ScienSage

Journal of Advanced Scientific Research

Available online through http://www.sciensage.info/jasr

ISSN
0976-9595
Research Article

A STUDY OF AN EFFICIENCY OF HANDLING OVERDISPERSION USING POISSON REGRESSION AND ZERO INFLATED POISSON REGRESSION: A THALASEMIA CASE STUDY

Wan Muhamad Amir W Ahmad, Nur Syabiha Zafakali, Nurfadhlina Halim, Nor Azlida Aleng, Syerrina Zakaria

Department of Mathematics, Faculty of Science and Technology, Universiti Malaysia Teregganu (UMT), 21030 Kuala Terengganu, Malaysia

**Corresponding author: wmamir@umt.edu.my*

ABSTRACT

Thalassemia is a blood related illness though descendant. Increasing number of patients suffering from thalassemia, especially among children has been reported year by year and has been identified ahead of other hereditary diseases in many parts of the world especially in Malaysia. Therefore, the increasing numbers of this illness annually, attracts the interest among the researchers to put an extra effort in order to overcome this illness. Other than that, most patients were also exposed other chronic diseases such as hemorrhagic illness, health problems, heart failure, influenza, anemia, pneumonia, acute bronchitis, asthma, acute tonsillitis, jaundice and tuberculosis. Data from patients especially among children has been successfully collected and it has been manifested in the form of statistic for analysis. The collected data then has been categorized according to certain scales appropriate to the analysis conducted. Due to the existence of many zero values in the data, the data is said to be suffering from over dispersion. This problem can be overcome by means of a zero-inflated Poisson regression model to reduce the value of zero. These models involved in this research are the Poisson regression model and the zero-inflated Poisson regression model. These methods are widely used to analyze count data. Actually these models are part of class of models in generalized linear models (GLM). One requirement of the Poisson distribution model is that the mean is equal to the variance. While the zero-inflated Poisson regression model applied when the mean is smaller than the variance. The results showed that, the percentage of male patients is slightly higher than the female patients and patients consist of Malays (91.0%), Chinese (6.1%), Indians (0.4%) and other races (2.4%). Apart from the research done, it showed that the thalassemia patients have a significant relation to the predictor variables such as health problems 1.0768 (p=0.0001), heart failure 0.9911(p=0.0001), influenza 0.5576 (p=0.0012), anemia 0.3868 (p=0.0001), pneumonia 0.6962 (p=0.0001), acute bronchitis 0.7090 (p=0.0001), asthma 0.8172 (p=0.0001), acute tonsillitis 0.5715 (p=0.0002), jaundice 1.7287 (p=0.0001) and tuberculosis 1.0139 (p=0.0002). However, for variable of hemorrhagic illness was found that the value of p is greater than the significant level and in conclusion, this variable is not significant to the case studies conducted. Based on the selection of the best model, the value of p for Vuong test is positive and statistically significant (p<0.0001, n=930). Thus, the final result from the analysis showed the zero-inflated Poisson regression model is more suitable in order to analyze the data of thalasemia patients among children. This is because, Poisson regression model is most appropriate for data that has no overdispersion while zero-inflated Poisson regression model is most appropriate for data that has overdispersion. This research can be used as the strategy of a better health management especially in patient management, to cut down the number of thalassemia patients among children.

Keywords: Count data, zero-inflation models, overdispersion.

