Analysing count data of Butterflies communities in Jasin, Melaka: A Poisson regression analysis.

Siti Afiqah Muhamad Jamil, M. Asrul Affendi Abdullah, Kek Sie Long, Maria Elena Nor, Maryati Mohamed, and Norradihah Ismail

Faculty of Science, Technology and Human Development, Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor

Email: afendi@uthm.edu.my

Abstract. Counting outcomes normally have remaining values highly skewed toward the right as they are often characterized by large values of zeros. The data of butterfly communities, had been taken from Jasin, Melaka and consists of 131 number of subject visits in Jasin, Melaka. In this paper, considering the count data of butterfly communities, an analysis is considered Poisson regression analysis as it is assumed to be an alternative way on better suited to the counting process. This research paper is about analysing count data from zero observation ecological inference of butterfly communities in Jasin, Melaka by using Poisson regression analysis. The software for Poisson regression is readily available and it is becoming more widely used in many field of research and the data was analysed by using SAS software. The purpose of analysis comprised the framework of identifying the concerns. Besides, by using Poisson regression analysis, the study determines the fitness of data for accessing the reliability on using the count data. The finding indicates that the highest and lowest number of subject comes from the third family (Nymphalidae) family and fifth (Hesperidae) family and the Poisson distribution seems to fit the zero values.

1. Introduction

Count data often have an exposure variable, which indicates the number of times the event could happen. Extra zeros are common features of count data in many disciplines including epidemiology, public health, biology, sociology, psychology, engineering, agriculture and especially for ecological data sets [1]. The characteristic of ecological data consists of measuring binary presence or absence, counts of abundance, proportional occupancy rates, or continuous population densities which have high tendency to contain many zero values [2]. Zero-inflated models attempt to account for excess zeros and there is thought to be two kinds of zeros, which are “true zeros” and “false zeros”. Zero-inflated models estimate two equations simultaneously; one for the count model and the other one are the excess zeros which are caused by the real ecological effect of interest. The analysis data with accessing high zero comprised the model of Poisson [3-5], Negative Binomial (NB) regression, zero-Inflated Poisson (ZIP) [6-9], and Zero-Inflated Negative Binomial (ZINB) is widely used [10].

There are two objectives in this research. The first objective is to identify the highest and lowest number of butterflies’ visit in Jasin, Melaka in order to control the extinction for conservation concerns. The second objective is to determine whether Poisson distribution fit the zero value.
In this study, the data set refers to the number of subject visits in that area. Since the researchers are considering on the number of subject, in epidemiologic studies, count data are generally used as the data involves zeros at some risk of the outcome of interest [11]. In terms of butterfly communities, analysing zero value of count data would roughly detect the appropriate condition or factors in order to control the extinction of the subjects. Hence, the researcher could identify which families of butterfly in certain habitat that needs to be preserved and thus, develop an appropriate solution to control the subjects from extinction.

From the previous study, the researchers using Poisson Regression, NB, ZIP model, zero hurdles Poisson (ZHP), ZINB and zero hurdles negative binomial (ZHNB) in the prediction of the number involve nodes. These applications of models involve the health fields and for the scope of the previous study, the researchers examined the data from 1152 patients who underwent axillary dissections in a tertiary hospital in India during January 1993-January 2005. The researchers fit and compare various count models in order to test the model abilities to predict the number of involving nodes. As the results, negative binominal regression model fits the data well compared to zero-inflated Poisson and zero hurdles Poisson regression model but, in spite of that, zero-inflated negative binomial and zero hurdles negative binomial predict the number of nodes more accurately than negative binomial regression model. Hence, in order to get nearest equal performances, zero-inflated negative binomial model was suggested for describing the nodal frequency in breasts cancer patients [4].

In addition, previous researcher [9] proposed a cautionary note regarding count models of alcohol consumption in randomized controlled trials. In this paper, the researchers considered Poisson, overdispersed Poisson, NB, ZIP and ZINB. Moreover, the researchers compared Type-1 error rate in a series of simulation studies of a randomized clinical trial and applied the methods to Addressing the Spectrum of Alcohol Problems (ASAP) trial. As a result, models that provide a poor fit for alcohol consumption data was when using the Poisson model. It did not preserve Type-1 error rates for the randomize group comparison when the true distribution was over-dispersed Poisson. Overall, in simulation studies and in the motivating example, the standard example, Poisson was not robust when fit to overdispersed count data.

