Determination of factors associated with motor complications frequency in people with early parkinson's disease: bayesian method for zero-inflated poisson regression

A Awaliah¹, S Abdullah² and A Bustamam³

¹,²,³ Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia

Abstract. Parkinson’s disease (PD) is the second most common neurodegenerative disease worldwide that mainly affects motor system. Treatment given to PD patients may have further complications effect such as dyskinesias. People with PD receiving medication often experience complications. It is interest to identify factors associated with the complications. Data on 215 people with PD obtained from the Parkinson’s Progression Markers Initiative (PPMI) database were analysed. Total scores for the Movement Disorder Society-Unified Parkinson Disease Rating Scale (MDS-UPDRS) Part 1, 2, and 3 were used as the explanatory variables. We proposed Zero-Inflated Poisson (ZIP) to model the frequency of motor complications in people with PD. Therefore, the parameters in ZIP regression were estimated using Bayesian approach. Sampling from the posterior distribution of the parameters is conducted using Monte Carlo Markov Chain-Gibbs Sampling (MCMC-GS). The result shows that total score of MDS-UPDRS Part 1 and 2 are negatively associated with people for no need medication while the opposite is observed for total score of Part 3. Furthermore, in the second stage of the model, total score of Part 3 is negatively associated with frequency of complications, while the opposite trend is observed for total score of Part 1 and 2.

1. Introduction

Parkinson’s disease (PD) is known as the second neurodegenerative disease in the world after Alzheimer's disease and affects 1% of the population aged 65 years and over. PD causes disruption of human motor system. In pathology, PD categorized as a disease caused by a dopamine reduction in the brain substance. These substances serve to organize and coordinate movements by the muscles [1].

According to the Parkinson's Foundation, the early symptoms of PD are the stiffness of the face, tremors in parts of fingers causing difficulty in writing, moving slowly with small steps, loss the ability of smelling, and often experience falls when standing or walking. The standard treatments to cure this disease have not found yet. Every Parkinson's sufferers have different treatment depending on the symptoms he felt. The treatment can be treated with medication or physical therapy.

Therapy given to patients with PD can cause more motor complications such as dyskinesias. This happens due to the influence of excessive medicines effects on patients with PD [7]. Data 215 patients with Parkinson's obtained from the Parkinson's Progression Markers Initiative (PPMI) database were analyzed.
The result showed that 144, or about 67% of patients do not experience motor complications and the rest suffer complications with different severity levels. The result is obtained based on the total score of count data from the test motor complications Movement Disorder Society-Unified Parkinson’s Disease Rating Scales (MDS-UPDRS) Part 4. There were no complications demonstrated through zero response while complications are characterized by a response greater than zero. In other words, data has excess zero condition. Therefore, this aim research was to determined the factors are associated with the frequency of motor complications in PD using Zero-Inflated Poisson regression (ZIP). The parameters in ZIP regression were estimated using Bayesian approach.

2. Data and variables
The selected respondents for this study are Parkinson’s patients with the following criteria.

- Early Parkinson’s or Parkinson’s patient is diagnosed with not more than 6 months.
- Level Hoehn and Yahr scale (HY) not more than 2.

The variables used in this research are following.

1. The response variable ($Y$)
   $Y$ is total score of MDS-UPDRS Part 4 in the second year of treatment.

2. The predictor variables ($X$)
   - $X_1$ is total score of MDS-UPDRS Part 1. That is a test of non-motor experiences daily living such as dizziness, urinating problems, and so on.
   - $X_2$ is total score of MDS-UPDRS Part 2. That is a test of motor experiences daily living such as talking, eating, writing, and etc.
   - $X_3$ is total score of MDS-UPDRS Part 3. It assesses the motor sign of PD such as facial expression, rigidity, hand movements and etc.

3. Zero-inflated poisson (ZIP) regression model
ZIP regression splits the responses into two stages by identifying structural zeros at the first stage and the Poisson counts at the second stage. Zero response in Poisson counts is known as sampling zeros. Structural zeros are the response obtained from an object that the response prediction will always be zero, while sampling zeros are obtained as a response to sampling variability [8]. The first stage is done by assuming the probability of structural zeros as success ($1$) is $\pi$ and probability the Poisson counts as fail ($0$) is $1 - \pi$. The second stage is done by assuming the mean for Poisson counts is $\lambda$. If $Y$ is a random variable distributed ZIP ($\pi$, $\lambda$) then the pdf of $Y$ is

$$Pr(Y = y) = \begin{cases} 
\pi + (1 - \pi)e^{-\lambda}, & y = 0 \\
(1 - \pi)e^{-\lambda}\frac{\lambda^y}{y!}, & y > 0.
\end{cases}$$

ZIP regression models have two sets of parameters corresponding to $\pi$ and $\lambda$. There are $\alpha$ and $\beta$. The regression model with covariates associated with $\pi$ is called the logistic regression function while dealing with $\lambda$ is the Poisson regression function [6].

In many cases, researchers often can not determine the number of structural zeros or sampling zeros. If he knows the predictor variables that affect the zero response, he could define the logistic regression function and Poisson regression function by adjusting the independent variable information he obtained. Conversely, if the information about the variables that determine the response of structural
zeros or Poisson counts, the researchers were able to define the function of a logistic regression equation and Poisson regression function using the same variables.

