Estimation of Zero-Inflation Parameter in Zero-Inflated Poisson Model

K.M.Sakthivel[#1], C.S.Rajitha[*2]

Department of Statistics, Bharathiar University, Coimbatore 641046, Tamil nadu, India

Abstract — *The modelling of count data is extensively used in many fields of research. There is handful of zero-inflated probability models available in literature. Among these models, zero inflated Poisson distribution is one of the widely used models for modelling data with excess number of zeros. In all the zero-inflated models, one can have parameter called zero-inflation parameter which is in addition to the number of parameters in underlying distribution. The estimation of the zero-inflation parameter of the zero-inflated Poisson (ZIP) models by MLE do not have an explicit expression and solved iteratively by using modern computing techniques. In this paper, we proposed a probability based inflation estimator (PBIE) for making inferences about the inflation parameter of the ZIP model and also studied the performance of the proposed estimator for simulated data.*

Keywords — *mean squared error, MLE, moment estimators, zero-inflation Parameter, zero-Inflated Poisson Model.*

I. INTRODUCTION

Choosing a suitable distribution for count data plays a significant role in data analysis. Usually Poisson distribution is used for modelling the count data. But in many applications count data shows over dispersion and long tail behaviour. So that other distributions such as negative binomial and generalized Poisson (Cosul and Jain, 1973) distributions were utilized by many researchers for modelling the over dispersed count data. And the existence of excess number of zero counts is a very specific type of over dispersion. Zero inflated models can be utilized in the situation where excess number of zero counts is generated in the count data. In zero inflated models we considered that data are coming from a mixture of two distributions in such a way that one generates only zero counts and other generates non negative counts from a Poisson or negative binomial distribution. So mixtures of degenerate or an odd distribution at zero with a non negative integer valued distribution are most suitable models in this particular situation. The result of these mixture distributions are suitable for over dispersed count data and these models are called zero inflated models. One of the most popular count models for modelling the zero inflated data is the zero inflated Poisson model (Lambert, 1992). The efficiency of zero inflated Poisson (ZIP) model over Poisson distribution was discussed by Rideout et.al (1998). For analysing the radio audience data, Couturier et.al (2010) used the zero inflated truncated generalized Pareto distribution. For dealing with the inflated count data, Davidson (2012) showed that ZIP distribution is the adequate model.

For estimating the parameters of the ZIP, various methods are used in the literature. Nanjundan and Naika (2012) discussed maximum likelihood estimation (MLE) and moment estimation (ME) methods for estimation of parameters of the ZIP distribution and compared the performance of MLE and ME asymptotically through simulation study. Similarly Becket et al (2014) discussed the estimation of parameters of ZIP distribution using MLE and ME via a simulation study with regard to standardized bias and standardized mean squared error. And for measuring the amount of zero inflation exhibits in the count data from Poisson distribution, Puig and Valero (2006) introduced a measure called zero inflation index for selection of distribution.

In this paper we introduced a new estimator for estimating the inflation parameter of ZIP distribution and provide suitable indication to promote the use of new estimator. The organization of the paper is as follows. In section 2, we provide an overview of ZIP model, in section 3, we briefly overview the zero inflation index and establish the behaviour of zero inflation index regarding various values of zero inflation parameter. In section 4, we proposed an estimator for zero inflation parameter of ZIP named as probability based inflation estimator (PBIE). In section 5 we compare the performance of PBIE with MLE via simulation study and showed that PBIE performs as good as MLE in terms of MSE, bias and SE. And in section 6 we provide a conclusion to the study.

II. ZERO INFLATED POISSON (ZIP) DISTRIBUTION

Poisson distribution is usually used for modelling the count data and this distribution assumes that mean and variance of the distribution are equal. However in real life applications most of the count data are obtained with

variance greater than mean. Usually this happened due to the unobserved heterogeneity of the data and hence the data exists with long tails. Another cause for excess variability is the occurrence of extra zero counts. In this situation zero inflated models provides better fit to the data. One of the most popular zero inflated models are the zero inflated Poisson distribution. It is suggested in a particular state of affairs, when the observed data shows two kinds of zeros. It is defined as a modified count model with two parts. The first part generates excess zero counts and zeros coming from the Poisson distribution and the second part generates positive counts from a truncated Poisson distribution.

