Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2023, 2023:50
https://doi.org/10.28919/cmbn/7933
ISSN: 2052-2541

# A FOUR-PARAMETER NEGATIVE BINOMIAL-LINDLEY REGRESSION MODEL TO ANALYZE FACTORS INFLUENCING THE NUMBER OF CANCER DEATHS USING BAYESIAN INFERENCE

UNCHALEE TONGGUMNEAD, KITTIPONG KLINJAN, EKAPAK TANPRAYOON, SIRINAPA ARYUYUEN*

Department of Mathematics and Computer Science, Rajamangala University of Technology Thanyaburi,
Pathum Thani, 12110, Thailand

**Abstract:** In this paper, factors influencing the number of cancer deaths in Thailand that is a heavy-tailed data with overdispersion, were analyzed. A new mixed negative binomial (NB) regression model derived from a four-parameter negative binomial-Lindley (NBL) distribution called a four-parameter NBL regression model was developed, with number of cancer deaths analyzed using different factors. Factor importance affecting the number of cancer deaths was also considered to construct an optimal model describing the number of cancer deaths in Thailand. The four-parameter NBL, NB and Poisson regression models were used to describe the data, with parameters in each model estimated using the Bayesian approach. Results showed that the four-parameter NBL model had the highest efficiency compared to the NB and Poisson models. The number of cancer deaths in Thailand was influenced by population size in each province at midyear 2021, province population per doctor, percentage of poor people in each province, number of deaths from cancer caused by smoking behavior from age 15 years and over and number of deaths from cancer as a result of drinking behavior from age 15 years and over.

*Corresponding author
E-mail address: sirinapa_a@rmutt.ac.th

Received March 01, 2023

**Keywords:** mixed NB model; Bayesian inference; generalized linear model; count data regression model; cancer deaths.

**2020 AMS Subject Classification:** 97K80, 91G70.

## 1. INTRODUCTION

Statistical models are essential in engineering, medical, biological science, management and public health fields, providing helpful information to draw conclusions and make decisions. Linear models describe a continuous response variable as a function of one or more predictor variables. Although widely used, linear models have certain limitations; for example, the dependent variable must be continuous or only quantitative, discrepancies are assumed to be normally distributed and each observation is independent of other variables. Generalized linear models (GLM) as more flexible, were developed from the general linear model to offer better coverage. A general linear model can be described as a linear regression model for continuous response variables that defines continuous or categorical predictors. For a count response variable, the general linear model is not appropriate because the dependent variable must be continuous or only quantitative. Therefore, the GLM is used to model discrete response variables. Several statistical models have been developed to better understand the count response variable such as the Poisson (Pois) and negative binomial (NB) models. The Pois model is an appropriate choice for repeated count data. However, this model is not realistic because of the restriction that the mean and variance are equal, while the NB model effectively manages overdispersion of longitudinal data [1] and also overcomes constraints of problematic distribution of count data with overdispersion [2-5]. The NB distribution is proper for count data when there is an overdispersion problem, without necessarily being heavy-tailed but heavy-tailed distributions often present overdispersion [6].

Later, new distributions were developed to provide more flexibility and coverage. One of the most widely used is the mixed NB distribution. Many mixed NB distributions have been introduced such as the NB-Lindley [7], NB-beta exponential [8], NB-generalized exponential [9], NB-gamma [10], NB-Sushila [11] and NB-generalized Lindley [12]. Recently, [13] proposed a new mixed NB distribution, namely a four-parameter negative binomial-Lindley (NBL) distribution for describing over and underdispersed count data. Mixed NB distributions are applied to statistical model events for count data in real life such as actuarial and insurance models [7, 10, 14, 15], medical or industrial models [14] or in the fields of ecology and biodiversity [16-18]. Parameters in mixed NB regression models were estimated using the Bayesian framework as a more flexible approach than the maximum likelihood estimation [19-20]. The difference between the maximum likelihood and Bayesian methods is that the underlying parameters are considered random variables characterized by a prior distribution [21]. Bayesian inference for mixed NB models has also been studied such as NB-Lindley [22], NB-generalized exponential [23], NB-Sushila [24], NB-generalized Lindley [12], NB-Quasi Lindley [17], NB-modified Quasi Lindley [16] and exponential [21] linear regression models.

This article analyzed factors influencing the number of cancer deaths in Thailand using the regression model for one mixed NB distribution, namely a four-parameter negative binomial-Lindley (NBL) distribution proposed by [13]. Their results showed that the four-parameter NBL distribution outperformed the Pois and NB distributions when fitting count data with overdispersion and a large number of zeros. However, the four-parameter NBL distribution has never been developed as a regression model. Therefore, this study applied the four-parameter NBL distribution under the GLM framework. Parameters of the proposed regression model were estimated using the Bayesian approach to compare the efficiency against some traditional regression models.

The rest of this paper is organized as follows. Section 2 presents an overview of the Pois, NB and four-parameter NBL distributions and also describes the generalized linear regression model. Section 3.1 develops a regression model derived from the four-parameter NBL distribution, while Section 3.2 addresses the Bayesian approach for parameter estimation and lists model selection criteria. Section 3.3 summarizes and analyzes the numerical application results using a real dataset of cancer deaths in Thailand, illustrating the practicability of the four-parameter NBL regression model. Finally, Section 4 presents the conclusions.

## 2. PRELIMINARIES

This section introduces an overview of traditional distributions for count data such as the Pois, NB and four-parameter NBL distributions. The generalized linear regression model is also described.

