

## Bayesian inference for zero-inflated negative binomial lindley model of overdispersed count data with excess zeros

Cenyu Hu<sup>a</sup>, Ling Fang<sup>b</sup>, Xianming Shi<sup>a\*</sup> and Yalong Wang<sup>a</sup>

<sup>a</sup>Army Engineering University of the PLA Shijiazhuang Campus, Hebei 050003, China

<sup>b</sup>Army Logistics Academy, Chongqing, China

### CHRONICLE

#### Article history:

Received March 16 2025

Received in Revised Format

May 19 2025

Accepted July 11 2025

Available online July 11 2025

#### Keywords:

*Zero-inflated negative binomial-Lindley distribution generalized Linear Model Bayesian inference Regression analysis*

### ABSTRACT

This article aims to develop the zero-inflated negative binomial-Lindley regression model to address the complexity of count data with zero excess and over-dispersion. The proposed compound distribution combines the zero generation mechanism with the Lindley distribution process, and the Bayesian hierarchical framework with MCMC sampling is adopted for parameter estimation, overcoming the limitations of traditional count models in handling complex data structures. The model is applied to two real datasets, one of which is characterized by a large number of zero observations. Its performance is compared with that of the NB-L and NB model. The results show that when the dataset presents the large number of zero values and the long tail feature, the ZINB-L GLM describes the dataset better than the other models.

© 2025 by the authors; licensee Growing Science, Canada

## 1. Introduction

Count data analysis forms the foundation of statistical modeling methodologies that span multiple disciplines including economics, epidemiology, and the social sciences (Moksony & Hegedus 2014; Altun 2019). According to Lord and Menting (2010), researchers developed and implemented innovative modeling techniques specifically tailored to count data frameworks in their analysis. In these research contexts, the dependent variable typically quantifies event frequency or occurrence rates. However, frequently present significant methodological challenges that complicate statistical inference. Two common issues are zero inflation, where observed zeros occur more frequently than expected under standard distribution assumptions, and overdispersion, where the variance of the count variable exceeds its mean, violating core count model assumptions. Traditional methods, like the Poisson and negative binomial distributions, address these complexities, but the Poisson model often produces biased estimates and large standard errors when applied to overdispersed data, as the assumption of equal mean and variance is rarely met in practice (Lord & Menting, 2010; Lord & Geedipally, 2011).

The negative binomial model provides more flexibility for overdispersion, but research shows it may still be inadequate for comprehensive count data modeling. Datasets may exhibit characteristics like excessive zeros, heavy tails, or low means, which can significantly affect model performance and inference validity. To address the first critical issue of excessive zero values in count data, researchers have developed specialized statistical frameworks. Various innovative models have emerged to handle datasets characterized by zero-inflation phenomena. Zamani and Ismail (2010) pioneered the negative binomial-Lindley distribution (Dzinyela et al., 2024), a sophisticated approach specifically designed to accommodate datasets with disproportionate zero observations. The NB-L model integrates the negative binomial model with the Lindley distribution. These two parameter distributions are not exactly the same as the traditional zero-inflated models in terms of their convergence to zero. Lord and Geedipally (2011) studied the NB-L distribution and discovered a large number of zero-inflated collision frequency datasets. Their result analysis indicated that in the face of excessive zero observations, the NB-L distribution

\* Corresponding author

E-mail [x.m.shi@126.com](mailto:x.m.shi@126.com) (X. Shi)

ISSN 1923-2934 (Online) - ISSN 1923-2926 (Print)

2025 Growing Science Ltd.

doi: 10.5267/j.ijiec.2025.7.001

outperformed the standard negative binomial model and Poisson distribution in terms of goodness-of-fit indicators. Moreover, Geedipally et al. (2012) extended and innovated the NB-L distribution through GLM. The evaluation showed that for the ZINB model, the NB-L GLM model was more suitable for fitting empirical count data.

To further address the issue of excessive zero values in technical data, we embedded the NB-L model into the zero-inflated framework. Although the Lindley distribution (Zamani & Ismail, 2010; Feng, 2019) has a closed form in various statistical software, there are still computational difficulties when combined with the NB distribution in the GLM framework. To solve this problem, researchers adopted the Bayesian approach (Fu, 2015) to construct a hierarchical model of the ZINB-L distribution within the GLM framework (Geedipally et al., 2012; Rahman Shaon and Qin, 2016). Bayesian inference, in the case of small samples, determines the posterior distribution through the prior distribution, which helps to improve the accuracy of parameters. The ZINB-L GLM model can significantly enhance the fitting degree and interpretability of the model by differentiating the systematic zero-inflation process from the underlying count value mechanism. It is particularly suitable for datasets with extreme values and excessive zeros.

This paper employs the Bayesian inference method for parameter estimation, and simplifies the complexity of the model through the MCMC method. The Bayesian approach integrates prior information to determine the posterior distribution, thereby achieving uncertainty quantification and stability in hierarchical models. For maximum likelihood estimation, a large amount of data is often required to support it, which may lead to data non-convergence and large biases. Therefore, in small sample events, the Bayesian method is adopted for parameter estimation.

The main objective of this paper is to elucidate a new model - ZINB-L GLM, and to compare it with other models by applying real datasets. Through model evaluation, it can be concluded that the ZINB-L GLM regression model outperforms other models when dealing with situations where there are too many zero observations.

**2. The Zero-Inflated Negative Binomial-Lindley Distribution Model**

In this section, we propose a new mixed model, called the zero-inflated negative binomial-Lindley (ZINB-L) distribution. ZINB-L distribution is an innovative model which is derived by blending the ZINB distribution with the Lindley distribution.