A random variable Y is said to have a zero inflated Poisson distribution its probability mass function (PMF) is obtained as

$$P(Y=y/\theta,\pi)=\begin{cases}\pi+(1-\pi)e^{-\theta} & when \quad y=0\\ (1-\pi)\dfrac{e^{-\theta}\theta^{y}}{y!} & when \quad y>0\end{cases}$$

where $\theta>0$ and $0<\pi<1$. And

$$P(y;\theta,\pi)=\pi P_0(y)+(1-\pi)P_1(y,\theta)$$

Where $P_0(y)=\begin{cases}1 \ if \ y=0\\ 0 \ if \ y\neq 0\end{cases}$ and $P_1(y,\theta)=\dfrac{e^{-\theta}\theta^{y}}{y!}, y=0,1,2,\ldots$

So that this distribution can be considered as a convex combination of a distribution which is degenerate at zero and the Poisson distribution with mean θ.

III. ZERO INFLATION INDEX

The concept of measuring the zero inflation from the Poisson distribution was first introduced by Puig and Valero (2006) and they named the measure as zero inflation index. The zero inflation index of a count random variable can be defined as

$$z_i=1+\log(p_0)/\theta$$

where p_0 is the proportion of zero counts and θ is the mean of the count random variable. If the value of z_i is zero then it is considered as Poisson distributed random variable. Otherwise it is considered that the random variable follows zero inflated Poisson distribution.

We generate 500 random samples from a zero inflated Poisson distribution with $\theta=2$ and $\pi_0=0.1,0.2,\ldots,0.8$ for different sample sizes n = 25, 50, 75 and 100. Table 1 shows the average zero inflation indexes of the above mentioned sample sizes across π_0. From the table it can be seen that for almost all sample sizes the zero inflation parameter is positive and overestimate the true parameter value of π_0.

TABLE 1

AVERAGE ZERO INFLATION INDEX OF 500 RANDOM SAMPLES FOR $\theta=2$ ACROSS π_0

Sample size(n)	Average zero inflation index							
	0.1	0.2	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
25	0.1971	0.3931	0.5369	0.6263	0.7141	0.7781	0.8571	0.8978
50	0.2432	0.4014	0.5492	0.6471	0.7264	0.7821	0.8413	0.9027
75	0.2645	0.4123	0.5362	0.6377	0.7262	0.7841	0.8571	0.8837
100	0.2765	0.4231	0.5298	0.6362	0.7173	0.7831	0.8461	0.9089

IV. ESTIMATION OF PARAMETERS OF ZIP DISTRIBUTION

A. *Method of moments*

For the zero inflated Poisson distribution the mean and variance are given by

$$E(Y)=\bar{Y}=\pi\theta \text{ and } Var(Y)=S^2=\pi\theta(1+\theta(1-\pi))$$

By solving these equations we get the moment estimators of ZIP are

$$\hat{\theta}_{ME}=\bar{Y}+(s^2/\bar{Y})-1$$

And

$$\hat{\pi}_{ME}=(s^2-\bar{Y})/(\bar{Y}^2+(s^2-\bar{Y})$$

Where $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$, $s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$

B. Method of Maximum likelihood

Let $y_1, y_2, \ldots, y_n$ be a random sample taken from ZIP distribution, then the likelihood function of the distribution can be written in the form

$$L(\theta / y_1, y_2, .. y_n) = [\pi + (1-\pi)e^{-\theta}]^{n_0} [(1-\pi)e^{-\theta}]^{n-n_0} \theta^S \left(\prod_{i=1}^{n} y_i! \right)$$

where $n_0 = \sum_{i=1}^{n} I\{y_i = 0\}$, $S = \sum_{i=1}^{n} y_i$ and $I\{y_i = 0\}$ is indicator variable defined on $\{y_i = 0\}$.

According to el-Shaarawi (1985), the maximum likelihood estimators of the ZIP distribution are found as

$$\hat{\theta}_{MLE} = \frac{S(1 - e^{-\hat{\theta}_{MLE}})}{n - n_0}, \quad \hat{\pi}_{ME} = \frac{\hat{p}_0 - e^{-\hat{\theta}_{MLE}}}{1 - e^{-\hat{\theta}_{MLE}}}$$

where $\hat{p}_0 = \frac{n_0}{n}$ which represents the empirical probability of $\{y_i = 0\}$ and is an unbiased estimator of $p_0 = P(Y = 0)$. $\hat{\theta}_{MLE}$ can be found by using Newton-Raphson method or some of the iterative methods like this, which can converge quickly to true MLE.

C. Probability based inflation estimator (PBIE)

For making inferences about the inflation parameter π of the ZIP distribution, we introduced a new estimator using the non-zero probabilities of Poisson and ZIP distribution. And the advantage of using the proposed estimator is very simple, since this estimator doesn't need any of the iterative methods for estimating the inflation parameter. The proposed estimator can be defined as follows

$$\hat{\pi}_{PBIE} = \left| \frac{(1-\pi)K_1 - K_2}{K_2} \right|$$

where $K_1 = P_{ZIP}(Y \neq 0)$ and $K_2 = P_{Pois}(Y \neq 0)$. $P_{ZIP}(Y \neq 0)$ is the probability of ZIP distribution at $Y \neq 0$ and $P_{Pois}(Y \neq 0)$ denotes the probability of Poisson distribution at $Y \neq 0$.