### 2.1 The Pois distribution

Let $Y$ be a random variable following the Pois distribution with parameter $\mu,$ denoted by $Y$ ~Pois$(\mu)$. Then its probability mass function (pmf) is

$$f(y;\mu)=\frac{\exp(-\mu)\mu^{y}}{y!},\quad y=0,1,2,\ldots \text{ and } \mu>0. \tag{1}$$

Its mean and variance are, respectively

$$\mathrm{E}(Y)=\mu \text{ and } \mathrm{Var}(Y)=\mu.$$

### 2.2 The NB distribution

Let $Y$ be a random variable following as the NB distribution with parameters $r$ and $p$, denoted by $Y$ ~NB$(r,p)$. Then its pmf is given by

$$f(y;r,p)=\binom{y+r-1}{r}p^{r}(1-p)^{y};\quad y=0,1,2,\ldots,\quad r>0 \text{ and } 0<p<1. \tag{2}$$

Its mean and variance are, respectively

$$\mathrm{E}(Y)=r\left(\frac{1-p}{p}\right) \text{ and } \mathrm{Var}(Y)=r\left(\frac{1-p}{p^{2}}\right).$$

For $Y\sim$ NB$(r,p)$, the parameter $p$ can be represented in terms of $r$ as $p=\frac{r}{\mu+r},$ where $\mu$ is the mean response and $r$ is the inverse of the dispersion parameter. The pmf of $Y$ can be written as follows:

$$f(y;r,p)=\frac{\Gamma(r+y)}{\Gamma(r)\Gamma(1+y)}\left(\frac{r}{\mu+r}\right)^{r}\left(\frac{\mu}{\mu+r}\right)^{y};\quad y=0,1,2,\ldots \tag{3}$$

where $\mu\geq 0,$ $r>0$ and $\Gamma(\cdot)$ denotes a complete gamma function. Then, its mean and variance can be defined as:

$$\mathrm{E}(Y)=\mu \text{ and } \mathrm{Var}(Y)=\mu+\left(\frac{1}{r}\right)\mu^2.$$

### 2.3 The four-parameter NBL distribution

The four-parameter NBL distribution was proposed by [13] in 2022 as a mixture between the NB distribution and a three-parameter Lindley ($L_3$) distribution with parameters $a,$ $b$ and $c$, i.e. $Y \mid r,\lambda \sim \mathrm{NB}(r, p=\exp(-\lambda))$ and $\lambda \sim L_3(a,b,c)$. The $L_3$ distribution was proposed by [25] with a probability density function (pdf) as follows:

$$g(\lambda;a,b,c)=\frac{c^2}{ac+b}\left(a+b\lambda\right)\exp(-c\lambda), \text{ for } \lambda>0, \tag{4}$$

where $a>0,$ $b>0,$ and $c>0.$

Let $Y$ be a random variable following the four-parameter NBL distribution with parameters $r,$ $a,$ $b$ and $c.$ Then its pmf is given by

$$f(y;r,a,b,c)=\binom{y+r-1}{y}\sum_{j=0}^{y}\binom{y}{j}(-1)^j\frac{\left[a(c+r+j)+b\right]c^2}{(ac+b)(c+r+j)^2}, \tag{5}$$

where $y=0,1,2,...,$ $r>0,$ $a>0,$ $b>0$ and $c>0$. Its mean and variance are, respectively

$$\mathrm{E}(Y)=\frac{r}{\delta_0}\left(\delta_1-\delta_0\right), \text{ and } \mathrm{Var}(Y)=\frac{r}{\delta_0}\left[(r+1)(\delta_2-\delta_1)-r(\delta_1-\delta_0)-\frac{r}{\delta_0}(\delta_1-\delta_0)^2\right],$$

where $\delta_k=\dfrac{a(c-k)+b}{(c-k)^2}$ for $k=1,2,3$ and $c\neq 0,1,2$.

The four-parameter NBL distribution is versatile as it nests several distributions when specific parameters are fixed [13]. These special cases are (i) a three-parameter NBL distribution (for $a=1$) proposed by [26] and (ii) a NBL distribution (for $a=b$) proposed by [7].

### 2.4 The generalized linear regression model

The GLM, originally introduced by [27], allows modeling of a wide range of probability distributions for response variables such as binomial, Pois and exponential distributions. The difference between a traditional linear regression model and the GLM is that a response variable in a GLM is related to the linear predictor through a link function rather than being assumed to be normally distributed. This link function allows for modeling non-normal response variables while using a linear combination of the predictor variables.

The GLM consists of three main components as follows:

(1) A random component which specifies that the conditional distribution of a response variable $Y_i$, for the $i$ th of $n$ independently sampled observations, is given values of explanatory variables with the mean $\mathrm{E}(Y_i) = \mu_i$ for $i = 1, \ldots, n$. A response variable $Y_i$ is assumed to follow a certain probability distribution such as a binomial, Pois or NB. The choice of distribution depends on the type of response variable and the nature of the data.

(2) A systematic component that specifies a linear predictor, which is a linear combination of the explanatory variables with $k$ predictors, denoted by $X_{ik}$, replaced by a linear predictor, $\eta_i$. The relationship is assumed as $\eta_i = \beta_1 + \beta_2 X_{i1} + \beta_3 X_{i2} + \cdots + \beta_{k+1} X_{ik}$, where $\beta_1, \beta_2, \ldots, \beta_{k+1}$ are unknown regression coefficients to be estimated.

(3) A link function that connects the linear predictor to the mean of $Y_i$. This is a monotonically increasing function, and the choice of link function depends on the distribution of $Y_i$. When $Y_i$ is a positive integer, $\mathrm{E}(Y_i) = \mu_i$ is also non-negative and the log-linearity for the mean is used as a link function. The log-link maps $\mu_i$ to the whole real line. Thus, the link function is $\log(\mu_i)$ that relates $\mu_i$ to the linear predictors. In the regression model, the log-linearity for the mean is commonly used as a link function: $\log(\mu_i) = \beta_1 + \beta_2 X_{i1} + \beta_3 X_{i2} + \cdots + \beta_{k+1} X_{ik}$. Its corresponding inverse transformation is

$$\mu_i = \exp(\beta_1 + \beta_2 X_{i1} + \beta_3 X_{i2} + \cdots + \beta_{k+1} X_{ik}) = \exp(\mathbf{X}_i^T \boldsymbol{\beta}), \tag{6}$$

where $\mathbf{X}_i^T = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$ is a vector of length $(k+1)$ where the $i$ th row of the $n \times (k+1)$

matrix $\mathbf{X}$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_{k+1})^T$ is a $(k+1)\times 1$ are unknown vectors of the regression coefficients.

