Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2022, 2022:103

https://doi.org/10.28919/cmbn/7705

ISSN: 2052-2541

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

SIRINAPA ARYUYUEN, UNCHALEE TONGGUMNEAD*

Department of Mathematics and Computer Science, Rajamangala University of Technology Thanyaburi,

Pathum Thani, 12110, Thailand

Copyright © 2022 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: This article aims to develop a new mixed distribution which mixes the negative binomial (NB) distribution and the modified quasi-Lindley (MQL) distribution. The new mixed NB distribution is called the negative binomial-modified quasi Lindley (NB-MQL) distribution. Parameters of the distribution and its regression coefficient are estimated using a Bayesian approach. The NB-MQL linear regression model is applied with an actual dataset with the characteristics of overdispersion of the count response variable. The results show that the NB-MQL model describes the factors influencing the number of patients with respiratory disease and long-term effects of lung cancer better than the NB and Poisson regression models.

Keywords: mixed NB distribution; Bayesian approach; generalized linear model; overdispersion; count data model.

2010 AMS Subject Classification: 97K80, 91G70.

*Corresponding author

E-mail address: unchalee_t@rmutt.ac.th

Received August 30, 2022

1. INTRODUCTION

The general linear model (GLM) is a tool used by researchers in many fields to analyze data such as regression analysis, independent t-test, analysis of variance, analysis of covariance, etc. It often refers to linear regression models for continuous response variables that define continuous and/or categorical predictors. Although it is widely used, there are some limitations that make the GLM inflexible; for example, the dependent variable must be continuous or only quantitative. The strict preliminary assumption on discrepancies is normally distributed, and each observation is independent of the others. The GLM was developed to be more flexible and offer better coverage in the form of generalized linear models (GLMs). The vital GLMs include the logistic regression model, the Poisson regression model, and the negative binomial (NB) regression model. The NB model was developed to overcome the constraints of the problematic distribution for count data with overdispersion [5,8,14,15]. Although the NB distribution is proper for count data when there is an overdispersion problem, the NB distribution is appropriate for count data, presenting overdispersion without necessarily being heavy-tailed; heavy-tailed distributions have a tendency toward overdispersion [23]. There may be a very high probability that no events of interest will occur in some situations, resulting in a higher frequency of zero values. When the value of zero is high, the frequency will cause the problem of overdispersion to become more severe. Therefore, the Poisson and NB distributions are not suitable for such data. Subsequently, new distributions were developed to provide more flexibility and coverage because flexibility and coverage issues could reflect the effectiveness of the development, reduce existing constraints, or create new approaches that are more flexible or more comprehensive and relevant to different contexts. One of the most widely used distribution developments was mixed NB distributions. Many mixed NB distributions are introduced, such as the NB-Lindley [26], NB-generalized exponential [4], NB-gamma [11], NB-Sushila [26], and NB-generalized Lindley [2]. Mixed NB distributions are applied to the statistical model events for count data in real life, such as actuarial and insurance models [3,11,26], medical or industrial models [3], or the fields of ecology and biodiversity [20].

In the past, the solution of GLMs in a regression framework was usually used by maximizing the

nonlinear log-likelihood. The Newton-Raphson method can be applied to iteratively find the maximum likelihood (ML) [15]. The ML method is limited because it only provides a point estimate that is not robust. In some situations, such as if the data are small or when the dispersion parameter is much larger than the mean, this method fails to converge. Moreover, the ML method does not consider the prior information, which may be helpful in the case of missing observations. One solution that solves such problems is the Bayesian method. Parameter estimation by the Bayesian approach is done by posterior processing distribution, which multiplies the prior distribution with the likelihood. Moreover, Bayesian inference can account for prior expert knowledge on variables of interest, especially in a small sample size. It provides a sample of estimators, which may be helpful for the uncertainty analysis [7]. In addition, the advantage of the Bayesian method in practice is its flexibility and coverage, as it can solve complex problems [10,25].

This study first proposed a new mixture NB distribution to be a flexible alternative for analyzing heavy-tailed count data with overdispersion. We will apply the GLMs framework with actual datasets of two response variables; i.e., the number of patients with respiratory disease and long-term effects of lung cancer. The data were collected from 77 provinces of Thailand in 2021 [1]. Finally, the conclusion is presented.

2. PRELIMINARIES

In this section, we introduce the Poisson, NB, and modified quasi-Lindley (MQL) distributions. In addition, the generalized linear regression model and criteria for model evaluation are provided.

2.1 The Poisson distribution

Let Y be a random variable distributed as the Poisson distribution with a parameter $\lambda > 0,$ denoted by $Y \sim \text{Pois}(\lambda)$, then its probability mass function (pmf) is

$$g_1(y) = \frac{\exp(-\lambda)\lambda^y}{y!};\ y = 0,1,2,.... \tag{1}$$

The mean and variance of the Poisson distribution are λ.

2.2 The NB distribution

Let Y be a random variable distributed as the NB distribution [13] with parameters $r > 0$ and $0 < p < 1$, denoted by $Y \sim \text{NB}(r, p)$. Its pmf is

$$g_2(y) = \binom{y+r-1}{y} p^r (1-p)^y;\ y = 0,1,2,\ldots. \tag{2}$$

Its mean and variance are $r(1-p)/p$ and $r(1-p)/p^2$ respectively.

2.3 The MQL distribution

In 2022, Tharshan and Wijekoon proposed the MQL distribution, which is derived as a finite mixture of the exponential (Exp) and gamma (Gam) distributions with the mixing proportion $m = c^3/(c^3+1)$. Its probability density function (pdf) is obtained by

$$h(\lambda) = mh_1(\lambda; b) + (1-m)h_2(\lambda; a, b), \tag{3}$$

where $\lambda > 0$, for $a > 0$, and $c^3 > -1$ are shape parameters, $b > 0$ is a scale parameter, and $h_1(\lambda; b)$ and $h_2(\lambda; a, b)$ are the pdf of the Exp and Gam distributions, respectively [22]. Finally, the pdf of the MQL distribution is

$$h(\lambda) = \frac{be^{-b\lambda}\left[c^3\Gamma(a) + (b\lambda)^{a-1}\right]}{(c^3+1)\Gamma(a)}, \tag{4}$$

where $\Gamma(.)$ is a complete gamma function. Its moment generating function (mgf) is

$$M_\lambda(t) = \frac{b\left[c^3(b-t)^{a-1} + b^{a-1}\right]}{(c^3+1)(b-t)^a}. \tag{5}$$

2.4 The generalized linear regression model

Linear regression is a statistical method used to create a linear model, in which the model describes the relationship between a response variable Y_i with the number of observations n, and a set of independent variables $(X_{i1}, X_{i2}, ..., X_{ik})$ for $i = 1, 2, 3, ..., n$ and k is the number of independent variables in the model. When we consider the response variable

that is a positive integer, the expected count response μ_i, is also non-negative. The log-link maps μ_i to the whole real line. Thus, the link function is the logarithm of the mean that is $\log \mu_i$ that relates μ_i to the linear predictors. On the regression model, the log-linearity for the mean is commonly used as a link function

$$\log \mu_i = \beta_1 + \beta_2 X_{i1} + \beta_3 X_{i2} + ... + \beta_{k+1} X_{ik} \tag{6}$$

From the above equation, the covariates can be linked to the mean of Y_i by the means of the log-link function, given by

$$\mu_i = \exp(\beta_1 + \beta_2 X_{i1} + \beta_3 X_{i2} + ... + \beta_{k+1} X_{ik}) = \exp(\mathbf{X_i^T \beta}), \tag{7}$$

where $\mathbf{X_i^T} = (1, x_{i1}, x_{i2}, ..., x_{ik})$ is a vector of length $(k+1)$ where the i th row of $n \times (k+1)$ matrix $\mathbf{X}$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_{k+1})^T$ is a $(k+1) \times 1$ unknown vector of the regression coefficients.