1. INTRODUCTION

An event count refers to the number of times an event occurs. In many fields such as in social, behavioral and biomedical sciences, as well as in public health, marketing, education, biological and agricultural sciences and industrial quality control, the response variable of interest is often measured as a nonnegative integer or count. Count data is very common in various fields such as in biomedical science, public health and marketing. Poisson regression is widely used to analyze count data. Poisson regression is a part of class of models in generalized linear models (GLM). One requirement of the Poisson distribution is that the mean equals the variance. In real-life application, count data often exhibits overdispersion. Overdipersion occurs when the variance is significantly larger than the mean. When this happens, the data is said to be overdispersed. Overdispersion can cause underestimation of standard errors which consequently leads to wrong inference. Besides that, test of significance result may also be overstated. Overdispersion can be handled by using zero-inflated Poisson regression. The results show that Poisson regression is most appropriate for data that has no overdispersion while zero-inflated Poisson regression is most appropriate for data that has overdispersion. The Poisson model is the model which is emphasizing on count data. However, count data often have variance exceeding the mean. In other words, count data usually shows greater variability in the response counts than one would expect if the response distribution truly were Poisson. The phenomenon where the

variance is greater than the mean is called overdispersion. Overdispersion implies that there is more variability around the model's fitted values than is consistent with a Poisson formulation. The overdispersion must be accounted by the analysis methods appropriate to the data. Poisson regression is not adequate for analyzing overdispersed data. Therefore, Osgood claimed that when Poisson fails because of overdispersion, turning to the zero-inflated Poisson regression model can provide a remedy [1]. Zero-inflated Poisson regression is more adequate for overdispersed data. This is because zero-inflated Poisson regression allows for overdispersion since its variance is naturally greater than its mean.

2. POISSON DISTRIBUTION AND ZERO-INFLATED GENERALIZED POISSON REGRESSION

A random variable *y* is said to have a Poisson distribution with parameter μ if it takes integer values $y = 0,1,2,\ldots$ with probability as below.

$$P(Y = y;\mu) = f(y) = \frac{e^{-\mu}\mu^{y}}{y!}, \qquad \mu > 0$$

Where, e=2.7182, μ is the mean of poisson distribution

Poisson equation is expressed as

$$p(y_i;\hat{\beta}) = \frac{[t_i\mu(x_i;\hat{\beta})]^{y_i}\, e^{-t_i[\mu(x_i;\hat{\beta})]}}{y_i!}$$

With

$\mu(x_i;\hat{\beta})$ is the mean of poisson

$\hat{\beta}$ is a vector parameters

Properties of Poisson Probability Distribution

i. The variance is equal to the mean, $E(y) = \mathrm{var}(y) = \mu.$ this property is called equidispersion.
ii. The distribution tends to be skewed to the right.
iii. Poisson distribution with a large mean is often well-approximated by a normal distribution.

2.1. Poisson Regression and Zero Inflated Poisson Regression

Poisson regression is a form of regression analysis used to model count data. For the analysis, the poisson regression model is given as follows:

$$\ln E(y \mid x_i) = \ln(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$

and

$$\mu_i = \exp(x_i^T\beta) = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki})$$

While the equation for the mean and variance for the poisson regression models is $\mu_i = t_i\mu(X_i;\hat{\beta}) = t_i \exp(x_i^T\hat{\beta})$ and $\mathrm{var}(Y_i) = t_i\mu(X_i;\hat{\beta}) = t_i \exp(x_i^T\hat{\beta})$. Next, the generalized oisson regression (GPR) model can be written as the equation below.

$$f(\mu_i,\alpha,y_i) = \left(\frac{\mu}{1+\alpha\mu_i}\right)^{y_i} \frac{(1+\alpha y_i)^{y_i-1}}{y_i!} \exp\left[\frac{-\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right] \qquad (1.1)$$

Where;

$$\mu_i = \mu_i(x_i) = \exp(\textstyle\sum x_{ij}\beta_j)$$

In the model (1.1), α is called the dispersion parameter. When $\alpha = 0$ the probability model in (1.1) reduces to the poisson regression model and this is a caseof equi-dispersion. When $\alpha > 0$ the GPR model in the (1.1) represents count data with over-dispersion. When $\alpha < 0$ the GPR model in the (1.1) represents count data with under-dispersion. A zero-inflated poisson regression (ZIP) is defined as

$$\begin{aligned} P(Y = y_i \mid x_i, z_i) &= \varphi_i + (1-\varphi_i) f(\mu_i,\alpha;0), \quad y_i = 0 \\ &= (1-\varphi_i) f(\mu_i,\alpha;0), \quad y_i > 0 \end{aligned}$$

