International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 23(2), November, 2023, pp 31-37
DOI: https://doi.org/10.3329/ijss.v23i2.70105

Outliers as a Source of Overdispersion in Poisson Regression Modelling: Evidence from Simulation and Real Data

Sohel Rana[1*], Abu Sayed Md. Al Mamun[2], F. M. Arifur Rahman[1], and Hanaa Elgohari[3]

[1]Department of Mathematical & Physical Sciences, East West University, Dhaka-1212, Bangladesh
[2]Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh
[3]Department of Statistics, Faculty of commerce, Mansoura University, Egypt

[*]Correspondence should be addressed to Sohel Rana
(Email: srana@ewubd.edu)

[Received August 24, 2023; Accepted October 12, 2023]

Abstract

The Poisson regression model is a well-known technique for modelling count data. However, it is necessary to satisfy the overdispersion assumption in order to fit the Poisson regression model. Due to the overdispersion problem in the Poisson regression model, standard errors might be underestimated, which may lead to a highly misleading inference. There are several tests in the literature to check the presence of overdispersion in the Poisson model. In this study, we apply a regression-based *t* test to identify the overdispersion. The simulation study and real data example clearly show that the overdispersion in the Poisson model is caused by the existence of outliers.

Keywords and Phrases: Generalized regression, Outlier, Overdispersion, Poisson regression.

AMS Classification: 62J05, 62G32.

1. Introduction

The Poisson regression model has received great attention in which the dependent variable defines the number of occurrences of some rare event. In many fields, such as biomedical science (Du *et al.* 2012), social science (Moksony and Hegedűs 2014), Metallurgical and Materials Engineering (Ajibade *et al.* 2019) and environmental science (Tobías *et al.* 2001), which involve count variables, Poisson regression is widely used. The Poisson regression model is one of the members of the family of generalized linear models (Hoffman 2004; Agresti 2012). The model has an *equidispersion property in which* the mean is equal to the variance. In many cases, however, the assumption of *equidispersion* does not hold, and the variance exceeds the mean. In the statistical literature, this is known as overdispersion (Agresti 2012).

Overdispersion is a common problem with count data and is the source of several other problems in analyzing count data (Hardin and Hilbe 2014, Hilbe, 2014). Moksony and Hegedűs (2014) stated that regression coefficients remain unbiased in the presence of overdispersion. However, the standard errors are underestimated, and hence the confidence intervals become unduly narrow, and

significance tests provide overly optimistic results. Therefore, it is important to identify the sources of overdispersion and to know the method to address the overdispersion problem. Overdispersion may arise from different sources, such as extra population-heterogeneity, omission of key predictors, zero inflation, and so on (King 1989; Osgood 2000). The objective of this study is to investigate other sources of overdispersion in the analysis of count data.

It is now evident that outliers may arise in almost all types of data (Sukmak and Thongkam 2013; Sleabi *et al.* 2015; Rana *et al.* 2015, 2018). Thus, it is not uncommon that outliers may arise in count data. Overdispersion may arise due to outliers in count data other than population-heterogeneity, omission of key predictors, and zero inflation. Both simulated data and real-life data are used in this study to support this argument.

This article is organized as follows- First, in Section 1, we reviewed the literature related to overdispersion in the Poisson regression model. The mathematical model of Poisson regression, the outlier detection in the Poisson regression model by residual analysis, and the testing procedure for overdispersion is described in Section 2. The next two sections discuss the results of the simulation study and real-world data analysis, followed by a conclusion in the final section.

2. Materials and Methods

The Poisson Regression Model

Suppose the random variable *Y* follows Poisson distribution with mean μ. Assume that the mean and the variance of Y are equal and observations are independent. Then the Poisson regression model is expressed as

$$ln(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$
$$\mu = e^{\beta_o} e^{\beta_1 X_1} e^{\beta_2 X_2} \ldots e^{\beta_p X_p} \quad (1)$$

In the simple linear regression model, it is in the form of $\mu_i = x'_i\beta$. This model takes any real values; however, count data can only take on nonnegative values. The solution to overcome this problem is to use log-linear model.

