Soft Computing (2024) 28:10911–10928
https://doi.org/10.1007/s00500-024-09889-4

Multiple arbitrarily inflated negative binomial regression model and its application

Ihab Abusaif[1] · Coşkun Kuş[1]

Accepted: 7 May 2024 / Published online: 26 July 2024

Abstract

This paper introduces a novel modification of the negative binomial distribution, which serves as a generalization encompassing both negative binomial and zero-inflated negative binomial distributions. This innovative distribution offers flexibility by accommodating an arbitrary number of inflation points at various locations. The paper explores key distributional properties associated with this modified distribution. Additionally, this study proposes several estimators designed to obtain estimates for the unknown parameters. Furthermore, the paper introduces a new count regression model that utilizes the modified distribution. To assess the performance of the proposed distribution and the count regression model, a comprehensive Monte Carlo simulation study is conducted. In the final stage of the paper, a real-world dataset is scrutinized to ascertain the superiority of the proposed model. This empirical analysis contributes to validating the practical applicability and effectiveness of the newly introduced distribution in comparison to existing models.

Keywords Count data · EM algorithm · Fisher scoring algorithm · Zero-inflated poisson distribution · Zero-inflated negative binomial distribution

1 Introduction

Count data is prevalent across various fields, spanning biomedical research, criminology, environmental economics, political science, industry, and sociology. However, conventional statistical models, such as Poisson or negative binomial (NB) regression, may fall short in adequately capturing the excess observed at certain points in the data. This inflation can be attributed to structural factors that give rise to a distinct process generating points in addition to the primary count-generating process.

To address this issue, inflated counting models incorporate several components. One such component involves the counting process to model the remaining observations, while another employs binary operations to model the excess points. Numerous researchers have enhanced both inflated negative binomial NB and inflated Poisson distributions. For example, Greene (1994) explored the zero-inflated negative binomial (ZINB) as a modified model for the Poisson regression model. Additionally, the zero-inflated Poisson distribution (ZIP) was introduced by Lambert (1992) to handle count data exhibiting both zeros and large counts.

Indeed, real data derived from various studies may exhibit excesses at certain points, and this excess is not exclusively confined to zeros. As a result, researchers have directed their focus towards identifying distributions that are well-suited for points with both excess zeros and non-zero excesses. One solution proposed is the zero–one inflated negative binomial (ZOINB) distribution, an inflated version of the nb distribution, as suggested by Alshkaki (2017). Another alternative is the zero–one inflated Poisson (ZOIP) distribution, an inflated version of the Poisson distribution, introduced by Melkersson and Olsson (1999). Recently Sun et al. (2021) proposed the zero–one–two inflated Poisson distribution (ZOTIP) as a model for count data characterized by excess zeros, ones, and twos. This distribution encompasses the zero-inflated Poisson (ZIP) and zero-and-one-inflated Poisson (ZOIP) distributions. These distributions are inadequate in fitting the data when the number of observations with inflation exceeds two. In such cases, Su et al. (2013) proposed a Multiple Inflation Poisson (MIP) model to analyze data with multiple inflated values, where these inflated values extend consecutively from zero to any value up to m. Saboori and

✉ Ihab Abusaif
censtat@gmail.com

[1] Department of Statistics, Faculty of Science, Selcuk University, 42250 Konya, Turkey

Doostparast (2023) introduced the Zero to k Inflated Poisson Regression Model, enabling inflated points to be three or even more. Another inflated version of the Poisson distribution is the zero- and k-inflated Poisson distribution model (ZkIP), proposed by Arora and Chaganty (2021). Additionally, Serra and Polestico (2023) put forth a zero- and k-inflated negative binomial (ZkINB) distribution. This innovative inflated version of the distributions is suitable when, beyond zero, the frequency of another count, denoted as k, tends to be higher in the data.

The assumption of zero, one, two inflation, or inflation at consecutive values of m may often not hold for real-world data. For example, an individual's daily cigarette consumption could exhibit inflation at instances of 0, 1, 10, and 20. In such cases, conventional methodologies may prove ineffective. Therefore, there is a demand for models that can accommodate inflation at arbitrary counts and positions. It is also noteworthy that in a prior paper, Serra and Polestico (2023) presented a new modification of the Poisson distribution, encompassing the generalizations of the Poisson, zero-inflated Poisson, zero–one inflated Poisson, and zero–one–two inflated Poisson distributions. In this current study, we introduce a flexible negative binomial (NB) distribution that permits an arbitrary number of inflation points at various positions simultaneously.

In this paper, we consider a new flexible extension of the NB distribution. The paper is organized as follows: In Sect. 2, the extension of the NB distribution is described and some distributional properties are studied. Section 3 discusses parameter estimation using several methods. The Monte Carlo simulation study is also performed to observe the performance of the estimators. Numerical examples are provided based on real data. In Sect. 4, a count regression model based on the proposed distribution is presented. A Monte Carlo simulation study is conducted to observe the performance of the maximum likelihood (ML) estimators. Numerical examples are also provided based on real data. Section 5 provides concluding remarks.

2 Proposed distribution and distributional properties

In this section, we introduce the Multiple Arbitrarily Inflated Negative Binomial (MAINB) distribution and discuss some distributional properties. Let W be a random variable from a degenerate distribution at a single point b denoted by $\omega \sim \text{Degenerate}\,(b)$. For $a_i, i = 1, 2, \ldots, k$, let $W_i \sim \text{Degenerate}\,(a_i)$, $X \sim \text{NB}\,(\theta, \mu)$ with probability mass function (pmf)

$$Pr(X = x) = \frac{\Gamma\,(x+\theta)}{x!\Gamma\,(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{x}, \quad x = 0, 1, 2, \ldots \tag{1}$$

and they are assumed to be independent. A discrete random variable Y is said to follow a MAINB distribution, denoted by $Y \sim \text{MAINB}_D\,(\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu)$, if its pmf is

$$\begin{aligned} f\,(y|\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu) &= \sum_{i=1}^{k} \phi_i \Pr\,(W_i = y) + \phi^* \Pr\,(X = y) \\ &= \begin{cases} \phi_1 + \phi^* \frac{\Gamma(a_1+\theta)}{a_1!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_1} & , y = a_1 \\ \phi_2 + \phi^* \frac{\Gamma(a_2+\theta)}{a_2!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_2} & , y = a_2 \\ \vdots & \vdots \\ \phi_k + \phi^* \frac{\Gamma(a_k+\theta)}{a_k!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_k} & , y = a_k \\ \phi^* \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{y} & , y \in \mathbb{N} - D \end{cases} \end{aligned}$$

or equivalently

$$\begin{aligned} f\,(y|\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu) &= \sum_{i=1}^{k} \left\{ \left(\phi_i + \phi^* \frac{\Gamma\,(a_i+\theta)}{a_i!\Gamma\,(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \right.\right. \\ &\quad \left.\left. \left(\frac{\mu}{\theta+\mu}\right)^{a_i} \right) I\,(y = a_i) \right\} \\ &\quad + \left(\phi^* \frac{\Gamma\,(y+\theta)}{y!\Gamma\,(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{y} \right) \\ &\quad I\,(y \in \mathbb{N} - D)\,, \end{aligned} \tag{2}$$

where $D = \{a_1, a_2, \ldots, a_k\}$, $\phi_1, \phi_2, \ldots, \phi_k$ are the proportions of structured inflated points, μ is the mean of the negative binomial distribution, θ is the dispersion parameter, $\phi^* = 1 - \sum_{i=1}^{k} \phi_i$ and $\phi^*, \phi_1, \phi_2, \ldots, \phi_k \in [0, 1)$.

In particular, when $a_1 = 0$, $a_2 = 1$ and $\phi_3 = \phi_4 = \cdots = \phi_k = 0$, the MAINB distribution is reduced to the ZIONB distribution given in Alshkaki (2017), when $a_1 = 0$ and $\phi_2 = \phi_3 = \cdots = \phi_k = 0$, the MAINB distribution is reduced to the ZINB distribution given in Greene (1994), when $\phi_1 = \phi_2 = \cdots = \phi_k = 0$, the MAINB distribution is reduced to the traditional NB distribution.

The cumulative distribution function (CDF) of $Y \sim \text{MAINB}_D\,(\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu)$ is given by

$$\Pr\,(Y \leqslant y) = \sum_{i=1}^{k^*} \phi_i + \phi^* \sum_{i=0}^{\lfloor y \rfloor} \frac{\Gamma\,(i+\theta)}{i!\Gamma\,(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{i}, \tag{3}$$

where $\lfloor a \rfloor$ is the floor function of a and $k^* = \#\Big\{ j : a_j \leq y, j = 1, 2, \ldots, k \Big\}$.

To derive the moments of the random variable $Y \sim \mathrm{MAINB}_D(\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu)$, we present a stochastic representation (SR), which is a useful tool to obtain the raw moments of Y.

Let the random vector $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_k, Z^*)^{\mathrm{T}} \sim \mathrm{Multinomial}_{k+1}(1; \phi_1, \phi_2, \ldots, \phi_k, \phi^*)$ and the random variable $X \sim \mathrm{NB}(\theta, \mu)$. Here, $\boldsymbol{Z}$ and X are mutually independent. The (SR) of the random variable $Y \sim \mathrm{MAINB}_D(\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu)$ is

$$
\begin{aligned}
Y &\overset{d}{=} \sum_{i=1}^{k} Z_i a_i + Z^* X \\
&= \begin{cases} a_1, & \text{with probability } \phi_1 \\ a_2, & \text{with probability } \phi_2 \\ \vdots & \qquad \vdots \\ a_k, & \text{with probability } \phi_k \\ X, & \text{with probability } \phi^*, \end{cases}
\end{aligned} \tag{4}
$$

where $\sum_{i=1}^{k} Z_i + Z^* = 1$, $\Pr(Z_i = 1) = \phi_i$, $i = 1, \ldots, k$ and $\Pr(Z^* = 1) = \phi^*$. Now we can re-write the pmf of Y by using SR in Eq. (4) as follows:

$$
\begin{aligned}
&\Pr(Y = y) \\
&= \begin{cases} \Pr(Z_1 = 1) + \Pr(Z^* = 1, X = a_1) \\ \Pr(Z_2 = 1) + \Pr(Z^* = 1, X = a_2) \\ \vdots \\ \Pr(Z_k = 1) + \Pr(Z^* = 1, X = a_k) \\ \Pr(Z^* = 1, X = y) \end{cases} \\
&= \begin{cases} \phi_1 + \phi^* \frac{\Gamma(a_1+\theta)}{a_1!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_1} & , y = a_1 \\ \phi_2 + \phi^* \frac{\Gamma(a_2+\theta)}{a_2!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_2} & , y = a_2 \\ \vdots & \quad \vdots \\ \phi_k + \phi^* \frac{\Gamma(a_k+\theta)}{a_k!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_k} & , y = a_k \\ \phi^* \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{y} & , y \in \mathbb{N} - D \end{cases}
\end{aligned} \tag{5}
$$

