Modelling Zero-Inflated Count Data Using Generalized Poisson and Ordinal Logistic Regression Models in Medical Research

Bijesh Yadav[1], Lakshmanan Jeyaseelan[2]*, S Marimuthu[1], Thenmozhi Mani[1], Sebastian George[3] and Shrikant I Bangdiwala[4]

[1]Department of Biostatistics, Christian Medical College, Vellore, Tamil Nadu, India

[2]Mohammed Bin Rashid University of Medicine and Health Sciences College of Medicine, Dubai, UAE United Arab Emirates

[3]Department of Statistical Sciences, Kannur University, Kannur, Kerala, India

[4]Population Health Research Institute and Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada

Received: 13 October 2022

Accepted: 16 July 2023

*Corresponding author: prof.ljey@gmail.com

DOI 10.5001/omj.2024.41

Abstract

Objectives: In medical research, the design of a study and the statistical methods are important since these direct the interpretation and conclusion. The selection of appropriate statistical models depends on the distribution of the outcome measure. Count data are commonly used in medical research, but are often over-dispersed or zero inflated. In some situation count data are also considered as ordinal (maximum value of a outcome is 5), and ordinal regression models can be applied. There are various models available for the analysis of over-dispersed data, such as negative binomial, generalized Poisson, and ordinal regression model, but which one is more appropriate has not been formally evaluated. The objective of this study is to examine whether the generalized Poisson (GP) model can be a better alternative to the ordinal logistic regression model, in Zero inflated Poisson model using simulated and real-time data.

Methods: In this study, we simulated data by providing different estimates of regression coefficients along with various sample sizes and various proportions of zeros and compared GP model and OLR model using the fits statistics. The comparison was also done for the real time datasets.

Results: The bias and mean squared error (MSE) from the simulated results showed consistently lower values in the GP compared to the OLR. The comparison was also done for real-time datasets where the standard error was consistently lower in the generalized Poisson model. The Bayesian information criterion (BIC) values were consistently lower in the GP model than in the OLR model except when the sample size is 1000 and the proportions of zeros were 30% and 40%.

Conclusions: This study has shown that the proposed generalized Poisson model would be a better alternative to the OLR. In addition, the modeling part and interpretation are easier in the GP model as compared to the OLR model.

Keywords: Over-dispersion, count data, ordinal logistic regression, generalized Poisson, simulation

Introduction

Count data is a data that is generated by a process that results in only non-negative integers. In other words, data of this type can only be zero or positive and must increment in whole numbers (eg: 0,1,2,…). Also count referred as the number of times an event occurs within a fixed interval of time (examples: cholera counts reported per week per locality, number of emergencies in referral or secondary care hospitals, number of Antenatal care (ANC) visits).

The usual analysis for count data is Poisson model but it is constrained by the equality of mean and variance. In medical research, it is difficult to get a count outcome with equal mean and variance. An alternative to Poisson model is negative binomial (NB) which account for over-dispersion (variance is more than the mean).When there is over-dispersion due to zeros, Generalized Poisson (GP) model and other models like hurdle and zero-inflated models are recommended.[1-2] The over-dispersion due to zeros are handled using a mixture model and two-part models such as hurdle and zero-inflated as well.[3-4] GP can be the replacement of mixture model and two part-part models for over-dispersed count responses due to zeros. GP model is based on maximum likelihood estimation and useful for both over-dispersion or under-dispersion data.[5-6]

In medical research, ordinal variables frequently express a patient's characteristics, attitude, behavior or status. These variables have natural ordering among the levels such as cancer stages (I, II, III, IV), level of pain (0-10 likert scale), level of satisfaction (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied) and Likert scales measures in the questionnaire survey research (strongly disagree, disagree, agree, strongly agree) etc. A major assumption of ordinal data is the effect of an independent variable on the response variable is constant for unit increase in the level of independent variable. That is, the values are assumed as equally distanced and ordered.