V. SIMULATION STUDY

For comparing the performance of the PBIE and MLE of the inflation parameter π of ZIP distribution, we conducted a simulation study and compared the performances of these two estimators using different measures such as bias, mean square error (MSE) and standard error (SE).

We have conducted the simulation study using the following algorithm:

1. Generate m=500 random sample for each of size n=25,50 and 100 from the $ZIP(\theta, \pi)$ for $\theta = 2$ and $\pi = \pi_0$ where $\pi_0 = 0.1, 0.2, \ldots, 0.8$
2. We calculate $\hat{\theta}_{MLE}$, $\hat{\theta}_{ME}$, $\hat{\pi}_{MLE}$, $\hat{\pi}_{ME}$ and $\hat{\pi}_{PBIE}$ with the help of R software.
3. Estimate $MSE(\hat{\pi}_{MLE})$ *and* $MSE(\hat{\pi}_{PBIE})$ as follows

$$MSE(\hat{\pi}_{MLE}) = \frac{\sum_{i=1}^{m} (\hat{\pi}_{MLE}^{(i)} - \pi)^2}{m}$$

$$MSE(\hat{\pi}_{PBIE}) = \frac{\sum_{i=1}^{m} (\hat{\pi}_{PBIE}^{(i)} - \pi)^2}{m}$$

The relative efficiency of $\hat{\pi}_{PBIE}$ over $\hat{\pi}_{MLE}$ are evaluated as follows.

$$RE(\hat{\pi}_{PBIE} / \hat{\pi}_{MLE}) = \frac{MSE(\hat{\pi}_{MLE})}{MSE(\hat{\pi}_{PBIE})}$$

Since the efficiency of $\hat{\pi}_{MLE}$ is always better than $\hat{\pi}_{ME}$ from the past study. Hence we ignored the comparison of $\hat{\pi}_{PBIE}$ with $\hat{\pi}_{ME}$.

4. The sample variances $v(\hat{\pi}_{MLE})$, and $v(\hat{\pi}_{PBIE})$ are computed as follows.

$$v(\hat{\pi}_{MLE}) = \frac{\sum_{i=1}^{m} (\hat{\pi}_{MLE}^{(i)} - \bar{\hat{\pi}}_{MLE})^2}{m},$$

$$v(\hat{\pi}_{PBIE}) = \frac{\sum_{i=1}^{m} (\hat{\pi}_{PBIE}^{(i)} - \bar{\hat{\pi}}_{PBIE})^2}{m}$$

Where $\bar{\hat{\pi}}_{MLE} = \frac{\sum_{i=1}^{m} \hat{\pi}_{MLE}^{(i)}}{m}$, $\bar{\hat{\pi}}_{PBIE} = \frac{\sum_{i=1}^{m} \hat{\pi}_{PBIE}^{(i)}}{m}$

5. The average biases of the estimates for MLE and PBIE are calculated as follows.

$$bias(\bar{\hat{\pi}}_{MLE}) = \frac{\sum_{i=1}^{m} (\hat{\pi}_{MLE}^{(i)} - \pi)}{m}$$

$$bias(\bar{\hat{\pi}}_{PBIE}) = \frac{\sum_{i=1}^{m} (\hat{\pi}_{PBIE}^{(i)} - \pi)}{m}$$

6. For m = 500 samples, the steps (1) to (5) are repeated for $\pi_0 = 0.1, 0.2, \ldots, 0.8$. For sample sizes n=25, 50 and 100, a simulation study is carried out and the results are presented in Table 2 only for n=100. (The results corresponding to 25, 50 are not provided due to almost similar behaviour). By observing the relative efficiency $RE(\hat{\pi}_{PBIE} / \hat{\pi}_{MLE})$, it can be seen that PBIE performs as good as the MLE.

The following figures (Fig 1 to 3) shows the performance of MLE and PBIE for making inference about the inflation parameter π of ZIP. It is observed that the MSE of MLE and PBIE of inflation parameter are almost equal. In terms of SE also MLE and PBIE of inflation parameter are almost equal. The behaviour (diagram) of MLE and PBIE for sample sizes 25 and 50 etc not included in the paper due to the similar pattern. From Table 2 and the diagrams it can be seen that the PBIE performs equally good as MLE of inflation parameter π of ZIP.

