



A Lindley–binomial model for analyzing the proportions with sparseness and excessive zeros

Dianliang Deng  and Xiaoqing Zhang

Department of Mathematics and Statistics, University of Regina, Sask, Canada

ABSTRACT

Proportional data arise frequently in a wide variety of fields of study. Such data often exhibit extra variation such as over/under dispersion, sparseness and zero inflation. For example, the hepatitis data present both sparseness and zero inflation with 19 contributing non-zero denominators of 5 or less and with 36 having zero seropositive out of 83 annual age groups. The whitefly data consists of 640 observations with 339 zeros (53%), which demonstrates extra zero inflation. The catheter management data involve excessive zeros with over 60% zeros averagely for outcomes of 193 urinary tract infections, 194 outcomes of catheter blockages and 193 outcomes of catheter displacements. However, the existing models cannot always address such features appropriately. In this paper, a new two-parameter probability distribution called Lindley–binomial (LB) distribution is proposed to analyze the proportional data with such features. The probabilistic properties of the distribution such as moment, moment generating function are derived. The Fisher scoring algorithm and EM algorithm are presented for the computation of estimates of parameters in the proposed LB regression model. The issues on goodness of fit for the LB model are discussed. A limited simulation study is also performed to evaluate the performance of derived EM algorithms for the estimation of parameters in the model with/without covariates. The proposed model is illustrated through three aforementioned proportional datasets.

ARTICLE HISTORY

Received 6 December 2022

Accepted 3 July 2023


KEYWORDS

Proportional data; EM algorithm; Lindley distribution; binomial distribution; overdispersion; sparseness; zero inflation

1. Introduction

Discrete data in the form of proportions have been used to construe occurrences in many potential fields, including biology, clinical trials, engineering, insurance, public health, engineering, ecology, econometrics, etc. Generally, the binomial models are often used to analyze such kind of discrete data. However, in many practical situations, these data often exhibit extra-variation (over/under dispersion). Other issues arisen from such kind of data include the excessive zeros and sparse observations. When those issues are not properly addressed, the analysis using usual binomial models may not provide a good fit to the proportional data [13] and fail to explain the kinds of variation to the actual data.

CONTACT Dianliang Deng  deng@uregina.ca  Department of Mathematics and Statistics, University of Regina, Sask, Canada S4S 0A2

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/02664763.2023.2237212>.

Therefore, the statisticians have widely addressed the phenomenon of large variation in the proportional data, and the most popular model used for over-dispersed proportional data is the beta-binomial (BB) model, which was proposed originally by Williams [33]. Afterward, Crowder [5,6] analyzed the proportions using the beta-binomial ANOVA. Paul [25] applied the beta-binomial model to the analysis of the proportions of affected fetuses. Recently, Menssen and Schaarschmidt [21] obtained the prediction intervals for overdispersed binomial data. Najera-Zuloaga *et al.* [24] analyzed health-related quality of life data using the beta-binomial regression model approaches. Furthermore, Deng and Paul [7] proposed score tests for zero inflation and overdispersion based on the zero-inflated beta-binomial models. Luo and Paul [18] considered the estimation for zero-inflated beta-binomial regression model with missing response data. Ascari and Migliorati [2] proposed a new beta-binomial model for overdispersed binomial data with outliers and an excess of zeros. Generally, the BB model can account for overdispersion, allowing the binomial probability to vary in terms of a beta distribution. In particular, the BB model enriches (and encompasses) the binomial model with an additional precision/dispersion parameter, which admits an interesting interpretation in terms of intraclass correlation as well [26,27].

However, it is no model that works well for the various proportional data at all time. For example, the features such as sparseness and zero-inflation presented in proportional data cannot be appropriately addressed using the existing models. There are such features in the hepatitis, whitefly and catheter management data sets. The hepatitis data set was given by Keiding [15] and exhibits the sparseness with 19 out of 83 annual age groups contributing non-zero denominators of 5 or less out of 83 groups. Also, there are 36 zero seropositives out of 83 annual age groups in this data set. Therefore, this dataset indicates not only extreme sparseness but also excessive zeros. The whitefly data are obtained from the experiment that is about the efficacy of the pesticide on whiteflies. van Iersel *et al.* [32] studied the purpose of controlling silver leaf whiteflies by using a subirrigation system. They conducted this study to determine the effectiveness of controlling silver leaf whiteflies on poinsettia with imidacloprid, which was delivered by a subirrigation system. The number of surviving whiteflies combining with the total number of whiteflies in each experiment formed as the proportional data. Whitefly data set consists of 640 observations with 339 zeros (53%). Compared with other observations (each of the other observations has an average of 3.5%), obviously there exist excessive zeros in the whitefly data. The data on catheter management study are collected in the randomized clinical trial, in which the indwelling urinary catheter users were taught the awareness and self-monitoring skills. Each patient was asked up to six times about three binary outcomes and the total number of times asked varies from 1 to 6. Therefore, the outcomes can be considered as the times of urinary tract infections (UTIs), catheter blockages, and catheter displacements during total six asking times and thus three sets of proportional data with binomial denominator six (6) were obtained. Further, there are 83 (43.0%), 127 (65.5%), and 140 (72.5%) zeros in the 193 outcomes of urinary tract infections (UTIs), 194 outcomes of catheter blockages, and 193 outcomes of catheter displacements, respectively. Thus there are very high volumes of zeros in catheter management data set. The results for analyzing these three datasets using the binomial and zero-inflated binomial models have shown that the fits of these models to hepatitis, whitefly and catheter management data are not very appropriate. Therefore, instead of the existing models, an alternative model should be considered

to fit the proportional data with extra variation, which the existing models may not be able to handle appropriately. Such model can be obtained by compound binomial distribution with a nonnegative distribution. The target distribution is Lindley distribution, which was first introduced by Lindley [17]. This distribution is quite popular for modelling lifetime data and has a wide applicability in survival and reliability because of its closed forms for the survival and hazard functions and also its good flexibility of fit. Furthermore, many researchers have proposed and studied new classes of distributions which compound with a family of Lindley distributions. For example, Sankaran [28] proposed a compound Poisson distribution, which is known as the Poisson–Lindley distribution, by mixing the Poisson distribution with Lindley distribution. He gave some examples of real data sets that the Poisson–Lindley distribution provided a good fit. Zamani and Ismail [35] proposed the negative binomial–Lindley distribution, by mixing the distributions of negative binomial and Lindley, and found that this two-parameter negative binomial–Lindley distribution is particularly suitable in explaining count data with excess zeros, based on the application to accident and insurance claims data. Then Calderin–Ojeda and Gómez–Déniz [4] extended the negative binomial Lindley distribution from univariate to multivariate, which provides a tractable model with attractive properties that makes it suitable for application in any field where overdispersion is observed in count data. Bhati *et al.* [3] introduced a new generalized Poisson–Lindley distribution, which is compounding Poisson distribution with two-parameter Lindley distribution. They proved that this new distribution is a good alternative to Poisson distribution and Poisson–Lindley distribution for the right-tailed data set. Tajuddin *et al.* [30] proposed a four-parameter negative binomial–Lindley distribution to model over- and under-dispersed count data with excess zeros. There are many other applications with compound Lindley distributions, which indicate the Lindley distribution is definitely popular and useful.

Nevertheless, to our knowledge, there is no research that focuses on the model, which is derived by compounding the binomial model with a member of Lindley distribution family. The purpose of this paper is to propose such distribution for proportional data, which is called as Lindley–binomial (LB) distribution by compounding the binomial distribution with a Lindley distribution. The distribution that we actually used for compounding with binomial distribution is a two-parameter Lindley distribution, which is also used for compounding with other distributions by many researchers. Therefore, the proposed distribution in this paper is a new generalized two-parameter Lindley–binomial distribution and is a good alternative to BB distribution. This distribution allows to enrich the variance structure so as to account for multiple causes of overdispersion. The great variety of possible shapes of the LB distribution with right/left-tailed behaviors directly demonstrate the flexibility of the corresponding model and address the presence of outliers, sparseness as well as excessive zero observations without requiring ad hoc extra components accounting for them. This is possible because the two-parameter Lindley distribution is the mixture of two gamma distributions and thus the LB model dedicates one of its mixture components to a particular group of observations (e.g. zero values and/or outliers) automatically and provides interesting information about the possible sources of extra variation.

The remainder of this paper is organized as follows. In Section 2, we propose a Lindley binomial distribution, which is inspired by good statistical properties of Poisson Lindley distributions. We then derive the probabilistic properties such as probability mass function, mean, variance and moment generating function for Lindley binomial distribution. The

likelihood-based statistical inference about parameters of interest is studied in Section 3. Moreover, the Fisher scoring algorithm and EM algorithm are given to compute the estimates of parameters for the proposed Lindley binomial regression model. Meanwhile, Pearson chi-squared residuals and deviance residuals are presented to assess the goodness of fit for the proposed and existing models. In Section 4, simulation studies are performed to evaluate the performance of proposed EM algorithm for the computation of estimates of parameters in the proposed LB model with/without covariates. Hepatitis data, whitefly data and Catheter management study data are analyzed as an application of the proposed methodology in Section 5 with the concluding remarks in Section 6.

2. Lindley–binomial distribution and its properties

In this section, to define the Lindley binomial distribution, we briefly review the two-parameter Lindley distribution. As first given by Shanker [29], a random variable X follows a two-parameter Lindley distribution, denoted as $X \sim L_2(\alpha, \theta)$ if the probability density function $f(x; \alpha, \theta)$ of X has the following form:

$$f(x; \alpha, \theta) = \frac{\theta^2}{\theta + \alpha} (1 + \alpha x) e^{-\theta x}; \quad x > 0, \theta > 0 \quad \text{and} \quad \alpha + \theta > 0 \quad (1)$$

The pdf of this two-parameter Lindley distribution can be also shown as a mixture of exponential distribution (θ) (or gamma distribution $\Gamma(1; \theta)$) and gamma distribution $\Gamma(2; \theta)$ as follows:

$$f(x; \alpha, \theta) = \pi f_1(x; \theta) + (1 - \pi) f_2(x; \theta) \quad (2)$$

with a different mixture proportion $\pi = \frac{\theta}{\alpha + \theta}$, $f_1(x; \theta) = \theta e^{-\theta x}$ and $f_2(x; \theta) = \theta^2 x e^{-\theta x}$. It can easily be seen that at $\alpha = 1$, the two-parameter Lindley distribution reduces to the one parameter Lindley distribution, which was first proposed by Lindley [17]. Furthermore, this distribution reduces to exponential distribution with mean θ^{-1} and gamma distribution with mean $2\theta^{-1}$ for $\alpha = 0$ and $\alpha = \infty$, respectively. Therefore, for two-parameter Lindley distribution, the parameter α can account for the mixture proportions if this distribution is considered as the mixture of two gamma distributions.

2.1. The definition of Lindley–binomial distribution

Now in what follows, based on the binomial distribution and two-parameter Lindley distribution, we give the definition for Lindley–binomial distribution as follows.

Definition 2.1: A random variable Y is said to follow a two-parameter Lindley–binomial distribution if it obeys the following stochastic representation:

$$Y|\Lambda \sim \text{Binomial}(m, e^{-\Lambda})$$

and

$$\Lambda \sim L_2(\alpha, \theta)$$

where $y = 0, 1, \dots, m, \theta > 0$ and $\alpha + \theta > 0$. This two-parameter Lindley–binomial distribution will be represented as $LB_2(m, \alpha, \theta)$.

From Definition 2.1, Lindley–binomial distribution is induced by compounding the binomial distribution with two-parameter Lindley distribution and the probability mass function with corresponding properties are presented in Proposition 2.1. The derivation of Proposition 2.1 is given in Supplemental Material.

