Journal of the Korean Data & Information Science Society
2013, **24**(4), 895–900

http://dx.doi.org/10.7465/jkdi.2013.24.4.895
한국데이터정보과학회지

A simple zero inflated bivariate negative binomial regression model with different dispersion parameters[†]

Dongseok Kim[1]

[1]Department of Mathematics, Kyonggi University

Received 16 May 2013, revised 27 May 2013, accepted 1 June 2013

Abstract

In this research, we propose a simple bivariate zero inflated negative binomial regression model with different dispersion for bivariate count data with excess zeros. An application to the demand for health services shows that the proposed model is better than existing models in terms of log-likelihood and AIC.

Keywords: Bivariate negative binomial, correlation, dispersion, zero inflation.

1. Introduction

Regression models are widely used in most fields in order to verify the causality between variables. But the normal regression model has limits in the research when categorical variables, especially such as count data, are observed as response variables. Poisson regression models have an important role in data analysis for count response variables as the normal regression models occupy a prominent place in data analysis of continuous response variables. The Poisson regression model has a strong assumption that its mean should be equal to its variance, although it is hard for a mean and a variance to be equal in real data analysis. Especially, the overdispersion, which means that the variance is greater than the mean, occurs very often. Negative binomial distributions are used instead of Poisson distribution in the count regression models as a way of solving the overdispersion. The negative binomial distribution is a gamma mixture of Poisson distribution, and thus it can be adapted easily due to an explicit marginal likelihood form.

In this paper, we extend the negative binomial regression models to a bivariate case in addition to zero-inflation. Since Li *et al.* (1999) proposed the zero-inflated bivariate count regression models, Walhin (2001) and Wang *et al.* (2003) showed the several zero inflated Poisson, denoted by BZIP models, could be extended to the zero inflated bivariate models using various Poisson mixtures. But, it is often for the zero-inflation to connect with the overdispersion, and it is more natural to use negative binomial distributions instead of Poisson distributions. In this direction, Wang (2003) first proposed the bivariate version of negative binomial models to control zero-inflation and overdispersion together. However, it is pointed out as a weakness that the bivariate zero-inflated negative binomial, denoted by

[†] This work was supported by Kyonggi University Research Grant 2010.
[1] Professor, Department of Mathematics, Kyonggi University, Gyeonggi-do 443-760, Korea.
E-mail: dongseok@kgu.ac.kr

BZINB models suggested by Wang, use the common dispersersion parameter, which is also correlation parameter, between two response variables. When we apply BZINB model for the analysis of bivariate zero-inflated count data which have different dispersions for each of response variables, the parameter estimates or their standard errors would be inefficient as the overdispersion leads to underestimation of standard errors in the the univariate Poisson regression model (Cox, 1983). It is due to the fact that the different dispersions are not properly taken into account in the BZINB model with common dispersion parameter.

A univariate model naturally extends for bivariate and multivariate model as Choi (2008) and Hong and Jung (2011). There have been various domestic research on bivariate and multivariate zero inflated Poisson models. Kim (1998, 2004) studied these models for the changepoint and Kim *et al.* (1999) found the moments of the bivariate zero-inflated Poisson distributions. As an expansion of multivariate zero inflated Poisson models, Kim (2003) provided an application of multivariate zero-inflated Poisson regression model.

In the present article, we propose a simple bivariate zero inflated negative binomial model, denoted by BZINBDD, allowing different dispersion parameters for two response variables. The proposed model is an extension of the Model 1 of BZIP by Walhin (2001). In an application to the analysis of health-care utilization data described in Cameron *et al.* (1988), the proposed BZINBDD model dominates either the generalized bivariate negative binomial, denoted by GBIVARNB model (Gurmu and Elder, 2000), or the BZINB model (Wang, 2003).

2. Model

Let Y_1 and Y_2 be the correlated random variables representing event counts, and let (y_{1i}, y_{2i}), $i = 1, \cdots, n$ be an observed vector of (Y_1, Y_2). We propose the following joint probability function of Y_{1i} and Y_{2i}

$$\begin{aligned} P(Y_{1i}=0, Y_{2i}=0) &= \phi_i + (1-\phi_i)(1+\tau_1\mu_{1i})^{-\tau_1^{-1}}(1+\tau_2\mu_{2i})^{-\tau_2^{-1}} \\ P(Y_{1i}=y_{1i}, Y_{2i}=y_{2i}) &= (1-\phi_i) f(y_{1i}, \mu_{1i}, \tau_1) f(y_{2i}, \mu_{2i}, \tau_2), \text{ if } (Y_{1i}, Y_{2i}) \neq (0,0), \end{aligned} \quad (2.1)$$

