International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:7, No:8, 2013

Zero Inflated Models for Overdispersed Count Data

Y. N. Phang and E. F. Loh

***Abstract*—**The zero inflated models are usually used in modeling count data with excess zeros where the existence of the excess zeros could be structural zeros or zeros which occur by chance. These types of data are commonly found in various disciplines such as finance, insurance, biomedical, econometrical, ecology, and health sciences which involve sex and health dental epidemiology. The most popular zero inflated models used by many researchers are zero inflated Poisson and zero inflated negative binomial models. In addition, zero inflated generalized Poisson and zero inflated double Poisson models are also discussed and found in some literature. Recently zero inflated inverse trinomial model and zero inflated strict arcsine models are advocated and proven to serve as alternative models in modeling overdispersed count data caused by excessive zeros and unobserved heterogeneity. The purpose of this paper is to review some related literature and provide a variety of examples from different disciplines in the applications of zero inflated models. Different model selection methods used in model comparison are discussed.

***Keywords*—**Overdispersed count data, model selection methods, likelihood ratio, AIC, BIC.

I. Introduction

COUNT data with excess zeros are commonly found in many disciplines such as health sciences, medicine, ecology, econometric, finance and road safety. Examples of data with too many zeros from various disciplines including agriculture, econometrics, patent applications, species abundance, medicine and use of recreational facilities are cited in [1]. Zero inflated Poisson (ZIP) model has received a lot of attention and has always been used in modeling count data where the extra variations are solely caused by the extra zeros. For overdispersed count data where the extra variability is caused by excess zeros and also unobserved heterogeneity, recommended models are zero inflated negative binomial (ZINB), zero inflated generalized Poisson (ZIGP), zero inflated double Poisson (ZIDP), zero inflated inverse trinomial (ZIIT) and zero inflated strict arcsine (ZISA) models. The most popular zero inflated models used by many researchers in published literature are zero inflated Poisson and zero inflated negative binomial followed by zero inflated generalized Poisson and zero inflated double Poisson models. ZIIT and ZISA models are newly advocated models which are not widely used and published in the literature. They are suitable for data with extra variability caused by both excess zeros and also unobserved heterogeneity. In addition, ZISA is recommended for overdispersed count data where the distribution is more than a mode. In this paper, we will discuss some published literature concerning the applications of zero inflated models including ZIP, ZINB, ZIGP, ZIDP, ZIIT and ZISA models, and provide some examples in different disciplines. Section II discusses the applications of various zero inflated models. Section III describes the properties of ZIP, ZINB, ZIGP, ZIDP, ZIIT and ZISA models. Section IV explains some model selection methods. A short summary is given in Section V.

Y. N. Phang is with Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Kampus Bandaraya Melaka, Malaysia (e-mail: phang@melaka.uitm.edu.my)

E. F. Loh is with Academy of Language Studies, Universiti Teknologi MARA, Kampus Bandaraya Melaka, Malaysia (e-mail: david_loh@ melaka.uitm.edu.my

II. Applications of Zero Inflated Models

The applications of Poisson model on count data are based on the assumption that the mean and variance are the same. However, in real life, most of the collected count data are found with variation bigger than the mean. There are many contributing factors to the extra variability. The more common one is due to the unobserved heterogeneity where usually these are the data with long tails. Negative binomial model is recommended in modeling overdispersed count data because it is able to accommodate variance with quadratic function. Generalize Poisson, Poisson inverse Gaussian, inverse trinomial and strict arcsine models are suggested for modeling data where the variance is a cubic function of mean.

Another factor which contributes to the extra variability is the occurrence of extra zeros which results in the observed zeros more than the expected one. Zero inflated models are popularly used to model data with excess zero. They are mixture models that combine a count component and a point mass at zero. Zero inflated models take into consideration the structural zeros and zeros which exist by chance. An overview of count data in econometrics including zero inflated models is provided in [2], [3]. Zero inflated Poisson model is applied when the count data possess the equality of mean and variance. For data with heavy zeros and long tails, ZINB, ZIDP, ZIPG, ZIIT and ZISA are suggested. These models are particularly suitable for data with excess zeros and variances with quadratic functions or cubic functions.