Proposition 2.1: *Let Y be a random variable which follows a two-parameter Lindley binomial distribution $LB_2(m, \alpha, \theta)$. Then the probability mass function of Y has the following form:*

$$P(Y = y) = \binom{m}{y} \frac{\theta^2}{\theta + \alpha} \sum_{k=0}^{m-y} \binom{m-y}{k} (-1)^k \frac{\theta + y + k + \alpha}{(\theta + y + k)^2},$$

where $y = 0, 1, 2, \dots, m; \theta > 0$ and $\theta + \alpha > 0$.

Note that for $\alpha = 0$ ($\pi = 1$), Lindley binomial distribution becomes a beta-binomial distribution $BB(m, \theta, 1)$ with the pmf as

$$P(Y = y) = \binom{m}{y} \frac{B(y + \theta, m - y + 1)}{B(\theta, 1)} \quad y = 0, 1, \dots, m$$

where $B(\beta_1, \beta_2)$ is the beta function defined as

$$B(\beta_1, \beta_2) = \int_0^1 x^{\beta_1-1} (1-x)^{\beta_2-1} dx.$$

In this sense, Lindley binomial distribution can be considered as the partial generalization of beta binomial distribution.

2.2. The probabilistic properties of Lindley–binomial distribution

In this section, we present the probabilistic properties of Lindley binomial distribution. At first, we may see the shapes of this probability mass functions for various values of parameters.

The probability mass function of $LB_2(m, \alpha, \theta)$ distribution with different values of parameters are given in Figure 1 and Figure S1 of the supplemental file. These figures are plotted to explore the effect of one parameter to the probability value given that other parameter is fixed. Based on Figure 1, the pmf of $LB_2(m, \alpha, \theta)$ distribution has the lowest mass at zero and the probability is significantly small at zero when θ is large such as $\theta = 10$ or 100 . Moreover, $LB_2(m, \alpha, \theta)$ distribution has the ability in fitting data with large frequency at the right endpoint. However, when θ is quite small, the pmf of $LB_2(m, \alpha, \theta)$ is right-tailed and has the highest mass at zero. Thus this proposed distribution is an alternative model to adequately fit the proportional data with the large frequency at left endpoint or at right endpoint. Also when the value of α is fixed, the maximum point of pmf changes from zero to the binomial denominator m as the value of θ changes from small to large. From Figure S1 in the supplemental file, when the value of θ is fixed, the shapes of pmf almost keep same even the value of α changes from small to large, which means θ is a shape parameter in $LB_2(\alpha, \theta)$ distribution.

Next, we discuss the probabilistic properties. We first give the following proposition.

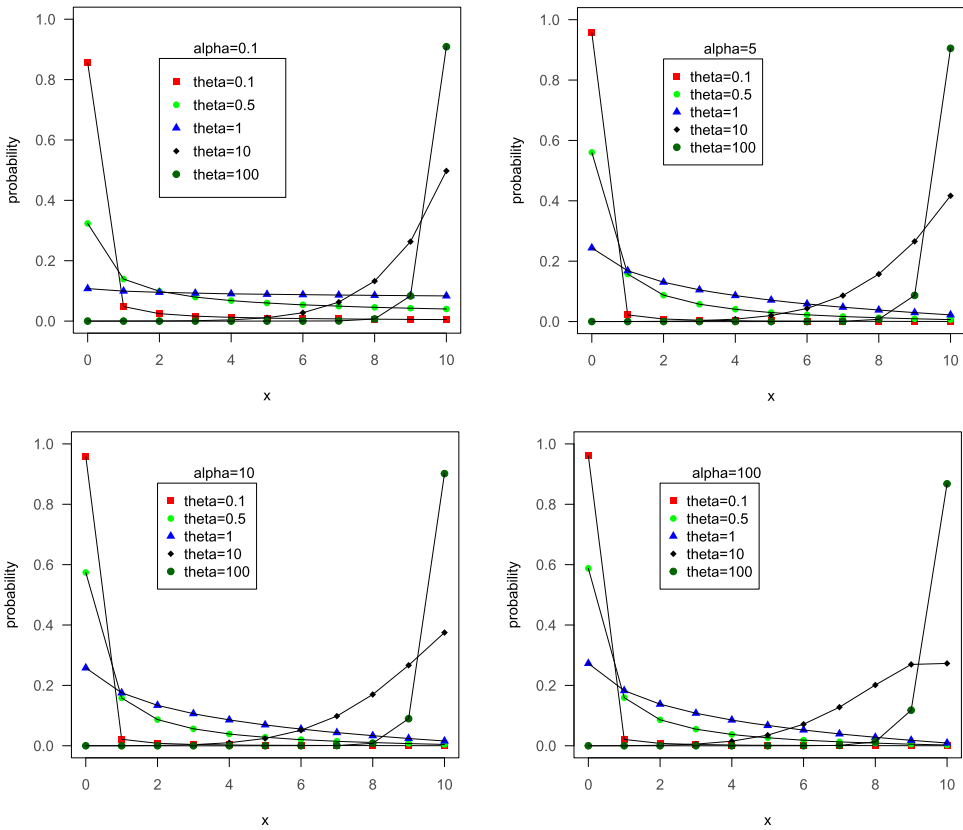


Figure 1. Pmf plots of LB_2 for different θ values with fixed α values.

Proposition 2.2: Let Y be a random variable which follows a two-parameter Lindley binomial distribution $LB_2(m, \alpha, \theta)$. Then the r th factorial moment of Y is given by

$$\mu_{(r)} \doteq E[Y(Y - 1) \cdots (Y - r + 1)] = \frac{m!}{(m - r)!} \frac{\theta^2(\theta + \alpha + r)}{(\theta + \alpha)(\theta + r)^2}.$$

Now from Proposition 2, we have the expectation and variance of Lindley binomial random variable Y as follows:

$$E(Y) = \frac{m\theta^2(\theta + \alpha + 1)}{(\theta + \alpha)(\theta + 1)^2}$$

and

$$\text{var}(Y) = \frac{m\theta^2(\theta + \alpha + 1)}{(\theta + \alpha)(\theta + 1)^2} \left(1 - \frac{m\theta^2(\theta + \alpha + 1)}{(\theta + \alpha)(\theta + 1)^2} \right) + m(m - 1) \frac{\theta^2(\theta + \alpha + 2)}{(\theta + \alpha)(\theta + 2)^2}.$$

Further, the moment generating function of Y can be derived using the law of total expectation:

$$M_Y(t) = \sum_{k=0}^m \binom{m}{k} \frac{\theta^2(\theta + \alpha + k)}{(\theta + \alpha)(\theta + k)^2} (e^t - 1)^k$$

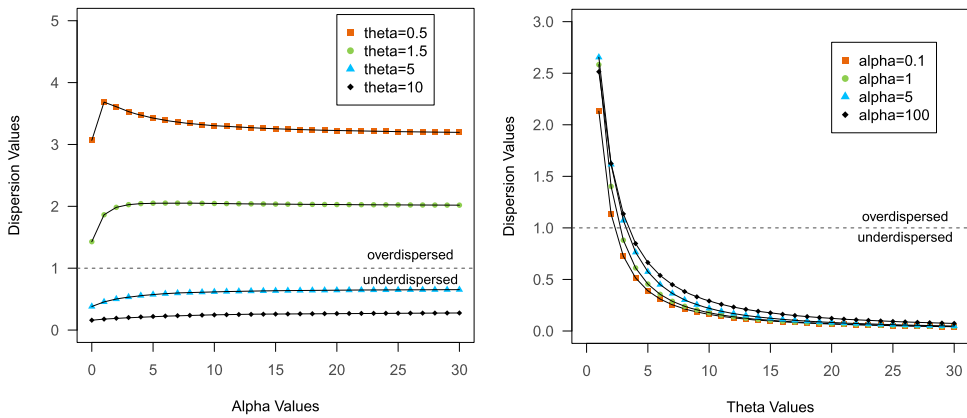


Figure 2. The plots for the index of dispersion Δ with different values of α and θ .

and the probability generating function and characteristic function of Y can be obtained in the same way.

Since the proposed Lindley binomial distribution will be used to fit the extra dispersed proportional data, the index of dispersion for this distribution should be addressed. The index of dispersion, which is also called variance-to-mean ratio, is a very useful tool to indicate whether a set of observations is clustered or dispersed compared to a standard statistical model. The index of dispersion for the $LB_2(m, \alpha, \theta)$ distribution comes out to be

$$\Delta = \frac{\text{var}(Y)}{E(Y)} = 1 + (m - 1) \frac{(\theta + \alpha + 2)(\theta + 1)^2}{(\theta + \alpha + 1)(\theta + 2)^2} - m \frac{\theta^2(\theta + \alpha + 1)}{(\theta + \alpha)(\theta + 1)^2} \quad (3)$$

The index of dispersion Δ given in (3) cannot be directly used for finding out whether LB_2 can adequately fit under-dispersed data, over-dispersed data or both of them. Thus, values of Δ are plotted as a function of parameters α and θ in Figure 2 to show the ability of the LB_2 distribution in fitting the data with either over-dispersion or under-dispersion. Based on the plot in the left panel of Figure 2, it is obvious that the index of dispersion for the LB_2 distribution can be either greater than one or less than one depending on the choice of the parameters. The plot in the right panel of Figure 2 indicates that as the value of θ increases, the value of Δ decreases and approaches to zero. Therefore, from these plots one could conclude that the LB_2 distribution can adequately fit over-dispersed data or under-dispersed data by choosing different values of parameters θ and α .

3. Likelihood based inferences for Lindley–binomial regression model

In Section 2, Lindley binomial distribution is defined based on the two-parameter Lindley distribution $L_2(\alpha, \theta)$ given in (1). However, the probability mass function has a little complicated form and thus results in the complexity of statistical inference for Lindley binomial model. To simplify the procedure of inferences, by reparameterizing to Lindley binomial distribution based on the mixture model (2) and setting $\pi = \theta/(\theta + \alpha)$, $\phi = 1/\theta$, we can have a following expression for the probability mass function of Lindley–binomial random

variable Y :

$$p(y; m, \pi, \phi) \triangleq P(Y = y) = \binom{m}{y} \sum_{k=0}^{m-y} \binom{m-y}{k} (-1)^k \frac{1 + (y+k)\pi\phi}{[1 + (y+k)\phi]^2} \tag{4}$$

where $y = 0, 1, 2, \dots, m; 0 < \pi < 1$ and $\phi > 0$. For convenience, we denote above form of Lindley binomial distribution as $LB(m, \pi, \phi)$. Furthermore, in terms of new parameters π and ϕ , we have the expressions for the factorial moment, mean, variance and moment generating function for Lindley binomial random variable Y as follows:

$$\begin{aligned} \mu_{(r)} &= \frac{m}{(m-r)} \frac{1+r\pi\phi}{(1+r\phi)^2}, \quad E(Y) = m \frac{1+\pi\phi}{(1+\phi)^2} \\ \text{var}(Y) &= m \frac{1+\pi\phi}{(1+\phi)^2} \left(1 - \frac{1+\pi\phi}{(1+\phi)^2} \right) + m(m-1) \left[\frac{1+2\pi\phi}{(1+2\phi)^2} - \frac{(1+\pi\phi)^2}{(1+\phi)^4} \right] \end{aligned}$$

and

$$M_Y(t) = \sum_{k=0}^m \binom{m}{k} \frac{1+k\pi\phi}{(1+k\phi)^2} (e^t - 1)^k.$$

Now, let Y_1, \dots, Y_n be random variables and $Y_i, i = 1, \dots, n$ follow Lindley binomial distribution $LB(m_i, \pi_i, \phi_i)$, where $m_i, i = 1, \dots, n$ are known binomial denominators, $(\pi_i, \phi_i), i = 1, \dots, n$ are unknown parameters. Suppose that y_i is the realization of random variable Y_i , then the observed data and associated binomial denominators would be represented by $\mathbf{y}_{\text{obs}} = \{y_1, \dots, y_n\}$ and $\mathbf{m}_{\text{obs}} = \{m_1, \dots, m_n\}$. Based on the probability mass function of Y given by (4), the likelihood function for the parameters $(\boldsymbol{\pi}, \boldsymbol{\phi}) = (\pi_1, \dots, \pi_n, \phi_1, \dots, \phi_n)$ can be obtained as

$$L(\boldsymbol{\pi}, \boldsymbol{\phi} | \mathbf{y}_{\text{obs}}, \mathbf{m}_{\text{obs}}) = \prod_{i=1}^n \left[\binom{m_i}{y_i} \sum_{j=0}^{m_i-y_i} \binom{m_i-y_i}{j} (-1)^j \frac{1+(y_i+j)\pi_i\phi_i}{[1+(y_i+j)\phi_i]^2} \right]$$

and thus the log-likelihood function is

$$\ell(\boldsymbol{\pi}, \boldsymbol{\phi} | \mathbf{y}_{\text{obs}}, \mathbf{m}_{\text{obs}}) = \sum_{i=1}^n \left\{ \ln \binom{m_i}{y_i} + \ln \left[\sum_{j=0}^{m_i-y_i} \binom{m_i-y_i}{j} (-1)^j \frac{1+(y_i+j)\pi_i\phi_i}{[1+(y_i+j)\phi_i]^2} \right] \right\} \tag{5}$$

3.1. MLEs of parameters for LB regression model

Based on the discussion above, we now derive the maximum likelihood estimates of parameters for LB regression model.

