İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 11, 2018, 2, 73-80

Geliş/Received:15.05.2018, Kabul/Accepted: 23.10.2018

www.istatistikciler.org

Araştırma Makalesi / Research Article

A new zero-inflated regression model with application

Emrah Altun
Bartın Üniversitesi
İstatistik Bölümü, Bartın 74100, Türkiye
emrahaltun@bartin.edu.tr
0000-0001-5065-2523

Abstract

In this paper, a new zero-inflated regression model, called Zero-Inflated Poisson-Lindley regression model, is proposed for count data modeling. Poisson-Lindley distribution arises from the Poisson distribution when its parameter follows a Lindley distribution. Contrary to Poisson distribution, Poisson-Lindley distribution allows for over-dispersion. Therefore, a new model is good candidate to model the over-dispersed and zero-inflated data sets. Application of proposed model to real data set is given and compared with Poisson and Zero-Inflated Poisson regression models. Empirical findings reveal that the Zero-Inflated Poisson-Lindley regression model provides better fits than Zero-Inflated Poisson regression model for zero-inflated and over-dispersed data set.

Keywords: Poisson-Lindley distribution; Over-dispersion; Inflated models; Power series distribution.

Öz

Yeni sıfır yığılmalı regresyon modeli ve uygulaması

Bu çalışmada, sayım verilerinin modellenmesi için sıfır yığılmalı Poisson-Lindley regresyon modeli olarak adlandırılan yeni bir sıfır yığılmalı regresyon modeli önerilmiştir. Poisson-Lindley dağılımı, Poisson dağılımının parameteresinin Lindley dağılımına sahip olduğu durumda ortaya çıkmaktadır. Poisson dağılımının aksine, Poisson-Lindley dağılımı aşırı yayılıma izin verir. Bu nedenle, yeni model aşırı yayılımlı ve sıfır yığılmalı veri kümelerini modellemek için iyi bir seçenektir. Önerilen modelin gerçek veri seti üzerine uygulaması verilmiş, Poisson ve Zero-Inflated Poisson regresyon modelleriyle karşılaştırılmıştır. Elde edilen bulgular, sıfır yığılmalı Poisson-Lindley regresyon modelinin, sıfır yığılmalı ve aşırı yayılım gösteren veri seti için, sıfır yığılmalı Poisson regresyon modelinden daha iyi uyum sağladığını göstermektedir.

Anahtar sözcükler: *Poisson Lindley dağılımı, Aşırı yayılım, Yığılmalı modeller, Güç serisi dağılımı.*

1. Introduction

Researches on the modeling of the count data in many fields such as insurance, public health, epidemiology, psychology etc. have been shown a great interest in the past two decades. Poisson regression model is widely used model for modeling the count data. It is widely documented that the count data displays over-dispersion. Using the Poisson distribution as an assumptional distribution on dependent variable causes the underestimated standard errors and damages the significance of regression parameters. In this case, quasi-Poisson regression model can be used to model the over-dispersion. Besides the quasi-Poisson regression model, researchers have been proposed alternative models for solving the over-dispersion problem such as Poisson-inverse Gaussian, Poisson-Lognormal and Negative Binomial regression models (see for details, Denuit et al. (2007)).

