AIMS Mathematics, 8(10): 23272–23290.
DOI: 10.3934/math.20231183
Received: 26 April 2023
Revised: 09 June 2023
Accepted: 25 June 2023
Published: 21 July 2023

http://www.aimspress.com/journal/Math

Research article

A novel approach for zero-inflated count regression model: Zero-inflated Poisson generalized-Lindley linear model with applications

Emrah Altun[1,*], **Hana Alqifari**[2,*] **and Mohamed S. Eliwa**[2,3]

[1] Department of Mathematics, Bartin University, Bartin, Turkey
[2] Department of Statistics and Operation Research, College of Science, Qassim University, Buraydah 51482, Saudi Arabia
[3] Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt

* **Correspondence:** Email: emrahaltun123@gmail.com, hn.alqifari@qu.edu.sa.

Abstract: Count regression models are important statistical tools to model the discrete dependent variable with known covariates. When the dependent variable exhibits over-dispersion and inflation at zero point, the zero-inflated negative-binomial regression model is used. The presented paper offers a new model as an alternative to the zero-inflated negative-binomial regression model. To do this, Poisson generalized-Lindley distribution is re-parametrized and its parameter estimation problem is discussed via maximum likelihood estimation method. The proposed model is called as zero-inflated Poisson generalized Lindley regression model. The results regarding the efficiency of parameter estimation of the proposed model are evaluated with two simulation studies. To evaluate the success of the proposed model in the case of zero inflation, two datasets are analyzed. According to the results obtained, the proposed model gives better results than the negative-binomial regression model both in case of over-dispersion and in the case of zero inflation.

Keywords: negative-binomial regression; over-dispersion; Poisson regression; Poisson generalized-Lindley; zero-inflation distribution; statistical model; simulation
Mathematics Subject Classification: 62E15

1. Introduction

The Poisson distribution is a well-known distribution to model the count data sets. As widely known, the mean and variance of the Poisson distribution are equal to each other. This property of the Poisson distribution causes some problems in modeling the real-life data sets. In real-life data modeling, the data sets are generally over-dispersed which means that the empirical variance is greater than the empirical mean. In this case, the use of the Poisson distribution for these type

of data sets yields the misspecification of the underlying probability distribution. Negative-Binomial (NB) distribution is the first choice for modeling the over-dispersed count data sets. However, we need more flexible discrete distributions to model highly over-dispersed count data sets. In the last decade, several authors have proposed alternative discrete distributions to handle this problem, such as Shoukri et al. [28], Shmueli et al. [26], Rodríguez-Avi et al. [25], Mahmoudi and Zakerzadeh [24], Lord and Geedipally [23], Cheng et al. [11], Gómez-Déniz [17], Sáez-Castillo and Conde-Sánchez [27], Zamani et al. [30], Gencturk and Yigiter [16], Bhati et al. [9], Imoto et al. [20], Wongrin and Bodhisuwan [31], Altun [2–6], El-Morshedy et al. [14, 15], Altun et al. [1], Eliwa et al. [13].

The other phenomena in count data modeling is inflation. Inflation is seen generally at zero point and called as zero-inflation. Zero-inflation means that the underlying data set contains too many zero observations that cannot be represented by the corresponding distribution, such as Poisson and NB. This situation is commonly seen in insurance and health sciences, such as loss frequency, number of physicians visits, daily coronavirus cases etc. In this case, zero-inflated version of the Poisson and NB regression models are used. These are called as zero-inflated Poisson (ZIP) and zero-inflated negative-binomial (ZINB). Thanks to their software support, these models have been applied to real-life problems by many researchers. For instance, a comparison of over-dispersed count data sets were studied by Avci et al. [8]. Besides, Ismail and Zamani [19] conducted a study for applications of the ZIP and ZINB models on the Malaysian own damage claim data. One can also visit the work of Lord et al. [22] to see the application of these models on the crash data. Also, Ayati and Abbasi [7] investigated the suitability of these models for the accidents on urban highways.

Here, the main purpose is to develop a new sophisticated model for the zero-inflated and/or over-dispersed data sets. To do this, we use the Poisson generalized-Lindley (PGL) distribution, introduced by Wongrin and Bodhisuwan [34]. The suitable re-parametrized version of the PGL distribution is introduced and its statistical properties are studied. The maximum likelihood estimation (MLE) method is preferred to estimate the unknown model parameters. The suitability of the MLE method for estimating the parameters of the proposed model is discussed with simulation study. Two real data sets are analyzed to prove the importance of the proposed distribution against the existing models such as Poisson, NB regression models and their zero-inflated models.

The other parts of the study are organized as follows: The re-parametrized PGL distribution is studied in Section 2. In Section 3, the parameter estimation problem is addressed with MLE and the simulation study is given. Section 4 is devoted to introduce a new regression model for both zero-inflated and over-dispersed cases. Section 5 contains the empirical results of the study. Some conclusion remarks are given in Section 6.

2. Re-parametrization of Poisson generalized-Lindley distribution

Wongrin and Bodhisuwan [34] introduced PGL distribution by using the generalized-Lindley distribution of Elbatal et al. [12]. Let the random variable Y follow the PGL distribution with probability mass function (pmf)

$$P(y;\alpha,\beta,\theta) = \frac{1}{y!(\theta+1)^{y+1}}\left[\left(\frac{\theta}{\theta+1}\right)^{\alpha}\frac{\theta\Gamma(y+\alpha)}{\Gamma(\alpha)} + \left(\frac{\theta}{\theta+1}\right)^{\beta}\frac{\Gamma(y+\beta)}{\Gamma(\beta)}\right], y = 0, 1, 2, ...,$$

where $\Gamma(\cdot)$ is the gamma function and $\alpha, \beta, \theta > 0$. The mean and variance of Y are

$$E(Y) = \frac{\alpha\theta + \beta}{\theta(\theta + 1)},$$

and

$$\text{Var}(Y) = \frac{\beta[\theta(-2\alpha + \theta + 2) + 1] + \alpha\theta\left[\alpha + (\theta + 1)^2\right] + \beta^2\theta}{\theta^2(\theta + 1)^2}.$$

Now, we introduce a re-parametrized version of the PGL distribution to make it suitable distribution for count regression model.

Proposition 1. *Let $\beta = \theta(\mu\theta + \mu - \alpha)$, then, the pmf of PGL distribution is*

$$P(y; \alpha, \theta, \mu) = \frac{1}{y!(\theta + 1)^{y+1}}\left[\left(\frac{\theta}{\theta + 1}\right)^{\alpha}\frac{\theta\Gamma(y + \alpha)}{\Gamma(\alpha)} + \left(\frac{\theta}{\theta + 1}\right)^{\theta(\mu\theta+\mu-\alpha)}\frac{\Gamma(y + \theta(\mu\theta + \mu - \alpha))}{\Gamma(\theta(\mu\theta + \mu - \alpha))}\right], \tag{2.1}$$

where $y = 0, 1, 2, ...$, $\alpha > 0$, $\theta > 0$ and $\mu > 0$. The mean and variance of Y are

$$E(Y) = \mu,$$

and

$$V(Y) = \frac{\theta(\mu\theta + \mu - \alpha)[\theta(-2\alpha + \theta + 2) + 1] + \alpha\theta\left[\alpha + (\theta + 1)^2\right] + \{\theta(\mu\theta + \mu - \alpha)\}^2\theta}{\theta^2(\theta + 1)^2}.$$

Hereinafter, the random variable Y refers to the re-parametrized PGL distribution given in (2.1) and shortly denoted as PGL(α, θ, μ). Following the results of Wongrin and Bodhisuwan [34], the statistical properties of the re-parametrized PGL distribution could be obtained easily. The pmf shapes of the re-parametrized PGL distribution are displayed in Figure 1. From these figures, we conclude that this distribution could be used to model zero-inflated, bimodal and right skewed count data sets.

