CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 10(1) (2025), Pages 457-468
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Zero Inflated Negative Binomial (ZINB) Regression: Application to the Pneumonia Study and Simulation under Several Scenarios

Santi Wahyu Salsabila*, Achmad Efendi, Nurjannah

Departement of Statistics, Faculty of Mathematics and Natural Science, Brawijaya University, Indonesia

Email: santisalsabilaa@gmail.com

ABSTRACT

This study aims at evaluating the performance of Zero Inflated Negative Binomial (ZINB) regression analysis using the Maximum Likelihood Estimation (MLE) approach through simulation study. The research data used are secondary data and simulations. Secondary data was obtained from the Ministry of Health of the Republic of Indonesia in 2023 regarding cases of under-five deaths due to pneumonia with a total of 38 samples. The simulation study is conducted to analyze the performance of ZINB regression based on various sample sizes and proportions of zero values. The results show that the ZINB regression model with the MLE approach produces parameter estimates that tend to be more sensitive to sample size, with improved performance at large sample sizes. Data with a large proportion of zeros reflects high variability as well as the presence of excess zeros, so the ZINB regression model can provide more stable and precise parameter estimates than those with a lower proportion of zeros. Therefore, the ZINB regression model is effective for data with a high proportion of zeros as it fits the characteristics of the data distribution, especially in cases of under-five deaths due to pneumonia.

Keywords: Excess Zeros; MLE; Pneumonia; Simulation; ZINB

INTRODUCTION

Regression method is used to explain the relationship between response variables and predictor variables [1]. One regression technique that can be used to analyze the relationship between response variables in the form of discrete data and predictor variables, which can be continuous, discrete, or a combination of both, is Poisson regression [2]. This model is used to model data with the assumption of equidispersion, which is when the mean and variance of the response variable have the same value [3]. However, in discrete data there is often overdispersion, which is a condition where the variance is greater than the mean. Overdispersion can be caused by several factors, such as missing data, the presence of outliers, or a higher-than-expected number of zero values.

Violation of the equidispersion assumption can cause bias in the standard errors of

Submitted: January 28, 2025 Reviewed: March 21, 2025 Accepted: May 25, 2025
DOI: https://doi.org/10.18860/cauchy.v10i1.32499

parameter estimates, such that the regression estimation results become inaccurate and do not reflect the true conditions [4]. One of the primary causes of overdispersion is the presence of excess zeros in the response variable [5]. The number of zeros indicates the presence of different mechanisms in the data [6]. In some cases, zero values have important meanings that cannot be ignored in the analysis. The greater the proportion of zero values in the response variable, the more likely it is that the parameter estimates in Poisson regression will deviate from the true values. Therefore, a more flexible regression model is needed to handle this condition.

One model that can overcome these problems is ZINB Regression. This model is a mixed distribution between Poisson and Gamma, so it can be applied to model data with a high number of zero values and overcome overdispersion, as it does not require equality between variance and mean [7]. Additionally, the ZINB model includes a dispersion parameter that helps describe the variability of the data. Since ZINB is a non-linear regression model, the parameter estimation method is an important aspect that needs to be studied. One of the commonly used methods in estimating ZINB regression parameters is Maximum Likelihood Estimation (MLE) [8]. Based on research conducted by Azizah and Novita Sari [9], the ZINB regression model is better used in overcoming overdispersion in data that has excess zero values compared to the ZIP regression model.

Simulation is a way of reproducing situation conditions using a model that aims for learning, testing or training, as well as evaluating and improving system performance [10]. The application of simulation studies to the ZINB regression model presents a novelty compared to previous research, which allows for more in-depth analysis of count data with excess zeros. Simulation studies can be conducted using various scenarios that represent variations in data conditions. Thus, this simulation can produce more appropriate predictions and can provide a more robust and data-based basis for decision-making [11].

The statistical methods can be applied to solve various problems, including in the field of health, especially the health of toddlers. Toddlerhood is an important period in human life and has a great influence on the level of public health, as it reflects the health of population as a whole. One of the health indicators is the under-five mortality rate. Pneumonia is the leading cause of under-five deaths in Indonesia, with the under-five mortality rate reaching 0.13% by 2023 [12]. World Health Organization (WHO) in 2021, revealed that pneumonia caused 740,180 under-five deaths or about 14% of total global under-five deaths [12]. It is an acute lung infection that can be caused by bacteria, viruses, fungi, chemical exposure, physical damage to the lungs, or indirect effects of other diseases.

