DIGITALCOMMONS
— @WAYNESTATE —

Journal of Modern Applied Statistical Methods

Volume 12 | Issue 1 Article 28

5-1-2013

Modeling and Handling Overdispersion Health Science Data with Zero-Inflated Poisson Model

Nur Syabiha binti Zafakali
Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia

Wan Muhamad Amir bin W Ahmad
Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation

Zafakali, Nur Syabiha binti and Ahmad, Wan Muhamad Amir bin W (2013) "Modeling and Handling Overdispersion Health Science Data with Zero-Inflated Poisson Model," *Journal of Modern Applied Statistical Methods*: Vol. 12 : Iss. 1 , Article 28.
DOI: 10.22237/jmasm/1367382420
Available at: http://digitalcommons.wayne.edu/jmasm/vol12/iss1/28

This Emerging Scholar is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

Journal of Modern Applied Statistical Methods
May 2013, Vol. 12, No. 1, 255-260

Copyright © 2013 JMASM, Inc.

1538 – 9472/13/$95.00

Emerging Scholars

Modeling and Handling Overdispersion Health Science Data with Zero-Inflated Poisson Model

Nur Syabiha Zafakali Wan Muhamad Amir W Ahmad
Universiti Malaysia Terengganu,
Kuala Terengganu, Malaysia

Health sciences research often involves analyses of repeated measurement or longitudinal count data analyses that exhibit excess zeros. Overdispersion occurs when count data measurements have greater variability than allowed. This phenomenon can be carried over to zero-inflated count data modeling. Referred to as zero-inflation, the Zero-Inflated Poisson (ZIP) model can be used to model such data. The Zero-Inflated Negative Binomial (ZINB) model is used to account for overdispersion detected in count data. The ZINB model is considered as an alternative for the Zero-Inflated Generalized Poisson (ZIGP) model for zero-inflated overdispersed count data. Consequently, zero-inflated models have been proposed for the situations where the data generating process results are overdispersed. This study considers modeling and handling overdispersion data among children with Thalassemia disease using the ZIP, ZINB and ZIGP models.

Key words: Count data; zero-inflation models; overdispersion; Thalassemia.

Introduction

Count data with too many zeros are common in a number of applications. Ridout, et al. (1998) cited examples of data with too many zeros from various disciplines including agriculture, econometrics, species abundance, medicine and recreational facility use. Several models have been proposed to handle count data with too many zeros. Lambert (1992) described Zero-Inflated Poisson (ZIP) models with an application to defect in manufacturing. Lee, et al. (2001) generalized the ZIP model to accommodate the extent of individual exposure and Hall (2000) described the Zero-Inflated Negative Binomial (ZINB) model and incorporated random effects into ZIP and ZINB models.

The Poisson model emphasizes count data. Overdispersion implies that there is more variability around a model's fitted values than is consistent with a Poisson formulation and Poisson regression can be useful in the analyses of such data. Tsou (2006) demonstrated that a Poisson regression model could be adjusted to become asymptotically valid for inference about regression parameters, even if the Poisson assumption fails. Because positive counts may still be overdispersed with respect to the zero-truncated Poisson distribution, in the last decade Zero-Inflated Generalized Poisson (ZIGP) models have been found useful for the analyses of count data with a large amount of zero-outcomes (Famoye & Singh, 2003). The generalized Poisson model has been used to model dispersed count data. It is a good competitor to the negative binomial model when

Nur Syabiha binti Zafakali is a final year postgraduate student in the Department of Mathematics within the Faculty of Science and Technology. She holds a Bachelor's of Computational Mathematics. Her research interests include Thalassemia disease among children in East Coast States of Malaysia. Email her at: syabiha_89@yahoo.com. Wan Muhamad Amir bin W Ahmad is a lecturer of Statistics in the Department of Mathematics within the Faculty of Science and Technology. He received his doctorate in Medical Statistics. Email him at: wmamir@umt.edu.my.

the count data is over-dispersed. ZIP and ZINB models have been proposed for the situations where the data generating process results into too many zeros (Famoye & Singh, 2006).