Firstly, we introduce the ZINB random variable as follows:

*2.1 The Zero-Inflated Negative Binomial Distribution*

The negative binomial distribution is the limiting form derived from a series of independent Bernoulli trials (Hilbe, 2011; Gijepilli et al., 2012). Let  $Y$  be a random variable with distribution parameterized by  $r$  and  $p$ , denoted as  $Y \sim NB(r, p)$ . The PMF of  $Y$  is

$$P(Y = y; r, p) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}(1-p)^y (p)^r \tag{1}$$

where  $r > 0$  and  $0 < p < 1$ . Its mean and variance are respectively:

$$E(Y) = \frac{r(1-p)}{p} \text{ and } Var(Y) = \frac{r(1-p)}{p^2} \tag{2}$$

Given the highly concentrated zero values and the significant heterogeneity of the data, we address this issue by using the zero-inflated approach and reparameterize it. We express the probability  $p$  as a function of the discrete parameter  $r$ , that is

$p = \frac{r}{\mu+r}$ , and the corresponding mixture distribution is:

$$f(y; r, \mu, \eta) = \begin{cases} \eta + (1-\eta)\left(\frac{r}{\mu+r}\right)^r & y = 0 \\ (1-\eta)\frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}\left(\frac{r}{\mu+r}\right)^r \left(\frac{\mu}{\mu+r}\right)^y & y > 0 \end{cases} \tag{3}$$

where  $r$  is the dispersion function,  $\eta$  is the zero-inflation parameter, and when  $\eta = 0$ , this distribution degenerates into the NB distribution.

The expectation and variance of the ZINB distribution are:

$$E(Y) = (1 - \eta)\mu \tag{4}$$

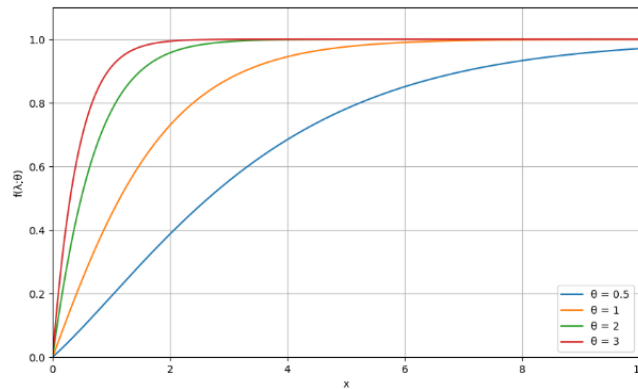
$$Var(Y) = (1 - \eta)\mu (1 + \eta\mu + r^{-1}\mu)$$

### 2.2 Lindley Distribution

The Lindley distribution was proposed by Lindley (1958) and is a univariate distribution that is interpreted as a mixture of the exponential distribution  $E(\theta)$  and the gamma distribution  $gamma(2, \theta)$  (Elbatal, Merovci and Elgarhy 2013). The PMF of this distribution is defined as follows:

$$f(\lambda; \theta) = \frac{(1 + \lambda)\theta^2}{1 + \theta} e^{-\theta\lambda} \tag{5}$$

where  $\theta$  represents the scale parameter, and the structural configuration endows the lindley distribution with a unique tailed characteristic. The plot of the lindley distribution are shown in Fig. 1.



**Fig. 1.** The pmf plots of  $\theta$  with the specified value of parameters

Compared with the traditional exponential distribution, it demonstrates superior performance when modeling random phenomena with a right-skewed pattern (Ghitany et al., 2008; Zamani & Ismail, 2010). Additionally, its moment generating function (mgf) of  $\lambda$  is

$$M_{\lambda}(t; \theta) = E[e^{t\lambda}] = \frac{\theta^2}{(1 + \theta)(\theta - t)} \left(1 + \frac{1}{\theta - t}\right) \quad t > 0 \tag{6}$$

As well as, the first and second moments of the Lindley distribution can be obtained from Eq. (6) as follows:

$$E(\lambda) = \frac{\theta + 2}{\theta(1 + \theta)} \quad \text{and} \quad E(\lambda^2) = \frac{3 + 2\theta}{\theta^2(1 + \theta)} \tag{7}$$

### 2.3 ZINB-L Distribution

Let Y be a random variable, that is  $Y \sim ZINB-L(r, \mu, \eta, \theta)$ . The ZINB-L distribution can be expressed in terms of the ZINB distribution and the Lindley distribution. According to Eq. (1), Eq. (2), and Eq. (3), the PMF of the ZINB-L distribution is:

$$\begin{aligned}
 f(y; \mu, r, \theta, \eta) &= \begin{cases} \eta + (1 - \eta)f(y = 0; \mu, r, \theta, \eta) & y = 0 \\ (1 - \eta)f(y; \mu, r, \theta, \eta), & y > 0 \end{cases} \\
 &= \int_0^{\infty} ZINB(y; r, \eta, \lambda\mu) \text{lindley}(\lambda; \theta) d\lambda \\
 &= \begin{cases} \eta + \frac{(1 - \eta)\theta^2}{1 + \theta} \int_0^{\infty} \left(\frac{\lambda\mu}{\lambda\mu + r}\right)^r (1 + r)e^{-\theta\lambda} d\lambda & y = 0 \\ (1 - \eta) \frac{\Gamma(y + r)\theta^2}{\Gamma(y + 1)\Gamma(r)(1 + \theta)} \int_0^{\infty} \left(\frac{r}{\lambda\mu + r}\right)^r \left(\frac{\lambda\mu}{\lambda\mu + r}\right)^y (1 + \lambda)e^{-\theta\lambda} d\lambda & y > 0 \end{cases} \tag{8}
 \end{aligned}$$

where  $y_i = 0, 1, 2, \dots, \mu_i > 0, i = 1, 2, \dots, n, r, \theta$  represents a positive parameter. Its mean and variance are respectively:

$$E(Y | \mu, r, \theta, \eta) = (1 - \eta)E(Y | \mu, r, \theta) = (1 - \eta)\mu_i E(\lambda) \tag{9}$$

$$Var(Y | \mu, r, \theta, \eta) = (1 - \eta)Var(Y | \mu, r, \theta) + (1 - \eta)\eta E(Y | \mu, r, \theta, \eta)^2$$

### 3. Bayesian inference of ZINB-L GLM regression model

#### 3.1 The framework of GLM

Within the ZINB-L GLM framework, the conditional mean is modeled as a nonlinear function of the explanatory variable vector. Specifically, this model assumes that the expected value of the response variable has a clear association with the covariates through a logit link function, where the systematic component can be expressed as:

$$g(\mu_i) = \ln(\mu_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \tag{10}$$

where the expected value of the sample is a non-negative number and it establishes a linear relationship between the expected value of the response variable and the explanatory variables through a connection function.  $w_i = g(\mu_i) = \ln(\mu_i)$ . The conditional expectation is associated with a linear predictor through an ln function, thereby establishing the GLM framework. Its expression is:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) = \exp(x_i^T \beta) \tag{11}$$

where  $\mu_i$  represents linear predictor,  $x_{ip}$  denotes the vector of covariates, and  $\beta_j$  corresponds to the regression coefficients. Similarly,  $\eta$  will be introduced into the relevant covariates and connected through the logit function.

