Applied Mathematical Sciences, Vol. 19, 2025, no. 4, 153 - 171
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ams.2025.919236

Bayesian Inference for Over-dispersed Count Data with Excess Zeros

Cenyu Hu [1], Fang Ling [2], Xianming Shi [*,1] and Yalong Wang [1]

Corresponding author: Xianming Shi

[1] Army Engineering University of the PLA Shijiazhuang Campus
Hebei 050003, China

[2] Army Logistics Academy, Chongqing, China

Abstract

This article introduces the Negative Binomial-Lindley (NB-L) Model, a novel framework addressing over-dispersed count data with excess zeros. The NB-L distribution enhances flexibility in modeling complex count structures. Parameter estimation employs the Bayesian hierarchical framework with Markov Chain Monte Carlo (MCMC) simulations, overcoming limitations of traditional models in capturing intricate data patterns.Empirical validation using two real-world datasets—one with prominent zero inflation—shows the NB-L generalized linear

model (GLM) outperforms Poisson and Negative Binomial (NB) models in accuracy and robustness for datasets with high zero frequencies and long-tailed distributions. These results establish the NB-L model as the powerful tool for analyzing challenging count data across disciplines.

Keywords: negative binomial-Lindley distribution;over-dispersion; MCMC; bayesian inference; regression analysis

1. Introduction

Count data analysis serves as a cornerstone of statistical modeling across diverse scientific domains, including economics, epidemiology, public health, and social sciences [1]. Such data—characterized by negative integer outcomes representing event frequencies—often exhibit complex structures that challenge conventional statistical frameworks[2]. As highlighted by [3], the unique nature of count data necessitates specialized modeling techniques to accurately capture phenomena such as over-dispersion (variance exceeding mean), excess zeros, or temporal/spatial dependencies.

In economic research, for instance, count models are pivotal for analyzing firm innovation outputs or financial market event, where over-dispersion frequently arises from heterogeneous firm capabilities or market volatility[4]. In epidemiology, these models play a critical role in quantifying disease incidence rates while accounting for zero-inflated datasets—such as non-reporting of mild infections or sampling biases in health surveys[5]. Social science studies, meanwhile, leverage

count data modeling to investigate behavioral outcomes like criminal incidents or educational participation, where structural zeros and unobserved heterogeneity complicate inference [6].

The foundational challenge in count data analysis lies in balancing model flexibility with theoretical rigor. Traditional approaches, such as the Poisson Model, assume equidispersion (variance = mean) and often fail to accommodate real-world data complexities[7]. This limitation has spurred the development of advanced frameworks, including the Negative Binomial model for over-dispersed data and zero-inflated models for excess zerosb by[8]. However, even these extensions face constraints when addressing datasets with both pronounced over-dispersion and non-trivial zero proportions, motivating the need for more nuanced modeling strategies.

Against this backdrop, the present study contributes to the methodological frontier by introducing the Negative Binomial-Lindley GLM. By integrating the Lindley distribution—a flexible discrete distribution with heavy-tailed properties—into the Negative Binomial framework, the proposed model aims to enhance representation of count data with complex zero-inflation patterns and long-tailed frequency distributions[9]. Through a Bayesian hierarchical estimation approach using Markov Chain Monte Carlo (MCMC) simulations[10], this research seeks to demonstrate the NB-L model's superior performance in capturing intricate data structures compared to conventional count models[11].

2. Negative Binomial-Lindley Distribution Model

In this section, we introduce a novel hybrid modeling framework: the Negative Binomial-Lindley (NB-L) distribution. This innovative distribution is formulated by compounding the Negative Binomial (NB) distribution with the Lindley distribution. The resultant NB-L model is specifically engineered to provide a flexible and robust framework for analyzing count data, particularly for concurrently addressing two prevalent and challenging characteristics:

(ii) a high incidence of zero observations (often termed 'excess zeros'),

(ii) significant over-dispersion, which frequently manifests as long-tailed empirical distributions.