2.5 Criteria for model evaluation

In the model comparison, three criteria are considered: (i) The deviance is $D(\Omega) = [-2 \log L(\mathbf{y} \mid \Omega)]$; where $L(\mathbf{y} \mid \Omega)$ is the likelihood function, the conditional joint pdf of the observations is given unknown parameters. (ii) The DIC is regarded as a generalization of Akaike's information criterion and the Bayesian information criterion, and is often and widely used as a goodness- of- fit measure when we use the Bayesian approach. The DIC is defined as $DIC = \bar{D}(\Omega) + p_D$, for $\bar{D}(\Omega) = \mathrm{E}[-2 \log L(\mathbf{y} \mid \Omega)]$ and $p_D = \mathrm{Var}[D(\Omega)]/2$, where the first term is the posterior mean of the deviance, and the second term is an alternative measure of the elective number of parameters [16]. The DIC is beneficial to Bayesian model comparison problems where the posterior distributions have been obtained by MCMC simulations [16,19]. Therefore, deviance and DIC are statistics to compare the models. The model which has the smallest value of deviance, DIC, and p_D is the best model.

3. MAIN RESULTS

In this paper, a new mixed NB distribution and its properties are proposed. Next, the proposed distribution is used to build the GLMs in the form of the regression analysis in cases of dependent variables for count data with overdispersion. The estimation of the model parameters with a Bayesian approach to derive the model for real data is provided.

3.1 A new mixed NB distribution

A new mixed NB distribution is obtained by mixing the NB and MQL distributions as follows.

Definition 1: Let Y be a random variable distributed as the NB distribution with parameters r and $p=e^{-\lambda}$ where λ is distributed as an MQL distribution with parameters a, b and c, i.e., $Y\sim \text{NB}(r,e^{-\lambda})$ and $\lambda\sim \text{MQL}(a,b,c)$. Then Y is a random variable distributed as a negative binomial-modified quasi Lindley (NB-MQL) distribution with parameters r, a, b and c, denoted by $Y\sim \text{NB-MQL}(r,a,b,c)$,

Theorem 1: Let $Y\sim \text{NB-MQL}(r,a,b,c)$, then its pmf is

$$f(y)=\binom{y+r-1}{y}\sum_{j=0}^{y}\binom{y}{j}(-1)^j\frac{b[c^3(b+r+j)^{a-1}+b^{a-1}]}{(c^3+1)(b+r+j)^a} \tag{8}$$

where $y=0,1,2,\ldots$, $r>0$, $a>0$, $b>0$ and $c^3>-1$

Proof: Let $Y\sim \text{NB}(r,p)$ and $\lambda\sim \text{MQL}(a,b,c)$, if $Y\mid\lambda\sim \text{NB}(r,p=e^{-\lambda})$ with the pmf as (2), we have the pmf of $Y\mid\lambda$ as follows:

$$f(y\mid\lambda)=\binom{y+r-1}{y}e^{-\lambda r}\left(1-e^{-\lambda}\right)^y=\binom{y+r-1}{y}\sum_{j=0}^{y}\binom{y}{j}e^{-\lambda r}(-1)^j e^{-\lambda(r+j)},$$

where $\lambda\sim \text{MQL}(a,b,c)$ with the pdf in (4), then the pmf of Y can be obtained by

$$f(y) = \int_0^\infty f(y \mid \lambda) g_3(\lambda) \mathrm{d}\lambda$$

$$= \binom{y+r-1}{y} \sum_{j=0}^{y} \binom{y}{j} (-1)^j \int_0^\infty e^{-\lambda(r+j)} \left[\frac{b e^{-b\lambda} \left[c^3 \Gamma(a) + (b\lambda)^{a-1} \right]}{(c^3+1)\Gamma(a)} \right] \mathrm{d}\lambda$$

$$= \binom{y+r-1}{y} \sum_{j=0}^{y} \binom{y}{j} (-1)^j \frac{b\left[c^3 (b+r+j)^{a-1} + b^{a-1} \right]}{(c^3+1)(b+r+j)^a}.$$

The shape of the pmf for the NB-MQL distribution is provided in Figure 1. Some basic properties of the NB-MQL distribution are obtained from the factorial moment as follows.

Theorem 2: Let $Y \sim \text{NB-MQL}(r, a, b, c)$, then its jth factorial moment is

$$\mu'_{[j]} = \frac{\Gamma(r+j)}{\Gamma(r)} \sum_{l=0}^{j} \binom{j}{l} (-1)^l \frac{b\left[c^3 (b-j+l)^{a-1} + b^{a-1} \right]}{(c^3+1)(b-j+l)^a} \tag{9}$$

where $j = 1, 2, 3, \ldots$, $r > 0$, $a > 0$, $b > 0$ and $c^3 > -1$.

Proof: The jth factorial moment of Y is

$$\mu_{[j]}(y; r, p) = \mathrm{E}[Y(Y-1)\cdots(Y-j+1)] = \frac{\Gamma(r+j)}{\Gamma(r)} \frac{(1-p)^j}{p^j}$$

where $j = 1, 2, 3, \ldots$. For $p = e^{-\lambda}$, we can write it as follows [12],

$$\mu_{[j]}(y; r, e^{-\lambda}) = \mathrm{E}_\lambda \left[\frac{\Gamma(r+j)}{\Gamma(r)} \frac{(1-e^{-\lambda})^j}{e^{-\lambda j}} \right] = \frac{\Gamma(r+j)}{\Gamma(r)} \mathrm{E}_\lambda \left(e^\lambda - 1 \right)^j$$

Using a binomial expansion in the term $(e^\lambda - 1)^j$, we can write the above equation as

$$\mu_{[j]}(y; r, e^{-\lambda}) = \frac{\Gamma(r+j)}{\Gamma(r)} \sum_{l=0}^{j} (-1)^l \mathrm{E}_\lambda \left(e^{\lambda(j-l)} \right) = \frac{\Gamma(r+j)}{\Gamma(r)} \sum_{l=0}^{j} \binom{j}{l} (-1)^l M_\lambda \left(j-l \right).$$

If $X \mid \lambda \sim \text{NB}(r, e^{-\lambda})$ and $\lambda \sim \text{MQL}(a, b, c)$ when substituting the mgf of λ as in (5) with $t = (j-l)$ into $\mu_{[j]}(.)$, we have the jth factorial moment of Y as in (9).

