CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 10 (2) (2025), Pages 605–615
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Bayesian Approach in Estimating Parameters of Zero-Inflated Negative Binomial Regression Model Using Cauchy Prior: Simulation Study on Pneumonia

Santi Wahyu Salsabila, Achmad Efendi*, and Nurjannah

Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Indonesia

Abstract

The Bayesian approach is one of the parameter estimation methods that can be applied to Zero-Inflated Negative Binomial (ZINB) regression analysis. The ZINB regression model is used to analyze over-dispersion data with excess zeros. This study aims to evaluate the performance of ZINB regression parameter estimation using a Bayesian approach with Cauchy prior in pneumonia studies. The analysis is applied to secondary data as well as to simulated data with various scenarios based on different sample sizes and proportions of zero values such that the optimal model can be determined. The results show that ZINB regression models using the Bayesian approach provide stable parameter estimates as sample sizes and proportions of zeros increase. In cases of under-five deaths due to pneumonia, the data often contains many zeros because not all regions report cases. The ZINB model effectively addresses over-dispersion and excess zeros through a combination of negative binomial and zero-inflation models. This provides more accurate modeling results to support policymaking. The Bayesian approach also provides flexibility in integrating prior information and handling small samples, making the ZINB model well suited for health data with rare events and many zeros.

Keywords: Bayesian; Over-dispersion; Pneumonia; Simulation; ZINB

Copyright © 2025 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (https://creativecommons.org/licenses/by-sa/4.0)

1. Introduction

Poisson regression is a statistical method applied to examine the relationship between discrete response variables and continuous, discrete, or mixed predictor variables. Over-dispersion, where the variance is greater than the mean, is often found in discrete data [1]. Some of the causes of over-dispersion are missing observations, outliers in the data, and having excess zero values. One alternative regression model in overcoming these problems is the Zero-Inflated Negative Binomial (ZINB) regression model. The ZINB model is based on a mixed Poisson and Gamma distribution [2]. Additionally, the ZINB model features a dispersion parameter that serves to explain the extent of variance in the data. Since the ZINB model is non-linear, parameter estimation is an important topic to study.

In ZINB regression, parameter estimation is typically carried out using the Maximum Likelihood Estimation (MLE) approach. Research conducted by [3] states that the MLE approach is used in constructing ZINB regression models. Another study by [4] also uses the

*Corresponding author. E-mail: a_efendi@ub.ac.id

Received: May 2, 2025 Revised: June 5, 2025 Accepted: June 29, 2025
DOI: https://doi.org/10.18860/cauchy.v10i2.33245

MLE approach in estimating parameters in ZINB regression models. Despite its frequent use, the MLE method has some limitations. one particular limitation is when dealing with small samples. A better alternative for estimating ZINB regression model parameters is the Bayesian approach.

The Bayesian approach is better to use because MLE is based solely on sample data, and sample size affects estimation results [5]. The Bayesian method considers sample data taken from population data. It also takes into account an initial distribution, called the prior distribution. Then, the prior distribution is integrated with the sample data information to produce parameter estimates known as the posterior distribution. [6] stated that the Bayesian regression model is superior to other regression models because it produces smaller standard errors and narrower confidence intervals. Bayesian regression models can also produce higher coverage probabilities and lower bias. [7] conducted a study explaining that ZINB regression parameter estimation was carried out using a Bayesian approach using normal priors.

In the context of estimating under-five pneumonia mortality cases, the Bayesian approach allows for the incorporation of supplementary information through the assignment of priors, yielding more precise analysis results. The posterior distribution is influenced by the selection of the prior distribution. Thus, assigning priors shapes a more representative posterior distribution. [8] conducted research to obtain a prior distribution for regression parameters that provides more stable estimations. The study used the Cauchy distribution as a prior on all regression parameters. The Cauchy distribution has a thick tail, making it more robust to outliers in the data. It can provide better parameter estimates and flexibility in Bayesian methods when used as a prior distribution.

In Bayesian methods, the combination of prior distributions and sample data is used to estimate parameters in the form of posterior distributions. In practice, however, posterior distributions are difficult to obtain as they are complex and cannot be easily solved analytically. In order to overcome this problem, the Markov Chain Monte Carlo (MCMC) algorithm was developed based on the Gibbs Sampling algorithm. MCMC generates samples from a given distribution using Markov chain properties. This method can solve complicated posterior distributions [9].

Previous research has primarily focused on developing regression models using MLE or Bayesian approaches using normal priors, but has rarely used simulation studies to determine the resulting performance. Simulation is a method of recreating situations using a model for the purposes of learning, testing, training, evaluation, and improvement of system performance [10]. Simulation is beneficial as a decision-making tool for designing systems with specific performance characteristics, at both the design and operational stages. Simulation is used not only for designing decisions but also for validating that the made decision are optimal [11]. Additionally, simulation can reduce costs and time while providing accurate predictions of complex system performance.

