Commun. Math. Res.
doi: 10.4208/cmr.2021-0045

Vol. **38**, No. 1, pp. 28-51
February 2022

The Binomial-Discrete Poisson-Lindley Model: Modeling and Applications to Count Regression

Christophe Chesneau[1], Hassan S. Bakouch[2], Yunus Akdoğan[3,*] and Kadir Karakaya[3]

[1] *LMNO, University of Caen-Normandie, Caen, France.*
[2] *Department of Mathematics, Faculty of Science, Tanta University, Tanta, Egypt.*
[3] *Department of Statistics, Faculty of Science, Selcuk University, Konya, Turkey.*

Received 13 April 2021; Accepted 6 September 2021

Abstract. On the basis of a well-established binomial structure and the so-called Poisson-Lindley distribution, a new two-parameter discrete distribution is introduced. Its properties are studied from both the theoretical and practical sides. For the theory, we discuss the moments, survival and hazard rate functions, mode and quantile function. The statistical inference on the model parameters is investigated by the maximum likelihood, moments, proportions, least square, and weighted least square estimations. A simulation study is conducted to observe the performance of the bias and mean square error of the obtained estimates. Then, applications to two practical data sets are given. Finally, we construct a new flexible count data regression model called the binomial-Poisson Lindley regression model with two practical examples in the medical area.

AMS subject classifications: 60E05, 62E10, 62E15

Key words: Binomial-discrete family, Poisson-Lindley distribution, estimation, data analysis, count regression.

*Corresponding author. *Email address:* yakdogan@selcuk.edu.tr (Y. Akdoğan)

1 Introduction

In physics, counts are encountered as radioactive decay, or photon counting using a Geiger tube, and for modelling such physical issue, Hu *et al.* [9] introduced the binomial-discrete family of discrete distributions characterized by the compounding of two discrete distributions: The binomial distribution and a generic discrete distribution with support $\mathbb{N}=\{0,1,\ldots\}$. It is defined by the following probability mass function (PMF):

$$g(x;p,\zeta)=\sum_{n=x}^{\infty}\binom{n}{x}p^x(1-p)^{n-x}P_\zeta(N=n), \tag{1.1}$$

where $p\in(0,1)$ and N denotes a discrete random variable with support $\mathbb{N}$, which may depend on a parameter vector denoted by ζ. Hu *et al.* [9] studied the special member of the family defined with N following the Poisson random variable with parameter λ, and showed that $g(x;p,\lambda)$ corresponds to the PMF of a Poisson distribution with parameter λp. But the Poisson distribution is not always suitable for modeling and analysis of data counts, because its mean and variance are equal. Therefore, there is a need to introduce other distributions for counts. Recently, Kus *et al.* [13] investigated another special member of the family that arises with N following the discrete Lindley distribution. The binomial-discrete Lindley distribution has seen the light. Then, it was proved that it could provide better fits than the former discrete Lindley distribution for six different data sets. Several extensions of the binomial-discrete family can be found in Akdogan *et al.* [1] and Deniz [7].

In this paper, we investigate a new special member of the binomial-discrete family from a statistical point of view and, based on it, we set up a new regression model. This member assumes that N follows a well-known extension of the discrete Lindley distribution: the so-called the Poisson-Lindley (PL) distribution. The motivations behind the PL distribution are recalled below. First of all, it is defined by the following PMF:

$$f(x;\theta)=\frac{\theta^2(x+\theta+2)}{(1+\theta)^{x+3}},\quad x\in\mathbb{N}, \tag{1.2}$$

where $\theta>0$. The PL distribution possesses tractable probability functions, as well as desirable properties, such as unimodality, overdispersion and increasing hazard rate function (HRF). Thanks to its skewness and kurtosis features, it can provide a better alternative to the Poisson, geometric and negative binomial distributions for modelling purposes. Further theoretical facts and applications for the PL distribution can be found in Sankaran [17] and Shanker and Fesshaye [18].

Hence, we introduce a new two-parameter discrete distribution with support $\mathbb{N}$, called the binomial Poisson-Lindley (Bin-PL) distribution with parameters p and θ. It is defined by the PMF in (1.1) with $P_\theta(N=n)=f(n;\theta)$. It has the PL distribution as a sub-distribution. Moreover, it is a mixture of geometric and negative binomial distributions. Hence, properties of the negative binomial distribution can be useful for determining those of the Bin-PL distribution. Further, the Bin-PL distribution has a bimodality feature besides the unimodal and decreasing ones.

Bimodal distributions can be seen in traffic analysis, where traffic peaks appear during the AM rush hour and then again during the PM rush hour. Also, this phenomenon is observed in medicine and in daily water distribution, as in the forms of showering and cooking in the morning and evening periods. Also, properties of the distribution recommend it for analyzing rightly skewed and leptokurtic data, as shall be shown later. Based on the introduced distribution, we propose a new flexible count data regression model called the Bin-PL model. Two practical examples are provided to show that the Bin-PL regression model works very well, and this is confirmed by comparing it with the classical Poisson, uniform-Poisson (UP) [7] and Bell [4] models.

The rest of the paper is structured as follows. Section 2 presents the basics of the Bin-PL distribution. Many of its properties, such as unimodality of the PMF, quantile function, stochastic orderings, and moments are described in Section 3. Several point estimation methods are considered for the Bin-PL model parameters and simulation studies are conducted to check the capacity of these methods in Section 4. Two applications are given to show the practicality of the Bin-PL model in Section 5. In Section 6, a new count regression model is introduced and an extra two practical data applications are carried out to show that the new model is useful for analyzing the practical data. Some concluding remarks end the paper in Section 7.

2 Basics on the Bin-PL distribution

First of all, let us discuss some basics of the Bin-PL distribution, as well as its immediate properties.

2.1 On the probability mass function

The result below presents a simple expression of the PMF of the Bin-PL distribution.

Proposition 2.1. *The PMF of the Bin-PL distribution with parameters p and θ can be expressed as*

$$g(x;p,\theta)=\frac{\theta^2}{1+\theta}\frac{p^x}{(\theta+p)^{x+2}}(x+\theta+p+1),\quad x\in\mathbb{N}. \tag{2.1}$$

Proof. For $|u|<1$, the following formulas hold:

$$\sum_{n=x}^{+\infty}\binom{n}{x}nu^n=\frac{1}{(1-u)^{x+2}}u^x(u+x),\quad \sum_{n=x}^{+\infty}\binom{n}{x}u^n=\frac{1}{(1-u)^{x+1}}u^x.$$

Therefore, based on (1.1) and (1.2), after some algebra, we get

$$\begin{aligned}
&g(x;p,\theta)=\sum_{n=x}^{+\infty}\binom{n}{x}p^x(1-p)^{n-x}\frac{\theta^2(n+\theta+2)}{(1+\theta)^{n+3}}\\
&=\frac{\theta^2}{(1+\theta)^3}p^x(1-p)^{-x}\left[\sum_{n=x}^{+\infty}\binom{n}{x}n\left(\frac{1-p}{1+\theta}\right)^n+(\theta+2)\sum_{n=x}^{+\infty}\binom{n}{x}\left(\frac{1-p}{1+\theta}\right)^n\right]\\
&=\frac{\theta^2}{(1+\theta)^3}p^x(1-p)^{-x}\left[\left(\frac{\theta+1}{\theta+p}\right)^2\left(\frac{1-p}{\theta+p}\right)^x\left(\frac{1-p}{1+\theta}+x\right)+(\theta+2)\frac{\theta+1}{\theta+p}\left(\frac{1-p}{\theta+p}\right)^x\right]\\
&=\frac{\theta^2}{1+\theta}\frac{p^x}{(\theta+p)^{x+2}}(x+\theta+p+1).
\end{aligned}$$

This ends the proof of Proposition 2.1. □

In the next of the study, the Bin-PL distribution will be sometimes denoted by Bin-PL (p,θ) when the parameters need to be specified.