The mean is given by $E(y_i \mid x_i) = (1-\varphi_i)\mu_i(x_i)$ and the variance is given by

$$V(y_i \mid x_i) = (1-\varphi_i)\left[\mu_i^2 + \mu_i(1+\alpha\mu_i)^2\right] - (1-\varphi_i)^2\mu_i^2 = E(y_i \mid x_i)\left[(1+\alpha\mu_i)^2 + \varphi_i\mu_i\right]$$

This poisson regression model is applied to analyze the number of count data. The poisson regression is a member of a class of generalized linear models (GLM), which is an extension of traditional linear models that allows the mean of a population to depend on a linear predictor through a nonlinear link function and allows the response probability distribution to be any member of exponential family distributions [2]. The use of Poisson regression is vast and the study of this type of regression is continuing. It has been an aid in many research areas such as in economy, epidemiology, sociology and medicine.Poisson regression is useful when the outcome is a count, with large-count outcomes being rare events [3]. It is the most widely used regression model for multivariate count data. Counts are integer and can never be negative. The distribution of counts is discrete and not continuous. Thus, they tend to be positively skewed. Ordinary least squares regression uses the normal distribution as its probability model. Hence, it is fundamentally not a good fit for discrete type data, because the Normal distribution is symmetric and extends from negative to positive infinity [4]. The Poisson distribution is a much better fit for count data. It characterizes the probability of observing any discrete number of events [1]. When the mean count is low, the Poisson distribution is skewed. As the mean count grows however, the Poisson distribution increasingly approximates the normal. Poisson regression uses the Poisson distribution as its probability model. Therefore it is one of the alternatives that can be used for analyzing count data. Poisson regression is a powerful analysis tool but as with all statistical methods, it can be used inappropriately if its limitations are not fully understood. There are three problems that might exist in Poisson regression analysis:

i. the data might be truncated or censored
ii. the data might contain excess zeroes

iii. the mean and variance are not equal, as required by Poisson distribution. This problem is called overdispersion

3. OVERDISPERSION

There are certain phenomena where an observation of zero events during the observation period can arise from two qualitatively different conditions, that is, the condition may result from failing to observe an event during the observation period or an inability to ever experience an event. For example, consider the number of crimes ever committed by each person in a community. In this case, most people are hardly involved in a crime. Therefore, there will be too many zero counts in the data. The basic Poisson regression model is appropriate only if the probability distribution matches the data [1]. Overdispersion (also called extra-Poisson variation) occurs if $Var(y) > \mu(y)$. If standard Poisson model is applied to overdispersed data, the efficiency of parameter estimates remains reasonably high, yet their standard errors are underestimated. To address phenomena with zeroinflated counting processes, the zero-inflated Poisson regression model has been developed.

In order to determine the overdispersion, there are some aspects need to be considered in the problem of excessive dispersion measures.

i. A formal test for excess the overdispersion measures
ii. Standard error for the variable accounting for over dispersion measures
iii. Test statistics for the variables that account for excess dispersion measures
iv. Using the most suitable model for data with excessive dispersion measures