Based on the relevance of the research and the high mortality rate of children under five due to pneumonia in East Java, this study aims to evaluate the performance of the ZINB regression model with a Bayesian approach in estimating parameters from discrete data with an excess proportion of zeros. The results of this study are expected to contribute to the development of more appropriate methods for estimating parameters of discrete data with excess zero values, particularly in the health sector. The study is organized by first providing preliminary information on the development of ZINB regression models using Bayesian approach for under-five mortality cases due to pneumonia. A brief road-map of the paper's structure, as the following: Section 2 details the methodology, Section 3 presents the results and discussion, and Section 4 provides conclusion.

2. Methods

This section provides an overview of the data, modeling methods, and simulation scenarios used, which are organized systematically to facilitate understanding of the analytical approach applied.

2.1. Data

This research uses secondary data and simulations. The secondary data were obtained from the East Java Provincial Health Office in 2023 [12]. The data set consists of one response variable and three predictor variables: 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).

The low rate of exclusive breastfeeding, both globally and nationally, increases the risk of infants contracting infectious diseases such as pneumonia. Research shows that infants who are not exclusively breastfed are seven times more likely to contract pneumonia than those who are [13]. Additionally, immunization has been proven to influence the incidence of pneumonia in infants [14]. Nutritional status is also a key factor, with malnourished infants having a 162 times higher risk of developing pneumonia compared to well-nourished infants [15]. With the appropriate statistical approach, it is hoped that the infant mortality rate due to pneumonia can be reduced, thereby increasing the chances of children surviving into the future.

2.2. Over-dispersion

A common issue in Poisson regression is over-dispersion, a condition in which the variance of the response variable exceeds its mean value [2]. When Poisson regression is applied to discrete data showing over-dispersion, the estimated regression coefficient parameters are consistent but inefficient because they produce relatively large standard error values. This can be expressed mathematically through Equation (1).

$$\chi^2 = \sum_{i=0}^{n} \frac{(y_i - \mu_i)^2}{\mu_i} \quad \text{with} \quad \phi = \frac{\chi^2}{n - p - 1} \tag{1}$$

where μ_i is the estimated mean of the i-th observation, ϕ is the dispersion value, n is the total number of observations, and p is the number of predictor variables. If the dispersion value is greater than one, it indicates that the response variable in the data is over-dispersed.

2.3. Excess Zero

One of the problems in the application of Poisson regression is the presence of zeros, also known as excess zero. This condition can lead to over-dispersion, a situation in which the variance of the data exceeds its mean. Using appropriate statistical models, such as Zero-Inflated Negative Binomial (ZINB), allows researchers to obtain more precise results and conclusions. Excess zeros can be identified when the response variable contains a proportion of zero that is significantly higher than the proportion of other discrete values. If more than 0.5 or 50% of the response variables have zero data, this often indicates over-dispersion [16].

2.4. Zero-Inflated Negative Binomial (ZINB) Regression

ZINB regression is a regression model derived from a Poisson-Gamma mixture distribution [2]. In the ZINB regression model, the response random variable y_i is a free random variable with $i = 1, 2, \ldots, n$ that can take two states, namely the zero inflation state and the negative binomial state [17]. The ZINB distribution function is written as in Equation (2) [18].

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

where $0 \leq \pi_i \leq 1$, $\mu_i \geq 0$, κ is the dispersion parameter with $1/\kappa > 0$, and $\Gamma(\cdot)$ is the gamma function. When $\pi_i = 0$, the random variable y_i follows $Y_i \sim \text{NB}(\mu_i, \kappa)$. Both μ_i and π_i are

considered to depend on the vector of covariates $\mathbf{x}_i$, which can be defined as in Equation (3) [19].

$$\mu_i = e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}, \quad \pi_i = \frac{e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}}}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}}} \quad \text{or} \quad 1-\pi_i = \frac{1}{1+e^{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}}} \tag{3}$$

where the superscript T in $\mathbf{x}_i^{\mathrm{T}}$ denotes the transpose, which means converting rows into columns or vice versa, β and γ are the parameters of the ZINB regression model.

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

$$\ln \hat{\mu}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}, \quad i = 1, 2, \ldots, n \quad \text{and} \quad j = 1, 2, \ldots, p \tag{4}$$

Model for logit (zero inflation state): $\hat{\pi}_i$

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

where p is the number of predictor variables, n is the number of observations, $\hat{\beta}$ and $\hat{\gamma}$ are the parameters of the ZINB regression model for the negative binomial state and zero inflation state.