Based on the relevance of the research and the high mortality rate of children under five due to pneumonia in Indonesia, this study aims at evaluating the performance of the Zero Inflated Negative Binomial (ZINB) regression model with the Maximum Likelihood Estimation (MLE) approach in estimating discrete data parameters with an excess proportion of zeros. This research is expected to contribute to the development of more suitable parameter estimation methods for discrete data with excess zero values, particularly in the health sector. The following parts include literature review, methodology, analysis and results, and the conclusion.

METHODS

Data

The data used in this study are secondary data and simulations. Secondary data was obtained from the Ministry of Health of the Republic of Indonesia in 2023 [12]. The data consists of one response variable and 3 predictor variables including the number of under-five deaths due to pneumonia (Y), under-fives aged 0-59 months with malnutrition status (X_1), exclusive breastfeeding in infants (X_2), and complete basic immunization coverage in infants (X_3).

Overdispersion

A common issue in Poisson regression is overdispersion, where the variance of the response variable exceeds the mean [7]. If the Poisson regression model is still used on discrete data that experiences overdispersion, then the estimation of the resulting regression coefficient parameters remains consistent but inefficient because it has an impact on the high standard error value. Overdispersion testing can be done using the Chi-Square approach [13]. Mathematically, it can be written in Equation (1).

$$\chi^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \sim \chi^2_{(n-m)} \tag{1}$$

where $\hat{\mu}_i = \exp(\hat{\beta}_0 + \sum_{j=1}^{k} x_{ij}\hat{\beta}_j)$, n is the number of observations, m is the number of parameters $(p+1)$, x_{ij} is the value of the predictor variable at the j^{th} observation to the i^{th}. If the Chi-Square value divided by the degrees of freedom which is greater than 1, it indicates that the response variable in the data is overdispersed.

Excess Zero

Excess zero is one of the problems that occur in Poisson regression. This condition can also cause overdispersion. In response variables with discrete data, the presence of zero values is common and has an important meaning, so the data must still be included in the analysis. In certain studies, there may be situations where the number of zeros is excessively high. A large proportion of zero values can affect the accuracy of calculations [14]. By including zeros in appropriate statistical models, such as Zero Inflated Negative Binomial, researchers can obtain better and more accurate conclusions. Excess zeros can be identified when the proportion of zero values in the response variable exceeds that of other discrete data. When the proportion of zeros in the response variable is greater than 50%, it certainly leads to overdispersion [15].

Zero Inflated Negative Binomial (ZINB) Regression

ZINB regression is a regression model derived from the Poisson Gamma Mixture distribution. In the ZINB regression model, the response random variable (y_i) is a free random variable with $i = 1,2,\dots,n$ that can form two states, namely zero state and negative binomial state. According to [6], the ZINB distribution function is written in Equation (2).

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\dfrac{1}{1+\kappa\mu_i}\right)^{\frac{1}{\kappa}}; \ for\ y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma\left(y_i + \frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right)y_i!}\left(\dfrac{\kappa\mu_i}{1+\kappa\mu_i}\right)^{y_i}\left(\dfrac{1}{1+\kappa\mu_i}\right)^{\frac{1}{\kappa}}; \ for\ y_i > 0 \end{cases} \tag{2}$$

where $0 \le \pi_i \le 1, \mu_i \ge 0$, κ is the dispersion parameter with $\frac{1}{\kappa} > 0$, $\Gamma(.)$ is the gamma function. When $\pi_i = 0$, the random variable y_i has a negative binomial distribution with mean μ_i nd dispersion parameter κ, so $Y_i \sim NB(\mu_i, \kappa)$. It is assumed that μ_i and π_i depend on the vector variables x_i which can be defined in Equation (3) as follows [16].

$$\begin{gathered}
\mu_i = e^{x_i^T\boldsymbol{\beta}} \\
\frac{\pi_i}{1-\pi_i} = e^{x_i^T\boldsymbol{\gamma}} \\
\pi_i = e^{x_i^T\boldsymbol{\gamma}} - \pi_i e^{x_i^T\boldsymbol{\gamma}} \\
\pi_i\left(1 + e^{x_i^T\boldsymbol{\gamma}}\right) = e^{x_i^T\boldsymbol{\gamma}} \\
\pi_i = \frac{e^{x_i^T\boldsymbol{\gamma}}}{1+e^{x_i^T\boldsymbol{\gamma}}}, \text{so } (1-\pi_i) = \frac{1}{1+e^{x_i^T\boldsymbol{\gamma}}}
\end{gathered} \qquad (3)$$

The ZINB regression model can generally be expressed in Equations (4) and (5).
Model for negative binomial state $\hat{\mu}_i$

$$ln\ \hat{\mu}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}\ , i = 1,2,\dots,n\ and\ j = 1,2,\dots,p \qquad (4)$$

Model for zero inflation $\hat{\pi}_i$

$$logit\ \hat{\pi}_i = \hat{\gamma}_0 + \sum_{j=1}^{p} \hat{\gamma}_j x_{ij}\ , i = 1,2,\dots,n\ and\ j = 1,2,\dots,p \qquad (5)$$

where p : number of predictor variables; n : number of observations; $\hat{\beta}$: ZINB regression model parameters; $\hat{\gamma}$: ZINB regression model parameters.