$$\text{logit}(\eta_i) = \ln\left(\frac{\eta_i}{1 - \eta_i}\right) = \gamma_0 + \gamma_1 x_{i1} + \dots + \gamma_t x_{it} \tag{12}$$

$$\eta_i = \frac{\exp(x_i^T \gamma)}{1 + \exp(x_i^T \gamma)}$$

The following is the construction of the regression model for the response variable  $Y$ . Within the framework of GLM, the ZINB-L GLM regression model is introduced. Let the conditional distribution of  $y_i | x_i^T$  to  $ZINB-L(y; \mu_i, r, \theta, \eta_i)$

$$f(y_i | x_i^T) = \begin{cases} \frac{e^{x_i^T \gamma}}{1 + e^{x_i^T \gamma}} + \frac{\theta^2}{(1 + e^{x_i^T \gamma})(1 + \theta)} \int_0^\infty \left(\frac{\lambda e^{x_i^T \beta}}{\lambda e^{x_i^T \beta} + r}\right)^r (1 + r)e^{-\theta \lambda} d\lambda & y = 0 \\ \frac{\theta^2 \Gamma(y_i + r)}{(1 + \theta)(1 + e^{x_i^T \gamma})\Gamma(y_i + 1)\Gamma(r)} \int_0^\infty \left(\frac{r}{\lambda e^{x_i^T \beta} + r}\right)^r \left(\frac{\lambda e^{x_i^T \beta}}{\lambda e^{x_i^T \beta} + r}\right)^{y_i} (1 + \lambda)e^{-\theta \lambda} d\lambda & y > 0 \end{cases} \tag{13}$$

The expectation and variance of the ZINB-L GLM distribution are:

$$E(Y_i | x_i^T) = (1 - \eta_i)E(Y | \mu_i, r, \theta) = (1 - \eta_i)\mu_i E(\lambda) \tag{14}$$

$$Var(Y_i | x_i^T) = (1 - \eta_i)Var(Y | \mu_i, r, \theta) + (1 - \eta_i)\eta_i E(Y | \mu_i, r, \theta)^2$$

By integrating prior knowledge with observational data, the Bayesian framework not only can adapt to highly uncertain situations, but also can dynamically update parameter estimates when new information is introduced, thereby demonstrating strong applicability in the modeling and analysis of complex systems (Lunn et al., 2013; Yamruboon et al., 2019; Gelman et al., 2013). Let  $\Omega = (r, \theta, \beta, \gamma)^T$  be the regression parameter vector. The likelihood function of  $\Omega$  is

$$L(\Omega | y, x^T) = \prod_{i=1}^n \left[ \frac{e^{x_i^T \gamma}}{1 + e^{x_i^T \gamma}} + \frac{\theta^2 \int_0^\infty \left(\frac{\lambda e^{x_i^T \beta}}{\lambda e^{x_i^T \beta} + r}\right)^r (1 + r)e^{-\theta \lambda} d\lambda}{(1 + e^{x_i^T \gamma})(1 + \theta)} \right]^{y_i=0} \tag{15}$$

$$\left[ \frac{\theta^2 \Gamma(y_i + r) \int_0^\infty \left(\frac{r}{\lambda e^{x_i^T \beta} + r}\right)^r \left(\frac{\lambda e^{x_i^T \beta}}{\lambda e^{x_i^T \beta} + r}\right)^{y_i} (1 + \lambda)e^{-\theta \lambda} d\lambda}{(1 + \theta)(1 + e^{x_i^T \gamma})\Gamma(y_i + 1)\Gamma(r)} \right]^{y_i>0}$$

### 3.2 Bayesian hierarchical modeling

The ZINB-L model proposed above can be regarded as an extension form of GLM. However, the likelihood function of the ZINB-L GLM regression model by Eq. (15) is not a closed-form expression but can be represented by a hierarchical Bayesian model.

Let the random effect  $\lambda \sim \text{Lindley}(\theta)$ . By introducing an auxiliary latent variable  $\theta$ , it can be decomposed as

$$\lambda \sim \text{Lindley}(\theta) \sim \frac{1}{1+\theta} \text{gamma}(2, \theta) + \frac{\theta}{1+\theta} \text{gamma}(1, \theta) \tag{16}$$

The hierarchical Bayesian framework of the ZINB-L GLM adopts a five-level structure. The expression is as follows:

$$f_Y(y_i | x_i^T) = \begin{cases} \eta_i + (1 - \eta_i) f(y = 0; \mu_i, r, \theta, \eta_i) \\ (1 - \eta_i) f(y; \mu_i, r, \theta, \eta_i) \end{cases}$$

$$f(y_i; \mu_i, \eta_i, r | \lambda) = \text{ZINB}(y_i; r, \lambda \mu_i, \eta_i)$$

$$\mu_i = \exp(x_i^T \beta)$$

$$\eta_i = \frac{\exp(x_i^T \gamma)}{1 + \exp(x_i^T \gamma)}$$

$$\lambda \sim \text{Lindley}(\theta) \tag{17}$$

### 3.3 Prior distributions and joint posterior density

This framework adopts the Bayesian statistical framework and systematically integrates the prior information through the setting of probability distributions. All unknown parameters in the study are taken into account. Suppose that the parameters  $r, \theta$  of the ZINB-L GLM regression model follow the gamma distribution, while  $\beta, \gamma$  follow the normal distribution, and all parameters are mutually independent. The joint prior distribution of the unknown parameters is as follows:

$$r \sim \text{gamma}(\alpha_r, z_r)$$

$$\theta \sim \text{gamma}(\alpha_\theta, z_\theta)$$

$$\beta \sim N(v_0, \sigma_\beta)$$

$$\gamma \sim N(b_0, \sigma_\gamma) \tag{18}$$

where  $\alpha_r, z_r, \alpha_\theta, z_\theta$  are all known positive parameters,  $v_0, b_0$  is a vector of hyperparameters, and  $\sigma_\beta, \sigma_\gamma$  is a  $(k + 1)$  order known non-negative specific matrix. Assuming each parameter conforms to independent and identically distributed, that is, the joint prior distribution of all unknown parameters is:

$$\pi(\Omega) = \pi(r)\pi(\theta)\pi(\beta)\pi(\gamma) \tag{19}$$

Bayesian theorem states that the posterior distribution can be obtained by multiplying the likelihood function with the prior distribution. Combining the likelihood function in Eq. (15) and the prior distribution in Equ. (19), the posterior distribution is obtained as

$$\pi(\Omega | X) \propto L(\Omega | y, X)\pi(r)\pi(\theta)\pi(\beta)\pi(\gamma) \tag{20}$$