Besides, the other researcher [7] proposed on using five regression models which are Poisson, over-dispersed Poisson, NB, ZIP, and ZINB. The models are fitted to data assessing predictors of vigorous physical activity (VPA) among Latina women. It has been discussed on the description of model, analytical, and graphical approach to aid the model selection. There was a little difference between ZIP and ZINB models but ZIP indicates more best fit model.

2. Materials and Methods of Study

In this research, the study used primary data of the butterfly communities in Malaysia. This primary data had been taken from Jasin, Melaka. This data set consists of five families of butterfly which are Papilionidae, Pieridae, Nymphalidae, Lycaenidae, and Hesperidae. Each family has a different number of subfamilies with different types of subjects. There are 131 types of subjects.

In order to perform an analysis from the observation of the data set, the researchers were using Poisson regression model, goodness of fit at 5% level of significance to achieve our objectives. The data was analysed by using SAS software [12]. Attributes information is about the variables of five different types of families which involve the Papilionidae, Pieridae, Nymphalidae, Lycaenidae, and Hesperidae. Each types of family comprised the observations on the number of subject visits at Jasin, Melaka.
Table 1. Quantitative and Qualitative Variables.

| Variable Name | Description | Data Type |
|---------------|-------------|-----------|
| NOOFSUBJECT   | Count data of the number of butterfly visits in Jasin, Melaka | Continuous |
| FAMILY        | 1= Papilionidae 2= Pteridae 3= Nymphalidae 4= Lycaenidae 5= Hesperidae | Nominal |

Table 1 show there are five independent variables and one dependent variable involve in the analysis of count data of butterfly communities.

Table 2. Data step description.

| DATA AFIQ.BUTTERFLY | The data statement has given the name for the output as ‘BUTTERFLY’ and was stored as a permanent file due to ‘AFIQ’ command. |

Table 2 shows the data step description using SAS procedure. The data steps used in order to call off the data from excel and make it as SAS permanent file for further analysis.

Table 3. Proc step description.

| PROC IMPORT | The external data of butterfly was imported from excel file. |
| PROC MEANS  | The number of observations, mean, standard deviation, minimum, and a maximum of the families of butterfly were identified. |
| PROC FREQ   | The frequency, per cent, cumulative frequency, cumulative frequency of origin between the families of a butterfly. |
| PROC UNIVARIATE | This step was used to visualize the count data of zero observations. |
| PROC GENMOD | This data step was used to analyse the Poisson and zero-inflated Poisson distribution. |

From Table 3, in order to analyse the data of butterfly, SAS permit estimation of dispersion parameter $\alpha$, to accommodate over-dispersion. In this study, proc freq have been used for the descriptive analysis while proc gchart used for the output of bar chart of families of butterfly visits in Jasin, Melaka. Moreover, SAS GENMOD procedure (proc genmod) will consider to be applied. It is utilized to produce the outcome of Poisson regression analysis so that the parameter estimates, the AIC and BIC values could be obtained.

3. Results and Discussion

Table 4. Descriptive Analysis.

| Family | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|--------|-----------|---------|----------------------|--------------------|
| 1      | 13        | 9.92    | 13                   | 9.92               |
| 2      | 15        | 11.45   | 28                   | 21.37              |
In general, by referring to the Table 4, the result shows the frequency, percentage, cumulative frequency and cumulative percentage of the families of butterfly visits in Jasin, Melaka. Those outcomes indicate the overall statistics about the visits of butterfly at Jasin, Melaka.

![Figure 1. The bar chart of frequency of families of butterfly visits at Jasin.](image)

Based on Table 4, the frequency of families of butterfly visits Jasin, the results shown the majority of families’ visits Jasin, Melaka was from the third family with the total of 67 number of subject out of 131 numbers of subject. Notably, the least number of subject visits was from the fifth family with 8 number of subject out of 131. The observation indicates that the highest number of subject and the lowest number of subject visits were from the third and fifth family respectively. Hence, the researchers should consider any factors that could increase the number of subject visits in Jasin that comes from the fifth family to control extinction for conservation concerns. On the other hand, the same goes with the results from Figure 1, which indicates the highest and the lowest frequency of butterflies’ visits in Jasin, Melaka that has been visualized by using the bar chart.