ZINB Regression with Maximum Likelihood Estimation (MLE) Approach

Parameter estimation for the ZINB regression model is typically estimated using the MLE method, with the EM (Expectation Maximization) algorithm employed to maximize the function. This method is applied to estimate a model when the density function is already known. Based on the function in Equation (2), Equation (3) is substituted so that Equation (6) is obtained as follows.

$$P(Y_i = y_i) = \begin{cases} \dfrac{e^{x_i^T\boldsymbol{\gamma}}}{1+e^{x_i^T\boldsymbol{\gamma}}} + \dfrac{1}{1+e^{x_i^T\boldsymbol{\gamma}}}\left(\dfrac{1}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\kappa}}; \ for\ y_i = 0 \\ \dfrac{1}{1+e^{x_i^T\boldsymbol{\gamma}}} \dfrac{\Gamma\left(y_i + \frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right) y_i!}\left(\dfrac{\kappa e^{x_i^T\boldsymbol{\beta}}}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(\dfrac{1}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\kappa}}; \ for\ y_i > 0 \end{cases} \qquad (6)$$

The likelihood function form of the ZINB regression model can be written in Equation (7). Furthermore, from Equation (7), the ln likelihood equation will be made and Equation (8) is obtained.

$$L(\boldsymbol{\theta}|y_i)=\begin{cases}\displaystyle\prod_{i=1}^{n}\frac{e^{x_i^T\boldsymbol{\gamma}}}{1+e^{x_i^T\boldsymbol{\gamma}}}+\frac{1}{1+e^{x_i^T\boldsymbol{\gamma}}}\left(\frac{1}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\kappa}};\ for\ y_i=0\\ \displaystyle\prod_{i=1}^{n}\frac{1}{1+e^{x_i^T\boldsymbol{\gamma}}}\frac{\Gamma\left(y_i+\frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right)y_i!}\left(\frac{\kappa e^{x_i^T\boldsymbol{\beta}}}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\kappa}};\ for\ y_i>0\end{cases}\tag{7}$$

$$\begin{aligned}&\ln L(\boldsymbol{\theta}|y_i)\\&=\begin{cases}\displaystyle\sum_{i=1}^{n}ln\left(\frac{e^{x_i^T\boldsymbol{\gamma}}}{1+e^{x_i^T\boldsymbol{\gamma}}}+\frac{1}{1+e^{x_i^T\boldsymbol{\gamma}}}\left(\frac{1}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\kappa}};\ for\ y_i=0\right)\\ \displaystyle\sum_{i=1}^{n}ln\left(\frac{1}{1+e^{x_i^T\boldsymbol{\gamma}}}\frac{\Gamma\left(y_i+\frac{1}{\kappa}\right)}{\Gamma\left(\frac{1}{\kappa}\right)\Gamma\left(y_i+1\right)}\left(\frac{\kappa e^{x_i^T\boldsymbol{\beta}}}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\kappa}};\ for\ y_i>0\right)\end{cases}\\&=\begin{cases}\displaystyle\sum_{i=1}^{n}ln\left(e^{x_i^T\boldsymbol{\gamma}}+\left(\frac{1}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\kappa}}\right)-\sum_{i=1}^{n}ln\left(1+e^{x_i^T\boldsymbol{\gamma}}\right)\ ;\ for\ y_i=0\\ \displaystyle-\sum_{i=1}^{n}ln\left(1+e^{x_i^T\boldsymbol{\gamma}}\right)+\sum_{i=1}^{n}ln\left(\Gamma\left(y_i+\frac{1}{\kappa}\right)\right)-\sum_{i=1}^{n}ln\left(\Gamma\left(\frac{1}{\kappa}\right)\right)\\ \displaystyle-\sum_{i=1}^{n}\ln\left(\Gamma(y_i+1)\right)+y_i\sum_{i=1}^{n}\left(\frac{\kappa e^{x_i^T\boldsymbol{\beta}}}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right)+\frac{1}{\kappa}\sum_{i=1}^{n}\left(\frac{1}{1+\kappa e^{x_i^T\boldsymbol{\beta}}}\right);\ for\ y_i>0\end{cases}\end{aligned}\tag{8}$$