The dispersion index (DI) is defined as DI = Var(X)/E(X). The DI shows the flexibility of the distribution in modeling over(under)-dispersed data sets. When the DI is greater than one, it means that the data set exhibits over-dispersion. The opposite case (DI < 1) indicates the under-dispersion. The variance and DI plots of the PGL distribution are displayed in Figure 2 (for fixed $\alpha = 0.5$). We conclude the following results from Figure 2: When the parameter μ increases, dispersion index and variance increase; when the parameter θ increases, dispersion index and variance decreases. The DI of the PGL is always greater than one. So, the PGL distribution is an appropriate choice to model over-dispersed data sets. Figure 3 shows the results of dispersion index and variance of the PGL distribution for $\alpha = 1.5$. The similar interpretation can be done as in the case of $\alpha = 0.5$.

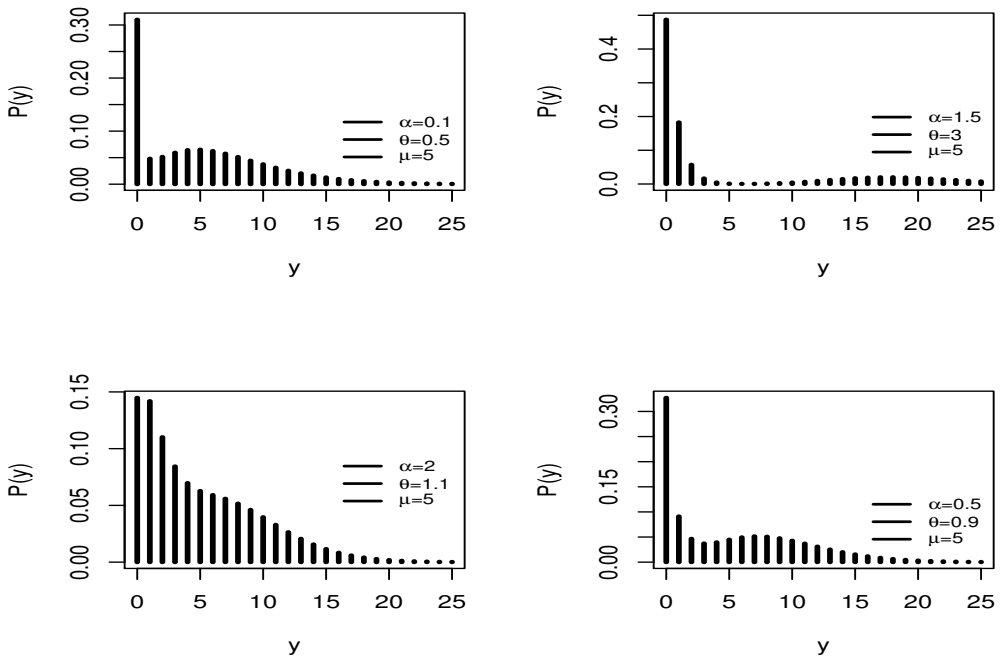

Figure 1. The pmf shapes of PGL distribution.

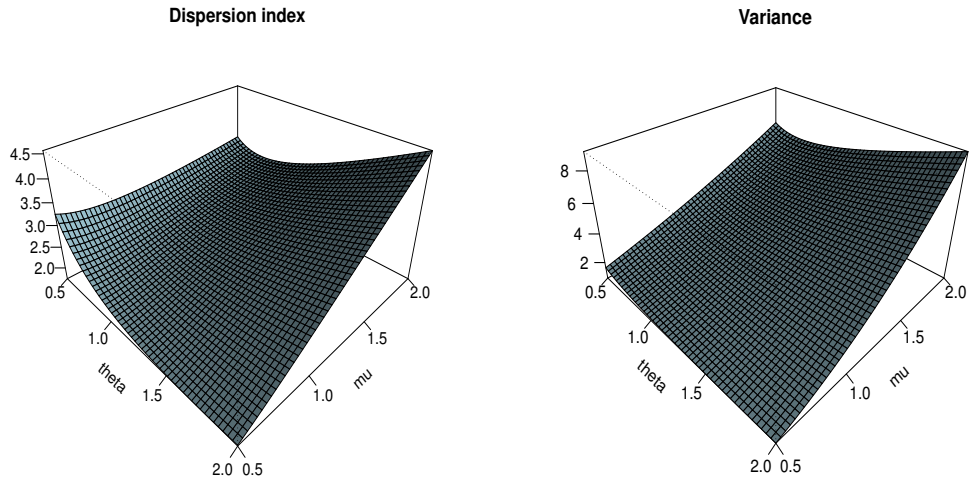

Figure 2. The dispersion index and variance of PGL distribution for $\alpha = 0.5$.

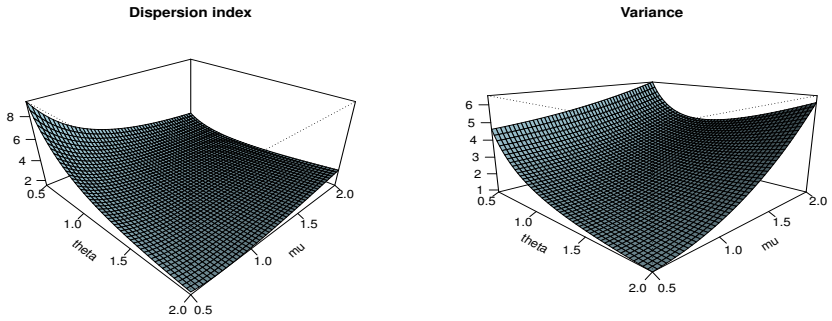

Figure 3. The dispersion index and variance of PGL distribution for for $\alpha = 1.5$.

3. Estimation

In this section, the parameters of PGL distribution are obtained by MLE method. The appropriateness of the MLE method is evaluated by simulation study.

3.1. Maximum likelihood estimation

Assume that we have a random sample, $y_1, y_2, \ldots, y_n$, from the PGL distribution. Then, the log-likelihood function of the PGL distribution is

$$\begin{aligned} \ell(\boldsymbol{\tau}) = & \sum_{i=1}^{n} \ln\left[\left(\frac{\theta}{\theta+1}\right)^{\alpha} \frac{\theta\Gamma(y_i+\alpha)}{\Gamma(\alpha)} + \left(\frac{\theta}{\theta+1}\right)^{\theta(\mu\theta+\mu-\alpha)} \frac{\Gamma(y_i+\theta(\mu\theta+\mu-\alpha))}{\Gamma(\theta(\mu\theta+\mu-\alpha))}\right] \\ & - \sum_{i=1}^{n} \ln\left[y_i!(\theta+1)^{y_i+1}\right] \end{aligned} . \tag{3.1}$$

where $\boldsymbol{\tau} = (\alpha, \theta, \mu)$ is the unknown parameter vector. The score vector components can be obtained by taking partial derivatives of (3.1) with respect to α, θ, μ. The likelihood equation does not have the explicit solution. In this case, we should prefer the direct maximization of the log-likelihood function given in (3.1). For this purpose, the optimization toolboxes of the R, S-Plus or Matlab can be used. The **nlm** function of the R software is used in this study. To construct the asymptotic confidence intervals, we need the observed information matrix whose elements are given by

$$\mathbf{I}_F(\boldsymbol{\tau}) = -\begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\theta} & I_{\alpha\mu} \\ I_{\theta\alpha} & I_{\theta\theta} & I_{\theta\mu} \\ I_{\mu\alpha} & I_{\mu\theta} & I_{\mu\mu} \end{pmatrix}. \tag{3.2}$$

Since the second derivatives of the log-likelihood function are complicated, these equations are omitted, however, these are upon request from the authors. The inverse of the observed information matrix evaluated at $\hat{\boldsymbol{\tau}}$ gives the asymptotic variance-covariance matrix. The asymptotic standard errors are obtained by the inverse of (3.2). Then, the asymptotic confidence intervals of the parameters are given by

$$\widehat{\alpha} \pm z_{p/2}\sqrt{Var(\hat{\alpha})}, \quad \widehat{\theta} \pm z_{p/2}\sqrt{Var(\hat{\theta})}, \quad \widehat{\mu} \pm z_{p/2}\sqrt{Var(\hat{\mu})},$$

where $z_{p/2}$ represents the left quantile value of the standard normal distribution at $p/2$.