that means the $\mathrm{MAINB}_D(\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu)$ is a mixture of $k+1$ distributions: Degenerate(a_1), Degenerate(a_2),..., Degenerate(a_k) and NB(θ, μ). From Eq. (4) it is clear that $E(Z_i) = \phi_i$, $E(Z^*) = \phi^*$, $Var(Z_i) = \phi_i(1 - \phi_i)$, $Var(Z^*) = \phi^*(1 - \phi^*)$, $E\left(Z_i Z_j\right) = 0$, $i \neq j$, $i, j = 1, \ldots, k$ and $E(Z_i Z^*) = 0$, $i = 1, \ldots, k$. Using these facts, the first two moments and the variance of the $\mathrm{MAINB}_D(\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu)$ distribution are obtained as

$$
E(Y) = \sum_{i=1}^{k} a_i \phi_i + \phi^* \mu, \tag{6}
$$

$$
E\left(Y^2\right) = \sum_{i=1}^{k} a_i^2 \phi_i + \phi^* \left(\mu + \mu^2 + \frac{\mu^2}{\theta}\right), \tag{7}
$$

and

$$
\begin{aligned}
Var(Y) = \sum_{i=1}^{k} a_i^2 \phi_i + \phi^* & \\
\left(\mu + \mu^2 + \frac{\mu^2}{\theta}\right) &- \left(\sum_{i=1}^{k} a_i \phi_i + \phi^* \mu\right)^2,
\end{aligned} \tag{8}
$$

respectively. Similarly, the n^{th} moment can be written as

$$
E\left(Y^n\right) = \sum_{i=1}^{k} a_i^n \phi_i + \phi^* E\left(X^n\right), \tag{9}
$$

where X has pmf given in Eq. (1). The moment generation function of the $\mathrm{MAINB}_D(\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu)$ distribution is also obtained as

$$
\begin{aligned}
M_Y(t) = e^{a_1 t} &\left\{\phi_1 + \phi^* P(X = a_1)\right\} + \cdots \\
&+ e^{a_k t} \left\{\phi_k + \phi^* P(X = a_k)\right\} \\
&+ \phi^* \sum_{y \in \mathbb{N} - D} e^{yt} P(X = y) \\
&= \sum_{i=1}^{k} \phi_i e^{a_i t} + \phi^* M_X(t) \\
&= \sum_{i=1}^{k} \phi_i e^{a_i t} + \phi^* \left(\frac{\theta}{\theta + \mu - \mu e^t}\right)^{\theta}.
\end{aligned}
$$

3 Parameter estimation

This section proposes several methods for estimating unknown parameters of the MAINB distribution. We discuss the traditional ML estimation method and derive Fisher scoring (FS) and expectation-maximization (EM) algorithms to obtain the ML estimates of parameters of interest. We also discuss all of the following estimation methods: least squares (LS) (Swain et al. 1988), weighted least squares (WLS), Cramér-von Mises (CvM) type (Choi and Bulgren 1968) and the three types of proportion (P_1), (P_2) and (P_3) estimation methods suggested by Bakouch et al. (2021). We compare their performance based on simulated samples.

3.1 Maximum likelihood estimation

Assume that $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \mathrm{MAINB}_D(\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu)$ and $y_1, y_2, \ldots, y_n$ denote their realizations. Let $\boldsymbol{y} = \{y_i\}_{i=1}^{n}$ be the observed data, and define $\mathbb{I}_j = \{i : y_i = a_j,$

$1 \leq i \leq n\}$ and $m_j = \sum_{i=1}^{n} \mathbb{I}\left(y_i = a_j\right)$ as the number of elements in $\mathbb{I}_j$, $j = 1, 2, \ldots, k$. Thus, the log-likelihood function can be written as

$$\begin{aligned}
&\ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right) \\
&\quad = \sum_{j=1}^{k} m_j \log\left(\phi_j + \phi^* \frac{\Gamma\left(a_j+\theta\right)}{a_j!\Gamma\left(\theta\right)}\right. \\
&\qquad \left.\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_j}\right) \\
&\qquad + \left(n - \sum_{j=1}^{k} m_j\right)\Big[\log\left(\phi^*\right) \\
&\qquad - \log\left(\Gamma\left(\theta\right)\right) + \theta \log\left(\frac{\theta}{\theta+\mu}\right)\Big] \\
&\qquad + \sum_{i \notin \cup_{j=1}^{k}\mathbb{I}_j} \Big[\log\left(\Gamma\left(y_i+\theta\right)\right) \\
&\qquad - \log\left(y_i!\right)\Big] + N \log\left(\frac{\mu}{\mu+\theta}\right), \qquad (10)
\end{aligned}$$

where $\boldsymbol{\Theta} = \left(\phi_1, \phi_2, \ldots, \phi_k, \theta, \mu\right)^{\mathrm{T}}$ is the parameter vector and $N = \sum_{i \notin \cup_{j=1}^{k}\mathbb{I}_j} y_i$.

Using the log-likelihood function given in Eq. (10), the ML estimator of the parameter vector $\boldsymbol{\Theta}$ is obtained by

$$\hat{\boldsymbol{\Theta}}_1 = \arg\max_{\boldsymbol{\Theta}} \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right). \qquad (11)$$

The ML estimator described in Eq. (11) can be obtained by utilizing basic search methods. For this task, the **R** function **optim** or some meta-heuristic algorithms can be considered suitable tools. The next two subsections provide two alternative search methods to these tools.

3.1.1 Fisher scoring algorithm

In this subsection, we can calculate the ML estimator of $\boldsymbol{\Theta}$ using FS. The elements of the score vector $\nabla\ell$ are obtained as follows:

$$\begin{aligned}
&\frac{\partial \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right)}{\partial \phi_h} \\
&\quad = \frac{m_h\left[1 - \frac{\Gamma(a_h+\theta)}{a_h!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_h}\right]}{\phi_h + \phi^* \frac{\Gamma(a_h+\theta)}{a_h!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_h}} \\
&\qquad - \sum_{\substack{i=1 \\ i \neq h}}^{k} \frac{m_i\left[\frac{\Gamma(a_i+\theta)}{a_i!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_i}\right]}{\phi_i + \phi^* \frac{\Gamma(a_i+\theta)}{a_i!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_i}} \\
&\qquad - \frac{n - \sum_{i=1}^{k} m_i}{\phi^*}, \; h = 1, 2, \ldots, k, \\
&\frac{\partial \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right)}{\partial \mu} \\
&\quad = \sum_{i=1}^{k} \frac{m_i \phi^* \Gamma\left(a_i+\theta\right)\theta^{\theta+1}\mu^{a_i-1}\left(\mu+\theta\right)^{-a_i-\theta-1}\left(a_i-\mu\right)}{\phi_i \Gamma\left(\theta\right) a_i! + \phi^* \Gamma\left(a_i+\theta\right)\theta^{\theta}\mu^{a_i}\left(\mu+\theta\right)^{-a_i-\theta}} \\
&\qquad - \frac{\left(n - \sum_{j=1}^{k} m_j\right)\theta}{\mu+\theta} + \frac{N\theta}{\mu\left(\mu+\theta\right)}, \\
&\frac{\partial \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right)}{\partial \theta} \\
&\quad = \sum_{i=1}^{k} \frac{m_i\left[\Psi\left(a_i+\theta\right) - \Psi\left(\theta\right) + \log\left(\frac{\theta}{\mu+\theta}\right) - \frac{a_i-\mu}{\mu+\theta}\right]}{1 + \frac{\phi\Gamma(\theta)a_i!(\mu+\theta)^{a_i+\theta}}{\phi^*\Gamma(a_i+\theta)\theta^{\theta}\mu^{a_i}}} \\
&\qquad + \left(n - \sum_{j=1}^{k} m_j\right)\left[\frac{\mu}{\mu+\theta} + \log\left(\frac{\theta}{\mu+\theta}\right) - \Psi\left(\theta\right)\right] \\
&\qquad + \sum_{i \notin \cup_{j=1}^{k}\mathbb{I}_j} \Psi\left(y_i+\theta\right) - \frac{N}{\left(\theta+\mu\right)},
\end{aligned}$$

where $\Psi\left(\cdot\right)$ is the digamma function. The elements of Hessian matrix $\nabla^2\ell$ are also given by