3.1.1. The formulation of LB regression model

Let Y_1, \dots, Y_n be independent random variables from the Lindley binomial distribution and Y_i follows the Lindley binomial distribution $LB(m_i, \pi_i, \phi_i)$, where for $i = 1, \dots, n, m_i$ are the known binomial denominators, π_i and ϕ_i are the unknown parameters. Further,

let \mathbf{w}_i and \mathbf{x}_i be the covariates associated with the proportional parameters π_i and scale parameter ϕ_i , respectively. Now suppose y_i is the realization of the random variable Y_i , then the observed data and associated binomial denominators would be represented by $\mathbf{y}_{\text{obs}} = \{y_1, \dots, y_n\}$ and $\mathbf{m}_{\text{obs}} = \{m_1, \dots, m_n\}$. To investigate the relationship between the parameters in LB model and covariates, the following regression model can be used to establish the association of parameters π_i and ϕ_i with covariates \mathbf{w}_i and \mathbf{x}_i ($i = 1, 2, \dots, n$):

$$\begin{cases} y_i \sim \text{LB}(m_i, \pi_i, \phi_i), & i = 1, 2, \dots, n \\ \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{w}_i^\top \boldsymbol{\alpha} \\ \log \phi_i = \mathbf{x}_i^\top \boldsymbol{\beta} \end{cases} \quad (6)$$

where $\mathbf{w}_i = (1, w_{1i}, \dots, w_{pi})^\top$ and $\mathbf{x}_i = (1, x_{1i}, \dots, x_{qi})^\top$ are not necessarily identical covariate vectors associated with the subject i ; ($i = 1, \dots, n$), and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_p)^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_q)^\top$ are the vectors of the regression coefficients associated with π_i and ϕ_i ($i = 1, 2, \dots, n$), respectively. Therefore, based on (5) and (6), the log-likelihood function for the regression coefficients $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ has the following form:

$$\begin{aligned} \ell(\boldsymbol{\theta} | \mathbf{y}_{\text{obs}}, \mathbf{m}_{\text{obs}}) &= \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y}_{\text{obs}}, \mathbf{m}_{\text{obs}}) = \sum_{i=1}^n \left\{ \ln \binom{m_i}{y_i} \right. \\ &\quad \left. + \ln \left[\sum_{j=0}^{m_i - y_i} \binom{m_i - y_i}{j} (-1)^j \frac{1 + \exp\{\mathbf{w}_i^\top \boldsymbol{\alpha}\} + (y_i + j) \exp\{\mathbf{w}_i^\top \boldsymbol{\alpha} + \mathbf{x}_i^\top \boldsymbol{\beta}\}}{(1 + \exp\{\mathbf{w}_i^\top \boldsymbol{\alpha}\}) [1 + (y_i + j) \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}]^2} \right] \right\} \quad (7) \end{aligned}$$

The primary purpose of the following sections is to estimate the parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

3.1.2. MLEs of parameters via Fisher scoring algorithm

In this section, the Fisher scoring algorithm is derived to calculate the MLEs of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Now, based on Equation (7), the first partial derivatives of log-likelihood with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y}_{\text{obs}}, \mathbf{m}_{\text{obs}})}{\partial \boldsymbol{\alpha}^\top} &= \sum_{i=1}^n \left[\frac{\sum_{j=0}^{m_i - y_i} \binom{m_i - y_i}{j} (-1)^j \frac{(y_i + j) \exp\{\mathbf{w}_i^\top \boldsymbol{\alpha} + \mathbf{x}_i^\top \boldsymbol{\beta}\}}{(1 + \exp\{\mathbf{w}_i^\top \boldsymbol{\alpha}\})^2 [1 + (y_i + j) \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}]} \right] \mathbf{w}_i^\top \quad (8) \end{aligned}$$

and

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y}_{\text{obs}}, \mathbf{m}_{\text{obs}})}{\partial \boldsymbol{\beta}^\top} &= \sum_{i=1}^n \left[\frac{\sum_{j=0}^{m_i - y_i} \binom{m_i - y_i}{j} (-1)^j \frac{-(y_i + j) \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\} (\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\} + 2 + (y_i + j) \exp\{\mathbf{w}_i^\top \boldsymbol{\alpha} + \mathbf{x}_i^\top \boldsymbol{\beta}\})}{(1 + \exp\{\mathbf{w}_i^\top \boldsymbol{\alpha}\}) [1 + (y_i + j) \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}]^3} \right] \mathbf{x}_i^\top \quad (9) \end{aligned}$$

The maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ are the solutions of Equations (8) and (9) equaling to zero.

However, there are no closed forms for these estimators and these non-linear equations do not seem to be solved directly. Their computations should be performed numerically using nonlinear optimization algorithms. Generally, Fisher scoring algorithm is a commonly used method to calculate maximum likelihood estimation and has a good stability even in multiparameter cases. The expected Fisher information matrix should always be positively definite, when the model is not over-parameterized Lauritzen [16]. Therefore, in what follows, we discuss Fisher scoring algorithm for computing the MLEs of parameters. The Newton–Raphson algorithm can be derived in similar way. To apply Fisher scoring algorithm, the Hessian matrix should be obtained first as follows:

$$\nabla^2 \ell(\alpha, \beta | y_{\text{obs}}, m_{\text{obs}}) = \begin{pmatrix} \frac{\partial^2 \ell(\alpha, \beta | y_{\text{obs}}, m_{\text{obs}})}{\partial \alpha^\top \partial \alpha} & \frac{\partial^2 \ell(\alpha, \beta | y_{\text{obs}}, m_{\text{obs}})}{\partial \alpha^\top \partial \beta} \\ \frac{\partial^2 \ell(\alpha, \beta | y_{\text{obs}}, m_{\text{obs}})}{\partial \beta^\top \partial \alpha} & \frac{\partial^2 \ell(\alpha, \beta | y_{\text{obs}}, m_{\text{obs}})}{\partial \beta^\top \partial \beta} \end{pmatrix} \quad (10)$$

Then the Fisher information matrix $J(\alpha, \beta) = -E\nabla^2 \ell(\alpha, \beta | y_{\text{obs}}, m_{\text{obs}})$ is

$$J(\alpha, \beta) = J(\alpha, \beta) = \begin{pmatrix} J_{\alpha\alpha} & J_{\alpha\beta} \\ J_{\beta\alpha} & J_{\beta\beta} \end{pmatrix}$$

Now let $(\alpha^{(0)}, \beta^{(0)})$ be the initial values of the MLEs $(\hat{\alpha}, \hat{\beta})$. If $(\alpha^{(t)}, \beta^{(t)})$ denote the t th approximation of $(\hat{\alpha}, \hat{\beta})$, then the $(t + 1)$ th approximation can be computed by

$$\begin{pmatrix} \alpha^{(t+1)} \\ \beta^{(t+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(t)} \\ \beta^{(t)} \end{pmatrix} + \begin{pmatrix} J_{\alpha\alpha}^{(t)} & J_{\alpha\beta}^{(t)} \\ J_{\beta\alpha}^{(t)} & J_{\beta\beta}^{(t)} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial \ell(\alpha^{(t)}, \beta^{(t)} | y_{\text{obs}}, m_{\text{obs}})}{\partial \alpha} \\ \frac{\partial \ell(\alpha^{(t)}, \beta^{(t)} | y_{\text{obs}}, m_{\text{obs}})}{\partial \beta} \end{pmatrix}$$

and the MLEs of (α, β) could be $(\alpha^{(t_0)}, \beta^{(t_0)})$ as $\|\alpha^{(t_0)} - \alpha^{(t_0-1)}\| + \|\beta^{(t_0)} - \beta^{(t_0-1)}\|$ is less than a threshold value. It should be pointed out that Fisher information matrix in LB model is intractable. However, it could be replaced with the observed information matrix in above Fisher scoring algorithm.

3.1.3. MLEs of parameters via the EM algorithm embedded with Fisher scoring algorithms at each M-step

In this section, we will develop the EM algorithm embedded with Fisher scoring algorithms at each M-step for MLEs of parameters in the proposed LB regression model. Note that the Fisher scoring algorithm possesses quadratic convergence and is sensitive to initial values. When the initial value $(\alpha^{(0)}, \beta^{(0)})$ of Fisher scoring algorithm is sufficiently near $(\hat{\alpha}, \hat{\beta})$, it converges very fast. However when the chosen initial value of $(\alpha^{(0)}, \beta^{(0)})$ is far from the true value of (α, β) , it might not converge. Furthermore, as we mentioned before, Fisher scoring algorithm does not work because Fisher information matrix is intractable from the likelihood function for the observed sample in LB model. Therefore in terms of the mixture property of two-parameter Lindley distribution, the EM algorithm can be

developed to compute the estimates of parameters for Lindley binomial model. In fact, the EM algorithm for maximum likelihood estimation uses the data likelihood as the objective function for choosing parameters. Sometimes this algorithm may not work well in all cases and the penalized methods may be used to modify the objective function. The typical penalized methods include the ridge penalty, the Bayes-inspired penalty and the logistic regression penalty (Hoerl and Kennard [14]; Hastie, Tibshirani and Friedman [11]; Morris [23]; Moreno and Lele [22]). Since the ridge penalty term only includes no-intercept parameters in the model and the logistic regression penalty involves the absolute values of parameters, we select the Bayes-inspired penalty for the estimation of parameters in the proposed EM algorithm. The modified objective function has the following form:

$$\log L(\boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{y}_{\text{obs}}, \mathbf{m}_{\text{obs}}) - \frac{\tau}{2} (\boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \boldsymbol{\beta})$$

where τ is a tuning parameter trading off the likelihood and penalty terms.

Therefore we will derive the EM algorithm for computing the MLEs of parameters in the proposed Lindley binomial model with Bayes-inspired penalty. As we know, the EM algorithm is a popular tool for estimating maximum likelihood estimation in joint statistical models by iterating between E-step and M-step. The E-step represents the expectation of the log-likelihood. The M-step computes parameters maximizing the expected log-likelihood found on the E-step. Then the unobserved latent variable is determined by these estimated parameters in the next E-step.

