International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:8, No:2, 2014

Statistical Analysis for Overdispersed Medical Count Data

Y. N. Phang, E. F. Loh

***Abstract*—**Many researchers have suggested the use of zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) models in modeling overdispersed medical count data with extra variations caused by extra zeros and unobserved heterogeneity. The studies indicate that ZIP and ZINB always provide better fit than using the normal Poisson and negative binomial models in modeling overdispersed medical count data. In this study, we proposed the use of Zero Inflated Inverse Trinomial (ZIIT), Zero Inflated Poisson Inverse Gaussian (ZIPIG) and zero inflated strict arcsine models in modeling overdispered medical count data. These proposed models are not widely used by many researchers especially in the medical field. The results show that these three suggested models can serve as alternative models in modeling overdispersed medical count data. This is supported by the application of these suggested models to a real life medical data set. Inverse trinomial, Poisson inverse Gaussian and strict arcsine are discrete distributions with cubic variance function of mean. Therefore, ZIIT, ZIPIG and ZISA are able to accommodate data with excess zeros and very heavy tailed. They are recommended to be used in modeling overdispersed medical count data when ZIP and ZINB are inadequate.

***Keywords*—**Zero inflated, inverse trinomial distribution, Poisson inverse Gaussian distribution, strict arcsine distribution, Pearson's goodness of fit.

I. Introduction

MANY medical count data are found overdispersed with variance bigger than the mean. In view of this, the commonly used Poisson model is no longer appropriate to be used in analysis. Negative binomial is recommended for overdispersed count data with extra variations caused by unobserved heterogeneity. For data with extra variations caused by the occurrence of extra zeros only, Zero Inflated Poisson (ZIP) is proposed assuming that the zeros come from structural zeros and sampling zeros. However, for overdispersed count data where the extra variations are contributed by the existence of extra zeros and unobserved heterogeneity, Zero Inflated Negative Binomial (ZINB) is recommended. Hurdle models are introduced when the zeros are contributed by structural zeros only. Sampling zeros are observed in the usual count distribution (Poisson, NB, etc) assuming that they occurred by chance. Structural zeros are observed due to some specific structure in the data. Hu et al. [1] provide a good explanation, and examples on sampling zeros and structural zeros. Poisson hurdle (PH) and negative binomial hurdle (NBH) are commonly applied in modeling health care utilization data with the assumption that the initial process brings individuals into the at risk population [2]-[4]. Baughman [5] provides a good discussion in deciding whether a zero inflated or hurdle model is appropriate for a given data set requires close collaboration with subject matter experts.

Rose et al. [6] applied Poisson, NB, ZIP, ZINB, PH and NBH models to a real life data set correspond to 4020 observed systemic adverse events for four injections for each of the 1005 study participants and found that ZINB and NBH provide the best fit among all the models. Results of the Akaike Information Criterion model selection method are consistent with the Pearson's chi-square goodness of fit suggesting that ZINB and NBH are the most suitable models in modeling this data set. Hu et al. [7] model count outcomes from HIV risk reduction intervention by comparing the competing models which involved, Poisson, NB, ZIP and ZINB claiming that ZIP and ZINB are not widely and effectively used in sex health research, especially in HIV prevention intervention and related studies.

ZINB is found to be able to address issues involving extra zeros and unobserved heterogeneity. Dwivedi et al. [8] applied the various models to compare and test the ability of these models in predicting the number of involved nodes in breast cancer patients and found that ZINB and HNB are able to account for the excess zeros and provide better prediction in comparison with Poisson, NB, ZIP and HP. However, ZINB is recommended in this study due to the nature of the data assuming that the zeros come from a 'high risk' group. Lee et al. [9] illustrate the use of Poisson, NB, ZIP and ZINB for overdispersed count data dealing with the number of incidents involving human papillomavirus infection (HPV). ZIP and ZINB are widely used to analyze dental caries with many zeros [10]-[12]. Most of the studies show that ZINB models are able to provide better fit than ZIP especially in modeling data with extra variations coming from unobserved heterogeneity and excess zeros.

