Bayesian Zero Inflated Negative Binomial Regression Model for The Parkinson Data

Shafira[1], Sarini Abdullah[2], Dian Lestari[3]
shfiiraa@sci.ui.ac.id[1], sarini@sci.ui.ac.id[2], dian.lestari@sci.ui.ac.id[3]

Department Mathematics, University of Indonesia, Depok, 16424,Indonesia[1,2,3]

Abstract. Excess zeros can be solved by Zero Inflated Poisson (ZIP). If over-dispersion still exists in the data, the ZIP model is no longer suitable. Replacing the Poisson distribution with negative binomial distribution in the counting process may provide an alternative solution. Zero Inflated Negative Binomial (ZINB) regression model is estimated using the Bayesian method. Conjugate non-informative priors were used. Sampling parameters from posterior distribution is conducted using Markov Chain Monte Carlo (MCMC) simulation with 50,000 burn-in and 150,000 iterations. The model was then implemented to Parkinson's disease data obtained from the Parkinson's Progression Markers Initiative (PPMI) program. The MCMC result showed the convergence of the parameters. The result of the inspection of motoric aspect was significant in explaining does Parkinson's patients have to consume drugs or not. The result of the inspection of non-motoric aspect and body response were significant in explaining motoric complication in Parkinson's disease sufferers.

Keywords: Count data, Excess zero, Markov Chain Monte Carlo, Over-dispersion, Parkinson's disease.

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290530

1 Introduction

In regression modeling for count data, problem of over-dispersion may arise due to excess zeros on the response variable. In this case, zero-inflated model might be suitable to model the data. Zero-inflated model assumes that zero value on the response variable were generated by two processes, namely the random zeros and structural zeros [1]. ZIP consists of two stages modeling: the first step to differentiate structural zeros with the counting process, and the second stage is to model the counting process, given the structural zeros have been sorted out from the data.

However, when observations included in the counting data still exhibits over-dispersion, then the expansion of Poisson distribution is required [2]. Zero Negative Binomial Inflated (ZINB) model might be the alternative for this condition, with the logistic regression model at the first stage and negative binomial regression for the second stage.

Bayesian method is more flexible for parameters estimation. In addition to the information from the data through the likelihood function, experts' judgement may be incorporated through specification of prior distribution for the parameters of interest. Combination of these two results in the posterior distribution, where sampling on the parameters can be done using Markov Chain Monte Carlo (MCMC)–Gibbs sampling technique. Therefore, in this paper, Bayesian method was implemented for the ZINB model parameters estimation, following a recommendation of Garay et al. (2015) [3].

2 Data

Data collected on 232 people with early Parkinson's disease (PD), taken in March 2019 from the Parkinson's Progression Markers Initiative (PPMI) database [4] were used in the analysis. This database named Movement Disorder Society-Unified Parkinson's Diseases Rating Scale (MDS-UPDRS) instrument. The subjects were on the scale of Hoehn and Yahr ranging from 0 to 3 in a period of one to five years. Before the data was filtered, data must be cleaned first. Patients whose data are incomplete will be eliminated. After that, data filtered by time period and Hoehn and Yahr scale.The response variableis the frequency of motoric complications experienced by people with PD. The subtotal scores from three parts of the MDS-UPDRS, namely Part I measuring the result of inspection of non-motoric aspect (X_1), Part II measuring the result of inspection of motoric aspects (X_2), and Part III measuring the result of inspection of body responses (X_3).

3 Statistical methods

3.1 Zero Inflated Negative Binomial Regression Model

Zero-inflated models assume that a response variable distributed zero-modified. Zero-modified distribution is a combination of a degenerate distribution at zero and such discrete distribution. For ZINB, the discrete distribution is a negative binomial distribution.