To establish the EM algorithm for the computation of MLEs for the parameters in LB regression model, we first set up the stochastic representation for the two-parameter Lindley distribution. Based on the reparameterization for the LB model, the pdf of two-parameter Lindley random variable Λ has the form:

$$f_{\Lambda}(\lambda) = \pi \frac{1}{\phi} e^{-\lambda/\phi} + (1 - \pi) \frac{\lambda}{\phi^2} e^{-\lambda/\phi} = \pi f_1(\lambda; \phi) + (1 - \pi) f_2(\lambda; \phi) \quad (11)$$

where $f_1(\lambda; \phi)$ and $f_2(\lambda; \phi)$ are the pdfs of gamma(1, ϕ) random variable U and gamma(2, ϕ) random variable V , respectively. Based on (11), the random variable Λ can be stochastically represented as

$$\Lambda = U^Z V^{1-Z}.$$

where Z follows the Bernoulli distribution with success probability π , that is $P(Z = 1) = 1 - P(Z = 0) = \pi$. This latent variable Z specifies to which mixture component each observation belongs. Therefore for a given Λ , there is an associated latent variable Z and the distribution function of Λ can be rewritten as

$$f_{\Lambda}(\lambda; z, \phi) = [f_1(\lambda; \phi)]^z [f_2(\lambda; \phi)]^{1-z}.$$

Let $p(z; \pi)$ represent the probability mass function of Z . Then the joint probability function is

$$f_{\Lambda, Z}(\lambda, z, \pi, \phi) = f_{\Lambda}(\lambda; z, \phi) p(z; \pi) = \{\pi f_1(\lambda; \phi)\}^z \{(1 - \pi) f_2(\lambda; \phi)\}^{1-z} \quad (12)$$

From the above expression, the full conditional distribution of Z is given by

$$Z | \pi, \phi, \lambda \sim \text{Bernoulli}(\pi^*)$$

where from Evin *et al.* [9],

$$\pi^* = \frac{\pi f_1(\lambda; \phi)}{\pi f_1(\lambda; \phi) + (1 - \pi)f_2(\lambda; \phi)} = \frac{\pi \phi}{\pi \phi + (1 - \pi)\lambda}.$$

For observed sample y_i with $i = 1, 2, \dots, n$ from $LB(m_i, \pi_i, \phi_i)$ distribution, based on (12) we introduce independent latent variables Z_i and Λ_i :

$$Z_i \sim \text{Bernoulli}(\pi_i), \quad \Lambda_i \sim L_2\left(\frac{1 - \pi_i}{\pi_i \phi_i}, \frac{1}{\phi_i}\right) \quad (i = 1, \dots, n) \tag{13}$$

We denote the latent/missing data by $Y_{\text{mis}} = \{z_i, \lambda_i\}_{i=1}^n$ and the complete data by $Y_{\text{com}} = \{Y_{\text{obs}}, Y_{\text{mis}}\} = Y_{\text{mis}}$, where z_i, λ_i are the realizations of Z_i and Λ_i , respectively. Thus the complete-data likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\pi}, \boldsymbol{\phi} | Y_{\text{com}}) &= \prod_{i=1}^n f_{\Lambda_i}(z_i, \lambda_i; \phi_i) p(z_i; \pi_i) = \prod_{i=1}^n \{\pi f_1(\lambda_i; \phi_i)\}^{z_i} \{(1 - \pi_i)f_2(\lambda_i; \phi_i)\}^{1-z_i} \\ &= \prod_{i=1}^n \left[\pi_i \frac{1}{\phi_i} e^{-\lambda_i/\phi_i} \right]^{z_i} \left[(1 - \pi_i) \frac{\lambda_i}{\phi_i^2} e^{-\lambda_i/\phi_i} \right]^{1-z_i} \end{aligned}$$

and the complete-data log-likelihood function is proportional to

$$\ell(\boldsymbol{\pi}, \boldsymbol{\phi} | Y_{\text{com}}) \propto \sum_{i=1}^n \left[z_i \log \pi_i + (1 - z_i) \log(1 - \pi_i) + (z_i - 2) \log \phi_i - \frac{\lambda_i}{\phi_i} \right]. \tag{14}$$

Hence, based on the regression model (6) and (14), the complete-data log-likelihood for regression parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is proportional to

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta} | Y_{\text{com}}) \propto \sum_{i=1}^n \left[z_i \mathbf{w}_i^\top \boldsymbol{\alpha} - \log(1 + e^{\mathbf{w}_i^\top \boldsymbol{\beta}}) + (z_i - 2) \mathbf{x}_i \boldsymbol{\beta} - \frac{\lambda_i}{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \right] \tag{15}$$

and the log-likelihood function with penalty has the form as

$$\begin{aligned} &\ell_\tau(\boldsymbol{\alpha}, \boldsymbol{\beta} | Y_{\text{com}}) \\ &\propto \sum_{i=1}^n \left[z_i \mathbf{w}_i^\top \boldsymbol{\alpha} - \log(1 + e^{\mathbf{w}_i^\top \boldsymbol{\alpha}}) + (z_i - 2) \mathbf{x}_i \boldsymbol{\beta} - \frac{\lambda_i}{e^{\mathbf{x}_i^\top \boldsymbol{\beta}}} \right] - \frac{\tau}{2} (\boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \boldsymbol{\beta}^\top \boldsymbol{\beta}) \end{aligned} \tag{16}$$

Now, the first and negative second partial derivatives of the complete-data penalized log-likelihood function (16) are given by

$$\begin{aligned} \frac{\partial \ell_p(\boldsymbol{\alpha}, \boldsymbol{\beta} | Y_{\text{com}})}{\partial \boldsymbol{\alpha}} &= \mathbf{w}^\top (\mathbf{z} - \boldsymbol{\pi}) - \tau \boldsymbol{\alpha}, \\ \frac{\partial \ell_p(\boldsymbol{\alpha}, \boldsymbol{\beta} | Y_{\text{com}})}{\partial \boldsymbol{\beta}} &= \mathbf{x}^\top (\mathbf{z} - 2\mathbf{1} + \boldsymbol{\lambda}/\boldsymbol{\phi}) - \tau \boldsymbol{\beta}, \\ -E \left(\frac{\partial^2 \ell_p(\boldsymbol{\alpha}, \boldsymbol{\beta} | Y_{\text{com}})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top} \right) &= \mathbf{w}^\top \text{diag}[\boldsymbol{\pi} (1 - \boldsymbol{\pi})] \mathbf{w} + \tau I_{p+1} \triangleq \mathbf{J}_{\text{com}}^{(\tau)}(\boldsymbol{\alpha}), \\ -E \left(\frac{\partial^2 \ell_p(\boldsymbol{\alpha}, \boldsymbol{\beta} | Y_{\text{com}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right) &= \mathbf{x}^\top \text{diag}[\boldsymbol{\phi}^{-1} E(\boldsymbol{\lambda})] \mathbf{x} + \tau I_{q+1} \triangleq \mathbf{J}_{\text{com}}^{(\tau)}(\boldsymbol{\beta}) \end{aligned}$$

where $\mathbf{1} = (1, \dots, 1)^\top$, $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_n)^\top$, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$, $\mathbf{z} = (z_1, \dots, z_n)^\top$, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)^\top$, $\boldsymbol{\lambda}/\boldsymbol{\phi} = (\lambda_1/\phi_1, \dots, \lambda_n/\phi_n)^\top$, $\text{diag}[\boldsymbol{\pi}(1 - \boldsymbol{\pi})] = \text{diag}[\pi_1(1 - \pi_1), \dots, \pi_n(1 - \pi_n)]$, $\text{diag}[\boldsymbol{\phi}^{-1}E(\boldsymbol{\lambda})] = \text{diag}[\phi_1^{-1}E(\lambda_1), \dots, \phi_n^{-1}E(\lambda_n)]$, I_{p+1} and I_{q+1} are the $(p+1) \times (p+1)$ and $(q+1) \times (q+1)$ identity matrices, respectively. Note that $\mathbf{J}_{\text{com}}^{(\tau)}(\boldsymbol{\alpha})$ is actually the complete-data Fisher information matrix associated only with the parameter vector $\boldsymbol{\alpha}$ and the covariate matrix \mathbf{w} , since it depends on neither the observed responses nor the latent/missing data. However, $\mathbf{J}_{\text{com}}^{(\tau)}(\boldsymbol{\beta})$ is associated not only with the parameter vector $\boldsymbol{\beta}$ and the covariate matrix \mathbf{x} but also with the latent variables $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$.

Now, the M-step is to separately calculate the MLEs of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ via two Fisher scoring algorithms as follows:

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + [\mathbf{J}_{\text{com}}^{(\tau)}(\boldsymbol{\alpha}^{(t)})]^{-1}[\mathbf{w}^\top(\mathbf{z} - \boldsymbol{\pi}(\boldsymbol{\alpha}^{(t)})) - \boldsymbol{\tau}\boldsymbol{\alpha}^{(t)}] \quad (17)$$

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + [\mathbf{J}_{\text{com}}^{(\tau)}(\boldsymbol{\beta}^{(t)})]^{-1}[\mathbf{x}^\top(\mathbf{z} - \mathbf{2}\mathbf{1} - \boldsymbol{\lambda}/\boldsymbol{\phi}(\boldsymbol{\beta}^{(t)})) - \boldsymbol{\tau}\boldsymbol{\beta}^{(t)}]. \quad (18)$$

The E-step is to replace the latent variables $\mathbf{z}, \boldsymbol{\lambda}$ in (17) and (18) by their conditional expectations:

$$E(\mathbf{z}|\mathbf{y}_{\text{obs}}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = (c_Z(y_1, m_1, \pi_1(\boldsymbol{\alpha}), \phi_1(\boldsymbol{\beta})), \dots, c_Z(y_n, m_n, \pi_n(\boldsymbol{\alpha}), \phi_n(\boldsymbol{\beta})))^\top \quad (19)$$

and

$$E(\boldsymbol{\lambda}|\mathbf{y}_{\text{obs}}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = (c_\Lambda(y_1, m_1, \pi_1(\boldsymbol{\alpha}), \phi_1(\boldsymbol{\beta})), \dots, c_\Lambda(y_n, m_n, \pi_n(\boldsymbol{\alpha}), \phi_n(\boldsymbol{\beta})))^\top \quad (20)$$

where $c_Z(y_i, m_i, \pi_i, \phi_i)$ and $c_\Lambda(y_i, m_i, \pi_i, \phi_i)$ ($i = 1, 2, \dots, n$) have the following expressions:

$$c_Z(y_i, m_i, \pi_i, \phi_i) \triangleq E(Z_i|y_i, m_i, \pi_i, \phi_i) = \frac{\sum_{k=0}^{m_i-y_i} \binom{m_i-y_i}{k} (-1)^k \frac{\pi_i}{1+(y_i+k)\phi_i}}{\sum_{k=0}^{m_i-y_i} \binom{m_i-y_i}{k} (-1)^k \frac{1+(y_i+k)\pi_i\phi_i}{[1+(y_i+k)\phi_i]^2}} \quad (21)$$

and

$$c_\Lambda(y_i, m_i, \pi_i, \phi_i) \triangleq E(\Lambda_i|y_i, m_i, \pi_i, \phi_i) = \frac{\sum_{k=0}^{m_i-y_i} \binom{m_i-y_i}{k} (-1)^k \frac{2\phi_i - \pi_i\phi_i + (y_i+k)\pi_i\phi_i^2}{[1+(y_i+k)\phi_i]^3}}{\sum_{k=0}^{m_i-y_i} \binom{m_i-y_i}{k} (-1)^k \frac{1+(y_i+k)\pi_i\phi_i}{[1+(y_i+k)\phi_i]^2}} \quad (22)$$

where $\pi_i = \pi_i(\boldsymbol{\alpha}) = \exp(\mathbf{w}_i^\top \boldsymbol{\alpha}) / (1 + \exp(\mathbf{w}_i^\top \boldsymbol{\alpha}))$ and $\phi_i = \phi_i(\boldsymbol{\beta}) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$. The derivation of (21) and (22) is given in Supplemental Material.