3.2 The NB-MQL regression model

In this paper, the regression model for a response variable distributed as the NB-MQL distribution is constructed. From the pmf in (2), we can parameterize p in terms of r as

$p = r/(\mu_i + r)$ for μ_i as the mean response variable while r is the reciprocal (or inverse of a dispersion parameter $\phi : \phi = 1/r$). The traditional NB distribution [5] can be rewritten to show its pmf as follows:

$$f(y_i; r, \mu_i) = \binom{y_i + r - 1}{y_i} \left(\frac{r}{\mu_i + r}\right)^r \left(\frac{\mu_i}{\mu_i + r}\right)^{y_i} \tag{10}$$

for $y_i = 0,1,2,...,$ $\mu_i > 0$ and $r > 0$. Then the mean and variance of Y_i are $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{Var}(Y_i) = \mu_i + \mu_i^2/r$ respectively.

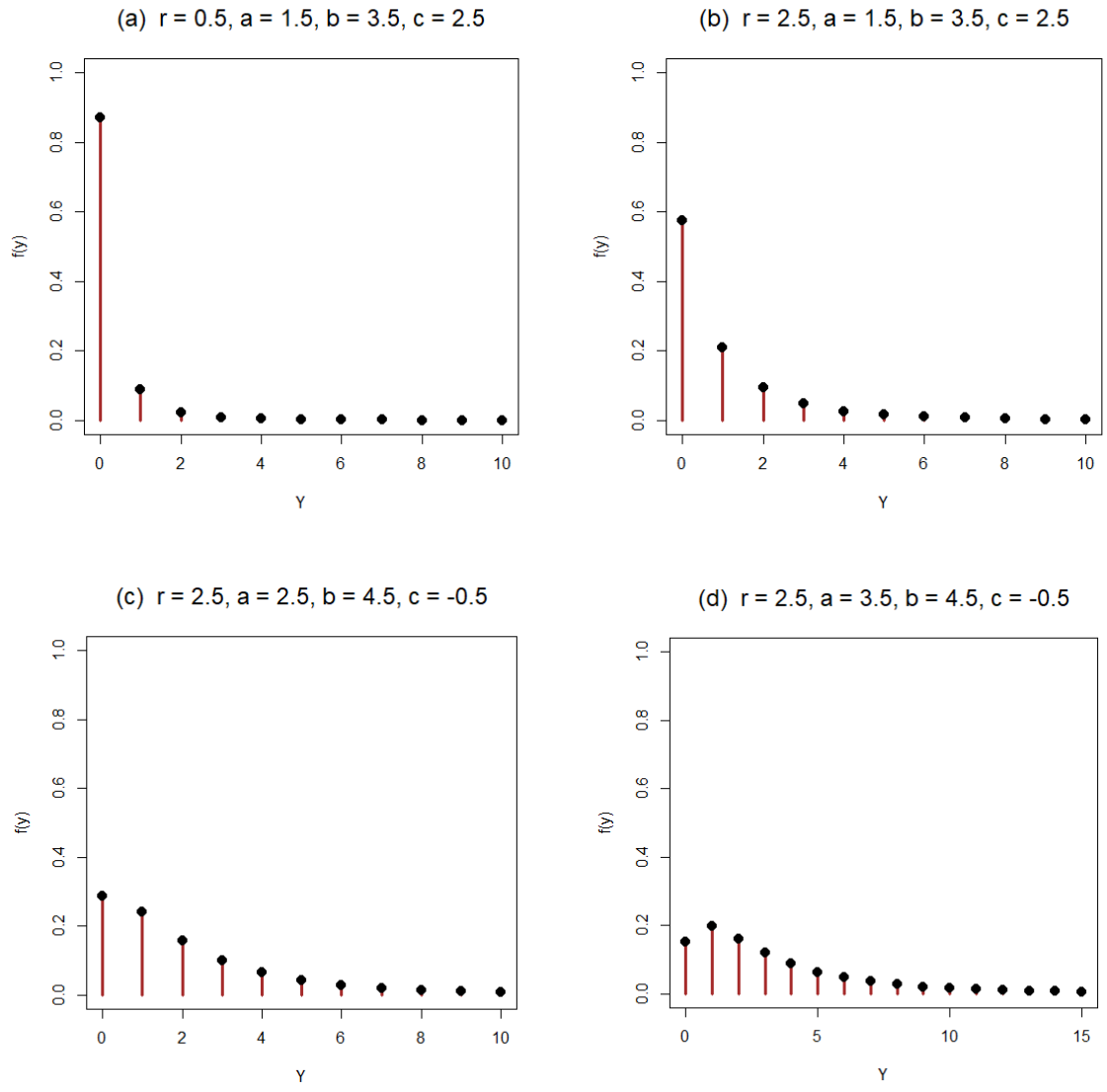


Figure 1. The pmf plots of the NB-MQL distribution with some specified parameter values.

The framework of the GLMs can be shown for deriving the NB-MQL model by considering the mixture between the NB and MQL distributions:

$$f(y_i;\boldsymbol{\theta}) = \int_0^{\infty} \mathrm{NB}(y_i; r, \lambda\mu_i)\mathrm{MQL}(\lambda; a, b, c)\mathrm{d}\lambda, \tag{11}$$

where $\boldsymbol{\theta} = (\mu_i, r, a, b, c)^T$ and the mean response μ_i is a similar parameter to a label in (10), and λ is distributed as the MQL distribution with the pdf in (4). The pmf of the NB-MQL distribution becomes:

$$f(y_i;\boldsymbol{\theta}) = \int_0^{\infty} \binom{y_i + r - 1}{y_i} \left(\frac{r}{\lambda\mu_i + r}\right)^r \left(\frac{\lambda\mu_i}{\lambda\mu_i + r}\right)^{y_i} \left(\frac{b\left[c^3(b+r+j)^{a-1} + b^{a-1}\right]}{(c^3+1)(b+r+j)^a}\right)\mathrm{d}\lambda\,. \tag{12}$$

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

$$f(y_i \mid \mathbf{x}_i^T) = \int_0^{\infty} \binom{y_i + r - 1}{y_i} \left(\frac{r}{\lambda e^{\mathbf{x}_i^T\boldsymbol{\beta}} + r}\right)^r \left(\frac{\lambda e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{\lambda e^{\mathbf{x}_i^T\boldsymbol{\beta}} + r}\right)^{y_i} \left(\frac{b\left[c^3(b+r+j)^{a-1} + b^{a-1}\right]}{(c^3+1)(b+r+j)^a}\right)\mathrm{d}\lambda \tag{13}$$

For $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ is a $(n \times 1)$ vector of the response variables which are n independent realizations of the NB- MQL model and $\boldsymbol{\Omega} = (r, a, b, c, \boldsymbol{\beta}^T)^T$ is a vector of the regression parameters. Thus, the likelihood function of $\boldsymbol{\Omega}$ is

$$L(\Omega \mid \mathbf{y}, \mathbf{x}) = \prod_{i=1}^{n} \binom{y_i + r - 1}{y_i} \int_0^{\infty} \left(\frac{r}{\lambda e^{\mathbf{x}_i^T\boldsymbol{\beta}} + r}\right)^r \left(\frac{\lambda e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{\lambda e^{\mathbf{x}_i^T\boldsymbol{\beta}} + r}\right)^{y_i} \left(\frac{b\left[c^3(b+r+j)^{a-1} + b^{a-1}\right]}{(c^3+1)(b+r+j)^a}\right)\mathrm{d}\lambda\,. \tag{14}$$