An ordinal outcome analyzed by ordinal logistic regression when certain conditions are met.[7-8] When the outcome of interest is ordinal that satisfies the proportionality assumption and the independent variables are either categorical or ordinal and continuous, then this could also be analyzed using Poisson, zero-inflated Poisson (ZIP), hurdle Poisson and negative binomial etc. This could also be analyzed using proportional odds models.[9] Many studies have reported that statistical data, especially ordinal data used in medical research are not often presented or analyzed according to structure of the data.[10-12]

Many types of ordinal models have been developed. Two most commonly used models are the proportional odds and the continuation ratio models. In this paper we are using only proportional odds model.[13] Both models use maximum likelihood methods to estimate odds ratio but different in dichotomizations of the data which is known as “cut-points”. In both models, the homogeneity effect across cut-points is assumed and a single odds ratio at all cut-points is calculated.[13] The proportional odds models is an extension of binary logistic regression for binary outcomes to allow for ordinal outcomes that have been involved in modeling cumulative logits.[14] It is also named as cumulative logit model. It is based on the assumption of identical log odds ratios across cut-points. At each cut-point, for example, the level of severity categorized as case and non-case. The odds ratio of exposure variable (level of severity) for any of these comparisons will be equal irrespective of cut-point is made. In other words, odds ratio is invariant to the dichotomization of the outcome. The proportional odds model assumes that in the hypothetical population from which this sample was drawn, the odds ratios from each of the two possible dichotomies are the same.

In ordinal responses, for the absence of symptom or disease activity such as none, never or normal provides high proportion of zeros which makes the outcome skewed. In such situations the traditional statistical techniques can provide biased findings. Count models can be appropriate in such situations.

In such situations, GP can be the replacement of mixture model and two part-part models for over-dispersed count responses due to zeros. GP model is based on maximum likelihood estimation and can be used for both over-dispersion or under-dispersion data.[5-6] A study done by Yusuf and Ugalahi (2015) [6], compared Poisson, NB and GP to determine the best fit for the number of antenatal care visits which is over-dispersed due to zeros. Among the three models, GP had the lowest fits statistics values such as log likelihood, AIC and BIC and suggested as a better model to identify the parameters related with the number of antenatal care visits.

Yadav et. al; (2021) [15] compared GP, mixture Poisson, mixture negative binomial and zero-inflated Poisson in terms of fits statistics such as bias, mean square error, AIC and BIC using real time and simulated data. This study shows that GP provides lower values of all fit statistics and identifies as a better model. Hence, the goal of this study is to test whether GP model can be a better alternative to the OLR model in zero-inflated outcomes of counts using simulated and real time data.

Methods

Poisson model has been commonly used for count data. The probability mass function (PMF) of Poisson model is given as:

$$f(y,\theta) = \frac{\theta^y e^{-\theta}}{y!}, y = 0,1,2,..., \theta > 0$$

Poisson model is a single parameter distribution. The mean and variance of Poisson model are equal. In Poisson regression model, interested to model the conditional mean E (θ|x). The expected outcome in terms of log function is given by θ = exp (xβ), Where θ is mean, x is independent variables and β is regression parameters.[16]

For the over-dispersed or under-dispersed count data generalized Poisson (GP) regression model may be also useful. It assumes the outcome variable Y_i has probability mass function (PMF):

$$f(y_i, \theta_i, \delta) = \frac{\theta_i(\theta_i + \delta y_i)^{yi-1} e^{-\theta_i - \delta y_i}}{y_i!}, y_i = 0,1,2,.....$$

Where $\theta_i > 0$ and max $(-1, -\theta_i/4) < \delta < 1$. The mean and variance of GP model are given by

$$Mean\ (\mu_i) = E(Y_i) = \frac{\theta_i}{(1-\delta)},$$

$$Variance\ (Y_i) = \frac{\theta_i}{(1-\delta)^3} = \frac{1}{(1-\delta)^2} E(Y_i) = \phi\, E(Y_i)$$