**The Pois regression model**: The regression model for $Y_i$ distributed as the Pois distribution can be rewritten to show its pmf as follows:

$$f(y_i;\mu_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}; \; y_i = 0,1,2,... \text{ and } \mu_i > 0. \tag{7}$$

The mean and variance of $Y_i$ are, respectively

$$\mathrm{E}(Y_i) = \mu_i \quad \text{and} \quad \mathrm{Var}(Y_i) = \mu_i.$$

**The NB regression model**: Let $Y_i$ follow a NB distribution with the pmf as (3). Then the traditional NB distribution [2] can be rewritten to show its pmf as:

$$f(y_i;r,\mu_i) = \binom{y_i + r - 1}{y_i}\left(\frac{r}{\mu_i + r}\right)^r \left(\frac{\mu_i}{\mu_i + r}\right)^{y_i}; \; y_i = 0,1,2,.... \tag{8}$$

For $\mu_i > 0$ and $r > 0$, its mean and variance are, respectively

$$\mathrm{E}(Y_i) = \mu_i \quad \text{and} \quad \mathrm{Var}(Y_i) = \mu_i + \left(\frac{1}{r}\right)\mu_i^2.$$

## 3. Main Results

### 3.1 Developing a regression model for the four-parameter NBL distribution

This section develops a new mixed NB regression model derived from the four-parameter NBL distribution called the four-parameter negative binomial-Lindley regression model to describe a count response. The framework of the GLM to derive the four-parameter NBL model involves a mixture of the NB and L3 distributions as:

$$f(y_i;\boldsymbol{\theta}) = \int_0^\infty f_{\mathrm{NB}}(y_i;r,\lambda\mu_i)\, f_{\mathrm{L}_3}(\lambda;a,b,c)d\lambda, \tag{9}$$

where $\boldsymbol{\theta} = (\mu_i, r, a, b, c)^T$ and the mean response $\mu_i$ is a similar parameter to a label in equation

(6). Based on equations (3) and (4), the pmf of the four-parameter NBL distribution becomes:

$$f(y_i;\boldsymbol{\theta})=\int_0^\infty \binom{y_i+r-1}{y_i}\left(\frac{r}{\lambda\mu_i+r}\right)^r\left(\frac{\lambda\mu_i}{\lambda\mu_i+r}\right)^{y_i}\frac{c^2\left(a+b\lambda\right)\exp(-c\lambda)}{ac+b}\,\mathrm{d}\lambda\,. \tag{10}$$

Suppose that a count response variable $Y_i$ and $\mathbf{X}_\mathbf{i}^\mathbf{T}$ are a set of covariates. The conditional distribution of $Y_i\,|\,\mathbf{X}_\mathbf{i}^\mathbf{T}$ can be written in the linear regression model as:

$$f(y_i\,|\,\mathbf{x}_i^T)=\int_0^\infty \binom{y_i+r-1}{y_i}\left(\frac{r}{\lambda\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)+r}\right)^r\left(\frac{\lambda\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}{\lambda\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)+r}\right)^{y_i}$$

$$\times\frac{c^2\left(a+b\lambda\right)\exp(-c\lambda)}{ac+b}\,\mathrm{d}\lambda, \tag{11}$$

where $\mathbf{y}=(y_1,y_2,...,y_n)^T$ is a $(n\times1)$ vector of response variables, which are $n$ independent realizations of the four-parameter NBL regression model and $\boldsymbol{\Omega}=(r,a,b,c,\boldsymbol{\beta}^T)^T$ is a vector of the regression parameters. Thus, the likelihood function of $\boldsymbol{\Omega}$ becomes:

$$L(\boldsymbol{\Omega}\,|\,\mathbf{y},\mathbf{x})=\prod_{i=1}^n\binom{y_i+r-1}{y_i}\int_0^\infty\left\{\left(\frac{r}{\lambda\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)+r}\right)^r\left(\frac{\lambda\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}{\lambda\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)+r}\right)^{y_i}\right.$$

$$\left.\times\frac{c^2\left(a+b\lambda\right)\exp(-c\lambda)}{ac+b}\right\}\mathrm{d}\lambda\,. \tag{12}$$

If $\mu_i=\exp(\mathbf{x}_i^T\boldsymbol{\beta})$, the mean of the response can be calculated using the conditional expectation as follows:

$$\mathrm{E}(Y_i\,|\,\mathbf{x}_i^T)=\mu_i\mathrm{E}(\lambda)\ \text{ and }\ \mathrm{Var}(Y_i\,|\,x_i^T)=\mu_i\mathrm{E}(\lambda)+\mu_i^2\left(\frac{1+r}{r}\right)\mathrm{E}(\lambda^2)-\left[\mu_i\mathrm{E}(\lambda)\right]^2, \tag{13}$$

where $\mathrm{E}(\lambda)$ and $\mathrm{E}(\lambda^2)$ are the first and second moments about the original L3 random variable [25], i.e.,

$$\mathrm{E}(\lambda)=\frac{ac+2b}{c(ac+b)}\ \text{ and }\ \mathrm{E}(\lambda^2)=\frac{2ac+6b}{c^2(ac+b)}.$$

### 3.2 Bayesian inference for the four-parameter NBL regression model

The vector of unknown parameters $\mathbf{\Omega}$ in equation (12) was estimated using the Bayesian approach, which considers prior information for parameter estimation. The Bayesian approach was implemented using a hierarchical Bayesian modeling approach relying on Markov Chain Monte Carlo (MCMC) techniques [23, 28] for the four-parameter NBL regression model. Accordingly, under a squared error loss function, the Bayesian estimator of $\mathbf{\Omega}$ will be $\mathrm{E}\left(\mathbf{\Omega} \mid y_i\right)$. Statistical efficiency programs for Bayesian analysis are available such as OpenBUGS [29-30], while techniques of Bayesian inference can be extended to hierarchical Bayesian analysis. However, most researchers are interested in studying the hierarchical Bayesian modeling approach that estimates model parameters in the regression model for a count response variable [1, 12, 16, 21, 24, 31].