Count data has another phenomena, called as zero-inflation. Zero-inflation is occurred when the number of zero outcomes are higher than the represented by Poisson distribution. It is widely seen in data sets related to insurance and health. Zero-inflated regression models, such as Zero-inflated Poisson (ZIP) and Zero-inflated Negative-Binomial (ZINB), have been proposed to model the zero-inflated data sets. ZINB regression model can be more appropriate choice than ZINB regression model when the count data still exhibits over-dispersion. There are several researches on the Poisson and ZIP regression models in recent years. For example Avcı et al. (2015), Ismail and Zamani (2013), Lord et al. (2005) and Ayati and Abbasi (2014). Avcı et al. (2015) compared the Poisson, Negative Binomial, Conway-Maxwell-Poisson regression models in modeling the over-dispersed alga data. Ismail and Zamani (2013) compared the generalized Poisson and Negative Binomial regression models in modeling the Malaysian claim data and German healthcare data. Lord et al. (2005) compared the Poisson and Negative Binomial regression models in modeling the crash data. Ayati and Abbasi (2014) used the zero-inflated regression model to investigate the effective factors on frequency and severity of accidents on urban highways. Recently, Xavier et al. (2017) introduced an extension of zero-inflated Poisson-Lindley distribution, called as zero-modified Poisson-Lindley and demonstrated the usefulness of the proposed distribution in case of inflation of zeros and deflation of zeros.

The goal of this paper is to introduce an alternative zero-inflated regression model for count data modeling. For this goal, Poisson-Lindley distribution is applied to count data regression models for zero-inflated case. The advantage of Poisson-Lindley distribution in comparison with Poisson distribution is to allow the modeling of the over-dispersed dependent variable. The usefulness of proposed regression model is illustrated by means of real data application.

The rest of the paper is organized as follows: In section 2, count data models are presented. In Section 3, real data application is given to demonstrate the usefulness of proposed model against to Poisson and Zero-inflated Poisson regression models. Some concluding remarks are given in Section 4.

2. Zero-Inflated Poisson-Lindley regression model

In Poisson regression, the dependent variable y_i is modeled by,

$$P(y_i) = \frac{e^{\lambda_i} \lambda_i^{y_i}}{y_i!} \tag{1}$$

where the conditional variance is equal to conditional mean, $E\left(y_i \middle| x_i\right) = V\left(y_i \middle| x_i\right) = \lambda_i = \exp(\mathrm{x}_i^T \beta)$. The log-likelihood function of Poisson regression model is given by,

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i x_i^T \beta - \exp(x_i^T \beta) - \ln(y_i!) \right] \tag{2}$$

The regression coefficients, β , can be estimated by maximum likelihood method. The derivative of log-likelihood function with respect to vector of coefficients, β, is set equal to zero,

$$\frac{d\ell(\beta)}{d\beta} = \sum_{i=1}^{y} (y_i - e^{x_i^T \beta}) x_i = 0 \tag{3}$$

Estimation of regression parameters in (3) can be obtained by means of the Newton-Raphson iteration procedure. Zero-inflated distributions come from the zero-inflated power series distribution. Zero-inflated power series distribution is given by,

$$P(y_i \mid w_0,\theta)=\begin{cases} w_0+(1-w_0)\,f(y_i \mid \theta), & y_i=0 \\ (1-w_0)\,f(y_i \mid \theta), & y_i>0 \end{cases} \tag{4}$$

where $f(y_i \mid \theta)$ is the probability mass function (pmf) of power series distribution and w_0 represents the degenerated part at zero. The most used discrete one parameter exponential families belong to the power series family is Poisson distribution. Using the Equation (4), the pmf of ZIP regression model is given by,

$$P(y_i)=\begin{cases} w_i+(1-w_i)\,\mathrm{e}^{-\lambda_i}, & y_i=0 \\ (1-w_i)\dfrac{\lambda_i^{y_i}\,\mathrm{e}^{-\lambda_i}}{y_i\,!}, & y_i>0 \end{cases} \tag{5}$$

where $0\le w_i<1$. In ZIP model, the mean of Poisson distribution, λ_i is linked to a regression of independent variables, x_i by log link function and the probability of zero process, w_i to a regression of independent variables, z_i by logit link function. Log link function and logit link function can be given as follows:

$$\begin{aligned} \log(\lambda_i) &= x_i^T\beta \\ \log\left(\frac{w_i}{1-w_i}\right) &= z_i^T\gamma \end{aligned} \tag{6}$$

where β and γ are vectors of regression parameters. The log-likelihood function of ZIP regression model is given by,

$$\begin{aligned} \ell(\beta,\gamma) &= \sum_{y_i=0}\ln\left[\exp(z_i'\gamma)+\exp\left(-\exp(x_i'\beta)\right)\right]+\sum_{y_i>0}\left[y_i x_i'\beta-\exp(x_i'\beta)-\ln(y_i\,!)\right] \\ &-\sum_{i=1}^{n}\ln\left(\left[1+\exp(z_i'\gamma)\right]\right) \end{aligned} \tag{7}$$

