Article ID: 1000-5641(2010)01-0017-07

Zero-truncated generalized Poisson regression model and its score tests

ZHAO Wei-hua[1], FENG Yu[2], LI Ze-an[1]

(1. *School of Sciences, Nantong University, Nantong Jiangsu* 226007, *China;*
2. *Nanjing University of Science and Technology, Nanjing* 210094, *China*)

Abstract: This paper extended the ZTP regression model to Zero-truncated generalized Poisson regression model. An algorithm for estimating parameters was obtained and two score tests were presented for testing the ZTP regression model against the ZTGP regression model, and for testing the significance of regression coefficients. A numerical example was given to illustrate our method and the power of score tests was investigated by Monte Carlo simulation.

Key words: generalized Poisson regression; zero-truncated; score test
CLC number: O212 **Document code:** A

零截尾广义 Poisson 回归模型及其 Score 检验

赵为华[1]， 冯　予[2]， 李泽安[1]

(1. 南通大学 理学院, 江苏 南通 226007; 2. 南京理工大学, 南京 210094)

摘要: 研究零截尾广义Poisson回归模型, 给出了模型的参数估计方法, 着重研究检验零截尾Poisson回归、零截尾广义Poisson回归以及检验回归系数显著性的score检验统计量, 并用一个数值实例来说明方法的有效性. 最后通过一个仿真模拟例子来研究score检验统计量的检验功效.

关键词: 广义Poisson回归; 零截尾; score检验

0 Introduction

For a random variable y representing counts where sample mean and sample variance are almost equal, the Poisson model is the standard approach to analysis. Quite often, count data exhibit substantial variations where the sample variance is either smaller or larger than the sample mean and it is classified as under- or over-dispersion, respectively. Various models and associated estimation methods have been proposed to deal with these dispersions, including

收稿日期: 2009-05
基金项目: 南通大学自然科学基金(08Z032); 江苏省自然科学基金(SBK200920379)
第一作者: 赵为华, 男, 讲师, 主要从事数理统计方面的研究. E-mail: zhaovenus@hotmail.com.

negative binomial models, mixed Poisson models, generalized Poisson models, hurdle Poisson models, Censored Poisson models and inflated Poisson models[1−10].

Meanwhile, we often find in some situations that count data not only has no zeros but also structurally excludes having 0 counts. For example, there is no zero length of hospital stay and zeros counts are structurally excluded from lengths of stay. Therefore, zero-truncated Poisson model has been used to analysis the hospital length of stay data[3,8]. In this paper, as a supplement of the above work on count regression models and score tests, our interest is to study zero-truncated generalized Poisson model(ZTGP) since it is a natural extension of the ordinary Poisson model.

The outline of the paper is as follows: In Section 1, we introduce the ZTGP regression model and its estimation method. Score tests for dispersion and regression parameters in the model are developed in Section 2. Section 3 presents an example to illustrate our methodology. A simulation study for powers of score test statistics will be presented in Section 4 and some conclusions are given in the last section.

1 Zero-truncated generalized Poisson regression model and estimation

Consider the generalized Poisson distribution with probability mass function[11].

$$f(y;\lambda)=\frac{1}{y!}\bigg(\frac{\lambda}{1+\alpha\lambda}\bigg)^{y}(1+\alpha y)^{y-1}\mathrm{exp}\bigg\{-\frac{\lambda(1+\alpha y)}{1+\alpha\lambda}\bigg\}, \tag{1.1}$$

with $y=0,1,2,\cdots$. In this distribution, the mean and variance of the distribution are, respectively, λ and $\lambda(1+\alpha\lambda)^2$, and the parameter α can be interpreted as a dispersion parameter. It is easily seen that $\alpha=0$ indicates the presence of equality of mean and variance (equi-dispersion), then the probability function in (1.1) reduces to the Poisson distribution, while $\alpha>0$ is over-dispersion and $\alpha<0$ is under-dispersion in the generalized Poisson distribution. Whenever $\alpha<0$, the value of α is such that $1+\alpha\lambda>0$ and $1+\alpha y>0$ so that the probability in(1.1) is non-negative. For more details, the reader is referred to Consul[11], Consul and Famoye[2].