The likelihood function of the four-parameter NBL regression model in equation (12) is not a closed form and can be executed using the representation of the hierarchical model implicit both in the integral and the definition of the L3 distribution. The L3 distribution is a mixture of exponential (Exp) and gamma (Gam) distributions, i.e., $\mathrm{Exp}(c)$ and $\mathrm{Gam}(2,c)$, therefore, the pdf of the L3 distribution in equation (3) can be written [25] as:

$$\lambda \sim \frac{ac}{ac+b}\mathrm{Exp}(c) + \frac{b}{ac+b}\mathrm{Gam}(2,c). \tag{14}$$

The four-parameter NBL distribution is conditional upon the unobserved site-specific frailty term $\lambda$ in equation (11), which describes the additional heterogeneity [22]. Consequently, the hierarchical framework can be represented as:

$$f(y_i;\mu_i,r \mid \lambda) = \mathrm{NB}(y_i;\lambda\mu_i,r);\quad \mu_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta}) \quad \text{and} \quad \lambda \sim \mathrm{L}_3(a,b,c). \tag{15}$$

The unknown parameters or $r,a,b,c$ and $\boldsymbol{\beta}$ in equation (12) are considered. If the parameters $r$, $a$, $b$ and $c$ of the four-parameter NBL regression model follow a gamma distribution and $\boldsymbol{\beta}$ is distributed as the normal distribution, they are mutually independently distributed in each parameter, and the joint prior distribution of all unknown parameters can be written as: $r \sim \mathrm{Gam}(\gamma_r,\tau_r)$, $a \sim \mathrm{Gam}(\gamma_a,\tau_a)$, $b \sim \mathrm{Gam}(\gamma_b,\tau_b)$, $c \sim \mathrm{Gam}(\gamma_c,\tau_c)$, and $\boldsymbol{\beta} \sim$

$\text{Normal}(\mathbf{b_0}, \mathbf{S_\beta})$, where the positive real values of $\gamma_r$, $\tau_r$, $\gamma_a$, $\tau_a$, $\gamma_b$, $\tau_b$, $\gamma_c$, $\tau_c$, $\mathbf{b_0}$ and $\mathbf{S_\beta}$ are known or fixed. Suppose that $\mathbf{b_0}$ is a $(k+1)\times 1$ hyper-parameter vector and $\mathbf{S_\beta}$ is a $(k+1)\times(k+1)$ known non-negative specific matrix. If each parameter is independently distributed, the joint prior distribution of all unknown parameters can be written as:

$$\pi(\mathbf{\Omega}) = \pi(r)\cdot\pi(a)\cdot\pi(b)\cdot\pi(c)\cdot\pi(\boldsymbol{\beta}). \qquad (16)$$

From equations (13) and (16), the posterior distribution is derived as follows:

$$\pi(\mathbf{\Omega}\,|\,\mathbf{X}) \propto \prod_{i=1}^{n} f(y_i\,|\,\mathbf{x}_i^T, \mathbf{\Omega})\cdot\pi(r)\cdot\pi(a)\cdot\pi(b)\cdot\pi(\boldsymbol{\beta}). \qquad (17)$$

The posterior distribution does not have an explicit form and the computational method called a Gibbs sampler was used in this study. The best known MCMC sampling algorithm was applied to find $\mathrm{E}(\mathbf{\Omega}\,|\,\mathbf{y})$. The model parameter $\mathbf{\Omega}$ was then estimated from the Bayesian method [32-33].

Model performance was compared based on the deviance, DIC and $p_D$ criteria. A good model must have lower values of these criteria. Details of each criterion are as follows:

1) The deviance is a criterion used to compare the suitability of a model when $D(\mathbf{\Omega}) = [-2\log L(\mathbf{y}\,|\,\mathbf{\Omega})]$, where $L(\mathbf{y}\,|\,\mathbf{\Omega})$ is the likelihood function, and the conditional joint pdf of the observations is given unknown parameters.

2) The DIC is regarded as a generalization of Akaike's information criterion and the Bayesian information criterion, and is widely used as a goodness-of-fit measure when using the Bayesian approach. The DIC is beneficial to Bayesian model comparison problems where the posterior distributions have been obtained by MCMC simulations [5, 34]. The DIC is defined as $\text{DIC} = \bar{D}(\mathbf{\Omega}) + \text{Var}[D(\mathbf{\Omega})]/2$ and $\bar{D}(\mathbf{\Omega}) = \mathrm{E}[-2\log L(\mathbf{y}\,|\,\mathbf{\Omega})]$.

3) The $p_D$ is associated with deviance and denoted by $p_D = \text{Var}[D(\mathbf{\Omega})]/2$.

### 3.3 Statistical modeling for empirical data

This section first describes the characteristics of the applications on a real dataset and then presents the modeling results using a Bayesian approach for the four-parameter NBL regression model by comparing the performances with the other models.

### 3.3.1 Data description

Data used as the dependent variable in this study were the number of cancer deaths in each province in Thailand in 2021 $(Y)$ [35]. The five independent variables are described below

- $X_1$ is the size of the population in each province at midyear 2021 (unit: people),
- $X_2$ represents the population in each province per doctor (unit: people),
- $X_3$ is the percentage of poor people in each province (unit: percent),
- $X_4$ represents the number of deaths from cancer caused by smoking behavior from age 15 years and over (unit: people), and
- $X_5$ represents the number of deaths from cancer as a result of drinking behavior from age 15 years and over (unit: people).