So far, ZINB model has been commonly used and suggested for modeling overdispersed medical count data where the ZIP model is no longer adequate and not able to accommodate the extra variability caused by excess zeros and unobserved heterogeneity. Not many studies are found in the literature, which discuss the use of the appropriate statistical models for overdispersed medical count data with very heavy tailed where ZIP and ZINB models are inappropriate. Recently, Ahmad [13] introduced zero inflated generalized Poisson model together with ZIP and ZINB in modeling and

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).

handling overdispersion count data among children with Thalassemia disease.

In this study, we proposed the use of Zero Inflated Inverse Trinomial (ZIIT), Zero Inflated Poisson Inverse Gaussian (ZIPIG) and zero inflated strict arcsine (ZISA) models in modeling overdispersed medical count data. Inverse Trinomial (IT), Poisson Inverse Gaussian (PIG), and Strict Arcsine (SA) models are generalized Poisson distribution which are natural extension of standard Poisson distribution. We applied these three models to a real life medical count data. Pearson's goodness of fit is used in assessing the general fit of the model. Section II discusses the characteristics of inverse trinomial, Poisson inverse Gaussian and strict arcsine models. Section III shows the characteristics of zero inflated models. Section IV explains the application of ZINB, ZIIT, ZIPIG and ZISA models to an overdispersed medical count data set. Section V provides the results and concluding remarks.

II. Properties of the Distributions

A. Inverse Trinomial Distribution

The inverse trinomial distribution may be derived from the Lagrangian expansion. The inverse trinomial is so named because its cumulant generating function is the inverse of that for the trinomial distribution [14].

B. Poisson Inverse Gaussian Distribution

The Poisson-inverse Gaussian distribution is a mixed Poisson distribution derived from the Poisson distribution using the inverse Gaussian as a mixed distribution. It has received much attention in modeling overdispersed count data such as species abundance data which are usually characterized by extremely long tails. The theory and applications of PIG distribution are discussed in [15]-[20].

C. Strict Arcsine Distribution

The strict arcsine distribution is introduced by [21]. Kokonendji and Khoudar [22] studied the properties of the SA distribution and found that the SA distribution is overdispersed, skewed to the right and leptokurtic. In addition, SA is a Poisson mixture. It can be obtained by introducing extra Poisson variation into a Poisson model.

III. Zero inflated models

Zero-inflated count models provide a way of modeling the excess zeros in addition to allowing for overdispersion. There are two data generating processes where process 1 generates only zeros with probability ω, and process 2 generates count from a statistical distribution such as Poisson, NB, IT, PIG or SA with probability 1-ω. In general:

$$y \sim \begin{cases} 0 & \text{with probability } \omega \\ g(y) & \text{with probability } 1-\omega \end{cases}$$

The probability of {Y=y} is

$$P_{ZI}(Y=y) = \begin{cases} \omega + (1-\omega)\,g(y) & \text{if } y = 0 \\ (1-\omega)\,g(y) & \text{if } y > 0 \end{cases}$$

where g(y) is the probability mass function for NB, IT, PIG or SA. Probability mass function for NB, IT, PIG and SA is given in Table I.

TABLE I
Probability Mass Function for NB, IT, PIG and SA

Models	*Pr(K=y)*
NB	$\binom{y+k-1}{y} t^k (1-t)$ where $t = \frac{k}{k+\lambda}$
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 $y = 0,1,2,\ldots$, where $\lambda>0$, $p \geq r$ and $p + q + r = 1$ $\binom{y+\lambda}{t, t+\lambda, y-2t} = \frac{(y+\lambda)!}{t!(t+\lambda)!(y-2t)!}$
PIG	$\frac{m_y^*}{y!} \exp\{(\nu/\mu) - (\nu/\mu_*)\}, \quad y = 0, 1, 2, \ldots..$ where $\mu_* = [\mu^{-2} + 2\nu^{-1}]^{-1/2}$ and $m_{y+1}^* = \mu_*^2 \{m_{y-1}^* + (2r-1) m_y^* / \nu\}$, $y = 0, 1, 2, \ldots..$ given that $m_0^* = 1$, $m_1^* = \mu_*$
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. Applications and Parameter Estimations

