Proceeding of ICSA 2019, p: 57-68
ISBN 978-979-19256-3-1 (PDF)

Hurdle Negative Binomial Regression Model

Ayu Andika[1], Sarini Abdullah[2*], Siti Nurrohmah[3]

[1,2,3]*Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia.*

[*)]Corresponding author: sarini@sci.ui.ac.id

Abstract

Poisson regression is a common regression model used for count data with equidispersion. However, in real data application, overdispersion often encountered, suggesting the seek for alternative model to the Poisson regression. In overdispersion data due to excess zeros and additional overdispersion in positive values, one of alternative model that can be used is hurdle negative binomial model. Hurdle negative binomial model is a two-part model consists of binary model and zero-truncated negative binomial model. In this study we discuss hurdle negative binomial model and Bayesian approach for the model's parameter estimation, then apply the method for modelling frequency of motoric complication in people with early Parkinson's disease. Markov Chain Monte Carlo with Gibbs Sampling (MCMC-GS) was implemented to sample the regression parameters from their posterior distribution. The result showed that hurdle negative binomial model fit the data satisfactorily, as implied by the convergence and unimodality of posterior density of the parameters of interest. We also identified risk factors for motoric complications.

***Keywords*:** Bayesian, Gibbs Sampling, Markov Chain Monte Carlo, Motoric Complication, Parkinson.

Introduction

Poisson regression is often used for modeling relationships between count variable based on several explanatory variables. Poisson regression requires the equidispersion assumption, that is, the variance and the mean have the same value. However, the variance in the data is often found greater than the mean, known as overdispersion. Overdispersion might be due to heterogeneity between subjects, positive correlation between responses, or excess zeros (Hilbe, 2011). Hurdle model can be an alternative to overcome over dispersed data. Hurdle model assume zero counts in response variable are generated from a different process than the positive counts (Hilbe, 2011). If positive counts of the data have extra overdispersion, then hurdle negative binomial model can be used as an alternative model.

Parameter estimation will be done using the Bayesian approach. For each parameter of interest, prior distribution -which represents the researcher's belief or initial judgement on the parameter- needs to be specified. Then basing on the sample data, sampling model or the likelihood is constructed. Combining the prior and the likelihood produced the posterior distribution, that we be used for the inference. From this process, Bayesian approach should provide more complete information to the inference, as it considers the expert judgement through the prior, not only based on the sample data. Therefore, we consider Bayesian approach for parameter estimation.

To showcase the use of the model, we used data on people with early Parkinson's disease, taken from the Parkinson's Progression Markers Initiative (PPMI) database (Song et al., 2019). This data is the version of March 19, 2018 and was accessed in May 28, 2019.

Parkinson disease is the second most common neurodegenerative movement disorder among elderly individuals (Shehzadi et al., 2018). One million people in the United States and over five million people in the world are estimated to be living with Parkinson's disease. There is no known cure for Parkinson's disease, and the current treatments can only mask some

symptoms (PPMI, 2019). Some current treatments are by providing dopaminergic and physical therapy. However, medication might have effects such as motoric complication in some patients.

Identifying characteristics of people who are at risk of developing motoric complications after medication may help in designing a better clinical treatment for these people. Therefore, in this study, we model the likelihood of people with Parkinson's disease to experience motoric complications, and further, for those who do have motoric complications, factors that explain the frequency of complications will be identified, simultaneously. Measurements from three parts of the Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) were used as the explanatory variables. This is due to the UPDRS as a golden standard for clinical assessment for people with Parkinson's disease (Song et al., 2019), and we argue that the revised version, i.e. MDS-UPDRS should also be able to explain the clinical condition of people with Parkinson's disease. Therefore, risk factors from motoric complication will be analyzed based on MDS-UPDRS data (PPMI, 2018).

Materials

Data

Data consists of observations from 385 people with early Parkinson's during their 10 visits within 5 years of study time. This data is part of the study conducted by the Parkinson's Progressive Markers Initiative (PPMI, 2019), an ongoing study on Parkinson's disease where the data were collected from the related participating clinics in several countries.

Among 385 observations, 241 of them did not experience motoric complication. It means 62% data has zero values and it indicates excess zeros condition. Furthermore, variance value of the entire sample is 7.592 greater than mean value of the entire sample that is 1.742. These all indicate the presence of overdispersion caused by excess zeros. Moreover, variance

value of positive counts is 6.702, greater than mean value of positive counts that is 4.659. So, it indicates the presence of overdispersion in positive counts. Therefore, hurdle negative binomial regression model can be used to model this Parkinson data.

