Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Detecting overdispersion in count data: A zero-inflated Poisson regression analysis

To cite this article: Siti Afiqah Muhamad Jamil *et al* 2017 *J. Phys.: Conf. Ser.* **890** 012170

View the article online for updates and enhancements.

Related content

- Analysing count data of Butterflies communities in Jasin, Melaka: A Poisson regression analysis.
  Siti Afiqah Muhamad Jamil, M. Asrul Affendi Abdullah, Sie Long Kek et al.
- Sparse estimation in high-dimensional zero-inflated Poisson regression model
  Mengmeng Jiang and Hang Zhang
- Modified Regression Correlation Coefficient for Poisson Regression Model
  Nattacha Kaengthong and Uthumporn Domthong

IOP ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

This content was downloaded from IP address 163.47.32.70 on 15/01/2020 at 03:19

Detecting overdispersion in count data: A zero-inflated 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. This study focusing on analysing count data of butterflies communities in Jasin, Melaka. In analysing count dependent variable, the Poisson regression model has been known as a benchmark model for regression analysis. Continuing from the previous literature that used Poisson regression analysis, this study comprising the used of zero-inflated Poisson (ZIP) regression analysis to gain acute precision on analysing the count data of butterfly communities in Jasin, Melaka. On the other hands, Poisson regression should be abandoned in the favour of count data models, which are capable of taking into account the extra zeros explicitly. By far, one of the most popular models include ZIP regression model. The data of butterfly communities which had been called as the number of subjects in this study had been taken in Jasin, Melaka and consisted of 131 number of subjects visits Jasin, Melaka. Since the researchers are considering the number of subjects, this data set consists of five families of butterfly and represent the five variables involve in the analysis which are the types of subjects. Besides, the analysis of ZIP used the SAS procedure of overdispersion in analysing zeros value and the main purpose of continuing the previous study is to compare which models would be better than when exists zero values for the observation of the count data. The analysis used AIC, BIC and Voung test of 5% level significance in order to achieve the objectives. The finding indicates that there is a presence of over-dispersion in analysing zero value. The ZIP regression model is better than Poisson regression model when zero values exist.

1. Introduction

Count data with excess zeros often occurs in areas such as engineering, epidemiology, psychology, sociology, public health and agriculture [1]. Zero-inflated Poisson (ZIP) regression [2-3] and zero-inflated negative binomial (ZINB) regression are suitable for modelling the count data. There are four scenarios of zero occurrences in ecological data and the modelling approach recommended for presence or absence and for count data, where zero inflation can be caused by false zeros, true zeros or a combination of both. In the absence of zero-inflation, a standard single distribution model is used such as binomial distribution. Besides, when true zeros lead to excess zeros observation, zero-inflated model (ZIP or ZINB) or mixture model are considered. However, when false zeros exist, zero-inflated mixture model should be considered. Overall, when both true and false zero exists, mixture of two or more distributions should be considered [4]. In term of butterfly communities, since the previous researcher

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

had taken the data in Jasin, Melaka, and the observation indicated zero observations, hence true zero exists. In spite of that, further analysis had been conducted to achieve the objectives of this study.

There are two objectives in this research paper. The first objective is to identify the presence of over-dispersion in analysing zero value and the second one is to compare which models would be better than when exists zero value for the observation of count data by extracting the result on the previous researchers and using Voung test of SAS procedure [4].

In this study, the data set refers to the number of subject visits Jasin, Melaka. Since the researchers are considering on the number of subjects, in epidemiologic studies, count data are generally used as the data involves zeros at some risk of the outcome of interest [5]. In addition, this study could help in identifying the appropriate model in analysing the count data of zero observation so that, any error or bias in the analysing process could be controlled and be avoided.

From the previous researcher [7], used ZIP distribution and hurdle models for modeling vaccine adverse event count data. As the data was characterized by excess zeros and heteroscedasticity, the researchers compared several modeling strategies for vaccine adverse event count data. Besides, in general, for public health studies, the researcher may conceptualize zero-inflated models ad allowing zeros to arise from the at-risk and not-at-risk population while hurdle conceptualized as having zeros only from at-risk populations. Using zero-inflated and hurdle models, the data is taken from anthrax vaccine absorbed (AVA) clinical trial study, whereby the number of systemic adverse events occurring after each of four injections was collected for each participant. The researchers assessed the model fits of Poisson, negative binomial (NB), ZIP model, ZINB model, Poisson hurdle (PH), and negative binomial hurdle (NBH) models. Finally, the researchers utilized the robust variance-covariance estimator to account for the repeated measurements on participants.