where ϕ_i represents an extra proportion of zero-zero cell for i^{th} observation, and $f(y_{ki}, \mu_{ki}, \tau_k)$, $k = 1, 2$ is the conventional negative binomial probability distribution given by

$$f(y_{ki}, \mu_{ki}, \tau_k) = \frac{\Gamma(y_{ki} + 1/\tau_k)}{y_{ki}!\Gamma(1/\tau_k)}(1+\tau_k\mu_{ki})^{-\tau_k^{-1}}(1+\tau_k^{-1}\mu_{ki}^{-1})^{-y_{ki}}, k = 1, 2, \quad (2.2)$$

where τ_1 and τ_2 are dispersion parameters for Y_{1i} and Y_{2i}, and μ_{1i}, μ_{2i} and ϕ_i depend on the vectors of covariates $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$ and $\mathbf{z}_i$ whose dimensions are $k_1 \times 1$, $k_2 \times 1$ and $k_3 \times 1$, respectively. We assume that

$$\mu_{1i} = \exp(\mathbf{x}_{1i}'\vec{\beta_1}), \quad \mu_{2i} = \exp(\mathbf{x}_{2i}'\vec{\beta_2}), \quad \text{and} \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{z}_i'\vec{\gamma},$$

where $\vec{\beta_1}$, $\vec{\beta_2}$ and $\vec{\gamma}$ are $k_1 \times 1$, $k_2 \times 1$ and $k_3 \times 1$ vector of parameters, respectively. The model (2.1) can be referred to as the bivariate zero inflated negative binomial model with different dispersion (BZINBDD). If $\tau_1 \to 0$ and $\tau_2 \to 0$, the model given in the proposed model (2.1) reduces to the Model 1 of Walhin (2001)'s BZIP model. The correlation coefficient of Y_{1i} and Y_{2i} is

$$Corr(Y_{1i}, Y_{2i}) = \frac{\phi_i\mu_{1i}\mu_{2i}}{\sqrt{\mu_{1i}\mu_{2i}[1+\mu_{1i}(\tau_1+\phi_i)][1+\mu_{2i}(\tau_2+\phi_i)]}}. \quad (2.3)$$

The correlation in the equation (2.3) is slightly different from the correlation of the BZINB model of Wang (2003, p. 375). In fact, there exist two sources to generate the correlation of Y_{1i} and Y_{2i} in the BZINB model. The first one is the common dispersion parameter of the bivariate negative binomial (BNB) distribution, and the second one is the inflated proportion of zero-zero cell of i^{th} observation. The correlation of Y_{1i} and Y_{2i} in the proposed model only depends on the proportion of zero inflation by introducing different dispersion parameters for two response variables. However, the proposed model may be more appropriate than the BZINB model for the bivariate zero inflated count data having different dispersions.

The log-likelihood function based on n independent sample is obtained by

$$\begin{aligned}\log L &= \sum_{i=1}^{n} I_{(y_{1i},y_{2i})=(0,0)} \log\Big(\exp(\mathbf{z}_i{}'\vec{\gamma}) + (1+\tau_1\mu_{1i})^{-\tau_1^{-1}}(1+\tau_2\mu_{2i})^{-\tau_2^{-1}}\Big) \\ &+ \sum_{i=1}^{n} I_{(y_{1i},y_{2i})\neq(0,0)} \sum_{k=1}^{2}\Big[\sum_{j=1}^{y_{ki}} \log(\tau_k y_{ki}+1-\tau_k j) - (\tau_k^{-1}+y_{ki})\log(1+\tau_k\mu_{ki}) \\ &\quad + y_{ki}\log(\mu_{ki}) - \log(y_{ki}!)\Big] - \sum_{i=1}^{n}\log(1+\exp(\mathbf{z}_i{}'\vec{\gamma})), \end{aligned} \tag{2.4}$$

where $I_{(\cdot)}$ is the indicator function taking the value 1 if the condition is true and 0 otherwise. Using the log-likelihood function in the equation (2.4), we obtain the ML estimator of each of the parameters. The asymptotic standard errors of the parameters can be obtained from the the outer product of the gradient (OPG) method (Davidson and MacKinnon, 1993).

3. Application

In this study, we apply the proposed model to the data of 1977-1978 Australia health survey given by Cameron *et al.* (1988). Gurmu and Elder (2000) and Wang (2003) already modelled health-care utilization by applying the GBIVARNB model and BZINB regression model from the afore-mentioned data. The data were obtained from the Journal of Applied Econometrics 1997 data archive. We consider the number of consultations with a doctor during the 2-weeks prior to survey (doctorcon, Y_1) and the number of consultations with non-doctor health professionals (chemist, optician, social worker etc) during the past 4 weeks prior to survey (nondoccon, Y_2) as the response variables. The frequency of non-users who never use any of services is 73.7 %, and the mean and the variance are 0.302 and 0.637 for doctor visits, and 0.215 and 0.932 for non-doctor visits. Therefore, two response variables look like having the different dispersions as well as zero inflation from descriptive statistics.