ZIP model was first introduced by [4] by considering an extra zero generating mechanism in a manufacturing process that switches between a perfect state and an imperfect state. Zero inflated negative binonial is designed to model data with population heterogeneity which may be caused by the occurrence of excess zeros and the overdispersion due to unobserved heterogeneity. Many studies show that ZINB model provides a better fit to the overdispersed count data when ZIP is inadequate. ZIP and ZINB models have been widely adopted by many researchers in various disciplines

which involve psychological research [5]; outcome from HIV risk reduction intervention [6], fitting the count data that are encountered in the field of biometrical, environmental, social sciences and transportation engineering [7]; demand of cigarettes in Turkey [8]; application to falls count data [9] and [10]; insurance claim [11], investigation of young drivers' motor vehicle crashes [12] and examination of sudden infant death syndrome [13]. Other related studies are found in [14]-[22].

GP, IT, SA models are discrete models with cubic variance functions. Therefore, ZIGP, ZIIT and ZISA are recommended for modeling data with too many zeros and heavy tails. It has been proven that they are able to incorporate overdispersion and excess zeros. ZIGP, ZIIT, and ZISA models always serve as alternatives models when the ZIP models were inadequate and the ZINB cannot be fitted to the data sets. ZIGP model has been used to model domestic violence data with too many zeros [23]. The applications of ZIGP to some overdispersed count data are reported in [24]-[26]. ZIIT and ZISA models are advocated by Phang and Loh and discussed in [27], and [28] but yet to be researched and used extensively by researchers. These two models are proven to be able to serve as alternative models in modeling excess zeros and overdispersed count data while other existing models are not able to provide a good fit. The examples are provided in [27], [28].

III. Properties of Zero Inflated Models

Zero inflated models are mixture models that combine a count component and a point mass at zero. If Y is an independent random variable having a zero-inflated Poisson distribution, the zeros are assumed to occur in two ways corresponding to distinct underlying states. The first state occurs with probability ω and the other state occurs with probability 1-ω and lead to a standard Poisson count. The zeros from the first state are called structural zeros and from the Poisson distribution are called sampling zeros [29].

The probability mass function (pmf) for ZIP model is given by

$$P_{ZIP}(Y = 0) = \omega + (1-\omega)\Pr_{Poi}(K = 0)$$

$$P_{ZIP}(Y = y) = (1-\omega)\Pr_{Poi}(K = y),\ y = 1,2,3,\ldots$$

where

$$\Pr_{Poi}(K = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

The pmf for ZINB, ZIGP, ZIDP, ZIIT and ZISA can be specified in the same way as the zero inflated Poisson model by simply replacing the Poisson distribution with the relevant count distributions such as NB, GP, IT and SA. The pmf for NB, GP, DP, IT and SA models are presented in Table I.

TABLE I
Probability Mass Function for NB, GP, DP, IT and SA

Models	Pr(K=y)
NB	$\binom{y+k-1}{y}t^{k}(1-t)$ where $t=\frac{k}{k+\lambda}$
GP	$\frac{(1+\alpha y)^{y-1}}{y!}\frac{\left(\lambda e^{-\alpha\lambda}\right)^{y}}{e^{\lambda}}$
DP	$\left(\theta^{1/2}e^{-\theta\lambda}\right)\left(\frac{e^{-y}y^{y}}{y!}\right)\left(\frac{e^{y}}{y}\right)^{\theta y}$
IT	$\frac{\lambda p^{\lambda}q^{y}}{y+\lambda}\sum_{t=0}^{\lfloor y/2\rfloor}\binom{y+\lambda}{t,t+\lambda,y-2t}\left(\frac{pr}{q^{2}}\right)^{t}$ for
	y = 0,1,2,…., where $\lambda>0$, $p\geq r$ and $p+q+r=1$
	$\binom{y+\lambda}{t,t+\lambda,y-2t}=\frac{(y+\lambda)!}{t!(t+\lambda)!(y-2t)!}$
SA	$\frac{A(y;\alpha)}{y!}p^{y}\exp\{-\alpha\arcsin(p)\}$
	where $0<\alpha$, $0<p<1$, and $A(x;\alpha)$ is defined as
	$A(x,\alpha)=\begin{cases}\prod_{k=0}^{z-1}(\alpha^{2}+4k^{2}) & \text{if } x=2z \text{ and } A(0,\alpha)=1\\ \alpha\prod_{k=0}^{z-1}(\alpha^{2}+(2k+1)^{2}) & \text{if } x=2z+1 \text{ and } A(1,\alpha)=\alpha\end{cases}$

