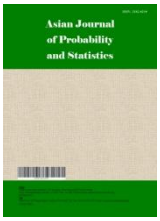

Asian Journal of Probability and Statistics

14(4): 14-20, 2021; Article no.AJPAS.73429
ISSN: 2582-0230

Modelling Claim Frequency in Insurance Using Count Models

A. Adetunji Ademola[1,2*] and Shamsul Rijal Muhammad Sabri[1]

[1]School of Mathematical Sciences, Universiti Sains Malaysia, Penang, Malaysia.
[2]Department of Statistics, Federal Polytechnic, Ile-Oluji, Nigeria.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2021/v14i430334
Editor(s):
(1) Dr. Halim Zeghdoudi, Badji-Mokhtar University , Algeria.
(2) Dr. Manuel Alberto M. Ferreira, Lisbon University,Portugal.
(3) Dr. Dariusz Jacek Jakóbczak, Koszalin University of Technology, Poland.
Reviewers:
(1) J. Visuvasam, India.
(2) Jollanda Shara, Univ. "Eqrem Cabej", Albania.
Complete Peer review History: https://www.sdiarticle4.com/review-history/73429

Original Research Article

Received 27 June 2021
Accepted 07 September 2021
Published 10 September 2021

Abstract

Background: In modelling claim frequency in actuary science, a major challenge is the number of zero claims associated with datasets.
Aim: This study compares six count regression models on motorcycle insurance data.
Methodology: The Akaike Information Criteria (AIC) and the Bayesian Information Criterion (BIC) were used for selecting best models.
Results: Result of analysis showed that the Zero-Inflated Poisson (ZIP) with no regressors for the zero component gives the best predictive ability for the data with the least BIC while the classical Negative Binomial model gives the best result for explanatory purpose with the least AIC.

Keywords: Claims frequency; count models; poisson model; negative binomial model; regression.

1 Introduction

A common task in social science, lifetime modelling, economics, and actuaries is the modelling of count variables. Because empirical dataset in economics and bio-medical sciences often exhibits over-dispersion with excess zeros responses, it is usually counterproductive to assume classical count-observations distributions (like

**Corresponding author: Email: adecap4u@gmail.com;*

Poisson) for the response variable since these distributions require the dataset to have equi-dispersion property [1]. Also, assuming Negative Binomial for the response variable requires the dataset to be over-dispersed i.e. higher variance than mean [2]. A number of modifications have however been made to classical Poisson and Negative Binomial distributions for modelling count observations when some of these assumptions are violated. An approach is the development of mixtures of Poisson distribution with other related distributions: Poisson-Lindley distributions [3], the Poisson-Exponential-Gamma [4], and Poisson-Exponential [4]. Another approach is the assumption of the so-called zero-augmented models that capture zero counts [5,6].

The zero-augmented models combine a zero point mass and a count component [6]. The technique was first used to model number of defects in manufacturing [6]. Since then, this model has been applied in different settings including insurance pricing. The variation had been applied to insurance data by [7] who studied the classical Poisson and logistic regression and compare the findings with a Zero Inflated Poisson (ZIP) model using insurance data from the French motor third party liability. The result shows that the ZIP outperforms the classical Poisson regression. Although it was reported that the logistic regression performs better than the ZIP model. The zero-inflated models was also used to model: the impact of lifestyle and motivations on car crashes [8]; claim frequency of car insurance [9]; number of claims and the number of accidents [10].

To model dataset with count observations, **R** [11], provided the Generalized Linear Models (GLMs) due to [12] with package **pscl** [13] that utilizes design and basic functions of **R** to implement a new function called **zeroinfl()**. The function implements some basic count observation regression models and their respective zero-augmented extensions.

A major challenge with modelling claims frequency in actuary science is the number of zero claims associated with dataset. Having a model that is capable of capturing these zero claims has always been an uphill task. In this study, the Zero Inflated Poisson (no regressors for zero component), Zero Inflated Negative Binomial (no regressors for zero component), Zero Inflated Poisson (with regressors for zero component), and Zero Inflated Negative Binomial (with regressors for zero component) are considered along with the classical Poisson and Negative Binomial distributions.