The twelve variables including Socio-economic variables and the insurance and health status variables are used as the explanatory variables for μ_{1i}, μ_{2i} and ϕ_i. For the detailed description and summary statistics for the explanatory variables, refer to Cameron *et al.* (1988).

Table 3.1 gives the results of parameter estimates, maximized value of the log-likelihood function and AIC value for the proposed model and the BZINB model of Wang (2003). The parameter estimates of BZINBDD and BZINB models show the similar results, while $|t|$ values of Wang's BZINB model are dramatically greater than those of the BZINBDD model. In BZINBDD model, τ_1 and τ_2 are estimated by 0.632 and 6.561 which are very

different. It is the evidence of having different dispersions for two response variables in this data. Therefore, we conjecture that Wang's BZINB model may underestimate the standard errors of parameters since it uses the common dispersion parameter, as the overdispersion leads to an underestimation of standard errors in the the univariate Poisson regression model (Cox, 1983). In addition, Table 3.1 also shows that the proposed BZINBDD dominates the BZINB model with respect to both the maximized value of log-likelihood and AIC.

For ϕ_i, all explanatory variables except income, freepoor, freerepa and chcond1 are significant at 5% significance level in the proposed BZINBDD model. For μ_{ki}, freepoor, illness and actdays are significant for doctor visits, while actdays and chcond1 are the only two important determinants of non-doctor health professional visits.

Table 3.1 Estimates from bivariate zero inflated negative binomial models

	BZINBDD						BZINB[b]					
	Doctorcon		Nondoccon		ϕ_i		Doctorcon		Nondoccon		ϕ_i	
Variable	Est.	$\|t\|$	Est.	$\|t\|$	Est.	$\|t\|$	Est.	$\|t\|$	Est.	$\|t\|$	Est.	$\|t\|$
Constant	-1.251	-4.84	-1.642	-3.21	0.699	1.08	-1.301	42.30	-1.508	43.80	0.852	11.10
Sex	0.055	0.77	0.110	0.84	-0.453	-2.47	0.036	0.94	0.205	4.89	-0.407	3.61
Age	1.205	0.90	-1.976	-0.66	8.131	2.12	1.087	18.40	-2.984	47.20	6.391	34.40
Agesq	-1.354	-0.93	2.576	0.82	-11.823	-2.63	-1.092	11.80	3.747	38.80	-9.503	29.60
Income	-0.193	-1.75	-0.068	-0.30	-0.243	-0.89	-0.155	3.12	-0.023	0.40	-0.131	1.27
Levyplus	-0.015	-0.15	0.148	0.70	-0.454	-2.17	-0.027	0.59	0.168	3.16	-0.457	4.02
Freepoor	-0.499	-2.59	-0.196	-0.43	0.021	0.04	-0.525	2.65	-0.126	0.58	0.102	0.24
Freerepa	-0.027	-0.21	0.375	1.31	-0.664	-1.70	0.029	0.55	0.377	6.86	-0.520	2.60
Illness	0.078	3.07	-0.018	-0.37	-0.604	-5.75	0.072	6.15	-0.051	3.86	-0.669	9.55
Actdays	0.111	15.84	0.095	5.43	-1.690	-2.52	0.111	22.70	0.095	17.80	-1.761	2.92
Hscore	0.021	1.53	0.051	1.80	-0.141	-2.43	0.022	2.55	0.038	4.04	-0.151	3.31
Chcond1	-0.039	-0.42	0.278	1.65	-0.268	-1.26	-0.025	0.55	0.324	6.45	-0.195	1.48
Chcond2	-0.007	-0.07	0.857	3.99	-0.961	-2.45	0.115	1.78	0.895	14.10	-0.622	2.21
Log(α)[a]							-0.029	0.44				
τ_1	0.632	7.73										
τ_2			6.561	11.43								
Log-likel.	-5254.0						-5715.9					
AIC	10590.0						11511.8					

[a] The coefficient log(α) is a common dispersion parameter of BZINB model.

[b] The result of BZINB model is reported in Wang (2003).

Table 3.2 gives the observed and fitted frequencies of the BZINBDD and BZINB models. The fitted cell frequencies of each model are calculated in the similar method of Gurmu and Trivedi (1996). Let $\hat{p}(c_{1i}, c_{2i}), i = 1, \cdots, n; c_{1i}, c_{2i} = 0, 1, \cdots$, denoted the fitted probability that (Y_{1i}, Y_{2i}) has (c_1, c_2). Then the fitted frequency in cell (c_1, c_2) is calculated as $\sum_{i=1}^{n} \hat{p}(c_{1i}, c_{2i})$, $c_{1i}, c_{2i} = 0, 1, \cdots$.