Where $\phi = 1/(1-\delta)^2$ is a dispersion factor. When δ=0, it becomes to equidispersion, and the GP distribution reduces to Poisson model with parameter θ_i. When $\delta > 0$, means over-dispersion and when $\delta < 0$, then under-dispersion.[17]

The often used OLR model in practice is the proportional odds (PO) model which is also said as cumulative logit model (CLM).The PO model is invariant to collapsing across categories which is frequently needed for the summarizing results.[18] For example, an outcome variable with 4 categories and assume 3 possible ways to divide four categories into two collapsed categories keeping natural order. We might compare group 0 to group 1 to 3, or group 0 and 1 to group 2 and 3, or group 0 through 2 to 3. However, we may not merge groups 0 and 3 for comparison with groups 1 and 2, since that would disrupt the natural ordering from 0 through 3. If an ordinal response variable D has G levels (D =0, 1, 2. . ., G-1), then there are G -1 ways to dichotomize the response outcome: (D ≥1 vs. D < 1; D ≥ 2 vs. D < 2, ... , D ≥ G-1 vs. D < G-1). With such categorization of D, the odds that D ≥ g is defined as:

$$odds(D \geq g) = \frac{P(D \geq g)}{P(D < g)}$$

Where (g = 1, 2, 3. . . G-1). An important assumption of this model is that the odds ratio is identical when odds ratio is calculated in different cut points. For example, OR (D ≥1) = OR (D ≥ 3).[19]

Akaike Information Criteria (AIC) and Bayesian Schwartz Information Criteria (BIC). The AIC and BIC are used for model evaluation. Both models are based on the maximum likelihood estimates. Lower the values, better the model. AIC is defined as:

$$AIC = -2L + 2k$$

Where, L is log-likelihood and k is number of parameters included in the model (number of variables plus intercept).

Similarly, BIC = -2L + k log (n)

Where, L is log-likelihood, k is number of parameters with intercept and n is number of rating classes or number of model observations.[20]

Model preference in terms of AIC and BIC:

Difference of AIC between models A & B	**Result if model A< model B**
>0.1 - ≤ 2.5	No difference in model
>2.5 - ≤ 6.0	Choose model A if n > 256
>6.0 - ≤ 9.0	Choose model A if n > 64
9+	Choose model A
Difference of BIC between two models	**Model Preference**
0 – 2	Weak
2 – 6	Positive
6 – 10	Strong
10 +	Very strong

The commonest criterion in evaluating the performance of a statistical model is based on its accuracy with respect to data fit. The widely used measure of accuracy is the mean squared error (MSE). Smaller value indicates the more accurate and reliable model.

$$MSE : \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$, where n is sample sizes

We first undertook a simulation study in order to evaluate the models under varying sample sizes. To assess how well generalized Poisson model and ordinal model fits data from a varies proportions of zero-inflation (10%, 20%, 30%, 40% & 50%), we generated 5 simulated datasets with varying sample sizes (100, 250, 500, 750 & 1000) for the fixed regression estimates (intercept = 0.5 and slope = 1). Independent variable x was generated from binomial distribution of n observations. Also, an error was generated with mean zero and variance equal to 0.03. Using this independent variable (x) and error, the dependent variable was generated as a linear combination with fixed regression estimates (intercept = 0.5 & slope = 1). This is assumed as the population mean. The antilog of predicted estimate from regression equation was the mean. Using this mean, the data (dependent variable) were generated using the package ZIP (gamlss.dist) under the function rZIP in R software and then combined the dependent variable (y) and independent variables (x). After that, the comparison between GP model and OLR model was done with varying proportions of zeros with varying sample sizes. This entire process was repeated for 1000 times and the median values of fit statistics were reported in the simulation table 1. The Model comparison was based on bias, MSE, AIC and BIC. The better model provides smaller AIC, BIC, Bias and MSE.