Now, let $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are the estimates of the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$, respectively. Then the asymptotic covariance matrices for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ can be obtained as $\text{cov}(\hat{\boldsymbol{\alpha}}) = \mathbf{J}_{\text{com}}^{-1}(\hat{\boldsymbol{\alpha}})$, $\text{cov}(\hat{\boldsymbol{\beta}}) = \mathbf{J}_{\text{com}}^{-1}(\hat{\boldsymbol{\beta}})$ and thus the corresponding confidence intervals for the components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be constructed by using the Wald-type method. For the value of τ , the EM algorithm can be carried out for each generated data set in simulation and τ can be chosen via maximizing the likelihood for the simulated data.

Remark: For the LB model without covariate, the MLEs of parameters π and ϕ can be easily obtained from aforementioned Fisher scoring algorithm and EM algorithm based

on the reduced model

$$\begin{cases} y_i \sim \text{LB}(m_i, \pi_i, \phi_i), & i = 1, 2, \dots, n \\ \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \alpha_0 \\ \log \phi_i = \beta_0 \end{cases}$$

and setting $\hat{\pi} = e^{\hat{\alpha}_0} / (1 + e^{\hat{\alpha}_0})$ and $\hat{\phi} = e^{\hat{\beta}_0}$. The details for these algorithms are omitted.

3.2. The issues for hypothesis testing and goodness-of-fit testing in LB model

In this section, we first discuss the hypothesis testing for Lindley binomial model. Since Lindley binomial distribution is derived via compounding the binomial model with the two parameter Lindley binomial distribution, which is the mixture of gamma distribution $\Gamma(1, \theta)$ and gamma distribution $\Gamma(2, \theta)$, one may like to know if Lindley binomial model is obtained via compounding the binomial distribution with a single gamma distribution. Therefore, the hypotheses we are interested in are

$$H_0^{(1)} : \pi = 0 \quad \text{versus} \quad H_1^{(1)} : \pi \neq 0 \tag{23}$$

and

$$H_0^{(2)} : \pi = 1 \quad \text{versus} \quad H_1^{(2)} : \pi \neq 1. \tag{24}$$

The hypothesis (23) is to test if the LB model is derived via the binomial model compounding with the gamma distribution $\Gamma(2, \theta)$ and (24) is to test if the model follows the specific beta binomial model $\text{BetaBin}(m, \theta, 1)$. Now, based on the likelihood method of model (4), the LRT statistics for testing the hypotheses (23) and (24) have the following forms, respectively:

$$\begin{aligned} T_1 &= 2l(\hat{\pi}, \hat{\phi}) - 2l(0, \hat{\phi}_1) \\ T_2 &= 2l(\hat{\pi}, \hat{\phi}) - 2l(1, \hat{\phi}_2) \end{aligned}$$

where $(\hat{\pi}, \hat{\phi})$ are the unconstrained MLEs of (π, ϕ) , which can be obtained via the Fisher scoring algorithm or EM algorithm given in Sections 3.1.1 and 3.1.2 for the LB regression model with $\hat{\pi} = \exp(\mathbf{w}^\top \boldsymbol{\alpha}) / (1 + \exp(\mathbf{w}^\top \boldsymbol{\alpha}))$ and $\hat{\phi} = \exp(\mathbf{x}^\top \hat{\boldsymbol{\beta}})$. $\hat{\phi}_1$, and $\hat{\phi}_2$ are the MLEs of ϕ under the null hypotheses $H_0^{(1)}$ and $H_0^{(2)}$, respectively, which can be derived in analogous algorithms as the unconstrained MLEs of (π, ϕ) .

Under $H_0^{(1)}$ and $H_0^{(2)}$, the LRT statistics T_1 and T_2 approximately follow the chi-squared distribution with one degree of freedom and the corresponding p -value can be computed as

$$p_1 = Pr(T_1 > t_1 | H_0^{(1)}) = Pr\{\chi^2(1) > t_1\}$$

and

$$p_2 = Pr(T_2 > t_2 | H_0^{(2)}) = Pr\{\chi^2(1) > t_2\}$$

where t_1 and t_2 are the observed values of T_1 and T_2 , respectively. Further, the LRT method can also be used to test the general null hypothesis $H_0 : \pi = \pi_0$.

Table 1. Pearson residuals for commonly used models for proportional data.

Model	Pearson residual
$BIN(m, \pi)$	$r_i^p = \frac{y_i - m_i \hat{\pi}}{\sqrt{m_i \hat{\pi} (1 - \hat{\pi})}}$
$BB(m, \alpha, \beta)$	$r_i^p = \frac{y_i - m_i \hat{\pi}_{BB}}{\sqrt{m_i \hat{\pi}_{BB} (1 - \hat{\pi}^*) [1 + (m - 1) \hat{\rho}_{BB}]}}$, $\pi_{BB} = \frac{\alpha}{\alpha + \beta}$, $\rho_{BB} = \frac{1}{1 + \alpha + \beta}$
$ZIB(m, \omega, \pi)$	$r_i^p = \frac{y_i - m_i (1 - \hat{\omega}) \hat{\pi}}{\sqrt{m_i (1 - \hat{\omega}) [\hat{\pi} (1 - \hat{\pi}) + m_i^2 \hat{\omega} \hat{\pi}^2]}}$
$LB(m, \pi, \phi)$	$r_i^p = \frac{y_i - m_i \hat{\pi}_{LB}}{\sqrt{m_i \hat{\pi}_{LB} (1 - \hat{\pi}_{LB}) + m_i (m_i - 1) [\frac{\hat{\pi}_{LB}}{2} (1 + \hat{\rho}_{LB})^2 - \hat{\rho}_{LB}^2 - \hat{\pi}_{LB}^2]}}$, $\pi_{LB} = \frac{1 + \pi \phi}{(1 + \phi)^2}$, $\rho_{LB} = \frac{1}{(1 + 2\phi)}$

Next we consider to assess the goodness of fit (GOF) for the LB model. There are numerous methods for testing the GOF of a model for proportional data. The most standard tests are residual deviance and Pearson’s χ^2 -test. We first discuss the Pearson residual for assessing the GOF in the commonly used models with proportional data. These models are binomial $BIN(m, \pi)$, beta binomial $BB(m, \alpha, \beta)$, zero-inflated binomial $ZIB(m, \omega, \pi)$ and Lindley binomial $LB(m, \pi, \phi)$ models. The Pearson residual, defined as the raw residual normalized by the estimated standard deviation of the response variable, can be expressed as

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(y_i)}}$$

where $\hat{\mu}_i$ is the fitted value for y_i and $\hat{V}(y_i)$ is the estimated value of variance for y_i .

The following table presents the Pearson residuals for aforementioned proportional models.

Note that all considered models are assumed to involve no covariates. For the case that the covariates are involved in the models, the corresponding parameters would depend on the covariates. Further, based on Pearson residuals given in Table 1, Pearson chi-squared statistic for GOF test is defined as

$$X_1^2 = \sum_{i=1}^n r_i^p{}^2.$$

Under a correctly specified model, X_1^2 follows an approximate chi-square distribution χ_{n-p}^2 , where n is the sample size, and p is the number of parameters.

Next we consider the deviance residuals for the proportional models. The general deviance statistic, which is defined as twice the difference between the log-likelihood for the saturated and fitted models, can be expressed as follows:

$$D(y, \hat{\mu}) = 2 \sum_{i=1}^n \left\{ \log[p(y_i | \hat{\theta}_s)] - \log[p(y_i | \hat{\theta})] \right\},$$

where $p(y_i | \hat{\theta}_s)$ is the log-likelihood function for the saturated model and $\hat{\theta}_s$ is the parameter estimates for the saturated model, in which there are as many estimated parameters as

Table 2. Deviance residuals for commonly used models for proportional data.

Model	Deviance residual
$BIN(m, \pi)$	$r_i^D = \text{sign}(y_i - m_i \hat{\pi}) \left\{ 2y_i \log \frac{y_i}{m_i \hat{\pi}} + 2(m_i - y_i) \log \frac{m_i - y_i}{m_i(1 - \hat{\pi})} \right\}^{1/2}$
$BB(m, \alpha, \beta)$	$r_i^D = \text{sign}(y_i - m_i \hat{\pi}_{BB}) \left\{ \sum_{j=0}^{y_i-1} 2 \log \frac{y_i}{m_i(\hat{\pi}_{BB} + j\hat{\theta}_{BB})} + \sum_{j=0}^{m_i-y_i-1} 2 \log \frac{m_i - y_i}{m_i(1 - \hat{\pi}_{BB} + j\hat{\theta}_{BB})} + \sum_{j=1}^{m_i-1} 2 \log(1 + j\hat{\theta}_{BB}) \right\}^{1/2}$ $\pi_{BB} = \frac{\alpha}{\alpha + \beta}, \theta_{BB} = \frac{1}{\alpha + \beta}$
$ZLB(m, \omega, \pi)$	$r_i^D = \text{sign}(y_i - m_i(1 - \hat{\omega})\hat{\pi}) \left\{ 2 \log \frac{y_i}{(\hat{\omega} + (1 - \hat{\omega})(1 - \hat{\pi})^{m_i})} I(y_i = 0) + [2y_i \log \frac{y_i}{m_i \hat{\pi}} + 2(m_i - y_i) \log \frac{m_i - y_i}{m_i(1 - \hat{\pi})} - 2 \log(1 - \hat{\omega})] I(y_i > 0) \right\}^{1/2}$
$LB(m, \pi, \phi)$	$r_i^D = \text{sign}(y_i - m_i \hat{\pi}_{LB}) \left\{ 2y_i \log \frac{y_i}{m_i} + 2(m_i - y_i) \log \frac{m_i - y_i}{m_i} - 2 \log \left(\sum_{j=1}^{m_i-y_i} \binom{m_i - y_i}{j} (-1)^j \frac{1 + (y_i + j)\hat{\pi}\hat{\phi}}{(1 + (y_i + j)\hat{\phi})^2} \right) \right\}^{1/2}$ $\pi_{LB} = \frac{1 + \pi\phi}{(1 + \phi)^2}$

data points [1,19]. By definition, a saturated model leads to a perfect fit to the data and has the highest log-likelihood among all models. $\log[p(y_i|\hat{\theta})]$ represents the log-likelihood function of the fitted model and $\hat{\theta}$ denotes the parameter estimate for the fitted model. Deviance residual represents the contribution of individual observation to the deviance $D(y, \hat{\mu})$, which is defined as the signed square root of the corresponding component for $D(y, \hat{\mu})$ and can be written as

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i},$$

where $d_i = 2\{\log[p(y_i|\hat{\theta}_s)] - \log[p(y_i|\hat{\theta})]\}$. Also, we list the deviance residuals for the commonly used proportional models in the following table.

Same as the Pearson Chi-squared statistic, the deviance statistic $D^2 = \sum_{i=1}^n (r_i^D)^2$ follows an approximate chi-squared distribution χ_{n-p}^2 under the correctly specified models. Therefore the goodness of fit in the proposed LB model can be assessed using Pearson chi-squared statistic and deviance statistic.