The log-likelihood function (7) can be maximized using the statistical softwares such as R, MATLAB, S-PLUS etc.

Here, we introduce the Zero-inflated Poisson-Lindley (ZIPL) regression model. Poisson-Lindley (PL) distribution arises from the Poisson distribution when its parameter follows a Lindley distribution. PL distribution was introduced by Sankaran (1970) for modeling the count data. The pmf of PL distribution is given by,

$$P(y_i)=\frac{\theta^2\left(y+\theta+2\right)}{\left(\theta+1\right)^{y+3}},\ y=0,1,2,,3...;\ \theta>0 \tag{8}$$

The mean and variance of PL distribution are given by,

$$\begin{aligned} \mu &= \frac{(\theta+2)}{\theta(\theta+1)} \\ \sigma^2 &= \frac{\theta^3+4\theta^2+6\theta+2}{\theta^2\left(\theta+1\right)^2} \end{aligned} \tag{9}$$

Variance of PL distribution can be rewritten as follows:

$$\sigma^2=\mu\left(1+\frac{\theta^2+4\theta+2}{\theta(\theta+1)(\theta+2)}\right)>\mu \tag{10}$$

As seen from (10), PL distribution is over-dispersed. Figure 1 displays the pmf shapes of PL distribution for various values of θ. It is clear that when the parameter θ increases, the degree of skewness to right increases.

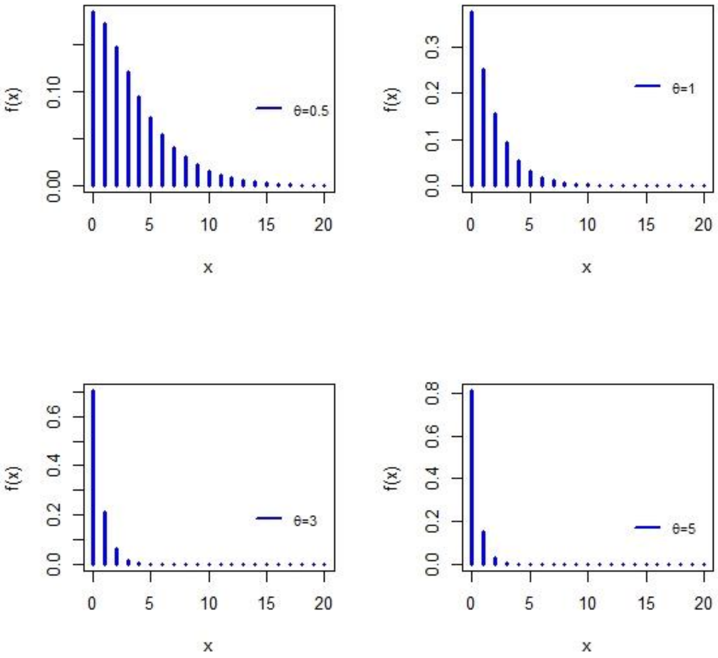

Figure 1. The pmf plots of Poisson-Lindley distribution for various parameters.