As a first remark, when $p=1$, we get the PL distribution, so our new distribution generalizes it. Also, for $\theta=1$, we see mixing portions are half. Furthermore, we can express $g(x;p,\theta)$ as

$$\begin{aligned}
g(x;p,\theta)&=\frac{\theta^2}{(1+\theta)(\theta+p)}\left(\frac{p}{\theta+p}\right)^x+\frac{\theta^2}{(1+\theta)(\theta+p)^2}(x+1)\left(\frac{p}{\theta+p}\right)^x\\
&=\frac{\theta}{1+\theta}g_1(x;p,\theta)+\frac{1}{1+\theta}g_2(x;p,\theta),
\end{aligned} \tag{2.2}$$

where for $i\in\{1,2\}$, $g_i(x;p,\theta)$ is the PMF of the negative binomial distribution denoted by $NB(i,p/(\theta+p))$, i.e.,

$$g_i(x;p,\theta)=(x+1)^{i-1}\left(1-\frac{p}{\theta+p}\right)^i\left(\frac{p}{\theta+p}\right)^x,\quad x\in\mathbb{N}.$$

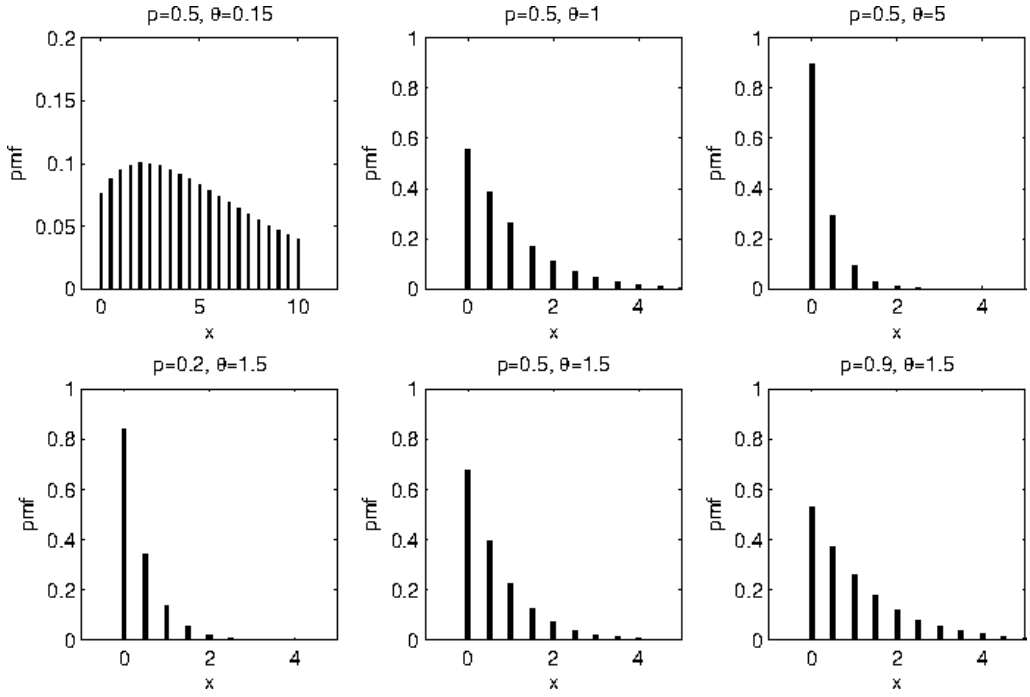

Figure 1: PMF of the Bin-PL distribution for different values of p and θ.

Hence, our Bin-PL distribution can be viewed as a mixture of negative binomial (geometric) $NB(1,p/(\theta+p))$ $(=G(p/(\theta+p)))$ and negative binomial $NB(2,p/(\theta+p))$ distributions, with mixing proportions $\theta/(1+\theta)$ and $1/(1+\theta)$, respectively. Consequently, the well-known properties of the negative binomial distribution can be useful to determine those of the Bin-PL distribution.

Fig. 1 presents the plots of the PMF of the Bin-PL distribution for some choices of p and θ. From Fig. 1, we observe that the PMF can be decreasing and unimodal when x is increasing.

3 On the properties of the new distribution

Some properties of the Bin-PL distribution are now discussed. First, we have

$$g(0;p,\theta)=\frac{\theta^2}{1+\theta}\frac{\theta+p+1}{(\theta+p)^2}$$

and

$$\frac{g(x+1;p,\theta)}{g(x;p,\theta)}=\frac{p}{\theta+p}\left(1+\frac{1}{x+\theta+p+1}\right),$$

which is clearly a decreasing function in x, implying the unimodality of the Bin-PL distribution. Furthermore, we have

$$\frac{g(x+2;p,\theta)g(x;p,\theta)}{[g(x+1;p,\theta)]^2}=1-\frac{1}{(x+\theta+p+2)^2}<1,$$

implying that $g(x;p,\theta)$ is log-concave. As an immediate consequence, the Bin-PL distribution has an increasing failure rate (see Johnson *et al.* [10, p. 209]).

Corollary 3.1. *Since log-concave probability mass functions are strongly unimodal (see Keilson and Gerber [11]), the Bin-PL distribution is unimodal.*

Corollary 3.2. *The mode of the Bin-PL distribution is given by*

$$mod=\begin{cases}\lfloor m\rfloor, & g\big(\lfloor m\rfloor;p,\theta\big)>g\big(\lfloor m\rfloor+1;p,\theta\big),\\ \lfloor m\rfloor+1, & g\big(\lfloor m\rfloor;p,\theta\big)<g\big(\lfloor m\rfloor+1;p,\theta\big),\end{cases}$$

where

$$m=\frac{-(p+\theta+1)-1}{\log(p)-\log(p+\theta)}$$

and $\lfloor x\rfloor$ denotes the integer part of x. When

$$\theta=w=\frac{1}{2}\left(\sqrt{p^2+10p+9}-(p+3)\right),$$

we have

$$g\big(\lfloor m\rfloor;p,\theta\big)=g\big(\lfloor m\rfloor+1;p,\theta\big),$$

and the PMF of the Bin-PL distribution is bimodal with modes given by $\lfloor m\rfloor$ and $\lfloor m\rfloor+1$. These facts are illustrated in Fig. 2.

The result below presents a simple expression of the cumulative distribution function (CDF) of the Bin-PL distribution.

Proposition 3.1. *The CDF of the Bin-PL distribution can be expressed as, for any integer t,*

$$F(t;p,\theta)=1-\frac{\big(\theta^2+p\theta+2\theta+t\theta+p\big)\,p^{t+1}}{(1+\theta)\,(p+\theta)^{t+2}}. \tag{3.1}$$