Furthermore, the recent researchers proposed on using Poisson, NB model, ZIP model and ZINB model to compare and demonstrate it with the incidental sexual behaviour of adolescent girls collected in HIV risk reduction. In this study, when it was applied to current HIV risk reduction intervention data, the researchers fitted the models for each four primary outcomes at three, six and twelve months. The first group is all vaginal sex episodes, followed by unprotected vaginal sex with steady partners, unprotected vaginal sex with other partners and lastly, any unprotected vaginal sex with steady or other partners. The researchers fitted Poisson and negative binomial regression models using the same six covariates and respective baseline measures of the dependent variable. For ZIP and ZINB, all covariates used in Poisson and NB were retained in both parts of the model which are logistic and Poisson for ZIP. Besides, in order to compare the performance of Poisson, negative binomial, zero-inflated negative binomial and zero-inflated Poisson, various indices such as likelihood ratio, Akaike information criterion (AIC), Bayesian information criterion (BIC), and Lagrange multiplier (LM) statistic can be utilised. In addition, the researchers compared the abilities to predict the number of zeros, observed versus predicted probability among the competing models and observed the difference between the predicted and actual counts by mean square error (MSE) performance measure [8].

Furthermore, recent research [9] has proposed on a demonstration of modeling count data with an application to physical activity. For example in this study, the counting outcomes such as days of physical activity or servings of fruits and vegetables often have a distribution that are highly skewed towards the right with excess zeros and posting analytical challenges. To attain the fitted models, the data of vigorous physical activity (VPA) among Latina women used five regression models which are Poisson, over dispersed Poisson, NB model, ZIP and ZINB model. Hence, in this study the ZIP model fit best.

2. Materials and methods of count data analysis

In this research, the researchers were using primary data about the butterfly communities in Malaysia. This primary data had been taken from Jasin, Melaka. This data set consisted of five families of butterfly, which were Papilionidae, Pieridae, Nymphalidae, Lycaenidae, and Hesperidae. Each family had a different number of subfamilies with different types of subjects. There were 131 types of subjects. For Papilionidae, the subjects assigned to be the first variable, which consist of one subfamily, which was

Papilionidae and the subfamily consist of 13 numbers of subjects. For Pieridae, the subjects assigned to be the second variables, which consist of two subfamilies that are Pierinae with 7 types of subjects and Coliadinae with 8 types of subject. For Nymphalidae, the subjects assigned to be the third variables, which consist of four subfamilies that were Danaine with 6 types of subjects, Satyrinae with 14 types of subjects, Morphinae with 7 types of subjects, Nymphalinae with 36 types of subjects and Charaxinae with 4 types of subjects. The fourth variable would be Lycaenidae, which consist of three subfamilies that were Riodininae by 4 types of subjects, Melitinae by 4 types of subjects and Lycaeninae by 20 types of subjects. The last variable was Hesperidae that has one subfamily with eight types of subjects.

In order to perform analysis from the observation of the data set, the researchers were using ZIP regression model, and analysed the Vuong test by using SAS software [10].

In this study, there is one dependent variable and five independent variables which consists of Papilionidae x_1, Pieridae x_2, Nymphalidae x_3, Lycaenidae x_4, and Hesperidae x_5. The number of subject comprised the count data on the types of butterfly visit in Jasin, Melaka while the family consist of five types of families of butterfly, which are the Papilionidae, Pieridae, Nymphalidae, Lycaenidae, and Hesperidae. The data step commands were applied to make it as a permanent file of SAS before proceeding with analysing the data. It is a basic SAS procedure analysing for butterfly communities of count data.

In this study, proc step of SAS procedure will be considered to be applied. There are two SAS procedures that can easily run a ZIP regression which are using **proc genmod** and **proc countreg.** Running **proc genmod** need to specify for both models. Firstly, the count **model** line, followed by the model predicting the certain zero in **zeromodel** line. In this study, the researchers were predicting number of subject with their family and predicting certain zeros with number of subject. Hence, the results of analysis could help in predicting the most preferred model in analysing count data of ecological inference. The dispersion parameter α, indicating that when $\alpha < 1$, the variance is less than its mean and it is under-dispersion while when $\alpha > 1$, the variance larger than its' mean and it is over-dispersion in the data. Besides, indicating that the result could be over-dispersion, proceed with identifying the excess zeros before testing for either NB model or zero inflated models. Hence, **proc univariate, proc means** and **proc freq** would be considered to be applied. However, univariate procedure does not fit the Poisson distribution as it is discrete while it is used for continuous, but Poisson can be visualized by combining SAS procedure.