$$
\begin{aligned}
\frac{\partial^2 \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right)}{\partial \phi_h^2} &= -\frac{m_h\left[1-\frac{\Gamma(a_h+\theta)}{a_h!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_h}\right]^2}{\left[\phi_h+\phi^*\frac{\Gamma(a_h+\theta)}{a_h!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_h}\right]^2} - \sum_{\substack{i=1\\ i\neq h}}^{k}\frac{m_i\left[\frac{\Gamma(a_i+\theta)}{a_i!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_i}\right]^2}{\left[\phi_i+\phi^*\frac{\Gamma(a_i+\theta)}{a_i!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_i}\right]^2} - \frac{n-\sum_{i=1}^{k} m_i}{(\phi^*)^2},\ \ h=1,2,\ldots,k,\\
\frac{\partial^2 \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right)}{\partial \phi_h \phi_s} &= \frac{m_h\left[1-\frac{\Gamma(a_h+\theta)}{a_h!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_h}\right]\left[\frac{\Gamma(a_h+\theta)}{a_h!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_h}\right]}{\left[\phi_h+\phi^*\frac{\Gamma(a_h+\theta)}{a_h!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_h}\right]^2}\\
&\quad+\frac{m_s\left[1-\frac{\Gamma(a_s+\theta)}{a_s!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_s}\right]\left[\frac{\Gamma(a_s+\theta)}{a_s!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_s}\right]}{\left[\phi_s+\phi^*\frac{\Gamma(a_s+\theta)}{a_s!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_s}\right]^2} - \sum_{\substack{i=1\\ i\neq h\\ i\neq s\\ h\neq s}}^{k}\frac{m_i\left[\frac{\Gamma(a_i+\theta)}{a_i!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_i}\right]^2}{\left[\phi_i+\phi^*\frac{\Gamma(a_i+\theta)}{a_i!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{a_i}\right]^2}\\
&\quad-\frac{n-\sum_{i=1}^{k} m_i}{(\phi^*)^2},\ \ h,s=1,2,\ldots,k,\\
\frac{\partial^2 \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right)}{\partial \mu^2} &= \sum_{i=1}^{k}\frac{m_i\phi^*\theta^{\theta+1}\mu^{a_i-2}\Gamma\left(a_i+\theta\right)\left[\left((a-\mu)^2-a\right)\theta-2a\mu+\mu^2\right]}{(\mu+\theta)^2\left(\phi_i\Gamma\left(\theta\right)a_i!\left(\mu+\theta\right)^{a_i+\theta}+\phi^*\Gamma\left(a_i+\theta\right)\theta^{\theta}\mu^{a_i}\right)}\\
&\quad-\sum_{i=1}^{k}\frac{m_i\left(\phi^*\left(a_i-\mu\right)\Gamma\left(a_i+\theta\right)\theta^{\theta+1}\mu^{a_i-1}\right)^2}{(\mu+\theta)^2\left(\phi_i\Gamma\left(\theta\right)a_i!\left(\mu+\theta\right)^{a_i+\theta}+\phi^*\Gamma\left(a_i+\theta\right)\theta^{\theta}\mu^{a_i}\right)^2}+\frac{\left(n-\sum_{j=1}^{k} m_j\right)\theta}{(\mu+\theta)^2}-\frac{\left(2\mu+\theta\right)N\theta}{\mu^2\left(\mu+\theta\right)^2},\\
\frac{\partial^2 \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right)}{\partial \mu\partial\phi_h} &= -\frac{m_h a_h!\left(\phi_h+\phi^*\right)\Gamma\left(a_h+\theta\right)\Gamma\left(\theta\right)\theta^{\theta+1}\mu^{a_h-1}\left(\mu+\theta\right)^{-a_i-\theta-1}\left(a_h-\mu\right)}{\left(\phi_h a_h!\Gamma\left(\theta\right)+\phi^*\Gamma\left(a_h+\theta\right)\theta^{\theta}\mu^{a_h}\left(\mu+\theta\right)^{-a_i-\theta}\right)^2}\\
&\quad-\sum_{\substack{i=1\\ i\neq h}}^{k}\frac{m_i a_i!\phi_i\Gamma\left(a_i+\theta\right)\Gamma\left(\theta\right)\theta^{\theta+1}\mu^{a_i-1}\left(\mu+\theta\right)^{-a_i-\theta-1}\left(a_i-\mu\right)}{\left(\phi_i a_i!\Gamma\left(\theta\right)+\phi^*\Gamma\left(a_i+\theta\right)\theta^{\theta}\mu^{a_i}\left(\mu+\theta\right)^{-a_i-\theta}\right)^2},\ \ h=1,2,\ldots,k,\\
\frac{\partial^2 \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right)}{\partial \theta\partial\phi_h} &= \frac{m_h a_h!\left(\phi_h+\phi^*\right)\Gamma\left(a_h+\theta\right)\Gamma\left(\theta\right)\theta^{\theta}\mu^{a_h}\left(\mu+\theta\right)^{-a_h-\theta}}{\left(\phi_h a_h!\Gamma\left(\theta\right)+\phi^*\Gamma\left(a_h+\theta\right)\theta^{\theta}\mu^{a_h}\left(\mu+\theta\right)^{-a_h-\theta}\right)^2}\left[-\Psi\left(a_h+\theta\right)+\Psi\left(\theta\right)-\log\left(\frac{\theta}{\mu+\theta}\right)+\frac{a_h-\mu}{\mu+\theta}\right]\\
&\quad+\sum_{\substack{i=1\\ i\neq h}}^{k}\left\{\frac{m_i a_i!\phi_i\Gamma\left(a_i+\theta\right)\Gamma\left(\theta\right)\theta^{\theta}\mu^{a_i}\left(\mu+\theta\right)^{-a_i-\theta}}{\left(\phi_i a_i!\Gamma\left(\theta\right)+\phi^*\Gamma\left(a_i+\theta\right)\theta^{\theta}\mu^{a_i}\left(\mu+\theta\right)^{-a_i-\theta}\right)^2}\left[-\Psi\left(a_i+\theta\right)+\Psi\left(\theta\right)-\log\left(\frac{\theta}{\mu+\theta}\right)+\frac{a_i-\mu}{\mu+\theta}\right]\right\},\ \ h=1,2,\ldots,k,\\
\frac{\partial^2 \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right)}{\partial \theta\partial\mu} &= \sum_{i=1}^{k}\left\{\frac{m_i\phi^*\theta^{\theta+1}\mu^{a_i-1}\left(a_i-\mu\right)\Gamma\left(a_i+\theta\right)}{\left(\theta+\mu\right)\left(\phi_i\Gamma\left(\theta\right)a_i!\left(\mu+\theta\right)^{a_i+\theta}+\phi^*\Gamma\left(a_i+\theta\right)\theta^{\theta}\mu^{a_i}\right)}\right.\\
&\quad\times\left.\left(\log\left(\frac{\theta}{\mu+\theta}\right)-\frac{\left(a_i-\mu\right)\theta-\mu}{\theta\left(\mu+\theta\right)}+\Psi\left(a_i+\theta\right)\right)\right\}+\sum_{i=1}^{k}\left\{\frac{m_i\theta\left(a_i-\mu\right)\left(\theta^{\theta}\mu^{a_i}\left(\mu+\theta\right)^{-\theta-a_i}\Gamma\left(a_i+\theta\right)\phi^*\right)^2}{\mu\left(\mu+\theta\right)\left(\phi^*\Gamma\left(a_i+\theta\right)\theta^{\theta}\mu^{a_i}\left(\mu+\theta\right)^{-\theta-a_i}+\phi\Gamma\left(\theta\right)a_i!\right)^2}\right.\\
&\quad\times\left.\left(-\log\left(\frac{\theta}{\mu+\theta}\right)-\Psi\left(a_i+\theta\right)+\frac{a_i-\mu}{\mu+\theta}\right)\right\}-\sum_{i=1}^{k}\frac{m_i\phi a_i!\phi^*\theta^{\theta+1}\mu^{a_i-1}\left(\mu+\theta\right)^{-\theta-a_i-1}\left(a_i-\mu\right)\Gamma\left(a_i+\theta\right)\Gamma\left(\theta\right)\Psi\left(\theta\right)}{\left(\phi^*\Gamma\left(a_i+\theta\right)\theta^{\theta}\mu^{a_i}\left(\mu+\theta\right)^{-\theta-a_i}+\phi\Gamma\left(\theta\right)a_i!\right)^2}\\
&\quad-\frac{\mu\left(n-\sum\limits_{i=1}^{k} m_i\right)-N}{\left(\mu+\theta\right)^2},\\
\frac{\partial^2 \ell\left(\boldsymbol{\Theta}|\boldsymbol{y}\right)}{\partial \theta^2} &= \sum_{i=1}^{k}\frac{m_i\left[\Psi\left(1,a_i+\theta\right)-\Psi\left(1,\theta\right)+\theta^{-1}-\left(\mu+\theta\right)^{-1}+\frac{a_i-\mu}{(\mu+\theta)^2}\right]}{1+\frac{\phi\Gamma(\theta)a_i!(\mu+\theta)^{a_i+\theta}}{\phi^*\Gamma(a_i+\theta)\theta^{\theta}\mu^{a_i}}}\\
&\quad+\sum_{i=1}^{k}\left\{\frac{m_i\phi_i\phi^*\theta^{\theta}\mu^{a_i}\left(\theta+\mu\right)^{\left(a_i+\theta\right)}\Gamma\left(\theta\right)a_i!\Gamma\left(a_i+\theta\right)}{\left[\phi_i\Gamma\left(\theta\right)a_i!\left(\theta+\mu\right)^{\left(a_i+\theta\right)}+\phi^*\theta^{\theta}\mu^{a_i}\Gamma\left(a_i+\theta\right)\right]^2}\left(-\Psi\left(a_i+\theta\right)+\Psi\left(\theta\right)-\log\left(\frac{\theta}{\mu+\theta}\right)+\frac{a_i-\mu}{\mu+\theta}\right)^2\right\}\\
&\quad-\left(n-\sum_{j=1}^{k} m_j\right)\left[\Psi\left(1,\theta\right)-\frac{\mu^2}{\theta\left(\mu+\theta\right)^2}\right]+\sum_{i\notin\cup_{j=1}^{k}\mathbb{I}_j}\Psi\left(1,y_i+\theta\right)+\frac{N}{\left(\theta+\mu\right)^2},
\end{aligned}
$$

where $\Psi(n, x)$ is the n-th polygamma function, which is the n-th derivative of the digamma function when n is a nonnegative integer.

We need the following expectations to obtain the elements of the Fisher Information matrix and the E-step of the EM algorithm. The expectations of m_j, $j = 1, \ldots, k$ and N are given by

$$E(m_j) = n\left(\phi_j + \phi^* \frac{\Gamma(a_j+\theta)}{a_j!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_j}\right), \tag{12}$$

$$\begin{aligned} E(N) &= E\left(\sum_{i=1}^{k} y_i - \sum_{i=1}^{k} a_i m_i\right) \\ &= nE(Y_1) - n\sum_{i=1}^{k} a_i \left(\phi_i + \phi^* \frac{\Gamma(a_i+\theta)}{a_i!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_i}\right) \\ &= n\left[\sum_{i=1}^{k} a_i\phi_i + \phi^*\mu - \sum_{i=1}^{k} a_i \left(\phi_i + \phi^* \frac{\Gamma(a_i+\theta)}{a_i!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_i}\right)\right] \\ &= n\phi^*\left[\mu - \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \sum_{i=1}^{k} a_i \frac{\Gamma(a_i+\theta)}{a_i!\Gamma(\theta)} \left(\frac{\mu}{\theta+\mu}\right)^{a_i}\right], \end{aligned} \tag{13}$$

respectively. Using Eqs. (12) and (13), the expected Fisher information matrix $\boldsymbol{I}(\boldsymbol{\Theta})$ is given by

$$\boldsymbol{I}(\boldsymbol{\Theta}) = E\left[-\nabla^2 \ell(\boldsymbol{\Theta}|\boldsymbol{y})\right].$$

Thus, the consistent estimators of standard errors of $\hat{\boldsymbol{\Theta}}$ are calculated by the square root of the diagonal elements of the inverse Fisher information matrix $\Sigma = \boldsymbol{I}^{-1}(\boldsymbol{\Theta})$ evaluated at $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}$. In order to obtain the estimate of our parameter vector $\boldsymbol{\Theta}$, let $\boldsymbol{\Theta}^{(0)} = \left(\phi_1^{(0)}, \phi_2^{(0)}, \ldots, \phi_k^{(0)}, \theta^{(0)}, \mu^{(0)}\right)^{\mathrm{T}}$ be an initial value and $\boldsymbol{\Theta}^{(t)} = \left(\phi_1^{(t)}, \phi_2^{(t)}, \ldots, \phi_k^{(t)}, \theta^{(t)}, \mu^{(t)}\right)^{\mathrm{T}}$ be the t-th approximation of $\hat{\boldsymbol{\Theta}}$, then the $(t+1)$-th approximation can be written using the FS algorithm:

$$\boldsymbol{\Theta}^{(t+1)} = \boldsymbol{\Theta}^{(t)} + \boldsymbol{I}^{-1}\left(\boldsymbol{\Theta}^{(t)}\right) \nabla\ell\left(\boldsymbol{\Theta}^{(t)}|\boldsymbol{y}\right). \tag{14}$$

The $(1-\alpha)100\%$ asymptotic confidence intervals (CIs) of ϕ_j, $j = 1, \ldots, k, \theta$ and μ are given by

$$\hat{\phi}_j \pm z_{\alpha/2}\sqrt{\Sigma^{jj}}, \tag{15}$$

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\Sigma^{(k+1)(k+1)}}, \tag{16}$$

and

$$\hat{\mu} \pm z_{\alpha/2}\sqrt{\Sigma^{(k+2)(k+2)}}, \tag{17}$$

respectively, where z_α is the αth upper quantile of the standard normal distribution.