3.2. Simulation

Now, we conduct a simulation study to see the finite-sample performance of the MLEs of the parameters of the PGL distribution. The below simulation steps are implemented.

(1) Determine the sample size n, simulation replication N and the parameter values of the PGL distribution, α, θ and μ.

(2) Using the parameter settings in Step 1, generate the random variables from the PGL distribution using the inverse transform method.

(3) Using the generated sample in Step 2, obtain the MLEs of the parameters α, θ and μ.

(4) Repeat N times the Steps 2 and 3.

(5) Using the MLEs and the true parameter values, compute the estimated values of biases, mean square errors (MSEs) and mean relative estimates (MREs). The required formulas for these measures can be found in Altun [5].

The simulation results are displayed in Figure 4. We determine the simulation replication $N = 10,000$ and the sample size $n = 50, 55, 60, \ldots, 500$. The true parameter values are $\boldsymbol{\tau} = (2, 2, 5)$. We expect to see that when the sample size becomes larger, the biases and MSEs should be near zero and MRE should be near one. From the results given in Figure 4, we conclude that the estimated biases and MSEs are near the zero. Also, as expected, the MREs are near the one. These results show that the MLE is an appropriate method to estimate the unknown parameters of the PGL distribution.

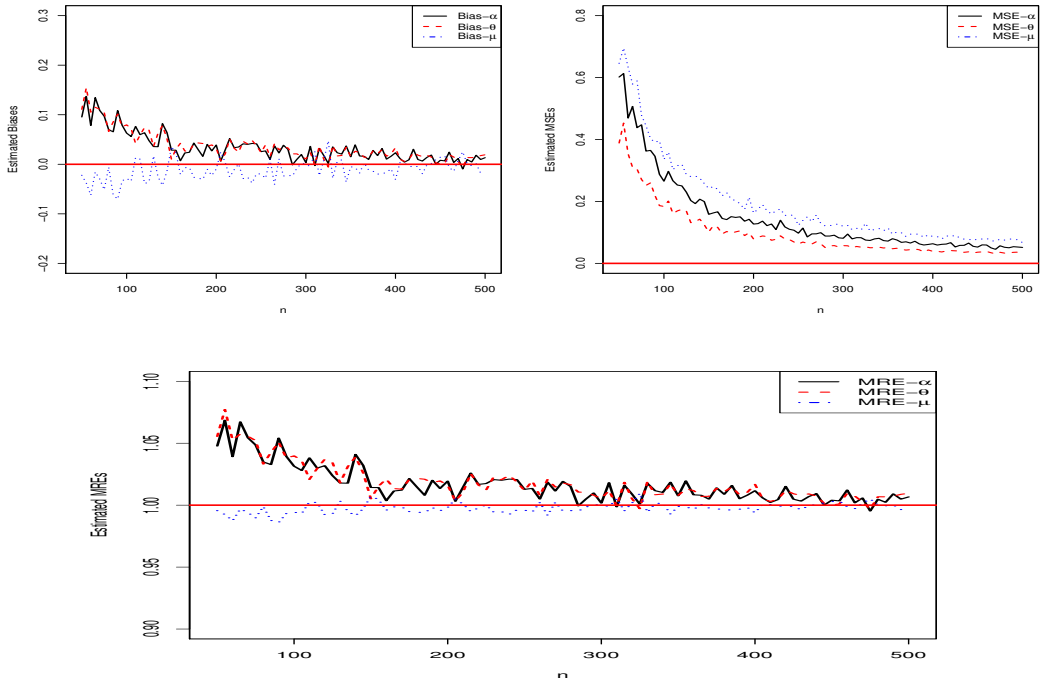

Figure 4. The graphical simulation results of the PGL distribution.

Additionally, two different parameter settings are evaluated to check whether the similar results are obtained for the different parameter vectors. The results are reported in Table 1. As in previous simulation study, the biases are near zero and MSE and MRE approach their desired values. Consequently, MLE is effective parameter estimation method for the PGL distribution.

Table 1. The simulation results of the PGL distribution for two different parameter settings.

Sample size	Parameters	$(\alpha = 2, \theta = 1, \mu = 5)$			$(\alpha = 3, \theta = 2, \mu = 4)$		
		α	θ	μ	α	θ	μ
n=50	Bias	0.4814	0.1297	-0.0321	0.2415	0.1503	-0.0631
	MSE	0.8389	0.3301	0.3394	0.9317	0.7125	0.3579
	MRE	1.2408	1.1297	0.9936	1.0805	1.0751	0.9842
n=250	Bias	0.0548	0.0134	-0.0202	0.0699	0.0423	-0.0172
	MSE	0.4491	0.0746	0.0761	0.3785	0.1035	0.0685
	MRE	1.0274	1.0134	0.9960	1.0233	1.0211	0.9957
n=500	Bias	0.0234	0.0047	-0.0146	0.0245	0.0209	-0.0164
	MSE	0.2180	0.0354	0.0366	0.1624	0.0445	0.0335
	MRE	1.0117	1.0047	0.9971	1.0082	1.0105	0.9959

4. Poisson generalized-Lindley regression model

As mentioned before, the Poisson regression model does not work well in case of over-dispersion. Dealing with the over-dispersed data set, the first choice is NB regression model. Now, we introduce an alternative regression model to the NB regression model for modeling the highly over-dispersed data sets. Assume that Y is a random variable distributed as a PGL distribution, given in (2.1). Since the mean of Y is $E(Y|\alpha, \theta, \mu) = \mu$, the log-link function can be used to link the covariates to the mean of the PGL distribution, as follows

$$\mu_i = \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right), i = 1, ..., n, \tag{4.1}$$

where $\boldsymbol{x}_i^T = (1, x_{i1}, x_{i2}, ...x_{ik})$ represents the covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ...\beta_k)^T$ represents the regression parameters. Note that the log-link function is used to link the covariates to the mean of the response variable. Since the mean of the response variable is defined on the positive domain, the link function should convert the observations defined on $\mathbb{R}$ to $\mathbb{R}_+$. However, the log-link function is not the only option to do this transformation. The softplus function, proposed by Weiss et al. [35], can be used as an alternative to the log-link function. Replacing μ in (2.1) with (4.1), the log-likelihood function of the PGL regression model is

$$\begin{aligned} \ell(\alpha, \theta, \boldsymbol{\beta}) = & \sum_{i=1}^{n} \ln \left[\begin{array}{l} \left(\frac{\theta}{\theta+1}\right)^{\alpha} \frac{\theta \Gamma(y_i+\alpha)}{\Gamma(\alpha)} + \left(\frac{\theta}{\theta+1}\right)^{\theta\left(\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\theta+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-\alpha\right)} \\ \\ \times \frac{\Gamma\left(y_i+\theta\left(\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\theta+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-\alpha\right)\right)}{\Gamma\left(\theta\left(\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\theta+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-\alpha\right)\right)} \end{array} \right] \\ & - \sum_{i=1}^{n} \ln\left[y_i!(\theta+1)^{y_i+1}\right]. \end{aligned} \tag{4.2}$$

The parameters α and θ are the distributional parameters and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ...\beta_k)$ is the vector of unknown regression parameters. These parameters are estimated by direct maximization of (4.2). The **nlm** function of R software is used to minimize the minus of (4.2), which is equivalent to the maximization of (4.2). The standard errors of the estimated parameters are obtained by means of hessian matrix evaluated at the MLEs of the parameters. The elements of the hessian matrix are computed with **fdHess** function of R software. The elements of the hessian matrix consist of the second-order partial derivatives of the log-likelihood function. Since these derivatives are complicated, they are omitted and not presented in the study.

