Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 4 (2015), pp. 2461-2467

http://www.ripublication.com

An Alternative Method For Fitting A Zero Inflated Negative Binomial Distribution

Zamira Zamzuri

School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia

ABSTRACT

Traffic accident data is usually over dispersed and has extra zeros. Zero-inflated negative binomial distribution (ZINB) has often been used to fit this type of data. A simulation study has been conducted to investigate the performance of the estimators of the ZINB with different proportions of zeros. It is found that the commonly used Maximum Likelihood Estimator (MLE) produce inaccurate estimates of the parameters for data with low proportion of zeros. Hence, the aim of this study is to propose an alternative method to fit the zero-inflated negative binomial distribution. The alternative method consists of two phases: a grid search and conditional maximum likelihood estimator (GCMLE). The empirical results indicate that this method produces better results than MLE in terms of smaller bias for dispersed data with a moderate proportion of zeros.

Keywords: count data, negative binomial, traffic accident, zero-inflated

1. INTRODUCTION

Recent studies reported that accident counts display an excess presence of zeros than would be expected from either a Poisson or a negative binomial distribution (Shankar et al. 1997; Kumara & Chin, 2003; Qin et al. 2004). Hence, a zero-inflated negative binomial distribution has been used to fit a traffic accident count data, since the data exhibits overdispersion and zero-inflation. The fundamental property of zero-inflated distributions is that there are two processes that generate the zero counts in the distribution; namely the structural zeros process and the random zeros process. The random zeros are generated from the count distribution, while structural zeros are from another process that generates only zero counts. The zero inflated negative binomial distribution is given as

$$g(x)=\begin{cases} \pi+(1-\pi)\left(\dfrac{1}{1+\theta\mu}\right)^{\frac{1}{\theta}} & x=0 \\ (1-\pi)\left(\dfrac{1}{\theta}+x-1\right)\left(\dfrac{1}{1+\theta\mu}\right)^{\frac{1}{\theta}}\left(\dfrac{\theta\mu}{1+\theta\mu}\right)^{x} & x=1,2,\ldots \end{cases} \qquad Eq.1$$

where π is the proportion of structural zeros, μ is the mean and θ is the dispersion parameter.

To estimate parameters of the zero-inflated negative binomial distribution, the maximum likelihood can be used, as provided by the R package 'zeroinfl' (Zeileis et al., 2008). However, it struggles to give an accurate estimate when π is small. To encounter this problem, we propose an alternative method to fit this distribution. This method combines two phases: a grid search and the conditional maximum likelihood method. Conditional maximum likelihood method has been shown as the best option to estimate parameters of a negative binomial distribution (Anraku & Yanagimoto, 1990). In the next section, we will describe the alternative method followed by the comparisons of the alternative method to maximum likelihood.

2.0 PROBLEM WITH THE MAXIMUM LIKELIHOOD METHOD

Table 1 is an extract of a simulation study that we conducted to compare the performance between `zeroinfl' and an alternative method that we introduced, conditional maximum likelihood with a grid search (GCMLE). The full results of this simulation study are presented in a later section. In Table 1, although `zeroinfl' seriously underestimates π, the *p*-value of the chi square goodness of fit test suggests that this is a good fit. If we inspect the histograms of the data compared to the fitted values (Figure 1), it is apparent that `zeroinfl' estimates provide a satisfactory fit. However, the true situation of the data has not been described; which is the data was generated by two different processes. Hence, there is a need for an alternative method that is able to provide accurate estimates of the parameters of this distribution. The last column of Table 1 provides estimates from an alternative method, conditional maximum likelihood method and a grid search (GCMLE). As can be seen in this table, GCMLE provides more accurate estimates. Details on the GCMLE method are provided in the next section.