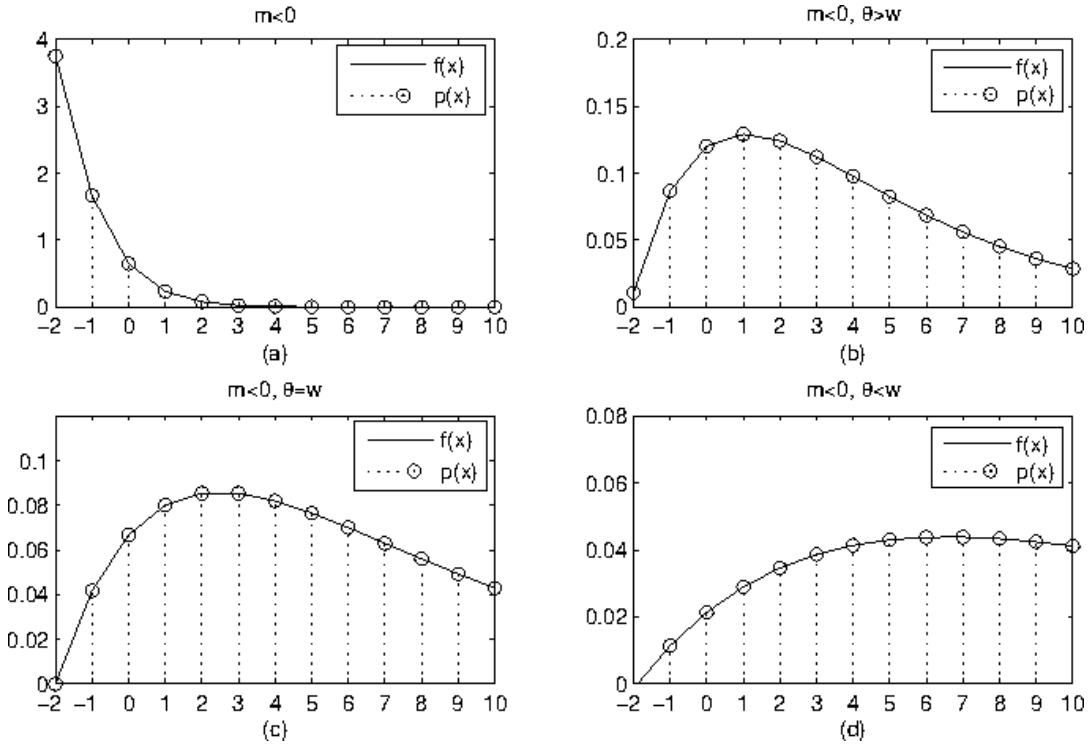

Figure 2: PMF for different values of w and m.

Proof. It follows from the mixture representation (2.2), geometric series expansions and some algebra, that

$$\begin{aligned}
&F(t;p,\theta)=\sum_{x=0}^{t}g(x;p,\theta)\\
&=\frac{\theta^2}{(1+\theta)(\theta+p)}\sum_{x=0}^{t}\left(\frac{p}{\theta+p}\right)^x+\frac{\theta^2}{(1+\theta)(\theta+p)^2}\sum_{x=0}^{t}(x+1)\left(\frac{p}{\theta+p}\right)^x\\
&=\frac{\theta}{1+\theta}\left[1-\left(\frac{p}{\theta+p}\right)^{t+1}\right]+\frac{1}{1+\theta}\left[1+(t+1)\left(\frac{p}{\theta+p}\right)^{t+2}-(t+2)\left(\frac{p}{\theta+p}\right)^{t+1}\right]\\
&=1-\frac{1}{1+\theta}\left[(t+\theta+2)\left(\frac{p}{\theta+p}\right)^{t+1}-(t+1)\left(\frac{p}{\theta+p}\right)^{t+2}\right]\\
&=1-\frac{\left(\theta^2+p\theta+2\theta+t\theta+p\right)p^{t+1}}{(1+\theta)(p+\theta)^{t+2}}.
\end{aligned}$$

This ends the proof of Proposition 3.1. □

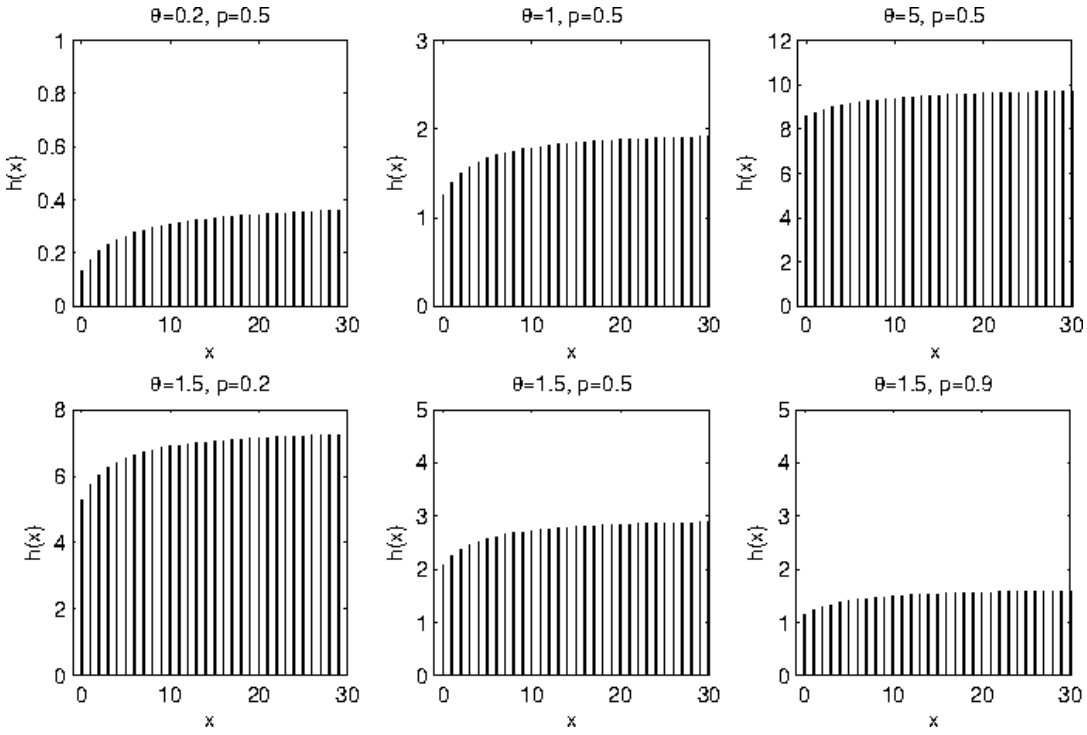

Figure 3: HRF of the Bin-PL distribution for different values of p and θ.

Using (2.1), the survival function of the Bin-PL distribution is given by, for any integer t,

$$S(t;p,\theta)=1-F(t;p,\theta)=\frac{\left(\theta^2+p\theta+2\theta+t\theta+p\right)p^{t+1}}{(1+\theta)(p+\theta)^{t+2}}. \tag{3.2}$$

Using (2.1) and (3.2), the HRF can be written as

$$h(x;p,\theta)=\frac{g(x;p,\theta)}{S(x;p,\theta)}=\frac{\theta^2(1+\theta+x+p)}{p(p\theta+p+x\theta+2\theta+\theta^2)}.$$

Bin-PL distribution has a non decreasing HRF for all values of p and θ. From Fig. 3, we observe that the HRF is non decreasing when x is increasing.

3.1 Quantile function

The quantile function of the Bin-PL distribution is obtained by

$$Q(u;p,\theta)=\frac{LambertW\left(-\xi(u-1)\exp(\xi)\right)}{\log(p)-\log(p+\theta)}-\frac{\theta^2+(p+2)\theta+p}{\theta}, \tag{3.3}$$

where

$$\xi=(\theta+1)(p+\theta)\log\left(\frac{p}{p+\theta}\right)$$

and

$$LambertW(z)=\sum_{n=1}^{\infty}\frac{(-n)^{n-1}z^n}{(n!)}.$$

From (3.3), the a-th quantile (x_a) of the Bin-PL distribution is written by

$$x_a=\begin{cases}\lfloor Q(u;p,\theta)\rfloor+1, & \lfloor Q(u;p,\theta)\rfloor\neq Q(u;p,\theta),\\ \lfloor Q(u;p,\theta)\rfloor, \lfloor Q(u;p,\theta)\rfloor+1, & \lfloor Q(u;p,\theta)\rfloor=Q(u;p,\theta).\end{cases}$$

That is x_a satisfies $F(x_a^-;p,\theta)\leq p\leq F(x_a;p,\theta)$. The median of the Bin-PL distribution is also obtained by equating a to 0.5.

