Proceeding of ICSA 2019, p: 143-154
ISBN 978-979-19256-3-1 (PDF)

Bayesian Method for Hurdle Regression

S S Hasanah[1], S Abdullah[2*], D Lestari[3].

[1,2,3)]*Department of Mathematics, Universitas Indonesia, KampusBaru UI, Depok, 16424, Indonesia*

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

Abstract

Hurdle model is an alternative model to overcome overdispersion caused by excess zero. The model consists of two stages: a binary process that determines whether the response variable has zero values or positive values, and the second stage to model only the positive counts. The first stage is modelled using binary logistic regression, and the next stage is modeled with the zero-truncated model using Poisson regression. Bayesian method was employed to estimate the models' parameters. Non-informative priors were specified for the parameters, and combined with the likelihood from the data, the non-closed form of posterior distributions were obtained, thus leading to the use of Markov Chain Monte Carlo (MCMC) with Gibbs Sampling to obtain samples from the posterior distributions. This method was applied to model the frequency of motoric complication in people with Parkinson's disease. The result showed that subtotal scores from the three parts of Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) could explain the frequency of motoric complication well, implied by the significance of the regression coefficients.

***Keywords*:** excess zero, Gibbs sampling, MCMC, over-dispersion, two-part model

Proceeding of ICSA 2019, p: 143-154
ISBN 978-979-19256-3-1 (PDF)

Introduction

Count data is an observation in the form of a frequency of an event, where the value is a non-negative integer. Poisson regression is used for the response variable in the form of count data. It assumes that the response variable has the same variance as the mean, known as equidispersion. However, in practice, the response variable has a greater variance than the mean, which is called over-dispersion. One of the causes of overdispersion is the number of zero values in the response, called excess zero (Winkelmann, 2008). Some examples of count data that have many zero values, i.e. the frequency of individuals infected by Escherichia coli (Jalava et al., 2011), frequency of cavities, caries, and broken teeth on a dental examination using the DMF index (Decayed, Missing, Filled) (Hofstetter et al., 2016), and frequency of cocoon population (Jenkinsetal, 2008).

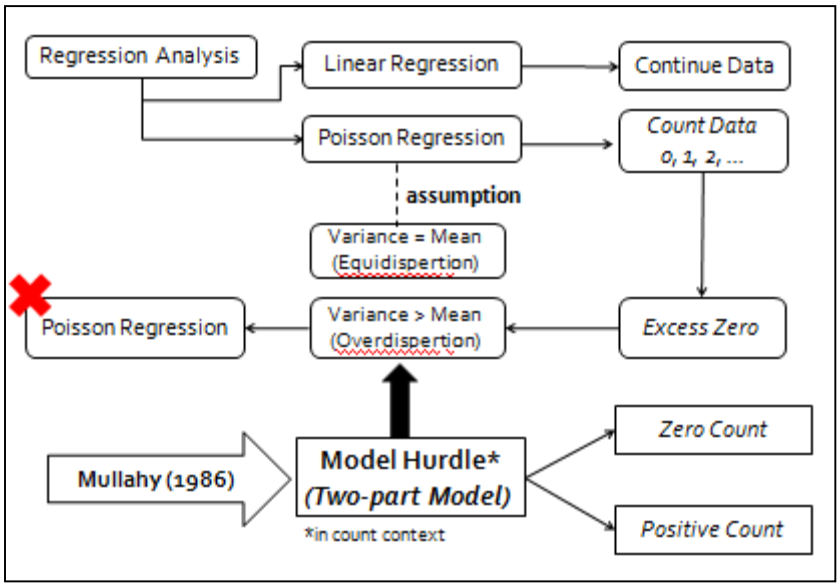

Figure 1. Hurdle model to overcome overdispersion problem in count data regression.

Mullahy (1986) introduced hurdle model that could be used to overcome overdispersion caused by excess zero. It was first developed for a count data context. Previously, the hurdle model was first developed by Cragg (1971) for problems in econometric, which deals with expenditure or consumption data. The modelling process is divided into two stages: a binary process that determines whether the response variable has a zero or a positive value through

binary logistic regression, followed by the process for the positive count only through Poisson regression (Congdon, 2005).

Bayesian approach will be used to estimate the model's parameters. This is due to the flexibility of the Bayesian method, where the inference is conducted based on the posterior information, which contain a relatively more complete information than other method, i.e. the maximum likelihood method. Information from the posterior were formed by likelihood from data combined with the expert judgement through prior distribution; therefore, we prefer the Bayesian approach. Moreover, it was shown that the use of prior information in the Bayesian method is expected to increase accuracy in estimating population parameters (Congdon, 2005).