Firstly, we introduce the NB random variable as follows:

2.1 The Negative Binomial Distribution

It is worth noting that the NB distribution presents two classic parametric forms: the first one originates from the Poisson-Beta mixture process, and the second one comes from the limit form of a series of independent Bernoulli trials. Based on the mathematical derivation of the latter parameterization, its PMF can be expressed as:

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

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 prevalence of excessive zero-event occurrences and the observed marked heterogeneity in data distribution characterized by pronounced overdispersion, we implement a reparameterization strategy that expresses probability p as a function of the dispersion parameter r and is given as:

$$p=\frac{r}{\mu+r} \tag{3}$$

2.2 Generalized Linear Model

Within the framework of negative binomial generalized linear models (NB GLM), the conditional mean is modeled as a nonlinear function of the explanatory variables, with the expected value of the response variable linked to the covariates via a log link function, thereby expressing the systematic component as:

$$g(\mu_i)=\ln(\mu_i)=\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} \tag{4}$$

The conditional expectation $\mu=E(Y)$ is connected to the linear predictor through a log-link function, thereby establishing the generalized linear model (GLM) framework as：

$$E(Y)=\mu_i=exp(\beta_0+\beta_1 x_i 1+\cdots+\beta_j x_i+\cdots+\beta_p x_i p)=exp(x_i^T \mid \beta) \tag{5}$$

where μ_i the linear predictor, x_{ip} represents the vector of covariates, β_j corresponds to regression coefficients.

Based on the mathematical derivation framework from Eqs.(1) and (3), the pmf of the negative binomial distribution can be re-parameterized as a Poisson-Gamma mixture model. Based on (4),the PMF of the variables becomes

$$f(y_i) = \frac{\Gamma(y_i + r)}{y_i!\ \Gamma(r)} \int_0^{\infty} \left(\frac{(\lambda e^{(x_i^T \beta)})}{(\lambda e^{(x_i^T \beta)} + r)}\right)^{y_i} \left(\frac{r}{\lambda e^{(x_i^T \beta)} + r}\right)^r d\lambda \tag{6}$$

2.3 Lindley distribution

The GL distribution is a continuous probability density function defined by parameters a and b. The random variable is modeled using the Lindley distribution, which was initially proposed by Lindley (1958) and is a single-parameter distribution. It can be interpreted as a mixture of the exponential distribution and the gamma distribution. When a = 1 and b = 2, this function precisely corresponds to the standard Lindley distribution. The PMF of the distribution is defined as follows:

$$f(\lambda;\theta) = \frac{(1+\lambda)\ \theta^2}{1+\theta} e^{(-\theta\lambda)} \tag{7}$$

where θ denotes the scale parameter. This structural configuration endows the Lindley distribution with distinctive heavy-tailed characteristics, demonstrating superior performance in modeling stochastic phenomena exhibiting right-skewed patterns compared to conventional exponential distributions [12].

The MGF of λ can be obtained by calculation as

$$M_{\lambda}(t;\theta) = E[e^{t\lambda}] = \frac{\theta^2}{(1+\theta)(\theta-t)}(1+\frac{1}{(\theta-t)}); t > 0 \quad (8)$$

In addition, the first and second moments of the lindley distribution are as follows:

$$E(\lambda) = \frac{1}{(1+\theta)}(1+\frac{2}{\theta}),\ and\ E(\theta^2) = \frac{(3+2\theta)}{(\theta^2(1+\theta))} \quad (9)$$

The integration of NB and Lindley distributions within a mixture framework enables the establishment of a GLM structure. The derivation of the NB-L GLM Model is as follows:

$$f(y_i;\mu_i,r,\theta) = \int_0^{\infty} NB(y_i;r,\lambda\mu_i)lindley(\lambda;\theta)d\lambda$$

where parameter λ is used as the parameter of Lindley distribution in the PDF of Eq.(7). Based on Eq.(1), (2) and (6), The pmf of the NB-L GLM distribution can be defined as follows:

$$f(y_i;\mu_i,r,\theta) = \frac{\Gamma(y_i+r)}{y_i!\ \Gamma(r)(1+\theta)}\int_0^{\infty} (\frac{(\lambda e^{(x_i^T\beta)})}{(\lambda e^{(x_i^T\beta)}+r)})^{(y_i)}(\frac{r}{\lambda e^{(x_i^T\beta)}+r})^r e^{-\theta\lambda}d\lambda \quad (10)$$