Moreover, for the proportional data with equal number of binomial denominators, other chi-squared statistic can be used to test the goodness of fit for the aforementioned models. This statistic has the following form:

$$X_2^2 = \sum_{y=0}^m \frac{(O_y - E_y)^2}{E_y},$$

where m is the value of binomial denominator for each experiment; O_y is the number of observations with y successes in all experiments; E_y is the expected number with y successes from the fitted model, which can be calculated as $E_y = n\hat{p}_y$ where n is the total number of

observations and \hat{p}_y is the expected probability that y successes are observed in m Bernoulli trials. Furthermore, \hat{p}_y can be obtained as

$$\hat{p}_y = \hat{P}(Y = y) = p(y|\hat{\theta}), \quad y = 0, 1, \dots, m$$

and $p(y|\hat{\theta})(y = 0, 1, \dots, m)$ is the expected probability calculated from the fitted model. Moreover, X_2^2 follows the approximate chi-squared distribution χ_{m-p}^2 under the correctly specified models, where p is the number of parameters in the fitted model.

On the other hand, the likelihood ratio or maximum likelihood statistical significance test G is increasingly being used in situations where chi-squared test X_2^2 is previously recommended [20]. The formula for G has the following form:

$$G = 2 \sum_{y=0}^m O_y \log \left(\frac{O_y}{E_y} \right)$$

where O_y and E_y are the same as that in chi-squared X_2^2 . Note that this G -test statistic is twice Kullback–Leibler divergence of the theoretical distribution from the empirical distribution. We may call it as Kullback–Leibler divergence statistic for the goodness of fit test. Also, the test statistic G has the same approximate chi-squared distribution as X_2^2 .

For diagnosing the models with proportional data, we can compare the residuals and calculate the values of Pearson chi-squared statistic X_1^2 , deviance statistic D^2 and statistic X_2^2 , Kullback–Leibler divergence statistic G (with equal number of binomial denominators) among all proportional models and thus find the best fitted model. Meanwhile, Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) can serve as the diagnosis tools for the selection of proportional models.

4. Simulation study

In this section, we carry out a limited simulation study to evaluate the performance of the proposed statistical methods in Section 3 for the LB model. We first examine the accuracy of EM algorithm for computing the MLEs for different parameter settings in the proposed LB models without covariates via simulation studies. Then we investigate its accuracy for the computation of regression parameters in the LB model with covariates.

4.1. Accuracy of MLEs for LB model without covariates

To evaluate the accuracy of EM algorithm for computing the MLEs of parameters π and ϕ in the Lindley binomial model without covariates, we consider 12 scenarios with $\pi = 0.25, 0.5$; $\phi = 0.2, 0.4, 0.6, 0.8, 1.0, 1.2$ and the binomial denominator $m = 6, 12$. The sample size is chosen as $n = 50, 75, 100, 200, 300, 400, 500$.

First, the procedure for generating the random number $\{y_i\}_{i=1}^n \stackrel{\text{iid}}{\sim} \text{LB}(m, \pi, \phi)$ is given as follows:

- (a) Use LindleyR to generate $\lambda_1, \dots, \lambda_n \stackrel{\text{iid}}{\sim} L_2((1 - \pi)/\pi\phi, 1/\phi)$;

(b) Generate

$$y_1 \sim \text{Binomial}(m, \exp\{-\lambda_1\}), \dots, y_n \sim \text{Binomial}(m, \exp\{-\lambda_n\}).$$

Then, $y_i \sim \text{LB}(m, \pi, \phi)$ for $i = 1, 2, \dots, n$.

From each generated sample the MLEs of parameters π and ϕ are calculated via EM algorithm (17)–(20) and the corresponding standard errors are obtained by using the asymptotic variances $\text{var}(\hat{\pi}) = \hat{\pi}(1 - \hat{\pi})/(n + \tau\hat{\pi}(1 - \hat{\pi}))$ and $\text{var}(\hat{\phi}) = \hat{\phi}^2/(2n - n\hat{\pi} + \tau\hat{\phi}^2)$. Next with repeating times $G = 1000$ for the parameters π and ϕ , the 1000 samples are independently generated and the corresponding 1000 EM MLEs and 1000 standard errors for parameters π and ϕ_1 are obtained. Further, in Table 3 and Table S1 of the supplemental file, MLE is the average of the 1000 estimates via the EM algorithm (17)–(20); MSE is the average of 1000 standard errors. As seen in Table 3 and Table S1 of the supplemental file, in most of scenarios the biases are small and the MLEs are very close to the corresponding true values of parameters in most cases, although the biases of EM estimates for some scenarios are a little large with sample size $n = 50, 75$ and 100 . However, by comparing the MLE and MSE, there is no significantly difference between the true values and the estimated values of parameters, which demonstrates that the proposed EM algorithm has very good performance.

4.2. Accuracy of MLEs for LB model with covariates

In this section, we perform the limited simulation study to investigate the performance of proposed EM algorithm for the estimation of regression parameters. We consider the following model:

$$Y_i \sim \text{LB}(m_i, \pi, \phi_i) \quad \text{and} \quad \log \phi_i = X_i\beta$$

where the mixture parameter π is assumed to be fixed, β is the vector of regression parameters and X_i is the vector of covariates ($i = 1, 2, \dots, n$). In the simulation, the sample size $n = 100, 200, 300, 400, 500$, the values of parameter vector $(\pi, \beta_0, \beta_1, \beta_2)$ are chosen to be $(0.25, -0.7, 1.3, -1.0), (0.50, -0.7, 1.3, -1.0), (0.75, -0.7, 1.3, -1.0); (0.25, 0.4, 0.6, -1.1), (0.50, 0.4, 0.6, -1.1), (0.75, 0.4, 0.6, -1.1); (0.25, -0.5, -0.9, 1.2), (0.50, -0.5, -0.9, 1.2), (0.75, -0.5, -0.9, 1.2)$ here β_0 is the intercept parameter. Two covariate variables are generated from uniform $U[0, 1]$ and Bernoulli(0.5). The binomial denominators are randomly generated from the integers 5–12. From the simulation study, Tables 4, S2 and S3 of the supplemental file are obtained.

From Table 4, Tables S2 and S3 of the supplemental file, one can see that the simulation results for LB regression model are consistent with that for LB model without covariate. The estimates for the parameters in LB regression model obtained via the EM algorithm are accurate although the biases are a little large for the scenarios that $\pi = 0.75$ and $n = 50, 75$.

5. Real-data application

In this section, we apply the proposed Lindley binomial model to analyze three proportional data sets with extreme sparseness and excessive zeros

Table 3. The estimates of parameters in Lindley binomial model with $\pi = 0.50, \phi = 0.2, 0.4, 0.6, 0.8, 1.0, 1.2$.

$\pi = 0.50$		$m = 6$				$m = 12$			
ϕ	n	Bias($\hat{\pi}$)	MSE($\hat{\pi}$)	Bias($\hat{\phi}$)	MSE($\hat{\phi}$)	Bias($\hat{\pi}$)	MSE($\hat{\pi}$)	Bias($\hat{\phi}$)	MSE($\hat{\phi}$)
0.2	50	-0.0338	0.0395	0.0098	0.0254	-0.0156	0.0456	0.0105	0.0254
	75	-0.0130	0.0358	0.0126	0.0211	-0.0108	0.0434	0.0066	0.0202
	100	-0.0434	0.0369	0.0028	0.0170	-0.0433	0.0411	0.0004	0.0166
	200	-0.0242	0.0296	0.0032	0.0120	-0.0033	0.0316	0.0038	0.0120
	300	-0.0040	0.0252	0.0047	0.0099	-0.0122	0.0270	0.0005	0.0095
	400	0.0013	0.0225	0.0046	0.0085	-0.0030	0.0238	0.0019	0.0083
	500	0.0035	0.0207	0.0046	0.0076	0.0067	0.0213	0.0025	0.0075
0.4	50	0.0000	0.0448	0.0260	0.0520	-0.0143	0.0504	0.0160	0.0499
	75	0.0157	0.0390	0.0295	0.0428	0.0048	0.0447	0.0196	0.0411
	100	-0.0277	0.0412	0.0080	0.0340	-0.0189	0.0433	0.0061	0.0337
	200	-0.0142	0.0310	0.0076	0.0240	-0.0044	0.0325	0.0079	0.0239
	300	-0.0036	0.0263	0.0064	0.0195	0.0069	0.0273	0.0088	0.0195
	400	0.0056	0.0235	0.0083	0.0169	0.0050	0.0240	0.0054	0.0167
	500	-0.0007	0.0213	0.0040	0.0149	0.0120	0.0216	0.0063	0.0150
0.6	50	-0.0102	0.0449	0.0363	0.0774	0.0060	0.0507	0.0448	0.0781
	75	0.0228	0.0405	0.0537	0.0652	0.0183	0.0443	0.0327	0.0625
	100	-0.0298	0.0416	0.0058	0.0503	-0.0222	0.0441	0.0072	0.0502
	200	0.0010	0.0318	0.0182	0.0364	0.0048	0.0330	0.0143	0.0360
	300	-0.0102	0.0269	0.0044	0.0288	-0.0008	0.0275	0.0072	0.0289
	400	0.0022	0.0236	0.0112	0.0253	-0.0032	0.0240	0.0028	0.0248
	500	0.0048	0.0214	0.0079	0.0224	0.0037	0.0217	0.0054	0.0223
0.8	50	0.0092	0.0454	0.0776	0.1077	0.0162	0.0495	0.0621	0.1050
	75	0.0236	0.0405	0.0746	0.0874	0.0322	0.0445	0.0603	0.0853
	100	-0.0266	0.0438	0.0038	0.0663	-0.0171	0.0451	0.0082	0.0667
	200	-0.0061	0.0324	0.0128	0.0478	-0.0026	0.0330	0.0105	0.0474
	300	-0.0060	0.0272	0.0100	0.0386	0.0012	0.0274	0.0112	0.0387
	400	-0.0076	0.0238	0.0011	0.0329	-0.0003	0.0241	0.0059	0.0331
	500	-0.0035	0.0215	0.0082	0.0297	-0.0005	0.0217	0.0047	0.0296
1.0	50	0.0213	0.0442	0.1016	0.1361	0.0222	0.0500	0.1008	0.1345
	75	0.0277	0.0401	0.1052	0.1109	0.0138	0.0442	0.0722	0.1059
	100	-0.0529	0.0442	-0.0184	0.0801	-0.0285	0.0455	-0.0041	0.0818
	200	-0.0298	0.0326	-0.0057	0.0577	-0.0137	0.0332	0.0005	0.0582
	300	-0.0096	0.0273	0.0039	0.0477	0.0034	0.0275	0.0161	0.0484
	400	-0.0020	0.0237	0.0149	0.0419	-0.0007	0.0242	0.0058	0.0413
	500	-0.0068	0.0215	0.0050	0.0369	0.0021	0.0217	0.0056	0.0370
1.2	50	0.0371	0.0436	0.1550	0.1687	0.0089	0.0480	0.1096	0.1602
	75	0.0391	0.0398	0.1404	0.1352	0.0267	0.0435	0.1072	0.1300
	100	-0.0629	0.0436	-0.0324	0.0950	-0.0521	0.0453	-0.0346	0.0947
	200	-0.0190	0.0326	0.0031	0.0700	-0.0181	0.0335	-0.0048	0.0692
	300	-0.0059	0.0273	0.0073	0.0574	-0.0094	0.0277	0.0036	0.0570
	400	-0.0045	0.0239	0.0055	0.0496	-0.0083	0.0241	0.0047	0.0494
	500	-0.0027	0.0215	0.0095	0.0445	-0.0023	0.0217	0.0065	0.0443

5.1. Incidence of hepatitis A in Bulgaria

The data set used in this section is the incidence of hepatitis A in Bulgaria by age. This data set was given by Keiding [15] and exhibits the sparseness with 19 out of 83 annual age groups contributing non-zero denominators of 5 or less out of 83 groups. Farrington [10] presented an analysis of this data set fitting to generalized linear models. They used the number of seronegatives as response variable with binomial errors. For the illustration of our proposed LB model, we consider the number of seropositives as binomial response

Table 4. The estimates of parameters in Lindley binomial model with $\pi = 0.25, 0.50, 0.75$ and $\beta = (0.4, 0.6, -1.1)$.