3.2 Stochastic ordering

In distribution theory, stochastic ordering is an important concept for evaluating the comparative behavior of random variables. If the distribution has the likelihood ratio order, then it has the stochastic order and hazard rate order. It is known that $X<_{lr}Y$ implies that $X<_{hr}Y$ which implies that $X<_{st}Y$, see Ramesh and Kirmani [16].

Theorem 3.1. *If $Y\sim$ Bin-PL(p,θ) and $X\sim NB(1,p/(\theta+p))$, then $X<_{lr}Y$ (which implies that $X<_{hr}Y$ which implies that $X<_{st}Y$).*

Proof. Let $p_Y(x)$ be the PMF of Y and $p_X(x)$ be the PMF of X given by (2.1). Then, the ratio function is given by

$$W(x)=\frac{p_Y(x)}{p_X(x)}=\frac{\left[\theta^2/(1+\theta)\right]\left[p^x/(\theta+p)^{x+2}\right](x+\theta+p+1)}{\left[\theta/(\theta+p)\right]\left[p/(\theta+p)\right]^x}=\frac{\theta(x+\theta+p+1)}{(\theta+1)(\theta+p)}.$$

Since $W(x)\leq W(x+1)$ for all $\theta>0$ and $p\in(0,1)$. The proof is complete. □

3.3 On the moments

Now, let $X\sim$ Bin-PL(p,θ). By using the representation (2.2), the moment generating function of X is given by

$$M(t;p,\theta)=E\left(e^{tX}\right)=\frac{\theta^2}{(1+\theta)}\left(\frac{1}{(\theta+p-pe^t)^2}+\frac{1}{\theta+p-pe^t}\right),\quad t<\log\left(1+\frac{\theta}{p}\right).$$

The mean and variance of X are, respectively, given by

$$\mu = E(X) = \frac{p(2+\theta)}{\theta(1+\theta)}, \quad \sigma^2 = V(X) = \frac{p\left(4p\theta+2p+p\theta^2+3\theta^2+2\theta+\theta^3\right)}{\theta^2(1+\theta)^2}.$$

The coefficient of variation is given by

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{4p\theta+2p+p\theta^2+3\theta^2+2\theta+\theta^3}}{(\theta+2)\sqrt{p}}. \tag{3.4}$$

Fig. 4 presents the plots of CV for different values of the parameters p and θ. From Fig. 4, we see that the coefficient of variation is increasing when p is fixed and θ is increasing. Also, it is decreasing when θ is fixed and p is increasing. Moreover, we observe that the Bin-PL distribution has a low-variance for $\theta < 1$ and high-variance for $\theta \geq 1$.

Figs. 5 and 6 present the plots of skewness and kurtosis of the Bin-PL distribution, respectively. From these figures, we observe that, when p and θ are increasing, the values of skewness and kurtosis are increasing. Furthermore, the Bin-PL distribution is rightly skewed and leptokurtic.

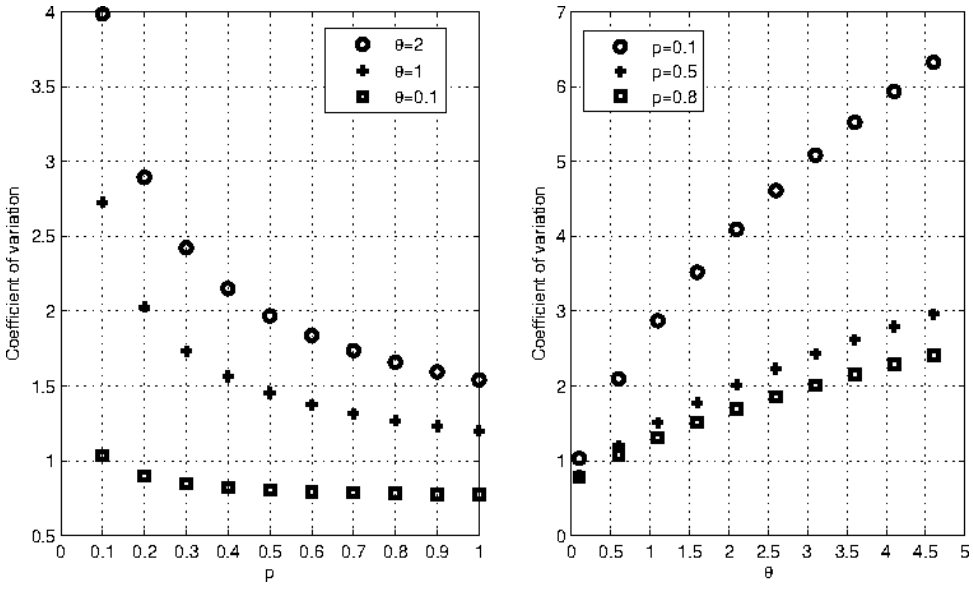

Figure 4: CV of the Bin-PL distribution for different values of p and θ.

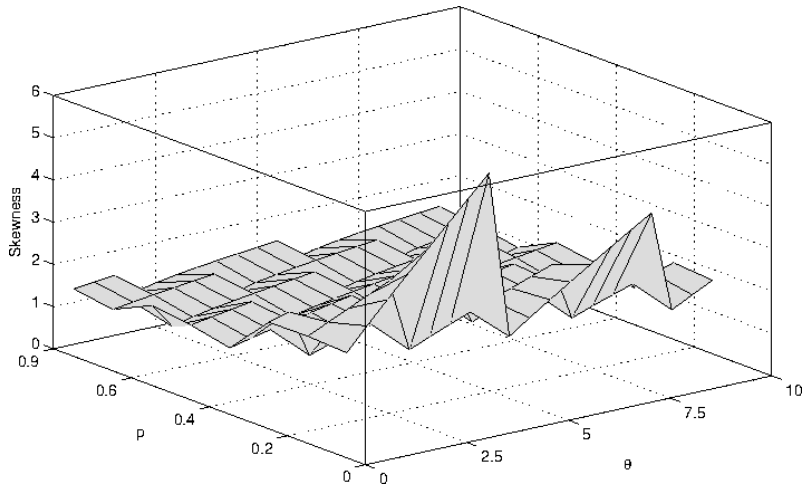

Figure 5: Skewness of the Bin-PL distribution for different values of p and θ.

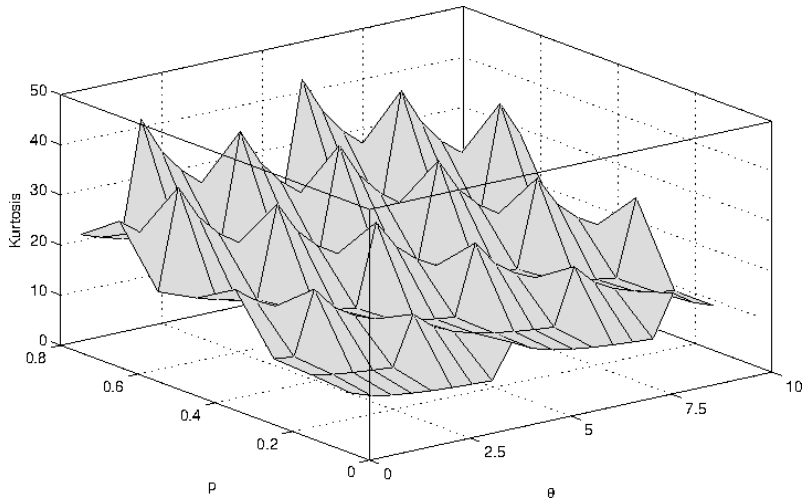

Figure 6: Kurtosis of the Bin-PL distribution for different values of p and θ.