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

$$\mathrm{E}(Y_i \mid \mathbf{x}_i^T) = \mu_i \mathrm{E}(\lambda) \quad \text{and} \quad \mathrm{Var}(Y_i \mid \mathbf{x}_i^T) = \mathrm{E}(Y_i \mid \mathbf{x}_i^T) + \mu_i^2 \left(\frac{1+r}{r}\right)\mathrm{E}(\lambda^2) - \mathrm{E}^2(Y_i \mid \mathbf{x}_i^T) \tag{15}$$

where $\mathrm{E}(\lambda)$ and $\mathrm{E}(\lambda^2)$ are the first second moments of the MQL distribution as follows:

$$\mathrm{E}(\lambda) = \frac{c^3 + a}{(c^3+1)b} \quad \text{and} \quad \mathrm{E}(\lambda^2) = \frac{2c^3 + a(a+1)}{(c^3+1)b^2}\,. \tag{16}$$

3.3 Bayesian inference for the NB-MQL regression model

In this paper, the vector of unknown parameters $\boldsymbol{\Omega}$ can be customarily estimated using the Bayesian approach, which allows the consideration of prior information for parameter estimation. Numerous researchers have shown interest in the study of the hierarchical Bayesian modeling approach relying on Markov Chain Monte Carlo (MCMC) techniques as referred to [10]. In this article, we implement the Bayesian approach using the MCMC technique for the NB-MQL regression model.

As shown in the likelihood function of the NB-MQL regression model in (14), it is not a closed form. It can be executed using the representation of the hierarchical model implicit both in the integral and the definition of the MQL distribution. Since the MQL distribution is mixed between the Exp distribution with scale parameter b, denoted by Exp(b), and the Gam distribution with shape parameter a and scale parameter b, denoted by Gam(a,b), therefore the pdf of the MQL distribution as in (3) can be written as proposed by [22].

$$\lambda \sim \frac{c^3}{c^3+1}\text{Exp}(b) + \frac{1}{c^3+1}\text{Gam}(a,b). \tag{17}$$

The NB-MQL distribution is conditional upon the unobserved site-specific frailty term λ, which describes the additional heterogeneity [9]. Consequently, the hierarchical framework can be represented as:

$$f(y_i;\mu_i,r\,|\,\lambda) = \text{NB}(y_i;\lambda\mu_i,r);\quad \mu_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta}) \quad \text{and} \quad \lambda \sim \text{MQL}(a,b,c) \tag{18}$$

In Bayesian inference, the prior distribution plays a defining role in the estimation of the unknown parameters in any distribution. In this study, all unknown parameters r, a, b, c and β are considered. Assuming the parameters of the NB-MQL regression model with parameters r, a and b are distributed as the Gam distribution, c^3 is distributed as the uniform (U) distribution, and $\boldsymbol{\beta}$ is distributed as the normal (N) distribution. They are mutually independently distributed in each parameter, and the joint prior distribution of all unknown parameters as follows:

$$r \sim \text{Gam}(\alpha_r, \theta_r),\ a \sim \text{Gam}(\alpha_a, \theta_a),\ b \sim \text{Gam}(\alpha_b, \theta_b),\ c \sim \text{U}(\alpha_c, \theta_c),\ \text{and}\ \ \boldsymbol{\beta} \sim \text{N}(\mathbf{b_0}, \mathbf{S_\beta}), \quad (19)$$

where the positive real values of $\alpha_r, \theta_r, \alpha_a, \theta_a, \alpha_b, \theta_b, \alpha_c, \theta_c, \mathbf{b_0}$ 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. Each parameter is supposed to be independently distributed, and the joint prior distribution of all unknown parameters can be written as

$$\pi(\boldsymbol{\Omega}) = \pi(r)\pi(a)\pi(b)\pi(c)\pi(\boldsymbol{\beta}). \quad (20)$$

From the likelihood function in (15) and the prior distribution in (19), we derive the posterior distribution as follows:

$$\pi(\Omega \mid \mathbf{X}) \propto \prod_{i=1}^{n} f(y_i \mid \mathbf{x}_i^T, \Omega)\, \pi(r)\pi(a)\pi(b)\pi(c)\pi(\boldsymbol{\beta}) \quad (21)$$

For the NB-MQL model, the full conditional posterior distributions for each parameter of $\boldsymbol{\Omega}$ derived from (21) are obtained as:

$$\pi(r \mid \mathbf{y}, \mathbf{x}, r, a, b, c) \propto \prod_{i=1}^{n} f(y_i \mid \mathrm{x}_i^T, \boldsymbol{\Omega})\pi(r), \quad \pi(a \mid \mathbf{y}, \mathbf{X}, r, a, b, c) \propto \prod_{i=1}^{n} f(y_i \mid \mathrm{x}_i^T, \boldsymbol{\Omega})\pi(a),$$

$$\pi(b \mid \mathbf{y}, \mathbf{x}, r, a, b, c) \propto \prod_{i=1}^{n} f(y_i \mid \mathrm{x}_i^T, \boldsymbol{\Omega})\pi(b), \quad \pi(c \mid \mathbf{y}, \mathbf{x}, r, a, b, c) \propto \prod_{i=1}^{n} f(y_i \mid \mathrm{x}_i^T, \boldsymbol{\Omega})\pi(c),$$

and $\pi(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{x}, r, a, b, c) \propto \prod_{i=1}^{n} f(y_i \mid \mathrm{x}_i^T, \Omega)\pi(\boldsymbol{\beta})$.

In this study, the model parameters $\boldsymbol{\Omega}$ can be estimated from the Bayesian method using the MCMC algorithm to produce the posterior inference for each parameter. Based on these prior densities, we generated three parallel independent MCMC chains for 30,000 iterations in each parameter, discarding the first 15,000 iterations as a burn-in for computation. In this paper, the expected posterior of the parameters is calculated using the jags function in the R2jags package of the R language [18, 21].