2.5. Zero-Inflated Negative Binomial (ZINB) Bayesian Regression

In the Bayesian method, a parameter treated as a random variable with a distribution is called a *prior distribution*. The prior distribution provides the initial information necessary for forming a posterior distribution. In ZINB regression, the parameters of the two models are assumed to have Cauchy distributions: $\boldsymbol{\beta} \sim \text{Cauchy}(l_\beta, s_\beta)$ for the negative binomial state model μ_i, and $\boldsymbol{\gamma} \sim \text{Cauchy}(l_\gamma, s_\gamma)$ for the zero inflation model π_i. Additionally, a dispersion parameter needs to be estimated and is denoted by κ. Since $\kappa > 0$, the dispersion parameter is modeled using a gamma distribution, namely $\kappa \sim \text{Gamma}(a, b)$. The parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and κ are assumed to be independent of each other; therefore, the product of their probability density functions is expressed in Equation (6).

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \kappa) = \prod_{j=0}^{m} \frac{s_{\beta_j}}{\pi\left(s_{\beta_j}^2 + (\beta_j - l_{\beta_j})^2\right)} \times \prod_{k=0}^{m} \frac{s_{\gamma_k}}{\pi\left(s_{\gamma_k}^2 + (\gamma_k - l_{\gamma_k})^2\right)} \times \frac{1}{b^a \Gamma(a)} \kappa^{a-1} e^{-\kappa/b} \tag{6}$$

The likelihood function is widely used in estimation processes, including the Bayesian method [20]. In the Bayesian method, the ZINB regression likelihood function is used to form the posterior distribution and is expressed in Equation (7).

$$\begin{aligned} f(Y \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \kappa) = & \prod_{i=1}^{n} \left[\frac{e^{\mathbf{x}_i^T \boldsymbol{\gamma}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\gamma}}} + \frac{1}{1+e^{\mathbf{x}_i^T \boldsymbol{\gamma}}} \left(\frac{1}{1+\kappa e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right)^{1/\kappa} \right] \\ & \times \prod_{i=1}^{n} \left[\frac{1}{1+e^{\mathbf{x}_i^T \boldsymbol{\gamma}}} \cdot \frac{\Gamma(y_i + 1/\kappa)}{\Gamma(1/\kappa)\, y_i!} \left(\frac{\kappa e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+\kappa e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right)^{y_i} \left(\frac{1}{1+\kappa e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right)^{1/\kappa} \right] \end{aligned} \tag{7}$$

The posterior distribution forms the basis for modeling using the Bayesian method, which combines two types of information: past data, which serves as a prior, and observation data, which serves as a constituent of the likelihood function. The posterior distribution used to

estimate ZINB regression parameters is given by Equation (8).

$$
\begin{aligned}
f(\boldsymbol{\beta},\boldsymbol{\gamma},\kappa \mid Y) \propto & \left[\prod_{j=0}^{m} \frac{s_{\beta_j}}{\pi\left(s_{\beta_j}^2+(\beta_j-l_{\beta_j})^2\right)} \cdot \prod_{k=0}^{m} \frac{s_{\gamma_k}}{\pi\left(s_{\gamma_k}^2+(\gamma_k-l_{\gamma_k})^2\right)} \cdot \frac{1}{b^a\Gamma(a)}\kappa^{a-1}e^{-\kappa/b}\right] \\
& \times \left[\prod_{i=1}^{n}\left(\frac{e^{\mathbf{x}_i^T\gamma}}{1+e^{\mathbf{x}_i^T\gamma}}+\frac{1}{1+e^{\mathbf{x}_i^T\gamma}}\left(\frac{1}{1+\kappa e^{\mathbf{x}_i^T\beta}}\right)^{1/\kappa}\right)\right. \\
& \left.\times \prod_{i=1}^{n}\left(\frac{1}{1+e^{\mathbf{x}_i^T\gamma}} \cdot \frac{\Gamma(y_i+1/\kappa)}{\Gamma(1/\kappa)y_i!}\left(\frac{\kappa e^{\mathbf{x}_i^T\beta}}{1+\kappa e^{\mathbf{x}_i^T\beta}}\right)^{y_i}\left(\frac{1}{1+\kappa e^{\mathbf{x}_i^T\beta}}\right)^{1/\kappa}\right)\right]
\end{aligned} \tag{8}
$$

2.6. Simulation Scenarios

Simulation method recreates situations and conditions using models for the purposes of learning, testing, training, evaluation, and improvement of system performance [10]. Simulation can be used to solve uncertain problems and consider possibilities that cannot be thoroughly accounted for. Simulation is a technique that uses a model of a real system to carry out experiments. It is beneficial as a decision-making tool for designing systems with specific performance requirements at both the design and operational stages.