3.1.2 Expectation-maximization algorithm

In this subsection, we give an EM algorithm as an alternative to the FS algorithm. The a_i observations from any MAINB distribution can be classed into the extra observations from the degenerate distribution at point a_i because of the population variability and the structural a_i resulting from an ordinary NB distribution. Thus, we can write

$$\mathbb{I}_j = \mathbb{I}_j^{\text{extra}} \cup \mathbb{I}_j^{\text{structral}}, \quad j = 1, 2, \ldots, k.$$

Let W_j be the number of elements in $\mathbb{I}_j^{\text{extra}}$ used to split m_j into w_j and $m_j - w_j$ for $j = 1, \ldots, k$, where w_j is the realization of W_j. Thus, the result of conditional predictive distribution for $W_j|(\boldsymbol{y}, \boldsymbol{\Theta})$ is

$$W_j|(\boldsymbol{y}, \boldsymbol{\Theta}) \sim \text{ Binomial } \left(m_j, \frac{\phi_j}{\phi_j + \phi^* \frac{\Gamma(a_j+\theta)}{a_j!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_j}}\right).$$

The complete data log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\Theta}|\boldsymbol{y}, \boldsymbol{w}) = &\sum_{i=1}^{k} w_i \log(\phi_i) + \left(n - \sum_{i=1}^{k} w_i\right) \left\{\log(\phi^*) - \log(\Gamma(\theta)) + \theta \log\left(\frac{\theta}{\theta+\mu}\right)\right\} \\ &+ \left(N + \sum_{i=1}^{k} a_i(m_i - w_i)\right) \log\left(\frac{\mu}{\theta+\mu}\right) \\ &+ \sum_{i=1}^{k} (m_i - w_i) \log\left(\frac{\Gamma(a_i+\theta)}{a_i!}\right) \\ &+ \sum_{i \notin \cup_{j=1}^{k} \mathbb{I}_j} \log\left(\frac{\Gamma(y_i+\theta)}{y_i!}\right), \end{aligned}$$

where $\boldsymbol{w} = (w_1, w_2, \cdots, w_k)$. The M-step is used to find the ML estimators of the complete data:

$$\hat{\phi}_i = \frac{w_i}{n}, \tag{18}$$

$$\hat{\mu} = \frac{N + \sum_{i=1}^{k} a_i \left(m_i - w_i\right)}{n - \sum_{i=1}^{k} w_i}, \tag{19}$$

and $\hat{\theta}$ can be determined by solving the following equation:

$$\begin{aligned} \frac{\partial \ell \left(\mathbf{\Theta} | \boldsymbol{y}, \boldsymbol{w}\right)}{\partial \theta} = & \left(n - \sum_{i=1}^{k} w_i\right) \\ & \left\{\frac{\mu}{\mu + \theta} + \log\left(\frac{\theta}{\mu + \theta}\right) - \Psi\left(\theta\right)\right\} \\ & + \frac{N + \sum_{i=1}^{k} a_i \left(m_i - w_i\right)}{\theta + \mu} \\ & + \sum_{i \notin \cup_{j=1}^{k} \mathbb{I}_j} \Psi\left(y_i + \theta\right) \\ & + \sum_{i=1}^{k} \left(m_i - w_i\right) \Psi\left(a_i + \theta\right) = 0. \end{aligned} \tag{20}$$

The E-step is used to replace w_i in Eqs. (18–20) by their conditional expectation

$$\begin{aligned} & E\left(W_i | \left(\boldsymbol{y}, \boldsymbol{\theta}\right)\right) \\ & \quad = \frac{m_i \phi_i}{\phi_i + \phi^* \frac{\Gamma(a_i+\theta)}{a_i!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{a_i}}. \end{aligned}$$

3.2 Least-squares and weighted least-squares estimation

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from the MAINB_D $(\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu)$ distribution, $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ denotes the corresponding order statistics and $y_{(j)}$ denotes the observed value of $Y_{(j)}$ for $j = 1, 2, \ldots, n$. Then, the LS estimate of $\mathbf{\Theta} = (\phi_1, \phi_2, \ldots, \phi_k, \theta, \mu)^{\mathrm{T}}$ is obtained by

$$\begin{aligned} \hat{\mathbf{\Theta}}_2 & = \underset{\mathbf{\Theta}}{\arg\min} \\ & \left\{\sum_{j=1}^{n} \left[F_Y\left(y_{(j)}; \mathbf{\Theta}\right) - \frac{\delta}{n+1}\right]^2\right\}, \end{aligned} \tag{21}$$

where $F_Y\left(y_{(j)}; \mathbf{\Theta}\right)$ is the estimate of the cdf at $y_{(j)}$ and $\delta = \sum_{i=1}^{n} I_{\{y_i \leq y_{(j)}\}}$. The WLS estimate of $\mathbf{\Theta}$ is

$$\begin{aligned} \hat{\mathbf{\Theta}}_3 = & \underset{\mathbf{\Theta}}{\arg\min} \\ & \left\{\sum_{j=1}^{n} \frac{(n+2)(n+1)^2}{\delta(n-\delta+1)}\right. \\ & \left.\left[F_Y\left(y_{(j)}; \mathbf{\Theta}\right) - \frac{\delta}{n+1}\right]^2\right\}. \end{aligned} \tag{22}$$

3.3 Cramér-von mises estimator

The CvM estimator is a type of minimum distance estimator. This approach attempts to estimate the parameters to minimize the squared distance between the cdf and ecdf estimations. The CvM estimate of $\mathbf{\Theta}$ is defined as

$$\hat{\mathbf{\Theta}}_4 = \underset{\mathbf{\Theta}}{\arg\min} \left\{\frac{1}{12n} \sum_{j=1}^{n} \left[F_Y\left(y_{(j)}; \mathbf{\Theta}\right) - \frac{2\delta - 1}{n+1}\right]^2\right\}. \tag{23}$$

3.4 Proportions estimator

To estimate the unknown parameters of the discrete CosPois distribution, Bakouch et al. (2021) presented three types of proportion estimators. A similar concept may be applied to estimate the MAINB parameters. Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from the MAINB_D $(\phi_1, \phi_2, \ldots, \phi_k; \theta, \mu)$ distribution. Also, let $y_1, y_2, \ldots, y_m, m \leq n$ be the realization of this sample with frequencies $b_1, b_2, \ldots, b_m$, respectively. Then, the proportion estimators can be defined as

$$\hat{\mathbf{\Theta}}_5 = \underset{\mathbf{\Theta}}{\arg\min} \left\{\sum_{j=1}^{m} \left[f_Y\left(y_{(j)}; \mathbf{\Theta}\right) - \frac{b_j}{n}\right]^2\right\}, \tag{24}$$

$$\hat{\mathbf{\Theta}}_6 = \underset{\mathbf{\Theta}}{\arg\min} \left\{\sum_{j=1}^{m} \left| f_Y\left(y_{(j)}; \mathbf{\Theta}\right) - \frac{b_j}{n}\right|\right\}, \tag{25}$$

and

$$\hat{\mathbf{\Theta}}_7 = \underset{\mathbf{\Theta}}{\arg\min} \left\{\max \left| f_Y\left(y_{(j)}; \mathbf{\Theta}\right) - \frac{b_j}{n}\right|\right\}, \tag{26}$$

where $f_Y\left(y; \mathbf{\Theta}\right)$ is the pmf given in Eq. (2). It is noted that we use the **optim** function of the **R** software to implement the optimization algorithm to find the direct maximization and minimization in Eq. (10) and Eqs. (21–26).

3.5 Simulation study

In this subsection, we implement a Monte Carlo simulation to examine the ML, LS, WLS, CvM, P_1, P_2 and P_3 estimation methods. The bias and the mean squared error (MSE) of parameters are estimated using 5000 trials. We study the case where the inflated values are at the points (0, 1, 3, 5).

The proportion of structured inflated points is set as $\phi_1 = 0.2$, $\phi_2 = 0.2$, $\phi_3 = 0.1$, and $\phi_4 = 0.1$. Various sample sizes are considered in this simulation study, with $n = 50, 100, 250, 500, 1000$. We set the true values of the μ and θ parameters to be 2 and 0.5, respectively.

Table 1 presents the bias and MSEs of the estimators. The results indicate that, as the sample size increases, the bias and MSEs of all estimation methods decrease. Furthermore,

Table 1 The bias and MSE of estimators for the MAINB$_{\{0,1,3,5\}}$ (0.2, 0.2, 0.1, 0.1; 2, 0.5) parameters using different methods