Table 1. Summary of empirical data

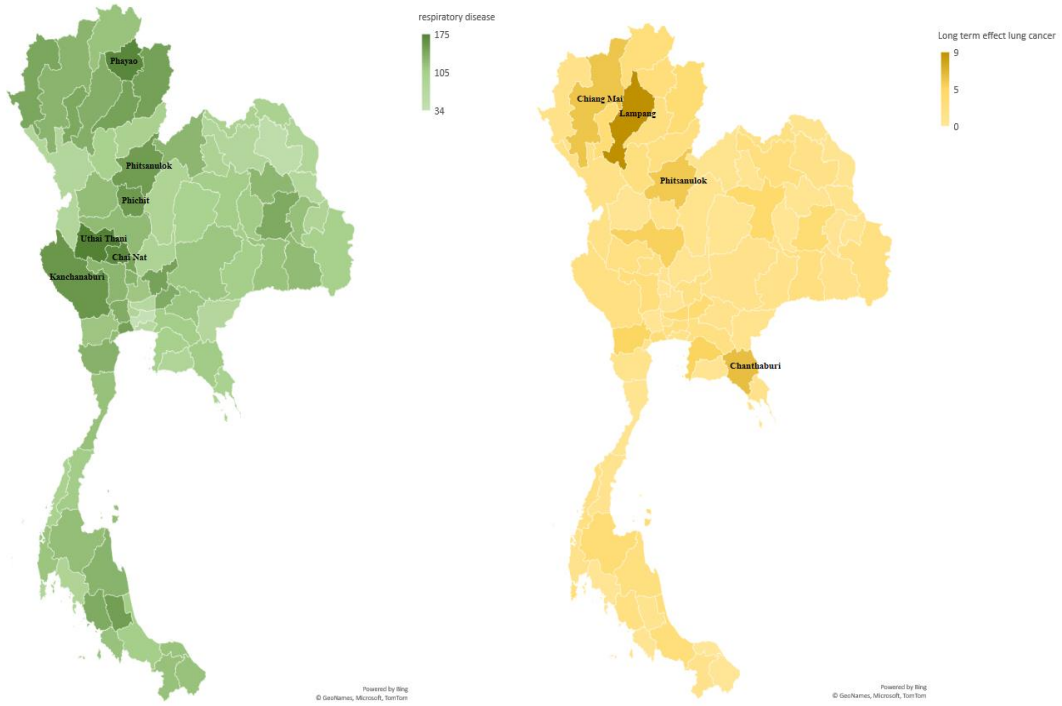
Variables	Minimum	Maximum	Median	Mean	Variance	Standard deviation
Y_1	19443	301,307	79,241	91,347.00	3.06×10^9	55,317.27
Y_2	210	13,617	1,118	2,109.00	6.07×10^6	2,463.74
X_1	8	32	16	17.12	40.51	6.36
X_2	21	76	66	64.65	68.99	8.31
X_3	14	97	29	30.73	200.25	14.15
X_4	1	17	6	6.01	14.22	3.77
X_5	0	9	2	2.55	5.12	2.26
X_6	162	837	271	309.50	14,454.60	120.23
X_7	191,049	5.53×10^6	676,105	859,422.00	5.20×10^{11}	721,110.26
X_8	548	19,948	2,600	3,073.00	5,198,139.00	2,279.94
X_9	0	43.96	6.60	8.61	63.51	7.97

3.4 Statistical modelling for empirical data

3.4.1. Empirical data

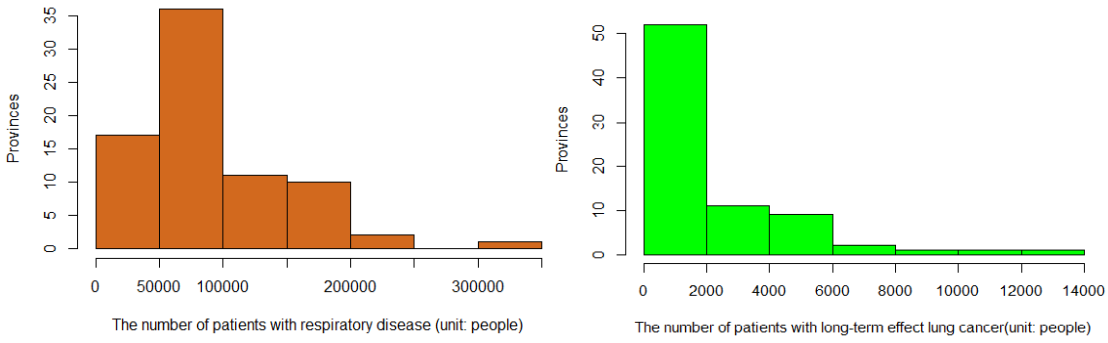
The data used in this study were the number of patients with respiratory disease in the dataset: 1) the number of patients with respiratory disease (Y_1; unit: people), and 2) the number of patients with the long-term effects of lung cancer (Y_2; unit: people). The data were collected from 77 provinces of Thailand in 2021 (Air Quality and Noise Management Bureau Pollution Control Department, 2021). All independent variables are as follows: X_1 is the average of $PM_{2.5}$ (μg/m³), X_2 is the average of O_3 (μg/m³), X_3 is the average of PM_{10} (μg/m³), X_4 is the average of SO_2 (μg/m³), X_5 is the average of NO_2 (μg/m³), X_6 is the average of CO (μg/m³), X_7 is the size of the population in each province at the midyear of 2021 (unit: people), X_8 is the ratio of doctors to population, and X_9 is the ratio of poor people in each province (unit: percent). According to Figure 2(a) on the analysis results of the disease incidence rate per 1000 population, the province with the highest incidence rate of respiratory disease is Uthai Thani (175 people per 1000 population). The mean

and variance of Y_1 are 91,347.00 and 3.06 x 10^9, respectively (see Table 1). At the same time, the province with the highest incidence rate of long-term effects of lung cancer is Lampang (9 people per 1000 population). The mean and variance of Y_2 are 2,109.00 and 6.07 x 10^6, respectively (see Table 1). Since the variance of the two data sets is greater than the mean, these data sets have an overdispersion problem. Figure 3(a) and Figure 3(b) show histograms of the number of patients with respiratory disease and the number of patients with long-term effects of lung cancer from 77 provinces in Thailand in 2021, respectively.



(a) (b)

Figure 2. The incidence rate in each province of Thailand in 2021 of (a) respiratory disease and (b) long-term effect lung cancer.



(a) (b)

Figure 3. The observed frequency (provinces) of (a) Y_1 and (b) Y_2.

3.4.2 Results of data analysis

In this section, the results of the data analysis are an illustrative method of applying the GLMs framework to build the regression model derived for the NB-MQL distribution. The dependent variables of Y_1 and Y_2 are provided in an NB-MQL distribution. The regression coefficient was estimated by the Bayesian approach. The posterior means (estimates), standard error (s.e.), 95% credible intervals (Cr.I.) of each parameter, and statistics for comparing the model's performance (the deviance, DIC, and p_D of the Poisson, NB and NB-MQL regression models of Y_1 and Y_2) are shown in Tables 2-3 respectively. In this study, the nine independent variables are standardized to standard scores.

When considering the performance of the models, the results indicate that the DIC and p_D values of the NB-MQL model are the smallest. Moreover, the density plots of the three MCMC chains with the MCMC plots package in R [6] from the NB-MQL regression model can be seen in Figure 3 and Figure 5. The results show that the density plots of all parameters in three parallel chains overlap well after the burn-in period. The trace plots of the NB-MQL regression model are displayed in Figure 4 and Figure 6. The trace plots show that graphs of the values of the simulated parameters against the drawn lines look almost vertical and dense. The motion of the trace plot reveals the characteristics of a converged manner, and the sequence seems stable. Therefore, it is confirmed from the results that the NB-MQL model can be fitted for this data set as well.

The results of the GLMs regression model for Y_1 are shown in Table 3. The results of the estimated parameters r, a, b and c for the NB-MQL regression model are:

$$\hat{r} = 5.752, \quad \hat{a} = 1.010, \quad \hat{b} = 1.038, \quad \text{and} \quad \hat{c} = 9.708.$$

From $\mathrm{E}(\lambda) = \frac{c^3 + a}{(c^3 + 1)b}$, we have $\mathrm{E}(\lambda) = \frac{9.708^3 + 1.010}{(9.708^3 + 1)1.038} = 0.963$. According to Table 2, the number of patients with respiratory disease in Thailand with the NB-MQL distribution can be represented as:

$$\hat{\mu} = 0.963\exp\{11.538 - 0.033Z_1 - 0.068Z_2 + 0.090Z_3 + 0.027Z_4 - 0.080Z_5$$
$$+0.033Z_6 + 0.543Z_7 + 0.050Z_8 + 0.052Z_9\}.$$

Where Z_i is standard normal score of a random variable X_i for $i = 1, 2, ..., 9$.