IV. Model Selection Methods

Model selection methods like Akaike information criterion (AIC) [30], Bayesian information criterion (BIC) [31], likelihood ratio test (LRT) and Vuong tests [32] are commonly used for model comparison especially among the non-nested models. Nested models are referring to situations where one model is a special case of another. Score test and likelihood ratio test are recommended for selecting nested models. All the zero inflated models discussed in this paper are categorized as non-nested models. AIC, BIC and LRT are model selection methods based on the log-likelihood. LRT is advocated by [33] and has become one of the most popular methods for testing restrictions on a statistical model. AIC and BIC are based on information theory which are commonly used in research because of its philosophical and computational advantages. In general, the smaller the AIC and BIC, the better is the model. The AIC and BIC are defined as follows:

$$\text{AIC} = -2\log\text{likelihood} + 2k$$

$$\text{BIC} = -2\log\text{likelihood} + k\ln(n)$$

where k = number of parameters and n = number of observations.

A model comparison among ZIP, ZIGP and ZIGP regression using AIC statistics and Vuong tests is discussed in [25]. AIC and BIC are adopted in comparing the suitability of

various models, including ZIP and ZINB for analysis of insect count data [34]. Xia [6] introduced the appropriate model-fit indices for comparing the performance of competing models, using data from a real study on HIV prevention intervention by employing Vuong test, AIC and BIC. Vuong's test is a suitable approach to be used in comparing models whether they are nested, overlapping or nonnested [32].

V. Summary

There are various zero inflated models found in the literature which have been advocated to model data which are overdispersed and with extra zeros. These data are found in various disciplines from public health, economics, epidemiology, psychology, sociology, political sciences, agriculture, species abundance and road safety. ZIP is suggested for data with the occurrence of structural zeros but the count component possesses equality of mean and variance. For data with too many zeros and heavy tails, we recommend ZINB, ZIGP, ZIDP, ZIIT, and ZISA models. The model comparison methods proposed and applied by many researchers are AIC, BIC and Vuong's test. These methods are found to be suitable in comparing nested and nonnested models.

Acknowledgment

This research is supported by the Fundamental Research Grant Scheme (FRGS), Ministry of Higher Education Malaysia, that is managed by the Research Management Institute, Universiti Teknologi MARA (600-RMI/ST/FRGS 5/3/Fst (217/2010).

References

[1] M. Ridout, C. G. B. Demetrio, and J. Hinde, "Models for count with many zeros", in: Invited Paper Presented at the 19th International Biometric Conference, CapeTown, South Africa, 1998, 178.

[2] A. C. Cameron and P. K. Trivedi, "Regression analysis of count data". Cambridge University Press. 1998

[3] A. C. Cameron and P. K. Trivedi, "Microenometrics: Methods and Applications". Cambridge University Press. 2005

[4] D. Lambert, "Zero-inflated Poisson regression, with an application to random defects in manufacturing". Technometrics, 34, 1992, 1-14

[5] L. Tom, M. Beatrijs, and D.S. Olivia, "The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression", British Journal of Mathematical and Statistical Psychology, 65, 163-180.

[6] Y. Xia, M. Dianne, J. Ma, C. Feng, C. Wendy, and X. Tu, "Modeling count outcomes from HIV risk reduction interventions: A comparison of competing statistical models for count responses", AIDS Research and Treatment, Vol. 2012, 1-11.

[7] B. M. Golam Kibria, "Applicaations of some discrete regression models for count data", Pakistan Journal of Statistics and Operation research, Vol11 No. 1, 2006, 1-16.

[8] A. Bilgic, W. J. Florkowski, and C. Akbay, "Demand for cigarettes in Turkey: an application of count data models", Empir, 39, 2010, 733-765.

[9] A. Khan, S. Ullah, and J. Nitz, J. (2011). "Statistical modelling of falls count data with excess zeros. Journal of the International society for Child and Adolescent Injury Prevention [Inj Prev]", 17(4), 2011, 266-270.