Method		ML		LS		WLS		CvM		P$_1$		P$_2$		P$_3$	
n	Par	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
50	$\hat{\phi}_1$	−0.0121	0.0176	−0.0466	0.0153	−0.0623	0.0198	−0.0243	0.0133	−0.0014	0.0152	0.0237	0.0096	−0.0765	0.0251
	$\hat{\phi}_2$	−0.0036	0.0071	−0.0140	0.0061	−0.0110	0.0063	−0.0096	0.0065	−0.0053	0.0068	0.0028	0.0059	−0.0222	0.0068
	$\hat{\phi}_3$	0.0046	0.0028	−0.0067	0.0027	−0.0034	0.0029	−0.0041	0.0029	0.0005	0.0028	0.0042	0.0028	−0.0135	0.0031
	$\hat{\phi}_4$	0.0048	0.0024	−0.0104	0.0026	−0.0087	0.0026	−0.0065	0.0027	0.0010	0.0024	0.0035	0.0024	−0.0122	0.0027
	$\hat{\mu}$	−0.0253	0.2862	0.2752	0.2672	0.3010	0.3415	0.1396	0.1972	0.1310	0.3384	0.0377	0.0830	0.4364	0.5140
	$\hat{\theta}$	0.1727	0.3975	0.0201	0.0892	−0.0065	0.1100	0.1149	0.1478	0.1511	0.2066	0.0877	0.0504	−0.0224	0.0615
100	$\hat{\phi}_1$	0.0031	0.0087	−0.0151	0.0072	−0.0208	0.0078	−0.0064	0.0068	−0.0041	0.0077	0.0188	0.0049	−0.0329	0.0091
	$\hat{\phi}_2$	0.0013	0.0033	−0.0038	0.0030	−0.0020	0.0032	−0.0020	0.0031	−0.0003	0.0032	0.0041	0.0030	−0.0084	0.0032
	$\hat{\phi}_3$	0.0016	0.0014	−0.0027	0.0014	−0.0005	0.0015	−0.0015	0.0014	−0.0001	0.0014	0.0017	0.0014	−0.0073	0.0015
	$\hat{\phi}_4$	0.0025	0.0012	−0.0044	0.0013	−0.0033	0.0013	−0.0025	0.0013	0.0006	0.0012	0.0020	0.0012	−0.0062	0.0013
	$\hat{\mu}$	0.0342	0.2312	0.1793	0.1528	0.1875	0.1806	0.0984	0.1264	0.1338	0.2478	0.0605	0.0465	0.3185	0.2798
	$\hat{\theta}$	0.1569	0.2286	0.0471	0.0699	0.0281	0.0880	0.0963	0.0910	0.0952	0.1477	0.0646	0.0262	−0.0123	0.0374
250	$\hat{\phi}_1$	0.0169	0.0058	0.0060	0.0026	0.0052	0.0027	0.0096	0.0027	0.0032	0.0036	0.0167	0.0021	−0.0035	0.0032
	$\hat{\phi}_2$	0.0027	0.0013	−0.0009	0.0012	−0.0005	0.0013	−0.0002	0.0012	0.0011	0.0012	0.0022	0.0012	−0.0020	0.0013
	$\hat{\phi}_3$	0.0004	0.0005	−0.0010	0.0005	−0.0006	0.0006	−0.0007	0.0005	0.0001	0.0005	0.0008	0.0005	−0.0027	0.0006
	$\hat{\phi}_4$	0.0001	0.0004	−0.0021	0.0005	−0.0019	0.0005	−0.0014	0.0005	−0.0004	0.0004	0.0003	0.0004	−0.0033	0.0005
	$\hat{\mu}$	0.1373	0.2047	0.1341	0.0687	0.1331	0.0739	0.1022	0.0578	0.1285	0.1212	0.0812	0.0274	0.2255	0.1460
	$\hat{\theta}$	0.1449	0.1837	0.0478	0.0294	0.0441	0.0350	0.0694	0.0344	0.0383	0.0435	0.0453	0.0119	0.0070	0.0198
500	$\hat{\phi}_1$	0.0184	0.0047	0.0101	0.0014	0.0108	0.0014	0.0117	0.0014	0.0074	0.0021	0.0154	0.0011	0.0081	0.0015
	$\hat{\phi}_2$	0.0027	0.0006	0.0001	0.0006	−0.0001	0.0006	0.0004	0.0006	0.0011	0.0006	0.0013	0.0006	0.0004	0.0006
	$\hat{\phi}_3$	−0.0006	0.0003	−0.0012	0.0003	−0.0016	0.0003	−0.0011	0.0003	−0.0005	0.0003	−0.0004	0.0003	−0.0016	0.0003
	$\hat{\phi}_4$	−0.0002	0.0002	−0.0010	0.0002	−0.0011	0.0003	−0.0007	0.0002	−0.0003	0.0002	0.0000	0.0002	−0.0015	0.0002
	$\hat{\mu}$	0.1648	0.1882	0.1193	0.0457	0.1189	0.0471	0.1024	0.0406	0.1117	0.0806	0.0854	0.0219	0.1714	0.0924
	$\hat{\theta}$	0.1199	0.0870	0.0427	0.0160	0.0493	0.0202	0.0532	0.0170	0.0379	0.0256	0.0439	0.0081	0.0253	0.0133
1000	$\hat{\phi}_1$	0.0159	0.0041	0.0122	0.0008	0.0132	0.0009	0.0133	0.0008	0.0114	0.0014	0.0155	0.0007	0.0127	0.0009
	$\hat{\phi}_2$	0.0026	0.0003	0.0008	0.0003	0.0007	0.0003	0.0010	0.0003	0.0015	0.0003	0.0014	0.0003	0.0011	0.0003
	$\hat{\phi}_3$	−0.0006	0.0001	−0.0009	0.0001	−0.0012	0.0001	−0.0009	0.0001	−0.0005	0.0001	−0.0005	0.0001	−0.0010	0.0001
	$\hat{\phi}_4$	−0.0004	0.0001	−0.0008	0.0001	−0.0009	0.0001	−0.0006	0.0001	−0.0003	0.0001	−0.0002	0.0001	−0.0008	0.0001
	$\hat{\mu}$	0.1500	0.1531	0.1047	0.0304	0.1085	0.0348	0.0982	0.0287	0.1028	0.0556	0.0860	0.0171	0.1320	0.0573
	$\hat{\theta}$	0.0937	0.0529	0.0370	0.0081	0.0450	0.0116	0.0437	0.0091	0.0400	0.0144	0.0419	0.0054	0.0341	0.0091

Table 2 Comparison results for emergency room grouped data

Count	Observed	NB			$\mathrm{MAINB}_{\{2\text{-}3,10\text{-}12,>12\}}$		
0	10046	10039.96			10047.002		
1	1466	1526.974			1460.767		
2–3	548	434.26			548		
4–5	92	142.652			103.685		
6–7	37	50.006			31.285		
8–9	12	18.191			9.766		
10–12	13	6.778			13		
> 12	9	2.568			9		
χ^2		77.52471			2.89116		
df		5			2		
p-value		<0.001			0.23561		
Par		Est	SE	p-value	Est	SE	p-value
ϕ_1					0.0150	0.00219	<0.001
ϕ_2					0.0008	0.00030	0.0068
ϕ_3					0.0007	0.00025	0.0079
μ		0.2607	0.00605	<0.001	0.2250	0.00656	<0.001
θ		0.3650	0.01887	<0.001	0.4109	0.02593	<0.001
$-\ell$		7719.474			7684.347		
AIC		15442.95			15378.69		

all estimates can be considered asymptotically unbiased and consistent.

3.6 Application with real data

In the previous subsection, we verify the validity and suitability of the proposed distribution via simulation. In this subsection, we consider the data given in Table 2 from the NHIS 2015 database on children under 18 years. The count variable is the number of children to visit the emergency room in one year. According to the adjusted data set, we have a total of $12,223$ children. Note that this data is also utilized in Arora and Chaganty (2021).

Previous studies about inflated distributions have suggested that inflation may occur at one or more points when the observed frequency is higher than expected or simply when the percentage of its presence in the data is quite large. In this study, to determine whether the data contains inflation, we developed the following algorithm: Utilizing Algorithm 1 and fitting the data to a NB distribution, we infer that the data exhibits inflation at counts (2-3,10-12,>12).

The Pearson chi-square statistic χ^2 is the most commonly employed statistic for conducting goodness-of-fit tests on count data. Table 2 shows that χ^2 of $\mathrm{MAINB}_{\{2\text{-}3,10\text{-}12,>12\}}$ is less than that of NB. Therefore, $\mathrm{MAINB}_{\{2\text{-}3,10\text{-}12,>12\}}$ is preferred to NB for fitting the grouped data. Furthermore, the p-value for the NB distribution is < 0.05, leading to the conclusion that the data is not consistent with an NB distribution. We employ the ML estimation method to fit both the NB and $\mathrm{MAINB}_{\{2\text{-}3,10\text{-}12,>12\}}$ distributions to the emergency room grouped data. The negative log-likelihood $(-\ell)$ and the Akaike information criterion (AIC) are used for result comparison.

The outcomes presented in Table 2 reveal that the $\mathrm{MAINB}_{\{2\text{-}3,10\text{-}12,>12\}}$ distribution offers the best fit to the data, surpassing the proposals in previous articles by ZOIP (Arora 2018). Additionally, it is demonstrated that all $\mathrm{MAINB}_{\{2\text{-}3,10\text{-}12,>12\}}$ parameters are significant, indicating the presence of inflation in the counts proposed in our study.

To determine the ML estimate of the parameters through the FS and EM algorithms (MLEs), the initial values of the parameters $\left(\phi_1^{(0)}, \phi_2^{(0)}, \phi_3^{(0)}, \mu^{(0)}, \theta^{(0)}\right)$ are set to $(0.05, 0.0005, 0.0005, 0.16, 0.1)$. We compute the FS algorithm in Eq. (14) and the EM algorithm in Eqs. (18–20), which converges to the MLEs estimates as reported in Table 3. The standard errors (SE) and 95% approximate CIs in Eqs. (15–17) are also presented in Table 3.

The iterations of the algorithms FS and EM are plotted in Fig. 1 and the FS algorithm appears to be faster than EM in achieving the ML estimates.

In addition to the suggested EM and FS algorithms, the following meta-heuristic algorithms were also employed in this example as alternatives: Grasshopper Optimization Algorithm (GOA) (Saremi et al. 2017), Ant Lion Optimizer (ALO) (Mirjalili 2015), Grey Wolf Optimizer (GWO) (Mirjalili

Algorithm 1

```
1: Compute the expected frequency according to the distribution we are interested in.
2: Compute the ratio (expected frequency / observed frequency).
3: Choose the appropriate tolerance ($t$); in this study we set $t = 0.2$.
4: Based on Step 3, inflation exists if the ratio is less than $1 - t$.
```

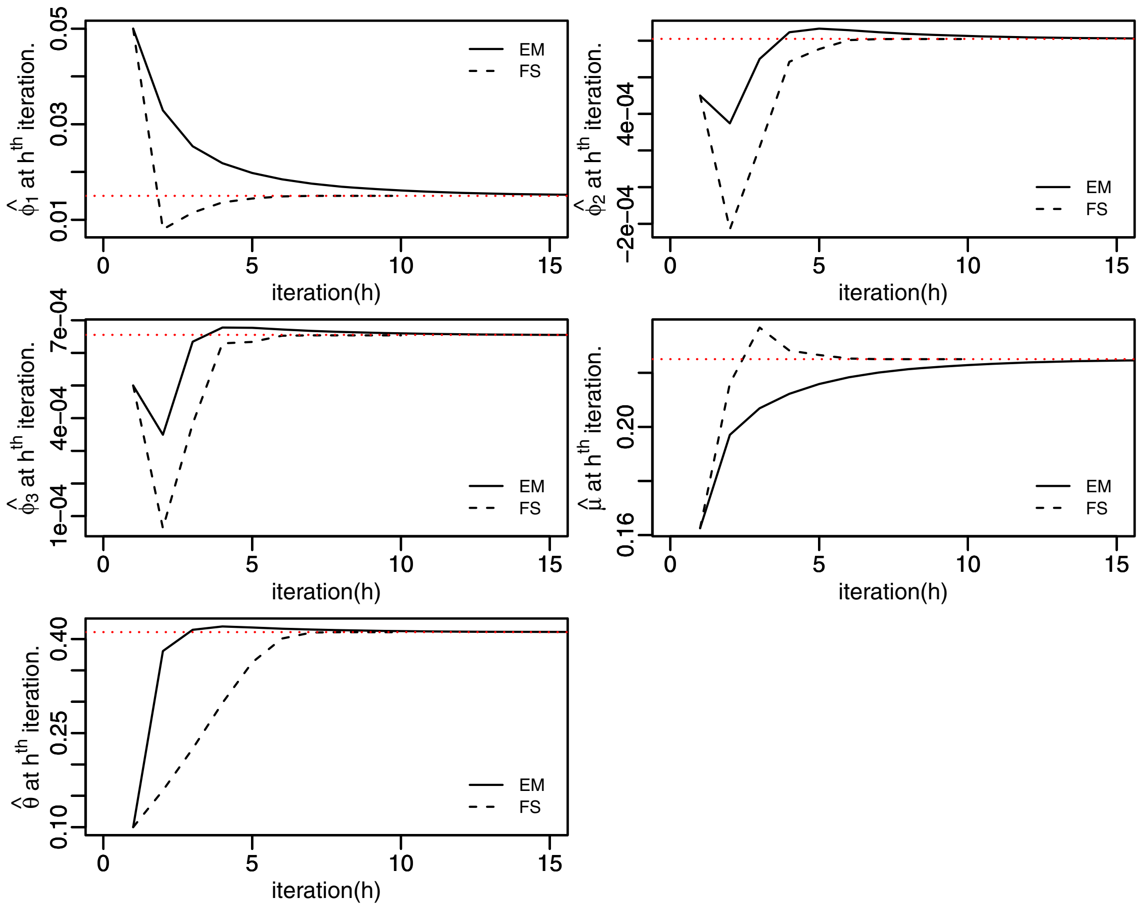

Fig. 1 Trace plot for MAINB$_{\{2\text{-}3,10\text{-}12,>12\}}$ parameters under EM and FS iterations

Table 3 MLEs, SE and approximate 95% CIs of parameters for the emergency room grouped data

Par	MLEs	SE	95% CIs
ϕ_1	0.01500	0.00219	(0.01457,0.01543)
ϕ_2	0.00081	0.00030	(0.00075,0.00087)
ϕ_3	0.00065	0.00025	(0.00061,0.00070)
μ	0.22501	0.00655	(0.22373,0.22630)
θ	0.41090	0.02592	(0.40582,0.41598)

et al. 2014), Moth Flame Optimizer (MFO) (Mirjalili 2015) and Black Hole Optimization Algorithm (BHO) (Hatamlou 2013). The parameter estimates and $-\ell$ values obtained from these algorithms are provided in Table 4. Note that the parameter values obtained through the meta-heuristic algorithms are closely similar to those achieved by the FS and EM algorithms.