The response variable (*Y*) is the number of motoric complications experienced by people with Parkinson's disease. While, for the explanatory variables we use the total score of non-motoric aspect of Parkinson's patient (total score of MDS-UPDRS Part I, X_{i1}), the total score of motoric aspect of Parkinson's patient (total score of MDS-UPDRS Part II, X_{i2}), and the total score of physical examination of Parkinson's patient (total score of MDS-UPDRS Part III, X_{i3}).

Statistical model

Hurdle negative binomial model can be used for modeling count data having excess zeros and overdispersion in the positive counts. Hurdle negative binomial model assumes that zero counts are generated from different process than the positive counts (Cheng, 2015). Zero counts in this model are assumed coming from one source. Hurdle negative binomial consists of two parts, the first is binary model to estimate binary process of zero counts versus positive counts. The second is zero-truncated negative binomial model to estimate over dispersed positive counts only.

Suppose relationship between count variable Y and p predictor variables $\{X_j, j = 1, 2, .., p\}$ will be modeled by regression model and given a random sample of size n. y_i represent i^{th}observation of response variable $(Y_i = y_i)$ and x_{ij} represent value of j^{th}predictor variable $(j = 1, 2, .., p)$ of i^{th}observation $(X_{ij} = x_{ij})$ where $i = 1, 2, .., n$. Suppose $f_1$is probability density function (pdf) of the first process, that is the pdf of Bernoulli distribution and f_2 is the pdf of the second process, that is pdf of negative binomial distribution.

Suppose Z is binary variable that determines if count variable is zero counts or not in the first process. This threshold is what referred as "hurdle" in the model hurdle (Cheng, 2015). If

the count variable has positive value then "hurdle" is crossed and if it has zero value than the "hurdle" is not crossed. Suppose Z has a random sample of size n $(Z_i = z_i)$ where $i = 1, 2, .., n$ and z_i represent the value of binary variable. Assume Z has Bernoulli distribution with the following definition,

$$Z = \begin{cases} 0, \text{for zero counts } (Y_i = 0), \\ 1, \text{ for positive counts } (Y_i > 0). \end{cases}$$

Pdf of binary variable Z is,

$$f_1(z_i) = p_i^{z_i}(1 - p_i)^{1 - z_i}$$

When "hurdle" is crossed then only the positive counts will be analyzed in the second process. Therefore, zero counts in the negative binomial (f_2) will be truncated.

$$P(Y_i = y_i | Y_i > 0) = \frac{f_2(y_i)}{1 - f_2(0)}, \; y_i = 1, 2, 3, \ldots$$

$$= \frac{1}{1 - (1 + a\mu_i)^{-a^{-1}}} \frac{\Gamma(y_i + a^{-1})}{\Gamma(y_i)\Gamma(a^{-1})} \left(\frac{a^{-1}}{\mu_i + a^{-1}}\right)^{a^{-1}} \left(\frac{\mu_i}{\mu_i + a^{-1}}\right)^{y_i}, y_i = 1, 2, 3, \ldots$$

By using probability law of total, then the pdf of negative binomial hurdle model is (Tian, 2018),

$$P(Y_i = y_i) = f_{HBN}(y_i) = \begin{cases} 1 - p_i, & y_i = 0 \\ \frac{p_i}{1 - (1 + a\mu_i)^{-a^{-1}}} \frac{\Gamma(y_i + a^{-1})}{\Gamma(y_i)\Gamma(a^{-1})} \left(\frac{a^{-1}}{\mu_i + a^{-1}}\right)^{a^{-1}} \left(\frac{\mu_i}{\mu_i + a^{-1}}\right)^{y_i}, & y_i = 1, 2, 3, .. \end{cases}$$

where:

p_i : Probability of count variable has a value other than zero.

a : Dispersion parameter.

μ_i : Mean of positive counts.

$\Gamma(.)$: Gamma function.

In the first process or binary process, relationship between binary variable and p predictor variables is modeled by logistic regression model.

$$logit\ (p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} = \boldsymbol{X_i^T \beta}$$

In the second process, relationship between positive counts of count variable and *k* predictor variables is modeled by zero-truncated negative binomial regression model.

$$\ln \mu_i = \gamma_0 + \gamma_1 G_{i1} + \cdots + \gamma_k G_{ik} = \boldsymbol{G_i^T \gamma}$$

where:

β_j : j^{th} regression coefficient of logistic regression model where *j = 0, 1, 2, .., p.*

X_{ij} : j^{th} predictor variable of logistic regression model where *j = 0, 1, 2, .., p* and ($X_{io} = 1$).