Let $Y$ is variable random distributed of ZIP. $Y = (y_1, y_2, ..., y_p) \sim ZIP(\pi, \lambda)$ by assuming $y_i$ for $i = 1, 2, ... p$ independent. Suppose that in the data, there are $k$ variable predictors in first stage and $l$ variables in second stage (Poisson counts). The regression model function for $Y \sim ZIP(\pi, \lambda)$ is

$$\logit(\pi) = \ln\left(\frac{\pi}{1 - \pi}\right) = X_1\alpha = \alpha_0 + \alpha_1X_{11} + \alpha_2X_{12} + \cdots + \alpha_mX_{1k}$$

(2)

$$\ln(\lambda) = X_2\beta = \beta_0 + \beta_1X_{21} + \beta_2X_{22} + \cdots + \beta_nX_{2l}$$

(3)

where

- $\pi$ : probability for the structural zeros
- $\lambda$ : mean for the Poisson counts
- $\alpha$ : regression parameters in stage 1 (structural zeros)
- $\beta$ : regression parameters in stage 2 (Poisson counts)
- $X_{kj}$ : the j variable predictor at the stage k

4. Bayesian method

The main foundation of Bayesian method is Bayes theorem [2].

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

(4)

Explanation:

- $p(\theta|y)$ is the posterior distribution which represents the probability of the parameter $\theta$ after given the data.
- $p(\theta)$ is a prior distribution which represents the probability of initial parameters $\theta$ before given the data.
- $p(y|\theta)$ is sampling distribution or likelihood. It represents the probability of the data $y$ by assuming the parameter $\theta$ is true
- $p(y)$ is marginal likelihood that represents as summation or integral between prior and likelihood.

Equation (4) can be expressed in other forms as

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

(5)

or

$$\text{posterior} \propto \text{prior} \times \text{likelihood}. \quad (6)$$

The main challenge in Bayesian method is determining the prior distribution. Prior determination is influenced by the parameters of interest. Selecting the prior types are characterized by the value of the used variance. The more information about the parameters, the smaller the variance of the prior distribution. Conversely, the less information, the greater variance.

5. Modelling data with bayesian method for zero-inflated poisson (ZIP) regression

Based on information from the data we obtained, the first stage of a structural model of zeros can be defined as patients who are not taking medications yet so they will not experience motor complications. The model of the second stage of Poisson counts refer to patients who are taking medication. Zero responses on Poisson counts indicate that patients do not experience the
complications. The responses greater than zero indicate the frequency of motor complications. The greater the value of Poisson counts, the higher rate of complications suffered by patients.

5.1. Likelihood for counts data in ZIP model
Suppose \( Y \) is frequency motor complications variable random for ZIP distribution of 215 patients with early PD. \( Y = (y_1, y_2, ..., y_{215}) \sim ZIP(\pi, \lambda). \) Assuming independent and identically \( y_i \) for \( i = 1, 2, ..., 215 \). Suppose that the observations contained \( m \) structural zeros responses and \( n \) Poisson counts, \( m + n = 215 \). Likelihood for \( Y \sim ZIP(\pi, \lambda) \) as follow.

\[
p(Y|\alpha, B) = \prod_{i=1}^{m} \left[ \frac{e^{x_{i\alpha}}}{1 + e^{x_{i\alpha}}} + \left( 1 - \frac{e^{x_{i\alpha}}}{1 + e^{x_{i\alpha}}} \right) e^{-e^{x_{2i\beta}}} \right] \times \prod_{i=1}^{n} \left[ \frac{1}{\lambda} e^{y_i \lambda} \left( \frac{e^{x_{2i\beta}}}{y_i!} \right) \right]. \tag{7}
\]

5.2. Prior distribution for parameters \( \alpha \) dan \( \beta \) in ZIP model
Without prior knowledge about the distribution of the parameter, we choose non-informative priors. In doing so, the role of the prior distribution is minimized and more weight is given to the data in determining the posterior distribution. Therefore, we determine normal distribution by assuming \( \mu = 0 \) and \( \sigma^2 = 1000 \). Prior distribution for \( \alpha_j, \beta_j \sim Normal(0,1000) \) with \( \pi = \frac{22}{7} \) and \( j = 0, 1, 2, 3 \) is

\[
p(\alpha, \beta) = \prod_{j=0}^{3} \left[ \frac{1}{\sqrt{2\pi \sigma_{\alpha j}}} e^{-\frac{-(\alpha_j - \mu_{\alpha j})^2}{2\sigma_{\alpha j}^2}} \right] \times \prod_{j=0}^{3} \left[ \frac{1}{\sqrt{2\pi \sigma_{\beta j}}} e^{-\frac{-(\beta_j - \mu_{\beta j})^2}{2\sigma_{\beta j}^2}} \right]. \tag{8}
\]

5.3. Posterior distribution for sampling parameters in ZIP model
Based on the likelihood and prior distribution we obtained in equation (7) and (8), the posterior distribution as follow.