The complete posterior distributions of each parameter of X by Eq. (22) are all obtained as

$$\pi(r | y, X, \theta, r, \gamma, \beta) \propto L(\Omega | y, X)\pi(r)$$

$$\pi(\theta | y, X, \theta, r, \gamma, \beta) \propto L(\Omega | y, X)\pi(\theta)$$

$$\pi(\beta | y, X, \theta, r, \gamma, \beta) \propto L(\Omega | y, X)\pi(\beta)$$

$$\pi(\gamma | y, X, \theta, r, \gamma, \beta) \propto L(\Omega | y, X)\pi(\gamma) \tag{21}$$

During the model estimation process, a total of three Markov chains were employed. Each chain had 30,000 iterations and the first 15,000 iterations were discarded. Therefore, the remaining 15,000 iterations are used for estimating the coefficients. The Gelman-Rubin (G-R) convergence statistic is employed to verify whether the simulation has converged correctly. The research team ensured that the GR statistic was less than 1.1. For comparison purposes, Mitra and Washington (2007) proposed that

when the GR statistic is less than 1.2, convergence can be achieved. This indicator requires the value to be as close to 1 as possible, and it should not exceed 1.1, and preferably not more than 1.05.

#### 4. Empirical analysis

This section is divided into two subsections and the first subsection describes the datasets used for analysis, the second subsection covers the application of the ZINB-L GLM in these datasets by constructing a multi-model comparison framework, systematically comparing the ZINB-L GLM with the standard negative binomial and the NB-L GLM.

##### 4.1 Data description

Two real public datasets have been cited. Aryuyuen (2021) also employed this dataset in his research to explore the Bayesian inference (Dao et al., 2025) method of the NB-GL regression model. Based on the same original data, this study independently completed data analysis and result derivation to verify the superiority of the model. Compared with the model (Aryuyuen & Bodhisuwan, 2021), the results of this study demonstrate superior performance in terms of model fit goodness and the explanatory power for zero-value data. All calculations and conclusions are re-generated based on the original public data and do not directly reference the processed results or data from previous studies.

##### 4.1.1 Strike Data Set

The dataset used in this study is StrikeNb, which can be freely accessed through the Ecdat package of R language (R Core Team, 2020). This dataset was initially compiled and recorded by Croissant and Graves (2020) and provided reliable data support for the study of labor strikes in the American manufacturing industry. The dataset used in this study is derived from the pioneering work of Croissant and Graves (2020), and it can be freely obtained through the R language econometrics analysis package Ecdat. A dataset containing 108 observations has been compiled, covering the contract strike events in the manufacturing industry of the United States from 1968 to 1976. The descriptive summary of the variables given in Table 1 indicates that the zero proportion of the response variable (defined as the number of strikes) is 4.63%, the expected value of the "strikes" column is 5.24, the variance is 13.94, and there is an issue of over-dispersion. The dispersion index is 2.685. Moreover, the explanatory variable "economic activity" (output) is used as a covariate in this analysis.

**Table 1**  
Descriptive Overview of Strike Data Variables (n = 108)

Variables	Min	Median	Max	Average (std. dev)
strikes (response)	0	5	18	5.24 (3.75)
output	-0.139	-0.00013	0.085	-0.003 (0.054)

##### 4.1.2 Doctor's visit Data Set

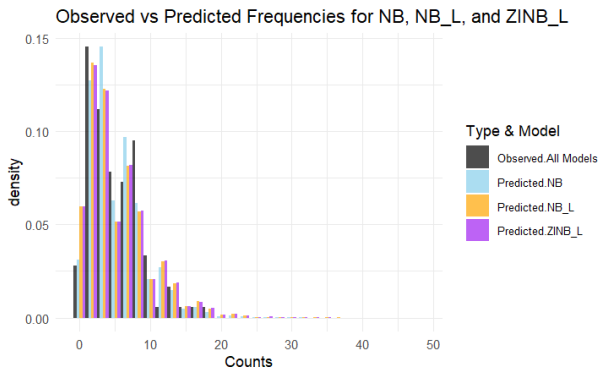
This study makes use of the Doctor Visits dataset, which is sourced from the AER package of R language (Kleiber and Zeileis, 2008). It was initially compiled by Cameron and Trivedi (2013) and was previously employed by Cameron et al. (1988) for the analysis of medical service demand. The dataset records the visiting behaviors of 5,190 single adults in Australia during the period from 1977 to 1978. This dataset takes the number of visits as the dependent variable, and includes covariates such as gender (1 = female, 0 = male), age (calculated by dividing years by 100), annual income (in ten thousand yuan), number of illnesses, days of reduced activity due to illness (reduced), and private insurance status (private: 1 = yes, 0 = no). The descriptive summary of the variables given in Table 2 indicates that the zero-proportion of the response variable "visits" is 0.798, the expected value of visits is 0.30, the variance is 0.63, and the dispersion index is 2.11. The preliminary analysis shows that there are zero-inflation and excessive dispersion characteristics in the number of visits. The summary of these data is presented in Table 2.

**Table 2**  
Descriptive Overview of Quantitative Variables in Doctor Visits Data Set

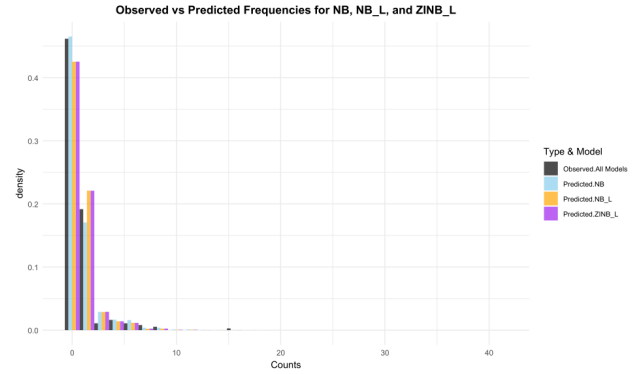
Variables	Max	Median	Min	Average (std. dev)
visits(response)	9	0	0	0.30 (0.79)
age	0.72	0.32	0.19	0.40 (0.20)
income	1.5	0.55	0	0.58 (0.36)
illness	5	1	0	1.43 (1.38)
reduced	14	0	0	0.86 (2.88)

##### 4.2 Modeling results

In this section, the analysis results of the ZINB-L GLM distribution and its regression model are presented. Specifically, in Section 4.2.1, the ZINB-L distribution is examined in terms of its performance in fitting the response variable Y; in Section 4.2.2, the application and effectiveness of the ZINB-L regression model are discussed.