π	n	Bias($\hat{\pi}$)	MSE($\hat{\pi}$)	Bias($\hat{\beta}_0$)	MSE($\hat{\beta}_0$)	Bias($\hat{\beta}_1$)	MSE($\hat{\beta}_1$)	Bias($\hat{\beta}_2$)	MSE($\hat{\beta}_2$)
0.25	50	0.0314	0.0380	0.0819	0.2506	-0.0078	0.3892	-0.0510	0.2235
	75	0.0382	0.0360	0.0663	0.2026	-0.0041	0.3142	-0.0203	0.1812
	100	-0.0095	0.0353	-0.0003	0.1701	-0.0009	0.2647	-0.0008	0.1527
	200	-0.0037	0.0281	0.0047	0.1201	-0.0078	0.1862	0.0011	0.1074
	300	-0.0019	0.0237	0.0008	0.0979	0.0023	0.1517	0.0016	0.0876
	400	0.0002	0.0209	0.0037	0.0847	0.0032	0.1312	-0.0051	0.0758
	500	-0.0023	0.0188	-0.0059	0.0756	0.0065	0.1172	0.0042	0.0677
0.50	50	-0.0277	0.0445	0.0018	0.2677	0.0158	0.4155	-0.0224	0.2386
	75	0.0057	0.0393	0.0221	0.2181	-0.0156	0.3394	-0.0208	0.1953
	100	-0.0531	0.0434	-0.0364	0.1798	-0.0195	0.2790	0.0243	0.1625
	200	-0.0017	0.0324	0.0054	0.1295	-0.0085	0.2007	0.0097	0.1164
	300	-0.0096	0.0273	-0.0008	0.1053	-0.0066	0.1631	0.0067	0.0945
	400	-0.0065	0.0238	-0.0035	0.0915	-0.0093	0.1416	0.0078	0.0819
	500	0.0045	0.0214	0.0097	0.0820	-0.0067	0.1270	-0.0015	0.0735
0.75	50	-0.1814	0.0441	-0.1397	0.2762	-0.0212	0.4307	0.0075	0.2469
	75	-0.1329	0.0389	-0.1101	0.2280	-0.0066	0.3540	0.0110	0.2030
	100	-0.0804	0.0348	-0.0652	0.1991	0.0061	0.3091	0.0014	0.1779
	200	-0.0344	0.0253	-0.0271	0.1419	-0.0076	0.2202	0.0095	0.1269
	300	-0.0147	0.0213	-0.0057	0.1162	0.0039	0.1800	-0.0037	0.1039
	400	-0.0093	0.0185	-0.0081	0.1006	0.0062	0.1561	0.0075	0.0901
	500	-0.0015	0.0166	0.0062	0.0902	-0.0037	0.1399	0.0046	0.0807

Table 5. Results of model fitting for the Hepatitis in Bulgaria dataset.

Model	Binomial	BB	LB	ZIB
Parameters	$\hat{p} = 0.2976$	$\hat{\alpha} = 0.3767,$ $\hat{\beta} = 1.3617$	$\hat{\pi} = 0.0384,$ $\hat{\phi} = 1.1375$	$\hat{\omega} = 0.3730$ $\hat{\pi} = 0.4009$
Log-likelihood	-240.4100	-155.8891	-154.9336	-191.8077
AIC	482.8200	315.7783	313.8672	387.6154
BIC	489.6577	320.6159	318.7048	392.4531
X^2_1	308.2910	78.4064	90.8661	105.9551
D^2	360.9004	191.8597	189.9486	263.6969

variable, for which, there are 36 zero seropositives out of 83 annual age groups. Therefore, this data set exhibits not only extreme sparseness but also excessive zeros. For the purpose of comparison, we perform the statistical analysis for this data set using the Lindley binomial (LB) model as well as the simple binomial model, beta-binomial (BB) model and zero-inflated binomial (ZIB) model and compare the proposed LB model with these existing proportional models via the GOF test statistics given in Section 3.2.

Table 5 presents the values of log-likelihood, AIC, BIC, Pearson chi-squared statistic X^2_1 and deviance statistic D^2 with the values of estimated parameters for the model fitting with the aforementioned four models.

The estimates of parameters in BB model are computed using ‘**bb.mle**’ package in R programming environment. However, it should be pointed out that the estimated values of parameters heavily depend on the initial values of parameters in ‘**bb.mle**’ package for beta binomial model. Based on the results in Table 5, although the Pearson chi-squares statistic for BB model is smallest among the four models, the LB model gives the largest log-likelihood -154.9336, and smallest AIC 313.8672, smallest BIC 318.7048 and smallest deviance statistic 189.9486. Therefore, the proposed Lindley–binomial model shows the

Table 6. Results for the analysis of Hepatitis data using LB and ZIB regression models.

Model	ZIB regression model	LB regression model
Estimates	$\hat{\omega} = 0.0556,$ $\hat{\gamma}_0 = 3.4493, \hat{\gamma}_1 = -1.3813$	$\hat{\pi} = 0.0000,$ $\hat{\beta}_0 = -3.5392, \hat{\beta}_1 = 1.0651$
log-likelihood	-122.0351	-118.6235
AIC	250.0702	243.5162
BIC	261.7456	254.9224
X_1^2	68.4555	30.2292
D^2	124.1516	117.3285

advantage for fitting the proportional data with sparseness and excessive zeros. Next, we consider assessing the mixture parameter π in LB model for this dataset. The likelihood ratio test (LRT) can be used to test if this parameter equals zero or one. For hepatitis data, the value of LRT is $2[\ell_{LB}(\hat{\pi}, \hat{\phi}) - \ell_{LB}(0, \tilde{\phi})] = 2(-154.9336 - (-155.4260)) = 0.9848$, which strongly support the null hypothesis $H_0 : \pi = 0$. This also shows that the data may come from the distribution which compounds the binomial with the single gamma(2, $1/\phi$) distribution. Furthermore, since the estimate of mixture parameter π in LB is 0.0384 and $\hat{\phi} = 1.1375$, the estimated values of corresponding original parameters are $\hat{\alpha} = (1 - \hat{\pi})/\hat{\pi}\hat{\phi} = 22.0166$ and $\hat{\theta} = 1/\hat{\phi} = 0.8791$. On the other hand, via LRT, we have $2[\ell_{ZIB}(\hat{\pi}, \hat{\phi}) - \ell_{BIN}(\hat{p})] = 2(-191.8077 - (-240.4100)) = 97.2046$, which strongly support the existence of zero-inflation in hepatitis data. From the above results, one can see the LB model can also be used to account for the zero-inflation in proportional data, which can be demonstrated in top two panels of Figure S1 in the supplemental file.

Now we further use the regression models to demonstrate the superiority of our proposed model. Since the data involve excessive zeros, we only compare the LB regression model with ZIB regression model. The LB regression model considered here has the form as

$$Y_i \sim \text{LB}(m_i, \pi, \phi_i) \quad \text{and} \quad \log \phi_i = \beta_0 + \log(\text{age}_i)\beta_1$$

and the ZIB regression model has the form as

$$Y_i \sim \text{ZIB}(m_i, \omega, \pi_i) \quad \text{and} \quad \log \frac{\pi_i}{1 - \pi_i} = \gamma_0 + \log(\text{age}_i)\gamma_1.$$

Here, the mixture parameter π in LB model and the zero-inflated parameter ω are assumed to be fixed for different subjects. The analysis of the data using these two regression models is given in Table 6, from which LB model obviously demonstrates the superiority to ZIB model.

Therefore based on the results given in Tables 5 and 6, we may conclude that the LB model has a better performance than binomial, BB and ZIB models for fitting hepatitis data.

5.2. Whitefly data

This whitefly data set is the result of the experiment that is about the efficacy of the pesticide on whiteflies given by [32]. In the whitefly data, the purpose of controlling silver leaf whiteflies was studied by using a subirrigation system. The study was designed to determine the

effectiveness of controlling silver leaf whiteflies on poinsettia with imidacloprid, which was delivered by a subirrigation system. Imidacloprid is a resilient and powerful chemical that has low toxicity to mammals and is used to control silver leaf whiteflies on poinsettia. At the first week of this experiment, researchers placed m adult whiteflies (here m is considered as the binomial denominators with range 6–15, mean = 9.5 and SD = 1.7) in clip-on leaf cages attached to one leaf per plant and then recorded the number of surviving whiteflies 2 days later, which is considered as the response variable. To measure reproductive inhibition, the fly cages were removed after obtaining the survival count but the position of each cage was marked. In the coming week, m adult whiteflies were placed in clip-on leaf cages attached to one leaf on the same plant and the number of surviving whiteflies were recorded. Therefore the number of surviving whiteflies combining with the total number of whiteflies in each experiment formed as the proportional data. There are 640 observations in the final data set with 339 zeros(53%). Compared with other observations (each of the other observations has an average of 3.5%), obviously there exist excessive zeros in the whitefly data. Therefore, this data set is an appropriate data set to be used to test the ability of the proposed LB model fitting data with zero inflation.

To consider the effects of covariates, we apply our proposed LB regression model and ZIB regression model to analyze the whitefly data. To compare our proposed model with ZIB mode, we consider the following covariates:

$$\mathbf{x} = (1, \text{plant}, \text{block}, \text{trt(i.e. treatment)}, \text{week}, \text{trt} \times \text{block}, \text{trt} \times \text{week})^\top$$

The ZIB regression model and LB regression model are as follows:

$$\text{LB model: } Y_i \sim \text{LB}(m_i, \pi, \phi_i) \quad \text{and} \quad \log \phi_i = \mathbf{x}^\top \beta_{\text{LB}}$$

and

$$\text{ZIB model: } Y_i \sim \text{ZIB}(m_i, \omega, \pi_i) \quad \text{and} \quad \log \frac{\pi_i}{1 - \pi_i} = \mathbf{x}^\top \beta_{\text{ZIB}}.$$

Here, the mixture parameter π in LB model and the zero-inflated parameter ω in ZIB model are assumed to be fixed for different subjects. The computational procedures of EM algorithm presented in Section 3.1.3 are used to calculate the MLEs of the regression coefficients. The calculation results based on LB regression model and ZIB regression model are summarized in Table 7. One can see that all indices of model diagnosis demonstrate that the proposed LB regression model outperforms the ZIB regression model.

5.3. Catheter management study

The data on catheter management study was used as an example in [12]. The purpose of this randomized clinical trial was to teach indwelling urinary catheter users the awareness and self-monitoring skills. It was conducted in New York state, and 202 subjects were recruited and randomized to the intervention and control groups. The primary outcomes of interest are whether the subjects experienced urinary tract infections (UTIs), catheter blockages, and catheter displacements during the last 2 months, as well as the corresponding counts of these experiences. Thus each patient was asked up to six times about three binary outcomes. Due to the death or dropout, some patients were asked for less than six times. So the total number of times asked varies from 1 to 6. However, the outcomes with the asking

Table 7. Results for the analysis of whitefly data using of ZIB and LB regression models.