3. Results of analysis

3.1 Identifying Overdispersion

Table 1. Identifying overdispersion.

Analysis Variable: No Of Subject							
Family	Subfamily	N. Obs	N	Mean	Var	Min	Max
1	A	13	13	**4.154**	**123.808**	0	40
2	B1	7	7	**3.286**	**7.238**	1	8
	B2	8	8	**1.625**	**6.554**	0	7
3	C1	6	6	1	0.4	0	2
	C2	14	14	**0.5**	**0.731**	0	3
	C3	7	7	0	0	0	0
	C4	36	36	**2.111**	**30.102**	0	33
	C5	4	4	2.25	2.25	1	4
4	D1	4	4	1.5	3	0	4
	D2	4	4	1.25	0.917	0	2
	D3	20	20	**1.7**	**18.853**	0	20
5	E1	8	8	**2.875**	**48.125**	0	20

From Table 1, the results show that more than one family have variance > mean which were for the first family, 123.808 > 4.154, the second family with two subfamilies, indicates, 7.238 > 3.286 and 6.554 > 1.625 respectively, for the third family of fourth subfamily by 30.102 > 2.111, the fourth family of the third subfamily with 18.853 > 1.700 and last but not least, the fifth family with 48.125 > 2.875. Since the results have variance > mean, these indicate that overdispersion exists for the count data of butterfly communities in Jasin, Melaka.

3.2 Excess Zeros Observations

Table 2. Detect excess zero observation.

Number of subject				
No Of-Subject	Freq	Percent	Cumulative Frequency	Cumulative Percent
0	56	42.75	56	42.75
1	42	32.06	98	74.81
2	13	9.92	111	84.73
3	7	5.34	118	90.08
4	3	2.29	121	92.37
5	1	0.76	122	93.13
6	2	1.53	124	94.66
7	1	0.76	125	95.42
8	1	0.76	126	96.18
10	1	0.76	127	96.95
20	2	1.53	129	98.47
33	1	0.76	130	99.24
40	1	0.76	131	100

Table 2 shows the excess zero observation. Excess zeros are needed in this analysis in order to gain precision on the validness of the count data so that the researcher could proceed with other analysis. By referring to the table above, the number of subject showed positive result on the presence of zeros observation. The frequency of zeros observation shows highest results by 56 out of 131 number of subject compared to the other subject such as 1 number of subject observed 42 types of butterfly comes to visit Jasin, 2 number of subject observed 13 types of butterfly comes to visit Jasin, 3 number of subject observed 7 types of subject comes to visit Jasin, 4 number of subject observed 3 types of subject comes to visit Jasin and so on.

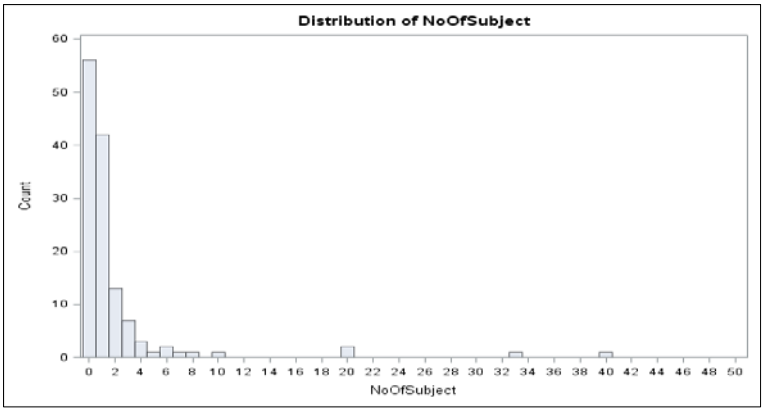

Figure 1: Bar Chart of the frequency of numbers of butterflies.

Therefore, from the previous results, since the observation shows the model does not fit the data well, then, the researchers proceed with detecting the presence of over-dispersion. From the results obtained shows over-dispersion exists. Hence, the researchers proceed with checking on the excess zeros observation for the count data of butterfly communities. Based on the results of analysis on Figure 1, the frequency of the zeros observation of the number of subject visit Jasin exceeds 50 number of subject compared to other subject. Hence, by comparing with table 11, zeros number of subject observed 56 types of subject comes to visit Jasin, Melaka out of 131 number of subject. These results indicate that zeros observation exists as they have a high number of observations. Since excess zero observation occurs, the researchers can precede on ZIP regression model. Otherwise, the researchers need to proceed on NB regression model.