The mean and variance of cancer deaths in each province ($n = 77$) of Thailand were 1,090.58 and 942,646.81, respectively. The variance of the number of cancer deaths in each province of Thailand was greater than the mean because this dataset had an overdispersion problem. Figure 2(a) and (b) show a histogram and a normal Q-Q plot, showing that the data had a heavy-tailed distribution.

**Table 1.** Summary of empirical data concerning cancer deaths in Thailand.

| Variable | Minimum | Maximum | Median | Mean | Standard deviation |
|---|---|---|---|---|---|
| $Y$ | 178.60 | 7,606.90 | 8227.96 | 1,090.58 | 970.90 |
| $X_1$ | 191,049.00 | 5,530,283.00 | 676,105.00 | 859,421.75 | 721,173.20 |
| $X_2$ | 548.00 | 19,948.00 | 2,600.00 | 3,073.13 | 2,279.94 |
| $X_3$ | 0.00 | 938.00 | 6.66 | 20.68 | 106.21 |
| $X_4$ | 25.22 | 890.38 | 121.89 | 152.52 | 121.51 |
| $X_5$ | 0.00 | 1,504.24 | 179.22 | 241.12 | 222.06 |

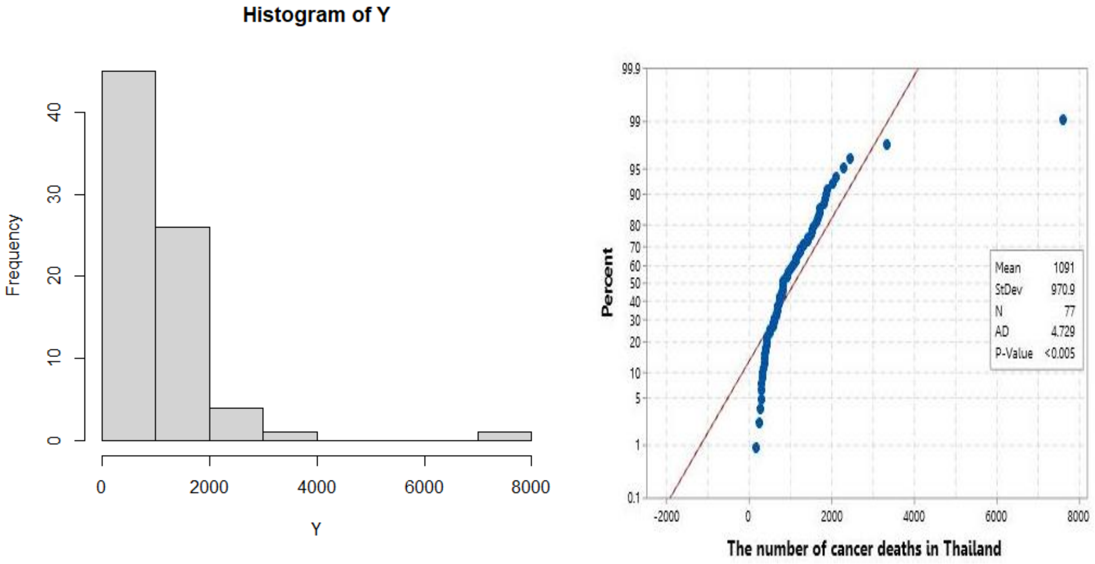

**Figure 2.** Empirical data: (a) The observed frequency (days) of the number of cancer deaths in Thailand and (b) The normal Q-Q plot of the number of cancer deaths in Thailand.

### 3.3.2 Modeling results

Three parallel independent MCMC chains for 100,000 iterations were generated for each parameter based on these prior densities, discarding the first 50,000 iterations as a burn-in for computation. The expected posterior of the parameters was calculated using the JAGS function in the R2jags package of R language [32-33]. Data analysis results illustrate applying the GLM framework to build the regression model derived from the four-parameter NBL distribution. The dependent variable, as the number of cancer deaths in Thailand, was used in three regression models including the Pois, NB and four-parameter NBL and the Bayesian approach was used to estimate the regression coefficient. The posterior means (estimates), standard error (s.e.) and 95% credible intervals (Cr.I.) of each parameter and statistics for comparing model performance including the deviance, DIC and $p_D$ of the three regression models are shown in Table 2.

The GLM results of the estimated parameters $r$, $a$, $b$ and $c$ for the proposed regression model were $\hat{r} = 8.820$, $\hat{a} = 0.126$, $\hat{b} = 215.75$, and $\hat{c} = 0.012$. From $\mathrm{E}(\lambda) = \dfrac{ac+2b}{c(ac+b)} = 166.67$. Since $\mathrm{E}(Y_i \mid \mathrm{X}_i^T) = \mu_i \mathrm{E}(\lambda)$, then the number of cancer deaths in

Thailand using the four-parameter NBL model can be represented as:

$$\hat{Y} = 166.67 \exp\{0.814 + 0.157Z_1 - 0.033Z_2 - 0.073Z_3 + 0.053Z_4 + 0.436Z_5\}, \qquad (18)$$

where $Z_i$ is a standard normal score of a random variable $X_{ij}$ for $i = 1, 2, 3, \ldots, n$ and $j = 1, 2, 3, \ldots, 5$. The value of deviance, DIC and $p_D$ of the four-parameter NBL regression model are 1,095.534, 1,102.7 and 7.1, respectively.