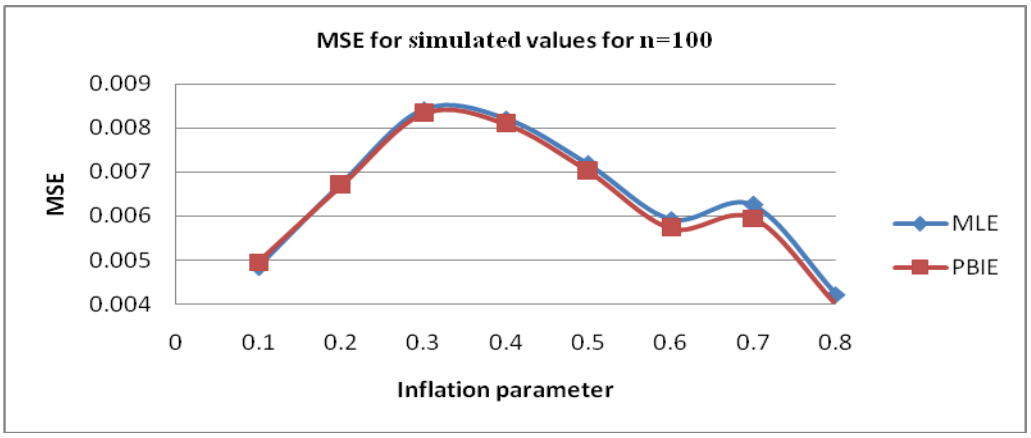

Fig. 1 MSE of estimated inflation parameter π for ZIP distribution for n= 100

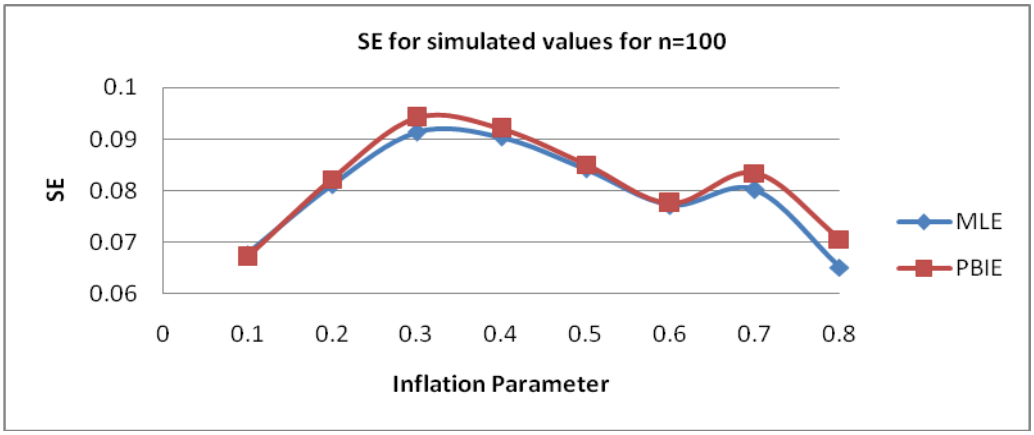

Fig. 2 SE of estimated inflation parameter π for ZIP distribution for n= 100

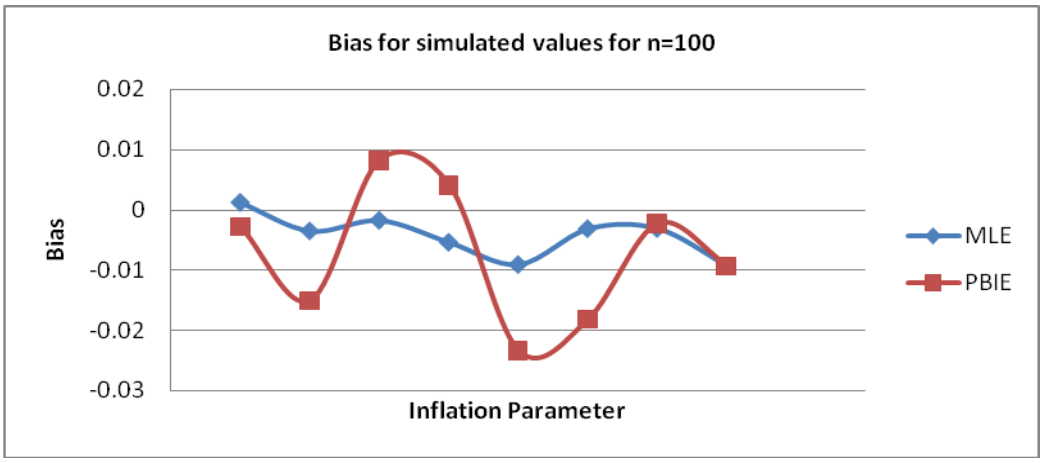

Fig. 3 Bias of estimated inflation parameter π for ZIP distribution for n= 100

TABLE 2

MSE, SE, BIAS AND RELATIVE EFFICIENCY FOR π OF ZIP DISTRIBUTION FOR n=100

Inflation parameter	n=100				
	Estimate type	MSE	SE	Bias	RE
0.1	MLE PBIE	0.00484 0.00495	0.0678 0.0672	0.00124 -0.00284	1.022727
0.2	MLE PBIE	0.00674 0.0067	0.0812 0.0822	-0.00352 -0.01512	0.994065
0.3	MLE PBIE	0.00842 0.00833	0.0914 0.0942	-0.00175 0.00821	0.989311
0.4	MLE PBIE	0.00821 0.0081	0.0904 0.0921	-0.00541 0.00412	0.986602
0.5	MLE PBIE	0.00717 0.00701	0.0842 0.0851	-0.00912 -0.02344	0.977685
0.6	MLE PBIE	0.00592 0.00572	0.0772 0.0776	-0.00314 -0.01821	0.966216
0.7	MLE PBIE	0.00625 0.00595	0.0802 0.0834	-0.00312 -0.00234	0.952
0.8	MLE PBIE	0.00421 0.00399	0.0652 0.0706	-0.00934 -0.00932	0.947743

VI. CONCLUSION

In this paper we briefly overviewed different method of estimation of zero inflation parameter of a ZIP distribution. Further zero inflation index is also discussed for detecting the over dispersion for the selection of Poisson distribution and zero inflated Poisson distribution. Usually MLE is used for estimation of the parameters of ZIP distribution, but most of the time it does not provide closed form expressions for the parameters of zero inflated models particularly zero inflation parameter. The proposed estimator called PBIE is simple and performs as efficient as MLE in case of zero inflation parameter and the performance of this estimator with MLE is proved empirically through simulation study. The study shows that the proposed estimator can be used as an alternative for estimating the zero inflation parameter.

REFERENCES

[1] S. Beckett, J. Jee, T. Ncube, , S. Pompilus, Q.Washington, A. Singh, and N. Pal, "Zero-inflated Poisson (ZIP) distribution: parameter estimation and applications to model data from natural calamities," *Involve: A Journal of Mathematics*, vol.7(6),pp. 751–767, 2014.