Table 1: Simulation based on zero-inflated poisson.

sample sizes	**Generalized Poisson model**				**Ordinal logistic regression model**			
	Bias	**MSE**	**AIC**	**BIC**	**Bias**	**MSE**	**AIC**	**BIC**
				10% zeros				
100	-2.42E-09	2.87	365	373	1.86	6.86	368	393
250	-1.49E-09	2.94	908	919	2.03	7.50	911	949
500	-1.08E-09	2.96	1813	1826	2.09	7.72	1817	1864
750	-7.85E-10	2.97	2714	2727	2.11	7.85	2718	2772

1000	-8.37E-10	2.97	3621	3635	2.10	7.84	3623	3682
20% zeros								
100	-7.23E-09	3.15	363	371	2.01	7.43	364	388
250	-2.08E-09	3.26	903	914	2.05	7.60	901	938
500	-1.35E-09	3.27	1805	1817	2.08	7.69	1794	1841
750	-7.09E-09	3.28	2703	2717	2.08	7.69	2686	2740
1000	-7.71E-10	3.28	3607	3622	2.08	7.70	3582	3641
30% zeros								
100	-3.68E-09	3.32	351	360	2.01	7.46	350	374
250	-2.47E-09	3.42	876	887	2.06	7.68	867	903
500	-1.55E-09	3.42	1749	1761	2.06	7.66	1726	1772
750	-1.32E-09	3.44	2621	2635	1.96	7.44	2586	2640
1000	-1.21E-09	3.43	3493	3508	1.95	7.44	3443	3500
40% zeros								
100	-4.08E-09	3.33	332	339	2.36	8.90	328	351
250	-2.63E-09	3.44	826	836	2.48	10.0	812	848
500	-1.87E-09	3.42	1647	1660	2.50	10.17	1615	1662
750	-1.43E-09	3.44	2464	2469	2.49	10.22	2415	2478
1000	-1.10E-09	3.44	3288	3302	2.50	10.25	3218	3276
50% zeros								
100	-3.48E-09	3.21	303	311	2.20	8.37	297	320
250	-2.49E-09	3.29	754	764	2.25	8.67	736	771
500	-1.75E-09	3.28	1515	1510	2.25	8.69	1464	1510
750	-1.48E-09	3.29	2246	2242	2.25	8.68	2190	2260
1000	-1.21E-09	3.28	2998	2976	2.25	8.70	2920	3012

There are three real time studies were used and these are 1. Fir 2. DMF (D= decayed, M= missing and F= filled teeth) and 3. Acute Diarrheal Disease (ADD). The reason for choosing multiple datasets was to apply GP and OLR models to situations with small numbers of observations to those with large number of observations.

The dataset "Fir" was obtained from R as a 'boot package'. The number of observations for fir data was 50. The objective of this study was to count the number of balsam-fir seedlings in all quadrant of grid of 50 five-foot square quadrants.[21] DMF study was a cross-sectional study on 440 children's dental caries. The objective was to explain oral health status and preventive dental habits of children at different age groups. The distribution of DMF count was positively skewed due to large number of a zeros counts for those children's without having caries experience.[22] ADD data reported (N = 3720) from the communicable disease hospital, Chennai during the period 2008 to 2010. The reported data was based on the case history of the patients across 155 wards of Chennai.

Results

The results of simulation of sample sizes 100, 250, 500,750 and 1000 at varying proportions of zeros at 10%, 20%, 30%, 40% and 50% of the fit statistics, such as bias, MSE, AIC and BIC between GP and OLR model are represented in table 1.

Sample Size 100: The bias, MSE, and BIC estimates of the GP model were lower than OLR model. The AIC values were slightly lower at the 40% and 50% of the zeros in the OLR model but not a significant difference. The details of the preference of better model based on AIC value were presented in the methods.

Sample Size 250: The GP model provides lower values of bias, MSE and BIC than the OLR model. The AIC values were slightly lower in OLR model.

Sample Size 500 and 750: For the size 500 and 750, the bias, MSE, and BIC estimates of the GP model were lower than OLR model. The AIC values were lower in the OLR model but not a significant difference but the other fits statistics were lower in the GP model.