Materials

Data on people with early Parkinson's disease taken from PPMI database was used to showcase hurdle model for this study (PPMI, 2018). Parkinson's disease is a slow progressive degenerative condition of the central nervous system that affects body movements in daily life caused by a lack of dopamine in the brain (Jenkinsetal, 2008). It has two kinds of symptoms, which is affect movement (motoricsymptoms) and not affect movement (non-motoricsymptom) (DeMaagd et al., 2015). In general, motor symptoms consist of tremors, musclestiffness, and slow movements (bradykinesia), whilen on motor symptoms include sleep problems, anxiety, depression and fatigue (Hayes, 2019).

Patients with Parkinson's disease undergo treatment to reduce Parkinson's symptoms and to prevent Parkinson's symptoms that appear so as not to become more severe. Some treatments that can be done include maintaining a healthy lifestyle, taking antiparkinsonian drugs; and undergo therapy. However, the drugs consumed also have side effects on motor of sufferers of Parkinson's disease, such as motor complications in patients with Parkinson's disease. Motor complications can occur after several years of treatment (APDA, 2017).

The variables used in Parkinson's data were the results of measurements through the MDS-UPDRS instrument such as the total score of MDS-UPDRS Part4 (as the response variable) measuring the frequency of motor complications, MDS-UPDRS Part 1 (X1) which is a test result of non-motor experiences in daily life, MDS-UPDRS Part 2 (X2) which is a test result of motor experiences in daily life, and MDS-UPDRS Part 3 (X3) which is assessment results from the motor sign of Parkinson's disease. This study aims to find out what factors that might explain the frequency of motor complications in people with early Parkinson's disease.

Data consists of observations from 300 patients. Among them, 126 observations (42%) were zero values, implying an excess zero problem, as shown in Figure 2. Therefore, analysis will be conducted separately for the zero counts and the positive counts, through the hurdle model.

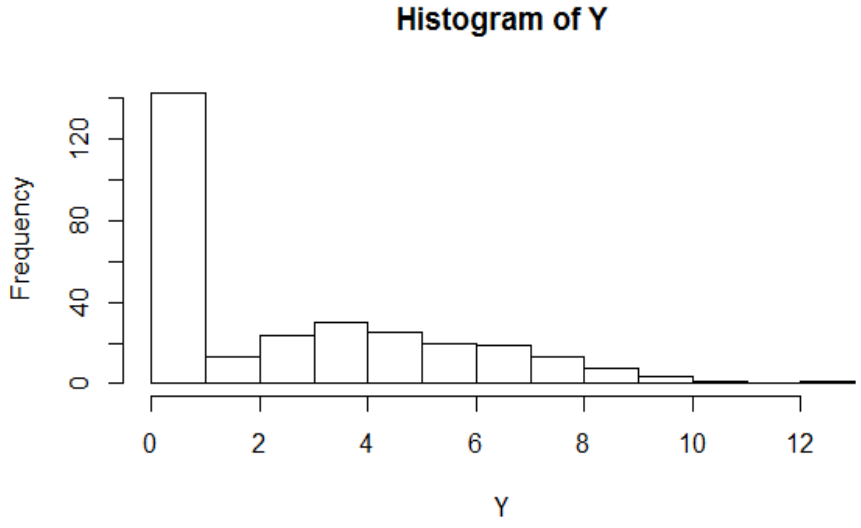

Figure 2. Histogram of number of motoric complications in people with early Parkinson's disease.

The hurdle model specifies two processes that generate zero and positive value. The first process is a determination whether observations are zero or positive. It can be modeled with a binary model. Furthermore, if the first process results in positive observations, then it

will be further analyzed in the second process. It only considers positive values so that observations that are zero will be truncated using a zero-truncated model (Tutz, 2012)

Assumes that f_1 and f_2 are the probability density function (pdf) with support {0, 1, 2, …} where f_1 is a pdf for the first process and f_2 is a pdf for the second process on the model hurdle (Tutz, 2012). In this paper, f_1 was derived from Bernoulli's distribution and f_2 was derived from Poisson's distribution.