where $y_i = 0,1,2,\ldots,\mu_i > 0, i = 1,2,\ldots,n$ and the positive parameters r,θ.Its mean and variance are respectively:

$$E(Y_i;\mu_i,r,\theta) = \mu_i E(\lambda) = exp(\beta_0 + \sum_{i=1}^{\beta} X_i)\frac{\theta+2}{\theta(\theta+2)} \quad (11)$$

$$Var(Y_i;\mu_i,r,\theta) = E(Y_i;\mu_i,r,\theta) + \mu_i^2 E(\lambda^2)\frac{1+r}{r} - E^2(Y_i;\mu_i,r,\theta) \quad (12)$$

By integrating prior knowledge with observational data, the Bayesian framework

not only can cope with highly uncertain situations, but also can dynamically update parameter estimation when new information is introduced. Thus, it demonstrates strong applicability in the modeling and analysis of complex systems [13].

Let $\Omega = (r, \theta, \beta)^T$ be the vector of the regression parameter. The likelihood function of Ω is

$$L(\Omega \mid y, x^T) = \Pi_{i=1}^{n} \frac{(\Gamma(y_i + r)\theta^2)}{y_i!\Gamma(r)(1+\theta)} \int_0^{\infty} \left(\frac{r}{\lambda e^{(x_i^T \beta)} + r}\right)^r \left(\frac{\lambda e^{(x_i^T \beta)}}{\lambda e^{(x_i^T \beta)} + r}\right)^{y_i} (1+\lambda) e^{(-\theta\lambda)} d\lambda \quad (13)$$

3. Bayesian Inference for NB-L GLM Model

3.1 Prior distributions and joint posterior density

This method adopts the Bayesian statistical framework, which systematically integrates prior information through the setting of probability distributions and takes all unknown parameters into account.Assuming that the parameters r, θ of the NB-L GLM follow gamma distribution, while β follow normal distribution, and all parameters are independent of each other. Then the joint prior distribution of the unknown parameters are as following

$$r \sim gamma(\alpha_r, z^r), \theta \sim gamma(\alpha_\theta, z^\theta), \beta \sim N(v_0, \sigma_\beta) \quad (14)$$

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

According to Bayes theorem, the posterior distribution is determined by multiplying the likelihood function by the prior distribution. The posterior distribution obtained is

$$\pi(\Omega \mid X) \propto L(\Omega \mid y, X)\pi(r)\pi(\theta)\pi(\beta) \tag{15}$$

For this complex model, the parameters of each component can be calculated. The complete posterior distributions of the parameters of Ωderived are all obtained

$$\begin{aligned} &\pi(r \mid y, X, \theta, r) \propto L(\Omega \mid y, X)\pi(r) \\ &\pi(\theta \mid y, X, \theta, r) \propto L(\Omega \mid y, X)\pi(\theta) \\ &\pi(\beta \mid y, X, \theta, r) \propto L(\Omega \mid y, X)\pi(\beta) \end{aligned} \tag{16}$$

3.2 Model evaluation

Based on the Bayesian framework, we adopt the MCMC method to conduct posterior inference on the model parameters. The observed values follow the NB distribution, and the site-specific fragility terms follow a gamma prior distribution.

The MCMC method transforms the complex high-dimensional posterior distribution sampling problem into a simple conditional distribution sampling problem, constructing a Markov chain with the stationary distribution as the target posterior distribution. It demonstrates significant advantages in high-dimensional integration and estimation problems.

This part adopts the Gibbs sampling technique by decomposing the joint update of high-dimensional parameters into the successive updates of individual parameter. In the specific implementation, we generated three parallel independent MCMC chains, each of which conducted 30,000 iterations and discarded the first 15,000

iterations as the aging period to ensure convergence and calculated the expected posterior values of the parameters to achieve a robust estimation of the target posterior distribution.