Methodology

Zero-Inflated Poisson Model

Consider the ZIP model, which is denoted by $\Pr(Y_i = y_i)$, in which the response variable $Y_i = (1,2,...,n)$ has a probability mass function (pmf) given by

$$\Pr(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)e^{-\theta_i}, & y_i = 0, \\ (1-\omega_i)\dfrac{\theta_i^{y_i} e^{-0_i}}{y_i!}, & y_i > 0, \end{cases} \tag{1}$$

Where $0 \le \omega_i < 1$ and $\theta_i > 0$. The random variable Y_i has a Poisson (θ_i) distribution when $(\omega_i) = 0$. The parameters θ_i and ω_i depend on vectors of covariates x_i and z_i, respectively. The ZIP model is given by

$$\log(\theta_i) = x_i^t \beta, \ \log(\frac{\omega_i}{1-\omega_i}) = z_i^t \gamma \tag{2}$$

and the mean and variance ZIP model are given by

$$E(Y_i) = (1-\omega_i)\theta_i,$$
$$Var(Y_i) = (1-\omega_i)\theta_i(1+\omega_i\theta_i) \tag{3}$$

Zero-Inflated Negative Binomial Model

For zero-inflated and overdispersed data a frequent modeling choice is the Zero-Inflated Negative Binomial (ZINB) model. The response variable $Y_i (i = 1,2,...,n)$ has a pmf given by

$$Pr(Y_i = y_i) =$$
$$\begin{cases} \omega_i + (1-\omega_i)(1+\kappa\theta_i^c)^{-\theta_i^{1-c}/\kappa}, \\ y_i = 0, \\ (1-\omega_i)\dfrac{\Gamma(y_i + \theta_i^{1-c}/\kappa)}{y_i!\,\Gamma(\theta_i^{1-c}/\kappa)}(1+\kappa\theta_i^c)^{-\theta_i^{1-c}/\kappa}(1+\theta_i^{-c}/\kappa)^{-y_i}, \\ y_i > 0 \end{cases} \tag{4}$$

Where $0 \le \omega_i \le 1$ and $\theta_i > 0$, κ is the dispersion parameter with $\kappa > 0$ and $\Gamma(.)$ is the gamma function. The mean and the variance of the model are defined as

$$E(Y_i) = (1-\omega_i)\theta_i$$
$$Var(Y_i) = (1-\omega_i)\theta_i(1+\theta_i\kappa^{-1}+\omega_i\theta_i) \tag{5}$$

The response variable Y_i has a negative binomial distribution with mean θ_i and dispersion parameter κ when $\omega_i = 0$ (Garay, et al., 2011). Ridout, et al. (2001) fitted various models to these data on the basis of the Poisson and negative binomial distributions and their zero-inflated counterpart.

Zero-Inflated Generalized Poisson Model

The generalized Poisson model (ZIGP) was proposed by Consul & Famoye (1992) and Famoye (1993). They emphasized the model to count data that are affected by a number of known predictor variables. Because positive counts may still be overdispersed with respect to the zero-truncated Poisson distribution, ZIGP models have been found to be useful for the analyses of count data with a large number of zero-outcomes. The generalized Poisson model has been used to model a household fertility data set (Wang & Famoye, 1997) and to model injury data (Wulu, et al., 2002). The generalized Poisson distribution for random variable Y_i takes the pmf given by

$$f(y_i;\lambda_i,\kappa)=\frac{\lambda_i(\lambda_i+\kappa y_i)^{y_i-1}e^{-\lambda_i-\kappa y_i}}{y_i!},$$
$$y_i=0,1,2,..., \tag{6}$$

where $\lambda_i > 0$ and max $(-1,-\lambda_i/4)<\kappa<1$. The mean and variance of Y_i are

$$\mu_i=E(Y_i)=\frac{\lambda_i}{1-\kappa},$$
$$Var(Y_i)=\frac{\lambda_i}{(1-\kappa)^3}=\frac{1}{(1-\kappa)^2}E(Y_i)=\varphi \tag{7}$$