γ_j : j^{th} regression coefficient of zero-truncated negative binomial regression model where *j = 0, 1, 2, .., k.*

G_{ij} : j^{th} predictor variable of zero-truncated negative binomial regression model where *j = 0, 1, 2, .., k.* and ($G_{io} = 1$).

Predictor variables used in logistic regression model and zero-truncated negative binomial regression model can be different, but this paper provide the same predictor variables in both models. So, this paper uses $k = p$ predictor variables.

Method of Analysis

Regression coefficients of hurdle negative binomial regression model and dispersion parameter are estimated by Bayesian method. Bayesian method assumes that the parameter has distribution for representing its uncertainty. Bayesian method consists of prior distribution, likelihood function, and posterior distribution. Information or belief about unknown parameter before data observed is represented by prior distribution. Prior belief will be updated by combining information of the data using Bayes theorem (Press, 2003). The belief held after updating prior belief is called posterior or posterior distribution. Posterior distribution is used to obtain parameter inference. The Bayes theorem can be written as (Hoff, 2009),

Proceeding of ICSA 2019, p: 57-68
ISBN 978-979-19256-3-1 (PDF)

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

where:

θ : Parameter of interest.

$p(\theta|y)$: Posterior distribution of .

$p(y|\theta)$: Distribution of data given θ or likelihood function.

$p(\theta)$: Prior distribution of θ.

$p(y)$: Marginal distribution of the data.

Assume response variable Y has hurdle negative binomial distribution. The hurdle negative binomial regression model used in this data is,

$$logit\ (p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} = \boldsymbol{X_i^T \beta}$$

and

$$\ln \mu_i = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3} = \boldsymbol{X_i^T \gamma}$$

Prior distribution used in hurdle negative binomial model is Normal (μ, σ^2) with $\mu = 0$ and $\sigma^2 = 10000$ for all regression coefficients and gamma (a_1, b_1) with $a_1 = 0.001$ and $b_1 = 0.001$ for dispersion parameter. These distributions are selected based on non-informative prior distribution. Suppose all these parameters are independent, then the prior distribution is,

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, a) = \left[\prod_{j=0}^{3} \frac{1}{\sigma_{\beta_j}\sqrt{2\pi}} \exp\left[-\frac{\left(\beta_j - \mu_{\beta_j}\right)^2}{2\sigma_{\beta_j}^2}\right]\right] \times \left[\prod_{j=0}^{3} \frac{1}{\sigma_{\gamma_j}\sqrt{2\pi}} \exp\left[-\frac{\left(\gamma_j - \mu_{\gamma_j}\right)^2}{2\sigma_{\gamma_j}^2}\right]\right] \times$$

$$\left[\frac{1}{\Gamma(a_1) b_1^{a_1}} a^{a_1 - 1} \exp\left(\frac{-a}{b_1}\right)\right]$$

Likelihood function used in this data is,

$$p(Y|\boldsymbol{\beta},\boldsymbol{\gamma},a) = \prod_{i=1}^{385} f(y_i;\boldsymbol{\beta},\boldsymbol{\gamma},a)$$

$$= \prod_{i=1}^{241} P(Y_i = y_i|y_i = 0) \times \prod_{i=242}^{385} P(Y_i = y_i|y_i > 0)$$

$$= \left[\prod_{i=1}^{241} 1 - \frac{\exp(\boldsymbol{X_i^T\beta})}{1+\exp(\boldsymbol{X_i^T\beta})}\right] \times$$

$$\prod_{i=242}^{385} \frac{\frac{\exp(\boldsymbol{X_i^T\beta})}{1+\exp(\boldsymbol{X_i^T\beta})}}{1-(1+a\exp(\boldsymbol{X_i^T\gamma}))^{-a^{-1}}} \frac{\Gamma(y_i+a^{-1})}{\Gamma(y_i+1)\Gamma(a^{-1})}$$

$$\left(\frac{a^{-1}}{\exp(\boldsymbol{X_i^T\gamma})+a^{-1}}\right)^{a^{-1}} \left(\frac{\exp(\boldsymbol{X_i^T\gamma})}{\exp(\boldsymbol{X_i^T\gamma})+a^{-1}}\right)^{y_i}$$