2 Data

A very pertinent concept in general insurance pricing involves classification of risks and identification of risk characteristics (like age, duration of policy, gender, type of policy, etc.) of the insured to estimate premium [7]. Since most insurance data have confidentiality issue, it is always difficult to obtain more recent data. This research uses data from Wasa (a Swedish Insurance Company). The data contains aggregated data on all insurance claims and policies from 1994 to 1998 on partial Casco Insurance for Motorcycle. The data was obtained from the **R-Package: insuranceData**, a package that contains dataset often used in claims frequency and claims severity models. The data had also been used by [14]. From the 64,548 observations, 6 variables were used to form different count models. From the data, the response variable is the number of claims (***antskad***) recorded by the insurance company within the years under review. The variable is a count variable that can be assumed to follow any of the count distributions (Poisson, Negative Binomial, and Geometric). The number of claims is modelled to be a function of the following risk characteristics: ***agarald*** (the owners age, between 0 and 99), ***zon*** (Geographic zone numbered from 1 to 7, in a standard classification of all Swedish parishes), ***mcklass*** **(**Classification by the EV ratio, defined as $\frac{\text{Engine power in kW x 100}}{\text{Vehicle weight in kg + 75}}$ rounded to the nearest lower integer. The 75 kg represent the average driver weight. The EV ratios are divided into seven classes), ***fordald*** (vehicle age, between 0 and 99), and ***duration*** **(**the number of policy years). Chart 1 shows the number of zeros in the data is highly dominating in comparison with other observations.

Table 1 below shows the frequency and percentages of the number of claims observed (the response variable) in the period under review. The table shows that over 98% of observed responses (number of claims observed) are **0**. This obviously presents a case of highly dispersed data set. Hence, assuming a Gaussian distribution for the response variable could be misleading. Then data gives variance of 0.011518 and mean of 0.010798, depicting an over-dispersed observation.

Table 1. Frequency of figures and their respective percentages

Figure	Frequency	Percentage
0	63878	98.9620
1	643	0.9962
2	27	0.0418
Total	64548	100.0000

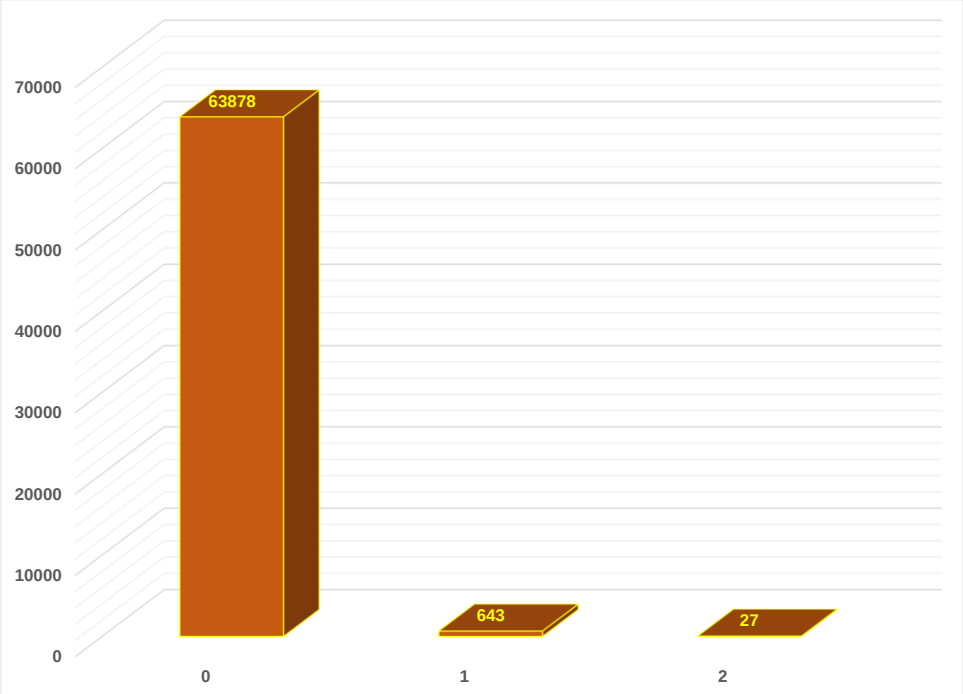

Chart 1. Bar chart showing frequency of observations

3 Models

Classical count data regression modelling involves using the GLM technique pioneered by [12] and extensively studied by [15]. **R language** gives a flexible framework for implementing the GLM framework with the **glm()**function that comes with the **stats package** [16].

Poisson: Poisson distribution is the simplest for modelling count observations. Its density function is defined as: $f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0,1,2,\ldots$ Generally, the distribution is used to describe the mean but usually underestimates the variance in the dataset.

Negative Binomial: This is more useful in modelling over-dispersed count observations. The Negative Binomial distribution with mean = λ and shape parameter θ is defined as: $f(x;\lambda,\theta) = \frac{\Gamma(x+\theta)}{\Gamma(\theta)x!}\frac{\lambda^x\theta^\theta}{(\lambda+\theta)^{x+\theta}}$. The Geometric distribution is a special case of the Negative Binomial distribution when $\theta = 1$. If θ is not specified apriori while using the GLM framework in **R language,** it can be estimated from the data by reusing the GLM technique to iterate the coefficients given θ and vice versa.