Let $\theta=\left(1-\mu+\sqrt{\mu^2+6\mu+1}\right)\big/2\mu$, then, PL distribution can be given by

$$P(y_i)=\frac{\left[\left(1-\mu+\sqrt{\mu^2+6\mu+1}\right)\big/2\mu\right]^2\left(y+\left(1-\mu+\sqrt{\mu^2+6\mu+1}\right)\big/2\mu+2\right)}{\left(\left(1+\mu+\sqrt{\mu^2+6\mu+1}\right)\big/2\mu\right)^{y+3}} \tag{11}$$

Using the Equation (4) and (11), Zero-inflated PL (ZIPL) distribution is given by

$$P(y_i)=\begin{cases} w_i+(1-w_i)\dfrac{\left[\left(1-\mu+\sqrt{\mu^2+6\mu+1}\right)\Big/2\mu\right]^2\left(\left[\left(1-\mu+\sqrt{\mu^2+6\mu+1}\right)\Big/2\mu\right]+2\right)}{\left[\left\{\left(1-\mu+\sqrt{\mu^2+6\mu+1}\right)\Big/2\mu\right\}+1\right]^3}, & y_i=0 \\ (1-w_i)\dfrac{\left[\left(1-\mu+\sqrt{\mu^2+6\mu+1}\right)\Big/2\mu\right]^2\left(y+\left(1-\mu+\sqrt{\mu^2+6\mu+1}\right)\Big/2\mu+2\right)}{\left(\left(1+\mu+\sqrt{\mu^2+6\mu+1}\right)\Big/2\mu\right)^{y+3}}, & y_i>0 \end{cases} \tag{12}$$

Note that when the parameter $w_i=0$, ZIPL distribution reduces to PL distribution. In ZIPL regression model, the mean of PL distribution, μ_i is linked to a regression of independent variables, x_i by log link function, $\log(\mu_i)=x_i^T\beta$, and the probability of zero process, w_i to a regression of independent variables, z_i by logit link function $\operatorname{logit}(w_i)=z_i^T\gamma$. The link function is chosen by considering the domain of the mean of response variable. Since the mean of PL distribution, μ_i, is defined on $\mathbb{R}^+$, the log-link function is suitable choice for linking the covariates to mean of the response variable. The log-likelihood function of ZIPL regression model is obtained as

$$\begin{aligned}\ell(\beta,\gamma)=\sum_{y_i=0}\ln\left(\begin{array}{c}\dfrac{\exp(z_i'\gamma)}{1+\exp(z_i'\gamma)}+\left(1-\dfrac{\exp(z_i'\gamma)}{1+\exp(z_i'\gamma)}\right)\\ \times\dfrac{\left[\left(1-\exp(x_i'\beta)+\sqrt{\exp(x_i'\beta)^2+6\exp(x_i'\beta)+1}\right)\Big/2\mu\right]^2\left(\left[\left(1-\exp(x_i'\beta)+\sqrt{\exp(x_i'\beta)^2+6\exp(x_i'\beta)+1}\right)\Big/2\exp(x_i'\beta)\right]+2\right)}{\left[\left\{\left(1-\exp(x_i'\beta)+\sqrt{\exp(x_i'\beta)^2+6\exp(x_i'\beta)+1}\right)\Big/2\exp(x_i'\beta)\right\}+1\right]^3}\end{array}\right)+\\ \sum_{y_i>0}\ln\left(\dfrac{\left[\left(1-\exp(x_i'\beta)+\sqrt{\exp(x_i'\beta)^2+6\exp(x_i'\beta)+1}\right)\Big/2\exp(x_i'\beta)\right]^2\left(y+\left(1-\exp(x_i'\beta)+\sqrt{\exp(x_i'\beta)^2+6\exp(x_i'\beta)+1}\right)\Big/2\exp(x_i'\beta)+2\right)}{\left(1+\exp(z_i'\gamma)\right)\left(\left(1+\exp(x_i'\beta)+\sqrt{\exp(x_i'\beta)^2+6\exp(x_i'\beta)+1}\right)\Big/2\exp(x_i'\beta)\right)^{y+3}}\right)\end{aligned} \tag{13}$$

The log-likelihood function (13) can be maximized using the statistical softwares. The **optim()** function of the R software is used to minimize the minus log-likelihood function given in (13). The function **optim()** is a function that can provide basic optimization capabilities and widely used functions in R. It contains several algorithms such as Nelder–Mead, quasi-Newton, conjugate-gradient and stochastic annealing algorithms. The parameter estimates in the model can be obtained using several initial values to guarantee the convergence to the global optimum. The estimated parameters of ZIP model is used as initial values for parameter estimation of ZIPL model.