At the same time, we often find that our count response model fails to have 0 counts. That is, we discover not only that the count has no zeros but also that it structurally excludes having 0 counts. A standard method of dealing with count models that excludes zero counts is to use a model typically referred to as a zero-truncated count model[8]. We do this by determining the formula to calculate 0 counts, subtract it from one, and then divide the count distribution probability function by the resultant value. Then the zero-truncated generalized Poisson distribution can be expressed as

$$f(y;\lambda|y>0)=\frac{1}{y!(\exp(\lambda/(1+\alpha\lambda))-1)}\bigg(\frac{\lambda}{1+\alpha\lambda}\bigg)^{y}(1+\alpha y)^{y-1}\mathrm{exp}\bigg\{-\frac{\alpha\lambda y}{1+\alpha\lambda}\bigg\}, \tag{1.2}$$

with $y=1,2,\cdots$. The model (1.2) is denoted by $ZTGP(\alpha,\lambda)$. When $\alpha=0$, the distribution reduces to zero-truncated Poisson model.

Assume that each observation $y_i, i = 1, 2, \cdots, n$ submits to zero-truncated generalized Poisson distribution, i.e., $y_i \sim ZTGP(\alpha, \lambda_i)$. Following the generalized linear model approach, we relate parameters λ_i to covariates $x_i \in R^p$ through the log-link function so that

$$\log \lambda_i = x_i^{\mathrm{T}} \beta. \tag{1.3}$$

Then we call model (1.2) and (1.3) the zero-truncated generalized Poisson regression model, where β is a p-dimension regression coefficient, and $x_i^{\mathrm{T}} = (x_{i1}, x_{i2}, \cdots, x_{ip}), i = 1, 2, \cdots, n$. The log-likelihood function of the ZTGP regression model based on a sample of n independent observations is expressed as

$$\begin{aligned}\log L = \quad & l(\alpha, \beta | y > 0) = \sum_{i=1}^{n} \Bigg[y_i (\log \lambda_i - \log(1 + \alpha\lambda_i)) + (y_i - 1)\log(1 + \alpha y_i) - \frac{\alpha \lambda y_i}{1 + \alpha\lambda_i} \\ & -\log y_i! - \log(\exp(\lambda_i/(1 + \alpha\lambda_i)) - 1) \Bigg]. \end{aligned} \tag{1.4}$$

Following the approach generalized Poisson regression model, we can obtain the maximum likelihood estimators by using the Newton-Raphson iterative method. By differentiating the log-likelihood function (1.4) with respect to α, β, we have

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} \left(-\frac{y_i \lambda_i (2 + \alpha\lambda_i)}{(1 + \alpha\lambda_i)^2} + \frac{y_i (y_i - 1)}{1 + \alpha y_i} + \frac{\lambda_i^2 \exp(\lambda_i/(1 + \alpha\lambda_i)}{\exp(\lambda_i/(1 + \alpha\lambda_i) - 1)(1 + \alpha\lambda_i)^2} \right), \tag{1.5}$$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \left(\frac{y_i}{(1 + \alpha\lambda_i)^2} - \frac{\lambda_i (\lambda_i/(1 + \alpha\lambda_i)}{\exp(\lambda_i/(1 + \alpha\lambda_i) - 1)(1 + \alpha\lambda_i)^2} \right) x_i, \tag{1.6}$$

so the score function is $U = \left(\frac{\partial l}{\partial \alpha}, \frac{\partial l}{\partial \beta} \right)$, and by differentiating twice the log-likelihood function with respect to α, β, we have the observed Fisher information matrix

$$\boldsymbol{K} = - \left(\begin{array}{cc} \dfrac{\partial^2 l}{\partial \alpha^2} & \dfrac{\partial^2 l}{\partial \alpha \partial \beta^T} \\ \dfrac{\partial^2 l}{\partial \beta \partial \alpha} & \dfrac{\partial^2 l}{\partial \beta \partial \beta^T} \end{array} \right). \tag{1.7}$$

The inverse of matrix $\boldsymbol{K}$ is partitioned as follows.

$$\boldsymbol{K}^{-1} = \left(\begin{array}{cc} \boldsymbol{K}^{\alpha\alpha} & \boldsymbol{K}^{\alpha\beta} \\ \boldsymbol{K}^{\beta\alpha} & \boldsymbol{K}^{\beta\beta} \end{array} \right), \tag{1.8}$$