**Table 2**. Statistics and interval estimations of each parameter of the Pois, NB and four-parameter NBL regression models for the number of cancer deaths in Thailand.

| Parameter | Pois | | NB | | Four-parameter NBL | |
|---|---|---|---|---|---|---|
| | Estimate<br>(s.e.) | 95% Cr.I. | Estimate<br>(s.e.) | 95% Cr.I. | Estimates<br>(s.e.) | 95% Cr.I. |
| Intercept ( $\beta_1$ ) | 6.917<br>(0.0005) | (6.909,6.924) | 6.810<br>(0.039) | (6.734,6.886) | 0.814<br>(0.1542) | (-1.771,3.533) |
| $Z_1(\beta_2)$ | -0.181<br>(0.0021) | (-0.218, -0.145) | 0.169<br>(0.257) | (-0.330,0.685) | 0.157<br>(0.0281) | (-.323,0.321) |
| $Z_2(\beta_3)$ | -0.008<br>(0.0003) | (-0.014, -0.002) | -0.032<br>(0.044) | (-0.116,0.057) | -0.033<br>(0.0050) | (-0.115,0.058) |
| $Z_3(\beta_4)$ | -0.008<br>(0.0005) | (-0.017, 0.000) | -0.073<br>(0.042) | (-0.153,0.013) | -0.073<br>(0.0048) | (-0.154,0.014) |
| $Z_4(\beta_5)$ | 0.161<br>(0.0016) | (0.133, 0.188) | 0.046<br>(0.174) | (-0.284,0.390) | 0.053<br>(0.0195) | (-0.281,0.387) |
| $Z_5(\beta_6)$ | 0.426<br>(0.0014) | (0.403, 0.449) | 0.432<br>(0.142) | (0.145,0.708) | 0.436<br>(0.0155) | (0.167,0.702) |
| $r$ | - | - | 8.766<br>(1.453) | (6.134,11.476) | 8.820<br>(0.168) | (6.185,11.933) |
| $a$ | - | - | - | - | 0.126<br>(0.043) | (0.000,1.647) |
| $b$ | - | - | - | - | 215.75<br>(94.300) | (0.000,2723.4) |
| $c$ | - | - | - | - | 0.012<br>(0.0033) | (0.012, 0.075) |
| Deviance | 9,720..893 | | 1,095.693 | | 1,095.534 | |
| DIC | 10,401.3 | | 1103.4 | | 1,102.7 | |
| $p_D$ | 680.4 | | 7.7 | | 7.1 | |

The performance of the proposed model represented as density and trace plots is shown in Figures 3 and 4, respectively and the suitability of the model was also considered. The trace plots show values of the relevant parameter during the runtime of the chain, while the density plots represent values of the trace plot of the appropriate parameter in the chain. In trace plots with no specific patterns and entangled chains, the convergence can be considered good. Density plots showing significant differences for the same number of simulations reflect poor convergence. Trace plots of three concurrent chains were applied for model convergence using MCMC methods [36]. Results showed that density plots of the three MCMC chains with the MCMC plots package in R [37] from the proposed model of all parameters in three parallel chains overlapped well after the burn-in period.

Trace plots of the four-parameter NBL regression model showed vertical and dense graphs of the values of the simulated parameters against the drawn lines. The motions of the trace plots revealed the characteristics of a converged stable sequence. Therefore, the four-parameter NBL model could be fitted to this dataset.

The parameter estimate of the traditional NB distribution was obtained as r = 8.766. Since $\mathrm{E}(Y_i \mid \mathrm{X}_i^T) = \mu_i$, then the GLM for the count data approach with the NB distribution can be represented as:

$$\hat{Y}_{NB} = \exp\left[6.810 + 0.169Z_1 - 0.032Z_2 - 0.073Z_3 + 0.046Z_4 + 0.432Z_5\right]. \qquad (19)$$

The deviance, DIC and $p_D$ of the generalized linear models of the NB model were 1,095.693, 1103.4 and 7.7, respectively. In the same way, the GLM for the count data approach using the Pois distribution can be represented as:

$$\hat{Y}_{Pois} = \exp\left[6.917 - 0.181Z_1 - 0.008Z_2 - 0.008Z_3 + 0.161Z_4 + 0.426Z_5\right]. \qquad (20)$$

The deviance, DIC and $p_D$ of the generalized linear models of the Pois model were 9,720.893, 10,401.3 and 680.4, respectively.

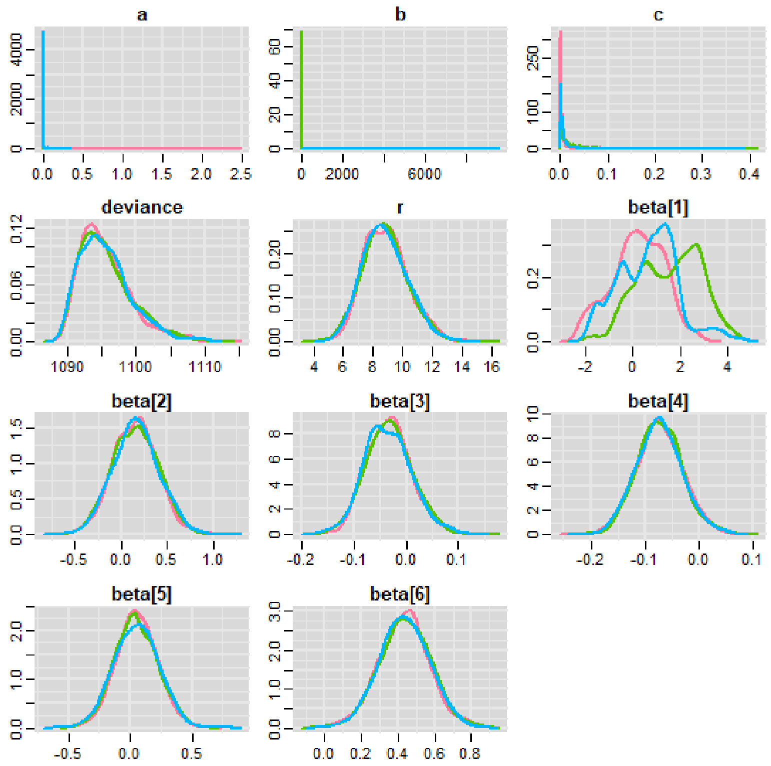

**Figure 3.** Density plots of the three MCMC chains for $r, a, b, c$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_6)^T$ from the four-parameter NBL regression model for the number of cancer deaths in Thailand.