Zero-Inflated Poisson (ZIP) Regression Model

Table 3. Analysis of ZIP regression model.

Criteria For Assessing Goodness Of Fit	
Criterion	Value
Deviance	630.4382
Scaled Deviance	630.4382
AIC (smaller is better)	644.4382
BIC (smaller is better)	664.5646

Table 3 shows the analysis of ZIP distribution with the list of the goodness of fit statistics. From the findings, degrees of freedom for this model were 124. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were observed to compare the best model for the analysis of count data. Continuously, the researchers are focusing on the results of AIC and BIC values, which are 644.438 and 664.565 respectively.

Table 4. Estimation of AIC and BIC score.

Models	Poisson regression model	ZIP model
AIC	865.9848	644.4382
BIC	880.3608	664.5646

As shown in the Table 4, estimation of AIC and BIC are important to indicate the better model in analysing count data of butterfly communities. In this study, since the researchers are focusing on the Poisson regression model and ZIP regression model, starting from the AIC values, for Poisson regression model, it shows 865.985 while for ZIP regression model shows 644.438. On the other hand, for BIC values, Poisson regression model shows 880.361 compared to ZIP regression model by 664.565. Comprising both observations, from AIC and BIC values, ZIP regression model fits better in analysing count data of butterfly communities.

Table 5. Estimation score for ZIP regression model.

Parameter	**Estimate**	**Std. Error**	**95% confidence interval**	
Intercept	1.755	0.209	1.346	2.165
PAPILIONIDAE	0.848	0.249	0.360	1.337
PIERIDAE	-0.578	0.269	-1.104	-0.051
NYMPHALIDAE	-0.787	0.234	-1.247	-0.328
LYCAENIDAE	-0.963	0.259	-1.47	-0.456
HESPERIIDAE	0	0	0	0

Based on Table 5, output shows the estimation score of ZIP 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, Table 5 shows the 95% confidence interval for the butterflies' communities.

Hence, to compare which models would be better when zero value exists could be observed based on Table 5 of the estimation using AIC and BIC value. The AIC value for Poisson regression model was 865.985 while for ZIP regression model was 644.438. These results clearly show that AIC value is much smaller in ZIP regression model compared to Poisson regression model. Besides, the BIC value for Poisson regression model was 880.361 while for ZIP regression model was 664.565. Hence, ZIP regression model shows the smaller value of BIC compared to Poisson regression model. In spite of that, it can be concluded that ZIP regression model would be better when estimating zero value of count data compared to Poisson regression model since both estimations of AIC and BIC show smaller values when estimating data on ZIP regression model.

Therefore, our estimated models for ZIP regression model have been concluded as below;
For ZIP regression model;

$$\text{Logit [Pr (Y = 1)]} = 1.755 + 0.8484\mathbf{X_1} - 0.578\mathbf{X_2} - 0.787\mathbf{X_3} - 0.963\mathbf{X_4} + 0\mathbf{X_5}$$

Where the variables represent the families of butterfly:

$\mathbf{X_1}$ = Papilionidae
$\mathbf{X_2}$ = Pieridae
$\mathbf{X_3}$ = Nymphalidae
X_4 = Lycaenidae
$\mathbf{X_5}$ = Hesperidae

Voung Test and Clarke Sign Test

Table 6. Voung test.

$\boldsymbol{H_0}$: models are equally close to the true value model $\boldsymbol{H_1}$: one of the models is closer to the model			
Vuong statistics	Z-value	Pr>|Z|	Preferred Model
Unadjusted	2.3941	0.00167	ZIP
Akaike Adjusted	2.3941	0.0167	ZIP
Schwarz Adjusted	2.3941	0.0167	ZIP

Based on the Table 6, the Voung test indicates that, if **Pr > |Z|,** we accept H_0. Since **Pr** = 0.0167 < **Z** = 2.394, hence, we reject H_0. This results shows that ZIP regression model is closer to the true model.

Table 7. Clarke Sign test.

H_0: models are equally close to the true model Ha: one of the models is closer to the true model			
Clarke Statistic	M	Pr>=|M|	Preferred Model
Unadjusted	8.5	0.1619	ZIP
Akaike Adjusted	8.5	0.1619	ZIP
Schwarz Adjusted	8.5	0.1619	ZIP

Based on the Table 7, the Clarke Sign test indicates that, if **Pr > |M|,** we accept H0. Since **Pr** = 0.0162 < **M** = 8.5, hence, we reject H_0. This result shows that ZIP regression model is closer to the true model.