where

$$\begin{aligned}\frac{\partial^2 l}{\partial \alpha^2} = \sum_{i=1}^{n} \Bigg\{ & y_i \lambda_i^2 \frac{3 + \alpha\lambda_i}{(1 + \alpha\lambda_i)^3} - \frac{y_i^2 (y_i - 1)}{(1 + \alpha y_i)^2} - \lambda_i^2 \exp(\lambda_i/(1 + \alpha\lambda_i) \cdot \\ & \frac{(1 - \alpha^2 \lambda_i^2)(\exp(\lambda_i/(1 + \alpha\lambda_i) - 1) - \lambda_i}{(1 + \alpha\lambda_i)^4 (\exp(\lambda_i/(1 + \alpha\lambda_i) - 1)^2} \Bigg\}, \end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial\beta\partial\beta^{\mathrm{T}}} = & \sum_{i=1}^{n}\Bigg\{-\frac{2\alpha\lambda_i y_i}{(1+\alpha\lambda_i)^3}-\lambda_i\exp(\lambda_i/(1+\alpha\lambda_i)\cdot \\ & \frac{(1-\alpha^2\lambda_i^2)(\exp(\lambda_i/(1+\alpha\lambda_i)-1)-\lambda_i}{(1+\alpha\lambda_i)^4(\exp(\lambda_i/(1+\alpha\lambda_i)-1)^2}\Bigg\}x_i x_i^{\mathrm{T}}, \\ \frac{\partial^2 l}{\partial\beta\partial\alpha} = & \sum_{i=1}^{n}\Bigg\{-\frac{2\lambda_i y_i}{(1+\alpha\lambda_i)^3}-\lambda_i\exp\lambda_i/(1+\alpha\lambda_i)\cdot \\ & \frac{\lambda_i^2-(2\lambda_i+2\alpha\lambda_i^2)(\exp(\lambda_i/(1+\alpha\lambda_i)-1)}{(1+\alpha\lambda_i)^4(\exp(\lambda_i/(1+\alpha\lambda_i)-1)^2}\Bigg\}x_i.\end{aligned}$$

The Newton-Raphson iterative algorithm used above requires the specification of initial values. Our suggestion is setting $\alpha = 0$ and β using ML estimation obtained from the Poisson regression model. Let $\xi = (\alpha, \beta^{\mathrm{T}})^{\mathrm{T}}$, under the usual regularity conditions for maximum likelihood estimation, when the sample size is large, $\hat{\xi} \sim N_p\left(\xi, K^{-1}\right)$ approximately.

2 Score tests for dispersion and regression parameters

In many applications, it is important to assess whether the assumed model is indeed appropriate. Gupta et al.[12], Feng-Chang Xie, et al.[13] developed score tests to detect the dispersion and zero inflation in a zero-inflated generalized Poisson regression model and a zero-inflated generalized Poisson mixed regression model. In this section, we derive two methods to test significance of dispersion and regression coefficients in ZTGP model. The test for significance of dispersion is equivalent to test the hypothesis

$$H_{01}: \alpha = 0, \quad H_{11}: \alpha \neq 0. \tag{2.1}$$

Let $\hat{\xi}_1 = (0, \hat{\beta}^{\mathrm{T}})^{\mathrm{T}}$ be the restricted maximum likelihood estimates(REML) under H_{01}. Then the score test statistic for testing H_{01} is

$$SC_1 = \left\{\left(\frac{\partial l}{\partial\alpha}\right)^2 \boldsymbol{K}^{\alpha\alpha}\right\}_{\hat{\xi}_1}, \tag{2.2}$$

where $\boldsymbol{K}^{\alpha\alpha}$ is the block matrix corresponding to the parameter α for the inverse of Observed Fisher information matrix $\boldsymbol{K}$. The standard asymptotic statistic suggests that the score statistic is asymptotically distributed as $\chi^2(1)$. Zhao Yang et al.[14] suggested this statistic is more appropriate in practical application than likelihood ratio test statistics due to the higher empirical power of the score test and only requires the parameter of interest test be estimated under null hypothesis(zero-truncated Poisson model). Meanwhile, the covariate effects on the parameter λ of ZTGP should also be considered. We may test the following hypothesis about regression coefficients

$$H_{02}: \beta^* = 0, \quad H_{12}: \beta^* \neq 0, \tag{2.3}$$

where β^* is a subset of β without the intercept β_0. Let $\hat{\xi}_2 = (\hat{\alpha}, \hat{\beta}_0, 0^{\mathrm{T}})^{\mathrm{T}}$ be the REML estimates of parameter ξ under null hypothesis H_{02}. Based on the log-likelihood function l and partitioning the x_i as $(1,(x_i^*)^{\mathrm{T}})^{\mathrm{T}}$, we can get the score function of β^* as

$$\frac{\partial l}{\partial \beta}=\sum_{i=1}^{n}\left(\frac{y_i}{(1+\alpha\lambda_i)^2}-\frac{\lambda_i(\lambda_i/(1+\alpha\lambda_i)}{\exp(\lambda_i/(1+\alpha\lambda_i)-1)(1+\alpha\lambda_i)^2}\right)x_i^*, \tag{2.4}$$