Estimation with MLE is calculated by maximizing the ln likelihood function in Equation (8). The summation of the log-likelihood function is not linear so that the likelihood function cannot be solved by ordinary numerical methods [17]. The EM algorithm is an alternative iterative approach used to maximize the likelihood function for data that includes latent variables, which arise from the definition of new variables such as the w_i variables. Suppose the response variable (Y) is related to the indicator variable (w) as follows.

$$w_i=\begin{cases}1,\text{if } y_i \text{ from } zero\ state\\ 0,\text{if } y_i \text{ from } NB\ state\end{cases}\tag{9}$$

The EM algorithm consists of two stages, namely the expectation stage and the maximization stage. The expectation stage is the calculation stage of the expectation of the ln likelihood function, then the maximization stage is the calculation stage to find the parameter estimate that maximizes the ln likelihood function resulting from the previous expectation stage with the Newton Raphson method [18].

The first step is the E-Step (expectation stage). In this stage, (m) is used to symbolize iteration [19].

$$w_i^{(m)}=\begin{cases}\left[1+e^{-x_i^T\boldsymbol{\gamma}^{(m)}}\left(\frac{1}{1+\boldsymbol{\kappa}^{(m)}+e^{-x_i^T\boldsymbol{\beta}^{(m)}}}\right)^{\frac{1}{\kappa}^{(m)}}\right]^{-1},jika\ y_i=0\\ 0,jika\ y_i>0\end{cases}\tag{10}$$

so

$$Q\left(\boldsymbol{\beta},\boldsymbol{\gamma};\boldsymbol{\beta}^{(m)},\boldsymbol{\gamma}^{(m)}\right)=\sum\nolimits_{i=1}^{n} ln\,L\,(\boldsymbol{\gamma}^{(m)}|y_i,w_i^{(m)})+\sum\nolimits_{i=1}^{n} ln\,L\,(\boldsymbol{\beta}^{(m)}|y_i,w_i^{(m)}) \tag{11}$$

where

$$ln\,L\,(\boldsymbol{\gamma}^{(m)}|y_i,w_i^{(m)})=\sum\nolimits_{i=1}^{n}\left[w_i^{(m)}\boldsymbol{x}_i^T\boldsymbol{\gamma}-ln(1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}})\right] \tag{12}$$

and

$$ln\,L\,(\boldsymbol{\beta}^{(m)}|y_i,w_i^{(m)})$$
$$=\sum\nolimits_{i=1}^{n}\left(1-w_i^{(m)}\right)\left[\frac{\Gamma\left(y_i+\frac{1}{\boldsymbol{\kappa}}\right)}{\Gamma\left(\frac{1}{\boldsymbol{\kappa}}\right)\Gamma\,(y_{\mathrm{i}}+1)}\left(\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1+\boldsymbol{\kappa}e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\boldsymbol{\kappa}}}\right] \tag{13}$$

The next stage is M-Step (maximization stage). The M-Step stage is carried out using the Newton Raphson iteration method to maximize $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from the results of the E-Step stage by calculating $\boldsymbol{\beta}^{(m+1)}$ and $\boldsymbol{\gamma}^{(m+1)}$.

Simulation Framework

Simulation is a way of reproducing situation conditions using a model that aims for learning, testing or training, as well as evaluating and improving system performance [10]. Simulation is a technique used to carry out experiments using a model of a real system [20]. Simulation can be used as an approach to solve various problems that are uncertain and possibilities that cannot be carefully calculated. According to [11], simulation is not only used to design decisions, but also to validate that the decision taken is the best decision.

The simulation data used in this study refers to processed data obtained from secondary data. Data generation is done using R version 4.4.2 with various proportions of zero values, namely $p=0.3, 0.5, 0.8$ and various sample sizes, namely $n=38, 100, 500$. Simulation studies were performed to evaluate the performance of the ZINB regression with the MLE approach in handling data containing overdispersion and excess zero at various proportions of zero values and various sample sizes.

Best Model Selection

According to [21], determining the best model can be done by looking at the AIC (Akaike Information Criterion) value. The AIC value is computed based on the maximum likelihood value and the number of parameters in the constructed regression model. Equation (14) shows the formula for calculating the AIC value.

$$AIC=-2\,ln\,L(\hat{\theta})+2p \tag{14}$$

where $L(\hat{\theta})$ is the likelihood value and p is the number of parameters. The optimal regression model is the one with the lowest AIC value.

RESULTS AND DISCUSSION

Overdispersion

Testing overdispersion in Poisson regression can be done using the Chi-Square value divided by the degree of freedom [22]. Overdispersion conditions occur when the Chi-Square value divided by the free degree has a value of more than 1. Based on the test results, the dispersion value is $4.03 > 1$. Therefore, it can be concluded that the response variable exhibits overdispersion.