Table 1: Problem with `zeroinfl' function and comparison between `zeroinfl' and GCMLE estimates

Parameter	True	'Zeroinfl'	GCMLE
π	0.300	0.004	0.200
μ	3.000	2.680	2.980
θ	2.860	3.570	2.380
p-value		0.749	0.858

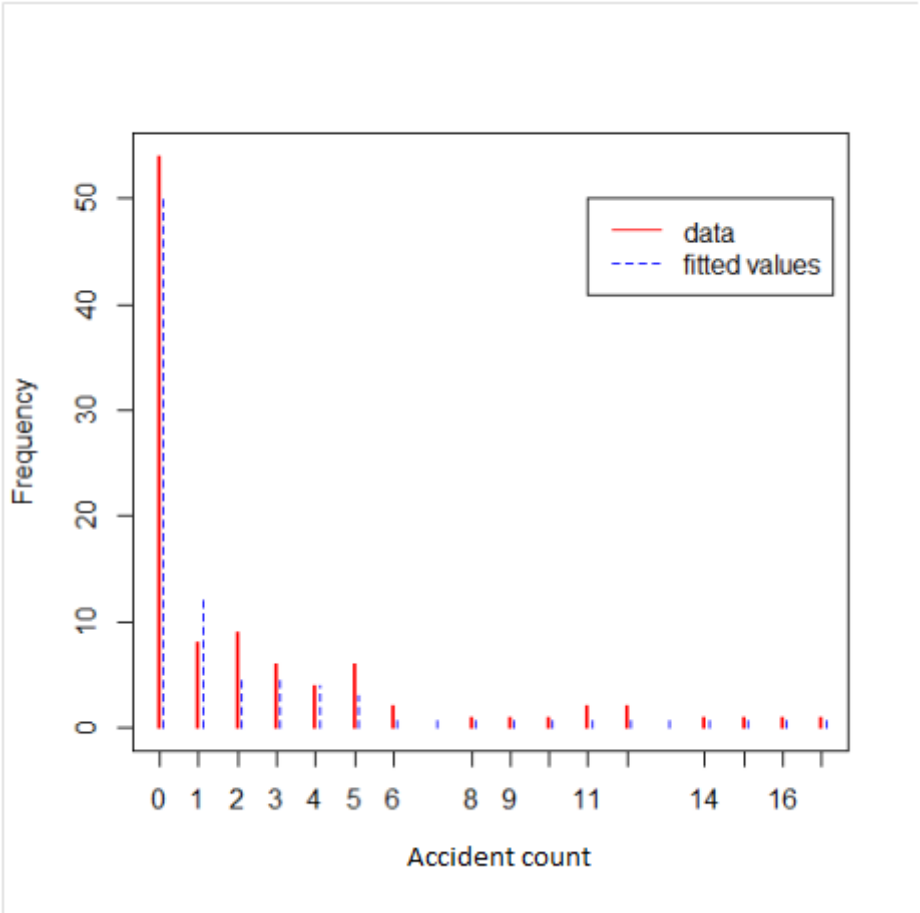

Figure 1: Comparison of histograms of the data and fitted values of `zeroinfl' solution

3. THE METHOD

We introduce an alternative method to estimate parameters of the zero-inflated negative binomial distribution by combining the conditional maximum likelihood method with a grid search. First, we explain the method overall then we look on the conditional maximum likelihood method. To apply the alternative method, we start with a grid search for π. Since π is a proportion, it ranges between zero and one. By using a range from 0.1 to 0.9 with step size 0.1, we extract the true zeros component from the data, leaving the data distributed as a negative binomial. We then use the conditional maximum likelihood method to estimate k and μ. We have nine sets of parameter estimates, and for each set of estimates, we test the goodness of fit of the fitted distribution to the data. By doing this, we can identify the interval with length 0.1 that produces the best fit, i.e the one with the smallest value of χ^2. Following this step, we repeat the same procedure for another grid of possible parameter values in the selected interval, with step size 0.01. The set of parameter estimates that produces the best fit after this second phase is the best estimate up to two decimal places for the parameters of the zero-inflated negative binomial distribution.

Conditional maximum likelihood is one of the methods available for estimating parameters of the negative binomial distribution. According to Anraku and Yanagimoto (1990), the conditional maximum likelihood method provides a better estimate of parameters of the negative binomial distribution than maximum likelihood and moment estimators. The negative binomial distribution is given as

$$g(x)=\binom{\frac{1}{\theta}+x-1}{x}\left(\frac{1}{1+\theta\mu}\right)^{\frac{1}{\theta}}\left(\frac{\theta\mu}{1+\theta\mu}\right)^{x} \quad x=0,1,2,... \qquad Eq.2$$

where $E(X)=\mu$ and $V(X)=\mu+\frac{\mu^2}{\theta}$.