**Fig. 2.** The bar chart plots of observed frequency and expected frequencies of each distributions from Strike Data Bayesian inference for the ZINB-L distribution



**Fig. 3** The bar chart plots of observed frequency and expected frequencies of each distributions from doctor's visit Data

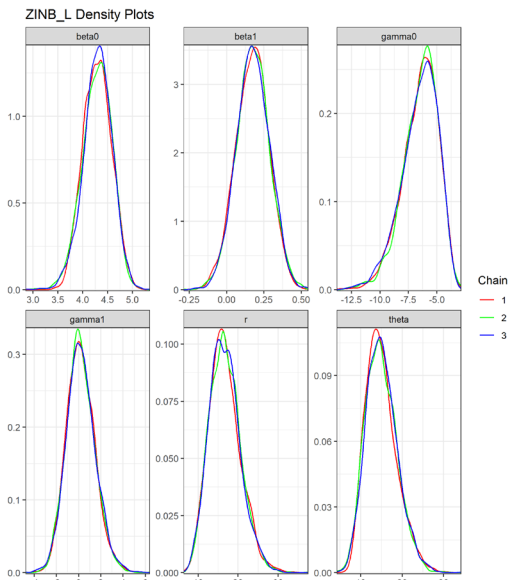
In this section, we evaluated the fitting performance of the ZINB-L GLM distribution for two datasets and compared it with the NB distribution and NB-L GLM distribution. The parameters of each distribution were estimated using Bayesian inference in order to assess the goodness of fit, we employed the KS test (Arnold & Emerson, 2011), Deviance, and DIC. The distribution that provided the best fit was identified by minimizing the KS statistic, Deviance, and DIC values.

**Table 3**

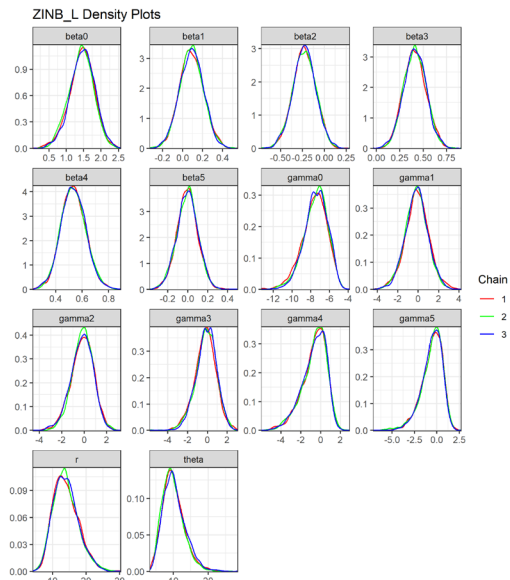
Posterior Distribution Summaries for NB, NB-L, and ZINB-L GLM distribution for Strike Data.

Parameter	NB		NB-L		ZINB-L	
	Mean(s.e.)	95% Cr.I.	Mean(s.e.)	95% Cr.I.	Mean(s.e.)	95% Cr.I.
$p$	0.36(0.05)	(0.26,0.47)	-	-	-	-
$r$	3.02(0.68)	(1.96,4.63)	3.98(0.69)	(1.89,4.57)	3.88(0.59)	(3.21,4.78)
$\theta$	-	-	18.99(22.7)	(0.00,57.96)	16.45(18.79)	(0.00,37.54)
Deviance	568.5		568.6		565.4	
DIC	570.5		570.8		568.6	
KS	0.0412		0.0478		0.0438	
p-value	0.9930		0.9659		0.9597	

The results for the "Strikes" dataset in Table 3 show that both the NB distribution and the ZINB-L distribution fit the data well. While the ZINB-L distribution shows slightly higher values in the KS test, Deviance, and DIC, these values are closely aligned with those of the NB distribution's KS statistic. Therefore, the ZINB-L distribution is a suitable model for this dataset, yielding a fit comparable to that of the NB distribution. In Table 4, the ZINB-L GLM distribution is shown to outperform both the NB and Poisson distributions when applied to the second dataset by Fig.3. It excels in key goodness of fit metrics, such as Deviance, and DIC, demonstrating its effectiveness in capturing data characteristics and optimizing the fit.



**Fig. 4.** Density plots of the three MCMC chain for  $r, \theta$  and  $\beta = (\beta_0, \dots, \beta_5)^T$  from ZINB-L regression model for the strikes data



**Fig. 5.** Density plots of the three MCMC chain for  $r, \theta$  and  $\beta = (\beta_0, \dots, \beta_5)^T$  from ZINB-L regression model for doctor's visit data

**Table 4**

Posterior Distribution Summaries for NB, NB-L GLM, and ZINB-L GLM distribution for the doctor visits data.

Parameter	NB		NB-L		ZINB-L	
	Mean(s.e.)	95% Cr.I.	Mean(s.e.)	95% Cr.I.	Mean(s.e.)	95% Cr.I.
$p$	0.56(0.02)	(0.52,0.59)	-	-	-	-
$r$	0.38(0.68)	(0.38,0.44)	0.38(0.03)	(0.33,0.44)	0.38(0.03)	(0.33,0.43)
$\theta$	-	-	12.99(18.7)	(0.00,47.28)	9.76(13.72)	(0.00,22.59)
Deviance	7174.0		7173.9		7170.8	
DIC	7176.0		7175.8		7171.3	

4.2.2 Bayesian inference for the ZINB-L GLM regression distribution

This section elaborates on the Bayesian approach using GLM for parameter estimation, and investigates the model efficiency of the ZINB-L GLM compared with NB-L GLM and NB distribution. The evaluation metrics include Deviance, DIC. The minimum values of deviance, DIC indicate that the model has the optimal fitting performance. Therefore, the optimization of the fitting effect is achieved by incorporating the  $\theta$  of the Lindley distribution into the adjustment of the ZINB-L GLM through Eq. (5). The Bayesian approach can be adopted to monitor and report the convergence of algorithms, which is a key aspect for ensuring the reliability of results. The sampling distribution of parameter values is analyzed by using the tracking graph and the posterior density graph to test the stationarity and convergence of the MCMC chain (Ntzoufras 2011).