Suppose Y is a discrete random variable that consisting of the counts on n subjects, $y_1, y_2, \dots, y_n$. Observations that go into structural zeros $(y_i = 0)$ have a degenerate distribution at zero with a probability of occurring is p. While the observations included in the NB counts $(y_i = 0, 1, 2, \dots)$follow a negative binomial distribution with probability of occurring is $(1 - p)$. Therefore, $Y$is ZINB distributed which defined by

$$Y = \begin{cases} structural\ zeros, & \text{with probability } p \\ counting\ process, & \text{with probability } (1-p) \end{cases} \quad (1)$$

Based on the probability function of the zero-modified distribution [5], then probability mass function (pmf) for ZINB distribution [3] is

$$\Pr(Y = y) = \begin{cases} p + (1-p)\left(\frac{\phi}{\mu+\phi}\right)^{\phi}, & y = 0 \\ (1-p)\frac{\Gamma(y+\phi)}{\Gamma(y+1)\,\Gamma(\phi)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y}, & y = 1, 2, \dots \end{cases} \quad (2)$$

where $(\phi)^{-1}, \mu,$ and $\Gamma(.)$ representing dispersion parameter, mean, and gamma function, respectively.

Assume that there are r predictors for logistic regression function [6], s predictors for negative binomial regression function [7]. Hence, ZINB regression model [3] can be written as follows.

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \boldsymbol{X_1'\beta} = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \cdots + \beta_s X_{r1} \quad (3)$$

$$ln(\mu) = \boldsymbol{X_2'\gamma} = \gamma_0 + \gamma_1 X_{12} + \gamma_2 X_{22} + \cdots + \gamma_s X_{s2} \quad (4)$$

where

X_{j1}	:	the j-th predictor at stage one, $j = 0, 1, 2, \dots, r$.
X_{k2}	:	the k-th predictor at stage two, $k = 0, 1, 2, \dots, s$.
β_j	:	the j-th regression parameters in stage one.
γ_k	:	the k-th regression parameters in stage two.

3.2 Bayesian Method for ZINB Regression

The main characteristic of the Bayesian method is that it uses a probability function to measure uncertainty in statistical inference, in other words its probability function can be used as a benchmark for a researcher's trust in an event.Applying the Bayes' rule [8], the posterior distribution $(p(\theta|y)$ for the model's parameters, θ, can be written as

$$p(\theta|y) \propto p(\theta)\, p(y|\theta) \quad (5)$$

where $p(\theta)$ is prior distribution, and $p(y|\theta)$ is the likelihood obtained from the data.

3.2.1 The Likelihood Function for ZINB Regression

Based on the explaining data in Data section above, this paper will use PPMI database, named MDS-UPDRS instrument. The response variable is the MDS-UPDRS Part IV, or the frequency of motoric complications experienced by people with PD (Y). The subtotal scores from three parts of the MDS-UPDRS, namely Part I measuring the result of inspection of non-motoric aspect (X_1), Part II measuring the result of inspection of motoric aspects (X_2), and Part III measuring the result of inspection of body responses (X_3). Suppose Y is a random variable with ZINB distribution containing 232 independent observations, then the likelihood function for ZINB regression is as follows.

$$p(Y|\boldsymbol{\beta},\boldsymbol{\gamma},\phi) = \prod_{y_i=0} \Pr(Y = y_i|\boldsymbol{\beta},\boldsymbol{\gamma},\phi) \times \prod_{y_i>0} Pr(Y = y_i|\boldsymbol{\beta},\boldsymbol{\gamma},\phi)$$

$$= \prod_{y_i=0}\left[\frac{e^{\boldsymbol{X_1'\beta}}}{1+e^{\boldsymbol{X_1'\beta}}} + \left(\frac{1}{1+e^{\boldsymbol{X_1'\beta}}}\right)\left(\frac{\phi}{e^{\boldsymbol{X_2'\gamma}}+\phi}\right)^{\phi}\right] \times \tag{6}$$

$$\prod_{y_i>0}\left[\left(\frac{1}{1+e^{\boldsymbol{X_1'\beta}}}\right)\frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}\left(\frac{\phi}{e^{\boldsymbol{X_2'\gamma}}+\phi}\right)^{\phi}\left(\frac{e^{\boldsymbol{X_2'\gamma}}}{e^{\boldsymbol{X_2'\gamma}}+\phi}\right)^{y}\right]$$