Next, we take logs in the form of $\eta_i = ln\,(\mu_i)$ and assume that the transformed mean follows a linear model $\eta_i = x'_i\beta$. Therefore, we consider a generalized linear model with log-link. We can combine these two steps and then express in log-linear model as $ln\,(\mu_i) = x'_i\beta$ and obtain a multiplicative model for the mean itself $\mu_i = exp\,(x'_i\beta)$.

Now the following normal equation, obtained by partially differentiating the log-likelihood function with respect to the model parameters, can be solved to estimate the parameters.

$$\sum_{i=1}^{n} [\, y_i - e^{x'_i \hat{\beta}} \,]\, x_i = 0 \quad (2)$$

Outlier diagnostic in the Poisson regression

Examining residuals is a primary approach for identifying the overall differences between the data and model, which helps to detect the outliers in the data. Observations that are not accommodated by the model are called outliers. Different types of residuals, such as Pearson residuals, Standardized Pearson residuals (or studentized Pearson residuals), Deviance residuals, Standardized deviance residuals have been used in the literature to detect the outliers (Mccullagh and Nelder 1989).Consider the raw residuals of Poisson regression, $r_i = y_i - \hat{\mu}_i$, then the Pearson residuals can be written as

$$p_i = \frac{r_i}{\sqrt{\widehat{\emptyset}\hat{\mu}_i}} \tag{3}$$

where $\widehat{\emptyset}$ is a dispersion parameter.

The Standardized Pearson residuals divide the Pearson residuals by the leverage factor $\sqrt{(1-h_i)}$. Where h_i is the i^{th} diagonal element of the leverage matrix, H= $V^{1/2}X(X'VX)^{-1}X'V^{1/2}$ and V is the diagonal matrix whose main diagonal contains the values μ_i . Thus, the formula of Standardized Pearson residuals becomes.

$$sp_i = \frac{p_i}{\sqrt{(1-h_i)}}, \quad i = 1,2,\dots,n \tag{4}$$

The Deviance residuals (McCullagh and Nelder, 1989) can be expressed as

$$d_i = sign(r_i) \sqrt{2\left\{y_i ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (r_i)\right\}} \tag{5}$$

Like the Standardized Pearson residuals, the standardized deviance residuals divide the deviance residuals by the leverage factor $\sqrt{(1-h_i)}$ as given below.

$$sd_i = \frac{d_i}{\sqrt{(1-h_i)}}, \quad i = 1,2,\dots,n \tag{6}$$

A large residual value ($|sd_i| > 3$) indicates the existence of outliers.

Test for Over dispersion in Poisson Regression Model

To test the overdispersion, a score test can be used (Dean and Lawless1989, Hilbe and Hardin 2014). However, Cameron and Trivedi (1990) proposed a regression based *t* test and found that the score test for the Poisson distribution is equivalent to the optimal regression-based *t* test.

Cameron and Trivedi (1990, 2005) considered the null hypothesis,

$$H_o: Var\ (Y_i) = \mu_i$$

The exact alternative hypothesis can be written as

$$H_1: Var\ (Y_i) = \mu_i + \alpha . g(\mu_i)$$

where, g(.) be a specified function. The over-dispersion or under-dispersion tests depends on whether α is different from zero. Specializing to H_0, the t-test statistic for $\alpha = 0$ is given by-

$$T_w = \left[\hat{g}'\widehat{W}\hat{\Sigma}\widehat{W}\hat{g}\right]^{-\frac{1}{2}}\hat{g}'\widehat{W}y^* \tag{7}$$

where $\hat{\Sigma}$ is a consistent estimator of Σ, $\hat{g}$ and y^* have *i*-th entries $\hat{g}(\hat{\mu}_i)$ and $(y_i - \hat{\mu}_i)^2 - y_i$ respectively, $\widehat{W}$ is a diagonal matrix with *i*-th diagonal entry $\widehat{w}_i$ such that $\widehat{W}$ is a consistent estimator of W and $w_i = w(\mu_i)$ is the weight of the weighted least-square estimate.