3.1. Poisson Regression Model

| Criteria For Assessing Goodness Of Fit | DF | Value | Value/DF |
|----------------------------------------|----|-------|----------|
| Deviance                               | 126| 664.014| 5.27     |
| Scaled Deviance                        | 126| 664.014| 5.27     |
| AIC (smaller is better)                |    | 865.9848|
| BIC (smaller is better)                |    | 880.3608|

| Models | Poisson Regression Model |
|--------|--------------------------|
| AIC    | 865.9848                 |
| BIC    | 880.3608                 |
Table 5 and Table 6 show the analysis of Poisson distribution with the list of the goodness of fit statistics. From the findings, degrees of freedom for this model were 126. In this analysis, the researchers are focusing on the results of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, which are 865.895 and 880.361 respectively.

Table 7. Goodness of fit test.

| Df | Chisq  | P-value |
|----|--------|---------|
| 126| 664.014| 0       |

As shown in the Table 7, the deviance follows a chi-square distribution with degrees of freedom equal to the model residual. From output, degrees of freedom is 126 with chi squared value 664.014 which comes from the estimation of the deviance. Since \( P\text{-value}=0.00<\alpha=0.05 \), hence the study reject \( H_0 \). Therefore, the goodness of fit test is statistically significant which makes the data does not fit the model very well. In this case this study may try to determine if there is an issue of overdispersion.

Table 8. Estimation score for Poisson regression model.

| Parameter   | Estimate | Std. Error | 95% confidence interval |
|-------------|----------|------------|-------------------------|
| Intercept   | 1.056    | 0.209      | 0.647 - 1.465           |
| PAPILIONIDAE| 0.368    | 0.249      | -0.120 - 0.856          |
| PIERIDAE    | -0.818   | 0.267      | -0.704 - 0.343          |
| NYMPHALIDAE| -0.676   | 0.232      | -1.130 - 0.222          |
| LYCAENIDAE  | -0.582   | 0.259      | -1.084 - 0.079          |
| HESPERIIDAE| 0        | 0          | 0                       |

Based on Table 8, the output shows the estimation score of Poisson regression model. The result indicates the output of estimation parameter, standard error and the 95% confidence interval of the analysis. The standard error for all parameters Papilionidae, Pieridae, Nymphalidae, Lycaenidae, and Hesperidae are between the ranges of \( 0.00 \leq \text{standard error} \leq 0.27 \). Since reliability can be expressed in standard error measurement, it is reliable as the size population of the number of subject is 131 number of subject. Besides, for 95% confidence interval, Papilionidae, has range between 0.647 \( \leq \text{confidence interval} \leq 1.456 \), for Pieridae between -0.120 \( \leq \text{confidence interval} \leq 0.856 \), for Nymphalidae, -0.704 \( \leq \text{confidence interval} \leq 0.343 \), for Lycaenidae, -1.130 \( \leq \text{confidence interval} \leq -0.079 \) and for 0.000 range for Hesperidae.

Therefore, final model for the Poisson regression analysis is;

\[
\text{Logit [Pr (Y = 1)]} = 1.056 + 0.368X_1 - 0.181X_2 + 0.676X_3 - 0.582X_4 + 0X_5
\]

Where the variables represent the families of butterfly:

- \( X_1 = \) Papilionidae
- \( X_2 = \) Pieridae
- \( X_3 = \) Nymphalidae
- \( X_4 = \) Lycaenidae
- \( X_5 = \) Hesperidae
4. Conclusion
In this study, we have demonstrated Poisson distribution in analysing count data from zero observation of ecological inference for butterfly communities at Jasin, Malacca.

Firstly, from the results of the frequency table, this study verify that the third family which was Nymphalidae showed the highest numbers of subjects’ visit at Jasin by 98 out of 131 observations with 38.28% while the fifth family which is Hesperiidae showed the lowest numbers of subjects’ visit at Jasin by 23 out of 131 observations with 8.98%. This indicates that, Hesperiidae have higher risk to extinction and need to control for conservation concerns.

