **8(1),** 31-34, January **(2020)** *Res. J. Mathematical and Statistical Sci.*

Short Communication

Zero inflated Poisson distribution in equidispersed data with excessive zeros

B.P. Tlhaloganyang[*] and R.M. Sakia
University of Botswana, Gaborone, Botswana
tlhaloganyangs@gmail.com

Available online at: www.iscamaths.com , www.isca.in , www.isca.me
Received 5th April 2019, revised 17th October 2019, accepted 16th November 2019

Abstract

From the literature, choosing the right model when the dependent variable is a count outcome remains a problem in literature. For count outcome variable with overdispersion due to excessive zero counts (zero inflation), Zero Inflated distributions such as Zero Inflated Poisson/Negative Binomial are usually considered to find better fitting models. Moreover, numerous studies suggested that if the data is characterized by equidispersion with signs of zero inflation, Zero Inflated Poisson (ZIP) distribution should be applied. Therefore, the aim of this paper is to investigate if ZIP distribution should substitute standard Poisson distribution if there are signs of zero inflation in equidispersed data. Equidispersed simulated and real life datasets with signs of zero inflation were used for the analysis. Evidence of equidispersion and zero inflation were tested and goodness-of-fit tests for both Poisson and ZIP distributions were obtained. Results revealed that for an equidispersed data with signs of zero inflation, standard Poisson performed better than ZIP distribution.

Keywords: Equidispersion, zero inflation, Poisson, zero inflated Poisson.

Introduction

Count data is encountered on daily basis in fields such as in transportation, health, manufacturing and many more. Choosing the right model when the dependent variable is a count outcome remains a problem in literature. In the past, one strategy was to treat count outcome as a continuous outcome and model it with Ordinary Least Square (OLS) regression. Another alternative was to dichotomize the count outcome (e.g event occurrence or not) and model it using binomial (logit or probit) regression. In nature, count outcome is non-linear, and as a result, linearity assumption of OLS is violated. Moreover, a lot of information is lost if we treat range of values as one value[1,2].

Recently, count outcome has been modelled using discrete distributions (e.g Poisson, Negative Binomial and many more) taking on non-negative integer values with lower bound of zero. The choice of which model to apply ranges from standard to modified distributions depending on the outcome's underlying distribution specifically the variance-mean relationship. From a range of available discrete distributions, Poisson distribution is considered the benchmark distribution in count data analysis and is suitable for application only when the data is equidispersed (i.e, equality of mean and variance). Violation of this assumption makes it inappropriate for analysis even though many practitioners usually apply it without first confirming validity of equidispersion[3]. When the data does not adhere to equality of mean and variance, we say it is under/overdispersed depending on the variance-mean relationship although overdispersion is the most common in analysis.

In the detection of overdispersion, Negative Binomial (NB) distribution is considered the best alternative feasible distribution as it allows the variance to exceed the mean[4]. Overdispersion can occur as a result of excessive zero counts that give smaller conditional mean than the true mean value. Data with excessive zeros are highly skewed to the right and it is termed as zero inflated data. When the data is characterized by overdispersion resulting from zero inflation both the Poisson and NB distribution under-predict the observed number of zeros[5]. As a result, Zero Inflated distributions (ZIDs) are usually considered to find better fitting models. Examples of the ZIDs include Zero Inflated Poisson (ZIP), Zero Inflated Negative Binomial (ZINB), Zero Inflated Negative Binomial-Crack (ZINB-CR) distribution and many more. These models are found to provide a statistically superior fit to the zero inflated data[6,7].

Based on these count data features, it is important for a researcher to first understand the type of data in hand before choosing the model to use[8]. Different datasets adhere to different distributions and for this reason choosing appropriate model remains a problem for most of researchers. In most cases, practitioners visually inspect descriptive characteristics to determine the model to use. Relying on this strategy without implementing statistical tests can sometimes result in the use of wrong models. Generally, use of inappropriate model can result in underestimation of the standard errors hence overstating the significance of regression coefficients which could bring uncertainty into research and practice[9].

For that reason, numerous studies proposed practical decision-making process of choosing the right model[1,10,11]. The logic of choosing appropriate distribution as suggested by the above mentioned studies can be represented as a summary in Figure-1. For explanation, the studies suggested that at the start, overdispersion should be checked. If overdispersion is detected, the researcher should proceed to check if there is zero inflation. If there is inflation, ZINB distribution should be applied, alternatively, Negative Binomial should be considered. On the other hand, if the dataset is observed to be equidispersed, these studies recommended that one should proceed to test for zero inflation. In the absence of inflation at zero, Poisson was suggested otherwise ZIP distribution should be applied.