[2] P.C. Consul, and G.C Jain, "On some interesting properties of the generalized Poisson distribution," *Biometrical Journal*, vol. 15, pp. 495–500, 1973b.

[3] F.Famoye, "Parameter estimation for generalized negative binomial distribution," *Communications in Statistics-Simulation and Computation*, vol. 26(1), pp. 269–279, 1997

[4] W. Feller, "On a general class of contagious distributions," *Annals of Mathematical Statistics*, *vol.*14 (4), pp.389–400, 1943.

[5] D. Lambert, "Zero-inflated Poisson regression with an application to defects in manufacturing," *Technometrics, vol*. 34(1),pp. 1–14, 1992.

[6] G. Nanjundan, and T. R. Naika,. "Asymptotic comparison of method of moments estimators and maximumlikelihood estimators of parameters in zero-inflated poissonmodel," *AppliedMathematics*, vol. 3, pp. 610–616, 2012.

[7] J. Neyman, " On a new class of contagious distributions applicable in entomology and bacteriology," *Annals of Mathematical Statistics*, vol. 10(1), pp. 35–57, 1939

[8] M. K. Patil, and D. T. Shirke, "Testing parameter of the power series distribution of a zero-inflated power series model," *Statistical Methodology*, vol. 4(4), pp.393–406, 2007.

[9] M. K. Patil, and D. T. Shirke, "Tests for equality of inflation parameters of two zero-inflated power series distributions," *Communications in Statistics-Theory and Methods*, vol. 40(14), pp. 2539–2553, 2011.

[10] P. Puig, and J.Valero, "Count data distributions: Some characterizations with applications," *Journal of American Statistical Association*, vol. 101(473), pp. 332–340, 2006.

[11] M. Ridout, C.G.B. Demetrio, and J. Hinde, "Models for count data with many zeros," In *International Biometric Conference*, Capetown, South Africa, December. 1998

[12] Y. S.Wagh, and K. K. Kamalja, "Comparison of methods of estimation for parameters of Generalized Poisson distribution through simulation study," *Communications in Statistics-Simulation and Computation*, vol. 46(5), pp. 4098–4112, 2017.