Under H_0, the asymptotic distribution of T_w is standard normal which can be used for either one-sided or two-sided tests of overdispersion ($\alpha > 0$) or/and underdispersion ($\alpha < 0$).

Simulation Study

A simulation study is carried out to investigate the source of over-dispersion problem in fitting the Poisson regression model. We consider the single and two predicators Poisson regression model to conduct the Monte Carlo simulation. For the single predicator Poisson regression model, the

predictor variable X is generated from the uniform distribution, i.e., X ~ Uniform (0, 1.5). The expected count of response variable Y is then generated by

$$\mathrm{E}(Y) = \mu = e^{1+X} \tag{8}$$

Thus, the random count of the response variable Y is then generated from the Poisson distribution with parameter μ. Like the single predicator Poisson regression model, the two-predictor model is generated in the same way, except that the expected count of response variable Y is generated by

$$\mathrm{E}(Y) = \mu = e^{1+X_1+X_2} \tag{9}$$

where X_1 ~ Uniform (0, 1.5), and X_2 ~ Uniform (0, 1.5).

In the next step, we insert different percentages of outliers, such as 5%, 10%, and 20%, in the Y values. Outliers are generated randomly by replacing some data points with the maximum value of Y. In the final stage, over-dispersion is identified by using the Cameron and Trivedi (2005) t-test (or z-test) through the AER package in R (Cameron and Trivedi 1990, 2005) for different sample sizes n = 30, 60, 100, and 200. The experiment is repeated 5,000 times in order to calculate the over-dispersion identification rate in the single and two predictors Poisson regression model.

Table 1: Identification of over-dispersion rate with single predictor Poisson regression model

	Different percentages of outlier			
Sample Size	**No outlier**	**5%**	**10%**	**20%**
30	0.012	0.135	0.380	0.820
60	0.019	0.320	0.853	0.996
100	0.027	0.737	0.992	1.000
200	0.026	0.992	1.000	1.000

It is observed from Table 1 that when there is no unusual observation in the data set, the identification of over-dispersion is only 1.2% for sample size n = 30 over 5,000 simulations. The percentage of over-dispersion identification increases slightly with the increase in sample sizes. However, when 5% contamination arises in the simulated data, identification of over-dispersion increases by more than 12% for sample size n = 30, and the rate of identification increases radically with the increase in sample size. An almost 100% over-dispersion rate is observed when the sample size *is* 200. Again, in the case of 10% contamination, the identification of over-dispersion increases from 13.50% to 38.08% for sample size n = 30, and a more than 80% identification rate is observed when the percentage (20%) of contamination is increased for the same sample size. About a 100% identification rate is observed for sample sizes greater than 100 and greater than 60 in cases of 10% and 20% contaminated data, respectively.

Table 2: Identification of over-dispersion rate with two predictors Poisson regression model

Sample Size	**Different percentages of outlier**			
	No outlier	**5%**	**10%**	**20%**
30	0.008	0.161	0.554	0.969
60	0.016	0.418	0.961	1.000
100	0.018	0.870	0.999	1.000
200	0.026	0.999	1.000	1.000

The percentage of overdispersion identification in case two predictors Poisson regression model with different sample sizes and different percentages of outliers is presented in Table 2. For sample

size n =30, the overdispersion identification rate is only 0.8% when no outlier is present in the dataset. The rate of identification increases to 16.1% when only a 5% outlier is inserted in the data set. The overdispersion identification rate increases to 55.4% in the case of 10% outlier, and a maximum (more than 95%) rate of identification is observed when 20% outlier is inserted for the same sample size (n= 30). When the sample size increases to 60, 1.6% over-dispersion identification rate is observed with no outlier situation, and the rate of identification increases to 41.8% when only 5% outlier is present in the data. For the same sample size, a more than 95% overdispersion detection rate is observed when 10 % or more than 10% of the outlier is inserted in the simulated data set. Again, the detection rate of the overdispersion increases slightly (1.6% to 1.8%) when the sample size increases from 60 to 100 with no inserted outliers. On the other hand, more than 85% over-dispersion identification rates are observed in cases of 5% or more than 5% outlier for the same sample sizes. For sample size n = 200, the identification of overdispersion rate is only 2.6% over 5,000 simulations in the no-outlier case, but almost 100% over-dispersion detection rate is observed in the case of any percentage of outliers in the data.