**Table 5**

Posterior Distribution Summaries for NB, NB-L GLM, and ZINB-L GLM for the strike Data.

Parameter	NB		NB-L GLM		ZINB-L GLM	
	Mean(s.e.)	95% Cr.I.	Mean(s.e.)	95% Cr.I.	Mean(s.e.)	95% Cr.I.
$\beta_0$	1.66(0.07)	(1.52,1.80)	3.57(1.92)	(-0.01,6.92)	4.10(1.22)	(1.45,6.47)
$\beta_1$	3.26(1.39)	(0.55,6.10)	3.02(1.48)	(0.25,5.94)	2.84(1.05)	(0.77,4.98)
$r$	2.89(0.63)	(1.93,4.31)	3.38(0.88)	(2.08,5.42)	4.72(1.30)	(2.79,7.85)
$\theta$	-	-	8.99(9.22)	(0.28,34.40)	11.77(10.13)	(0.96,38.89)
Deviance	563.80		563.47		558.6	
DIC	567.20		566.20		561.10	

The performance results of the ZINB-L GLM for the strike data and the doctor visit data are presented in Table 5 and Table 6 respectively. Taking the strike data as an example, it is a significant relationship between the response variable (strikes) and the covariate (output). The results in Table 5 show that the ZINB-L GLM outperforms the NB-L, and NB models in terms of deviance and DIC indicators.

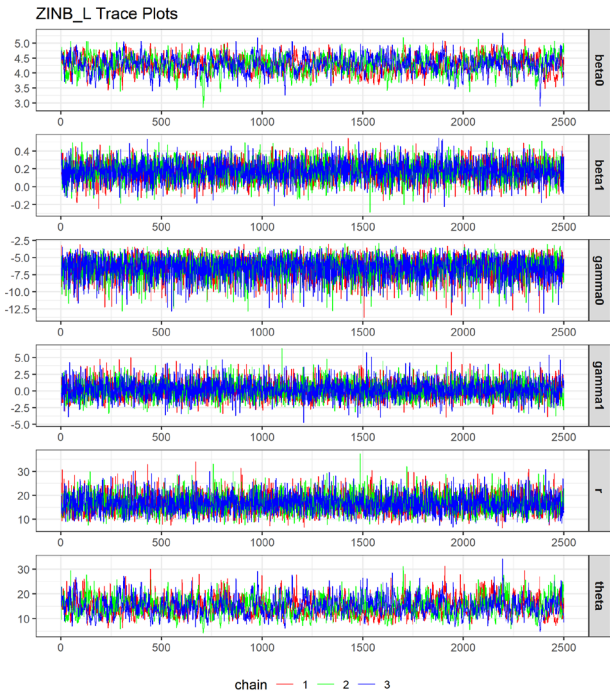
**Table 6**

Posterior Distribution Summaries for NB, NB-L GLM, and ZINB-L GLM distribution for the doctor visits data

Parameter	NB		NB-L GLM		ZINB-L GLM	
	Mean(s.e.)	95% Cr.I.	Mean(s.e.)	95% Cr.I.	Mean(s.e.)	95% Cr.I.
$\beta_0$	-2.28(0.13)	(-2.52,2.04)	-1.49(1.52)	(-4.02,1.61)	-1.56(0.83)	(-2.97,1.07)
$\beta_1$	0.22(0.07)	(0.08,0.35)	0.22(0.07)	(0.08,0.35)	0.22(0.07)	(0.08,0.35)
$\beta_2$	0.33(0.21)	(-0.09,0.72)	0.33(0.21)	(-0.08,0.75)	0.33(0.21)	(-0.08,0.75)
$\beta_3$	-0.16(0.10)	(-0.37,0.05)	-0.15(0.10)	(-0.35,0.05)	-0.14(0.10)	(-0.33,0.05)
$\beta_4$	0.22(0.02)	(0.17,0.26)	0.22(0.02)	(0.17,0.26)	0.24(0.02)	(0.19,0.26)
$\beta_5$	0.14(0.01)	(0.12,0.16)	0.14(0.01)	(0.13,0.16)	0.14(0.01)	(0.13,0.16)
$r$	0.92(0.09)	(0.77,1.10)	0.92(0.09)	(0.76,1.11)	0.92(0.07)	(0.79,1.08)
$\theta$	-	-	1.04(0.94)	(0.07,3.56)	0.89(0.89)	(0.01,3.21)
Deviance	6410.77		6410.73		6405.57	
DIC	6424.00		6423.70		6417.49	

For the second dataset, the predictive model for the first five covariates of the doctor's visit was studied. As shown in Table 6, through the model performance evaluation, it was found that the ZINB-L GLM exhibited the best goodness-of-fit, with its DIC and deviance values being significantly lower than those of other comparison models. In terms of covariate effect analysis, the regression coefficient directions of all models were consistent. The ZINB-L GLM model effectively captured the over-dispersion characteristics and zero-value clustering phenomenon of the visit data through the dual-process modeling mechanism (zero-inflation process and count process), indicating that different modeling methods have robustness in judging the trend of variable influence. To verify the applicability of the ZINB-L GLM model, we further evaluated the model performance by analyzing diagnostic plots (Fung, 2024). These diagnostic plots include posterior density plots and trace plots, which are used to check the quality of posterior distribution samples and the convergence of the model. The posterior density plot can intuitively display the characteristics of parameter distributions, while the trace plot reflects the stability and consistency of the sampling process. Fig. 4 and Fig. 5 respectively illustrate the posterior density distributions of all parameters of the ZINB-L GLM in the first and second datasets. The results can be seen that the posterior densities of the three parallel chains achieved a high degree of overlap after the burn-in period, indicating that the posterior distribution samples of the parameter estimation have good representativeness and consistency. Meanwhile, the trace plots in Fig 6 and Fig 7 both show the changing trends of all parameters in the sampling sequence, and the distribution of the simulated parameter values is dense