In the model selection analysis, this study adopts three core evaluation indicators: deviance, DIC, and the number of effective parameters (p_D). Among them, DIC is a generalized extension form of AIC and BIC , and has become a standardized tool for evaluating the goodness of fit of Bayesian model.

It is particularly suitable for model comparison studies based on posterior distributions obtained through MCMC simulation

4. Empirical applications

4.1 Data description

The dataset used can be freely accessed through the Ecdat package of R language [14].The dataset used in this study is derived from the pioneering work of [15], and it can be freely obtained through the R language econometrics analysis package Ecdat. 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.

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

Variables	Min	Median	Max	Average (std. dev)
strikes	0	5	18	5.24（3.75）
output	-0.13996	-0.00013	0.08554	-0.003（0.05456）

4.2 Bayesian inference for the NB-L GLM Distribution

In this section, we present the analysis results of the NB-L GLM distribution and its Model.

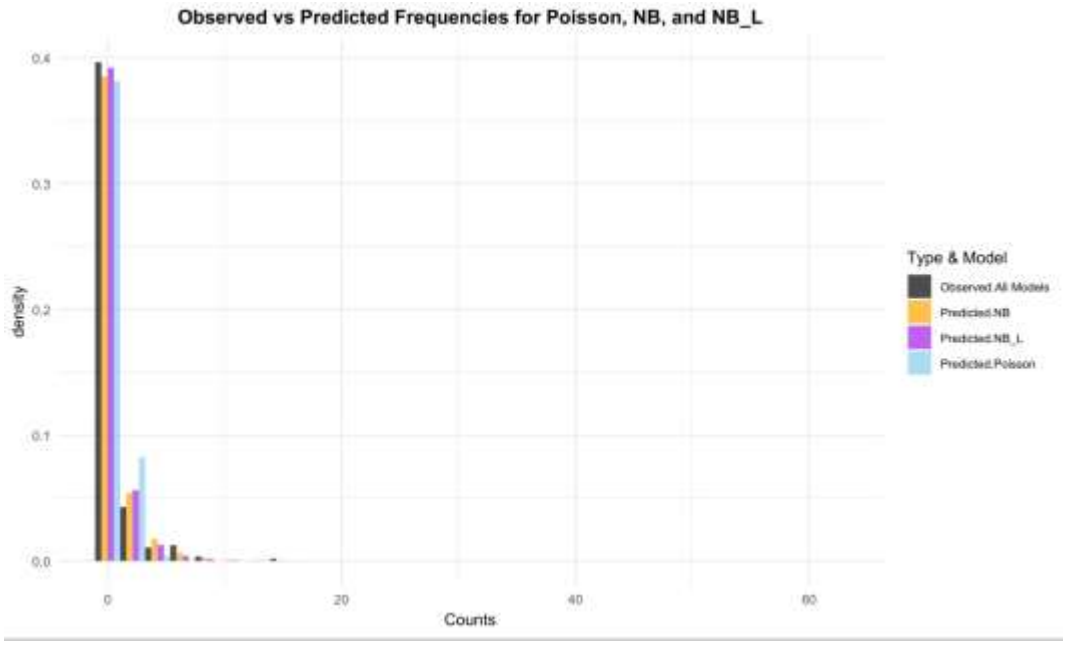


Figure 1 The bar chart plots of observed frequency and expected frequencies of each distributions.

We evaluated the fitting performance of the NB-L GLM distribution for the 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 [16], Deviance, and DIC. The distribution that provided the best fit was identified by minimizing the KS statistic, Deviance, and DIC values.

The results for the "Strikes" dataset in Table 3 show that both the NB distribution and the NB-L GLM distribution fit the data well. While the NB-L GLM 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 NB-L GLM distribution is a suitable model for this dataset, yielding a fit comparable to that of the NB distribution.