Real-life Data Example

In this section, we use low birth weight (LBW) data taken from Hosmer and Lemeshow (2000). The data consisted of 189 observations with 10 variables. The response variable is the number of physician visits in the 1st trimester which is count in nature. Therefore, the Poisson regression model is appropriate for describing this data. We consider the variables birth weight (1= low birth weight baby; 0=normal weight), hypertension (1=history of hypertension; 0 =no hypertension), weight (lbs) at last menstrual period: 80-250 lbs, age of mother: 14-45 as predictors in this model.

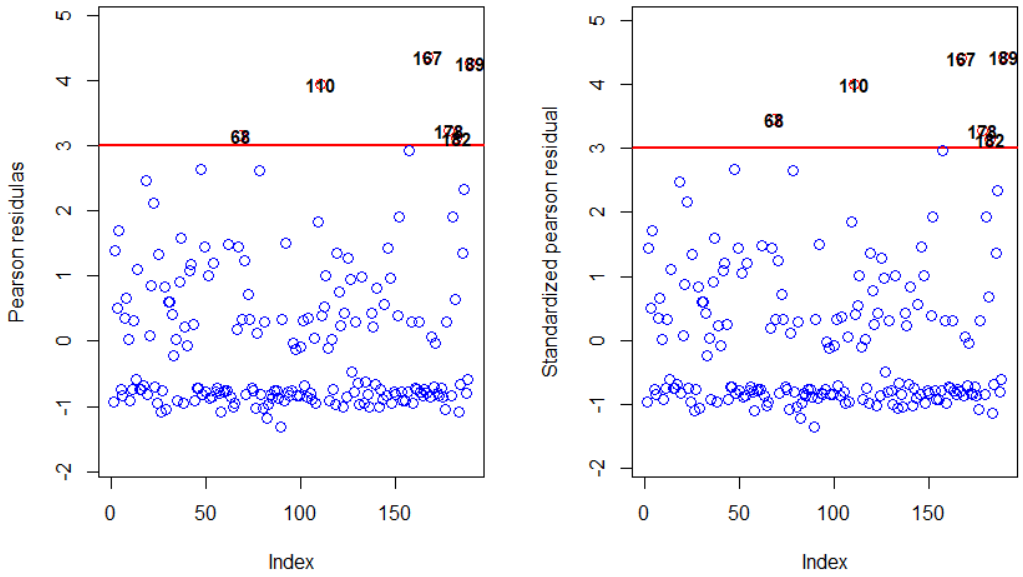

Figure 1 Detection of outlier in low birth weight (lbw) data

Pearson residuals and standardized Pearson residuals are used to identify outliers in low birth weight (lbw) data. From Figure 1, it is clearly observed that six observations exceed the cut-off

point of Pearson residuals and standardized Pearson residuals. Therefore, outliers exist in low birth weight (lbw) data.

Table 3: Result of t-test for overdispersion in low birth weight (lbw) data

Case	t-test	*p*-value	Decision
Original data (with outliers)	2.4545	0.007	Overdispersion
Modified data (after removing the outliers)	0.5598	0.288	*Equidispersion*

Table 3 represents the dispersion test for low birth weight (lbw) data in the Poisson regression model. First, we apply the Cameron and Trivedi (2005) t-test to the original low birth weight (lbw) data. Based on the p-value ($p<0.01$), we decide that overdispersion exists in the Poisson regression model when outliers are present in the data. However, when the t-test is applied to modified low birth weight (lbw) data, i.e., the data after removing the outliers from the original data, then the Poisson regression model follows the equidispersion assumption. Therefore, we conclude from the real dataset that outliers are an issue for the overdispersion problem in the Poisson regression model.