Excess Zero

The examination of zero inflation is conducted by calculating the proportion of zero values in the response variable. The results of the excess zero check for the response variable are shown in Table 1 below.

Table 1. Excess Zero Check Result

Number of under-five deaths due to pneumonia	Frequency	Percentage
0	30	78.95%
1	4	10.52%
2	2	5.26%
9	1	2.63%
11	1	2.63%

Based on the table above, it is clear that the data has excess zero in the response variable because the proportion of zero values is more than 50%, which is 78.95%. Thus, it can be concluded that the Poisson regression model is not appropriate for use in modeling.

Zero Inflated Negative Binomial (ZINB) Regression

The ZINB regression model is a model that can be used to overcome overdispersion and excess zero problems [23]. The ZINB regression model was applied to cases of under-five deaths due to pneumonia in each province in Indonesia. In this modeling, three predictor variables and one response variable were used.

Simultaneous and partial tests are performed to assess the significance level of the parameter estimation results in the ZINB regression model. Simultaneous testing can be done using the G test statistic, while partial testing is done using the Wald test statistic. The results of parameter estimation for the ZINB regression model, along with the simultaneous and partial test results, are shown in Table 2.

Table 2. Parameter Estimation Results of ZINB Regression

Parameter	Estimation	Z value	Pr(>|z|)
$\hat{\beta}_0$	-18.54	-2.89	0.00*
$\hat{\beta}_1$	2.80	1.82	0.07
$\hat{\beta}_2$	-0.19	-2.54	0.01*
$\hat{\beta}_3$	0.32	3.22	0.00*
$\hat{\gamma}_0$	-28.25	-1.41	0.16
$\hat{\gamma}_1$	8.23	1.31	0.19
$\hat{\gamma}_2$	-0.50	-1.51	0.13
$\hat{\gamma}_3$	0.58	1.51	0.13
Pr(>Chisq) = 0.0055*			

*) Significant with 5% significance level

The ZINB regression model is expressed as follows.
Model for negative binomial state $\hat{\mu}_i$

$$\begin{aligned} ln\,\hat{\mu}_i &= -18.54 + 2.80\,X_1 - 0.19\,X_2 + 0.32\,X_3 \\ \hat{\mu}_i &= exp(-18.54 + 2.80\,X_1 - 0.19\,X_2 + 0.32\,X_3) \end{aligned} \quad (15)$$

Model for zero inflation $\hat{\pi}_i$

$$logit\ \hat{\pi}_i = -28.25 + 8.23\, X_1 - 0.50\, X_2 + 0.58\, X_3$$
$$\hat{\pi}_i = \frac{exp(-28.25 + 8.23\, X_1 - 0.50\, X_2 + 0.58\, X_3)}{1 + exp(-28.25 + 8.23\, X_1 - 0.50\, X_2 + 0.58\, X_3)} \qquad (16)$$