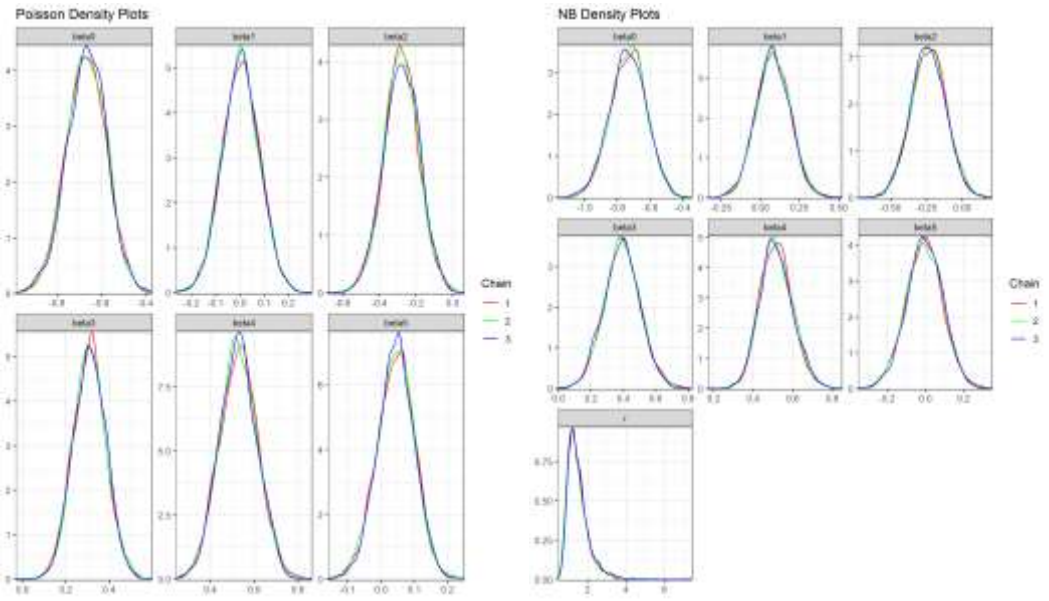


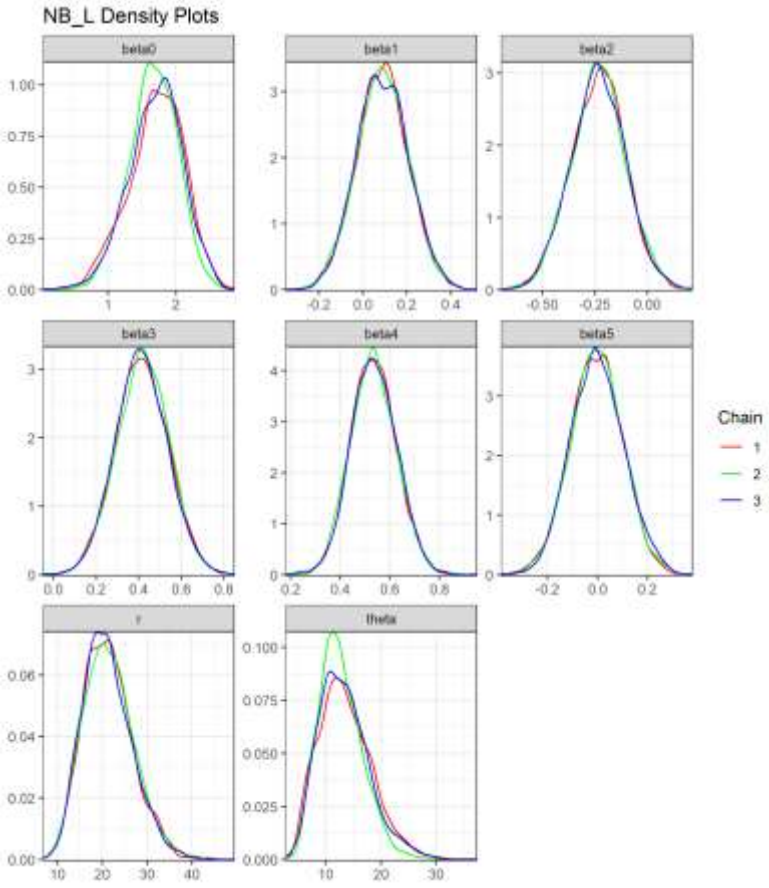
Figure 2 Density plots of the three MCMC chains for r, θ, and $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ from the Possion model and NB model for the strikes data.