Zero Inflated Models: These models give special attention to count data with excess zero (for overview, see [17]). The models combine a point mass to $\boldsymbol{0}$ with distributions like Poisson and Negative Binomial. For example, the Zero-inflated negative binomial regression models count observations with excessive zeros (usually for over-dispersed count observations).

3.1 Model Selection

Assuming six different models for the response variable (claim frequency) for the data used in this research, table 2 shows various values of the Akaike Information Criteria (AIC, [18]) and Bayesian Information Criteria (BIC, [19]). Selecting criterion to utilize in choosing the best among competing models had been discussed [20-23]. AIC and BIC are the two most used information criteria. AIC has been reported to be the optimal for prediction as it is asymptotically equivalent to cross-validation while BIC is reported to be the best for explanation as it allows for consistent estimation of the underlying data generating process [24].

The Bayesian Information Criterion (BIC) has been reported to show superiority in large samples in comparison to the Akaike Information Criterion [24,25]. The distribution with the lowest criterion is adjudge the best. If p is the number of estimated parameters and n is the number of observations and LL is the Loglikelihood, the criteria are defined as: $AIC = 2p - 2LL$; and $BIC = p\ln(n) - 2LL$

In this research, Table 2 shows that Zero Inflated Poisson (no regressors for zero component) is best if the intention is to use the result of the analysis for predictive purpose while the classical Negative Binomial gives best explanation to the data.

Table 2. Selection criteria for the competing models

Model	Distribution	AIC	Rank	BIC	Rank
1	Poisson	7097.581	6	7152.03	4
2	Negative Binomial	7073.283	3	7136.81	1
3	Zero Inflated Poisson (no regressors for zero component)	7078.446	5	7141.97	2
4	Zero Inflated Negative Binomial (no regressors for zero component)	7075.284	4	7147.89	3
5	Zero Inflated Poisson (with regressors for zero component)	7056.080	1	7164.98	5
6	Zero Inflated Negative Binomial (with regressors for zero component)	7058.070	2	7176.05	6

4 Results and Discussion

Table 3 gives parameter estimates for the Negative Binomial model (best for explanation of the observations) while tables 4a and 4b provides parameter estimates for the Zero-Inflated Poisson (with regressors for zero component).

Table 3. Coefficients for negative binomial

	Estimate	Std. Error	z-value	P-value
(Intercept)	-2.020323	0.187518	-10.774	< 2e-16 ***
Agarald	-0.043562	0.003370	-12.925	< 2e-16 ***
Zon	-0.339811	0.031197	-10.893	< 2e-16 ***
mcklass	0.155377	0.026917	5.772	7.82e-09 ***
fordald	-0.070375	0.006313	-11.147	< 2e-16 ***
duration	0.193598	0.013159	14.712	< 2e-16 ***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

The dispersion parameter for the Negative Binomial is *0.4463* with Null deviance: 5386.5 on 64547 degrees of freedom and the Residual deviance: 4748.5 on 64542 degrees of freedom. It is observed that all estimated parameters are highly significant.

Table 4a. Count model coefficients for ZIP (Poisson with log link)

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-0.880163	0.399226	-2.205	0.02748 *
Agarald	-0.042968	0.008264	-5.199	2.00e-07 ***
Zon	-0.386817	0.094949	-4.074	4.62e-05 ***
Mcklass	0.184411	0.064203	2.872	0.00408 **
Fordald	-0.082219	0.013575	-6.057	1.39e-09 ***
duration	0.107593	0.021129	5.092	3.54e-07 ***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 4b. Zero-inflation model coefficients (binomial with logit link)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.868989	0.601996	1.444	0.149
agarald	0.006871	0.012825	0.536	0.592
Zon	-0.049650	0.153132	-0.324	0.746
mcklass	0.045745	0.101286	0.452	0.652
fordald	-0.017934	0.021885	-0.820	0.412
duration	-0.589756	0.130774	-4.510	6.49e-06 ***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 4a contains the Poisson regression coefficients for the covariates while table 4b is the inflation model. All predictors are statistically significant in the Poisson model while only *duration* is significant in the inflation model. Neither of the models indicate superiority of one over the other. This can however be determined using the Vuong statistic [26]. Table 5 shows the comparison of the classical Poisson model and Zero-inflated Poisson Model. With significant probability values, the table shows that the ZIP model explains the reality in the data more than the classical Poisson model.

Table 5. Vuong non-nested hypothesis test-statistic

	Vuong z-statistics	H_A	P-value
Raw	-3.3523793	Model 2 > Model 1	0.0004006
AIC-corrected	-2.6004584	Model 2 > Model 1	0.0046550
BIC-corrected	0.8114448	Model 1 > Model 2	0.2085551

5 Conclusion

From the competing count regression models considered in this study, it is observed that different models compete in explaining observed claim frequency. It can therefore be concluded that exploring different models for any count data is essential as different data can be better explained by different models. Concise clarification should also be made from the onset about the objective of modelling as it has been shown in this study that a model gives better result when the intention is to explain the observations while another gives a better result when the objective is prediction.