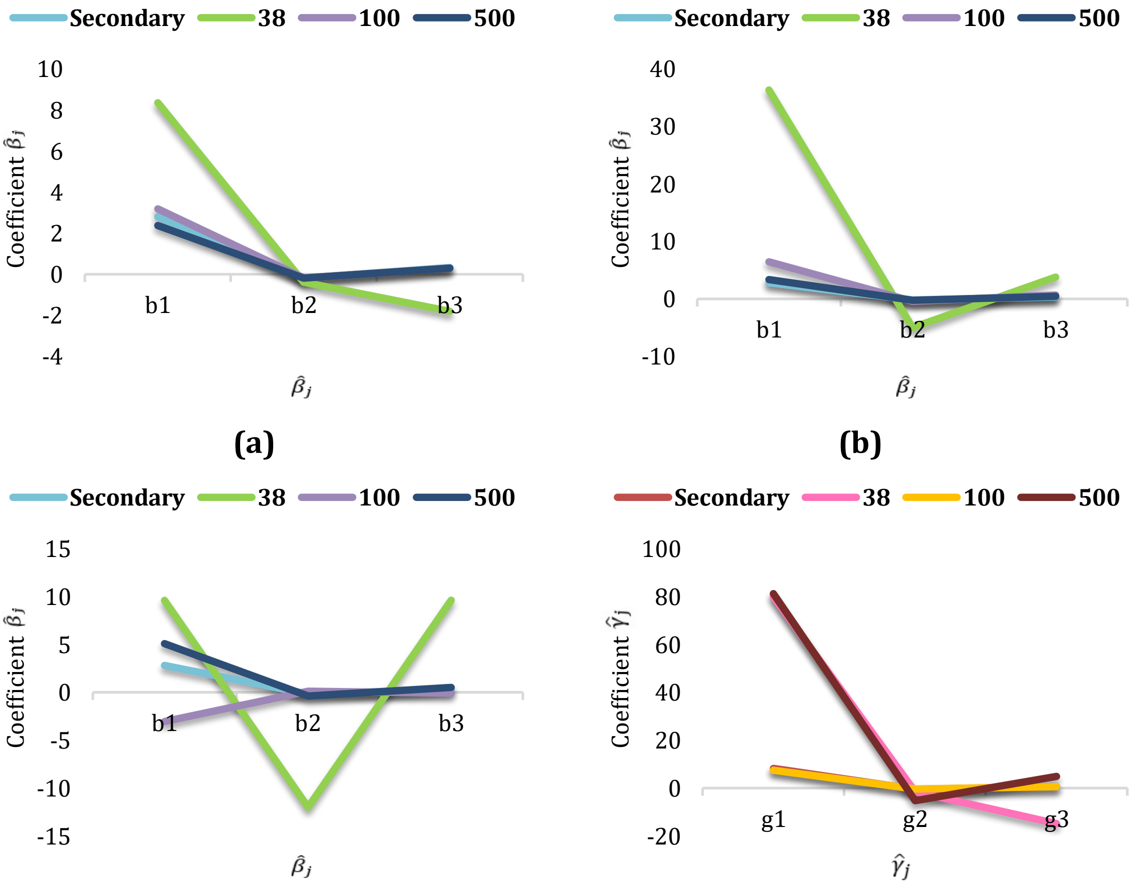
Based on the results of the simultaneous significance test for parameter estimates in the ZINB regression model, it is evident that the predictor variables X_1, X_2, and X_3 collectively have a significant effect on the response variable. Meanwhile, the partial significance test results indicate that the response variable is significantly influenced by X_2 and X_3.

Based on the partial significance results, the ZINB regression model is formulated as shown in Equation (17).

$$\hat{\mu}_i = exp(-18.54 - 0.19\, X_2 + 0.32\, X_3) \qquad (17)$$

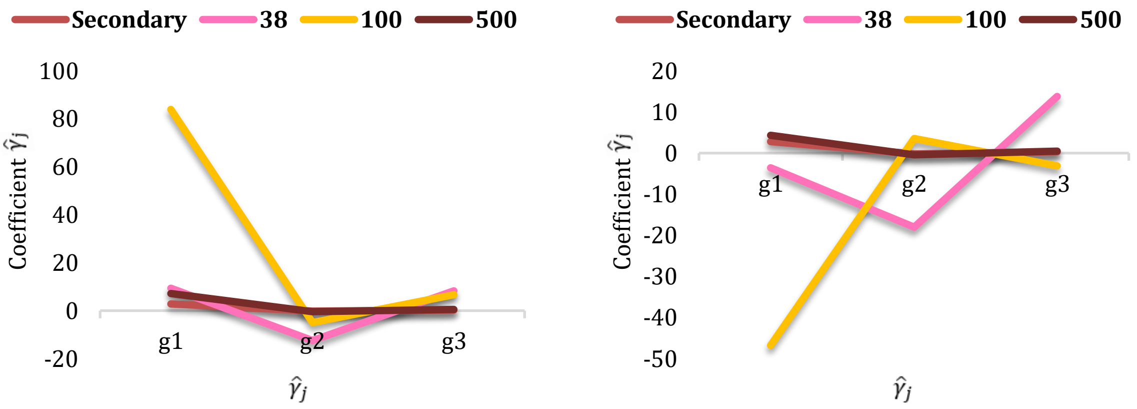
Simulation Results of ZINB Regression

Simulation studies are used to validate that the decision taken is the best decision [11]. Simulation data generation is performed based on the initial ZINB regression parameters. The simulation results for estimating parameters $\hat{\beta}_j$ and $\hat{\gamma}_j$ are presented in Figure 1.



(a) **(b)**

(c) **(d)**



(e) **(f)**

Figure 1. Plots of Parameter Estimation $\hat{\beta}_j$ at Various Sample Sizes and Proportion of Zero Values (a) p=0.3 (b) p=0.5 (c) p=0.8 and Parameter $\hat{\gamma}_j$ at Various Sample Sizes and Proportion of Zero Values (d) p=0.3 (e) p=0.5 (f) p=0.8