The term $\phi=1/(1-\kappa)^2$ is a dispersion parameter. When $\kappa=0$, the generalized Poisson distribution reduces to a Poisson distribution with parameter λ_i and is a case of equidispersion in the model, if $\kappa<0$, the generalized Poisson model represents count data with underdispersion, and if $\kappa>0$, generalized Poisson model represents count data with overdispersion. When there are more zero observations than expected, the generalized Poisson model will not provide good fit in general. Sampling zeros can be fitted into the ZIGP model, but not structural zeros. A good alternative to fit zero-inflated count data is a ZIGP model. The ZIGP model is defined as

$$Pr(Y_i=y_i)=$$
$$\begin{cases}\omega_i+(1-\omega_i)e^{-(1-\kappa)^{\theta_i}},\ y_i=0,\\ (1-\omega_i)((1-\kappa)\theta_i+\kappa y_i)^{y_i-1}\dfrac{(1-\kappa)\theta_i}{y_i!}exp\begin{bmatrix}(-(1-\kappa)\\ \times\theta_i-\kappa y_i)\end{bmatrix},\\ y_i>0\end{cases} \tag{8}$$

where $\theta_i>0$. The mean and variance of the ZIGP distribution are given by

$$E(Y_i)=(1-\omega_i)\theta_i$$
$$Var(Y_i)=(1-\omega_i)\theta_i[1/(1-\kappa)^2+\omega_i\theta_i]. \tag{9}$$

The parameters θ_i and ω_i depend on vectors of covariates x_i and z_i, respectively. ω_i specifies the probability of structural zero status and can be modeled using a logit link function in which

$$\log it(\omega_i)=\log\omega_i/(1-\omega_i)=z_i'\gamma \tag{10}$$

where z_i is the i^{th} row vector of the covariate matrix and γ is the parameter vector.

Results

To show the utility of the developed approach, the Zero-Inflated count models was applied to a real data set of underlying Thalassemia disease among children. Thalassemia is a genetic blood disorder in which the body makes an abnormal form of hemoglobin, the protein in red blood cells that carries oxygen. Thalassemia are common autosomal recessive disorders (Thursz, 2007). This study involved a sample of 930 for children age between 1-12 years. To build this dataset, the numbers of diagnoses among children aged 1-12 years who suffer from Thalassemia were counted. The data were collected at the Medical Record Unit in Hospital Universiti Sains Malaysia (HUSM), Kubang Kerian, Kelantan in north-east Malaysia 2005 to 2010. For the purpose of this study, the count data is used. The diagnosis is considered as the response variable and the selected variables are: disease of blood, health services, heart failure, Influenza, Anemia, Pneumonia, acute bronchitis, Asthma, Acute Tonsillitis, Jaundice and Tuberculosis.

To handle overdispersion in a zero-inflated dataset, a one-sided test was used and the level of significance was set as $\alpha=0.05$. The data analyses were performed in SAS 9.3, using *proc genmod* and *proc nlmixed*. In counting the number of responses to an exposure a patient may have no diagnosis response because of their immunity or resistance to a disease.

In the dataset, there were 930 patients and among these, 635 patients had no diagnosis (see Table 1); thus, there are 68% zero counts in the data. The overdispersion might have been due to many excess zeros for the case when $y = 0$ because 68% of observed counts are zeros. Frequency (percent) of patients who received a different type of diagnosis is a total of 125 (13.4%), while patients who received two different types of diagnosis was 95 (10.2%). For patients who received three different types of diagnosis the frequency (percent) was 58 (6.2%) and patients who received four different types of diagnosis was 17 (1.8%). Mean (standard deviation) of the variables diagnosis showed a value of 600 (1019), while the variance of 1039. Because the data has too many zeros and is over-dispersed the zero-inflation model can be applied.