We applied ZIIT, ZIPIG and ZISA to a vaccine adverse event count data [6]. The data are the frequencies which correspond to 4020 observed systemic adverse events for four injections for each of the 1005 study participants. Maximum likelihood estimation method is used in estimating the parameters for all the suggested models. Person's chi square of fit test is used in assessing the general fit of the models. The null hypothesis is that the model fits the data and the alternative hypothesis is that the model does not fit the data. All the four models yield very small chi-square value which leads to the acceptance of the null hypothesis. The results indicate that all the suggested models provide a good fit to the data. Table II provides the actual and predicted frequencies by models and goodness of fit results for fitted models.

TABLE II
THE ACTUAL FREQUENCIES CORRESPOND TO 4020 OBSERVED SYSTEMIC ADVERSE EVENTS FOR FOUR INJECTION FOR EACH OF TH E1005 STUDY PARTICIPANTS

	Actual	Expected frequency			
		ZINB	ZIIT	ZIPIG	ZISA
0	1437	1437	1437	1437.01	1437
1	1010	1001.94	999.06	988.56	980.25
2	660	684.02	687.96	703.61	721.32
3	428	411.69	412.73	416.24	415.49
4	236	231.15	230.46	226.82	220.90
5	122	124.17	123.22	119.24	115.53
6	62	64.70	64.09	61.83	60.39
7	34	32.96	32.72	31.96	31.90
8	14	16.51	16.49	16.56	16.98
9	8	8.16	8.23	8.61	9.15
10	4	3.99	4.08	4.50	4.96
11	4	1.93	2.02	2.37	2.72
12	1	1.77	1.94	2.67	3.41
Pearson's χ^2		2.44	2.58	4.62	9.13

V. RESULTS AND DISCUSSION

ZINB exhibits the best fit (χ^2 =2.44, p>0.05) among the fitted models followed by ZIIT (χ^2 =2.58, p>0.05). ZIPIG (χ^2 =4.62, p>0.05) and ZISA (χ^2 =9.13, p>0.05) also indicate good fit. ZINB is parsimonious in comparison with ZIIT, ZIPIG and ZISA. ZIIT, ZIPIG and ZISA are models with four parameters and they are more complicated than ZINB. In this study, ZINB appears as the best choice of model in fitting this vaccine adverse event count data due to its smallest chi-square value which indicates the best fit. However, the results suggest that ZIIT, ZIPIG and ZISA models can serve as alternative models in modeling overdispersed medical count data considering the extra variations come from the unobserved heterogeneity and excess zeros. They are recommended for data which are very heavy-tailed and with preponderance of zeros which ZINB cannot accommodate. This is because NB is a distribution where the variance is a quadratic function of mean, whereas IT, PIG, and SA are distributions where the variance is a cubic function of mean.

REFERENCES

[1] M.-C. Hu, M. Pavlicova, and E. V. Nunes, "Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial," *The American journal of drug and alcohol abuse,* vol. 37, pp. 367-375, 2011.

[2] C. J. Brown, J. A. Pagán, and E. Rodríguez-Oreggia, "The decision-making process of health care utilization in Mexico," *Health Policy,* vol. 72, pp. 81-91, 2005.

[3] K. J. Krobot, J. S. Kaufman, D. B. Christensen, J. S. Preisser, W. C. Miller, and M. A. Ibrahim, "Accessing a new medication in Germany: A novel approach to assess a health insurance-related barrier," *Annals of Epidemiology,* vol. 15, pp. 756-761, 2005.