3.2.2 Prior Distribution for Parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}$, and $\boldsymbol{\phi}$

Since there is no prior information from historical data or from previous experiment, then all parameters will use conjugate non-informative priors.Prior distribution for β and γ are assumed to be normal, while ϕ is assumed to be gamma distributed. So, the joint prior distribution for ZINB regression parameters is

$$p(\boldsymbol{\beta},\boldsymbol{\gamma},\phi) = \prod_{j=0}^{r}\left[\frac{1}{\sigma_{\beta_j}\sqrt{2\pi}}e^{\frac{-\left(\beta_j-\mu_{\beta_j}\right)^2}{2\sigma^2_{\beta_j}}}\right] \times \prod_{k=0}^{s}\left[\frac{1}{\sigma_{\gamma_k}\sqrt{2\pi}}e^{\frac{-\left(\gamma_k-\mu_{\gamma_k}\right)^2}{2\sigma^2_{\gamma_k}}}\right] \times \frac{1}{b^a\Gamma(a)}\phi^{a-1}e^{-\phi/b} \tag{7}$$

All parameters assumed have prior specification, that is β_j ~normal (0,1000), γ_k ~normal (0,1000), and ϕ~gamma (a, b) with $a = 0.001$ and $b = 0.001$.

3.2.3 Posterior Distribution for Parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}$, and $\boldsymbol{\phi}$

Combining the prior and the likelihood, the resulting posterior for the parameter β, γ, and ϕ is

$$
\begin{aligned}
p(\boldsymbol{\beta},\boldsymbol{\gamma},\phi|Y) = &\prod_{j=0}^{r}\left[\frac{1}{\sqrt{2\pi(1000)}}e^{-\frac{\beta_j^2}{2(1000)}}\right]\times\prod_{k=0}^{s}\left[\frac{1}{\sqrt{2\pi(1000)}}e^{-\frac{\gamma_k^2}{2(1000)}}\right]\\
&\times\frac{1}{0.001^{0.001}\Gamma(0.001)}\phi^{-0.999}e^{-\phi/0.001}\\
&\times\prod_{y_i=0}\left[\frac{e^{\boldsymbol{X_1'\beta}}}{1+e^{\boldsymbol{X_1'\beta}}}+\left(\frac{1}{1+e^{\boldsymbol{X_1'\beta}}}\right)\left(\frac{\phi}{e^{\boldsymbol{X_2'\gamma}}+\phi}\right)^{\phi}\right]\\
&\times\prod_{y_i>0}\left[\left(\frac{1}{1+e^{\boldsymbol{X_1'\beta}}}\right)\frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}\left(\frac{\phi}{e^{\boldsymbol{X_2'\gamma}}+\phi}\right)^{\phi}\left(\frac{e^{\boldsymbol{X_2'\gamma}}}{e^{\boldsymbol{X_2'\gamma}}+\phi}\right)^{y}\right].
\end{aligned}
\tag{8}
$$

Posterior distribution in (8) is difficult to be solved analytically. Therefore, a numerical simulation using the Markov Chain Monte Carlo-Gibbs sampling is used to update the parameters given initial values, and to sample the parameters given the simulation is convergent.

4Results and Discussion

In modeling the frequency of motoric complications data from Parkinson's patients, a computer program R version 3.5.3[9] with R2JAGS package[10] is used to run the MCMC-Gibbs sampling algorithms. **Table 1** reports the posterior means (mean), standard deviations (SD), and 95% credible intervals (2.5 percentile and 97.5 percentile) of the model parameters fitting the ZINB. Significant variables listed in **Table 1** with bold prints on each parameter. According to Liu & Power [11], a significant variable is a variable with the coefficient parameters that do not contain zero between 2.5 percentile and 97.5 percentile.

Table 1. The results of parameter estimate for 232 patients using MCMC-Gibbs sampling algorithm.