The simulation data in this study were generated based on previously processed secondary data. This dataset is created with various characteristics to evaluate the ability of the ZINB regression model in handling over-dispersion [21]. The characteristics used to generate the response variables include:

1. $Y_i \sim \text{ZINB}(\mu_i, \kappa, p)$: μ_i and κ are obtained from the estimated parameters of the ZINB model based on secondary data.
2. The proportion of zeros (p): 0.3, 0.5, and 0.8.
3. Sample sizes (n): 38, 100, and 500.

The research conducted by [21] used zero proportions of 0.4, 0.6, and 0.8. In this study, the zero proportions used were $p = 0.3$, 0.5, and 0.8, representing low, moderate, and high levels of zero inflation. These variations were used to evaluate the sensitivity and performance of the ZINB model in the context of under-five deaths due to pneumonia, where zero data is common. Meanwhile, the sample sizes used were $n = 38$, 100, and 500, respectively reflecting small (referring to actual data), medium, and large data conditions. The combination of zero proportions and sample sizes allows for a more comprehensive evaluation of the performance of the ZINB regression model under various data conditions. The predictor variables are obtained from the original data through resampling using the bootstrap method. The data generation process was conducted using `R` software version 4.4.2.

3. Results and Discussion

This section presents the research results and analysis in a structured way. First, it tests for over-dispersion and excess zeros in the data. Then, it presents the results of the Zero-Inflated Negative Binomial (ZINB) Bayesian model estimation, simulation results, and an evaluation of the accuracy of the parameter estimation.

3.1. Over-dispersion

Over-dispersion in Poisson regression can be tested by comparing the chi-square value to the degrees of freedom [22]. Over-dispersion occurs if the dispersion value is greater than one. The dispersion value obtained from the test results is 1.28, which is greater than one. Therefore, it can be inferred that the response variable shows over-dispersion.

3.2. Excess Zero

Zero inflation is assessed by evaluating the proportion of zero values present in the response variable. The results of the excess zero check are presented in Table (1) below.

Table 1: Excess Zero Checking Result

Number of under-five deaths due to pneumonia	Percentage
0	81.58%
1	13.16%
3	2.63%
9	2.63%

Referring to Table 1, it can be seen that the response variable experiences an excess zero because the proportion of zero values exceeds 50%, which is 81.58%. Therefore, it can be concluded that the Poisson regression model is not appropriate for use in modeling these data.

3.3. Zero-Inflated Negative Binomial (ZINB) Bayesian Regression

The Zero-Inflated Negative Binomial (ZINB) Bayesian regression model was applied to cases of under-five deaths due to pneumonia in East Java. In this modeling, three predictor variables and one response variable were used. The ZINB Bayesian model is formed in several stages. First, the likelihood function is determined and the prior distribution is set for each parameter in the model. Next, the posterior distribution is formed using Markov Chain Monte Carlo (MCMC) simulation through the Gibbs Sampling algorithm. After obtaining the posterior distribution, convergence testing is performed to ensure the simulated samples adequately represent the actual posterior distribution. These steps are all important in forming the Bayesian ZINB regression model, which consists of two main components.
Model for log (negative binomial state): $\hat{\mu}_i$

$$\begin{aligned} \ln \hat{\mu}_i &= 6.75 + 1.07X_1 - 0.01X_2 - 0.09X_3 \\ \hat{\mu}_i &= \exp(6.75 + 1.07X_1 - 0.01X_2 - 0.09X_3) \end{aligned} \tag{9}$$

Model for logit (zero inflation state): $\hat{\pi}_i$

$$\begin{aligned} \operatorname{logit} \hat{\pi}_i &= -132.79 + 1.16X_1 + 0.23X_2 + 1.21X_3 \\ \hat{\pi}_i &= \frac{\exp(-132.79 + 1.16X_1 + 0.23X_2 + 1.21X_3)}{1 + \exp(-132.79 + 1.16X_1 + 0.23X_2 + 1.21X_3)} \end{aligned} \tag{10}$$

Equation (9) shows that a 1% increase in the number of under-fives aged 0-59 months with a malnutrition status can enhance the average number of under-five deaths due to pneumonia by 1.07 or 2.92%. Conversely, a 1% increase in exclusive breastfeeding and complete basic immunization coverage in infants can decrease the average number of under-five deaths due to pneumonia. This interpretation shows that the regression model accurately reflects the relationship between health factors and pneumonia-related deaths in children under five.

3.4. Simulation Results of ZINB Bayesian Regression

Simulation studies are used to validate that the decision made is optimal [11]. Simulation data is generated based on initial ZINB parameters. Fig. 1 presents the simulation results for estimating the parameters $\hat{\beta}_j$ and $\hat{\gamma}_j$.