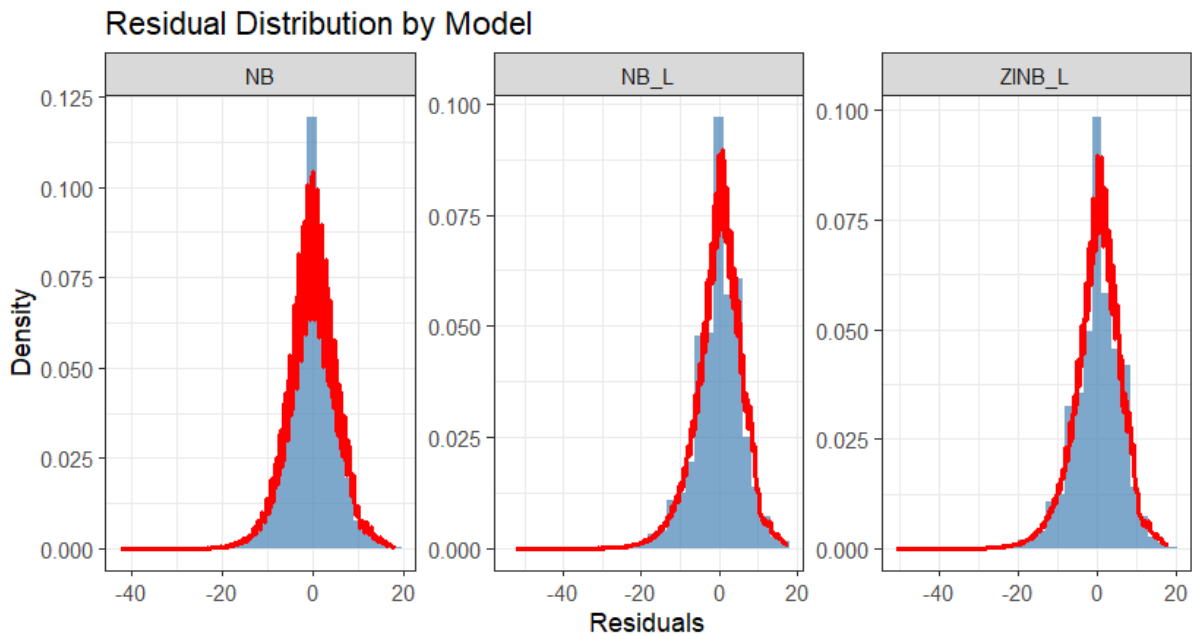
and close to vertical, further verifying the convergence and sampling stability of the model.



**Fig. 6.** Trace plots of the three MCMC chain for  $r, \theta$  and  $\beta = (\beta_0, \dots, \beta_5)^T$  from ZINB-L regression model for strikes data

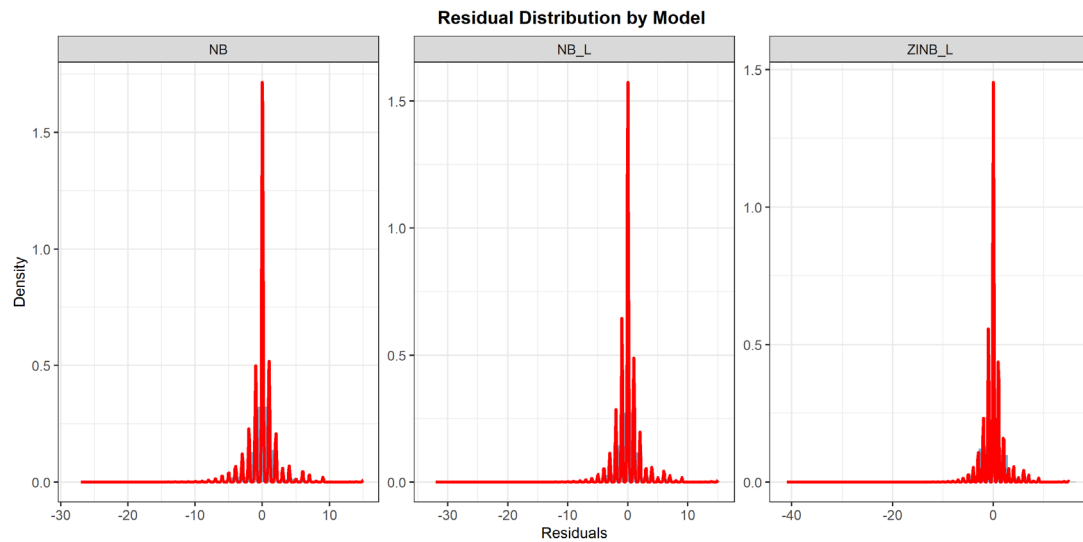


**Fig. 7.** Trace plots of the three MCMC chain for  $r, \theta$  and  $\beta = (\beta_0, \dots, \beta_5)^T$  from ZINB-L regression model for doctor's visit data



**Fig. 8.** Cumulative residual plot for the strikes data

The cumulative residual diagnosis method has been utilized to conduct an in-depth assessment of the model fitting situation. The analysis indicates the ZINB-L GLM significantly outperforms both the NB-L GLM and the standard NB model. As illustrated in Fig. 8 and Fig. 9, the cumulative residuals for the ZINB-L GLM closely converge to zero, indicating minimal systematic bias. In contrast, the NB-L GLM and NB models consistently exhibit negative residuals, suggesting persistent underestimation across the range of the response variable. Moreover, the maximum deviation observed in the ZINB-L GLM is considerably lower than that in the alternative models, confirming its superior predictive accuracy and robustness, consistent with previous dataset evaluations.



**Fig. 9.** Cumulative residual plot for the doctor's visit data

The residual plot analysis demonstrates that the ZINB-L GLM exhibits exceptional performance in fitting the dataset, outperforming both the NB-L and NB models. The ZINB-L GLM not only addresses zero-inflation and overdispersion effectively but also ensures robust parameter estimation and superior fit. These advantages highlight the model's substantial value and its potential for widespread application in the analysis of complex count data. Moreover, the absence of any discernible trends or patterns in the cumulative residuals further supports the model's validity, suggesting that no significant misspecification issues exist. This reinforces the model's overall effectiveness and its ability to accurately represent the data.