[10] S. Ullah, C. F. Finch, and L. Day, "Statistical modelling for falls count data". Accident Analysis and Prevention [Accid Anal Prev], 42(2), 2010, 384-392.

[11] K. K. W. Yau and K. C. H. Yip, "On modeling claim frequency data in general insurance with extra zeros". Insurance: Mathematics and Economics Vol. 36, Issue 2, 2005, 153-163.

[12] A. H. Lee, M. R. Stevenson, K. Wang, and K. K. W. Yau, "Modelling young driver motor vehicle crashes: data with extra zeros", Accident Analysis Prevention, 34, 2002, 515-521.

[13] M. L. Dalrymple, I. L. Hudson, and R. P. K. Ford, "Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS", Computational Statistics & Data Analysis, 41, 2003, 491-504

[14] A. C. Mehmet, "Zero-inflated regression models for modeling the effect of air pollutants on hospital admissions", Polish Journal of Environment Studies, Vol. 21, No. 3, 2012, 565-568.

[15] R. Winkelmann, Econometric Analysis of Count Data. Springe Verlag, Berlin, Heidelberg, 2008.

[16] R. Winkelmann, "Health care reform and the number of doctor visits – An econometric analysis," Journal of Applied Econometrics 19, 2004, 455-472

[17] K. K. W. Yau, K. Wang, and A. H.and Lee, "Zero-inflated negative binomial mixed regression Modeling of overdispersed count data with extra zeros". Biometrical Journal, 45, 4, 2003, 437-452.

[18] S. Gurmu and P. K. Trivedi, "Excess zeros in count models for recreational trips", Journal of Business and Economic Statistics, 14, 1996, 469-477.

[19] D. B. Hall, "Zero inflated Poisson and binomial with random effects: a case study," Biometrics, 56, 2000, 1030-1039

[20] D. Bohning, E. Dietz, P. Schlattman, L. Mendonca and U. Kirchner, "The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology". Journal of the Royal Statistical Society, Series A, 1999, 162-209

[21] P. J. W. Carrivick, A. H. Lee, and K. K. W. Yau, "Zero inflated Poisson modeling to evaluate occupational safety interventions", Safety Science, 41, 2002, 53-63.

[22] A. H. Welsh, R. B. cunningham, C. F. Donnelly and D. B. Lindenmayer, "Modelling the abundance of rare species: statistical models for counts with extra zeros", Ecolog Modell, 88, 1996, 297-308.

[23] F. Famoye and P. S. Karan, "Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data," J of Data Science 4, 2006, 117-130.

[24] Z. Yang, J. W. Hardin, and C. L. Addy, "Score test for Zero inflation in overdispersed count data", Cummunication in Statistics – Theory and Methods, 39, 2010, 2008-2030.

[25] C. Czado, V. Erhardt, A. Min, and S. Wagner, "Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates", Statiscal Modelling, 7, 2, 2007, 125-153.

[26] P. L. Gupta, R. C. Gupta, and R. C. Tripathi, Score test for zero inflated generalized model", Communication in Statistics – Theory and Methods, Vol.33, No.1, 2004, 47-64.

[27] Y. N. Phang, "Statistical inference for a family of discrete distribution with cubic variance functions", Unpublished PhD thesis, University Malaya, Malaysia, 2007

[28] Y. N. Phang, and E. R. Loh. Proceedings: IASC 2008: Joint Meeting of 4th World Conference of the IASC and 6th Conference of the IASC and 6th conference of the Asian Regional Section of the IASC on Computational Statistic and Data Analysis, Yokohama, Japan, 2008

[29] N. Jasakul, and P. H. John, " Score tests for extra-zero models in zero-inflated negative binomial models", Communications in Statistics-Simulation and Computation, 38, 2009, 92-108.

[30] H. Akaike, "A new look at the statistical model identification". *IEEE* Transaction on Automatic Control, 19(6), 1974, 716-724.

[31] G. Schwarz, "Estimating the dimensions of a model", Annals of Statistics, 6, 1978, 461-464

[32] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses", Econometrica, 57, 1989, 307-333.

[33] J. Neyman, and E. S. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference", Biometrika, 20, 1928, 175-240.

[34] G. Sileshi, "Selecting the right statistical model for analysis of insect count data by using information theoretic measures', Bulletin of Entomological Research, 96, 2006, 479-488.