## 4. Conclusions

A four-parameter negative binomial-Lindley (NBL) was developed and applied using the GLM framework to build a regression model. The four-parameter NBL regression model addressed the number of cancer deaths in Thailand using different factors when the dependent variable had overdispersion problems and was heavy-tailed. The proposed model was compared with the Pois and NB traditional models. Results showed that the four-parameter NBL model had the highest efficiency and was more suitable than the NB and Pois models, with the lowest values for deviance, DIC and $p_D$. The deviance, DIC and $p_D$ values of the four-parameter NBL model were lower than the NB model at 0.01, 0.06 and 8.45%, respectively and significantly lower than the Pois model at 787.32, 843.26 and 9,483.10%, respectively. The four-parameter NBL model fitted better to overdispersed and heavy-tailed count data than the classical distributions, indicating that the number of cancer deaths in Thailand was influenced by the size of the population in each province at midyear 2021, the population in each province per doctor, the percentage of poor people in each province, the number of cancer deaths from smoking age 15 years and over and number of cancer deaths from drinking from age 15 years and over.

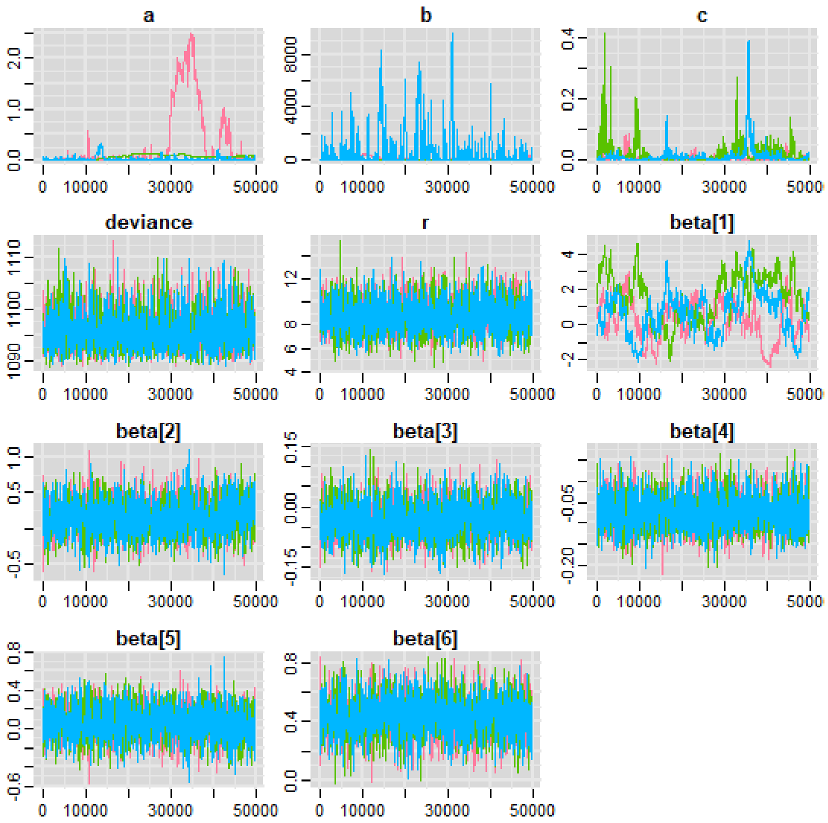

**Figure 4**. Trace plots of the three MCMC chains for $r,a,b,c$ and $\boldsymbol{\beta}=(\beta_1,\beta_2,...\beta_6)^T$ from the four-parameter NBL regression model for the number of cancer deaths in Thailand.

## ACKNOWLEDGMENTS

This research was supported by The Science, Research and Innovation Promotion Funding (TSRI) (Grant no. FRB660012/0168). This research block grants was managed under Rajamangala University of Technology Thanyaburi (FRB66E0640N.1). The authors gratefully acknowledge this funding. We are also thankful to those who could not be mentioned here for their kindness and encouragement. And finally, the authors would like to thank the anonymous reviewers for their comments and suggestions.

## CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

## REFERENCES

[1] A.A. Yirga, S.F. Melesse, H.G. Mwambi, et al. Negative binomial mixed models for analyzing longitudinal CD4 count data, Sci. Rep. 10 (2020), 16742. https://doi.org/10.1038/s41598-020-73883-7.

[2] A.C. Cameron, P. Johansson, Count data regression using series expansions: with applications, J. Appl. Econ. 12 (1997), 203–223. https://doi.org/10.1002/(sici)1099-1255(199705)12:3<203::aid-jae446>3.0.co;2-2.

[3] W. Gardner, E.P. Mulvey, E.C. Shaw, Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models, Psychol. Bull. 118 (1995), 392–404. https://doi.org/10.1037/0033-2909.118.3.392.

[4] H. He, W. Tang, W. Wang, et al. Structural zeroes and zero-inflated models, Shanghai Arch. Psychiatry, 26 (2014), 236-242.

[5] J.S. Long, Regression models for categorical and limited dependent variables, Sage Publications, Thousand Oaks, 1997.

[6] Z. Wang, One mixed negative binomial distribution with application, J. Stat. Plan. Inference. 141 (2011), 1153-1160. https://doi.org/10.1016/j.jspi.2010.09.020.

[7] H. Zamani, N. Ismail, Negative binomial-Lindley distribution and its application, J. Math. Stat. 6 (2010), 4-9.

[8] C. Pudprommar, W. Bodhisuwan, P. Zeephongse, A new mixed negative binomial distribution, J. Appl. Sci. 12 (2012), 1853–1858. https://doi.org/10.3923/jas.2012.1853.1858.

[9] S. Aryuyuen, W. Bodhisuwan, The negative binomial-generalized exponential (NB-GE) distribution, Appl. Math. Sci. 7 (2013), 1093-1105.