4.1. Simulation of PGL regression model

Now, we evaluate the suitability of the MLE method for estimating the parameters of the PGL regression model. The simulation replication number N is determined as $10,000$ and four sample sizes are used: $n = 50, 250, 500, 1000$. Using the log-link function, we generate random variables using the $\ln(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ with parameters $\beta_0 = 0.5, \beta_1 = 0.5, \beta_2 = 0.5$ and $\theta = 1$, $\alpha = 1$. The covariates are generated from the standard uniform distribution. The response variable, y_i, is generated based on the values of μ_i, α and θ. The simulation results are listed in Table 2. The average of estimates (AEs) are near the true parameter values for small and large sample sizes. The biases and MSEs are near the desired values. These results confirm the consistency property of the MLEs.

Table 2. The simulation results of PGL regression model.

Sample size	Parameters	β_0	β_1	β_2	θ	α
n=50	AE	0.4967	0.4973	0.4778	1.1607	1.3993
	MSE	0.0860	0.1402	0.1349	0.2728	0.8066
n=250	AE	0.5056	0.4935	0.4944	1.0620	1.1438
	MSE	0.0270	0.0452	0.0419	0.1381	0.2302
n=500	AE	0.4999	0.4826	0.5031	1.0359	1.0569
	MSE	0.0135	0.0225	0.0212	0.0687	0.1110
n=1000	AE	0.5012	0.4986	0.4953	1.0057	1.0109
	MSE	0.0064	0.0118	0.0097	0.0293	0.0413

Additionally, we compare the standard deviations of the estimators based on the simulated samples and fdHess function. The **sample** function of R software is use to generate bootstrap samples. The bootstrap samples are used to calculate the bootstrap standard errors of the model parameters. The bootstrap replication number is determined as $1,000$. The model parameters are $\beta_0 = 0.5, \beta_1 = 0.5, \beta_2 = 0.5$ and $\theta = 2$, $\alpha = 2$. The simulation results are reported in Table 3. The results show that the obtained standard errors using two different approaches are close to each other. Thus, it is verified that the fdHess function works well to obtain the asymptotic standard errors of the model parameters.

Table 3. Comparison of standard errors of the estimated parameters using bootstrap methodology and fdHess function.

Sample size	Standard Errors	Parameters				
		β_0	β_1	β_2	θ	α
n=50	fdHess	0.2514	0.3089	0.3157	0.6405	0.8302
	Bootstrap	0.2834	0.3528	0.3320	0.5448	0.7137
n=250	fdHess	0.1573	0.1900	0.1921	0.4134	0.5405
	Bootstrap	0.1607	0.1975	0.1767	0.3969	0.5589
n=500	fdHess	0.1035	0.1261	0.1255	0.2773	0.3624
	Bootstrap	0.1284	0.0798	0.1555	0.1938	0.2992

4.2. Zero-inflated PGL regression model

The ZIP and ZINB regression models are the most widely used models in case of the zero-inflation. ZINB regression model could be more appropriate choice in most cases since Poisson distribution does not model the over-dispersion. The zero-inflated Poisson distribution is given by

$$P(y;\lambda) = \begin{cases} w + (1-w)e^{-\lambda}, & y = 0, \\ (1-w)\frac{\lambda^y e^{-\lambda}}{y!}, & y > 0, \end{cases} \tag{4.3}$$

where $0 \leq w \leq 1$. It is easy to see that when the $w = 0$, the zero-inflated Poisson distribution reduces to Poisson distribution. As in PGL regression model, the mean of Poisson distribution, λ_i, is linked to covariates by means of log-link function. The probability of zero counts, w_i is linked to covariates by means of logit-link function which is given by

$$\ln\left(\frac{w_i}{1-w_i}\right) = \boldsymbol{z}_i^T\boldsymbol{\gamma}, \tag{4.4}$$

where $\boldsymbol{z}_i^T = (1, z_{i1}, z_{i2}, ...z_{ik})$ is the vector of covariates and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, ...\gamma_k)^T$ is the unknown vector of regression coefficients for zero process. The log-likelihood function of ZIP regression model is given by

$$\begin{aligned} \ell(\boldsymbol{\beta},\boldsymbol{\gamma}) = &\sum_{y_i=0} \ln\left[\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right) + \exp\left(-\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right)\right] + \sum_{y_i>0}\left[y_i\boldsymbol{x}_i^T\boldsymbol{\beta} - \exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) - \ln(y_i!)\right] \\ &- \sum_{i=1}^{n} \ln\left(\left[1 + \exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)\right]\right). \end{aligned} \tag{4.5}$$

The log-likelihood function given in (4.5) can be maximized by means of **nlm** function of R. As mentioned before, when the corresponding data displays over-dispersion, the negative-binomial regression model should be used. The pmf of zero-inflated negative-binomial distribution is given by

$$P(y;w,\lambda,\tau) = \begin{cases} w + (1-w)\left(1+\frac{\lambda}{\tau}\right)^{-\tau}, & y = 0, \\ (1-w)\frac{\Gamma(y+\tau)}{y!\Gamma(\tau)}\left(1+\frac{\lambda}{\tau}\right)^{-\tau}\left(1+\frac{\tau}{\lambda}\right)^{-y}, & y > 0, \end{cases} \tag{4.6}$$

where τ is the shape parameter. When $w = 0$, the zero-inflated negative-binomial distribution reduces to negative-binomial distribution. The log-likelihood function of ZINB regression model is given by

$$\begin{aligned}\ell(\boldsymbol{\beta},\boldsymbol{\gamma},\tau) &= \sum_{i=1}^{n}\ln\left(1+\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)\right) - \sum_{y_i=0}\ln\left(\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)+\left(\frac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)+\tau}{\tau}\right)^{-\tau}\right)\\ &+ \sum_{y_i>0}\left(\tau\ln\left(\frac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)+\tau}{\tau}\right)+y_i\ln\left(1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\tau\right)\right)\right)\\ &+ \sum_{y_i>0}\left(\ln\Gamma(\tau)+\ln\Gamma(y_i+1)-\ln\Gamma(y_i+\tau)\right).\end{aligned} \tag{4.7}$$