3. Conclusion

In this paper, we investigated whether outliers are a cause of the overdispersion problem in the Poisson regression model. The results of the simulation study show that when there is no outlier in the data set, the identification of the overdispersion rate is very small for different sample sizes. However, when contamination of outliers arises in the data set, the identification rate of overdispersion increases dramatically, and the rate of identification also increases with the increase in sample size for both the single predictor and two predictors Poisson regression models. It becomes clearer from the analysis of real-life data that outliers cause overdispersion. Therefore, it is concluded that outliers are another source of overdispersion in the Poisson regression model. Finally, we recommend identifying the source of overdispersion in the Poisson regression model. If the overdispersion problem arises as a result of outliers, we need to treat the outliers first instead of applying alternative methods for solving this problem. As for the treatment of outliers, we may remove the outliers from the data or apply other suitable methods.

References

[1] Agresti A. (2012). An Introduction to categorical data analysis. New York: John Wiley & Sons.

[2] Ajibade O. A, Agunsoye J. O. and Oke S. A. (2019). Poisson distribution: How tensile properties of particulate polymer composites are enhanced in a Poisson - motivated Taguchi method. Engineering and Applied Science Research. 46 (2):130- 141.

[3] Cameron A. C and Trivedi P. K. (1990). Regression-based tests for overdispersion in the Poisson model. Journal of Econometrics. 46: 347–36.

[4] Cameron A.C and Trivedi P. K. (2005). Microeconometrics: Methods and applications. Cambridge: Cambridge University Press.

[5] Du J, Park Y. T, Theera-Ampornpunt N, Mccullough J. S and Speedie S. M. (2012). The use of count data models in biomedical informatics evaluation research. Journal of the American Medical Informatics Association. 19(1): 39-44.

[6] Hardin, J, Hilbe and J. M. (2014). Regression Models for Count Data Based on the Negative Binomial(p) Distribution. The Stata Journal, 14(2), 280–291.

[7] Hilbe, J. (2014). Modeling Count Data. Cambridge University Press, New York.

[8] Hoffman J. P. (2004). Generalized linear models. Boston: Pearson Education Inc.

[9] Hosmer D and Lemeshow S. (2000). Applied logistic regression. New York: Wiley.

[10] King G. (1989). Variance specification in event count models, from restrictive assumptions to a generalized estimator. American Journal of Political Science. 33: 762–784.

[11] Mccullagh P and Nelder J. A. (1989). Generalized linear models. 2nd ed. London: Chapman & Hall.

[12] Moksony F and Hegedűs R. (2014). The use of Poisson regression in the sociological study of Suicide. Corvinus Journal of Sociology and Social Policy. 5(2): 97-114.

[13] Osgood D. W. (2000). Poisson-based regression analysis of aggregate crime rates. Journal of Quantitative Criminology. 16(1): 21-43.

[14] Rana S, Sleabi W. D and Midi H. (2018). Fixed parameters support vector regression for outlier detection. Economic Computation and Economic Cybernetics Studies and Research. 52(2): 267-292.

[15] Rana S, John A. H, Midi H and Imon. A.H.M.R. (2015). Robust regression imputation for missing data in the presence of outliers. Far East Journal of Mathematical Sciences. 97(2): 183-195.

[16] Sleabi W. D, Rana S and Midi, H. (2015). Non-sparse-insensitive support vector regression for outlier detection. Journal of Applied Statistics. 42(8): 1723–1739.

[17] Sukmak V, Thongkam J. (2013). Improving quality of breast cancer data through pre-processing. Engineering and Applied Science Research. 40(4): 493-504.

[18] Tobías A, Díaz J, Saez M, Alberdi J. C. (2001). Use of Poisson regression and Box–Jenkins models to evaluate the short-term effects of environmental noise levels on daily emergency admissions in Madrid, Spain. European journal of epidemiology. 17(8):765-71.