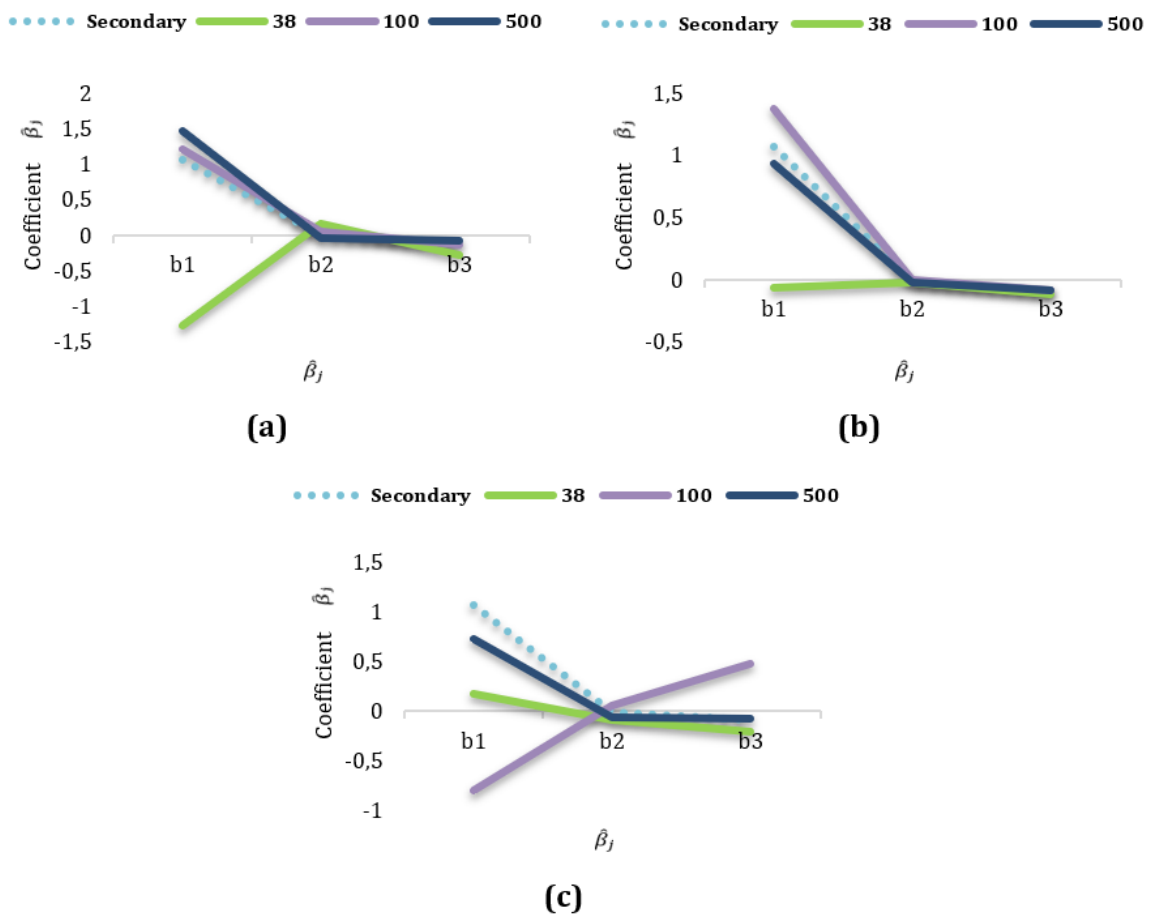


Fig. 1: Plots of parameter estimation $\hat{\beta}_j$ at various sample sizes and proportions of zero values: (a) $p = 0.3$, (b) $p = 0.5$, and (c) $p = 0.8$

As can be seen in Fig. 1, the parameter estimates $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ tend to be inconsistent in small sample sizes ($n = 38$). This reflects the instability of parameter estimates in small samples. Meanwhile, the parameters are more stable and closer to their initial values when $n = 100$ and 500. Furthermore, simulations with various proportions of zeros ($p = 0.3, 0.5, 0.8$) show that the results of parameter estimation tend to be better at a proportion of 0.5. This is because, in proportion of 0.5, each parameter shows a smaller variance than in other proportions.