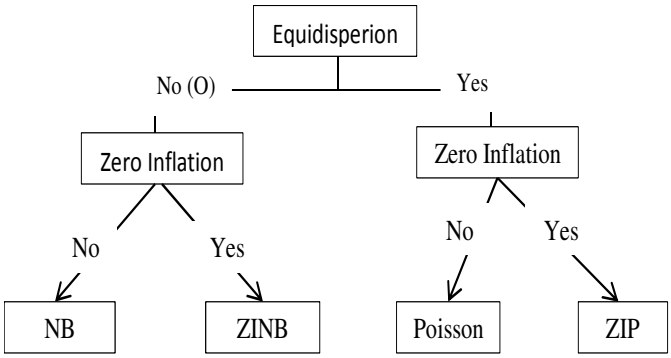

Figure 1: Selection of a distribution under Overdispersion/Zero Inflation (O: Overdispersion).

Therefore, the aim of this paper was to investigate if ZIP distribution should substitute standard Poisson distribution if there are signs of zero inflation in equidispersed data. The next section gives a review of methods used to accomplish the objective. In the succeeding section, application was conducted based on the use of simulated and real life data. The paper ends with discussion of results and recommendations. To study this in a practical sense, equidispersed count data signaling zero inflation was used.

Methodology

Poisson distribution: Assuming Y is a count random variable under all distributions, if it follows Poisson distribution its pmf is given by

$$P(Y=y)=\frac{e^{-\lambda}\lambda^{y}}{y!},\quad y\in\ \{0\}\cup\mathbb{Z}^{+}$$

where $\lambda>0$ and $E(Y)=Var(Y)=\lambda$. Thus, data simulated from Poisson distribution is always equidispersed. A famous real life application of this distribution can found from Bortkiewicz[12] who studied number of deaths by horse kicks in the Prussian Army from 1875 to 1894.

Zero Inflated Poisson (ZIP) distribution: The pmf of ZIP distribution as shown by Edwin[9] is given as

$$P(Y=y)=\begin{cases}\phi+(1-\phi)e^{-\lambda} & for y=0\\ (1-\phi)\dfrac{e^{-\lambda}\lambda^{y}}{y!} & for y\in\mathbb{Z}^{+}\end{cases}$$

where $0<\phi<1$ and it is termed as the zero inflation parameter. When$\phi=1$ it implies mass is concentrated at 0 (excluded) and $\phi=0$ implies absence of zero inflation. ZIP is a finite-mixture distribution of two distributions being Poisson distribution and Degenerate distribution with their probabilities concentrated at zero. The mean and variance of ZIP distribution are given as $E(Y)=(1-\phi)\lambda$ and $Var(Y)=\lambda(1-\phi)(1+\phi\lambda)$ respectively.

Overdispersion testing: It should be noted that the data simulated from a Poisson distribution does not need to be tested for overdispersion. To test if a real life dataset is overdispersed, we assume a Poisson mixture in outcome variable signifying presence of overdispersion. Assuming Y~Poisson(v,λ) where v is a continuous r.v taking on non-negative values with $E(v)=1$ and $V(v)=\tau$. The random variable Y then becomes a r.v of mixed Poisson with mean and variance given as $E(y)=\lambda$ and $V(y)=\lambda+\tau\lambda$ respectively[13,14]. In relation to the variance of Poisson distribution, extra variation can be examined based on the hypotheses $H_0;\tau=0$ vs $H_1;\tau>0$. This test will be implemented using AER-package based on the test statistic developed by Cameron and Trivedi[15]. When $\tau>0$ there is overdispersion, when $\tau=0$ there is equidispersion and $\tau<0$ implies there is underdispersion.

Score test for zero inflation: Based on mean and variance estimate, $\hat{\lambda}=\bar{x}$, Muoka[3] developed the score test for zero inflation under the hypotheses $H_0:\phi=0$ vs $H_1:\phi>0$ as

$$U=\frac{(n_0-n\bar{P}_0)^2}{n\bar{P}_0(1-\bar{P}_0)-n\bar{x}\bar{P}_0}\sim\chi_1^2$$

Where: $\bar{P}_0=\ exp\ (-\hat{\lambda})$, n and n_0 represent the total number of observations and number of zeros in the data respectively. Failure to reject the null hypothesis will indicate that the data is not inflated at zero.

Results and discussion

In this section, equidispersed data of different sizes was simulated from Poisson using small mean values. Datasets with mean values closer to zero were considered as they entailed more zero counts. Real life application was done using the famous Poisson distributed dataset studying deaths by horse kicks in the Prussian Army from 1875 to 1984 as it had a more zero counts[12]. At the start, the data was tested if it is really equidispersed. Score test was also applied to check for zero inflation. Moreover, Akaike Information Criterion (AIC) was obtained to evaluate the goodness-of-fit of Poisson and ZIP distributions. Minitab and R-software were used for analysis.