4. CASE STUDY AND RESULTS

To show the utility of the developed approach, we applied the zero-inflated count models to a real data set in underlying Thalassemia disease among children. Thalassemia is a blood disorder passed down through families (inherited) in which the body makes an abnormal form of hemoglobin, the protein in red blood cells that carries oxygen. Thalassemia is common autosomal recessive disorders [5]. This study involved a sample of 930 patients with Thalassemia, representative for children age between 1-12 years. Consider a dataset on children, we may count how many diagnosis children aged 1-12 years had while suffering from Thalassemia. Patient's data with Thalassemia disease were collected at the Medical Record Unit in Hospital Universiti Sains Malaysia (HUSM), Kubang Kerian, Kelantan in north-east Malaysia. For the purpose of this study, we have employed the count data. Due to missing observations for many variables, we have used only some selected variables that are associated with the diagnosis of patients. The diagnosis is considered as the response variable, and the selected variables are: disease of blood, health services, beta Thalassemia, Influenza, Anemia, Pneumonia, Hemoglobin H, Asthma, Acute Tonsillitis, Jaundice and Tuberculosis. The main objective of this paper is to handling overdispersion in a zero-inflated dataset, it is a one-sided test and the level of significance was set as $\alpha = 0.05$. The analysis of these data was performed in SAS 9.3, by using *proc genmod*. In counting the number of response to an exposure, the patient may have no diagnosis response because of his/her immunity or resistence to the desease. Table 1 show the analyses of Kolmogorov-smirnov test towards respons variable (*Y*).

Table 1: Kolmogorov-smirnov test

Poisson parameter (a,b)	Response variable (*y*)
number, *n*	930
Mean	0.9817
Kolmogorov-smirnov	1.047
p value	0.233

Note: significant at level 0.05

Hypothesis for this test:

H_0 : Data distributed according to the poisson distribution

H_1 : Data do not distribute according to the poisson distribution

Based on Table 1, the value of *p* from this test is 0.233 and this value shows greater than significant level 0.233<0.05. Therefore, H_0 is accepted at significant level, α=0.05. The conclusion that we can made are response variable (y) is distributed according to the poisson distribution. The result of implementing poisson regression model is presented in Table 2.

Table 2: Summary of Model Fitting PoissonRegression

Variable	Estimate	Standard error	*P* value
Disease of Blood	0.4518	0.3584	0.2074
Health Problem	1.0768	0.0699	0.0001*
Heart Failure	0.9911	0.0841	0.0001*
Influenza	0.5576	0.1724	0.0012*
Anemia	0.3868	0.0863	0.0001*
Pneumonia	0.6962	0.0989	0.0001*
Hemoglobin H	0.7090	0.0765	0.0001*
Asthma	0.8172	0.0804	0.0001*
Acute Tonsillitis	0.5715	0.1557	0.0002*
Jaundice	1.7287	0.2209	0.0001*
Tuberculosis	1.0139	0.2679	0.0002*
Deviance	1439.2684		
Pearson Chi-square	2454.1498		
AIC[a]	2299.0770		
BIC[b]	2366.7696		
Log likelihood	-1135.5385		
DF	916		
Scale	1.000		

Note: significant at level 0.05

[a]*Akaike's information criteria*; [b]*Bayesian information criterion*

Table 2 shows summary of model fitting poisson regression. From the analysis it is found that the value of *p* for each variable is significant at level 0.05 except for variable tuberculosis. From table 3, it can be seen that the value for deviance is D=1439.2684 and the value for Pearson Chi-square $\chi^2_{0.05} = 2454.1498$. The value of both deviance and Pearson Chi-square are very much larger than their degrees of freedom. Furthemore, dividing each value by their degree of freedom give value greater than 1, which is 1.5713 and 2.6792. Therefore, overdispersion exists in this data and Poisson regression model is clearly not adequate to describe the counts of diagnosis. Another count model which allows for overdispersion is the zero-inflated Poisson regression. Table 3 shows the analysis of data by implementing the model of zero-inflated Poisson regression.