Additionally, to enhance the fitting effect, the NB-L GLM distribution needs to be adjusted by incorporating the mean of the GL distribution as per Eq.(8).

Table 2 Posterior Distribution Summaries for NB, NB-L GLM, and NB-L GLM distribution for Strike Data.

Parameter	Possion		NB		NB-L	
	Mean(s.e.)	95%Cr.I.	Mean(s.e.)	95%Cr.I.	Mean(s.e.)	95%Cr.I.
p	0.36(0.05)	(0.26,0.48)	—	—-	—	—-
r	3.05(0.71)	(1.91,4.67)	25.92(6.02)	(15.98,39.19)	31.28(5.57)	(21.42,43.27)
θ	—	—-	6.56(1.31)	(4.33,9.55)	7.89(2.04)	(5.55,10.78)
Deviance	568.56		563.27		520.77	
DIC	570.53		557.59		550.25	
KS	0.1574		0.1966		0.1957	
p_d	0.6587		0.6787		0.6773	

Figure 3 Density plots of the three MCMC chains for r, θ, and $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ from the NB model for the strikes data.



Applying the Bayesian method, monitoring and reporting the convergence of the algorithm is a key aspect for ensuring the reliability of the results. We utilize trace plots and posterior density plots to analyze the sampling distribution of parameter values, in order to examine the stationarity and convergence of the MCMC chain. These plots visually reflect the stability of the model during the sampling process, thereby ensuring the accuracy of the inference results.

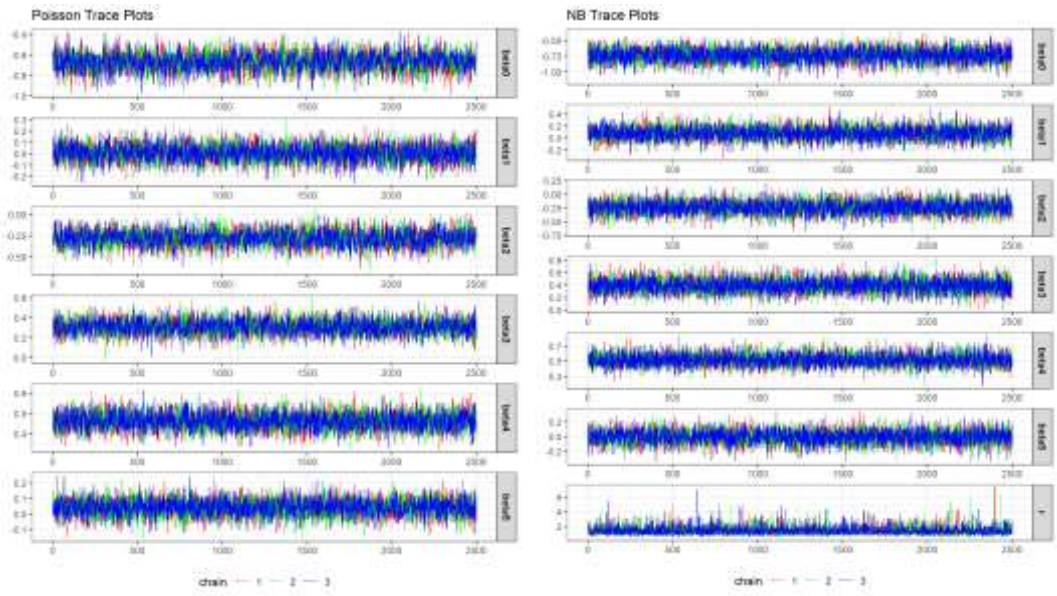


Figure 4 Trace plots of the three MCMC chains for r, θ, and $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_5)^T$ from the Possion model and NB model for the data.

Figures 2 and 3 respectively illustrate the posterior density distributions of all parameters of the Poisson, NB, and NB-L models. From the results, it can be seen that after the burning period, the posterior densities of the three parallel chains achieved a high degree of overlap, indicating that the posterior distribution samples of the parameter estimation have good representativeness and consistency.

At the same time, the trajectory graphs in Figures 4 and 5 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.parameter distribution.