Sample Size 1000: Similarly, for the sample size 1000, the bias, MSE, and BIC estimates of the GP model were lower than OLR model but The AIC values were lower in the OLR model than the GP model.

The bias, MSE provided by GP model was consistently lower than OLR model irrespective of the sample sizes and the proportions of zeros. On an average the MSE provided by the GP model and OLR model were 3.3 and 8.1 respectively with varying sample sizes and proportions of zeros. The BIC estimates were also lower in GP except for sample size 1000 and zeros at 30% and 40%. The pictorial representation of BIC and MSE were presented in the figure 1 and figure 2. The AIC values were slightly lower in OLR or almost equal in both models.

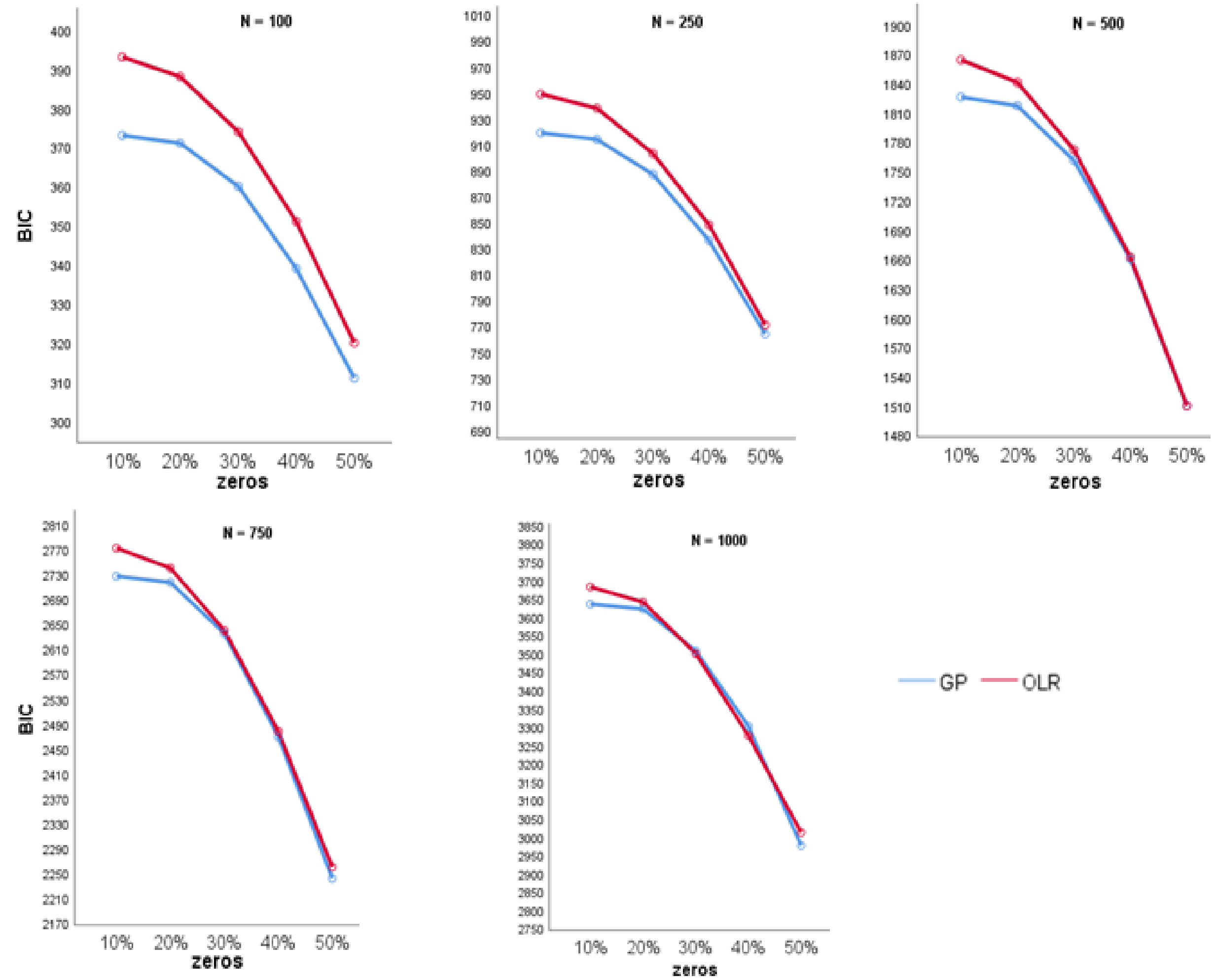

Figure 1: BIC with varying sample sizes and varying proportions of zeros.

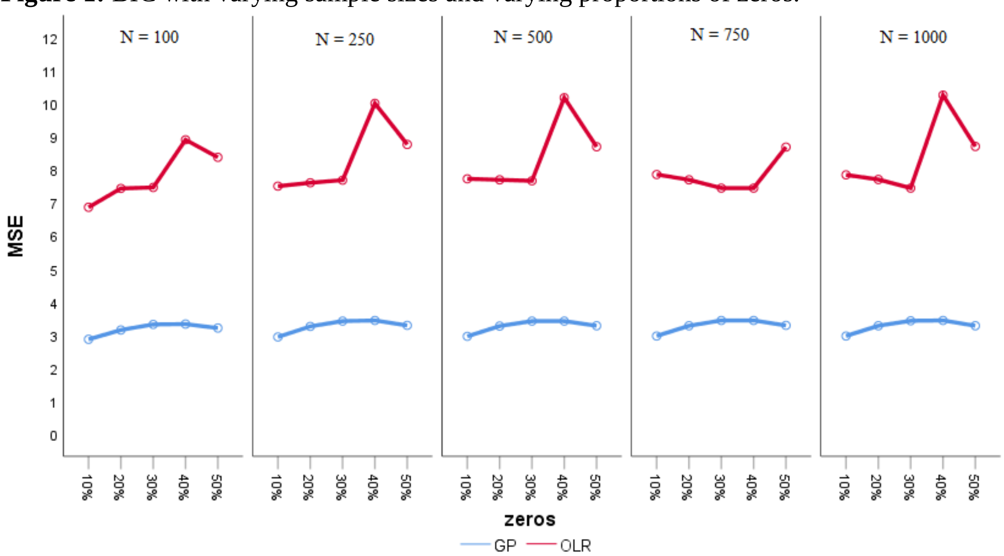

Figure 2: MSE with varying sample sizes and varying proportions of zeros.