Partitioning the matrix $\boldsymbol{K}^{\beta\beta}$ as

$$\boldsymbol{K}^{\beta\beta}=\begin{pmatrix}\boldsymbol{K}^{\beta_0\beta_0} & \boldsymbol{K}^{\beta_0\beta^*}\\ \boldsymbol{K}^{\beta^*\beta_0} & \boldsymbol{K}^{\beta^*\beta^*}\end{pmatrix}, \tag{2.5}$$

where $\boldsymbol{K}^{\beta\beta}$ is the block matrix corresponding to the parameter β. Then the score test statistic for testing β^* is

$$SC_2=\left\{\left(\frac{\partial l}{\partial \beta^*}\right)^{\mathrm{T}}\boldsymbol{K}^{\beta^*\beta^*}\left(\frac{\partial l}{\partial \beta^*}\right)\right\}_{\hat{\xi}_1}. \tag{2.6}$$

3 Example: hospital length of stay data

To illustrate our methodology for fitting a ZTGP model, we first consider a data set from the 1997 MedPar dataset (available at http://www. gseis.ucla.edu/courses/data/medpar). The response variable y denote the length of hospital stay which does not and cannot have any zero values. Length of stay begins with a value of one and grows from there. There are 1 495 observations in the MedPar dataset, and the minimum count is 1 and the maximum count is 116, with mean 9.854 2 and median 8. The dispersion index (the ratio of variance to mean) is 7.917 5. so the data exhibit over-dispersion. There are 9 variables in the dataset. Here we select five important explanatory variables, i.e., hmo(x_1), white(x_2), type2(x_3), type3(x_4), died (x_5), from the variables and using the zero-truncated generalized Poisson regression model to fit the data for illustrating our results:

$$y_i \sim ZTGP(\alpha,\lambda_i)$$

with $\log\ \lambda_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}$. The estimations are $\hat{\beta}=(2.346\,2,-0.067\,6,-0.117\,9,0.251\,9,0.734\,8,-0.242\,0)$ and $\hat{\alpha}=0.154\,1$. To test the significance of the dispersion and regression coefficients, using the results obtained in Section 2, we get the score test statistics $SC_1=1\,259.3$, $SC_2=136.9$. The corresponding p-values are all smaller than 0.000 1. There we should reject the hypothesis H_{01} and H_{02}. ZTGP regression model is suitable for this dataset. To compare the goodness of fit, we compute the values of AIC and the minus log-likelihood for ZTGP model and some other alternative models to fit these data. The results are provided in Tab. 1, which shows that among these models, the ZTGP regression model, is better than others to fit this data set.

Tab. 1 Compare of different models for MedPar dataset

model number	model type	$-\log L$	AIC
1	Poisson	6 835.0	9.135 7
2	zero-truncated Poisson	6 834.7	9.135 3
3	generalized Poisson	4 768.6	6.370 0
4	zero-truncated generalized Poisson	4 738.2	6.329 4

where $AIC = -2(\log L + k)/n$, k is the total number of parameters in the model.

4 Simulation study

In this section, we examine the performances of score test statistics via Monte Carlo simulations to provide finite-sample properties of the proposed statistics. The model used for simulation study is

$$y_i \sim ZTGP(\alpha, \lambda_i), \quad i = 1, \cdots, n,$$

where $\log \lambda_i = \beta_0 + \beta_1 x_i$, and the true values under the null hypothesis is chosen as $\beta_0 = -0.8$, $\beta_1 = 1.2$.

We first generate a set of random numbers from a uniform distribution in the interval [1,3] as the value of x_i. To get values of y_i, a random variate is drown from a ZTGP model with the true values of parameters, the value of x_i, and a given β_0, β_1. Repeating this procedure n times, we get a set of simulated data $y_i, i = 1, 2, \cdots, n$.

In brief, here we only list the result of score test statistics SC_1, i.e., the performance for testing of dispersion. The values of score test statistics SC_1 are computed by formulas (2.2) shown in Section 3. We take $\alpha = 0.01, 0.025, 0.05, 0.075, 0.1$. For given values of parameters, we do 1 000 replications(the values of x_i are fixed for each replication). We then obtain the empirical power of the tests by calculating the proportion of times that the test value is greater than the $\chi^2_\alpha(1)$ critical value at $\alpha = 0.05$ level. The simulation are performed for different n to get the simulated powers of test statistics. The results are shown in Tab. 2 and Fig. 1.