The likelihood function can be partitioned into the conditional *(LC)* and residual likelihood *(LR),*

$$L(x;\theta,\mu)=\prod_{i=1}^{n}\left\{\frac{\binom{\frac{1}{\theta}+x_i-1}{x_i}\binom{\frac{n}{\theta}+t-1}{t}}{\binom{\frac{n}{\theta}+t-1}{t}}\right\}\left\{\left(\frac{1}{1+\theta\mu}\right)^{\frac{n}{\theta}}\left(\frac{\theta\mu}{1+\theta\mu}\right)^{t}\right\}$$

$$=LC(x;\theta\,|\,t).LR(t;\theta,\mu)$$

where $t=\sum_{i=1}^{n}x_i$. The conditional likelihood estimator (*CLE)* equation for the dispersion parameter θ is

$$CLE(x;\theta)=\sum_{i=1}^{n}\sum_{j=1}^{x_i}\frac{j-1}{1+\theta(j-1)}-\sum_{i=1}^{t}\frac{i-1}{n+\theta(i-1)}$$

which is solved iteratively.

4. RESULTS & DISCUSSIONS

We ran a simulation to compare the performance of GCMLE and the `zeroinfl' function in R. We generated data with sample size *n=100* from the zero-inflated negative binomial, based on all combinations of these sets of parameters:

- π*: 0.1, 0.5, 0.8*
- μ*: 0.1, 0.5, 1, 5*
- θ*: 0.1, 0.5, 1, 2*

Then a zero-inflated negative binomial distribution was fitted to these data sets by using both GCMLE and the `zeroinfl' function. We recorded the goodness of fit test result and compared the performance of these two approaches. For each combination of parameters, we repeated this process 100 times.

Changes in values of the parameters change the features of the distribution. A

higher value of π implies a higher proportion of zeros in the data set. Increasing the value of θ makes the range of the data wider and increases the count of zeros in the data set. A smaller value of μ implies a higher proportion of zeros and smaller range of data.

Figure **2** summarizes the results comparing the performance of GCMLE and `zeroinfl'. The yellow regions represent where `zeroinfl' has a better fit than GCMLE; 'zeroinfl' has a poorer fit than GCMLE for the green regions. It is clear that the `zeroinfl' dominates the result; however there are some points worth discussing.