The results from the 3 real time studies are represented in Table 2. The mean (variance) of the studies fir, DMF and ADD were 2.14(2.41), 3.46(25.93) and 1.75(19.50)) respectively. The proportion of observed zeros of the outcome was 14%, 45.5% and 60% in the studies fir, DMF and ADD respectively. The outcome of studies were over-dispersed and inflated by zeros according to Poisson model.

Table 2: Real time data.

Datasets		**Generalized Poisson model**				**Ordinal logistic regression model**			
	N	**b**	**SE**	**AIC**	**BIC**	**b**	**SE**	**AIC**	**BIC**
Fir	50	- 0.021	0.073	184	190	- 0.130	0.183	188	201
DMF	440	- 0.052	0.123	1937	2009	- 0.083	0.174	1997	2031
ADD	3720	- 0.092	0.051	11604	11623	- 0.117	0.064	11635	11909

Fir data: The regression coefficient and standard error (SE) for the variable row were lower in GP model than OLR model. Similarly, AIC and BIC values were also lower in GP model.

DMF data: The regression coefficient and SE for gender were lower in GP model than OLR model. Similarly, AIC and BIC values were also lower in GP model.

ADD data: Likewise other data, in ADD data also, the regression coefficient and SE for gender were lower in GP than OLR. Similarly, AIC and BIC estimates were also lower in GP model than OLR model.