The parameter estimation values of $\hat{\gamma}_1$, $\hat{\gamma}_2$, and $\hat{\gamma}_3$ demonstrate that increasing the sample size impacts the stability of the estimation results across the tested scenarios, as shown respectively in Fig. 2.a, Fig. 2.b, and Fig. 2.c. At small ($n = 38$) and medium ($n = 100$) sample sizes, the parameter estimates exhibit greater variability. This suggests that increasing the sample size will improve classification accuracy. When viewed based on the proportion of zero values, the parameter estimates are stable at $p = 0.5$. With a small proportion of zero values, zero inflation becomes less dominant and can cause fluctuations in the parameter estimates. However, this effect can be reduced at larger sample sizes. In general, it can be concluded that the Bayesian approach to simulation requires a large sample size and an appropriate proportion of zero values to produce stable and accurate parameter estimates when applying the ZINB regression model.

3.5. Accuracy of Parameter Estimation

An estimator is considered good if the value it produces is nearly equal to the true parameter value. The accuracy of parameter estimation can be evaluated using a measure of bias. Bias can be calculated from the difference between the mean value of the parameter estimator and the true parameter value, expressed as $E(\hat{\beta}) - \beta$ or $E(\hat{\gamma}) - \gamma$. This bias reflects how far the

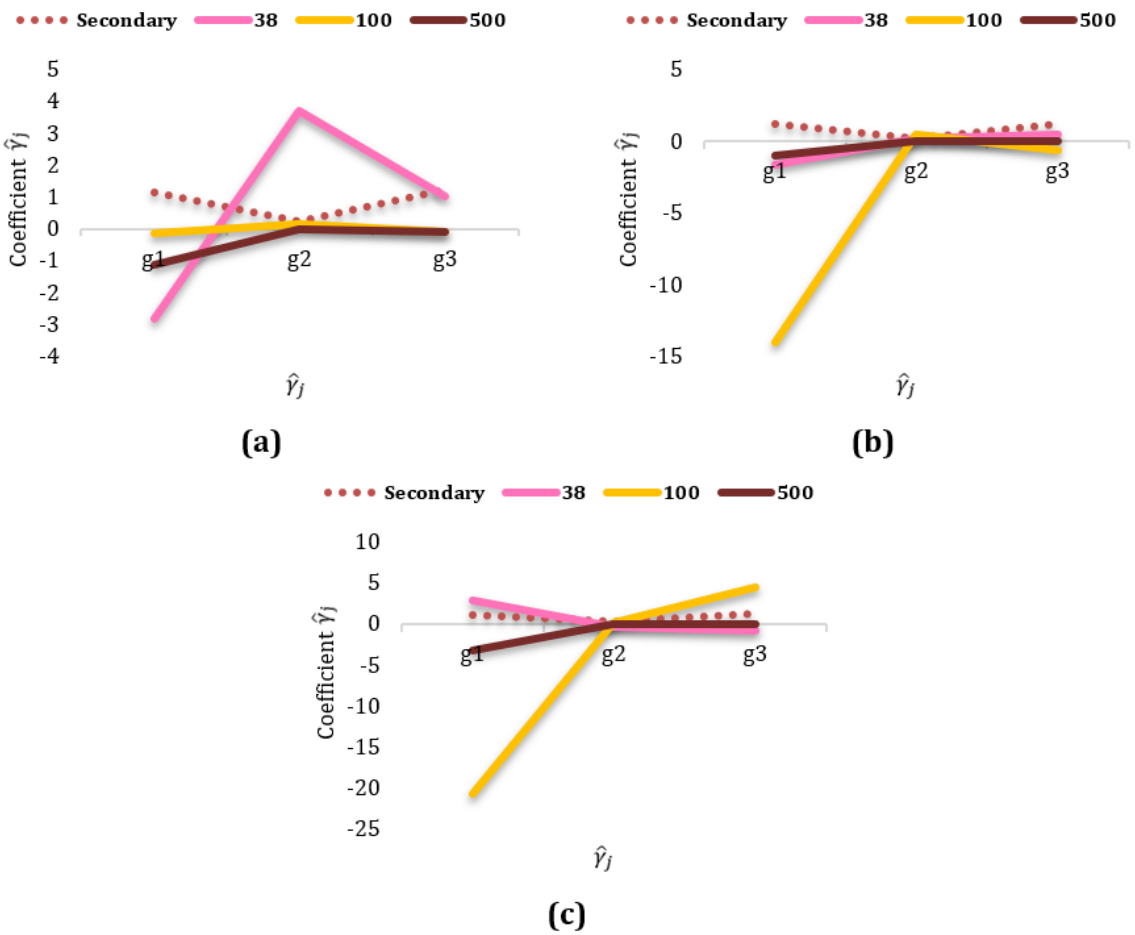


Fig. 2: Plots of parameter estimation $\hat{\gamma}_j$ at various sample sizes and proportions of zero values: (a) $p = 0.3$, (b) $p = 0.5$, and (c) $p = 0.8$

estimation results deviate from the actual parameter. A bias value close to zero indicates that the parameter estimator has a small bias, suggesting that the estimation method is capable of producing values that are representative of the actual parameter. The following table shows the regression coefficient bias values for each simulation scenario.