Therefore, output above shows the Voung test followed by the Clarke sign test. The positive values of Z statistics for Voung test indicate that it is the first model of ZIP regression model which is closer to the true model. Both of these have the same null hypothesis and it happens that the two tests are consistent since both tests leading to a strong support for the ZIP regression model.

4. Conclusion

In this study, ZIP regression has been analysed and by comparing the results with previous studies, Poisson regression analysis was utilised as the subject to be compared with ZIP model.

By referring to the result of mean procedure, the values of mean and variance have been examined in order to detect the presence of overdispersion. Since most of the results have variance > mean, these results indicates that overdispersion exists for the count data of butterfly communities in Jasin. Therefore, it was clear that overdispersion affects the model fit of zero value.

Moreover, by using SAS univariate procedure and SAS freq procedure, the results shows that there are excess zeros existed with 42.75% by 56 out of 131 number of subject visits in Jasin, Melaka. In this situation, a zero inflated model should be considered than using NB model as the result indicates that the count data for the number of subjects visits Jasin contain excess zero.

Since the data have excess zero, the study proceed with ZIP regression model. Based on the output, the value of AIC and BIC have been taken out for comparing the best model. The result showed the values for AIC = 644.438 and the BIC = 664.565.

Taken together the values of AIC and BIC for both Poisson regression model and ZIP regression model, to address this concern, the same SAS genmod procedure was utilised which was the proc genmod. As a result, the smallest value of AIC and BIC is considered as the best preferred model. Hence, in this study, ZIP distribution indicates the best-preferred model compared to Poisson distribution.

On the other hand, the Voung test and Clarke sign test was used by comprising both Poisson regression analysis and ZIP regression analysis in one procedure of SAS macro procedure. Therefore, the result shows that the most preferred model is ZIP regression analysis because ZIP shows as one of the models which closer to the true model.

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 [11], 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 two types of model in analysing count data of zero observations, which were Poisson regression model and ZIP regression model, there are not enough evidence to use ZIP regression model in analysing the count data of zero observations. 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 compared models that need to be considered in order to analyse zero observation of ecological data. Based on the objective respectively, since the data less fit towards the Poisson regression model, then, overdispersion exists, this study proceed with ZIP regression model as checking excess zeros shows positive results. Thus, ZIP regression model fit best towards count data of butterfly communities. Overall, all statistical analyses and plots were performed using SAS GENMOD, SAS FREQ, SAS MEANS 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] Moghimbeigi, A., Eshraghian, M. R., Mohammad, K., and McArdle, B. 2008. *Journal of Applied Statistics*; **35**(10): 1193-1202.
[2] Lambert, D. 1992. *Technometrics*; 34: 1-14.
[3] Liu, Y. and G.-L. Tian (2015). *Computational Statistics & Data Analysis*; 83: 200-222.
[4] 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.
[5] Siti Afiqah, Muhamad Jamil, M. Asrul Affendi Abdullah, Kek Sie Long, Maria Elena Nor, Maryati Mohamed, and Norradihah Ismail. 2017. Analysis Count Data of butterflies communities in Jasin, Melaka: A Poisson regression analysis.
[6] Dwivedi, A.K., Dwivedi, S.N., Deo, S., Shukla, R., and Kopras, E. 2010. *National Institute of Health*; **2**(7): 641-651
[7] Rose, C. E., Martin, S.W., Wannemuehler, K.A. and Plikaytis, B. D. 2006. *J Biopharm Stat***;** **16**(4): 463-481.
[8] Xia, Y., Beedy, D.M., Ma, J., Feng, C., Cross, W., and Tu, X. 2012. Modeling Count Outcomes from HIV Risk Reduction Interventions: A Comparison of Competing Statistical Models for Count Responses. AIDS Research and Treatment; **11**.
[9] Slymen, D.J., Ayala, G.X., Arredondo, E.M., and Elder, J.P. 2006 *Epidemiologic Perspectives & Innovation*; **3:** 3.
[10] SAS Institute Inc.: *SAS/STAT User's Guide, Version 9.3 Cary, NC*: SAS Institute Inc, 2014.
[11] Preston, K.L., Redak, R.A., Allen, M.F., and Rotenberry, J.T. 2012. *Biological Conservation*; **152**: 280-290.