Furthermore, we estimate the MAINB$_{\{2\text{-}3,10\text{-}12,>12\}}$ parameters by the methods in Sect. 3 and the results are presented in Table 5.

4 A count regression model

Count regression models are frequently employed in the analysis of count data to elucidate the behavior of dependent variables, which can be either metric or dummy variables. Commonly used count regression models include Poisson and NB regression models. Suppose the independent responses $Y_i, i = 1, 2, \ldots, n$ constitute a sample from the distribution MAINB$_D$ $(\phi_{1i}, \phi_{2i}, \ldots, \phi_{ki}; \theta_i, \mu_i)$. In the following, the parameters μ_i and $\phi_{1i}, \phi_{2i}, \ldots, \phi_{ki}$ are

Table 4 Result of MAINB$_{\{2\text{-}3,10\text{-}12,>12\}}$ parameter estimation using different meta-heuristic algorithms

Par	GOA	ALO	GWO	MFO	BHO
ϕ_1	0.0150024	0.0150019	0.0149963	0.0150020	0.0166253
ϕ_2	0.0008087	0.0008086	0.0008085	0.0008086	0.0021583
ϕ_3	0.0006557	0.0006537	0.0006502	0.0006537	0.0020133
μ	0.2250087	0.2250112	0.2250356	0.2250110	0.2256817
θ	0.4109142	0.4109026	0.4109648	0.4109025	0.4487292
$-\ell$	7684.347	7684.347	7684.347	7684.347	7698.34

Table 5 Result of MAINB$_{\{2\text{-}3,10\text{-}12,>12\}}$ parameter estimation using different methods

Par	ML	LS	WLS	CvM	P$_1$	P$_2$	P$_3$
$\hat{\phi}_1$	0.015	0.10761	0.09026	0.12556	0.01579	0.01509	0.01611
$\hat{\phi}_2$	0.00081	−0.01108	−0.00357	−0.01057	0.00095	0.00064	0.00052
$\hat{\phi}_3$	0.00065	0.00011	0.00084	−0.00029	0.00074	0.00066	0.00068
$\hat{\mu}$	0.22501	0.33994	0.29582	0.34602	0.2211	0.22452	0.22157
$\hat{\theta}$	0.4109	0.33649	0.39638	0.34782	0.4301	0.41691	0.42014

linked to the covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and $\boldsymbol{z}_{ji} = (z_{ji1}, \ldots, z_{jiq})^{\mathrm{T}}$, $j = 1, 2, \ldots, k$, respectively. Here p and q are the number of independent variables, usually $x_{i1} = z_{ji1} = 1$ for all i. Thus, the log-linear and logit forms can be defined as

$$\log(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \quad i = 1, 2, \ldots, n, \tag{27}$$

and

$$\operatorname{logit}\left(\phi_{ji}\right) = \boldsymbol{z}_{ji}^{\mathrm{T}}\boldsymbol{\gamma}_j, \;\; j = 1, 2, \ldots, k; \quad i = 1, 2, \ldots, n, \tag{28}$$

where $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \ldots, \beta_p\right)^{\mathrm{T}}$ and $\boldsymbol{\gamma}_j = \left(\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jq}\right)^{\mathrm{T}}$ are vectors of the regression parameters. Substituting Eqs. (27) and (28) in Eq. (10), the log-likelihood function is written by

$$\begin{aligned}
\ell_R(\Xi|\boldsymbol{y}) = {} & \sum_{i=1}^{n} I\{y_i = a_1\} \log\left(\phi_{1i} + \phi_i^* \frac{\Gamma(a_1+\theta)}{a_1!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{a_1}\right) \\
& + \sum_{i=1}^{n} I\{y_i = a_2\} \log\left(\phi_{2i} + \phi_i^* \frac{\Gamma(a_2+\theta)}{a_2!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{a_2}\right) \\
& + \cdots + \sum_{i=1}^{n} I\left\{y_i = a_j\right\} \log\left(\phi_{ji} + \phi_i^* \frac{\Gamma\left(a_j+\theta\right)}{a_j!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{a_j}\right) \\
& + \sum_{i=1}^{n} I\{y_i \in \mathbb{N} - D\} \log\left(\phi_i^* \frac{\Gamma(y_i+\theta)}{y_i!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{y_i}\right),
\end{aligned}$$

where $\Xi = \left(\boldsymbol{\beta}, \boldsymbol{\gamma}_j\right)$, μ_i, ϕ_{ji} and D are given by Eqs. (27), (28) and (2) respectively, and $\phi_i^* = 1 - \sum_{j=1}^{k} \phi_{ji} \in [0, 1)$. The ML estimator of Ξ is given by

$$\widehat{\Xi} = \underset{\Xi}{\arg\max} \quad \ell_R(\Xi|\boldsymbol{y}).$$

The ML estimate of the unknown parameter vector Ξ can be obtained using a search algorithm. It is noted that the **R** function **optim** can be used to maximize the log-likelihood function.

4.1 Simulation study

In this subsection, we assess the performance of the MAINB regression model through the ML estimation method.

In this subsection, we evaluate the performance of the MAINB regression model using the ML estimation method. The assessment involves 2000 replications, and the sample size is varied as $n = 50$, 100, 250, 500, and 1000. Log-linear and logit linear forms in Eq. (27) and Eq. (28) are employed for regression structures. The covariates for the parameter μ are generated from a standard normal distribution, while the covariates for ϕ_1, ϕ_2, and ϕ_3 parameters are generated from a uniform distribution with parameters $(-1.5, -0.5)$.

Table 6 The AE, bias and MSE of estimators for the MAINB$_{\{0,2,4\}}$ regression model parameters

n		$\hat{\gamma}_{01}$	$\hat{\gamma}_{11}$	$\hat{\gamma}_{21}$	$\hat{\gamma}_{02}$	$\hat{\gamma}_{12}$	$\hat{\gamma}_{22}$	$\hat{\gamma}_{03}$	$\hat{\gamma}_{13}$	$\hat{\gamma}_{23}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\theta}$
50	AE	−0.3078	2.2947	1.1020	−0.0031	1.2576	2.4513	0.8353	2.3562	0.9140	0.4779	0.5273	0.8235	1.9062
	Bias	−0.8078	0.2947	0.1020	−0.5031	0.2576	0.4513	−0.1647	0.3562	−0.0860	−0.0221	0.0273	0.0235	1.1062
	MSE	23.1255	8.0466	7.0556	8.7718	6.3046	9.7011	4.9642	9.6601	5.7801	0.1371	0.1673	0.1579	4.1860
100	AE	0.0697	2.0361	1.1097	0.4316	1.1648	2.1492	1.0642	2.1564	1.0529	0.4953	0.4999	0.8131	1.2997
	Bias	−0.4303	0.0361	0.1097	−0.0684	0.1648	0.1492	0.0642	0.1564	0.0529	−0.0047	−0.0001	0.0131	0.4997
	MSE	3.4937	3.5993	3.2284	2.4416	3.0591	3.2984	1.7855	1.9819	1.7077	0.0597	0.0472	0.0451	1.0091
250	AE	0.2582	2.0651	1.0454	0.4961	1.0494	2.0707	1.0638	2.0964	1.0251	0.4987	0.4995	0.7987	0.9570
	Bias	−0.2418	0.0651	0.0454	−0.0039	0.0494	0.0707	0.0638	0.0964	0.0251	−0.0013	−0.0005	−0.0013	0.1570
	MSE	1.4288	2.1895	1.9486	0.8069	1.0966	1.1749	0.5344	0.5786	0.5214	0.0256	0.0135	0.0153	0.1479
500	AE	0.3746	2.0395	0.9671	0.5241	1.0389	2.0336	1.0503	2.0439	1.0389	0.5068	0.4998	0.8023	0.8777
	Bias	−0.1254	0.0395	−0.0329	0.0241	0.0389	0.0336	0.0503	0.0439	0.0389	0.0068	−0.0003	0.0023	0.0777
	MSE	0.6306	1.1963	1.0865	0.4348	0.4976	0.5771	0.2178	0.2426	0.2120	0.0118	0.0073	0.0069	0.0513
1000	AE	0.4717	2.0603	0.9890	0.5353	1.0245	2.0451	1.0390	2.0304	1.0214	0.4996	0.5010	0.7993	0.8359
	Bias	−0.0284	0.0603	−0.0110	0.0353	0.0245	0.0451	0.0390	0.0304	0.0214	−0.0005	0.0010	−0.0008	0.0359
	MSE	0.2819	0.6631	0.5101	0.2050	0.2276	0.2609	0.1017	0.1098	0.0924	0.0053	0.0032	0.0037	0.0187

Table 7 Frequency distribution of the number of cigarettes smoked per day

Count	0	1	2	3	4	5	6	7	8	9	10
Frequency	2730	17	20	26	22	29	24	17	17	4	118
Count	11	12	13	14	15	16	17	18	20	22	23
Frequency	3	10	6	3	61	2	2	3	212	1	1

The true values of the regression parameters for ϕ_1, ϕ_2, ϕ_3, and μ are pre-specified as ($\gamma_{01} = 0.5, \gamma_{11} = 2.0, \gamma_{21} = 1.0$), ($\gamma_{02} = 0.5, \gamma_{12} = 1.0, \gamma_{22} = 2.0$), ($\gamma_{03} = 1.0, \gamma_{13} = 2.0, \gamma_{23} = 1.0$), and ($\beta_0 = 0.5, \beta_1 = 0.5, \beta_2 = 0.8$) respectively. The true value of θ is set to 0.8.