The dispersion parameters of all models correspond to 0.599 for the dispersion index. This value indicates that the dispersion in the data is not large and could be corrected by the use of a model that incorporates a dispersion parameter such as ZIP, ZINB or ZIGP. The models of ZIP, ZINB and ZIGP are positively associated ($p < 0.0001$) by implementing each of the models to the data using *Proc Nlmixed* (see Table 2). The analyses of fitting zero-inflated models variable include: Disease of Blood, Health Services, heart failure, Influenza, Anemia, Pneumonia, acute bronchitis, Asthma, Acute Tonsillitis, Jaundice and Tuberculosis.

A fit of these models gives a deviance of 1137.4476 on 930 d.f. The deviance (D) is equivalent to the likelihood-ratio test statistic G^2 that is defined as:

$$G^2 = 2\sum_{i=1}^{n} y_i \log(\frac{y_i}{\mu_i}) \qquad (11)$$

Consequently, a test of overdispersion of the data, which is measured by the ratio of deviance/d.f. = 1.233. Because this value is greater than 1, there is strong evidence that the data is overdispersed and therefore $E(Y_i) \neq Var(Y_i)$. All three models fit the data well with corresponding log-likelihoods of: −559.8034 (ZIP), −568.7238 (ZINB) and −558.3192 (ZIGP) obtained from output *ProcGenmod* (see Table 3). Clearly, the ZIGP model seems to fit the data best as it has the smallest Akaike Information Criterion (AIC=1148.6384). Although the ZIP, ZINB and ZIGP all fit the data well, the effect of having a diagnosis in patients does not appear to have any significant effect on the number of diagnoses under these models.

Conclusion

This article focused on handling overdispersion data that involves zero-inflation models. Overdispersion can be modeled when counts show more variability than previously assumed models. However, the consideration of zero-inflation for sample size less than 50 is not encouraged, it is recommended to evaluate scores for sample sizes ≥ 100. The parametric bootstrap method is recommended for sample size between 50 and 100 for a reliable performance (Jung, Jhun & Lee, 2005).

Three different methods were used in this study: (i) ZIP model, (ii) ZINB model and (iii) ZIGP model with covariate dependence. The ZIP model described in Lambert's (1992) seminal work provides a sufficient fit to data when overdispersion in raw data is caused by zero-inflation. The ZINB model should be considered if data continue to suggest additional overdispersion (overdispersion can be the result of excess zeroes): The ZIP model is not appropriate for these data, because the Poisson model does not accommodate the remaining overdispersion and not accounted for through zero-inflation. The ZIGP model is applied in different fields to model zero-inflated and overdispersed data. The ZIGP could provide a better fit than the ZINB when there is a large zero-fraction; this implies that the ZIGP model could be a reasonable alternative to the ZINB model.

It is surprising that in all these models, it appears that the ZIGP model provided a good fit to the data because it had the smallest value of Akaike Information Criterion (AIC). Although the ZIGP model appears to be a good competitor to the ZIP and ZINB models, it is unknown under what conditions, if any, which model would perform best.

Table 1: Diagnosis Count from 930 Thalassemia Patients in HUSM

Diagnosis	Patients	
	Frequency	Percent (%)
0	635	68.3
1	125	13.4
2	95	10.2
3	58	6.2
4	17	1.8
Total	930	100