Then, proceed with Poisson regression model. Based on the result of Poisson, by taking out the AIC and BIC values, the AIC = 865.985 while for the BIC = 880.361. The smallest values of AIC and BIC would be the most preferred model in analysing zero observations of count data.

Next, for the second objectives, this study test for the goodness of fit by the Poisson distribution model for significant of the study. To assess the fit of the model, this study used the chi-squared value. This assumes the deviance follows a chi-square distribution with degrees of freedom equal to the model residual. Based on the result of goodness of fit, since $P$-value = 0.00 < $\alpha$ = 0.05, hence, this study reject $H_0$. As expected, from the test, the model was statistically significant. However, this indicates that the data do not fit the model well.

Overall, the Poisson regression model is significant but less fit in analysing the zero observation count data of butterfly communities.

5. Discussion
In general, this study has some limitation on the data being observed for the number of butterflies’ visit in Jasin, Melaka. Since the data does not involve the factors of butterflies’ visit, hence, the studies can only detect for the most predicted families or subfamilies of butterfly that lead to extinction. Therefore, according to the previous literature [13], habitat fragmentation is one of the major causes of local extinction of butterfly communities besides turnover increase and immigration decreases on small isolated compared to large connected islands also affects the extinction of butterflies communities.

Since the current studies are only using Poisson regression model there are not enough evidence to only used Poisson regression model in analysing the count data of zero observations as it can be compared more precisely using other analysis. Hence, further research should try to analyse this count data of butterfly communities’ using different types of analysis such as, NB model, ZINB, hurdle model, and zero-inflated hurdle model in analysing the count data of zero observation if necessarily.

In conclusion, this study only analysed zero observation of ecological data using Poisson regression analysis. Based on the objective respectively, the highest and lowest number of subject have been detected so that higher expected families of butterfly that lead to extinction could be preserved. Overall, all statistical analyses and plots were performed using SAS GENMOD, SAS FREQ, and SAS UNIVARIATE procedures.

Acknowledgement
We would like to thank the Ministry of Higher Education and Universiti Tun Hussien Onn Malaysia, UTHM for supporting this research project.

References
[1] Liu, Y. and G.-L. Tian. 2015. *Computational Statistics & Data Analysis*; 83: 200-222.
[2] Martin, T.G., Wintle, B.A, Rhodes, J.R., Kuhnert, P. M., Field, S.A., Low-Choy, S.A., Tyre, A.J and Possingham, H.P. 2005. *Ecology Letters*; 8(11): 1235-1246.
[3] Coxe, S., West, S.G., Aiken, L.S. 2009. *J Pers Assess*; 91(2): 121-136.
[4] Ahmed, S.A., Sadiqqa, J.S., Quaiser, Saghir. and Kamal, S. 2015. *Journal of Basic and Applied Sciences*; 11:6679.
[5] Lambert, D. (1992). *Technometrics*; 34: 1-14.
[6] Rose, C. E., Martin, S.W., Wannemuehler, K.A. and Plikaytis, B. D. 2006. *J Biopharm Stat*; 16(4): 463-
[7] D. J. Slymen, G. X. Ayala, E. M. Arredondo, and J. P. Elder, Epidemiologic Perspectives & Innovation 3, 2006. SAS Institute.

[8] Neelon, B.H., O’Malley, A.J., and Normand, S.L.T. 2010. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. National Institute of Health; 10(4): 421-439.

[9] Horton, N.J., Kim, E., and Saitz, R. 2007. BioMed Central Medical Research Methodology; 7: 9.

[10] Mohd Asrul, A.A., Naing, N.N., and Kudus, A. 2012. Qualitative and Quantitative Economics Research.

[11] Dwivedi, A.K., Dwivedi, S.N., Deo, S., Shukla, R., and Kopras, E. 2010. National Institute of Health; 2(7): 641-651.

[12] Inc.: SAS/STAT User’s Guide, Version 9.3 Cary, NC: SAS Institute Inc, 2014.

[13] Preston, K.L., Redak, R.A., Allen, M.F., and Rotenberry, J.T. 2012. Biological Conservation; 152: 280-290.