4 Estimation methods with simulation

In this section, the performance of various strategies for estimating model parameters is examined using simulation studies.

4.1 Maximum likelihood method

Estimation of the Bin-PL model parameters, i.e., p and θ, is investigated by the maximum likelihood (ML) method. Let $x_1,\ldots,x_n$ be n (integer) values generated

from the Bin-PL distribution. Thus, $(x_1,\ldots,x_n)$ is a sample from the Bin-PL distribution. Then, based on Proposition 2.1, the likelihood and log-likelihood functions of the Bin-PL model are, respectively, given by

$$L(p,\theta)=\prod_{i=1}^{n} g(x_i;p,\theta)=\frac{\theta^{2n}}{(1+\theta)^n}\frac{p^{\sum_{i=1}^n x_i}}{(\theta+p)^{2n+\sum_{i=1}^n x_i}}\prod_{i=1}^{n}(x_i+\theta+p+1)$$

and

$$\ell(p,\theta)=\log[L(p,\theta)]=2n\log(\theta)-n\log(1+\theta)+\log(p)\sum_{i=1}^{n}x_i$$
$$-\log(\theta+p)\left(2n+\sum_{i=1}^{n}x_i\right)+\sum_{i=1}^{n}\log(x_i+\theta+p+1).$$

Then, the ML estimates of p and θ are defined by

$$\left(\hat{p},\hat{\theta}\right)=\operatorname{argmax}\left(\ell(p,\theta)\right)_{(p,\theta)\in(0,1)\times(0,\infty)}.$$

They also satisfied the following non-linear equations:

$$\frac{\partial\ell(\widehat{p},\widehat{\theta})}{\partial p}=0,\quad \frac{\partial\ell(\widehat{p},\widehat{\theta})}{\partial\theta}=0,$$

simultaneously, with

$$\frac{\ell(p,\theta)}{\partial p}=\frac{1}{p}\sum_{i=1}^{n}x_i-\frac{1}{\theta+p}\left(2n+\sum_{i=1}^{n}x_i\right)+\sum_{i=1}^{n}\frac{1}{x_i+\theta+p+1},$$
$$\frac{\ell(p,\theta)}{\partial\theta}=\frac{2n}{\theta}-\frac{n}{1+\theta}-\frac{1}{\theta+p}\left(2n+\sum_{i=1}^{n}x_i\right)+\sum_{i=1}^{n}\frac{1}{x_i+\theta+p+1}.$$

In particular, the MLEs are linked by the following simple expression:

$$\frac{1}{\widehat{p}}\sum_{i=1}^{n}x_i=\frac{2n}{\widehat{\theta}}-\frac{n}{1+\widehat{\theta}}.$$

Based on the existing theory of the ML method, under well established regularity conditions, the asymptotic distribution of each of the proposed random version of the estimators is the normal distribution, with the mean equal to the related unknown parameter, and with variance equal to the corresponding component of the inverse of the observed information matrix. With this asymptotic property, asymptotic confidence intervals and statistical tests can be constructed. See, for instance, [3].

4.2 Least and weighted least squares methods

Let $x_{(1)}<x_{(2)}<\cdots<x_{(n)}$ be the ordered observations from the Bin-PL distribution. Using the CDF given in (3.1), for $i=1,\ldots,n$, we have

$$F\big(x_{(i)};p,\theta\big)=1-\frac{\big(\theta^2+p\theta+2\theta+t\theta+p\big)\,p^{x_{(i)}+1}}{(1+\theta)\,(p+\theta)^{x_{(i)}+2}}. \tag{4.1}$$

The classical empirical CDF, denoted by $F^*(x_{(i)})$, can be used in order to estimate $F(x_{(i)};p,\theta)$. Substituting the empirical CDF in (4.1), we obtain the following model:

$$F^*\big(x_{(i)};p,\theta\big)=\left(1-\frac{\big(\theta^2+p\theta+2\theta+t\theta+p\big)\,p^{x_{(i)}+1}}{(1+\theta)\,(p+\theta)^{x_{(i)}+2}}\right)+\varepsilon_i,$$

where ε_i is the error term for i-th observation. Now, least squares error (LS) estimators of the parameters can be obtained by minimizing the following function with respect to p and θ:

$$L(p,\theta)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\Big(F^*(x_{(i)})-F(x_{(i)};p,\theta)\Big)^2.$$

Similarly, the weighted least squares (WLS) estimates of p and θ are obtained by minimizing the following function:

$$L_W(p,\theta)=\sum_{i=1}^{n}\frac{(n+2)\,(n+1)^2}{i(n-i+1)}\varepsilon_i^2=\sum_{i=1}^{n}\frac{(n+2)\,(n+1)^2}{i(n-i+1)}\Big(F^*(x_{(i)})-F(x_{(i)};p,\theta)\Big)^2.$$

All the minimization problems can be done via some numerical methods, such as the Nelder-mead or BFGS methods.

4.3 Method of proportions

The method of proportions (MP) was proposed by Khan *et al.* [12] to estimate the parameters of the discrete Weibull distribution. Here, we use the same method for estimating the parameters of the Bin-PL distribution. We define the indicator function by

$$v(x_i)=\begin{cases}1, & x_i=0,\\ 0, & x_i>0.\end{cases}$$

Then $y=(1/n)\sum_{i=1}^{n}v(x_i)$ denotes the proportion of 0's in the sample. It is clear that the random version of y is a consistent and an unbiased estimator of

the probability $f(0;p,\theta)=[\theta^2/(1+\theta)][(\theta+p+1)/(\theta+p)^2]$. Similarly, the proportion of 1's in the sample, say z, is suitable estimate of the probability $f(1;p,\theta)=[\theta^2/(1+\theta)][p(\theta+p+2)/(\theta+p)^3]$. Therefore, the proportion estimates of p and θ parameters, say $\hat{p}$ and $\hat{\theta}$, are obtained from the solution of the following equations:

$$\frac{\hat{\theta}^2}{1+\hat{\theta}}\frac{(\hat{\theta}+\hat{p}+1)}{(\hat{\theta}+\hat{p})^2}=y, \tag{4.2}$$

$$\frac{\hat{\theta}^2}{1+\hat{\theta}}\frac{\hat{p}(\hat{\theta}+\hat{p}+2)}{(\hat{\theta}+\hat{p})^3}=z. \tag{4.3}$$

Eqs. (4.2) and (4.3) can be solved numerically using Newton-Raphson method.

4.4 Method of moments

To estimate the parameters of the Bin-PL distribution by the method of moments (MM), we equate the first and second sample moments with their corresponding population moments to get

$$\frac{p(2+\theta)}{\theta(1+\theta)}=y \tag{4.4}$$

and

$$\frac{\left(\theta^2+2p\theta+2\theta+6p\right)p}{\theta^2(\theta+1)}=z, \tag{4.5}$$

where

$$y=\frac{1}{n}\sum_{i=1}^{n}x_i, \quad z=\frac{1}{n}\sum_{i=1}^{n}x_i^2$$

are the first and second sample moments, respectively. Therefore, by solving (4.4) and (4.5), we get

$$\hat{\theta}=\frac{2z-2y-4y^2+\sqrt{y^2(4y^2+2y-2z)}}{2y^2+y-z}, \quad \hat{p}=\frac{y\hat{\theta}\left(\hat{\theta}+1\right)}{2+\hat{\theta}}.$$

Those are estimates of θ and p, respectively.

4.5 Simulation study

To gain some information about the performance of the estimates obtained by the preceding estimation methods, we consider a set of simulation studies. In

the simulation study, 5000 trials were used to estimate the biases (Bias) and mean square errors (MSEs) of the ML, LS, WLS, MM and MP estimates. Different sample sizes are considered. Three parameter settings are considered. The results are given in Tables 1-3.