Table 2: Summary of Zero-Inflated Model Fit

Variable	Estimate (Standard Error)		
	ZIP	ZINB	ZIGP
Disease of Blood	1.0144 (0.3594)	0.4763 (0.3546)	0.4737 (0.3594)
Health Services	1.2116 (0.0883)	0.5075 (0.0722)	0.5075 (0.0925)
Heart failure	0.9702 (0.1048)	0.5222 (0.1009)	0.5222 (0.1012)
Influenza	0.9429 (0.1852)	0.5852 (0.1781)	0.5852 (0.1783)
Anemia	0.7778 (0.1004)	0.4996 (0.0873)	0.4997 (0.0944)
Pneumonia	0.9062 (0.1147)	0.5174 (0.1142)	0.5173 (0.1143)
Acute bronchitis	0.9116 (0.0913)	0.5072 (0.0713)	0.5071 (0.0904)
Asthma	1.1563 (0.0939)	0.5869 (0.0910)	0.5869 (0.0963)
Acute Tonsillitis	0.7234 (0.1743)	0.3794 (0.1743)	0.3794 (0.1746)
Jaundice	1.2890 (0.3675)	0.5771 (0.3823)	0.5767 (0.3888)
Tuberculosis	0.6550 (0.3857)	0.3791 (0.3702)	0.3789 (0.3702)

References

Famoye, F., & Singh, K. P. (2003). On inflated generalized Poisson models. *Advances in Applied Statistics*, *2*, 145-158.

Ridout, M., Hinde, J., & Dem´etrio, C. G. B. (2001). A score test for testing a ZIP model against ZINB alternatives. *Journal of Biometrics*, *57*, 219-223.

Ridout, M., Demetrio, C. G. B., & Hinde, J. (1998). *Models for count data with many zeros*. Invited paper presented at the Nineteenth International Biometric Conference, Cape Town, South Africa, 179-190.

Lambert, D. (1992). ZIP with an application to defects in manufacturing. *Journal of Technometrics*, *34*, 1-14.

Table 3: Parameter Estimates, Standard Errors (in parentheses) for All Models

Parameters	ZIP	ZINB	ZIGP
a_0	-17.4510 (246.90*)	3.9808 (0.4553*)	3.9811 (0.4615*)
b_0	-1.9236 (0.08624*)	-0.5294 (0.0517*)	-0.5294 (0.1125*)
Log likelihood	-569.8034	-568.7238	-558.3192
AIC[a]	1151.6069	1153.6069	1148.6384
BIC[b]	1228.9698	1235.8050	1226.0013
Pearson Chi-Square χ^2	416.0162		
Deviance	393.0901		
d.f	918		

*significant at level 0.05, a = Akaike's information criteria (AIC), b = Bayesian information criterion

Hall, D. B. (2000). ZIP and binomial with random effects: A case study. *Journal of Biometrics*, *56*, 1030-1039.

Lee, A. H., Wang, K., & Yau, K. K. W. (2001).Analyses of ZIP data incorporating extent of exposure. *Journal of Biometrics*, *43*, 963-975.

Famoye, F. (1993). Restricted generalized Poisson model. *Communications in Statistics, Theory and Methods*, *22*, 1335-1354.

Consul, P. C., & Famoye, F. (1992). Generalized Poisson model. *Communications in Statistics, Theory and Methods*, *21*, 89-109.

Wang, W., & Famoye, F. (1997). Modeling household fertility decisions with generalized Poisson. *Journal of Population Economics*, *10*, 273-283.

Wulu, J. T., Singh, K. P., Famoye, F., & McGwin, G. (2002). Analyses of count data. *Journal of the Indian Society of Agricultural Statistics*, *55*, 220-231.

Garay, A. M., Hashimoto, E. M., Ortega, E. M. M., & Lachos, V. H. (2011). On estimation and influence diagnostics for ZINB models. *Journal of Computational Statistics and Data Analyses*, *55*, 1304-1318.

Thursz M. (2007). Iron, hemochromatosis and thalassaemia as risk factors for fibrosis in hepatitis C virus infection. *Journal of Gut*, 56, 613-614.

Jung, B. C., Jhun, M., & Lee, J. W. (2005). Bootstrap tests for overdispersion in a zero-inflated Poisson model. *Journal of Biometrics*, *61*, 626-629.

Tsou, T. S. (2006). Robust Poisson regression. *Journal of Statistical Planning and Inference*, *136*, 3173-3186.