The log-likelihood given in (4.7) can be maximized with **nlm** function of R software. Here, an alternative zero-inflated regression model is introduced based on the zero-inflated PGL distribution. The pmf of zero-inflated PGL distribution is given by

$$P(y;w,\alpha,\theta,\mu) = \begin{cases} w+(1-w)\left[\left(\frac{\theta}{\theta+1}\right)^{\alpha+1}+\frac{\theta^{\theta(\mu\theta+\mu-\alpha)}}{(\theta+1)^{\theta(\mu\theta+\mu-\alpha)+1}}\right], y=0,\\ (1-w)\frac{1}{y!(\theta+1)^{y+1}}\left[\begin{array}{l}\left(\frac{\theta}{\theta+1}\right)^{\alpha}\frac{\theta\Gamma(y+\alpha)}{\Gamma(\alpha)}+\\ \left(\frac{\theta}{\theta+1}\right)^{\theta(\mu\theta+\mu-\alpha)}\frac{\Gamma(y+\theta(\mu\theta+\mu-\alpha))}{\Gamma(\theta(\mu\theta+\mu-\alpha))}\end{array}\right], y>0,\end{cases} \tag{4.8}$$

where $0 \le w \le 1$ and $\alpha > 0$, $\theta > 0$ and $\mu > 0$. Inserting (4.1) and (4.4) in (4.8), the log-likelihood function of zero-inflated PGL (ZIPGL) regression model is given by

$$\begin{aligned}\ell(\boldsymbol{\beta},\boldsymbol{\gamma},\alpha,\theta) &= \sum_{y_i=0}\ln\left(\frac{\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)}{1+\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)}+\frac{1}{1+\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)}\left[\left(\frac{\theta}{\theta+1}\right)^{\alpha+1}+\frac{\theta^{\theta\left(\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\theta+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-\alpha\right)}}{(\theta+1)^{\theta\left(\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\theta+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-\alpha\right)+1}}\right]\right)\\ &\sum_{y_i>0}\ln\left(\frac{1}{\left[1+\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)\right]y_i!(\theta+1)^{y_i+1}}\left[\begin{array}{l}\left(\frac{\theta}{\theta+1}\right)^{\alpha}\frac{\theta\Gamma(y_i+\alpha)}{\Gamma(\alpha)}\\ +\left(\frac{\theta}{\theta+1}\right)^{\theta\left(\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\theta+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-\alpha\right)}\\ \times\frac{\Gamma\left(y_i+\theta\left(\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\theta+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-\alpha\right)\right)}{\Gamma\left(\theta\left(\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\theta+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)-\alpha\right)\right)}\end{array}\right]\right).\end{aligned} \tag{4.9}$$

The unknown parameters, α, θ, $\boldsymbol{\beta} = (\beta_0,\beta_1,\beta_2,...\beta_k)$ and $\boldsymbol{\gamma} = (\gamma_0,\gamma_1,\gamma_2,...\gamma_k)^T$ are obtained by maximizing the (4.9) with **nlm** function of R software.

5. Empirical study

In this section, two real data sets are analyzed to show the flexibility of the PGL regression model against the Poisson and NB regression models and also their zero-inflated counterparts. Also, we compare the PGL model with NPGL model, proposed by Altun [2]. In statistics literature, there are many discrete distributions to models the over or under-dispersed count data sets. Some of these distributions can be cited as follows: Mean Conway-Maxwell-Poisson distribution by Huang [18], zero-modified geometric by Kang et al. [21] and generalized COM-Poisson by Qian and Zhu [33]. The best model for the fitted data is chosen according to the results of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The lowest values of the AIC and BIC indicate the best model for the data used.

5.1. Absences of high school students

The first data set comes from the daily number of absences of 316 high school students. We model the daily number of absences with some covariates such as gender and type of instructional program.

The same data set was analyzed by Altun [5]. The female individuals are coded 1 and male individuals are coded 0. The gender is represented by the covariate (x_1). The instructional program variable has three categories. These are general, academic and vocational. Therefore, one of them is determined as the baseline category and two dummy variables are created. The baseline category is selected as the vocational program. The general program (x_2) and academic program (x_3) are used as two dummy variables. The below regression structure is fitted to the data set.

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}\right). \tag{5.1}$$

The probability distribution of the response variable, number of absence, is displayed in Figure 5(a). The mean and variance of the response variable are 5.955 and 49.518, respectively. Since the DI is greater than one, it is concluded that the response variable exhibits over-dispersion. Cameron and Trivedi [10] proposed a test to evaluate the over-dispersion. The **dispersiontest** function of R software is used to perform over-dispersion test of Cameron and Trivedi [10]. The obtained test statistic value is $z = 6.679$ and corresponding p value is < 0.001. This result verifies the over-dispersion problem in response variable.

(a) (b)

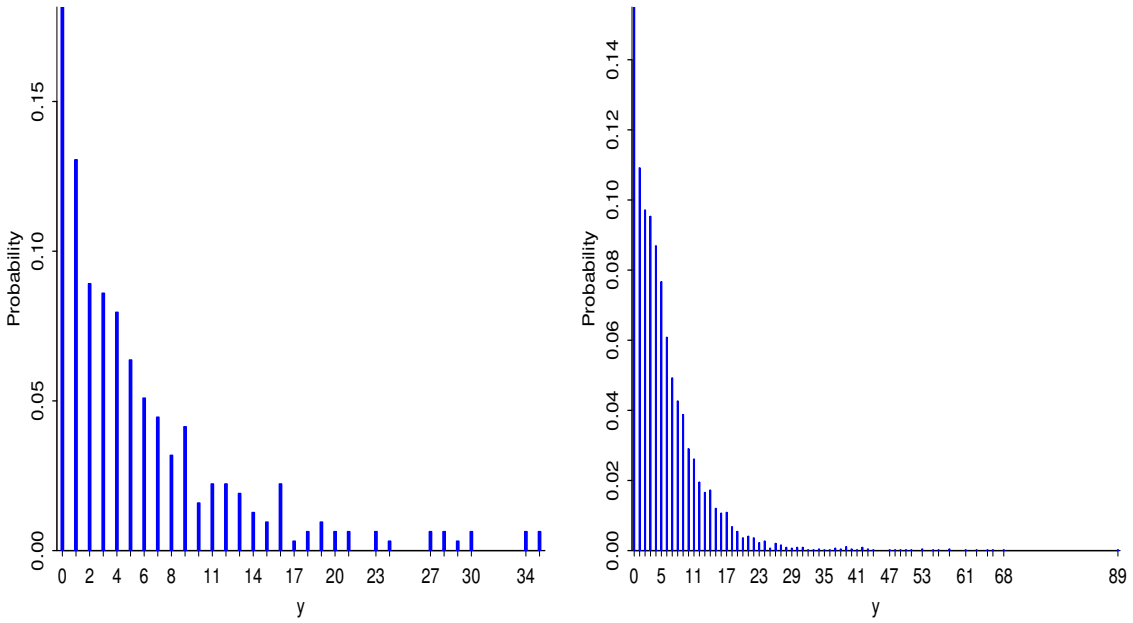

Figure 5. The probability distributions of (a) days of absence of students and (b) number of physician visits.

The estimated parameters with corresponding standard errors (SEs) and goodness of fit statistics are listed in Table 4. From Table 4, since the PGL regression model has the lowest values of AIC and BIC, we conclude that the PGL performs better than Poisson, NB and NPGL regression models.

Table 4. The estimated parameters of the fitted count regression models for the number of absence data.

Covariates	Poisson			NB			NPGL			PGL		
	Estimate (SE)	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
β_0	1.323	0.089	<0.001	1.271	0.214	<0.001	1.272	<0.001		1.515	0.192	<0.001
β_1	-0.234	0.047	<0.001	-0.193	0.123	0.118	-0.195	0.011		-0.233	0.101	0.022
β_2	1.374	0.076	<0.001	1.362	0.199	<0.001	1.365	<0.001		1.206	0.156	<0.001
β_3	0.957	0.066	<0.001	0.949	0.140	<0.001	0.951	<0.001		0.708	0.152	<0.001
τ	-	-	-	1.017	0.104	-	-	-	-	-	-	-
α	-	-	-	-	-	-	-	-		0.177	0.022	-
θ	-	-	-	-	-	-	1.059	0.136	-	1.040	0.459	-
$-\ell$	1343.250			867.225			881.409			862.480		
AIC	2694.500			1744.449			1772.818			1736.960		
BIC	2709.498			1763.196			1791.565			1759.456		

As seen from estimated regression coefficients of PGL regression model, we conclude that the gender, academic and instructional programs variables have statistically significant effects (at 1% level) on the days of absence for high school students. The number of absences for female students are $\exp(-0.233) = 0.792$, that is 20.8% lower than male students. It means that male students have higher absences than female students. However, the number of absences for general and academic instructional program students are $\exp(1.206) = 3.340$ that is 234% and $\exp(0.708) = 2.029$,that is 102.9% higher than the vocational instructional program students.