The result of the GLMs model for Y_2 is shown in Table 3. The results of the estimated parameters of the NB-MQL distribution in the GLMs models are: $\hat{r} = 1.375$, $\hat{a} = 1.005$, $\hat{b} = 0.859$, $\hat{c} = 4.642$, and $\mathrm{E}(\lambda) = 1.164$. According to Table 3, the number of patients with long-term effects of lung cancer in Thailand with the GLMs approach with the NB-MQL distribution can be represented as:

$$\hat{\mu} = 1.164\exp\{5.817 + 0.238Z_1 - 0.017Z_2 + 0.155Z_3 - 0.095Z_4 - 0.084Z_5$$
$$+0.342Z_6 + 0.504Z_7 - 0.203Z_8 + 0.075Z_9\}.$$

Table 2. Parameter estimates and various statistics of fitting models for Y_1.

Parameters	Poisson		NB		NB-MQL	
	Mean (s.e.)	95% CR.I	Mean (s.e.)	95% CR.I	Mean (s.e.)	95% CR.I
Intercept (β_1)	11.436 (0.000)	(11.436, 11.437)	11.326 (0.048)	(11.231, 11.423)	11.538 (0.216)	(8.704, 15.597)
Z_1 (β_2)	-0.023 (0.001)	(-0.025, -0.021)	-0.030 (0.106)	(-0.237, -0.182)	-0.033 (0.012)	(-0.238, 0.177)
Z_2 (β_3)	-0.087 (0.001)	(-0.088, -0.086)	-0.065 (0.073)	(-0.218, 0.069)	-0.068 (0.008)	(-0.222, 0.071)
Z_3 (β_4)	0.060 (0.001)	(0.058, 0.061)	0.087 (0.094)	(-0.089, 0.278)	0.090 (0.011)	(-0.091, 0.281)
Z_4 (β_5)	-0.026 (0.001)	(-0.028,-0.025)	0.023 (0.085)	(-0.138, 0.195)	0.027 (0.010)	(-0.137, 0.192)
Z_5 (β_6)	-0.125 (0.001)	(-0.126,-0.123)	-0.076 (0.081)	(-0.238, 0.084)	-0.080 (0.009)	(-0.237, 0.078)
Z_6 (β_7)	0.067 (0.001)	(0.066, 0.068)	0.031 (0.073)	(-0.019, 0.172)	0.033 (0.008)	(-0.110, 0.176)
Z_7 (β_8)	0.274 (0.000)	(0.273, 0.275)	0.544 (0.078)	(0.388, 0.702)	0.543 (0.009)	(0.399, 0.701)
Z_8 (β_9)	0.050 (0.000)	(0.050, 0.051)	0.022 (0.054)	(-0.079, 0.137)	0.021 (0.006)	(-0.074, 0.132)
Z_9 (β_{10})	0.052 (0.000)	(0.051, 0.053)	0.018 (0.058)	(-0.095, 0.134)	0.019 (0.007)	(-0.090, 0.136)
r	-	-	5.762 (0.949)	(4.061, 7.810	5.752 (0.949)	(4.045, 7.755)
a	-	-	-	-	1.010, (0.953)	0.047, 3.616)
b	-	-	-	-	1.038 (1,091)	0.057, 3.997)
c	-	-	-	-	9.708 (5.973)	-0.196, 19.407)
Deviance	1.271×10^6		1,812.585		1,812.604	
DIC	1.456×10^6		1,829.7		1,828.3	
p_D	1.853×10^5		17.1		15.7	

Table 3. Parameter estimates and various statistics of the fitting models for Y_2.

Parameters	Poisson		NB		NB-MQL	
	Mean (s.e.)	95% CR.I	Mean (s.e.)	95% CR.I	Mean (s.e.)	95% CR.I
Intercept (β_1)	6.563 (0.007)	(6.559, 6.573)	6.202 (0.102)	(6.010, 6.402)	5.817 (1.450)	(3.416, 9.174)
Z_1 (β_2)	0.177 (0.012)	(0.169, 0.201)	0.231 (0.274)	(-0.325, 0.749)	0.238 (0.270)	(-0.295, 0.782)
Z_2 (β_3)	0.076 (0.009)	(0.070, 0.094)	-0.017 (0.191)	(-0.412, 0.333)	-0.017 (0.191)	(-0.412, 0.336)
Z_3 (β_4)	0.091 (0.011)	(0.084, 0.112)	0.163 (0.247)	(-0.278, 0.700)	0.155 (0.243)	(-0.290, 0.656)
Z_4 (β_5)	0.010 (0.010)	(0.004, 0.031)	-0.099 (0.180)	(-0.425, 0.290)	-0.095 (0.180)	(-0.428, 0.277)
Z_5 (β_6)	-0.030 (0.010)	(-0.037, -0.011)	-0.079 (0.192)	(-0.463, 0.298)	-0.084 (0.193)	(-0.460, 0.300)
Z_6 (β_7)	0.259 (0.007)	(0.254, 0.273)	0.346 (0.164)	(0.020, 0.665)	0.342 (0.164)	(0.017, 0.664)
Z_7 (β_8)	0.178 (0.003)	(0.176, 0.184)	0.503 (0.150)	(0.226, 0.609)	0.504 (0.148)	(0.231, 0.798)
Z_8 (β_9)	-0.188 (0.005)	(-0.192, -0.178)	-0.201 (0.091)	(-0.367, -0.014)	-0.203 (0.089)	(-0.367, -0.010)
Z_9 (β_{10})	0.092 (0.006)	(0.087, 0.103)	0.077 (0.119)	(-0.144, 0.316)	0.075 (0.119)	(-0.154, 0.311)
r	-	-			1.375 (0.206)	(0.992, 1.811)
a	-	-	-	-	1.005 (1.010)	(0.033, 3.751)
b	-	-	-	-	0.859 (0.910)	(0.044, 3.353)
c	-	-	-	-	4.642 (3.022)	(-0.494, 9.730)
Deviance	32,302.4		1,105.9		1,105.8	
DIC	49,698.0		1,119.1		1,118.3	
p_D	17,395.6		13.1		12.5	