From Tables 1-3, it can be said that all estimates are asymptotically unbiased, when the sample size n increases, that is, the biases close to zero and the MSEs decrease to zero. Moreover, LS, MM, and WLS estimates of p and θ are better than the others in terms of bias and MSE in almost small sample sizes.

Table 1: Bias and MSE of the estimates for some sample size and parameters $p=0.2$ and $\theta=2$.

		Bias		MSE	
Methods	n	p	θ	p	θ
ML estimates	100	0.0162	-0.0264	0.0038	0.0641
	200	0.0029	0.0036	0.0026	0.0340
	400	0.0026	0.0028	0.0009	0.0266
	500	0.0022	0.0022	0.0009	0.0066
	750	0.0020	-0.0021	0.0008	0.0056
LS estimates	100	0.0010	-0.0007	0.0036	0.0001
	200	0.0010	-0.0005	0.0018	0.0001
	400	0.0008	-0.0005	0.0009	0.0001
	500	0.0003	0.0000	0.0007	0.0000
	750	-0.0002	-0.0000	0.0005	0.0000
WLS estimates	100	0.0010	-0.0007	0.0036	0.0001
	200	0.0010	-0.0005	0.0018	0.0001
	400	0.0008	-0.0005	0.0009	0.0001
	500	0.0003	0.0000	0.0007	0.0000
	750	-0.0002	-0.0012	0.0005	0.0000
MP estimates	100	0.0023	-0.0008	0.0042	0.0001
	200	0.0017	-0.0005	0.0021	0.0000
	400	0.0013	-0.0003	0.0011	0.0000
	500	0.0006	-0.0002	0.0008	0.0000
	750	0.0001	-0.0001	0.0006	0.0000
MM estimates	100	-0.0026	0.0000	0.0036	0.0000
	200	-0.0009	-0.0000	0.0018	0.0000
	400	-0.0003	-0.0001	0.0009	0.0000
	500	-0.0003	-0.0000	0.0007	0.0000
	750	-0.0005	-0.0000	0.0005	0.0000

Table 2: Bias and MSE of the estimates for some sample size and parameters $p=0.7$ and $\theta=0.5$.

		Bias		MSE	
Methods	n	p	θ	p	θ
ML estimates	100	0.0017	0.0733	0.0232	0.0226
	200	0.0963	0.0482	0.0197	0.0188
	400	0.0498	0.0364	0.0173	0.0106
	500	0.0296	0.0321	0.0164	0.0081
	750	0.0255	0.0313	0.0138	0.0074
LS estimates	100	0.0151	0.0161	0.0493	0.0221
	200	-0.0117	0.0111	0.0212	0.0099
	400	-0.0106	-0.0050	0.0194	0.0080
	500	-0.0016	-0.0003	0.0176	0.0074
	750	0.0162	0.0102	0.0087	0.0065
WLS estimates	100	0.0151	0.0161	0.0493	0.0221
	200	-0.0017	0.0111	0.0212	0.0099
	400	-0.0106	-0.0050	0.0134	0.0080
	500	-0.0016	-0.0003	0.0176	0.0074
	750	0.0162	0.0102	0.0087	0.0065
MP estimates	100	-0.0233	-0.0119	0.0353	0.0188
	200	-0.0120	-0.0060	0.0195	0.0098
	400	-0.0070	-0.0037	0.0094	0.0046
	500	-0.0042	-0.0021	0.0074	0.0038
	750	-0.0030	-0.0014	0.0050	0.0025
MM estimates	100	-0.0322	0.0405	0.0066	0.0119
	200	0.0032	0.0042	0.0064	0.0021
	400	0.0038	0.0039	0.0048	0.0016
	500	0.0031	0.0030	0.0038	0.0013
	750	0.0032	0.0026	0.0031	0.0010

5 Fitting data examples

In this section, two practical data examples are carried out to show the applicability of the Bin-PL model compared to the other models. The Poisson-Lindley (PL) (Sankaran [17]), uniform-geometric (UG) (Akdogan *et al.* [1]), negative binomial, geometric and Poisson models are used to fit the two real-life data sets. In order to specify the best model, we calculate the Kolmogorov-Smirnov (KS),

Table 3: Bias and MSE of the estimates for some sample size and parameters $p=0.5$ and $\theta=4$.

		Bias		MSE	
Methods	n	p	θ	p	θ
ML estimates	100	0.0084	-0.0135	0.0184	0.0201
	200	-0.0065	-0.0079	0.0106	0.0194
	400	0.0037	-0.0018	0.0048	0.0152
	500	0.0034	-0.0016	0.0036	0.0067
	750	0.0025	-0.0015	0.0026	0.0014
LS estimates	100	0.0050	-0.0023	0.0212	0.0005
	200	0.0021	-0.0014	0.0103	0.0003
	400	0.0014	-0.0007	0.0052	0.0001
	500	0.0012	-0.0004	0.0042	0.0001
	750	-0.0002	-0.0004	0.0027	0.0001
WLS estimates	100	0.0050	-0.0023	0.0212	0.0005
	200	0.0021	-0.0014	0.0103	0.0003
	400	0.0014	-0.0007	0.0052	0.0001
	500	0.0012	-0.0004	0.0042	0.0001
	750	-0.0002	-0.0004	0.0027	0.0001
MP estimates	100	0.0085	-0.0044	0.0253	0.0007
	200	0.0037	-0.0020	0.0122	0.0003
	400	0.0022	-0.0011	0.0061	0.0001
	500	0.0018	-0.0009	0.0049	0.0001
	750	0.0002	-0.0004	0.0032	0.0001
MM estimates	100	-0.0052	-0.0003	0.0202	0.0004
	200	-0.0022	-0.0002	0.0100	0.0002
	400	-0.0011	-0.0001	0.0052	0.0001
	500	-0.0008	-0.0001	0.0041	0.0001
	750	-0.0007	0.0001	0.0027	0.0001

Anderson-Darling (AD) and Cramer von Mises (CVM) goodness-of-fit statistics and the related p-value for all models. The MM is used in practical data applications since it is observed better than the other methods in cases of bias and MSE in simulation studies. Computations of the MM are obtained by the optim routine and all goodness-of-fit statistics are calculated by the goftest routine in the software R.

Data set 1: The first data set consists of survival times in days of 72 guinea pigs and is given in Table 4. These data are taken from Bjerkedal *et al.* [2].

Table 4: Data set 1.

15 22 24 24 32 32 33 34 38 38 43 44 48 52 53 54 54 55 56
57 58 58 59 60 60 60 60 61 62 63 65 65 67 68 70 70 72 73 75
76 76 81 83 84 85 87 91 95 96 98 99 109 110 121 127 129 131
143 146 146 175 175 211 233 258 258 263 297 341 341 376.

Data set 2: The second data set is given in Table 6 and consists of the 2003 final examination marks of 48 slow space students in mathematics at the Indian Institute of Technology at Kanpur. The data set is taken from Gupta and Kundu [8].

According to Tables 5-7, the Bin-PL distribution is more appropriate for analyzing the considered data than the PL, UG, negative binomial, geometric and Poisson distributions. The empirical CDF and estimated CDF of the Bin-PL distribution are provided in Figs. 7 and 8. As a result, it is observed that the Bin-PL distribution provides a better fit for the considered data than the other models.

Table 5: Some results for data set 1.