Posterior distribution proportional to multiplication of prior distribution and likelihood function is,

$$p(\boldsymbol{\beta},\boldsymbol{\gamma},a|Y) \propto p(\boldsymbol{\beta},\boldsymbol{\gamma},a) \times p(Y|\boldsymbol{\beta},\boldsymbol{\gamma},a)$$

$$\propto \left[\prod_{j=0}^{3} \frac{1}{\sigma_{\beta_j}\sqrt{2\pi}} \exp\left[-\frac{\left(\beta_j - \mu_{\beta_j}\right)^2}{2\sigma_{\beta_j}^2}\right]\right] \times \left[\prod_{j=0}^{3} \frac{1}{\sigma_{\gamma_j}\sqrt{2\pi}} \exp\left[-\frac{\left(\gamma_j - \mu_{\gamma_j}\right)^2}{2\sigma_{\gamma_j}^2}\right]\right] \times$$

$$\left[\frac{1}{\Gamma(a_1)b_1^{a_1}} a^{a_1-1} \exp\left(\frac{-a}{b_1}\right)\right] \times \left[\prod_{i=1}^{241} 1 - \frac{\exp(\boldsymbol{X_i^T\beta})}{1+\exp(\boldsymbol{X_i^T\beta})}\right] \times$$

$$\prod_{i=242}^{385} \frac{\frac{\exp(\boldsymbol{X_i^T\beta})}{1+\exp(\boldsymbol{X_i^T\beta})}}{1-(1+a\exp(\boldsymbol{X_i^T\gamma}))^{-a^{-1}}} \frac{\Gamma(y_i+a^{-1})}{\Gamma(y_i+1)\Gamma(a^{-1})}$$

$$\left(\frac{a^{-1}}{\exp(\boldsymbol{X_i^T\gamma})+a^{-1}}\right)^{a^{-1}} \left(\frac{\exp(\boldsymbol{X_i^T\gamma})}{\exp(\boldsymbol{X_i^T\gamma})+a^{-1}}\right)^{y_i}$$

Posterior distribution is in a non-closed form, so it will be difficult to do analytical calculation. Computational technique is needed to do sampling values from posterior

distribution. Markov Chain Monte Carlo-Gibbs Sampling (MCMC-GS) is one of sampling algorithm that can be used.

Results and Discussion

Parameter estimations obtained by Bayesian method with MCMC-GS implemented in R (R Core Team, 2019). After discarding 50.000 iterations as burn-in, the next 100.000 iterations were drawn as the posterior sample for each of the parameters of interest. Summary of the results is presented in Table 1.

Table 1. Summary of posterior parameter of hurdle negative binomial model.

Parameter	Mean	Std. Deviation	Percentile 2.5	Median	Percentile 97.5
a	**0.435**	0.128	0.240	0.416	0.739
β_0	-0.002	0.009	-0.021	-0.002	0.017
β_1	-0.0002	0.009	-0.018	-0.0002	0.017
β_2	0.007	0.008	-0.009	0.007	0.024
β_3	**-0.012**	0.004	-0.019	-0.011	-0.003
γ_0	0.003	0.010	-0.016	0.004	0.024
γ_1	0.016	0.008	-0.0005	0.016	0.033
γ_2	**0.024**	0.007	0.009	0.024	0.039
γ_3	**0.025**	0.003	0.018	0.025	0.032

Parameter without zero value between percentile 2.5 and percentile 97.5 means that the parameter is significant to the model. Based on Table 1, the significant parameters are $a, \beta_3, \gamma_2,$ and γ_3. The hurdle negative binomial regression model based on parameter estimation is,

$$logit\ (p_i) = \ln\left(\frac{p_i}{1-\ p_i}\right) = -0.002 - 0.0002\ X_{i1} + 0.007\ X_{i2} - 0.012\ X_{i3} \qquad (13)$$

$$\ln \mu_i = 0.003 + 0.016\ X_{i1} + 0.024 X_{i2} + 0.025\ X_{i3}$$

In Table 1, the variable that has a significant parameter in the first model is X_{i3} and based on Equation (13), X_{i3} has parameter with negative value, it means that the greater total score of physical examination of Parkinson's patient (X_{i3}), then the risk or the probability of patient experiencing motoric complication is decreasing by assuming other variables are held constant. In the second model, the variables that have significant parameters are X_{i2} and X_{i3} based on Table 1. Based on Equation (14), each of X_{i2} and X_{i3} have parameters with positive values, it means that the greater total score of motoric aspect (X_{i2}) then the mean of motoric complication frequency is increasing by assuming other variables are held constant. This also applies to physical examination of Parkinson's patient (X_{i3}).

Sample distributions obtained from posterior distribution are shown below.