Based on Figure 1, the estimation results of parameters $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ are relatively more stable at sample sizes of 100 and 500. At small sample size ($n = 38$), the parameter $\hat{\beta}_1$ shows large fluctuations, namely $\hat{\beta}_1 = 8.36$ for $p = 0.3$ and $\hat{\beta}_1 = 36.34$ for $p = 0.5$. Whereas, at medium ($n = 100$) and large ($n = 500$) sample sizes, the parameter $\hat{\beta}_1$ tends to stabilize close to the initial value of the parameter. This can occur due to data imbalance in certain scenarios. Parameter estimation results at large sample sizes ($n = 500$) tend to be more stable across various scenarios p. This is in line with statistical theory stating that larger sample sizes produce more precise parameter estimates. Additionally, simulations that are conducted with various proportions of zero ($p = 0.3; 0.5; 0.8$) show that parameter estimation results tend to be better at a proportion of 0.8. This occurs as at a larger proportion, the data satisfies the overdispersion and excess zero conditions, which are the main problems addressed by the ZINB regression model.

The parameter estimation results of $\hat{\gamma}_1, \hat{\gamma}_2$, and $\hat{\gamma}_3$ show that the sample size (n) and the proportion of zero values (p) significantly affect the probability of zero inflation and the stability of the parameter estimates. At small ($n = 38$) and medium ($n = 100$) sample sizes, the parameter estimates show larger fluctuations with zero inflation probabilities that vary depending on the proportion. A higher proportion of zero values ($p = 0.8$) may cause the parameter estimates to be more influenced by the number of zeros. At smaller sample sizes, the presence of zero inflation becomes more dominant and can lead to high fluctuations in parameter estimates. However, this influence decreases at larger sample sizes. Overall, it can be concluded that the simulation results using the MLE approach require a sufficiently large sample size and an appropriate proportion of zero values to produce stable and precise parameter estimates in the ZINB regression model.

Best Model Selection

Determining the best model between one method and another can be done by the AIC (Akaike Information Criterion) value [21]. The optimal model is the one with the lowest AIC value. The following presents the AIC value for each method.

Table 3. AIC Results of Each Model

n	Proportion of Zero Values	AIC
38	0.3	38.98
	0.5	61.71
	0.8	32.90
100	0.3	96.76
	0.5	74.87
	0.8	56.81
500	0.3	571.43
	0.5	441.00
	0.8	251.22

Based on Table 3, it is evident that the smallest AIC value occurs at a small sample size of $n = 38$. As the sample size increases, the AIC value also increases. This indicates that the model is not flexible enough to capture the complexity of data in large samples. When viewed at the proportion of zero values, the smallest AIC value occurs at a high proportion of zero values, namely $p = 0.8$. The ZINB regression model is suitable for data containing a high proportion of zero values and fulfilling overdispersion and excess zero conditions because the zero inflation component has a more significant role in capturing data patterns.

CONCLUSIONS

It can be concluded that the Zero Inflated Negative Binomial (ZINB) regression analysis with the Maximum Likelihood Estimation (MLE) approach shows that the parameter estimation results tend to be sensitive to the sample size, with performance increasing at large sample sizes. However, the MLE method is able to produce accurate parameter estimates on data with a large proportion of zero values. Data with a large proportion of zeros reflects high variability as well as the presence of excess zeros, so the ZINB regression model can provide more stable and precise parameter estimates than those with a lower proportion of zeros. Therefore, the ZINB regression model is effectively used on data with a high proportion of zeros because it is more in line with the characteristics of the data distribution.

This study has limitations related to the number of sample sizes, the proportion of zero values, and the best model selection method. Thus, future research is expected to explore simulations using a variety of sample sizes and proportions of zero values as well as other best model selection methods. Additionally, adding predictor variables that affect under-five mortality due to pneumonia in Indonesia will provide more in-depth results.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

This study was entirely conducted by the authors without the use of generative AI or AI-assisted technologies. All activities related to the study, including data collection, analysis, and manuscript preparation, were carried out independently by the authors.

DECLARATION OF COMPETING INTEREST

The authors declare that there are no competing interests related to this study. The entire research process was conducted independently, without any external influence that could affect the results or conclusions.

FUNDING

This study was conducted without any specific grant from funding agencies. All activities related to the study, including data collection, analysis, and manuscript preparation, were carried out independently by the authors.

DATA AVAILABILITY

The data of this study can be accessed through the following link: https://kemkes.go.id/app_asset/file_content_download/172231123666a86244b83fd8.51637104.pdf.

REFERENCES

[1] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2012.

[2] L. Anisa and N. A. K. Rifai, "Analisis Regresi Logistik Biner dengan Metode Penalized Maximum Likelihood pada Penyakit Covid-19 di RSUD Pringsewu," *J. Ris. Stat.*, pp. 129–136, 2022, doi: 10.29313/jrs.v2i2.1425.

[3] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*, no. 53. Cambridge university press, 2013.