\[
\text{posterior} \propto \text{prior} \times \text{likelihood}
= \prod_{j=0}^{3} \left[ \frac{1}{\sqrt{2\pi \sigma_{\alpha j}}} e^{-\frac{-(\alpha_j - \mu_{\alpha j})^2}{2\sigma_{\alpha j}^2}} \right] \times \prod_{j=0}^{3} \left[ \frac{1}{\sqrt{2\pi \sigma_{\beta j}}} e^{-\frac{-(\beta_j - \mu_{\beta j})^2}{2\sigma_{\beta j}^2}} \right] \times \prod_{i=1}^{m} \left[ \frac{e^{x_{i\alpha}}}{1 + e^{x_{i\alpha}}} + \left( 1 - \frac{e^{x_{i\alpha}}}{1 + e^{x_{i\alpha}}} \right) e^{-e^{x_{2i\beta}}} \right] \times \prod_{i=1}^{n} \left[ \frac{1}{\lambda} e^{y_i \lambda} \left( \frac{e^{x_{2i\beta}}}{y_i!} \right) \right]. \tag{9}
\]

6. The results for modelling data by bayesian method for ZIP regression model
We apply algorithm of Markov Chain Monte Carlo-Gibbs sampling (MCMC-GS) with 200,000 iterations for sampling the parameters from the posterior distribution. The results shows in Table 1. The bolds typing show the variables associated with parameters affect the model significantly. The significant parameters is showed by no zero between 2.5 percentil and 97.5 percentil [5].

**Table 1.** The results of parameter estimates for 215 patients with early PD using algorithm MCMC-GS for ZIP regression model
Based on Table 1, it can be seen that the total score of the MDS-UPDRS Part 2 and 3 associated with the frequency of motor complications at the first stage regression. It means, whether or not the consumption of medicines by patients is associated with these variables. While the total score of the MDS-UPDRS Part 1 associated with the second stage regression. This suggests that for patients taking the medicines, the variables associated with the frequency of experiencing motor complications. Thus, the logistic regression function and Poisson regression function for ZIP regression model is

$logit (\pi) = X_1 \alpha = 0.648 - 0.03436X_{11} - 0.1053X_{12} + 0.0643X_{13}$. \hspace{1cm} (10)

$ln(\lambda) = X_2 \beta = 1.585 + 0.02077X_{21} + 0.009587X_{22} - 0.005303X_{23}$. \hspace{1cm} (11)

Logistic regression function in equation (10) indicates that the relation between the total score of MDS-UPDRS Part 1 and 2 inversely related with logit (\pi). It suggest, the smaller the variable is, the greater probability for structural zeros. Thus the probability of patients do not experience the complications are increase. While the total score of the MDS-UPDRS Part 3 relates positively to logit (\pi). That is, the greater the variable, the bigger the possibility of PD patients not suffering the motor complications. In equation (11), the correlation of total score MDS-UPDRS Part 1 and 2 are inversely related to ln (\lambda). It shows that the smaller the value of the variables, the greater the probability of experiencing the frequency of complications. While the total score of the MDS-UPDRS Part 3 relates positively to ln (\lambda). Thus, the greater the value of the variable then the frequency of experiencing complications are also getting bigger.

The full posterior distributions of the estimated regression coefficients from the ZIP regression model with covariates are displayed in Figure 1.

**Figure 1.** Posterior density plots of estimated regression coefficients from the ZIP regression model with covariates. “a” represents parameters coefficients in the structural zeros stage and “b” represents parameters coefficients in the Poisson counts stage.
7. **Conclusion**

215 data with early PD Patients were analysed to determine factors that associated with frequency of motor complications. The conclusion of this research is the patients who consume or not medicines are associated with total score of MDS-UPDRS Part 2 and 3. As for the patients consume the medicines, the frequency of motor complications associated with total score of MDS-UPDRS Part 1.

8. **References**

[1] Gibrat C et al 2009 *Differences between Subacute and Chronic MPTP Models: Investigation of Dopaminergic Neural Degeneration and α-synuclein Inclusions* (Journal of Neurochemistry: no 109) p 1469-1482

[2] Hoff P D 2009 *A First Course in Bayesian Statistical Method* (New York: Springer)

[3] Hogg R V and Craig A T 1995 *Introduction to Mathematical Statistics* (New Jersey: Pearson Prentice Hall)

[4] Lambert D 1992 *Zero-Inlates Poisson Regression with an Application to Defects in Manufacturing* (Technometrics vol 34 no 1) p 1-14

[5] Liu H and Powers D A 2012 *Bayesian Inferences for Zero-Inlates Poisson Regression Model* vol 7 no 2 (Journal of Statistics: Advances in Theory and Applications) p 1-34

[6] Montgomery D C et al 2012 *Introduction to Linear Regression Analysis* (New Jersey: John Wiley and Sons)

[7] Parkinson’s Progressive Markers Initiative https://www.ppmi.org

[8] Tang W et al 2014 *Structural Zeros and Zero-Inflated Models* (Journal of Mathematics vol 3)

9. **Acknowledgments**

This research is funded by Universitas Indonesia via PITTA 2018.