The profile log-likelihood plots of the PGL regression model are displayed in Figure 6. These figures are helpful to evaluate the suitability of the estimated model parameters. Thanks to the profile log-likelihood plots, it is obvious that the estimated parameters of the PGL model are the maximizers of the log-likelihood function.

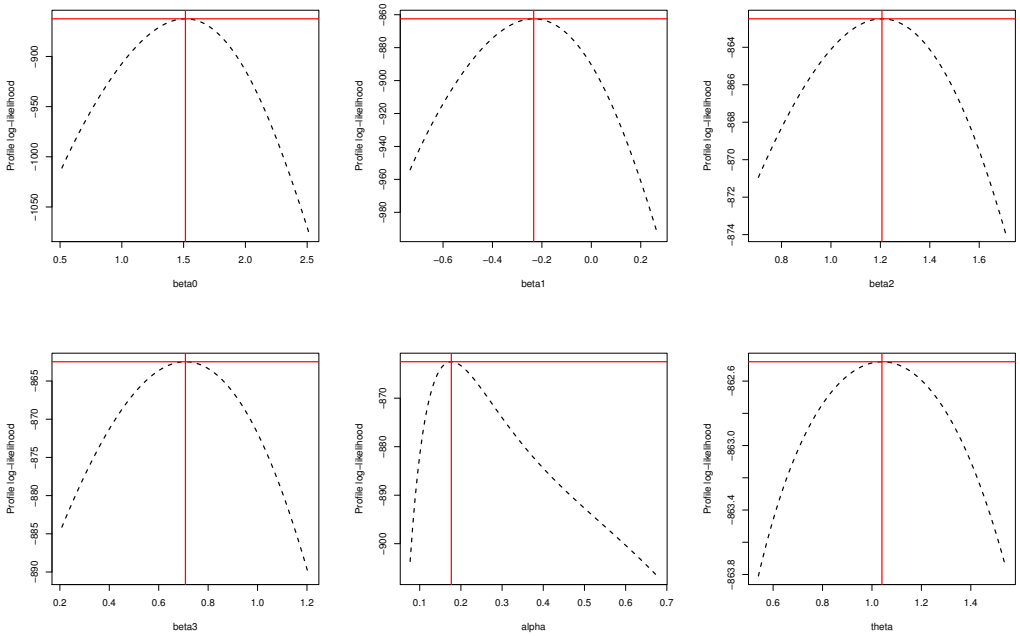

Figure 6. Profile log-likelihood plots of the PGL regression model.

The residual analysis is also applied to check the model suitability for the fitted data set. To do this, the randomized quantile residuals (rqs) are used. The rqs is

$$r_{q,i} = \Phi^{-1}(u_i), \tag{5.2}$$

where $u_i = F(y_i; \hat{\mu}_i)$ is uniformly distributed random variable between $a_i = \lim_{y \uparrow y_i} F(y; \hat{\mu}_i)$ and $b_i = F(y; \hat{\mu}_i)$ (Altun, [2]). Note that $F(y; \mu)$ is the cdf of PGL distribution. If the model is statistically valid for the data set, the rqs follows the standard normal distribution. The index and quantile-quantile plots of the PGL regression model are displayed in Figure 7. These figures show that the PGL regression model provides perfect fit to the data.

(a) (b)

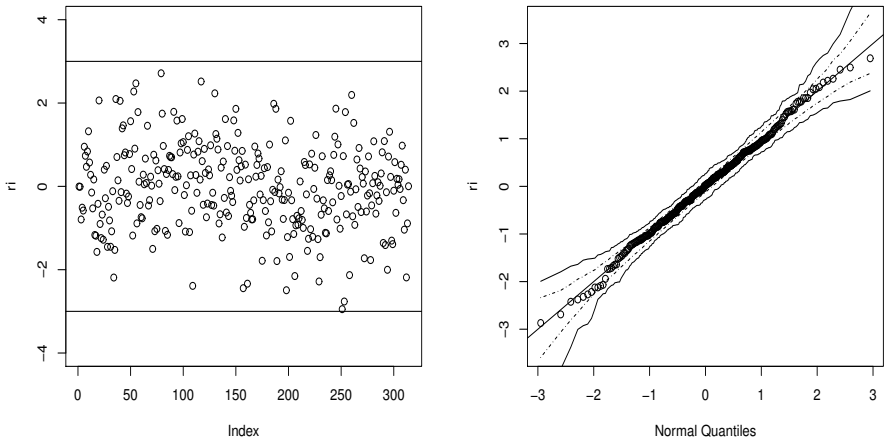

Figure 7. Residual analysis of PGL regression model.

5.2. NMES

The second data comes from the US National Medical Expenditure Survey (NMES) in the years of 1987–1988. The data set has 4406 observations. It can also be found in the **countreg** package of R software (see Zeileis et al. [32]). Here, our goal is to model the number of physician visits y, with following variables: number of hospital stay, (x_1), number of chronic conditions, (x_2), gender (female=0, male=1) (x_3), number of years of education, (x_4) and indicator of private insurance (yes=1, no=0), (x_5). The following model is fitted to NMES data set using the zero-inflated Poisson, negative binomial and PGL regression models.

$$\begin{aligned} \log(\mu_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}, \\ \text{logit}(w_i) &= \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \gamma_4 z_{4i} + \gamma_5 z_{5i}. \end{aligned} \tag{5.3}$$

The histogram of the number of physician visits are displayed in Figure 5(b). The mean and variance of the number of physician visits are 5.774 and 45.687, respectively. Since the DI of the response variable is greater than one, it exhibits over-dispersion. As in Section 5.1, the over-dispersion test of Cameron and Trivedi [10] is performed. The obtained test statistic value is $z = 6.679$ and corresponding p value is < 0.001. Therefore, the response variable has a over-dispersion. As seen from Figure 5(b),

the response variable is highly peaked at zero. To assess the zero-inflation, the test proposed by Van den Broek [29] is used. The **zero.test** is used for this purpose. The test statistic follows a chi-square distribution with one degree of freedom. The obtained test statistic value is $\chi^2 = 33438.09$ and its p-value is < 0.001. This result verifies that the frequency of zero process in response variable cannot be modeled by Poisson regression model. Therefore, zero-inflated regression models are needed to model such data sets. The AIC and BIC of the fitted count regression models are listed in Table 5. Since the data exhibits both over-dispersion and zero-inflation, Poisson and NB regression models do not perform well. According to the AIC and BIC values, PGL and ZIPGL models perform better than other models for NMES data set.

Table 5. Comparison of models for NMES data.

	Poisson	NB	PGL	ZIP	ZINB	ZIPGL
AIC	36314.70	24430.48	24263.38	32611.14	24286.48	24244.30
BIC	36353.04	24475.22	24314.51	32687.83	24369.56	24333.77

The estimated parameters of the fitted models as well as their standard errors are summarized in Table 6. Zero-inflated regression models have two parts to interpret. These parts are related the non-inflated and zero-inflated processes. As mentioned before, non-inflated process is modeled with log-link function and zero-inflated process is modeled by logit-link function. Therefore, the regression coefficients of zero-inflated process can be interpreted as odd ratio. As seen from estimated regression coefficients of ZIPGL regression model, for non-inflated process, all variables have statistically significant (at 1% level) effect on number of physician visits. According to zero-inflated process, number of chronic conditions and privative insurance variables have statistically significant effects on number of physician visits. Having private insurance decreases the odds of not having the opportunity of physician visits by $\exp(-1.174) = 0.309$, which is 69.1%.