Suppose the random variable Y is count data which states the number of events with non-negative integer values. In the first process, suppose S is a binary variable that determines whether the observation is zero values or positive value $S = 0$ means that the zero value is observed, while $S = 1$ means that the positive values are observed (Congdon, 2005). So, the result of the binary process specified by f_1 is as follows

$$\begin{aligned} \Pr(S = 0) = \Pr(Y = 0) = f_1(0) = 1 - \pi \\ \Pr(S = 1) = \Pr(Y > 0) = 1 - f_1(0) = \pi \end{aligned} \tag{1}$$

If a positive value is resulted by the first process, then the observation with positive value is further analyzed in the second process by truncated at zero count model using f_2 with the conditional distribution such that

$$\Pr(Y = y|y > 0) = \frac{\Pr(Y = y, y > 0)}{\Pr(y > 0)} = \frac{\mathrm{f}_2(y)}{1 - \mathrm{f}_2(0)} = \frac{\lambda^y}{(e^\lambda - 1)y!}, y = 1,2, \ldots \tag{2}$$

From this, there is a normalization which show by $1 - f_2(0)$. It tells the truncation at zero of the models (Liu & Powers, 2012).

With use the law of total probability,

$$\Pr(Y = y) = \Pr(S = 0)\Pr(Y = y|S = 0) + \Pr(S = 1)\Pr(Y = y|S = 1) \tag{3}$$

so, pdf for a hurdle model with f_1 from Bernoulli and f_2 from Poisson could be written as (Hilbe, 2014)

$$\Pr(Y = y) = \begin{cases} 1 - \pi & , \quad y = 0 \\ \pi \dfrac{\lambda^y}{(e^\lambda - 1)y!}, & y > 0 \end{cases} \tag{4}$$

where $0 < \pi < 1$ states the probability when the observation crossed zero ($Y > 0$) dan $\lambda > 0$ states the mean frequency of the event.

Suppose Y is the response variable that describes how many occurrences of an observation value are non-negative integers. So, the equation of hurdle regression model can be written as follows

$$logit\ (\pi_i) = \ln \frac{\pi_i}{1 - \pi_i} = \alpha_0 + \ \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_k x_{ik} = \boldsymbol{x_{1i}^T \alpha} \tag{5}$$

and

$$\ln(\lambda_i) = \beta_0 + \ \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \boldsymbol{x_{2i}^T \beta} \tag{6}$$

Where π represent probability for the hurdle at zero is crossed, λ represent mean frequency of event, α represent regression coefficient in model stage 1, β represent regression coefficient in model stage 2, X_{1j} represents the j-th predictor variable in model stage 1, $j = 1,2, \dots, k$, and X_{2j} represent the l-thpredictor variable in model stage 2, $l = 1,2, \dots, p$.

The set of explanatory variables for the model at both stages could be set different. In this paper, it is assumed that the model at both stages uses the same predictor variable, so $k = p$.

Method of Analysis

The Bayesian approach assumes that parameters, for example, θ, are random variables that have a certain distribution. Parameter estimation in Bayesian considering the prior information. Prior information will be updated after information from the sample is obtained. Both of them will be combined with Bayes theorem as follows (Gelman, 2014)

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{7}$$

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{8}$$

where$p(\theta)$is prior distribution that states an information about θ before there is information from data, $p(y|\theta)$is a likelihood function that connects observations y to θ, $p(\theta|y)$is a posterior distribution that states information about θ after there is information from the data. Posterior distribution has an important role in Bayesian inference and then, $p(y)$ is a marginal likelihood.

Suppose Y is a response, that is the total score of MDS-UPDRS part 4 consisting of 300 observations, $Y_i = y_i$ where $i = 1,2,\dots,300$. Assume that each observation is mutually independent with $i = 1,2,\dots,n_1$ is the observation of zero counts and $i = (n_1 + 1),\dots,300$ is a positive count observation (Congdon, 2005). Thus, it can be obtained the likelihood function as follows.

$$L(\alpha,\beta) = \left[\prod_{i=1}^{300} \frac{1}{1+e^{x_{1i}^T\alpha}} \prod_{i=n_1+1}^{300} e^{x_{1i}^T\alpha}\right]\left[\prod_{i=1}^{300} \frac{e^{x_{2i}^T\beta y_i}}{\left(e^{x_{2i}^T\beta}-1\right)y_i!}\right] \tag{9}$$

In this study, the frequency of motor complications in people with Parkinson's disease, previous information about the parameters to be assessed (α and β) is not much known so that the prior chosen is non-informative prior.