The response variable Y_i is generated using the specified parameters. Then, the **optim** function in **R** is employed to obtain the ML estimates of the parameters of the MAINB regression model.

The simulation results are interpreted based on the average of estimates (AE), bias, and MSE. The results, as presented in Table 6, indicate that as the sample size increases, the bias and MSE of all estimates decrease. Additionally, the AE remains stable for all sample sizes and is close to the true parameter values.

4.2 Application with real data

In this subsection, we consider a data set to demonstrate the purposed MAINB regression model. The data is collected by TurkStat “http://www.tuik.gov.tr” for the Turkey Health Survey in 2014. This data set was analyzed by Tüzen and Erbaş (2018) and they saw that the data set contains a very large percentage of zeros, indicating excess zeros in the data. In our example, we define the dependent variable as the number of daily smoked cigarettes. Table 7 shows the frequency of the dependent variable, it is clear that we have a high proportion of zeros, tens, fifteens, and twenties. The covariates used in this example are presented in Table 8.

The NB, ZIP, ZINB and MAINB$_{\{0,10,15,20\}}$ regression models are used to analyze the cigarette data. To evaluate the performance of the model, we employ AIC and negative log-likelihood ($-\ell$). Table 10 summarizes the results from this model, including the AIC and $-\ell$. In this example, the

Table 8 Description of covariates

Variable name	Definition	variable label (Frequency)
Gender(female)	What is your sex? “Male” and “Female”	= 1 if respondent is female (1794)
Education(primary)	What is your completed level of education? “No school completed”, “Primary education”, “High school” and“Higher education”	= 1 if primary education (1853)
Education(high school)		= 1 if high school (963)
Education(higher education)		= 1 if higher education (351)
Marital(married)	What is your marital status? “Single” and “Married”	= 1 if married (472)
Employment(not.working)	In the last week, have you ever worked¿‘Working” and “Not working”	= 1 if not working(2469)
Income(1081-1550.TL)	What is your household’s total net income per month? “0-1080 TL”, “1081-1550 TL”,“1551-2170 TL”, “2171-3180 TL” and “3181+ TL”	= 1 if 1081-1550 TL (653)
Income(1551-2170.TL)		= 1 if 1551-2170 TL (573)
Income(2171-3180.TL)		= 1 if 2171-3180 TL (656)
Income(3181+.TL)		= 1 if 3181+ TL (464)

Table 9 Model validation results of the models

	NB model	ZIP model	ZINB model	MAINB model
MSE	27.761403	26.193524	26.204056	26.188256
RMSE	5.268909	5.117961	5.118990	5.117446
MAE	3.168169	3.135915	3.135717	3.144213

Table 10 Results for the numerical data

Par	Reg.par.	NB		ZIP		ZINB		MAINB	
		Est	SE	Est	SE	Est	SE	Est	SE
μ	β_0	2.1610*	0.3471	2.8933*	0.0526	2.8779*	0.1083	2.1055*	0.2454
	β_1	−1.8359*	0.1441	−0.3860*	0.0339	−0.3955*	0.0611	−0.3195*	0.1226
	β_2	−0.2085	0.3258	−0.1202*	0.0527	−0.0974	0.1081	−0.0915	0.2340
	β_3	0.0612	0.3405	−0.1863*	0.0552	−0.1598	0.1130	−0.3211	0.2496
	β_4	−0.1293	0.3804	−0.2414*	0.0635	−0.2287*	0.1272	−0.3256	0.2828
	β_5	0.7608*	0.2017	−0.0554	0.0356	−0.0575	0.0684	−0.0082	0.1425
	β_6	−1.0581*	0.1583	−0.1196*	0.0242	−0.1232*	0.0475	−0.0251	0.1081
	β_7	−0.1288	0.1974	−0.1319*	0.0349	−0.1428*	0.0688	−0.2011	0.1584
	β_8	−0.0801	0.2065	−0.0728*	0.0348	−0.0777	0.0695	0.1353	0.1496
	β_9	−0.0194	0.1990	−0.0497	0.0341	−0.0476	0.0684	0.1443	0.1627
	β_{10}	−0.4357*	0.2259	−0.1619*	0.0395	−0.1837*	0.0768	−0.0588	0.1738
ϕ_1	γ_{01}			−0.0019	0.2449	−0.0019	0.2451	−0.0104	0.2456
	γ_{11}			1.6574*	0.1183	1.6527*	0.1184	1.6225*	0.1173
	γ_{21}			0.3217	0.2375	0.3206	0.2377	0.3188	0.2361
	γ_{31}			0.0531	0.2461	0.0506	0.2463	0.0730	0.2455
	γ_{41}			0.2877	0.2752	0.2843	0.2755	0.3028	0.2743
	γ_{51}			−0.7863*	0.1526	−0.7886*	0.1529	−0.7576*	0.1513
	γ_{61}			1.1509*	0.1020	1.1505*	0.1021	1.1443*	0.1034
	γ_{71}			0.0011	0.1459	−0.0010	0.1461	−0.0449	0.1535
	γ_{81}			−0.1029	0.1498	−0.1040	0.1499	−0.0907	0.1531
	γ_{91}			0.0645	0.1473	0.0639	0.1474	0.1019	0.1482
	γ_{101}			0.0073	0.1620	0.0048	0.1622	−0.0295	0.1634
ϕ_2	γ_{02}							−3.8785*	1.2993
	γ_{12}							−1.3414*	0.2784
	γ_{22}							1.0458	1.2303
	γ_{32}							1.7948	1.2000
	γ_{42}							1.4940	1.2321
	γ_{52}							0.5551*	0.2959
	γ_{62}							−0.9196*	0.2179
	γ_{72}							0.5275	0.3221
	γ_{82}							−0.4406	0.4102
	γ_{92}							0.2934	0.3165
	γ_{102}							−0.0895	0.4078
ϕ_3	γ_{03}							−4.4497*	0.9670
	γ_{13}							−1.5847*	0.3973
	γ_{23}							0.9100	0.8741
	γ_{33}							0.6000	0.8274
	γ_{43}							1.2214	0.9909
	γ_{53}							0.5487	0.4076
	γ_{63}							−0.7793*	0.2926
	γ_{73}							0.2729	0.4624
	γ_{83}							0.6036	0.4605
	γ_{93}							0.9429*	0.4229
	γ_{103}							0.7471	0.4705

Table 10 continued

Par	Reg.par.	NB		ZIP		ZINB		MAINB	
		Est	SE	Est	SE	Est	SE	Est	SE
ϕ_4	γ_{04}							-0.2976	0.2772
	γ_{14}							-2.3812^*	0.2510
	γ_{24}							-0.7996^*	0.2865
	γ_{34}							-0.7461^*	0.3045
	γ_{44}							-1.2011^*	0.3832
	γ_{54}							0.3510	0.2363
	γ_{64}							-1.1888^*	0.1534
	γ_{74}							-0.3437	0.2191
	γ_{84}							-0.2616	0.2139
	γ_{94}							-0.5521^*	0.2127
	γ_{104}							-0.4357^*	0.2513
θ		0.0717^*	0.0037			4.4852^*	0.0878	3.5381^*	0.6862
AIC		7611.48		7229.49		6602.464		5251.161	
$-\ell$		3793.74		3592.745		3278.232		2569.58	

*significant at the level $\alpha = 0.05$

Table 11 LRT test results of models

Models	χ^2	p value
NB - ZINB	1031	< 0.001
ZINB - MAINB	1417.303	< 0.001

values taken by the MSE, RMSE, and MAE criteria as alternatives to $-\ell$ are also presented in Table 9.

According to the results of the four models, the zero-inflated models fitted the cigarette data better than the underlying models. Also, the proposed model is preferred over the others.

A likelihood ratio test (LRT) compares the goodness of fit of two nested regression models. We perform LRT to determine if zero-inflated and underlying models and also if MAINB$_{\{0,10,15,20\}}$ and ZINB are significantly different. The results in Table 11 confirm that the proposed model is more suitable than the other models for fitting the data.

4.2.1 Residual analysis

Analyzing residuals is crucial for assessing and validating the accuracy of a regression model. Residuals provide insights into detecting outliers and assessing the correctness of assumptions about the error distribution. A method proposed by Atkinson (1981) involves constructing simulated envelopes for residuals with an unknown distribution, using Pearson residuals for the residual analysis.

Given that the distribution of standardized residuals is unknown, simulated envelopes are generated to identify potential misspecifications in the error distribution or the presence of outliers. Consequently, residuals that fall outside the envelope can be considered outliers or indicative of a model mis-specification. Therefore, the adequacy of the fitted model improves as the proportion of points outside the envelope decreases.

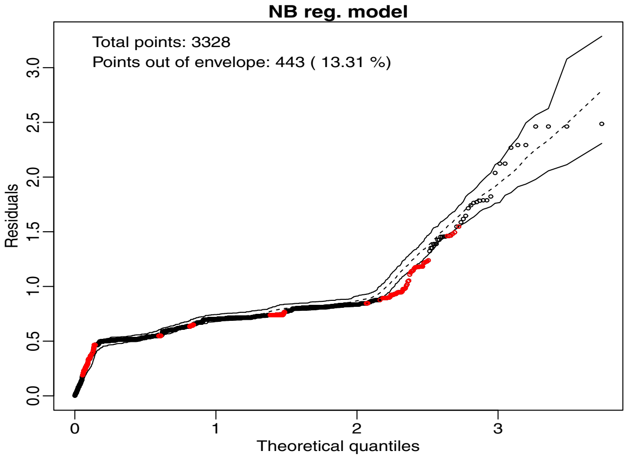

Fig. 2 Simulated envelopes for the Pearson residuals in the NB regression model

Figures 2, 3, 4 and 5 show the simulated envelopes for the Pearson residuals of the NB, ZIP, ZINB and MAINB$_{\{0,10,15,20\}}$ regression models, respectively. The graphs' lines display the median (dashed line), 5th percentile, and 95th percentile values (solid lines) for each observation out of 100 simulated points. The algorithm presented in Lemonte et al. (2019) is used to plot the simulated envelopes for the Pearson residuals.