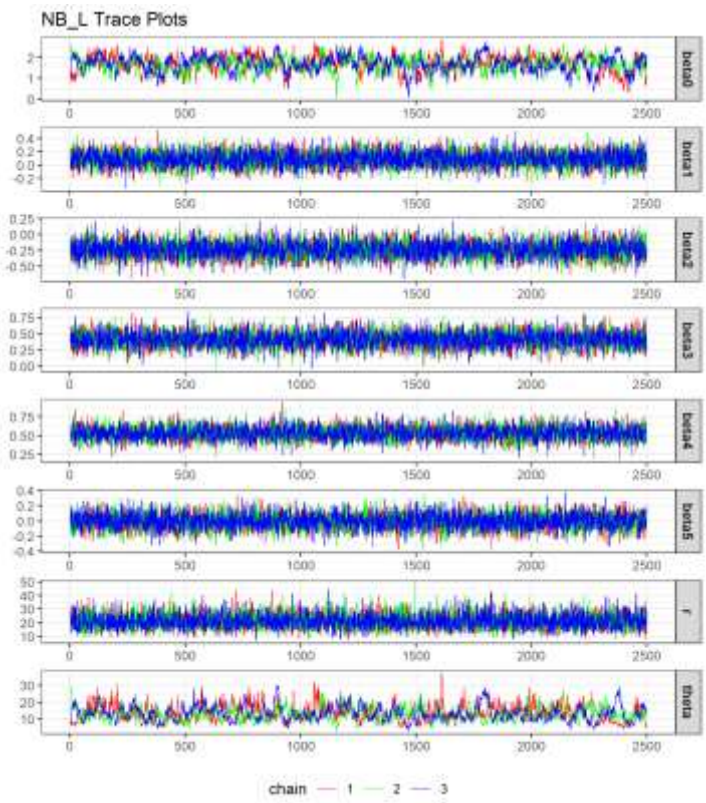


Figure 5 Trace plots of the three MCMC chains for r, θ, and $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_5)^T$ from the NB-L model for the data.

To verify the applicability of the NB-L GLM model, we further evaluated the model performance by analyzing residual density plots. These residual density 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.

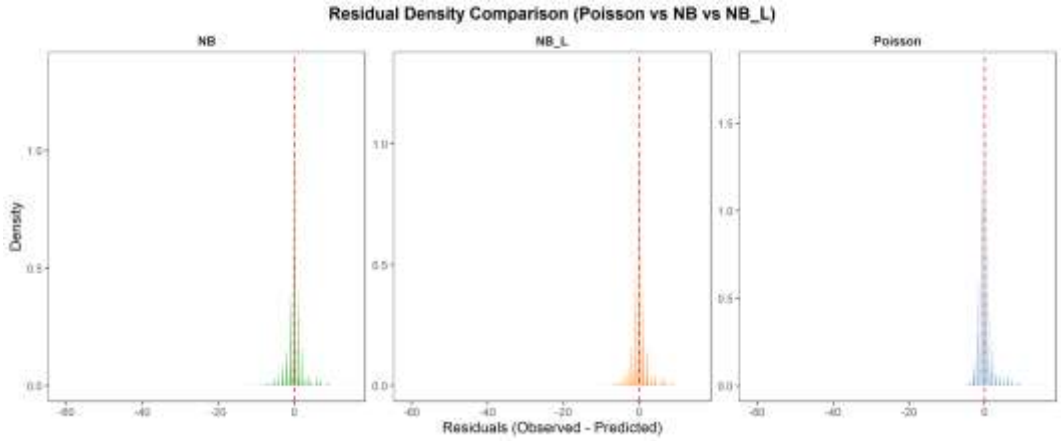


Figure 6 Residual density comparison plots

The Residual density plots analysis demonstrates that the NB-L exhibits exceptional performance in fitting the dataset, outperforming both the Possion and standard NB models. The NB-L model 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 lindley distribution with mixed negative binomial distribution to establish the NB-L framework. This innovative methodological integration significantly improves predictive accuracy in datasets characterized by zero-inflation.Notably, when datasets contain minimal zero values, the NB-L framework naturally converges to the NB model, ensuring that its worst-case performance remains equivalent to the NB approach.