Parameter	*Mean*	SD	2.5 Percentile	Median	97.5 Percentile
ϕ	3.43698	1.22132	1.74547	3.22431	6.38276
β_0	0.00599	0.03144	-0.05601	0.00581	0.06753
β_1	-0.00215	0.02519	-0.05211	-0.00164	0.04704
$\boldsymbol{\beta_2}$	**-0.06189**	0.02421	-0.10950	-0.06204	-0.01463
β_3	-0.00180	0.01027	-0.02272	-0.00156	0.01789
γ_0	0.02364	0.03141	-0.03828	0.02373	0.08543
$\boldsymbol{\gamma_1}$	**0.04209**	0.01565	0.01079	0.04245	0.07256
γ_2	0.02778	0.01560	-0..00227	0.02733	0.05824
$\boldsymbol{\gamma_3}$	**0.01895**	0.00519	0.00854	0.01905	0.02901

From **Table 1**, we can write the estimates of ZINB regression model.

$$logit(p) = ln\left(\frac{p}{1-p}\right) = 0.00599 - 0.00215X_{11} - 0.06189X_{21} - 0.00180X_{31} \quad (9)$$

$$ln(\mu) = 0.02364 + 0.04209X_{12} + 0.02778X_{22} + 0.01895X_{32} \quad (10)$$

In **Table 1**, the variables having a significant effect on logistic regression is variable with the total inspection score of motoric aspect, with the estimated parameter is -0.06189. That is, the separation of observation whether patients need to take drugs or not can be affected by variable MDS-UPDRS Part II. Parameter estimates of variable MDS-UPDRS Part II is -0.06189, means that the risk of the patient not taking the drug and not experiencing motoric complications is equal to $\exp(-0.06189X_{21})$, where the other variables are assumed to be constant. Then for negative binomial regression, the variables that affect significantly is the total score of non-motoric aspects of the examination and the body's responses, with the estimated parameter is 0.04209 and 0.01895 respectively. That is the frequency of motoric complications influenced by variable MDS-UPDRS Part I and Part III. Parameter estimates of variable MDS-UPDRS Part I is 0.04209 and variable MDS-UPDRS Part III is 0.01895, means that the mean of frequency of motoric complications each patient equal to $\exp(0.04209X_{12} + 0.01895X_{32})$, where the other variables assumed to be constant.

In (9), all variables have negative regression coefficients. Because of logit function, model in (9) means that the smaller the MDS-UPDRS Part I, II, and III scores, the greater the probability of observations included in structural zeros. In other words, if a patient has a small score for the MDS-UPDRS Part I, II and III variables, then the probability that the patient will not has a motor complication due to drugs will be even greater, because this patient did not consume the drug indeed.

Model (10) shows a relationship between the probability of how often the patient experiences motoric complications with total score of each variables. It is seen that the regression coefficient is positive. This means that if score of each variables has increase, then motor complications will often occur because of the increased probability. In other words, patients taking drugs with large MDS-UPDRS Part I, II, and III scores have a high probability too for the occurrence of motor complications.

Random values generated from the posterior distribution can be described through a density plot in **Figure 1**. Beta represents parameter coefficients in the first stage (structural zeros), and gamma represents parameter coefficients in the second stage (NB counts).