From Table 3.2, one can see that the BZINB model tends to overpredict or underpredict the observed frequencies, especially when $Y_1 \leq 1$ or $Y_2 \leq 1$. On the contrary, the BZINBDD model provides an adequate fit. In general the BZINBDD fits the data better than the BZINB model does. Let us remark that Table 3.1 and Table 3.2 are obtained by SAS/IML.

Table 3.2 Observed and fitted frequencies

Model	Doctorcon (Y_1)	Nondoccon (Y_2)					
		0	1	2	3	4	5+
Observed		3826	196	57	9	17	36
BZINBDD	0	3879.0	167.8	62.0	29.2	15.5	26.9
BZIND		3888.1	284.2	52.5	11.1	2.7	1.1
Observed		670	66	18	4	6	18
BZINBDD	1	575.9	56.4	21.5	10.6	5.8	12.3
BZIND		419.3	139.4	39.9	11.7	3.7	2.1
Observed		148	11	4	1	1	9
BZINBDD	2	169.0	17.4	7.0	3.6	2.1	5.8
BZIND		105.8	54.6	21.8	8.4	3.3	2.6
Observed		25	2	2	0	0	1
BZINBDD	3	53.6	5.9	2.6	1.4	0.9	3.0
BZIND		28.7	20.7	10.8	5.3	2.6	2.7
Observed		19	1	1	0	0	3
BZINBDD	4	20.1	2.4	1.1	0.7	0.4	1.8
BZIND		8.5	8.1	5.4	3.2	1.9	2.6
Observed		28	2	2	0	2	5
BZINBDD	5+	20.3	2.6	1.3	0.8	0.6	2.6
BZIND		4.6	6.4	6.1	5.1	3.9	11.4

4. Conclusions

This paper proposed a simple bivariate zero inflated negative binomial regression model with different dispersions on two response variables. In the application of health-care utilization, the proposed model is better than the BZINB model in the sense of maximized value of log-likelihood, AIC and fitted frequencies. The proposed methodology may be further explored by a suitable simulation for the subsequent research.

The proposed model can be extended to more general BZINBDD model. In fact, the proposed model ignored the natural correlation by allowing different dispersions. The general BZINBDD model can be constructed using the underlying BNB distribution from the trivariate reduction technique of three independent negative binomial random variables. The further research may include the estimation and testing problem (for example, testing for zero inflation, testing for covariance parameter) in the general BZINBDD model.

References

Cameron, A. C., Trivedi, P. K., Milne, F. and Piggott, J. (1988). A microeconometric model of the demand for health care and health insurance in Astralia. *Review of Economic Studies*, **55**, 85–106.

Choi, J. (2008). A marginal logit mixed-effects model for repeated binary response data. *Journal of Korean Data & Information Science Society*, **19**, 413–420.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika* **70**, 269–274.

Davidson, R. and MacKinnon, J.G. (1993). *Estimation and inference in econometrics*, Oxford University Press, New York.

Gurmu, S. and Elder, J. (2000). Generalized bivariate count data regression models. *Economics Letters*, **68**, 31–36.

Hong, C. S. and Jung, M. H. (2011). Undecided inference using bivariate probit models. *Journal of Korean Data & Information Science Society*, **22**, 1017–1028.

Kim, K. M. (1999). Inferences for the changepoint in bivariate zero-inflated Poisson model. *Journal of Korean Data & Information Science Society*, **10**, 319–327.

Kim, K. M. (2003). An application to multivariate zero-inflated Poisson regression model. *Journal of Korean Data & Information Science Society*, **14**, 177–186.

Kim, K. M. (2004). Tests for the change-point in the zero-inflated Poisson distribution. *Journal of Korean Data & Information Science Society*, **15**, 387–394.

Kim, K. M., Lee, S. H. and Kim, J. T. (1998). Moments of the bivariate zero-inflated Poisson distributions. *Journal of Korean Data & Information Science Society*, **9**, 47–56.

Li, C. S., Lu, J. C., Park, J., Kim, K. and Brinkley, P. A. and Peterson, J. (1999). Multivariate zero-inflated Poisson models and their applications. *Technometrics*, **41**, 29–38.

Walhin, J. F. (2001). Bivariate ZIP models. *Biometrical Journal*, **43**, 147–160.

Wang, K., Lee, A. H., Yau, K. and Carivick, P. (2003). A bivariate zero-inflated Poisson regression model to analyze occupational injuries. *Accident Analysis and Prevention*, **35**, 625–629.

Wang, P. (2003). A bivariate zero-inflated negative binomial regression model for count data with excess zeros. *Economics Letters*, **78**, 373–378.