## 5. Conclusion

This paper introduces a novel statistical distribution that combines zero-inflation with mixed NB distribution to establish the ZINB-L GLM framework. The innovative methodological integration significantly improves predictive accuracy in datasets characterized by zero-inflation. Operating within the generalized linear model paradigm, the study meticulously formulates the ZINB-L GLM regression approach. For parameter estimation, the research employs Bayesian techniques implemented through MCMC simulations, providing robust statistical inference for the proposed model. Empirical analysis of strike data and doctor visit records demonstrates the ZINB-L distribution superior capability in handling zero-inflated data compared to both NB-L and NB models. Cumulative residual plots confirm that ZINB-L GLM consistently delivers enhanced fit quality while preserving robust theoretical properties. Statistical evaluation metrics including Deviance, DIC, indicate the ZINB-L GLM regression framework outperforms conventional NB and NB-L regression approaches. The findings reveal that ZINB-L GLM exhibits exceptional performance not only with datasets characterized by zero-inflation and heavy-tailed distributions, but also in contexts with elevated sample means. The ZINB-L GLM framework will naturally converge to the NB-L GLM model (Lord and Geedipally, 2011), where there are relatively few zero values in the dataset, thereby ensuring that its worst-case performance is comparable to that of the NB-L GLM method. The ZINB-L distribution was originally conceived to address the analytical challenges posed by zero-inflated datasets, offering efficient modeling capabilities for count data characterized by excess zeros. Empirical evidence across both examined datasets confirms the ZINB-L substantive improvement over alternative approaches, with the deviance and DIC values establishing the clear hierarchy of model effectiveness: ZINB-L GLM demonstrating superior performance, followed by NB-L and NB regression frameworks. In summary, The ZINB-L GLM preserves the fundamental properties of traditional negative binomial approaches while significantly enhancing adaptability to both overdispersion and zero-inflation phenomena. Our findings demonstrate the ZINB-L GLM's capacity to effectively supersede conventional NB and NB-L GLM methodologies across diverse applied contexts. The advancement provides substantive theoretical foundations and methodological innovations for statistical inference in fields confronting the analytical challenges of zero-inflated count data, establishing the robust platform for future methodological developments in this domain.

## Declaration of competing interest

The authors declare that they have no known competing interest to report regarding the present study.

## References

- Altun, E. (2019). A new model for over-dispersed count data: Poisson quasi-Lindley regression model. *Mathematical Sciences*, 13(3), 241-247. doi: 10.1007/s40096-019-0293.
- Arnold, T., & Emerson, J. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2), 34–

9. doi: 10.32614/RJ-2011-016.
- Aryuyuen, S., & Bodhisuwan, W. (2013). The negative binomial-generalized exponential (NB-GE) distribution. *Applied Mathematical Sciences*, 7(22), 1093–105. doi: 10.12988/ams.2013.13099.
- Aryuyuen, S. (2021). Bayesian inference for the negative binomial-generalized Lindley regression model: properties and applications. *Communications in Statistics - Theory and Methods*, 52(13), 4534–4552. doi: 10.1080/03610926.2021.1995434.
- Croissant, Y., & Graves, S. (2020). Ecdat: Data sets for econometrics. R package version 0.3-9. <https://CRAN.R-project.org/package=Ecdat>.
- Dao, V. H., Gunawan, D., Kohn, R., Tran, M. N., Hawkins, G. E., & Brown, S. D. (2025). Bayesian inference for evidence accumulation models with regressors. *Psychological Methods*. doi: 10.1037/met0000669.
- Dzinyela, R., Shirazi, M., Das, S., & Lord, D. (2024). The negative Binomial-Lindley model with Time-Dependent Parameters: Accounting for temporal variations and excess zero observations in crash data. *Accident Analysis & Prevention*, 207, 107711. doi: 10.1016/j.aap.2024.107711.
- Feng, C. X. (2019). Zero-augmented accelerated spatial failure model for modeling hospital length of stay data. *Spatial and Spatio-temporal Epidemiology*, 29, 121–137. doi: 10.1016/j.sste.2018.05.001.
- Fu, S. (2015). A hierarchical Bayesian approach to negative binomial regression. *Methods and Applications of Analysis*, 22(4), 409–28. doi: 10.4310/MAA.2015.v22.n4.a4.
- Fung, T. C. (2024). Robust estimation and diagnostic of generalized linear model for insurance losses: a weighted likelihood approach. *Metrika*, 87(3), 333–366. doi: 10.1007/s00184-023-00917-6.
- Geedipally, S. R., Lord, D., & Dhavala, S. S. (2012). The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. *Accident Analysis and Prevention*, 45, 258–65. doi: 10.1016/j.aap.2011.07012
- Ghitany, M. E., Atieh, B., & Nadarajah, S. (2008). Lindley distribution and its application. *Mathematics and Computers in Simulation*, 78(4), 493–506. doi: 10.1016/j.matcom.2007.06.007.
- Lord, D., & Mannering, F. L. (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research – Part A*, 44(5), 291–305.
- Lord, D., & Geedipally, S. R. (2011). The negative binomial-Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. *Accident Analysis and Prevention*, 43(5), 1738–1742.
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book: A practical introduction to Bayesian analysis*. London: Chapman Hall.
- Moksony, F., & Hegedűs, R. (2014). The use of Poisson regression in the sociological study of suicide. *Corvinus Journal of Sociology and Social Policy*, 5(2), 97–114. doi: 10.14267/cjssp.2014.02.04.
- R Core Team (2020). R: A Language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Su, S., & Yajima, M. (2021). R2jags: Using R to Run JAGS. R package version 0.7-1. <https://CRAN.R-project.org/package=R2jags>.
- Yamrubboon, D., Thongteeraparp, A., Bodhisuwan, W., Jampachaisri, K., & Volodin, A. (2019). Bayesian inference for the negative binomial-Sushila linear model. *Lobachevskii Journal of Mathematics*, 40, 42–54. doi: 10.1134/S1995080219010141.
- Zamani, H., & Ismail, N. (2010). Negative binomial-Lindley distribution and its application. *Journal of Mathematics and Statistics*, 6(1), 4–9. doi: 10.3844/jmssp.2010.4.9.



© 2025 by the authors; licensee Growing Science, Canada. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).