[4] A. R. Nasution, K. Sadik, and A. Rizki, "Perbandingan Kinerja Regresi Conway-Maxwell-Poisson dan Poisson-Tweedie dalam Mengatasi Overdispersi Melalui Data Simulasi," *Xplore J. Stat.*, vol. 11, no. 3, pp. 215–225, 2022, doi: 10.29244/xplore.v11i3.1018.

[5] R. Cahyandari, "Pengujian Overdispersi pada Model Regresi Poisson (Studi Kasus: Laka Lantas Mobil Penumpang di Provinsi Jawa Barat)," *Statistika*, vol. 14, no. 2, pp. 69–76, 2014, [Online]. Available: https://ejournal.unisba.ac.id/index.php/statistika/article/view/1204

[6] A. M. Garay, V. H. Lachos, and H. Bolfarine, "Bayesian estimation and case influence diagnostics for the zero-inflated negative binomial regression model," *J. Appl. Stat.*, vol. 42, no. 6, pp. 1148–1165, 2015, doi: 10.1080/02664763.2014.995610.

[7] J. M. Hilbe, "Negative Binomial Regression Second Edition-Negative Binomial Regression: Second Edition Joseph M. Hilbe Frontmatter More information," pp. 1–18, 2011, [Online]. Available: www.cambridge.org

[8] D. A. Pradana, "Estimasi parameter regresi zero-inflated negative binomial dengan metode algoritma Expectation Maximization (EM)(studi kasus: penyakit difteri di Jawa Barat tahun 2016)," 2019, *Universitas Negeri Malang*.

[9] Azizah and D. Novita Sari, *The Implementation of ZIP Regression and ZINB Regression to Overcome Overdispersion of Death Cases Due to Filariasis in Indonesia*, vol. 20, no. 1. 2023. doi: 10.31851/sainmatika.v20i1.11707.

[10] B. K. Ghosh, R. Bowden, B. Gladwin, and C. Harrell, *Simulation Using ProModel*. McGraw-Hill, 2004.

[11] A. M. Law, W. D. Kelton, and W. D. Kelton, *Simulation modeling and analysis*, vol. 3. Mcgraw-hill New York, 2007.

[12] Kementrian Kesehatan, *Profil Kesehatan Indonesia 2023*. Jakarta: Kementerian Kesehatan RI, 2024.

[13] J. R. Wilson and K. A. Lorenz, *Modeling binary correlated responses using SAS, SPSS and R*, vol. 9. Springer, 2015.

[14] J. Stachurski, *A primer in econometric theory*. Mit Press, 2016.

[15] F. Famoye and J. S. Preisser, "Marginalized zero-inflated generalized Poisson regression," *J. Appl. Stat.*, vol. 45, no. 7, pp. 1247–1259, 2018, doi: 10.1080/02664763.2017.1364717.

[16] H. Zamani and N. Ismail, "Functional form for the zero-inflated generalized poisson regression model," *Commun. Stat. - Theory Methods*, vol. 43, no. 3, pp. 515–529, 2014, doi: 10.1080/03610926.2012.665553.

[17] B. Ariawan, S. Suparti, and S. Sudarno, "Pemodelan Regresi Zero-Inflated Negative Binomial (Zinb) Untuk Data Respon Diskrit Dengan Excess Zeros," *J. Gaussian*, vol. 1, no. 1, pp. 55–64, 2012.

[18] D. B. Hall, "Zero-inflated Poisson and binomial regression with random effects: a case study," *Biometrics*, vol. 56, no. 4, pp. 1030–1039, 2000.

[19] A. M. Garay, E. M. Hashimoto, E. M. M. Ortega, and V. H. Lachos, "On estimation and influence diagnostics for zero-inflated negative binomial regression models," *Comput. Stat. Data Anal.*, vol. 55, no. 3, pp. 1304–1318, 2011.

[20] D. Siagian, "Sugiarto,'Metode Statistika untuk Bisnis dan Ekonomi', Jakarta: PT," *Gramedia Pustaka Utama*, 2002.

[21] M. I. Bhatti and H. Al-Shanfari, *Econometric analysis of model selection and model testing*. Routledge, 2017.

[22] E. H. Payne, M. Gebregziabher, J. W. Hardin, V. Ramakrishnan, and L. E. Egede, "An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data," *Commun. Stat. Comput.*, vol. 47, no. 6, pp. 1722–1738, 2018.

[23] Y. Gençtürk and A. Yiğiter, "Modelling claim number using a new mixture model: negative binomial gamma distribution," *J. Stat. Comput. Simul.*, vol. 86, no. 10, pp. 1829–1839, 2016, doi: 10.1080/00949655.2015.1085987.