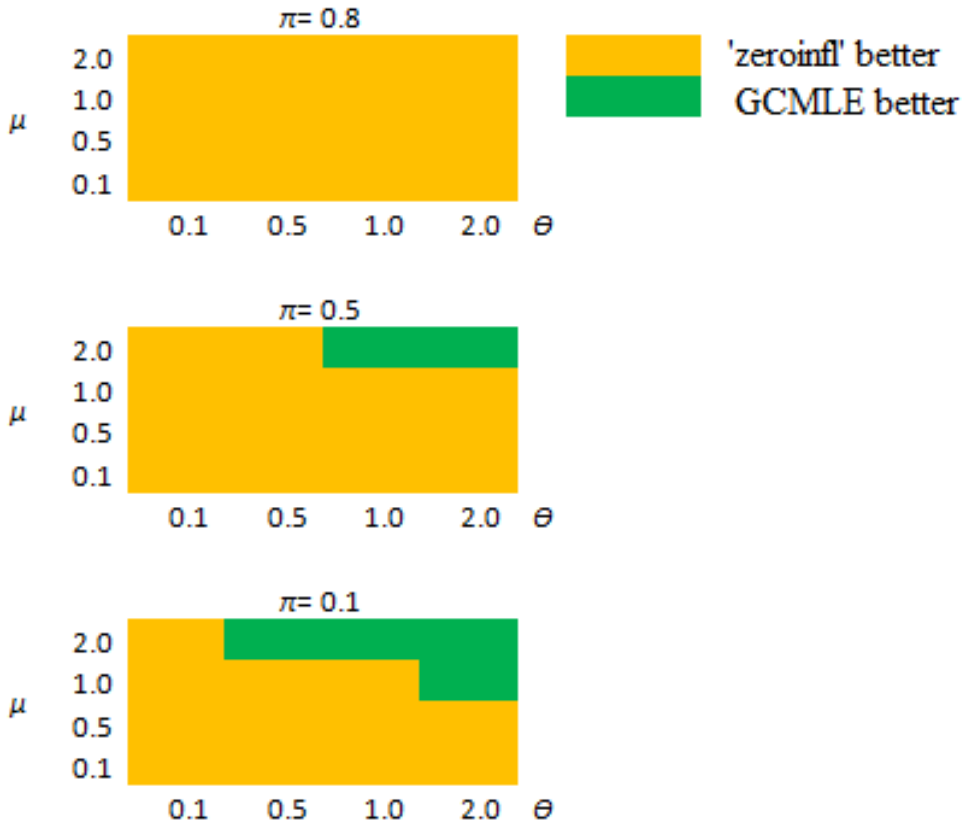

Figure 2: Simulation result of comparing GCMLE and `zeroinfl' function

By examining the green region, we see that the GCMLE approach can produce better estimates when the data has a wider range or is overdispersed with π= 0.1 and 0.5. When the data has too many zeros, resulting from a high value of π, a small value of μ or a small value of θ, GCMLE struggles to offer better estimation compared with the `zeroinfl' method. Hence, we can say that GCMLE offers a better fit to data that are overdispersed with a moderate level of zero-inflation. Although GCMLE did not give a set of parameter estimates with a better fit for other combinations of parameter values, this approach is more in control since we start the estimation process by

finding estimates of π from the grid search. Table **2** gives examples of parameter estimation by GCMLE and the `zeroinfl' function.

Table 2: Summary of comparisons between GCMLE and `zeroinfl' function

Example	Parameter	True	'Zeroinfl'	GCMLE	Comment
1	π	0.10	0.001	0.090	'Zeroinfl' provides better fit
	μ	0.10	0.078	0.112	but fails to identify the
	θ	0.10	1.190	0.085	presence of extra zeros.
	χ^2		0.502	0.745	
	p-value		0.778	0.689	
2	π	0.10	1E-4	0.103	For more dispersed data,
	μ	5.00	4.960	5.112	GCMLE provides estimates
	θ	2.00	2.840	2.350	with better fit and closer to
	χ^2		18.186	10.458	the true values of
	p-value		0.253	0.790	parameters.
3	π	0.50	0.510	0.530	Estimations by the two
	μ	5.00	5.939	5.097	methods are comparable
	θ	1.00	0.891	0.605	and provide a satisfactory
	χ^2		15.685	12.682	fit.
	p-value		0.109	0.242	
4	π	0.80	0.850	0.640	Both methods struggle with
	μ	5.00	8.743	2.263	the high proportion of zeros
	θ	2.00	0.559	0.485	and dispersed data.
	χ^2		52.078	71.673	
	p-value		5.2E-10	4.6E-14	

5. CONCLUSIONS

We proposed an alternative method for estimating parameters of a zero-inflated negative binomial distribution using conditional maximum likelihood with a grid search (GCMLE). Based on the comparison of this method with the maximum likelihood method, it is shown that the performance of both methods is comparable, but GCMLE works better with more dispersed data that has a moderate inflation of zeros.

ACKNOWLEDGEMENT

We would like to thank Profesor Graham Wood from Dept. of Statistics, Macquarie University, Australia and Dr. Ross Sparks from CSIRO for their invaluable input and comments to improve this paper

REFERENCES

2. Anraku, K. and Yanagimoto, T. (1990). Estimation for the negative binomial distribution based on conditional likelihood. *Communication in Statistics Simulation and Computation* 19, 771-786.
3. Kumara, S. S. P, Chin, H.C., (2003). Modelling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Injury Prevention 4(1),53-57.
4. Qin, X., Ivan, J.N., Ravishanker, N. (2004). Selecting exposure measures in crash rateprediction for two-lane highway segments. Accident Analysis & Prevention 36, 183-191.
5. Shankar, V., Milton, J., Mannering, F.L. (1997). Modelling accident frequency as zero-altered probability processes: an empirical enquiry. Accident Analysis & Prevention 29, 829-837.
6. Zeileis, A., Kleiber, C., Jackman, S. (2008). Regression Models for Count Data in R. Journal of Statistical Software 27(8), 1-25.