Table 6. The estimated parameters of the fitted models for the NMES data.

Covariates	Poisson			NB			PGL			ZIP			ZINB			ZIPGL		
	Est.	SE	p	Est.	SE	p	Est.	SE	p	Est.	Se	p	Est.	SE	p	Est.	SE	p
β_0	0.987	0.037	<0.001	0.931	0.086	<0.001	0.897	0.074	<0.001	1.446	0.037	<0.001	1.245	0.088	<0.001	1.108	0.111	<0.001
β_1	0.182	0.006	<0.001	0.240	0.020	<0.001	0.208	0.017	<0.001	0.175	0.006	<0.001	0.223	0.020	<0.001	0.207	0.017	<0.001
β_2	0.175	0.004	<0.001	0.204	0.012	<0.001	0.186	0.009	<0.001	0.129	0.004	<0.001	0.155	0.012	<0.001	0.160	0.011	<0.001
β_3	-0.116	0.013	<0.001	-0.136	0.031	<0.001	-0.135	0.026	<0.001	-0.065	0.013	<0.001	-0.090	0.031	0.003	-0.118	0.030	<0.001
β_4	0.022	0.002	<0.001	0.021	0.004	<0.001	0.021	0.004	<0.001	0.015	0.002	<0.001	0.016	0.004	<0.001	0.017	0.005	<0.001
β_5	0.183	0.017	<0.001	0.192	0.040	<0.001	0.237	0.034	<0.001	0.061	0.017	<0.001	0.096	0.042	0.021	0.171	0.059	0.003
τ				1.180	0.033	-							1.443	0.034	-			
α							0.228	0.008	-							0.233	0.009	-
θ							1.516	0.143	-							1.316	0.128	-
Zero-inflated																		
γ_0										0.287	0.222	0.196	0.587	0.449	0.191	0.447	0.825	0.588
γ_1										-0.310	0.091	<0.001	-0.815	0.411	0.047	-0.479	0.581	0.410
γ_2										-0.542	0.044	<0.001	-1.319	0.186	<0.001	-1.920	0.814	0.018
γ_3										0.418	0.089	<0.001	0.635	0.201	0.002	0.447	0.370	0.227
γ_4										-0.056	0.012	<0.001	-0.090	0.026	<0.001	-0.106	0.063	0.093
γ_5										-0.751	0.102	<0.001	-1.181	0.221	<0.001	-1.174	0.544	0.031

As in the previous section, the profile log-likelihood plots are displayed to check the correctness of the estimated model parameters. According to the plots in Figure 8, the estimated model parameters of the ZIPGL model are the maximizers of the function, given in (4.9).

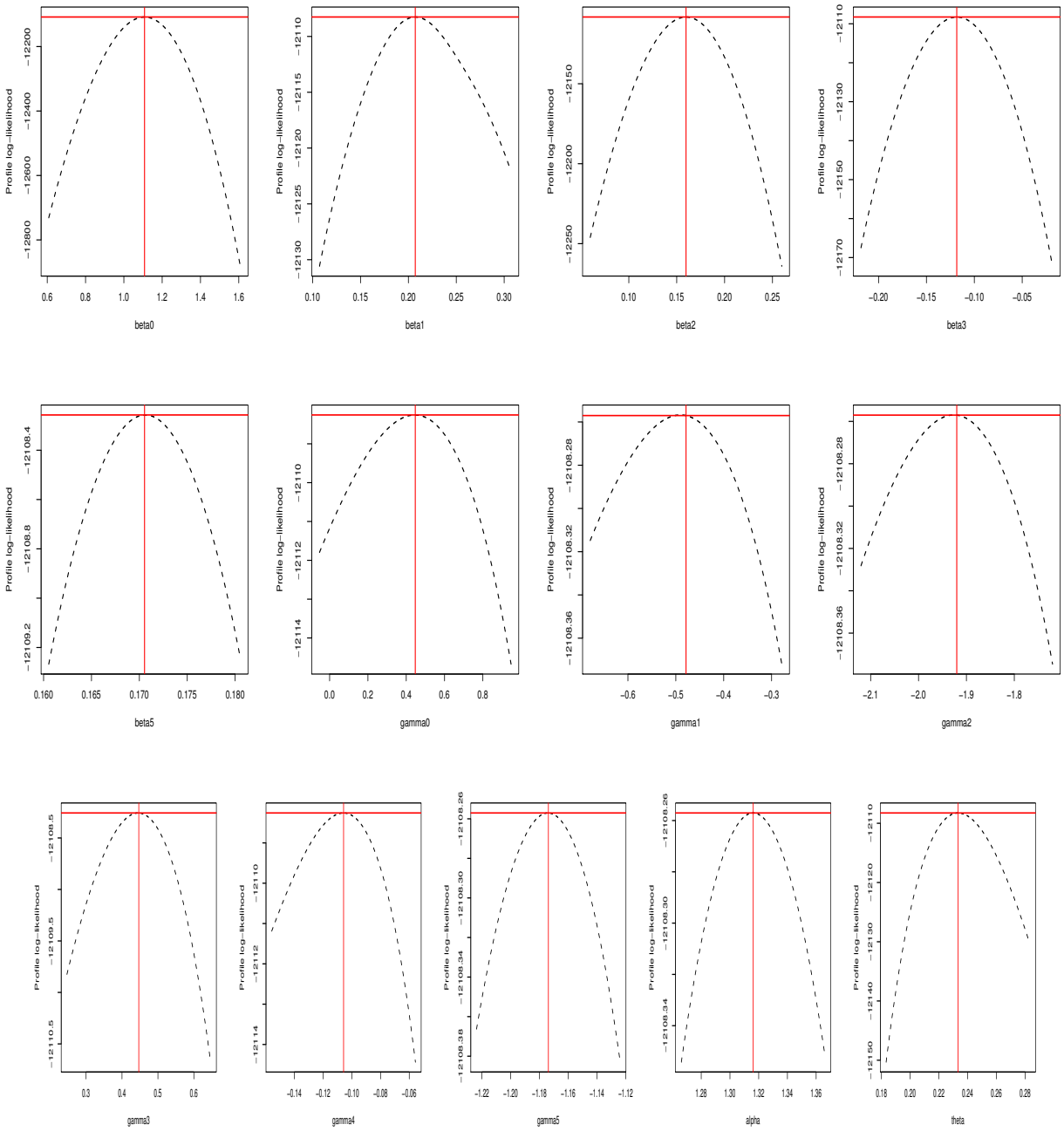

Figure 8. Profile log-likelihood plots of the ZIPGL regression model.

6. Conclusions

This study introduces a new count regression model for zero-inflated and over-dispersed count data sets based on the re-parametrization of the PGL distribution. The PGL regression model and its zero-inflated counterpart are studied. Two real data sets are analyzed to convince the readers in favor of the PGL regression model against the Poisson and NB regression models. Empirical findings show that PGL and ZIPGL regression models provide better fits than Poisson, ZIP, NB and ZINB regression models. As a future work of this study, one-inflated PGL regression model could be considered. The one-inflated regression models are useful for modeling the claim numbers in insurance. We hope that the PGL and ZIPGL regression models find a wider application area in different applied sciences.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The second author would like to thank the Deanship of Scientific Research at Qassim University for funding the publication of this project.

Conflict of interest

The authors have no conflicts of interest.