[10] Y. Gençtürk, A. Yiğiter, Modelling claim number using a new mixture model: negative binomial gamma distribution, J. Stat. Comput. Simul. 86 (2015), 1829–1839. https://doi.org/10.1080/00949655.2015.1085987.

[11] D. Yamrubboon, W. Bodhisuwan, C. Pudprommarat, et al. The negative binomial-Sushila distribution with

application in count data analysis, Thailand Statistician. 15 (2017), 69-77.

[12] S. Aryuyuen, Bayesian inference for the negative binomial-generalized Lindley regression model: properties and applications, Commun. Stat. - Theory Methods. (2021). https://doi.org/10.1080/03610926.2021.1995434.

[13] R.R.M. Tajuddin, N. Ismail, K. Ibrahim, et al. A four-parameter negative binomial-Lindley distribution for modeling over and underdispersed count data with excess zeros, Commun. Stat. - Theory Methods. 51 (2020), 414-426. https://doi.org/10.1080/03610926.2020.1749854.

[14] S. Aryuyuen, The negative binomial-new generalized lindley distribution for count data: properties and application, Pak. J. Stat. Oper. Res. 18 (2022), 167-177. https://doi.org/10.18187/pjsor.v18i1.2988.

[15] A. Khodadadi, M. Shirazi, S. Geedipally, et al. Evaluating alternative variations of negative Binomial-Lindley distribution for modelling crash data, Transportmetrica A: Transport Sci. (2022). https://doi.org/10.1080/23249935.2022.2062480.

[16] S. Aryuyuen, U. Tonggumnead, A new mixed negative binomial regression model to analyze factors influencing the number of patients with respiratory disease and long-term effects of lung cancer, Commun. Math. Biol. Neurosci. 2022 (2022), 103. https://doi.org/10.28919/cmbn/7705.

[17] S. Aryuyuen, U. Tonggumnead, Bayesian inference for the negative binomial-quasi Lindley model for time series count data on the COVID-19 Pandemic, Trends Sci. 19 (2022), 3171. https://doi.org/10.48048/tis.2022.3171.

[18] J. Stoklosa, R.V. Blakey, F.K.C. Hui, An overview of modern applications of negative binomial modelling in ecology and biodiversity, Diversity. 14 (2022), 320. https://doi.org/10.3390/d14050320.

[19] O.S. Adesina, D.A. Agunbiade, S.A. Osundina, Bayesian regression model for counts in scholarship, Math. Theory Model. 7 (2017), 45-57.

[20] A.C. Cameron, P.K. Trivedi, Regression analysis of count data, Second edition, Cambridge University Press, Cambridge, 2013.

[21] C. Jornsatian, W. Bodhisuwan, Bayesian inference for negative binomial-beta exponential distribution and its regression model, Lobachevskii J. Math. 43 (2022), 2501-2514. https://doi.org/10.1134/s1995080222120162.

[22] S.R. Geedipally, D. Lord, S.S. Dhavala, The negative binomial-Lindley generalized linear model: characteristics and application using crash data, Accident Anal. Prevention. 45 (2012), 258–265. https://doi.org/10.1016/j.aap.2011.07.012.

[23] P. Vangala, Negative binomial-generalized exponential distribution: generalized linear model and its applications, M. S. Thesis, Univ. Texas A&M, 2015.

[24] D. Yamrubboon, A. Thongteeraparp, W. Bodhisuwan, et al. Bayesian inference for the negative binomial-Sushila linear model, Lobachevskii J. Math. 40 (2019), 42-54. https://doi.org/10.1134/s1995080219010141.

[25] R. Shanker, K.K. Shukla, R. Shanker, et al. A three-parameter Lindley distribution. Amer. J. Appl. Math. Stat. 7 (2017), 15-26.

[26] S. Denthet, A. Thongteeraparp, W. Bodhisuwan, Mixed distribution of negative binomial and two-parameter Lindley distributions, in: 2016 12th International Conference on Mathematics, Statistics, and Their Applications (ICMSA), IEEE, Banda Aceh, Indonesia, 2016: pp. 104–107. https://doi.org/10.1109/ICMSA.2016.7954318.

[27] J.A. Nelder, R.W.M. Wedderburn, Generalized linear models, J. R. Stat. Soc. Ser. A (General). 135 (1972), 370-384. https://doi.org/10.2307/2344614.

[28] A. Gelman, J.B. Carlin, H.S. Stern, et al. Bayesian data analysis, CRC Press, New York, (2013).

[29] D.J. Lunn, A. Thomas, N. Best, et al. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility, Stat. Comput. 10 (2000), 325-337. https://doi.org/10.1023/a:1008929526011.

[30] M. Plummer, JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling, in: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, 20-22 March 2003, 1-10.

[31] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, et al. Bayesian Measures of Model Complexity and Fit, J. R. Stat. Soc. Ser. B: Stat. Methodol. 64 (2002), 583–639. https://doi.org/10.1111/1467-9868.00353.

[32] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, (2022). https://www.R-project.org/. [Access 10 January 2023].

[33] Y.S. Su, M. Yajima, M.Y.S Su, J.A.G.S System Requirements. Package `R2jags'; R package version 0.03-08, (2015), https://cran.r-project.org/web/packages/R2jags. [Access 10 January 2023].

[34] D. Lunn, C. Jackson, N. Best, et al. The BUGS book. A practical introduction to Bayesian analysis, Chapman Hall, London. (2013).

[35] Strategy and Planning Division Ministry of Public Health, ThaiHealthStat, https://www.hiso.or.th/thaihealthstat/topic/index-covid-map.php.[accessed February 2022].

[36] J.C. Doll, S.J. Jacquemin, Bayesian model selection in fisheries management and ecology, J. Fish Wildlife Manage. 10 (2019), 691-707. https://doi.org/10.3996/042019-jfwm-024.

[37] S.M. Curtis, mcmcplots: Create plots from MCMC output, R package version 0.4.3.(2018), https://cran.r-project.org/package=mcmcplots. [Access 12 October 2021].