Table 3: Summary of Model Fitting Zero-Inflated Poisson Regression

Variable	Estimate	Standard error	*p* value
Disease of Blood	0.3281	0.3634	0.3665
Health Problem	0.9379	0.0988	0.0001
Heart Failure	0.8854	0.0981	0.0001
Influenza	0.4876	0.1744	0.0052
Anemia	0.3352	0.0887	0.0002
Pneumonia	0.6114	0.1070	0.0001
Hemoglobin H	0.6329	0.0856	0.0001
Asthma	0.6863	0.1022	0.0001
Acute Tonsillitis	0.5058	0.1602	0.0016
Jaundice	1.6281	0.2348	0.0001
Tuberculosis	0.9237	0.2680	0.0006
Deviance	2263.9009		
Pearson Chi-square	2092.5274		
AIC[a]	2295.9009		
BIC[b]	2373.2638		
Log likelihood	-1131.9504		
DF	914		
Scale	2.2356		

Note: significant at level 0.05

[a]*Akaike's information criteria*; [b]*Bayesian information criterion*

From the table it is showed that the *p* value for each variable is significant at level 0.05 except for variable of tuberculosis which has the *p* value greater than 0.05. From the analysis, we found that the value of scale parameter, $(\sqrt{\phi})$ for the model is 2.2356. When $\phi > 1$, its proven that the data were overdispersed.

5. DISCUSSION AND CONCLUSION

In this pape, two different methods have been used (i) poisson regression and (ii) zero inflation poisson regression.This study focuses on the comparison analysis between poisson regression and zero inflation poisson regression. Overdispersion occures when $Var(Y) > \mu$. Overdispersion measures are rare event in analysis.

According to the analysis for modeling data thalassemia, the performance of zero inflation poisson regression is much better compared to the poisson regression patients. Through this comparison, it appears that the zero inflated poisson regression model is used instead of the Poisson regression model because the model has a value of AIC (Akaike Information Criterion) is the smallest. AIC values for the Poisson regression model and zero inflated poisson regression model respectively 2299.0770 and 2295.9009. Besides that, a positif value of statistical vuong test with a significant *p* value shows that the zero inflated poisson regression model is better fitted for the data for thalassemia patients among children. This may provide us with the improved understanding of how to deal with the overdispersion data.

6. REFERENCES

1. Osgood W. *Journal of Quantitative Criminology*, 2000; **16**: 21-43.
2. McCullagh P, Nelder JA. Generalized Linear Models. 2nd Edition. London: Chapman & Hall, 1989.
3. Kutner MH, Nachtsheim CJ, Neter J. Applied Linear Regression Models. 4th ed. New York: McGraw-Hill Companies, Inc. 2004.
4. Atkins DC, Gallop RJ. *Journal of Family Psychology*, 2007; **21**: 726-735.
5. Atifah N, Adam M, Bharu K. Thalassemia among Blood Donors at the Hospital, 2006; **37:** 549-552.

APPENDIX A. Sample SAS code to fit Poisson regression model

```
input Disease_ of_Blood  Health_Disease Heart_failure  Influenza Anemia Pneumonia bronkitis_Akut asma tonsillitis_Akut Jaundis tuberkulosis diagnosis;
cards;
```

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	1	0	0	0	0	0	0	3
0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	0	0	0

```
proc genmod data = thalassemia;
model diagnosis = Disease_ of_Blood  Health_Disease Heart_failure  Influenza Anemia Pneumonia bronkitis_Akut asma tonsillitis_Akut Jaundis tuberculosis/link=log dist = poisson;
run;
```

APPENDIX B. Sample SAS code to fit zero-inflated Poisson regression model

```
data thalassemia;
input Disease_ of_Blood  Health_Disease Heart_failure  Influenza Anemia Pneumonia bronkitis_Akut asma tonsillitis_Akut Jaundis tuberculosis diagnosis;
cards;
```

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	1	0	0	0	0	0	0	3
0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	0	0	0

```
proc genmod data = thalassemia;
model diagnosis = Disease_of_Blood Health_Disease Heart_failure Influenza Anemia Pneumonia bronkitis_Akut asma tonsillitis_Akut Jaundis tuberculosis diagnosis/
dist = zip;
zeromodel child/link = logit;
run;
```