	Bin-PL	Geometric	Poisson	NB	PL	UG
KS	0.1275	0.2159	0.5755	0.4115	0.1291	0.2865
AD	1.8774	4.6895	Inf	67.887	1.8812	8.2259
CVM	0.3371	0.8490	7.1848	4.2234	0.3366	1.6021
KS p-value	0.1885	0.0024	0.0000	0.0000	0.1810	0.0000
AD p-value	0.1095	0.0041	0.0000	0.0000	0.1070	0.0000
CVM p-value	0.1093	0.0054	0.0000	0.0000	0.1067	0.0000
$\widehat{p}_1$	0.7738	0.0099	99.8194	25.6075	0.0198	0.0046
$\widehat{p}_2$	0.0153			0.2093		

Table 6: Data set 2.

29 25 50 15 13 27 15 18 7 7 8 19 12 18 5 21 15 86 21 15 14 39 15 14
70 44 6 23 58 19 50 23 11 6 34 18 28 34 12 37 4 60 20 23 40 65 19 31

Table 7: Some results for data set 2.

	Bin-PL	Geometric	Poisson	NB	PL	UG
KS	0.1056	0.2222	0.3997	0.1483	0.1108	0.2768
AD	0.6438	3.2014	Inf	2.2753	0.7156	5.2776
CVM	0.0935	0.5706	2.4394	0.2616	0.1041	1.0079
KS p-value	0.6557	0.0174	0.0000	0.2415	0.5966	0.0012
AD p-value	0.6067	0.0218	0.0000	0.0654	0.5441	0.0021
CVM p-value	0.6195	0.0260	0.0000	0.1741	0.5667	0.0021
$\widehat{p}_1$	0.4999	0.0372	25.8958	4.8729	0.0745	0.0181
$\widehat{p}_2$	0.0379			0.1611		

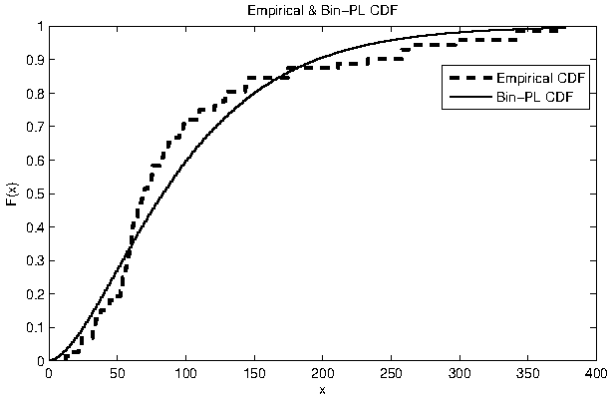

Figure 7: Empirical CDF and estimated CDF of the Bin-PL distribution based on data set 1.

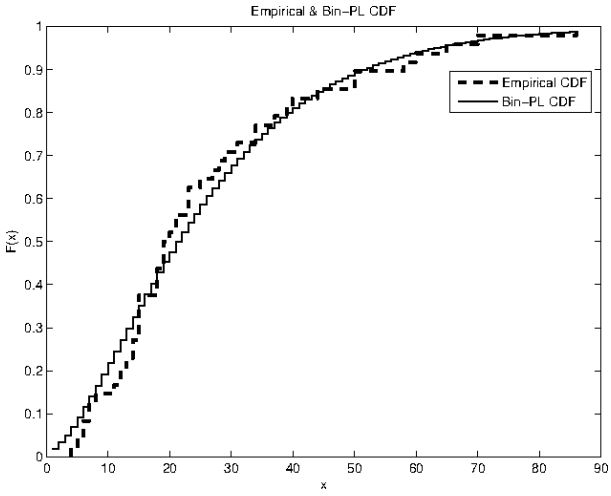

Figure 8: Empirical CDF and estimated CDF of the Bin-PL distribution based on data set 2.

6 Count regression analysis

In this section, we construct a count regression model based on the Bin-PL distribution, with applications.

6.1 Methodology

Let X be the response variable and y be associated $p\times 1$ vector of covariates. We consider that the response variable X follows the Bin-PL distribution with mean $\mu(y)$. Furthermore, the mean of response variable is linked with the explanatory variables by log linear form, i.e., $\mu_i=\exp(\beta\mathbf{y}_i^T)$, where $\beta=(\beta_1,\beta_1,\ldots,\beta_p)$ and $\mathbf{y}_i=(1,y_{1i},y_{2i},\ldots,y_{pi})$. By replacing θ with $(p-\mu+\sqrt{p^2+6\mu p+\mu^2})/(2\mu)$, we obtain the re-parameterize PMF as

$$p(x_i)=\frac{(p-\mu_i+\zeta_i)/(2\mu_i)^2}{1+(p-\mu_i+\zeta_i)/(2\mu_i)}\frac{p^{x_i}\left[(x_i+(p-\mu_i+\zeta_i)/(2\mu_i)+p+1)\right]}{\left((p-\mu_i+\zeta_i)/(2\mu_i)+p\right)^{x_i+2}}$$

and the corresponding log-likelihood equation is given as

$$\ell(p,\theta)=\sum_{i=1}^{n}\log\left\{\frac{\left((p-\exp\left(\beta\mathbf{y}_i^T\right)+\zeta_i)/(2\exp(\beta\mathbf{y}_i^T))\right)^2}{1+\left(p-\exp\left(\beta\mathbf{y}_i^T\right)+\zeta_i\right)/\left(2\exp(\beta\mathbf{y}_i^T)\right)}\right.$$
$$\left.\times\frac{p^{x_i}\left[x_i+(p-\exp(\beta\mathbf{y}_i^T)+\zeta_i)/(2\exp(\beta\mathbf{y}_i^T))+p+1\right]}{\left((p-\exp(\beta\mathbf{y}_i^T)+\zeta_i)/(2\exp(\beta\mathbf{y}_i^T))+p\right)^{x_i+2}}\right\},$$

where

$$\zeta_i=\sqrt{p^2+6\exp\left(\beta\mathbf{y}_i^T\right)p+\left(\exp(\beta\mathbf{y}_i^T)\right)^2}.$$

The above equations are not in closed form and can not be solved explicitly. Some numerical methods can be used to achieve solutions (see, Ma and Gui [15]). In the next, we give two examples of applications of Bin-PL count regression model by comparing it to some existing models, namely Poisson, Bell, and uniform-Poisson regression models.

6.2 Illustrative example 1

In this subsection, we show the Bin-PL regression application, taking into account the data obtained from Crawley [5]. The data set consists of one response and two explanatory variables. The number of infected blood cells (per mm2) belonging to

Table 8: Estimates from the Bin-PL regression and the other regression models.

Paremeter	Estimate (Bin-PL)	Estimate (Bell)	Estimate (Poisson)
β_1	0.4860	0.4946	0.5091
β_2	-1.1798	-1.1726	-1.1775
β_3	0.2225	0.2058	0.1846
$\hat{p}$	0.9985		
$\hat{\ell}$	-629.61	-632.02	-693.77
AIC	1267.22	1270.03	1393.53

an individual is dependent, smoking (yes: 0; no: 1) and gender (female: 0; male: 1) variables are explanatory variables. The demographic statistics of this data set and the results of the Bell regression model are given by (Lemonte *et al.* [14]). From Table 8, Bin-PL regression model outperforms better than the Poisson and Bell regression models on the basis of $\hat{\ell}$ and AIC.