	ZIB regression model		LB regression model	
	Logit(π)	ω	Log(ϕ)	π
Intercept	-1.3323 (0.2336)	0.4998 (0.0198)	1.1307 (0.2954)	0.9902 (0.0039)
Plant	-0.0833 (0.0206)	-	0.1353 (0.0261)	-
Block	0.2389 (0.0881)	-	0.0420 (0.1101)	-
trt	0.5449 (0.0588)	-	-0.2109 (0.0763)	-
Week	-0.0295 (0.0020)	-	0.0298(0.0027)	-
trt \times block	-0.0700 (0.0218)	-	-0.0127(0.0284)	-
trt \times week	0.0234 (0.0050)	-	-0.0304(0.0067)	-
Log-likelihood		-1361.6009		-1066.4729
AIC		2734.2017		2146.9458
BIC		2752.2036		2178.1761
Pearson χ^2		980.4065		746.8631
Deviance		2189.6592		1599.4033

time less than six are discarded. Therefore, the outcomes can be considered as the times of urinary tract infections (UTIs), catheter blockages, and catheter displacements during total six asking times and thus three sets of proportional data with binomial denominator six (6) were obtained. Further, there are 83 (43.0%), 127 (65.5%) and 140 (72.5%) zeros in the 193 outcomes of urinary tract infections (UTIs), 194 outcomes of catheter blockages, and 193 outcomes of catheter displacements, respectively. These high volumes of zeros suggest that there may be the zero-inflation issue and zero-inflated binomial model can be used to fit such data. Ye *et al.* [34] combined the repeated binary outcomes and applied Wald, LRT, score and a new statistic to test zero inflation for binomial responses. All testing results show that there are structural zeros in the three outcomes. For the purpose of illustration, we analyze these three proportional data sets by using our proposed LB model and compare it with ZIB and BB models. Table 8, Tables S4 and S5 of the supplemental file present the results from the analysis of three proportional datasets based on the ZIB, BB and LB models without covariates. In the analysis of these proportional data, the binomial denominators for all outcomes of urinary tract infections (UTIs), catheter blockages, and catheter displacements are same. Instead of Pearson chi-squared test X_1^2 and deviance test D^2 , the chi-squared test X_2^2 and Kullback–Leibler divergence G are used to assess the GOF for all three models.

From the results given in Table 8, Tables S4, S5 and Figure S2 of the supplemental file, one can see that in terms of likelihood, AIC, BIC, chi-squared X_2^2 and Kullback–Leibler divergence G , Lindley binomial model has the best performance among three models for fitting of outcomes of urinary tract infections (UTIs), catheter blockages and beta binomial model shows the superiority to other two models for fitting of outcomes from catheter displacements. These results also show that although there exist zero-inflations in three proportional data sets, the zero-inflated binomial model may not fit the data very well and the existence of zero inflation in the data does not means there exist the structural zeros.

6. Concluding remarks

In this paper, a new model for proportional data, called ‘Lindley binomial model’ has been proposed. The model is defined by compounding the binomial distribution with Lindley distribution. It can also be regarded as an extension of binomial model and can be used

Table 8. Results of model fitting for catheter blockages in Example 2.

Number of periods with positive responses	Observed values	Expected values of distributions		
		ZIB	LB	BB
0	127	127.0001	128.6495	127.9433
1	36	24.4049	30.4732	30.1008
2	16	24.6308	15.1612	15.8942
3	4	13.2581	8.9067	9.5671
4	5	4.0143	5.5072	5.8159
5	3	0.6482	3.3777	3.2703
6	3	0.0436	1.9245	1.4084
Log-likelihood		-233.8865	-215.7144	-217.5755
MLEs of parameters		$\hat{\omega} = 0.6027$ $\hat{\pi} = 0.2876$	$\hat{\pi} = 0.0663$ $\hat{\phi} = 2.1000$	$\hat{\alpha} = 0.2761$ $\hat{\beta} = 2.0419$
AIC		471.7726	435.4288	437.1510
BIC		478.3083	441.9645	443.6867
Chi-squared test χ^2_2		224.1656	4.4629	6.3388
Kullback–Leibler divergence G		41.3739	5.0277	6.7498

to fit the proportional data with extra variation such as the over/under dispersion, sparseness and zero inflation and thus is more flexible model for analyzing the proportional data. Specially, this model may be more appropriate to the binomial data with big probability at zero or at the binomial denominator. The Fisher scoring algorithm and EM algorithms are derived for the computation of the estimates of parameters in the proposed regression model. The simulation results demonstrate that the proposed EM algorithm has an excellent performance for the computation of MLEs of parameters in the proposed Lindley–binomial model with/without covariates. Hepatitis data, whitefly data and catheter management study data are used to demonstrate the proposed model and inferential methods in the proposed Lindley–binomial model. The results show the Lindley–binomial model has the advantage for the analysis of proportional data with sparseness and excessive zeros.

However, there is no model that can fit all kinds of proportional data. For example, the proportional data may display large frequencies of both zeros and right endpoints. Our proposed LB model can only address the single endpoint inflation (zero inflation or right-endpoint inflation). Therefore the proposed model has the limitation for the application. Deng and Zhang [8] and Tian *et al.* [31] proposed the endpoint inflated binomial model with the statistical properties to fit such data. We may consider the different way to account for the endpoint inflation. In fact, we are thinking about to extend the proposed Lindley binomial model by compounding the binomial model to an analogue of Lindley models. In current research, we actually compound the binomial distribution with the mixture of two gamma distributions with same rate parameter θ . Our idea for this new distribution is to compound the binomial distribution with mixture of two gamma distributions with different rate parameters like gamma distribution $(1, \theta_1)$ and gamma distribution $(2, \theta_2)$. If this is working for the bimodal data, it will be another alternative to fit the proportional data. We will be doing such research in the future.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors are very grateful to editor, associate editor and two referees for their careful reading and valuable comments which have greatly improved this paper. The research of first author is partially supported by Natural Sciences and Engineering Research Council of Canada(NSERC).

ORCID

Dianliang Deng  <http://orcid.org/0000-0002-4128-7920>

References

- [1] A. Agresti, *Foundations of Linear and Generalized Linear Models*, Wiley, New York, 2015.
- [2] R. Ascari and S. Migliorati, *A new regression model for overdispersed binomial data accounting for outliers and an excess of zeros*, *Stat. Med.* 40 (2021), pp. 3895–3914. <https://doi.org/10.1002/sim.9005>.
- [3] D. Bhati, D. Sastry, and P.M. Qadri, *A new generalized Poisson–Lindley distribution: applications and properties*, *Austrian J. Stat.* 44 (2015), pp. 35–51.
- [4] E. Calderin-Ojeda and E. Gómez-Déniz, *The multivariate negative binomial–Lindley distribution. Properties and new representation for the univariate case*, *J. Comput. Appl. Math.* 347 (2019), pp. 36–48.
- [5] M.J. Crowder, *Beta-binomial Anova for proportions*, *J. Royal Stat. Soc. Ser. C. (Appl. Stat.)* 27 (1978), pp. 34–37.
- [6] M.J. Crowder, *Inference about the intraclass correlation coefficient in the beta-binomial ANOVA for proportions*, *J. Royal. Stat. Soc. Ser. B. (Methodol)* 41 (1979), pp. 230–234.
- [7] D. Deng and S.R. Paul, *Score tests for zero-inflation and over-dispersion in generalized linear models*, *Stat. Sin.* 15 (2005), pp. 257–276.
- [8] D. Deng and Y. Zhang, *Score tests for both extra zeros and extra ones in binomial mixed regression models*, *Commun. Stat. Theory Methods* 44 (2015), pp. 2881–2897. <https://doi.org/10.1080/03610926.2013.809118>.
- [9] G. Evin, J. Merleau, and L. Perreault, *Two-component mixtures of normal, gamma, and Gumbel distributions for hydrological applications*, *Water Resour. Res.* 47 (2011), pp. W08525. <https://doi.org/10.1029/2010WR010266>.
- [10] C.P. Farrington, *On assessing goodness of fit of generalized linear models to sparse data*, *J. R. Stat. Soc. Series B(Methodol)*. 58 (1996), pp. 349–360.
- [11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science, New York, USA, 2009.
- [12] H. He, H. Zhang, P. Ye, and W. Tang, *A test of inflated zeros for poisson regression models*, *Stat. Methods Med. Res.* 28 (2019), pp. 1157–1169.
- [13] J. Hinde and C.G.B Demétrio, *Overdispersion: models and estimation*, *Comput. Stat. Data Anal.* 27 (1998), pp. 151–170.
- [14] A. Hoerl and R. Kennard, *Ridge regression: biased estimation for nonorthogonal problems*, *Technometrics* 12 (1970), pp. 55–67.
- [15] N. Keiding, *Age-specific incidence and prevalence: a statistical perspective(with discussion)*, *J. R. Stat. Soc. A* 154 (1991), pp. 371–412.
- [16] L. Lauritzen, *Bs2 Statistical Inference, Lecture 4*, University of Oxford, 2009.
- [17] D.V. Lindley, *Fiducial distributions and Bayes theorem*, *J. R. Stat. Soc. Ser. B* 20 (1958), pp. 102–107.
- [18] R. Luo and S.R. Paul, *Estimation for zero-inflated beta-binomial regression model with missing response data*, *Stat. Med.* 37 (2018), pp. 3789–3813.
- [19] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, 2nd ed., Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series, Chapman & Hall, 1989.
- [20] J.H. McDonald, *G-test of goodness-of-fit*, *Handbook of Biological Statistics (Third ed.)*. Baltimore, MD: Sparky House Publishing. 2014, pp. 53–58.

- [21] M. Menssen and F. Schaarschmidt, *Prediction intervals for overdispersed binomial data with application to historical controls*, Stat. Med. 38 (2019), pp. 2652–2663.
- [22] M. Moreno and S.R. Lele, *Improved estimation of site occupancy using penalized likelihood*, Ecology 91 (2010), pp. 341–346.
- [23] C. Morris, *Parametric empirical Bayes inference: theory and applications*, J. Am. Stat. Assoc. 78 (1983), pp. 47–55.
- [24] J. Najera-Zuloaga, D. Lee, and I. Arostegui, *Comparison of beta-binomial regression model approaches to analyze health-related quality of life data*, Stat. Methods Med. Res. 27 (2018), pp. 2989–3009.
- [25] S.R. Paul, *Analysis of proportions of affected fetuses in teratological experiments*, Biometrics 38 (1982), pp. 361–370.
- [26] R.L. Prentice, *Binary regression using an extended beta-binomial distribution with discussion of correlation induced by covariate measurement errors*, J. Am Stat. Assoc. 81 (1986), pp. 321–327.
- [27] K.K. Saha and S.R. Paul, *Bias-corrected maximum likelihood estimator of the intraclass correlation parameter for binary data*, Stat. Med. 24 (2005), pp. 3497–3512.
- [28] M. Sankaran, *The discrete Poisson–Lindley distribution*, Biometrics 26 (1970), pp. 145–149.
- [29] R. Shanker, S. Sharma, and R. Shanker, *A two-parameter Lindley distribution for modeling waiting and survival times data*, Appl. Math. 04 (2013), pp. 363–368.
- [30] R.R.M. Tajuddin, N. Ismail, K. Ibrahim, and S.A.A. Bakar, *A four-parameter negative binomial-Lindley distribution for modeling over and underdispersed count data with excess zeros*, Commun. Stat. Theory Methods 51 (2022), pp. 414–426. <https://doi.org/10.1080/03610926.2020.1749854>.
- [31] G.L. Tian, H. Ma, Y. Zhou, and D. Deng, *Generalized endpoint-inflated binomial model*, Comput. Stat. Data Anal. 89 (2015), pp. 97–114. <https://doi.org/10.1016/j.csda.2015.03.009>.
- [32] M.W. van Iersel, R.D. Oetting, and D.B. Hall, *Imidacloprid applications by subirrigation for control of silverleaf whitefly (Homoptera: Aleyrodidae) on poinsettia*, J. Econ. Entomol. 93 (2000), pp. 813–819.
- [33] D.A. Williams, *The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity*, Biometrics 31 (1975), pp. 949–952.
- [34] P. Ye, Y. Tang, L. Sun, W. Tang, and H. He, *Testing inflated zeros in binomial regression models*, Biom. J. 63 (2021), pp. 59–80. <https://doi.org/10.1002/bimj.202000028>.
- [35] H. Zamani and N. Ismail, *Negative binomial–Lindley distribution and its application*, J. Math. Stat. 6 (2010), pp. 4–9.