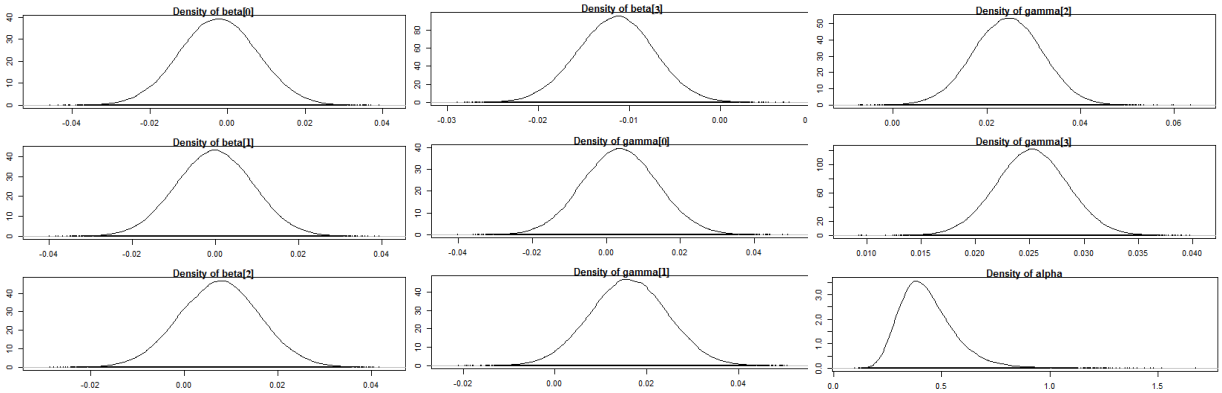

Figure 1. Density plot of sample value of $\beta_0, \beta_1, \beta_2, \beta_3, \gamma_0, \gamma_1, \gamma_2, \gamma_3$ and a.

In MCMC-GS, it is important to test the convergence. One of ways to know the convergence is by analyze the density plot. Based on Figure 1, the sample distribution is unimodal which indicates the parameter estimations are convergent.

Some researches about motoric complication has been discussed, one of the researches is only providing logistic regression model to model the relationship between the event of motoric

complication and some predictor variables (Kadastik-Eerme et al., 2017). Therefore, this paper proposed the hurdle negative binomial model to also model the relationship between the frequency of motoric complication and some predictor variables, so it could give an additional information.

Conclusion

We showed that hurdle negative binomial regression model can be used to handle the excess zeros condition in the data when there is overdispersion due to excess zeros and from other process in positive counts. Based on data application with hurdle negative binomial using Parkinson data, risk factor that explained the occurrence of motoric complication is the total score of physical examination of Parkinson's patient. As for the frequency of motor complications, the identified important risk factors are the total score of motoric aspect and the total score of physical examination of Parkinson's patient.

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

Cheng, Joyce H. (2015). *Bayesian Method for Hurdle Model*. Baylor University.

Hilbe, J. M. (2011*). Negative Binomial Regression (2nd ed.)*. Cambridge: Cambridge University Press.

Hoff, Peter D. (2009). *A First Course in Bayesian Statistical Method*. London, New York: Springer.

Kadastik-Eerme, L., Taba, N., Asser, T. & Taba, P. (2017). *Factors Associated with Motor Complication in Parkinson's Disease.* https://www.ncbi.nlm.nih.gov/pubmed/29075578. (Accessed on July 1, 2019).

Parkinson's Progressive Markers Initiative. (2018). *Motor Assessment: Motoric / MDS-UPDRS*. https://ida.loni.usc.edu/pages/access/studyData. (Accessed on May 28, 2019).

Parkinson's Progressive Markers Initiative. (2019). *About Parkinson Disease.* https://www.ppmi-info.org/about-ppmi/about-pd/. (Accessed on July 1, 2019).

Press, S. James. (2003). *Subjective and Objective Bayesian Statistics (2nd ed.),* Canada: John Wiley & Sons, Inc.

R Core Team. (2019*). R: A language and environment for statistical computing [Computer Software]*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Shehzadi, Anam & Razzaq, Kanza & Tahir, Muhammad & Anwar, Zahid & Raza, Akash.(2018). *Parkinson's disease*. International Journal of Applied Biology and Forensics. 2.

Song, J., Fisher, B. E., Petzinger, G., Wu, A., Gordon, J., & Salem, G. J. (2009). *The Relationships Between The Unified Parkinson's Disease Rating Scale and Lower Extremity Functional Performance in Persons with Early-Stage Parkinson's Disease*. Neurorehabilitation and neural repair, 23(7), 657-661.

Tian, Cheng. (2018). *Hurdle Models in Non-Life Insurance*. Faculty of Mathematics and Physics, Charles University.