Summary for real time data: In all three real time datasets, the standard error was consistently lower in GP model than in OLR model. The ordinal model gives overestimation of the regression estimates than the GP model. The model comparison parameters such as AIC and BIC values were consistently lower in GP model compared to OLR model. The lower AIC and BIC values, better the model.

Discussion

It is common that researchers use ordinal logistic regression for outcomes that are count variables. In nature, the count outcome is ordinal in structure.[9] In such situations, it is ideal to fit models that are suitable for ordinal data with or without zeros. While there are many models available for count data analysis, it is important to find which model fits the data well in terms of AIC, BIC and other goodness of fit statistics.

The ordinal regression models have been using in statistics quit a long time. They are used to flexibly form in count outcome that is ordinal and perform model selection.[23-24] We have suggested a new statistical GP model which is also used for over-dispersed count responses inflated by zeros.

(Manuguerra and Helle, 2016) [25] applied the ordinal regression to continuous outcome when the outcome from visual analog scales (VAS) used in pain assessment. A commonly applied method for VAS responses is to group them and analyze them as assuming as a discrete ordinal responses. This study also measured breast cancer patients health related quality of life after chemotherapy and it was measured using Linear Analog Self-Assessment (LASA) scales at each chemotherapy treatment. Pain assessment outcome was analyzed by two ways and results compared. In first it was assumed as ordinal response and analyzed using ordinal model and in second, assumed as continuous response and analyzed using parametric g function. Other hand, the quality of life outcome compared between parametric g function and non-parametric g function. GP model could be a suitable alternative to these types of responses.

(Kelley and Anderson, 2008)[18] proposed a mixture model for ordinal response when the response is zero-inflated due to never responses. The real time alcohol consumption study compared using proportional odds (PO), partial proportional odds (PPO) and zero-inflated proportional odds (ZIPO) models. The AIC value was slightly lower in the ZIPO model. Our earlier work on simulation and real time studies suggested that generalized Poison was a better model for zero-inflated outcome compared with mixture Poisson, mixture negative binomial, Zero-inflated Poisson in terms fit statistics (Yadav et. al, 2021). However, this study compared the performance of GP with OLR and we have shown that the GP model is better as compared to OLR model.

Our goal is to present a suitable method which allow researcher to use ordinal response outcome more accurately using GP model. GP could be a better alternative to ordinal model. There are not many studies that have been done for count data that use ordinal model. Our work has shown that GP model has been consistently better than OLR model, both in real and simulated data.

The limitations of the study is that, though the GP regression is a better model for ordinal data, the GP regression method is not incorporated in the standard software such as SPSS etc. Therefore, the researchers should be well versed in using the R or Python software. Though we have simulated data with various proportions and sample sizes, we could not find the reason why there is a different finding for 1000 observations with proportion of outcome 30% and 40%. Moreover, not many studies are done in this area to compare our findings.

Conclusion

The bias and MSE are consistently lower in the GP model. The BIC estimates were also consistently lower in the GP except when the sample size was 1000 and the proportions of zeros were 30% and 40%. The AIC estimates were almost similar or slightly lower in the OLR model. This study has shown that the proposed GP model would be a better choice to the OLR model. The modeling part and interpretation are easier as compared to OLR model. This study suggests that irrespective of the proportion of outcome and sample sizes, Generalized Poisson model is better as compared to Ordinal logistic regression model, though ordinal logistic regression is easy to use.

Disclosure

The authors declared that no conflicts of interest regarding authorship and publication of this article. The authors received no financial support.

References

1. Valle D, Ben Toh K, Laporta GZ, Zhao Q. Ordinal regression models for zero-inflated and/or over-dispersed count data. Sci Rep. 2019 Dec;9(1):3046.

2. Joe H, Zhu R. Generalized Poisson Distribution: the Property of Mixture of Poisson and Comparison with Negative Binomial Distribution. Biometrical Journal. 2005;47(2):219–29.