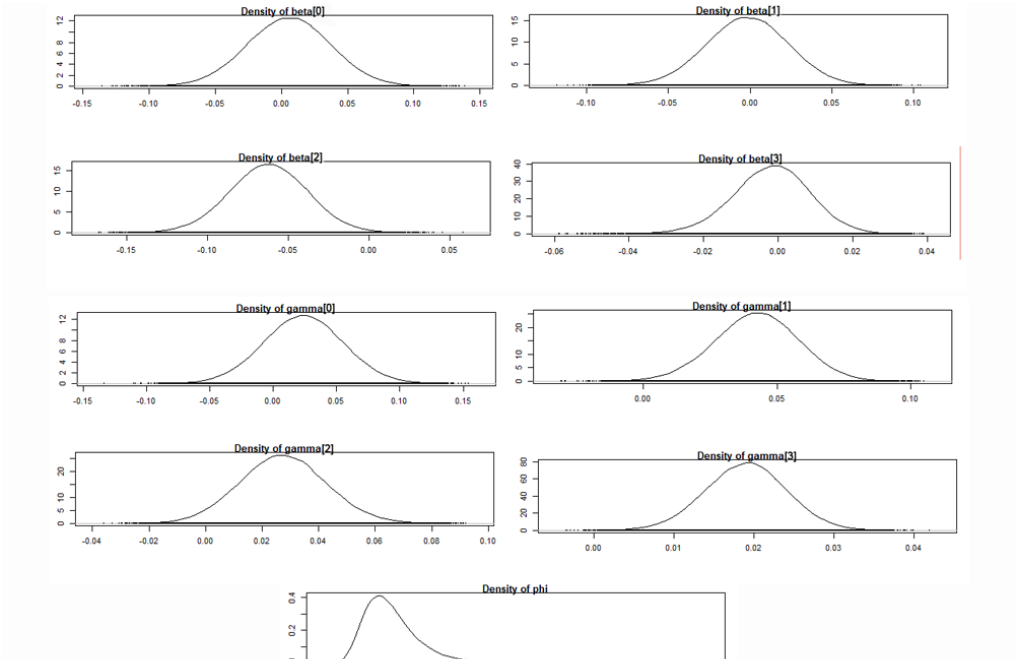

Fig. 1. Posterior density plot of estimated regression coefficients from ZINB regression.

5Conclusion

ZINB regression can overcome data with over-dispersion caused by excess zero and calculate data using two regression models, namely logistic regression and negative binomial regression. In estimating the parameters, Bayesian methods will be used. Complex calculations using the Bayesian method in the estimation parameters can be solved by the Markov Chain Monte Carlo (MCMC) simulation that can generate random values with the Gibbs-sampling algorithm. The application of ZINB regression using the Bayesian MCMC-Gibbs sampling method in 232 Parkinson data provides two conclusions. First, the variables that determine significantly on the need for patients to take MDS-UPDRS medicine are Part II. Second, in patients taking drugs the frequency of motor complications is calculated by the MDS-UPDRS variable Part II and Part III.

Acknowledgements

This research supported by the University of Indonesia with PITTA B 2019 research grant scheme, with ID number NKB-0665/UN2.R3.1/HKP.05.00/2019. We thank to all reviewers for the improvement of this article.

References

[1] Yang, S., Harlow, L. L., Puggioni, G., & Redding, C. A. A comparison of different methods of zero-inflated data analysis and an application in health surveys. Journal of Modern Applied Statistical Methods. Vol 16. No 1. pp. 518-543(2017)
[2] Lewsey, J. D., and Thomson, W. M.: The utility of the zero-inflated Poisson and zero-inflated negative binomial models: A case study of cross-sectional and longitudinal DMF data examining the effect of socioeconomic status. Community Dentistry and Oral Epidemiology.Vol 32. No 3. pp. 183-189 (2004)
[3]Garay, A. M., Lachos, V. H., Bolfarine, H.: Bayesian estimation and case influence diagnostics for the zero-inflated negative binomial regression model. Journal of Applied Statistics.Vol 42. No 6. pp. 1148-1165 (2015)
[4] Parkinson's Progressive Markers Initiative. 2018. PPMI Study CRF and Assessments. https://ida.loni.usc.edu/pages/access/studyData.jsp?categoryId=3&subCategoryId=4.
[5] Johnson, Norman Lloyd., and Samuel Kotz. Distributions in Statistics: Discrete Distributions. Wiley, 1969.
[6] Montgomery, D. C., Peck, E. A., Vining, G. G. Introduction to Linear Regression Analysis. Chicester: Wiley, 2012.
[7] Hilbe, Joseph. Negative Binomial Regression. Cambridge University Press, 2011.
[8]Gelman, Andrew, et al. Bayesian Data Analysis. Chapman & Hall/CRC, 2013.
[9] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2019.
[10] Su, Yu-Sung,Yajima, M.*R2jags:Using R to Run 'JAGS'*, 2015. URL https://CRAN.R-project.org/package=R2jags. R package version 0.5-7.
[11] Liu, H., Powers, D.A.: Bayesian Inference for Zero-Inflated Poisson Regression Models. Journal of Statistics: Advances in Theory and Applications. Vol 7. No 2. pp. 155-188 (2012)
[12]Hilbe, J. M., Ishida, E. E., dan S., D. S. Bayesian models for astrophysical data using R, JAGS, Python, and Stan. Cambridge: Cambridge University Press, 2017.
[13]Hogg, R. V., McKean, J. W., dan Craig, A. T. Introduction to mathematical statistics. Boston: Pearson, 2019.