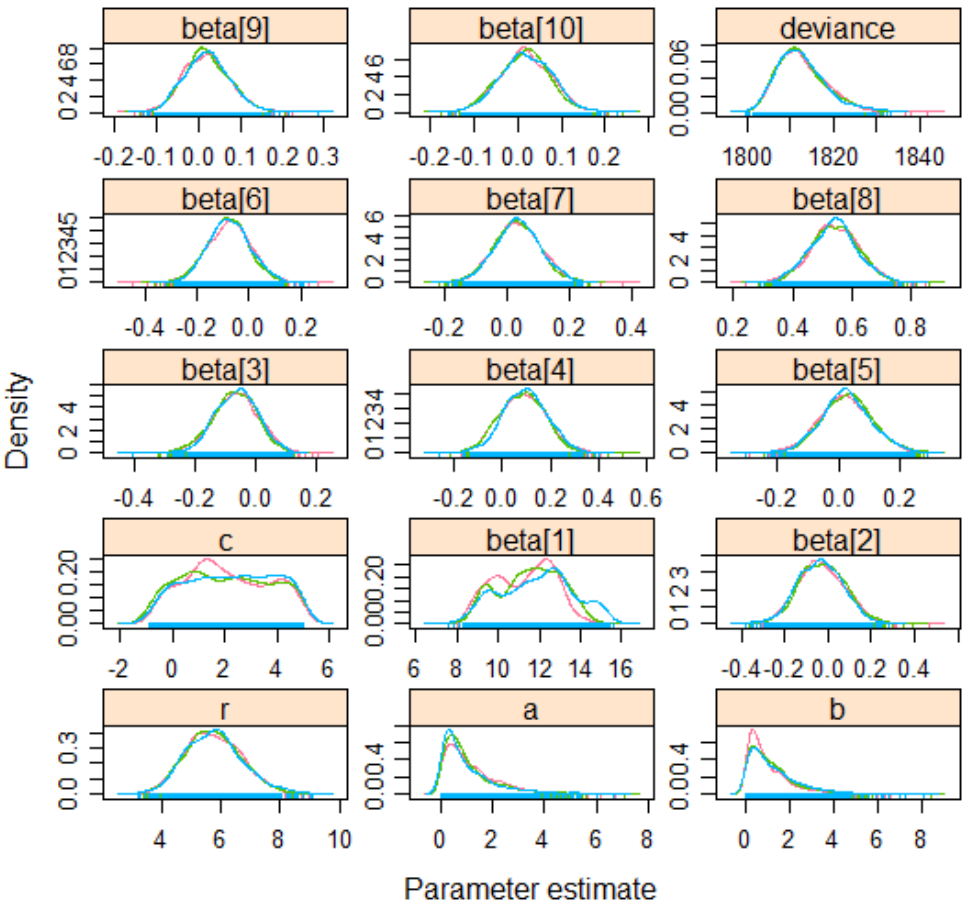


Figure 4. Density plots of three MCMC chains for r, a, b, c, deviance and $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_{10})^T$ from NB-MQL regression model for the first dataset.

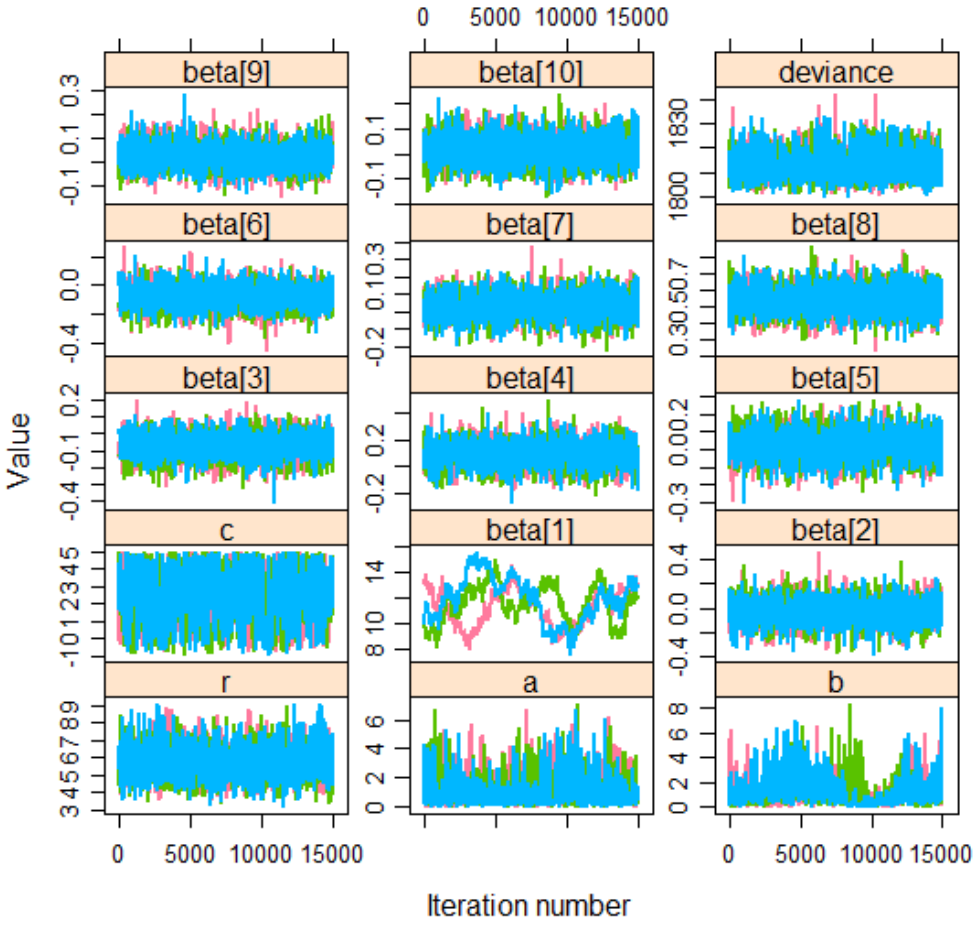


Figure 5. Trace plots of three MCMC chains for r, a, b, c, deviance and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_{10})^T$ from the NB-MQL regression model for Y_1.

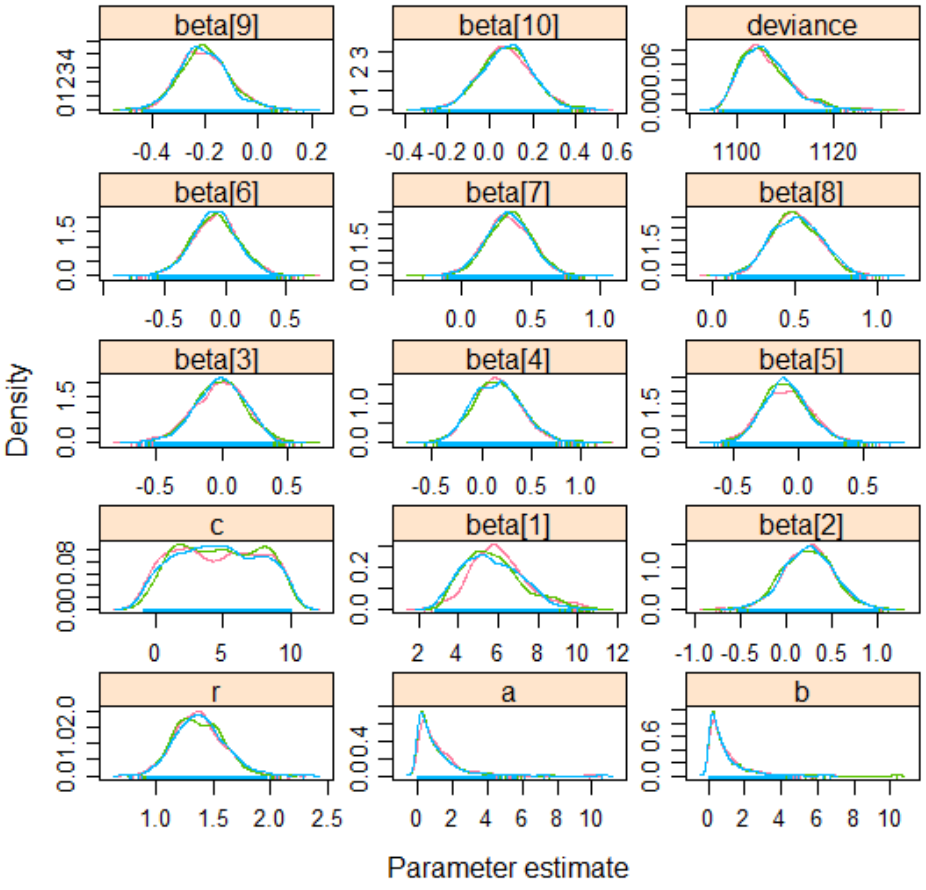


Figure 6. Density plots of three MCMC chains for r, a, b, c and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_{10})^T$ from NB-MQL regression model for Y_2.