Tab. 2 Simulation powers of SC_1

n	$\alpha = 0.01$	$\alpha = 0.025$	$\alpha = 0.05$	$\alpha = 0.075$	$\alpha = 0.1$
40	0.067	0.164	0.574	0.826	0.952
80	0.075	0.369	0.892	0.996	1
120	0.097	0.502	0.957	0.997	1
200	0.171	0.785	0.996	1	1

The results shown in Tab. 2 and Fig. 1 that the power of the test for detecting dispersion α increase slowly for small $n(n = 40)$ and small $\alpha(\alpha = 0.01)$; but for larger values of n, as α increases it approaches to 1 quickly.

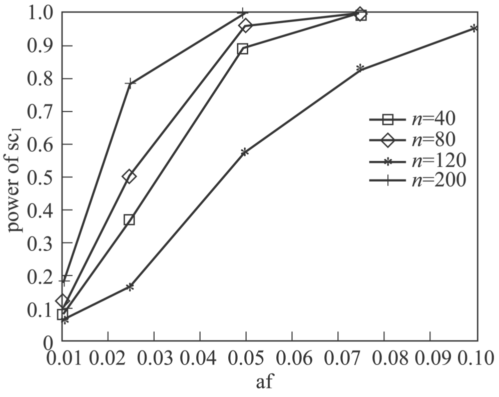

Fig. 1 Simulated power of test statistic SC_1

5 Conclusion

In this paper, we have presented a zero-truncated generalized Poisson regression model. An algorithm for estimating parameters is obtained and two score tests are presented for testing the ZTP regression model against the ZTGP regression model, and for testing the significance of regression coefficients. A numerical example and Monte Carlo simulation are given to illustrate our method.

Our main work has been focused on the ZTGP regression models without correlation between observations. However, it seems that it is reasonable to assume the correlation between observations. We will consider it in our future research.

Acknowledgements The authors are thankful to the editor and anonymous referees for valuable and constructive suggestions leading to improve quantity and presentations.

[References]

[1] CAMERON A C, PRAVIN K T. Regression Analysis of Count Data [M]. New York: Cambridge University Press, 1998.

[2] CONSUL P C, FAMOVE F. Generalized Poisson regression model[J]. Commumication in Statistics Theory and Methods, 1992, 21: 81-109.

[3] HILBE J M. Negative Binomial Regression[M]. Cambridge: Cambridge University Press, 2007.

[4] FAMOVE F, WANG W. Censored generalized Poisson regression model[J]. Computational Statistics and Data Analysis, 2004, 46: 547-560.

[5] XIE F C, WEI B C. Influence analysis for Poisson inverse Gaussian regression models based on the EM algorithm[J]. Metrika, 2008, 67: 49-62.

[6] XIE F C, WEI B C. Diagnostics analysis in censored generalized Poisson regression model[J]. Journal of Statistical Computation and Simulation, 2007, 77: 695-708.

[7] HALL D B. Zero-inflated Poisson and binomial regression with random effects : a case study[J]. Biometrics, 2000, 56: 1030-1039.

[8] HARDIN J W, HILBE J M. Generalized Linear Models and Extensions[M]. 2nd ed.[s.l.]: Stata Press, 2007.

[9] KARLIS D. A general EM approach for maximum likelihood estimation in mixed Poisson regression models[J]. Statist Modelling, 2001(1): 305-318

[10] WINKELMANN R. Econometric Analysis of Count Data[M]. 5th ed. Berlin: Springer-Verlag, 2008.

[11] CONSUL P C. Generalized Poisson Distribution: Properties and Application[M]. New York: Marcel Dekker, 1989.

[12] GUPTA P L, GUPTA R C. TRIPATHI R C. Score test for zero inflated generalized Poisson model[J]. Communcations in Statistics Theory and Methods, 2004, 33: 47-64.

[13] XIE F C, WEI B C, LIN J G. Score tests for zero-inflated generalized Poisson mixed regression models[J]. Computational Statistics and Data Analysis, 2009, 53: 3478-3489.

[14] YANG Z, HARDIN J W, ADDY C L. A score test for over dispersion in Poisson regression based on generalized Poisson-2 model[J]. Journal of Statistical Planning and Inference, 2009, 139: 1514-1521.