The prior is selected normal distribution with $\mu_\alpha = 0$ and $\sigma_\alpha^2 = 10000$ for the α parameter and similar is also done for the β parameter (HAS, 2009). If the variances used are large enough, then the prior is used in the form of prior non-informative (Ntzoufras, 2009). Thus, the prior distribution for $\alpha \sim N(0,10000)$ and $\beta \sim N(0,10000)$ are as follow

$$\begin{aligned} p(\alpha,\beta) &= p(\alpha)p(\beta)\, L(\alpha,\beta) \\ &= \left[\prod_{j=0}^{3} \frac{1}{100\sqrt{2\pi}} exp\left(-\frac{(\alpha_j)^2}{2(10000)}\right)\right]\left[\prod_{j=0}^{3} \frac{1}{100\sqrt{2\pi}} exp\left(-\frac{(\beta_j)^2}{2(10000)}\right)\right] \end{aligned} \tag{10}$$

Based on the likelihood function in equation (9) and the prior distribution in equation (10), the posterior distribution form is obtained as follows

$$posterior \propto prior\ x\ likelihood$$

$$p(\alpha,\beta) \propto \left[\prod_{j=0}^{3} \frac{1}{100\sqrt{2\pi}} exp\left(-\frac{(\alpha_j)^2}{2(10000)}\right)\right]\left[\prod_{j=0}^{3} \frac{1}{100\sqrt{2\pi}} exp\left(-\frac{(\beta_j)^2}{2(10000)}\right)\right]$$

$$x\left[\prod_{i=1}^{300} \frac{1}{1+e^{x_{1i}^T\alpha}} \prod_{i=n_1+1}^{300} e^{x_{1i}^T\alpha}\right]\left[\prod_{i=1}^{300} \frac{e^{x_{2i}^T\beta y_i}}{\left(e^{x_{2i}^T\beta}-1\right)y_i!}\right] \quad (11)$$

The posterior distribution in equation (11) is used to estimate the values of the parameter α and β. However, the result of the posterior distribution is not closed-form so it is difficult to calculate manually and computational techniques are needed to estimate the parameter values α and β from the posterior distribution.

Results and Discussion

Since the posterior distribution of the parameters of interest is in a non-closed form, numerical simulation using Markov Chain Monte Carlo (MCMC) was conducted. Applying the Gibbs sampling using JAGS (Su & Yajima, 2015) accessed through R (R Core Team, 2019), after 10.000 iterations as burn-in, the next 100.000 iterations were taken as the posterior samples for each parameter. Summary of the resultsare shown in Table 1.

Table 1. The results of parameter estimates with Bayesian method

Parameter	Mean	Standard Deviance	2.5 Percentile	Median	97.5 Percentile
α_0	0.00204	0.00997	-0.00471	0.00203	0.00872
α_1	**0.02058**	0.00915	0.01435	0.02060	0.02681
α_2	**0.01463**	0.00844	0.00890	0.00890	0.02034
α_3	**0.01435**	0.00434	0.01141	0.01434	0.01727
β_0	0.00909	0.01001	0.00231	0.00910	0.01582
β_1	**0.02730**	0.00714	0.02248	0.02726	0.03208

β_2	**0.01734**	0.00550	0.01364	0.01733	0.02106
β_3	**0.02678**	0.00234	0.02521	0.02684	0.02838

The parameterswhich the 95% credible interval that do not contain zero values could be signed as the significant parameters (Liu & Powers, 2012). From Table 1, the significant parameters are α_1,α_2, α_3, β_1, β_2, and β_3. It means, all of the explanatory variables (X_1, X_2, X_3) could significantly explain whether patients experienced motor complications of Parkinson's disease, and when they do, how many complications would likely to take place. The model fit for logistic and Poisson regression are

$$logit\ \pi = \ln\frac{\pi}{1-\pi} = 0.002014 + 0.02058X_{11} + 0.01463X_{12} + 0.01435X_{13} \quad (13)$$

$$\ln\lambda = 0.00909 + 0.02730X_{21} + 0.01734X_{22} + 0.02678X_{23} \quad (14)$$

From equation (13) and (14), it can be seen that all the predictor variables have positive signs. It means, all of these associates positively to $logit\ (\pi)$ and $\ln(\pi)$. That is, the greater the value of variables X_{11}, the probability of patients with Parkinson's disease who has experience in motor complications is increasing. The same inference also applies to variable X_{13} and X_{13}. Moreover, the greater value of variablesX_{21}, the frequency of motor complications experienced by patients with Parkinson's disease is also increasing, as well as for X_{22} and X_{23}.

Evaluation of the convergence of model parameters were conducted through the density plots, as depicted in Figure 3. It shows the full posterior distributions of the estimated parameter from the hurdle regression with their predictor variables. Unimodality of the distribution implies the convergence of parameters.