Table 2: Regression Coefficient Bias Values in Each Simulation Scenario

Parameter	$n = 38$			$n = 100$			$n = 500$		
	$p = 0.3$	$p = 0.5$	$p = 0.8$	$p = 0.3$	$p = 0.5$	$p = 0.8$	$p = 0.3$	$p = 0.5$	$p = 0.8$
$\hat{\beta}_1$	-2.35	0.17	-0.19	0.14	0.06	-0.04	0.40	-0.03	0.01
$\hat{\beta}_2$	-1.14	-0.01	-0.03	0.30	0.01	0.01	-0.14	-0.01	0.01
$\hat{\beta}_3$	-0.90	-0.08	-0.11	-1.15	0.07	0.57	-0.35	-0.05	0.01
$\hat{\gamma}_1$	-4.00	3.46	-0.21	-1.31	-0.09	-1.33	-2.30	-0.24	-1.29
$\hat{\gamma}_2$	-2.82	-0.02	-0.77	-15.19	0.19	-1.88	-2.16	-0.26	-1.25
$\hat{\gamma}_3$	1.68	-0.65	-2.08	-21.88	-0.06	3.27	-4.34	-0.29	-1.25

As presented in Table 2, the bias values of the $\hat{\beta}$ estimator tend to approach zero as the sample size increases, particularly at $n = 500$, with values ranging around 0.01. This indicates that the estimation method provides results that closely approximate the true parameter values, suggesting a low bias level. Moreover, the model demonstrates sufficient flexibility to accommodate the complexity of larger sample sizes. In terms of zero proportions, the bias value of the $\hat{\beta}$ estimator also increases at higher proportions of zero values. These findings support the suitability of the ZINB Bayesian model for datasets characterized by high zero-inflation and over-dispersion, as

the zero-inflation component plays a key role in capturing the underlying structure of such data.

4. Conclusion

Based on the research results, it can be inferred that Zero-Inflated Negative Binomial regression analysis with a Bayesian approach using Cauchy prior produces more stable parameter estimates as both the sample size and the proportion of zero values increase. In the context of under-five deaths due to pneumonia, data often contain many zeros because not all regions report cases. The ZINB model can handle over-dispersion and excess zeros with two components: the negative binomial model and the zero inflation model. This results in more accurate and relevant modeling for informed policymaking. Additionally, the Bayesian approach allows for greater flexibility in incorporating prior information and addressing small sample sizes, which are common in health research. These findings confirm that the ZINB model is suitable for health data with rare events and excess zeros.

The number and variety of simulation scenarios used in this research is limited, so the scenarios do not reflect all possible data conditions that may occur. Therefore, future research should expand the scope of simulation scenarios. Furthermore, incorporating other relevant variables in future research may improve the quality of the analysis and strengthen the model's ability to describe phenomena, particularly in pneumonia studies in the health sector.

Credit Authorship Contribution Statement

Santi Wahyu Salsabila: Conceptualization, Methodology, Writing–Original Draft, Validation, Visualization, Software, Data Curation. **Achmad Efendi:** Software, Writing–Review & Editing, Supervision. **Nurjannah:** Data Curation, Writing–Review & Editing, Supervision.

Declaration of Generative AI and AI-assisted technologies

No generative AI or AI-assisted technologies were used during the preparation of this manuscript.

Declaration of Competing Interest

The authors declare that no competing interests.

Funding and Acknowledgments

This research received no external funding.

Data Availability

The dataset analyzed during the current study is publicly available in the East Java Provincial Health Office website[1].

References

[1] Adrian Colin Cameron and Pravin K Trivedi. *Regression Analysis of Count Data*. Cambridge university press, 2013. DOI: 10.1017/CBO9781139013567.

[2] Joseph M Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2011. DOI: 10.1017/CBO9780511973420.

[1] https://dinkes.jatimprov.go.id

[3] Cindy Cahyaning Astuti, Agnes Ona Bliti Puka, and Akbar Wiguna. "Zero-Inflated Negative Binomial Modeling in Infant Death Case Due To Pneumonia in East Java Province". In: *BAREKENG: Jurnal Ilmu Matematika dan Terapan* 17.4 (2023), pp. 1835–1844. DOI: 10.30598/barekengvol17iss4pp1835-1844.

[4] Moh Irfan Agus Saputro and Mohammad Farhan Qudratullah. "Estimation of zero-inflated negative binomial regression parameters using the maximum likelihood method (case study: factors affecting infant mortality in Wonogiri in 2015)". In: *Proceeding International Conference on Science and Engineering*. Vol. 4. 2021, pp. 240–254. http://sunankalijaga.org/prosiding/index.php/icse/article/view/666.