[4] T.-C. Liu and C.-S. Chen, "An analysis of private health insurance purchasing decisions with national health insurance in Taiwan," *Social science & medicine,* vol. 55, pp. 755-774, 2002.

[5] A. Baughman, "Mixture model framework facilitates understanding of zero-inflated and hurdle models for count data," *Journal of Biopharmaceutical Statistics,* vol. 17, pp. 943-946, 2007.

[6] C. E. Rose, S. W. Martin, K. A. Wannemuehler, and B. D. Plikaytis, "On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data," *Journal of Biopharmaceutical Statistics,* vol. 16, pp. 463-481, 2006.

[7] M.-C. Hu, M. Pavlicova, and E. V. Nunes, "Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial," *The American journal of drug and alcohol abuse,* vol. 37, pp. 367-375, 2011.

[8] A. K. Dwivedi, S. N. Dwivedi, S. Deo, R. Shukla, and E. Kopras, "Statistical models for predicting number of involved nodes in breast cancer patients," *Health,* vol. 2, p. 641, 2010.

[9] J. Lee, G. Han, W. Fulp, and A. Giuliano, "Analysis of overdispersed count data: application to the Human Papillomavirus Infection in Men (HIM) Study," *Epidemiology and infection,* vol. 140, pp. 1087-1094, 2012.

[10] M. S. Gilthorpe, M. Frydenberg, Y. Cheng, and V. Baelum, "Modelling count data with excessive zeros: The need for class prediction in zero-inflated models and the issue of data generation in choosing between zero inflated and generic mixture models for dental caries data," *Statistics in medicine,* vol. 28, pp. 3539-3553, 2009.

[11] S. B. Javali and P. V. Pandit, "Using zero inflated models to analyze dental caries with many zeroes," *Indian Journal of Dental Research,* vol. 21, p. 480, 2010.

[12] B. T. Pahel, J. S. Preisser, S. C. Stearns, and R. G. Rozier, "Multiple imputation of dental caries data using a zero□inflated Poisson regression model," *Journal of Public Health Dentistry,* vol. 71, pp. 71-78, 2011.

[13] W. M. A. b. W. Ahmad, "Modeling and Handling Overdispersion Health Science Data with Zero-Inflated Poisson Model," *Journal of Modern Applied Statistical Methods,* vol. 12, p. 28, 2013.

[14] K. Shimizu and T. Yanagimoto, "The inverse trinomial distribution," *Japanese Journal of Applied Statistics,* vol. 20, pp. 89-96, 1991.

[15] M. S. Holla, "On a Poisson-inverse Gaussian distribution," *Metrika*, 11, 115-121, 1966.

[16] M. Sankaran, "Mixtures by the inverse Gaussian distribution," *Sanky B,* 30, 455-458,1968

[17] H. S. Sichel, 'On a family of discrete distributions particularly suited to represent long-tailed frequency data," in N.F. Laubscher (Ed.), *Proceedings of the third Symposium on Mathematical Statistics*, Pretoria, CSIR, 51-97,1971

[18] G. E. Willmot, "The Poisson-inverse Gaussian distribution as an alternative to the negative binomial," *Scandinavian Actuarial Journal*, 113-127, 1989.

[19] J. K. Ord and G. A. Whitmore, "The poisson-inverse gaussian disiribuiion as a model for species abundance," *Communications in Statistics-theory and Methods,* vol. 15, pp. 853-871, 1986.

[20] S. H. Ong, "A note on the mixed Poisson formulation of the Poisson-inverse Gaussian distribution," *Communications in Statistics-Simulations*, 27(1), 67-78, 1998.

[21] G. Letac and M. Mora, "Natural real exponential families with cubic variance functions," The Annals of Statistics, 18, 1990, 1-37.

[22] C. C. Kokonendji and M. Khoudar, "On Strict Arcsine Distribution" Communications in Statistics. Theory Methods,33(5), 2004, pg 993-1006.