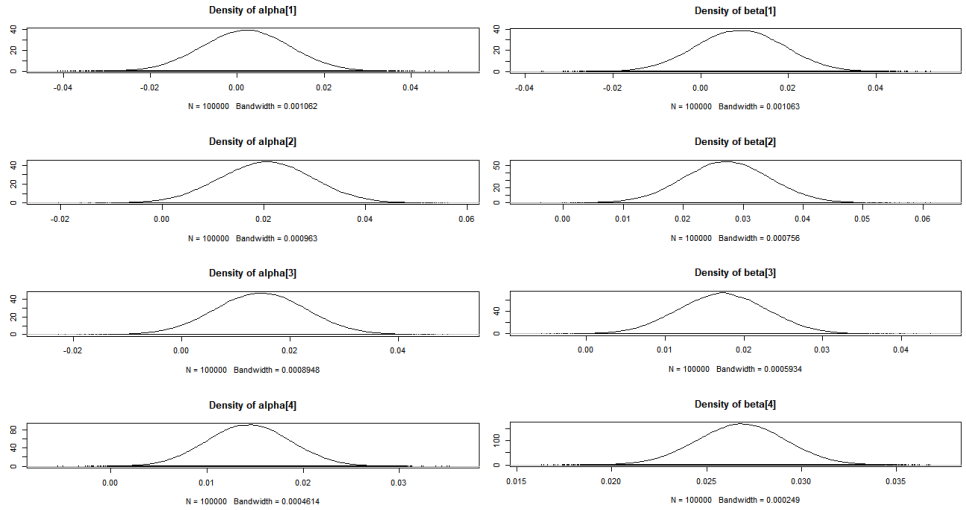

Figure 3. Posterior density plot of α (left column)- the regression coefficient for the 1st stage model- and β (right column)- the regression coefficient for the 2nd stage model.

Conclusions

In this study, hurdle model could successfully fit count data with overdispersion due to excess zeros, as shown by the convergence of the posterior estimates of regression parameters. MDS-UPDRS in both stages of hurdle regression was associated significantly with MDS-UPDRS Part 1; Part 2, and Part 3, showing the importance of the MDS-UPDRS in explaining the severity of Parkinson's disease, where severity in this study is represented by the frequency of motoric complications in people with Parkinson's disease. For future research, other measurements in addition to the MDS-UPDRS would benefit the study to obtain more comprehensive explanation on the motoric complications.

Acknowledgement

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

Aldstadt, J., Koenraadt, C. J. M., Fansiri, T., Kijchalao, U., Richardson, J., Jones, J. W., & Scott, T. W. (2011). Ecological Modeling of Aedes aegypti (L.) Pupal Production in Rural Kamphaeng Phet, Thailand. *PLOS Neglected Tropical Diseases*, 5, 940.

American Parkinson Disease Foundation (APDA). (2017). *Parkinson's Disease Handbook*. New York.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. Wiley.

DeMaagd et al. (2015). Parkinson's Disease and Its Management. *Pharmacy and Theurapetic*, 40(8): 504-510, 532.

Gelman, A. (2014). *Bayesian Data Analysis*. BocaRaton: CRC Press

Haute Autorite de Sante (HAS). (2009). *Indirect Comparison Methods and Validity*. France: Saint-Denis La Plaine Cedex.

Hayes, M. T. (2019). Parkinson's Disease and Parkinsonism. *The American Journal of Medicine*.

Hilbe, J. (2014). *Modeling Count Data*. New York: Cambridge University Press.

Hofstetter, H., Dusseldorp, E., Zeileis, A., & Schuller, A. A. (2016). Modeling Caries Experience: Advantages of the Use Hurdle Model. *Caries Research*, 50, 517-526.

Jalava, K., Ollgren, J., Eklund, M., Siitonen, A., & Kuus, M. (2011). Agricultural, Socioeconomic and Environmental Variables as Risks for Human Verotoxigen Escherichia coli (VTEC) Infection in Finland. *BMC Infectious Diseases*, 11.

Jenkinsetal. (2008). Parkinson's Disease. Edinburgh University.

Liu, H., & Powers, D. A. (2012). Bayesian Inference for Zero-Inflated Poisson Regression. *Journal of Statistics: Advanced in Theory and Applications*, 7(2), 155-188.

Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*, 12(3), 337-350.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Sokoban, NJ: Wiley.

Parkinson's Progressive Markers Initiative. (2018). Motor Assessment. *https://ida.loni.usc.edu/pages/access/studyData* (access on April 4th, 2019).

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/

Su, Y., & Yajima, M. (2015). R2jags: Using R to Run 'JAGS' [Computer Software]. *R package version 0.5-7*. https://CRAN. R-project.org/package=R2jags.

Tian, C. (2018). *Hurdle Model In Nonlife Insurance*. Charles University.

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Berlin: Springer.