3. Lachenbruch PA. Analysis of data with excess zeros. Stat Methods Med Res. 2002 Aug;11(4):297–302.

4. Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. J Health Econ. 1998 Jun;17(3):247–81

5. Wang W, Famoye F. Modeling household fertility decisions with generalized Poisson regression. J Popul Econ. 1997 Aug;10(3):273–83.

6. Yusuf OB, Ugalahi LO. On the Performance of the Poisson, Negative Binomial and Generalized Poisson Regression Models in the Prediction of Antenatal Care Visits in Nigeria. American Journal of Mathematics and Statistics. 2015;5(3):128–36.

7. Armstrong BG, Sloan M. ORDINAL REGRESSION MODELS FOR EPIDEMIOLOGIC DATA. Am J Epidemiol. 1989 Jan 1; 129(1):191–204.

8. Das S, Rahman RM. Application of ordinal logistic regression analysis in determining risk factors of child malnutrition in Bangladesh. Nutrition Journal. 2011 Nov 14;10(1):124.

9. Liu Q, Shepherd BE, Li C, Harrell FE. Modeling continuous response variables using ordinal regression: Modeling continuous response variables using ordinal regression. Statistics in Medicine. 2017 Nov 30;36(27):4316–35.

10. Forrest M, Andersen B. Ordinal scale and statistics in medical research. Br Med J (Clin Res Ed). 1986 Feb 22;292(6519):537–8.

11. LaValley MP, Felson DT. Statistical presentation and analysis of ordered categorical outcome data in rheumatology journals. Arthritis Rheum. 2002 Jun 15;47(3):255–9.

12. Avram MJ, Shanks CA, Dykes MH, Ronai AK, Stiers WM. Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg. 1985 Jun;64(6):607–11.

13. Scott SC, Goldberg MS, Mayo NE. Statistical assessment of ordinal outcomes in comparative studies. Journal of Clinical Epidemiology. 1997 Jan;50(1):45–55.

14. Ananth C. Regression models for ordinal responses: a review of methods and applications. International Journal of Epidemiology. 1997 Dec 1;26(6):1323–33.

15. Yadav B, Jeyaseelan L, Jeyaseelan V, Durairaj J, George S, Selvaraj KG, et al. Can Generalized Poisson model replace any other count data models? An evaluation. Clinical Epidemiology and Global Health. 2021;11:100774.

16. Cameron AC. Regression Analysis of Count Data, 2013.

17. Harris T, Yang Z, Hardin JW. Modeling Underdispersed Count Data with Generalized Poisson Regression: The Stata Journal, 2012 (https://journals.sagepub.com/doi/10.1177/1536867X1201200412)

18. Kelley ME, Anderson SJ. Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model. Stat Med. 2008 Aug 15;27(18):3674–88.

19. Kleinbaum D.G, Klein M. Logistic Regression. New York, New York: Springer New York; 2010. (http://link.springer.com/10.1007/978-1-4419-1742-3).

20. Hilbe, J.M.Negative Binomial Regression. Cambridge University Press, Cambridge, second edition, 2011, page: 70-73.

21. Davison AC, Hinkley DV. Bootstrap Methods and their Application, 2013. (https://www.cambridge.org/core/books/bootstrap-methods-and-their application/ED2FD043579F27952363566DC09CBD6A).

22. Hofstetter H, Dusseldorp E, Zeileis A, Schuller AA. Modeling Caries Experience: Advantages of the Use of the Hurdle Model. Caries Res. 2016;50(6):517–26.

23. Agresti A. Introduction to Categorical Data Analysis. 2019.

24. McCullagh P. Regression Models for Ordinal Data. Journal of the Royal Statistical Society: Series B (Methodological). 1980;42(2):109–27.

25. Manuguerra M, Heller GZ. Ordinal regression models for continuous scales. Int J Biostat. 2010;6(1): (doi: 10.2202/1557-4679.1230. PMID: 21969972).