Empirical evidence across both examined datasets confirms the NB-L substantive improvement over alternative approaches, with deviation metrics and DIC values establishing a clear hierarchy of model effectiveness: NB-L demonstrating superior performance, followed by NB and conventional Possion frameworks. In summary, The NB-L model preserves the fundamental properties of traditional negative binomial approaches while significantly enhancing adaptability to both overdispersion and zero-inflation phenomena.

References

[1] Altun, E. A new model for over-dispersed count data: Poisson quasi-Lindley regression model, *Mathematical Sciences*, **13** (2019), no. 3, 241–247. https://doi.org/10.1007/s40096-019-0293-5

[2] Moksony, F., and Hegedűs, R.. The use of Poisson regression in the sociological study of suicide, *Corvinus Journal of Sociology and Social Policy*, **5** (2014), no. 2, 97–114. https://doi.org/10.14267/cjssp.2014.02.04

[3] Lord, D., Mannering, F. L. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives, *Transportation Research – Part A: Policy and Practice*, **44** (2010), no. 5, 291–305. https://doi.org/10.1016/j.tra.2010.02.001

[4] Fung, T. C. Robust estimation and diagnostic of generalized linear model for insurance losses: a weighted likelihood approach, *Metrika*, **87** (2024), no. 3, 333–366. https://doi.org/10.1007/s00184-024-00952-6

[5] Feng, C. X. Zero-augmented accelerated spatial failure model for modeling hospital length of stay data, *Spatial and Spatio-temporal Epidemiology*, **29** (2019), 121–137. https://doi.org/10.1016/j.sste.2018.05.001

[6] Hilbe, J. X *Modeling Count Data*, New York: Cambridge University Press, 2014.

[7] Aryuyuen, S., and W. Bodhisuwan. The negative binomial-generalized exponential (NB-GE) distribution, *Applied Mathematical Sciences*, **7** (2013), no. 22, 1093–105. https://doi.org/10.12988/ams.2013.13099

[8] Aryuyuen, S. Bayesian inference for the negative binomial-generalized Lindley regression model: properties and applications, *Communications in Statistics - Theory and Methods*, **52** (2021), no. 13, 4534–4552. https://doi.org/10.1080/03610926.2021.1995434

[9] Yamrubboon, D., Thongteeraparp, A., Bodhisuwan, W., et al. Bayesian Inference for the Negative Binomial-Sushila Linear Model, *Lobachevskii Journal of Mathematics*, **40** (2019), 42–54. https://doi.org/10.1134/s1995080219010141

[10] Lord, D., Geedipally, S. R. The negative binomial-Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros, *Accident Analysis and Prevention*, **43** (2011), no. 5, 1738–1742. https://doi.org/10.1016/j.aap.2011.04.004

[11] Zhang, S., et al. Research on Amplification Algorithm of Small Sample Soil Composite Data Based on Probability Distribution, *Journal of Physics: Conference Series*, (2024), 2872:012009. https://doi.org/10.1088/1742-6596/2872/1/012009

[12] Ghitany, M. E., Atieh, B., and Nadarajah, S. Lindley distribution and its application *Mathematics and Computers in Simulation*, 78 (2008), no. 4, 493–506. https://doi.org/10.1016/j.matcom.2007.06.007

[13] Yamrubboon, D., Thongteeraparp, A., Bodhisuwan, W., et al. Bayesian Inference for the Negative Binomial-Sushila Linear Model, *Lobachevskii Journal of Mathematics*, **40** (2019), 42–54. https://doi.org/10.1134/s1995080219010141

[14] Croissant, Y., and S. Graves. 2020. Ecdat: Data sets for econometrics. R package version 0.3-9. https://CRAN.R-project.org/package=Ecdat

[15] R Core Team. 2020. R: A Language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
https://www.R-project.org/

[16] Arnold, T., and J. Emerson. Nonparametric goodness-of-fit tests for discrete null distributions, *The R Journal*, **3** (2011), no. 2, 34–9.
https://doi.org/10.32614/rj-2011-016

Received: April 30, 2025; Published: June 2, 2025