To test the closeness of the Bin-PL regression model to other competitive models, H_0, null hypothesis against the H_1, alternative hypothesis must be tested. Hypotheses are given as follows:

$$H_0: E\big(\ell_{Bin-PL}(\hat{\Theta}_1)-\ell_{Poisson}(\hat{\Theta}_2 t)\big)=0,$$
$$H_1: E\big(\ell_{Bin-PL}(\hat{\Theta}_1)-\ell_{Poisson}(\hat{\Theta}_2)\big)\neq 0.$$

For testing the hypotheses, we use the likelihood ratio test proposed by Vuong [19], with test statistics given as

$$Z=\frac{1}{w\sqrt{n}}\big(\ell_{Bin-PL}(\hat{\Theta}_1)-\ell_{Poisson}(\hat{\Theta}_2)\big),$$

where

$$w^2=\frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\frac{p(x_i;\hat{\Theta}_1)}{g(x_i;\hat{\Theta}_2)}\right)\right]^2-\left[\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p(x_i;\hat{\Theta}_1)}{g(x_i;\hat{\Theta}_2)}\right)\right]^2$$

and $p(.;.)$ and $g(.;.)$ represent the PMF of Bin-PL and Poisson models, respectively. Furthermore, Z is asymptotically normally distributed. The Vuong test statistic Z value with corresponding p-values are given in Table 9. It is worth mentioning that the p-values are also calculated using the Vuong test statistic Z.

From Table 9, in both cases, the p-value is less than 0.05 (significance level at 5%). Hence, we can strongly conclude that the proposed Bin-PL regression model is preferred over the Poisson and Bell models.

Table 9: Vuong test results for Example 1.

	Z	$p-value$
Bin-PL & Poisson	6.3283	0.000
Bin-PL & Bell	2.6456	0.012

6.3 Illustrative example 2

In this subsection, we illustrate the Bin-PL regression application, taking into account the data obtained from Deb and Trivedi [6]. The data set consists of one response and ten explanatory variables for Bin-PL regression application. The number of stays after hospital admission (HOSP) is the response variable and the explanatory variables are: EXCLHLTH (Self-perceived health status excellent: 1, else: 0), POORHLTH (Self-perceived health status poor: 1, else: 0), NUMCHRON (Number of chronic conditions), AGE, MALE (Male: 1, else: 0), MARRIED (Married: 1, else: 0), FAMINC (Equals family income in $10,000), EMPLOYED (employed: 1, else: 0), PRIVINS (Private health: 1, else: 0), MEDICAID (Medicaid: 1, else: 0) variables are also explanatory variables. Deb and Trivedi [6] give details about the definition of these variables and the summary statistics. From Table 10,

Table 10: Estimates from the Bin-PL regression and the other regression models.

		Bin-PL	Uniform-Poisson	Poisson
Variable	Parameter	Estimate(S.E)	Estimate(S.E)	Estimate(S.E)
	β_1	-3.548(0.40)	-3.530(0.37)	-3.376(0.34)
EXCLHLTH	β_2	-0.714(0.18)	-0.725 (0.18)	-0.726(0.17)
POORHLTH	β_3	0.612(0.08)	0.627 (0.07)	0.618(0.06)
NUMCHRON	β_4	0.275(0.02)	0.274 (0.02)	0.263(0.02)
AGE	β_5	0.200(0.05)	0.197 (0.04)	0.178(0.04)
MALE	β_6	0.164(0.07)	0.154 (0.06)	0.131(0.06)
MARRIED	β_7	-0.049(0.07)	-0.043 (0.07)	-0.039(0.06)
FAMINC	β_8	0.005(0.01)	0.005 (0.01)	0.007(0.01)
EMPLOYED	β_9	0.023(0.12)	0.023 (0.11)	0.022(0.10)
PRIVINS	β_{10}	0.178(0.09)	0.200 (0.08)	0.197(0.07)
MEDICAID	β_{11}	0.220(0.12)	0.227 (0.11)	0.236(0.09)
	$\hat{p}$	0.800(0.26)		
$\ell_{\max}$		-2875.49	-2951.33	-3042.83
AIC		5774.98	5924.66	6107.66

Table 11: Vuong test results for Example 2.

	Z	$p-value$
Bin-PL & Poisson	7.3309	0.000
Bin-PL & Uniform-Poisson	3.9687	0.000

Bin-PL regression model outperforms well than the Poisson and UP regression models on the basis of $\hat{\ell}$ and AIC.

For testing the closeness of Bin-PL regression model with other competitive models, hypotheses and Vuong test procedures used in the first illustrative application are also applied for the second application. The Vuong test statistic value with corresponding p-values are given in Table 11. The p-values are calculated using the Vuong test statistic Z.

From Table 11, in both cases, the p-value is less than 0.05 (significance level at 5%). Hence, we can strongly conclude that the proposed Bin-PL regression model is preferred over the Poisson and UP models.

7 Concluding remarks

In this paper, a new discrete distribution with support $\mathbb{N}$ was introduced. Some distributional properties are obtained. Several estimators were studied to estimate the two model parameters. Extensive simulation studies for three different parameter settings were carried out. Fitting two practical data sets by this distribution is considered. A new count regression model has been introduced. The regression model is applied to two medical data sets and it is observed that our model is competitive in modeling practical data.

Acknowledgments

The authors would like to express their gratitude to the referees for their insightful remarks, which aided in the improvement of the work.

References

[1] Y. Akdogan, C. Kus, A. Asgharzadeh, I. Kinaci, and F. Sharafi, *Uniform-geometric distribution*, J. Stat. Comput. Simul. 86 (2016), 1754–1770.

[2] T. Bjerkedal, *Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli*, Am. J. Trop. Med. Hyg. 72 (1960), 130–148.
[3] G. Casella and R. L. Berger, *Statistical Inference*, Brooks/Cole, 1990.
[4] F. Castellares, S. L. Ferrari, and A. J. Lemonte, *On the Bell distribution and its associated regression model for count data*, Appl. Math. Model. 56 (2018), 172–185.
[5] M. J. Crawley, *The R Book*, John Wiley and Sons, 2012.
[6] P. Deb and P. K. Trivedi, *Demand for medical care by the elderly: A finite mixture approach*, J. Appl. Econ. 12 (1997), 313–336.
[7] E. G. Deniz, *A new discrete distribution: Properties and applications in medical care*, J. Appl. Stat. 40 (2013), 2760–2770.
[8] R. D. Gupta and D. Kundu, *A new class of weighted exponential distributions*, Statistics 43 (2009), 621–634.
[9] Y. Hu, X. Peng, T. Li, and H. Guo, *On the Poisson approximation to photon distribution for faintlasers*, Phys. lett., A 367 (2007), 173–176.
[10] N. L. Johnson, S. Kotz, and A. W. Kemp, *Univariate Discrete Distributions*, John Wiley and Sons, 2005.
[11] J. Keilson and H. Gerber, *Some results for discrete unimodality*, J. Amer. Statist. Assoc. 66 (1971), 286–389.
[12] M. S. A. Khan, A. Khalique, and A. M Abouammoh, *On estimating parameters in a discrete Weibull distribution*, IEEE Trans. Reliab. 38 (1989), 348–350.
[13] C. Kus, Y. Akdogan, A. Asgharzadeh, I. Kinaci, and K. Karakaya, *Binomial-discrete Lindley distribution*, Commun. Fac. Sci. University of Ankara Ser. A1-Math. Stat. 68 (2019), 401–411.
[14] A. J. Lemonte, G. Moreno-Arenas, and F. Castellares, *Zero-inflated Bell regression models for count data*, J. Appl. Stat. 47 (2020) 2, 265–286.
[15] Y. Ma and W. Gui, *Pivotal inference for the inverse Rayleigh distribution based on general progressively Type-II censored samples*, J. Appl. Stat. 46 (2019), 771–797.
[16] C. G. Ramesh and S.N. Kirmani, *On order relations between reliability measures*, Stoch. Models 3 (1987), 149–156.
[17] M. Sankaran, *The discrete Poisson Lindley distribution*, Biometrics 26 (1970), 145–149.
[18] R. Shanker and H. Fesshaye, *On Poisson-Lindley distribution and its applications to biological sciences*, Biom. Biostat. Int. J., 2 (2015), 103–107.
[19] Q. H. Vuong, *Likelihood ratio tests for model selection and non-nested hypotheses*, Econometrica 57 (1989), 307–333.