References

1. E. Altun, D. Bhati, N. M. Khan, A new approach to model the counts of earthquakes: INARPQX (1) process, *SN Appl. Sci.*, **3** (2021), 1–17. https://doi.org/10.1007/s42452-020-04109-8
2. E. Altun, A new two-parameter discrete poisson-generalized Lindley distribution with properties and applications to healthcare data sets, *Comput. Stat.*, **36** (2021), 2841–2861. https://doi.org/10.1007/s00180-021-01097-0
3. E. Altun, A new generalization of geometric distribution with properties and applications, *Commun. Stat.-Simu. Comput.*, **49** (2020), 793–807. https://doi.org/10.1080/03610918.2019.1639739
4. E. Altun, A new one-parameter discrete distribution with associated regression and integer-valued autoregressive models, *Math. Slovaca*, **70** (2020), 979–994. https://doi.org/10.1515/ms-2017-0407
5. E. Altun, A new model for over-dispersed count data: Poisson quasi-Lindley regression model, *Math. Sci.*, **13** (2019), 241–247. https://doi.org/10.1007/s40096-019-0293-5
6. E. Altun, A new zero-inflated regression model with application, *J. Stat.-Stat. Actuar. Sci.*, **11** (2018), 73–80.
7. E. Ayati, E. Abbasi, Modeling accidents on Mashhad urban highways, *Open J. Safety Sci. Technol.*, **4** (2014), 22–35. https://doi.org/10.4236/ojsst.2014.41004

8. E. Avci, S. Alturk, E. N. Soylu, Comparison count regression models for overdispersed alga data, *Int. J. Recent Res. Appl. Stud.*, **25** (2015), 1–5.
9. D. Bhati, P. Kumawat, E. Gómez-Déniz, A new count model generated from mixed Poisson transmuted exponential family with an application to health care data, *Commun. Stat.-Theor. M.*, **46** (2017), 11060–11076. https://doi.org/10.1080/03610926.2016.1257712
10. A. C. Cameron, P. K. Trivedi, *Regression analysis of count data*, Cambridge University Press, Cambridge, 1998. https://doi.org/10.1017/CBO9780511814365
11. L. Cheng, S. R. Geedipally, D. Lord, The Poisson-Weibull generalized linear model for analyzing motor vehicle crash data, *Safety Sci.*, **54** (2013), 38–42. https://doi.org/10.1016/j.ssci.2012.11.002
12. I. Elbatal, F. Merovci, M. Elgarhy, A new generalized Lindley distribution, *Math. Theor. Model.*, **3** (2013), 30–47.
13. M. S. Eliwa, E. Altun, M. El-Dawoody, M. El-Morshedy, A new three-parameter discrete distribution with associated INAR (1) process and applications, *IEEE Access*, **8** (2020), 91150–91162. https://doi.org/10.1109/ACCESS.2020.2993593
14. M. El-Morshedy, E. Altun, M. S. Eliwa, A new statistical approach to model the counts of novel coronavirus cases, *Math. Sci.*, 2021, 1–14. https://doi.org/10.1007/s40096-021-00390-9
15. M. El-Morshedy, M. S. Eliwa, E. Altun, Discrete Burr-Hatke distribution with properties, estimation methods and regression model, *IEEE Access*, **8** (2020), 74359–74370. https://doi.org/10.1109/ACCESS.2020.2988431
16. Y. Gencturk, A. Yigiter, Modelling claim number using a new mixture model: Negative binomial gamma distribution, *J. Stat. Comput. Simu.*, **86** (2016), 1829–1839. https://doi.org/10.1080/00949655.2015.1085987
17. E. Gómez-Déniz, A new discrete distribution: Properties and applications in medical care, *J. Appl. Stat.*, **40** (2013), 2760–2770. https://doi.org/10.1080/02664763.2013.827161
18. A. Huang, Mean-parametrized Conway-Maxwell-Poisson regression models for dispersed counts, *Stat. Model.*, **17** (2017), 359–380. https://doi.org/10.1177/1471082X17697749
19. N. Ismail, H. Zamani, *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** (2013), 1–28.
20. T. Imoto, C. M. Ng, S. H. Ong, S. Chakraborty, A modified Conway-Maxwell-Poisson type binomial distribution and its applications, *Commun. Stat.-Theor. M.*, **46** (2017), 12210–12225. https://doi.org/10.1080/03610926.2017.1291974
21. Y. Kang, F. Zhu, D. Wang, S. Wang, A zero-modified geometric INAR (1) model for analyzing count time series with multiple features, *Can. J. Stat.*, 2023. https://doi.org/10.1002/cjs.11774
22. D. Lord, S. P. Washington, J. N. Ivan, Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory, *Accident Anal. Prev.*, **37** (2005), 35–46. https://doi.org/10.1016/j.aap.2004.02.004
23. D. Lord, S. R. Geedipally, The negative binomial-Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros, *Accident Anal. Prev.*, **43** (2011), 1738–1742. https://doi.org/10.1016/j.aap.2011.04.004

24. E. Mahmoudi, H. Zakerzadeh, Generalized Poisson-lindley distribution, *Commun. Stat.-Theor. M.*, **39** (2010), 1785–1798. https://doi.org/10.1080/03610920902898514

25. J. Rodríguez-Avi, A. Conde-Sínchez, A. J. Sáez-Castillo, M. J. Olmo-Jiménez, A. M. Martínez-Rodríguez, A generalized Waring regression model for count data, *Comput. Stat. Data Anal.*, **53** (2009), 3717–3725. https://doi.org/10.1016/j.csda.2009.03.013

26. G. Shmueli, T. P. Minka, J. B. Kadane, S. Borle, P. Boatwright, A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution, *J. Roy. Stat. Soc. C-Appl.*, **54** (2005), 127–142. https://doi.org/10.1111/j.1467-9876.2005.00474.x

27. A. J. Sáez-Castillo, A. Conde-Sánchez, A hyper-Poisson regression model for overdispersed and underdispersed count data, *Comput. Stat. Data Anal.*, **61** (2013), 148–157. https://doi.org/10.1016/j.csda.2012.12.009

28. M. M. Shoukri, M. H. Asyali, R. VanDorp, D. Kelton, The Poisson inverse Gaussian regression model in the analysis of clustered counts data, *J. Data Sci.*, **2** (2004), 17–32. https://doi.org/10.6339/JDS.2004.02(1).135

29. J. Van den Broek, A score test for zero inflation in a Poisson distribution, *Biometrics*, 1995, 738–743. https://doi.org/10.2307/2532959

30. H. Zamani, N. Ismail, P. Faroughi, Poisson-weighted exponential univariate version and regression model with applications, *J. Math. Stat.*, **10** (2014), 148–154. https://doi.org/10.3844/jmssp.2014.148.154

31. W. Wongrin, W. Bodhisuwan, Generalized Poisson-Lindley linear model for count data, *J. Appl. Stat.*, **44** (2017), 2659–2671. https://doi.org/10.1080/02664763.2016.1260095

32. A. Zeileis, C. Kleiber, S. Jackman, Regression models for count data in R, *J. Stat. Softw.*, **27** (2008), 1–25. https://doi.org/10.18637/jss.v027.i08

33. L. Qian, F. Zhu, A flexible model for time series of counts with overdispersion or underdispersion, zero-inflation and heavy-tailedness, *Commun. Math. Stat.*, 2023, 1–24. https://doi.org/10.1007/s40304-022-00327-1

34. W. Wongrin, W. Bodhisuwan, The Poisson-generalised Lindley distribution and its applications, *Songklanakarin J. Sci. Technol.*, **38** (2016), 654–656.

35. C. H. Weiss, F. Zhu, A. Hoshiyar, Softplus INGARCH models, *Stat. Sinica*, **32** (2022), 1099–1120. https://doi.org/10.5705/ss.202020.0353

AIMS Press