3. Empirical Study

3.1. Data description

The data set of United States National Medical Expenditure Survey (NMES) for 1987/1988 is used to compare the proposed model with other competitive models. The data set is available in **countreg** package of R software. The response variable, (y_i), number of physician visits is modeled by the following covariates: number of hospital stay (x_{i1}), number of chronic conditions, (x_{i2}), gender (x_{i3}), number of years of education, (x_{i4}), and indicator of private insurance (x_{i5}).The descriptive statistics of used data set is given in Table 1.

Table 1. Descriptive statistics of NMES data set.

Variables	Min	Max	Mean	Median
Number of physician visits	0	89	5.77	4
Number of hospital stay	0	8	0.295	0
Number of chronic conditions	0	8	1.542	1
Gender(0:female, 1:male)	0	1	-	1
Number of years of education	0	18	10.291	11
Private insurance(0:no, 1:yes)	0	1	-	1

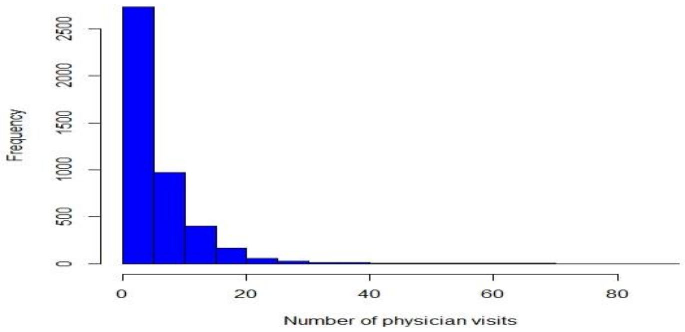

Figure 2. Histogram for the number of physician visits.

As seen from the Figure 2, the histogram of the number of physician visits is highly peaked at zero value.

3.2. Empirical Results

Here, NMES data set is modeled with Poisson, PL, ZIP and ZIPL regression models. Table 1 shows the estimated $-\ell$, Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). The smallest values of these statistics shows the best model. As seen from Table 1, ZIPL model has the smallest values of these statistics. Therefore, ZIPL regression model can be chosen as the best model for fitted data set.

Table 1. The goodness-of-fit statistics of fitted models for the NMES data set

Model Selection Criteria	P	PL	ZIP	ZIPL
$-\ell$	18151.35	12244.82	16293.57	**12143.77**
AIC	36314.70	24501.64	32611.14	**24311.54**
BIC	36353.04	24539.98	32687.83	**24388.23**

The test introduced by Cameron and Trivedi (1998) is used to test over-dispersion. The test statistic is asymptotically distributed as normal with zero mean and unit variance. The obtained test statistic is $z=12.131$ and corresponding p value is $<2.2\times10^{-16}$. Therefore, the dependent count data variable is over-dispersed. Van den Broek (1995) proposed the score test for testing whether the Poisson distribution provides sufficient representation for the frequency of zero observation. Score test statistic is given by

$$T_s(Y)=\left(\sum_{i=1}^{n}\frac{1_{\{y_i=0\}}-\tilde{p}_0}{\tilde{p}_0}\right)^2\left[\left(\frac{1-\tilde{p}_0}{\tilde{p}_0}\right)-n\bar{y}\right]^{-1}$$

where $\tilde{p}_0 = \exp(-\hat{\lambda})$ and n is the number of observations. Under the null hypothesis score test statistics is asymptotically distributed as χ^2 distribution with 1 degree of freedom. Score test statistic and corresponding p value are obtained as 33438.0888 and $< 2.2\times 10^{-16}$, respectively. According to score test result, it is concluded that Poisson distribution is insufficient to represent frequency of zero values in the dataset.