In this study using data on insurance data on motorcycle, the Zero-Inflated Poisson with no regressors for the zero component gives the best predictive ability for the data while the classical Negative Binomial model gives the best result for explanatory purpose among the six considered models

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Umar MA, Yahya WB. On the Applications of Some Poisson Related Distributions, Proceedings of 3rd International Conference of Professional Statisticians Society of Nigeria. 2019;(3):458-463.

[2] Shanker R, Hagos F. On Poisson-Lindley Distribution and its Applications to Biological Sciences. International Journal of Biometrics and Biolstatistics. 2015;2(4):00036.
Available:http://DOI:10.15406/bbij.2015.02.0036

[3] Sankaran M. The discrete Poisson-Lindley distribution. Biometrics. 1970;26: 145-149.

[4] Umar MA. A Zero-truncated Poisson–Exponential-Gamma Distribution and its Applications. An M.Sc. dissertation Submitted to the Department of Statistics, University of Ilorin, Ilorin, Nigeria. Unpublished; 2019.

[5] Mullahy J. Specification and Testing of Some Modified Count Data Models. Journal of Econometrics. 1986;33:341–365.

[6] Lambert D. Zero-inflated Poisson Regression, With an Application to Defects in Manufacturing. Technometrics. 1992;34:1–14.

[7] Qazvini M. On the Validation of Claims with Excess Zeros in Liability Insurance: A Comparative Study, Risks. 2019;7:71.
DOI: 10.3390/risks7030071

[8] Lee AH, Mark R, Stevenson KW, Kelvin KW, Yau. Modelling young driver motor vehicle crashes: Data with extra zeros. Accident Analysis and Prevention. 2002;34: 515-21.

[9] Yip Karen, Kelvin CH,Yau KW. On modeling claim frequency data in general insurance with extra zeros. Insurance: Mathematics and Economics. 2005;36:153-63.

[10] Boucher J, Michel D, Montserrat G. Number of accidents or number of claims? An approach with zero-inflated Poisson models for panel data. The Journal of Risk and Insurance. 2009;76:821-46.

[11] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2020.
Available:https://www.R-project.org/

[12] Nelder JA, Wedderburn RWM. Generalized Linear Models. Journal of the Royal Statistical Society A. 1972;135:370–384.

[13] Jackman S. pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University, Stanford, California. R package version 0.95; 2008.
Available:http://CRAN.R-project.org/package=pscl

[14] Ohlsson E, Johansson B. Non-life Insurance Pricing with Generalized Linear Models, EAA Lecture Notes. Springer, Berlin, Heidelberg; 2010.
DOI: https://doi.org/10.1007/978-3-642-10791-7_1

[15] McCullagh, P, Nelder, JA. Generalized Linear Models. 2nd edition. Chapman & Hall, London; 1989.

[16] Chambers JM, Hastie TJ. (eds.) Statistical Models in S. Chapman & Hall, London; 1992.

[17] Cameron AC, Trivedi PK. Regression Analysis of Count Data. Second Edition. Cambridge University; 2013.

[18] Akaike H. A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, AC. 1974;19:716-723.

[19] Schwarz G. Noop. The Annals of Statistics. 1978;6:461-464.

[20] Anderson DR, Burnham KP, White GC. AIC model selection in over-dispersed capture-recapture data, Ecology. 1994;(75):1780-1793.

[21] Anderson DR, Burnham KP, White GC. Comparison of AIC and CAIC for model selection and statistical inference from capture-recapture studies. Journal of Applied Statistics. 1998;25:263-282.

[22] Brewer MJ, Butler A, Cooksley SL. The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity, Methods in Ecology and Evolution. 2016;7:679-692.
DOI: 10.1111/2041-210X.12541

[23] Medel CA, Salgado SC. Does BIC Estimate and Forecast Better than AIC? MPRA Paper No. 42235; 2012.
Available:https://mpra.ub.uni-muenchen.de/42235/

[24] Shmueli G. To Explain or to Predict?" Statistical Science. 201025(3):289–310.
DOI: 10.1214/10-STS330

[25] Henry de-Graft Acquah. Comparison of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in selection of an asymmetric price relationship, Journal of Development and Agricultural Economics. 2010;2(1):1-6.

[26] Desmarais BA, Harden JJ. Testing for Zero Inflation in Count Models: Bias Correction for the Vuong Test, Stata Journal. 2013;(13)4:810-835.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://www.sdiarticle4.com/review-history/73429