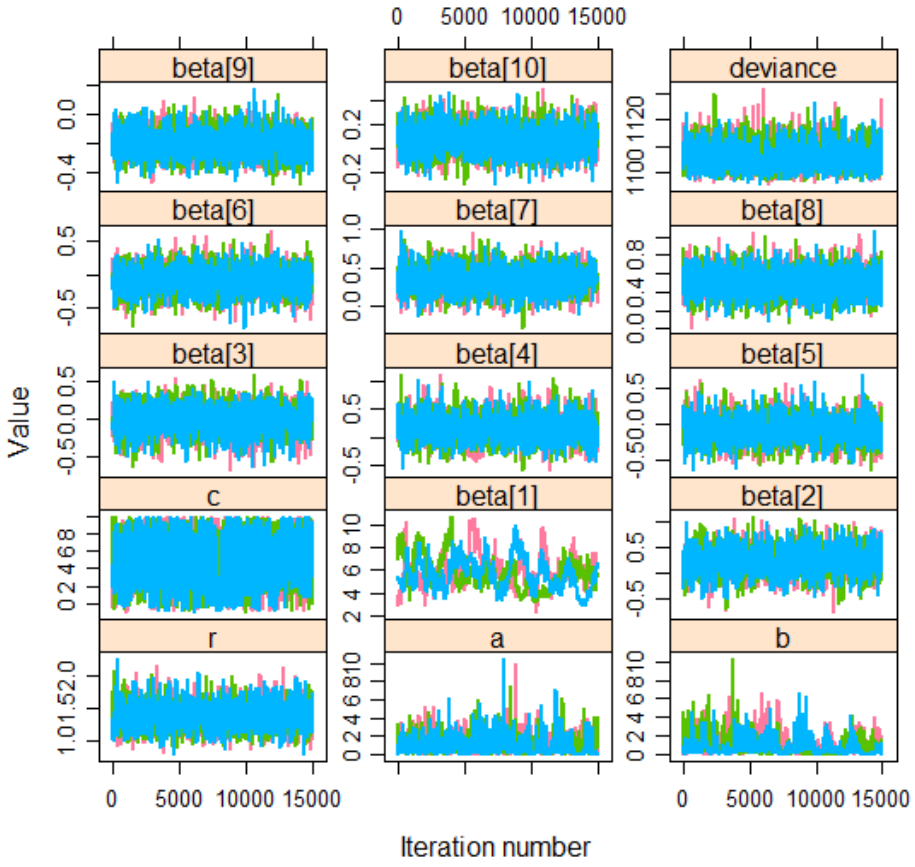


Figure 7. Trace plots of the three MCMC chains for *r*, *a*, *b*, *c*, and $\boldsymbol{\beta} = (\beta_1, \beta_2, ... \beta_{10})^T$ from NB-MQL regression model the for Y_2.

3.5. CONCLUSION

This study develops the new mixed NB distribution, which is called the NB-MQL distribution, and applies the newly created distribution with a GLMs framework for Y_1 Y_2, where the dependent variable is in the form of count data. In addition, Y_1 and Y_2 have overdispersion problems. The model efficacy study found that for Y_1 and Y_2 the Deviance, DIC, and p_D values of the NB-MQL model were significantly lower than those of the Poisson model. But when comparing the NB-MQL and NB models, the Deviance, DIC, and p_D values of the NB-MQL model were lower than the NB model in all situations. Except in the case of Y_1 being the dependent variable, the Deviance of the NB model is slightly lower than the NB-MQL model, at 0.001 %. According to the results,

the NB-MQL model seems to outperform other models. Therefore, the NB-MQL is an alternative in creating or developing a model related to an overdispersion count response variable and various covariates, and the model can be applied to accurate data in many fields.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the participation of the Faculty of Science and Technology, RMUTT University. 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] Air Quality and Noise Management Bureau Polluation Control Department, Air quality index (2021), http://air4thai.pcd.go.th/webV2/download.php. [Access 18 September 2021].

[2] S. Aryuyuen, Bayesian inference for the negative binomial-generalized Lindley regression model: properties and applications, Commun. Stat. – Theory Methods. (2001), 1-19. https:// doi:10.1080/03610926.2021.1995434.

[3] 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.

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

[5] 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.

[6] 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].

[7] S. Fu, A hierarchical Bayesian approach to negative binomial regression, Methods Appl. Anal. 22 (2015) 409–428. https://doi.org/10.4310/maa.2015.v22.n4.a4.

[8] 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.

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

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

[11] 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.

[12] E. Gómez-Déniz, J.M. Sarabia, E. Calderín-Ojeda, Univariate and multivariate versions of the negative binomial-inverse Gaussian distributions with applications, Insurance: Math. Econ. 42 (2008) 39–49. https://doi.org/10.1016/j.insmatheco.2006.12.001.

[13] M. Greenwood, G.U. Yule, An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, J. R. Stat. Soc. 83 (1920), 255-279. https://doi.org/10.2307/2341080.

[14] H. He, W. Tang, W. Wang, et al. Structural zeroes and zero-inflated models, Shanghai Arch. Psychiatry, 26 (2014), 236-242. https://doi.org/10.3969/j.issn.1002-0829.2014.04.008.

[15] J.S. Long, Regression models for categorical and limited dependent variables, Advanced quantitative techniques in the social sciences series 7, Sage Publication, Thousand Oaks, CA, (1997).

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

[17] 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..

[18] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, (2022). https://www. Rproject.org/. [Access 12 January 2022].

[19] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, et al. Bayesian measures of model complexity and fit, J. R. Stat. Soc. B. 64 (2002), 583–639. https://doi.org/10.1111/1467-9868.00353.

[20] 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.

[21] Y.S. Su, M. Yajima, M.Y.S Su, J.A.G.S System Requirements. Package `R2jags'; R package version 0.03-08, (2015), http://CRAN. R.project.org/package=R2jags. [Access 12 October 2021].

[22] R. Tharshan, P. Wijekoon, A modification of the quasi Lindley distribution, Open J. Stat. 11 (2021), 369-392. https://doi.org/10.4236/ojs.2021.113022.

[23] 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.

[24] D. Yamrubboon, W. Bodhisuwan, C. Pudprommarat, et al. The negative binomial-Sushila distribution with application in count data analysis, Thailand Statistician, 15 (2011), 69-77.

[25] 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.

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