Table 2 shows the estimated parameters of fitted models and corresponding standard errors and p values. It is observed that the estimates of all parameters, except γ_0 and γ_1, are found significant at 5% level of significance as p value for all the parameter estimates are less than 5% level of significance.

Table 2. The estimated parameters of fitted models and corresponding standard errors and p values for the NMES data set.

Covariates	Poisson			PL			ZIP			ZIPL		
	Est.	S.E	p-value	Est.	S.E	p-value	Est.	S.E	p-value	Est.	S.E	p-value
β_0	0.987	0.037	<0.001	0.925	0.076	<0.001	1.446	0.037	<0.001	1.235	0.085	<0.001
β_1	0.182	0.006	<0.001	0.233	0.018	<0.001	0.175	0.006	<0.001	0.218	0.019	<0.001
β_2	0.175	0.004	<0.001	0.200	0.010	<0.001	0.129	0.004	<0.001	0.154	0.011	<0.001
β_3	-0.116	0.013	<0.001	-0.135	0.028	<0.001	-0.065	0.013	<0.001	-0.090	0.030	<0.001
β_4	0.022	0.002	<0.001	0.022	0.004	<0.001	0.015	0.002	<0.001	0.016	0.004	<0.001
β_5	0.183	0.017	<0.001	0.195	0.036	<0.001	0.061	0.017	<0.001	0.101	0.040	<0.001
Zero-inflation model coefficients												
γ_0	-	-	-	-	-	-	0.287	0.222	0.196	0.564	0.455	0.216
γ_1	-	-	-	-	-	-	-0.310	0.091	<0.001	-0.756	0.405	0.062
γ_2	-	-	-	-	-	-	-0.542	0.044	<0.001	-1.303	0.181	<0.001
γ_3	-	-	-	-	-	-	0.418	0.089	<0.001	0.633	0.205	<0.001
γ_4	-	-	-	-	-	-	-0.056	0.012	<0.001	-0.089	0.027	<0.001
γ_5	-	-	-	-	-	-	-0.751	0.102	<0.001	-1.190	0.224	<0.001

4. Conclusion

In this study, Zero-inflated Poisson-Lindley regression model is proposed and applied to real data set. The modeling ability of proposed regression model is compared with Zero-inflated Poisson regression model. Empirical findings show that proposed model provides more accurate fitting performance than Poisson regression model and its zero-inflated case. We hope that the results given in this paper will be useful for researchers and practitioners studying in field of count data modeling.

References

[1] M. Sankaran, 1970, The Discrete Poisson-Lindley Distribution, *Biometrics*, 145-149.

[2] M. Denuit, X. Maréchal, S. Pitrebois, J. F. Walhin, 2007, *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*, John Wiley & Sons.

[3] E. Avcı, S. Alturk, E. N. Soylu, E, 2015, Comparison count regression models for overdispersed alga data, *IJRRAS*, 25(1), 1-5.

[4] N. Ismail, H. Zamani, 2013, Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models, *In Casualty Actuarial Society E-Forum*, 41(20),1-18.

[5] D. Lord, S.P. Washington, J.N. Ivan, 2005, Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory, *Accident Analysis and Prevention*, 37, 35-46.

[6] E. Ayati, E.Abbasi, 2014, Modeling Accidents on Mashhad Urban Highways, *Open Journal of Safety Science and Technology*, 4, 22-35.

[7] D. Xavier, M. Santos-Neto, M. Bourguignon, V. Tomazella, 2017, Zero-Modified Poisson-Lindley distribution with applications in zero-inflated and zero-deflated count data. *arXiv preprint arXiv:1712.04088.*

[8] J. Van den Broek, 1995, A score test for zero inflation in a Poisson distribution. *Biometrics*, 738-743.