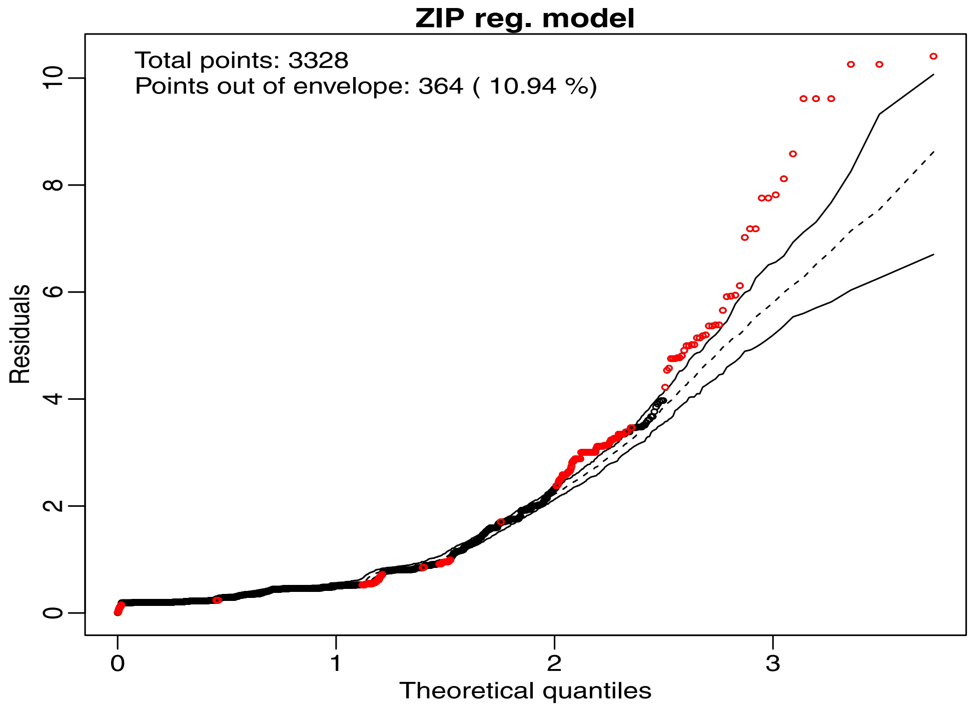

Fig. 3 Simulated envelopes for the Pearson residuals in the ZIP regression model

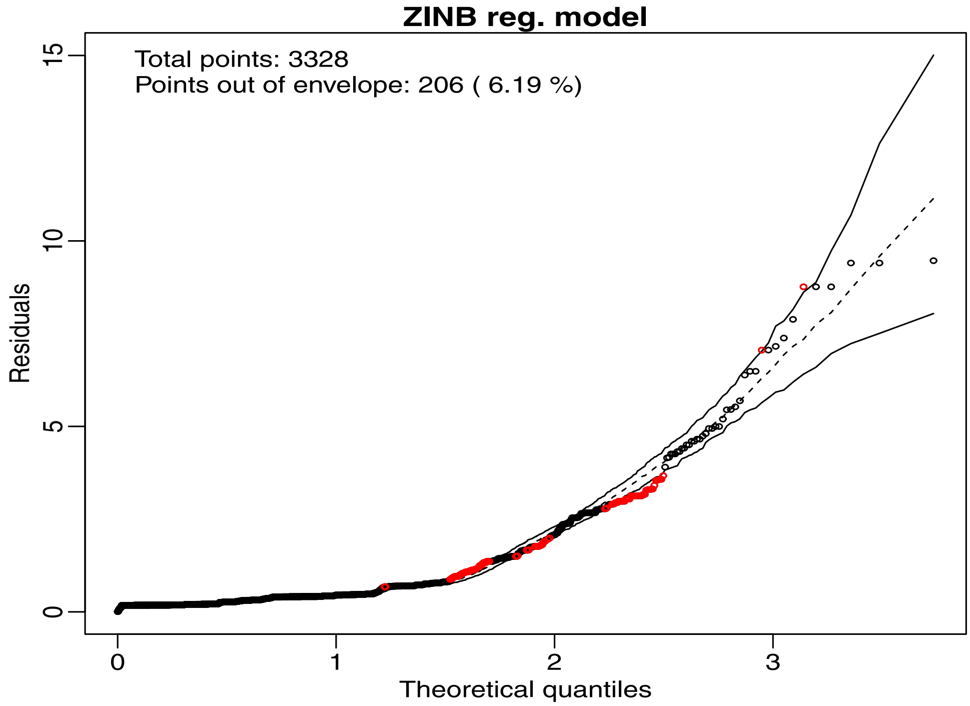

Fig. 4 Simulated envelopes for the Pearson residuals in the ZINB regression model

From Figs. 2, 3, 4 and 5, the proportions of the observations outside the envelopes (red points) are 13.31%, 10.94%, 6.19% and 0.48% for NB, ZIP, ZINB and MAINB regression models, respectively. We can conclude that the MAINB regression model gives a better fit than the other regression models.

5 Conclusion

In count data, there may be excess in some values, which may be located in arbitrary places. To analyze data like this, we proposed a new inflated distribution. Based on the existing literature so far, situations involving inflation at zero, one, two, or arbitrary points have been addressed. However, observations in the real world indicate that inflation can occur at

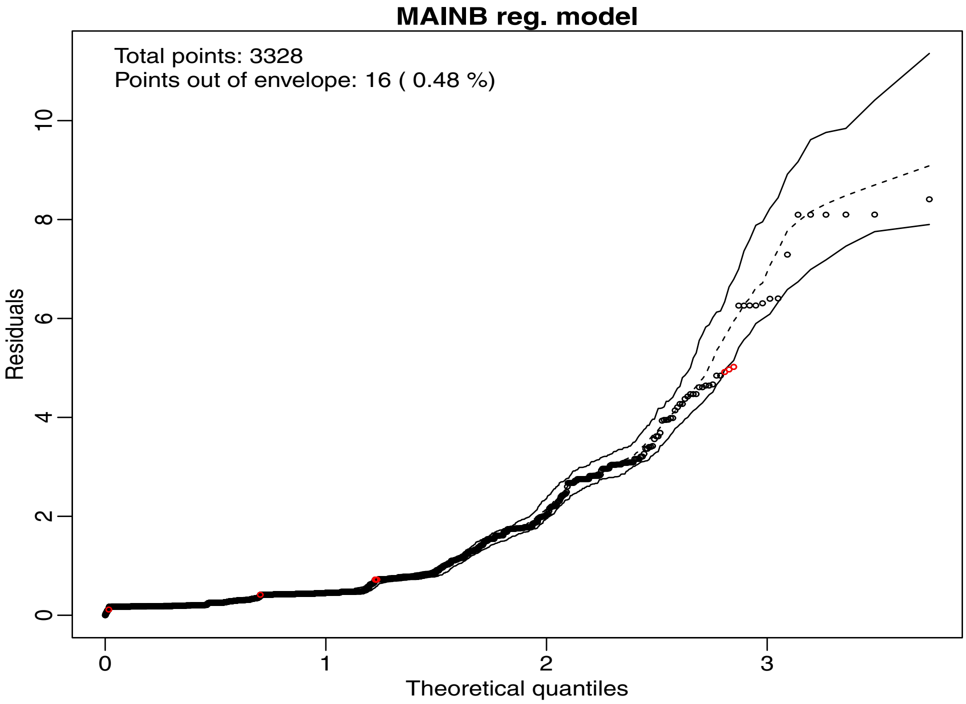

Fig. 5 Simulated envelopes for the Pearson residuals in the MAINB regression model

arbitrary points in an arbitrary number. The distribution and count regression analysis proposed by us introduces a new methodology that can be applied in such cases. For parameter estimation, we consider several methods. The performance of the MAINB distribution is investigated using a simulation study. These results verify that parameter estimations of the MAINB are asymptotically unbiased and consistent. The MAINB regression model is introduced in this paper, and the simulation results show that the proposed model is a promising alternative for modeling inflated count data. A real data set is used to validate the performance of the proposed model. Furthermore, our proposed regression model is a promising alternative for modeling inflated count data with excess values located in arbitrary places.

In some cases, when analyzing count data, inflation may occur at arbitrary points in an arbitrary number, but deflation can also occur. As a subject for future research, we can suggest the development of new methodologies capable of addressing such situations. Bayesian estimates and confidence intervals of the parameters of the MAINB regression model can also be examined in future studies.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Ihab Abusaif and Coşkun Kuş. The first draft of the manuscript was written by Ihab Abusaif and Coşkun Kuş and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The authors declare that no funds, grants, or other support was received during the preparation of this manuscript.

Data availability Enquiries about data availability can be directed to the authors.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Informed consent This is to state that we (Authors) give full permission for the publication, and reproduction in publication (including books, journals, online and internet). I declare, in consequence of granting this permission, that I have no claim on the ground of breach of confidence or any other ground in any legal system against the authors. I hereby agree to release and discharge authors and any editors or other contributors and their agents, publishers, successors and assigns from any claims, demands, or causes of action that I may now have or may hereafter have for libel, defamation, invasion of privacy, copyright or moral rights or violation of any other rights.

References

Abusaif I, Kuş C (2023) Multiple arbitrarily inflated poisson regression analysis. Submitted paper

Alshkaki RSA (2017) Moment estimators of the parameters of zero-one inflated negative binomial distribution. Int J Math Comput Sci 11(1):38–41

Arora M (2018) Extended poisson models for count data with inflated frequencies. Old Dominion University,

Arora M, Chaganty NR (2021) Em estimation for zero-and k-inflated poisson regression model. Computation 9(9):94

Arora M, Chaganty NR (2021) Em estimation for zero-and k-inflated poisson regression model. Computation 9(9):94

Atkinson AC (1981) Two graphical displays for outlying and influential observations in regression. Biometrika 68(1):13–20

Bakouch H, Chesneau C, Karakaya K, Kuş C (2021) The cos-poisson model with a novel count regression analysis. Hacettepe J Math Stat 50(2):559–578

Choi K, Bulgren W (1968) An estimation procedure for mixtures of distributions. J R Stat Soc Ser B Stat Methodol 30(3):444–460

Greene WH (1994) Accounting for excess zeros and sample selection in poisson and negative binomial regression models

Hatamlou A (2013) Black hole: a new heuristic optimization approach for data clustering. Inf Sci 222:175–184

Lambert D (1992) Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34(1):1–14

Lemonte AJ, Moreno-Arenas G, Castellares F (2019) Zero-inflated bell regression models for count data. J Appl Stat

Melkersson M, Olsson C (1999) Is Visiting the dentist a good habit?: Analyzing count data with excess zeros and excess ones. University of Umeå

Mirjalili S (2015) The ant lion optimizer. Adv Eng Softw 83:80–98

Mirjalili S (2015) Moth-flame optimization algorithm: a novel nature-inspired heuristic paradigm. Knowl-Based Syst 89:228–249

Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. Adv Eng Softw 69:46–61

Saboori H, Doostparast M (2023) Zero to k inflated poisson regression models with applications. J Stat Theory Appl 1–27

Saremi S, Mirjalili S, Lewis A (2017) Grasshopper optimisation algorithm: theory and application. Adv Eng Softw 105:30–47

Serra IJA, Polestico DLL (2023) On the zero and k-inflated negative binomial distribution with applications. Adv Appl Stat 88(1):1–23

Su X, Fan J, Levine RA, Tan X, Tripathi A (2013) Multiple-inflation poisson model with l 1 regularization. Statistica Sinica 1071–1090

Sun Y, Zhao S, Tian GL, Tang ML, Li T (2021) Likelihood-based methods for the zero-one-two inflated poisson model with applications to biomedicine. J Stat Comput Simul 1–27

Swain JJ, Venkatraman S, Wilson JR (1988) Least-squares estimation of distribution functions in johnson's translation system. J Stat Comput Simul 29(4):271–297

Tüzen MF, Erbaş S (2018) A comparison of count data models with an application to daily cigarette consumption of young persons. Commun Stat Theory Methods 47(23):5825–5844

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.