Table-1 and 2 gives data simulated from Poisson with mean values 0.1 and 0.85 respectively and sample sizes of 100, 1000 and 10000 in each of these datasets. Poisson simulated datasets does not need to be tested for equidispersion as their means are equal to their variances. Visually, Table-1 and 2 contains a lot of zero counts which could possibly signal zero inflation.

Table-1: Poisson simulated data at $\lambda = 0.1$.

	n=100	n=1000	n=10000
Count	Frequency	Frequency	Frequency
0	90	907	9051
1	10	88	903
2	0	5	44
3	0	0	2
U-statistic (p-val)	.457	.862	.989
AIC: Poisson ZIP	68.05 70.05	660.2 662.2	6661.5 6663.5

Table-2: Poisson simulated data at $\lambda = 0.85$.

Count	n=100	n=1000	n=10000
	Frequency	Frequency	Frequency
0	41	474	4318
1	39	349	3590
2	17	128	1546
3	3	40	443
4	0	8	88
5	0	1	14
6	0	0	1
U-statistic (p-val)	.303	.425	.553
AIC: Poisson ZIP	232.9 234.9	2321.4 2322.8	24181.8 24183.5

Based on the score test for zero inflation, U, all the corresponding p-values in Tables-1 and 2 for n=100, n=1000 and n=10000 supports the null hypothesis that there is no inflation at zero. Moreover, the AIC scores of Poisson distribution from are smaller than that of ZIP distribution at different sizes indicating that Poisson performs better than ZIP distribution. Considering the real life application shown in Table-3, approximately 55% of the counts are zeros.

Test of overdipersion and zero inflation p-values obtained from Table-3 point out that this data is equidispersed and it is not inflated at zero. AIC scores comparing Poisson and ZIP distribution marked Poisson as the best model.

Table-3: Number of Prussian soldiers accidentally killed by horse-kick[12].

Count	Frequency
0	109
1	65
2	22
3+	4
Overdispersion (p-val)	.970
U-statistic (p-val)	.928
AIC: Poisson ZIP	414.2 416.2

Conclusion

Poisson distribution is known to be a benchmark count data distribution. It is suitable for application when the mean is equal to the variance. If equidispersed data appears to have lot of zero counts, numerous studies have suggested that Zero Inflated Poisson (ZIP) distribution should be used. This paper therefore investigated whether ZIP distribution should replace Poisson distribution if equidispersed data have excessive zero counts. From the analysis, score test results showed that excessive zero counts is not a problem in equidispersed data. In addition, AIC scores revealed that as long as count data is equidispersed, Poisson distribution is the best model to consider even if there are lot of zero counts. Moreover, ZIP distribution have a variance greater than mean, therefore, it should be used in overdispersion cases.

References

1. 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.

2. Long J.S. (1997). Regression Models for Categorical and Limited Dependent Variables: Advanced Quantitative Techniques in the Social Sciences. *Sage Publications, Thousand Oaks*, 7.

3. Muoka A.K., Waititu A. and Ngesa O.O. (2016). Statistical models for count data. *Science Journal of Applied Mathematics and Statistics*, 4(6), 256-262.

4. Ismail N. and Zamani H. (2013). Estimation of claim count data using negative binomial, generalized poisson, zero-inflated negative binomial and zero-inflated generalized poisson regression models. *In Casualty Actuarial Society E-Forum*, 41, 1-28.

5. Akın D. (2011). Analysis of highway crash data by negative binomial and poisson regression models. Second International Symposium on Computing in Science and Engineering, Kusadasi, Izmir, Turkey, 1.

6. Chen F., Chen S. and Ma X. (2016). Crash frequency modeling using real-time environmental and traffic data and unbalanced panel data models. *International journal of environmental research and public health*, 13(6), 609.

7. Dong C., Nambisan S.S., Richards S.H. and Ma Z. (2015). Assessment of the effects of highway geometric design features on the frequency of truck involved crashes using bivariate regression. *Transportation Research Part A: Policy and Practice*, 75, 30-41.

8. Yaacob W.F.W., Lazim M.A. and Wah Y.B. (2010). A practical approach in modelling count data. In Proceedings of the Regional Conference on Statistical Sciences, Malaysia, 176-183.

9. Edwin T. (2014). Power series distributions and zero-infalted models. Doctorate Thesis. University of Nairobi.

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

11. 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.

12. Bortkiewicz L. (1898). Das Gesetz der kleinen Zahlen. BG Teubner, Leipzig.

13. Alam N. (2015). Detecting Overdispersion in Count Data: Comparison of Tests. Doctorate Thesis. East West University.

14. 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.

15. Cameron A.C. and Trivedi P.K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of econometrics*, 46(3), 347-364.