[5] Kristin Rulin Katianda, Rito Goejantoro, and Andi M Ade Satriya. "Estimasi Parameter Model Regresi Linier dengan Pendekatan Bayes". In: *EKSPONENSIAL* 11.2 (2021), pp. 127–132. DOI: 10.30872/eksponensial.v11i2.653.

[6] Sepideh Roohi, Mohammad Reza Baneshi, Alireza Noroozi, Ahmad Hajebi, and Abbas Bahrampour. "Comparing Bayesian regression and classic zero-inflated negative binomial on size estimation of people who use alcohol". In: *Journal of Biostatistics and Epidemiology* 2.4 (Oct. 2017), pp. 173–179. https://jbe.tums.ac.ir/index.php/jbe/article/view/130.

[7] Mekuanint Simeneh Workie and Abebaw Gedef Azene. "Bayesian negative binomial logit hurdle and zero-inflated model for characterizing smoking intensity". In: *Journal of Big Data* 8.1 (2021). DOI: 10.1186/s40537-021-00452-8.

[8] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu Sung Su. "A weakly informative default prior distribution for logistic and other regression models". In: *Annals of Applied Statistics* 2.4 (2008), pp. 1360–1383. DOI: 10.1214/08-AOAS191.

[9] Ioannis Ntzoufras. *Bayesian modeling using WinBUGS*. John Wiley and Sons, 2009. DOI: 10.1002/9780470434567.

[10] C. Harrell, B.K. Ghosh, and R. Bowden. *Simulation Using ProModel*. McGraw-Hill series in industrial engineering and management science. McGraw-Hill/Higher Education, 2004. https://books.google.co.id/books?id=ou8G2njfMzQC.

[11] Averill M Law, W David Kelton, and W David Kelton. *Simulation modeling and analysis*. Mcgraw-hill New York, 2007. https://archive.org/details/simulationmodeli0000lawa.

[12] Dinkes Jawa Timur. "Profil Kesehatan Provinsi Jawa Timur Tahun 2023". In: *Dinas Kesehatan Provinsi Jawa Timur* (2024). https://dinkes.jatimprov.go.id/userfile/dokumen/PROFIL%20KESEHATAN%20PROVINSI%20JAWA%20TIMUR%20TAHUN%202023.pdf.

[13] Dian Eka Puspitasari and Fariani Syahrul. "Faktor Risiko Pneumonia pada Balita Berdasarkan Status Imunisasi Campak dan Status ASI Eksklusif". In: *Journal of Universitas Airlangga* 3.1 (2015), pp. 69–81. https://e-journal.unair.ac.id/index.php/JBE/article/download/1315/1074.

[14] Riza Rigustia, Laura Zeffira, and Ade Teti Vani. "Faktor Risiko yang Berhubungan dengan Kejadian Pneumonia pada Balita di Puskesmas Ikur Koto Kota Padang". In: *Health and Medical Journal* 1.1 (2019), pp. 22–29. DOI: 10.33854/heme.v1i1.215.

[15] Yulianti Kristiani Banhae, Yohanes Mau Abanit, and Domianus Namuwali. "Faktor Risiko yang Berhubungan dengan Kejadian Pneumonia pada Balita di Kota Kupang". In: *Jurnal Ilmiah Permas: Jurnal Ilmiah STIKES Kendal* 13.3 (2023), pp. 1099–1106. DOI: 10.32583/pskm.v13i3.1138.

[16] Felix Famoye and John S. Preisser. "Marginalized zero-inflated generalized Poisson regression". In: *Journal of Applied Statistics* 45.7 (2018), pp. 1247–1259. DOI: 10.1080/02664763.2017.1364717.

[17] Aldo M Garay, Elizabeth M Hashimoto, Edwin MM Ortega, and Victor H Lachos. "On estimation and influence diagnostics for zero-inflated negative binomial regression models". In: *Computational Statistics & Data Analysis* 55.3 (2011), pp. 1304–1318. DOI: 10.1016/j.csda.2010.09.019.

[18] Aldo M. Garay, Victor H. Lachos, and Heleno Bolfarine. "Bayesian estimation and case influence diagnostics for the zero-inflated negative binomial regression model". In: *Journal of Applied Statistics* 42.6 (2015), pp. 1148–1165. DOI: 10.1080/02664763.2014.995610.

[19] Hossein Zamani and Noriszura Ismail. "Functional form for the zero-inflated generalized poisson regression model". In: *Communications in Statistics - Theory and Methods* 43.3 (2014), pp. 515–529. DOI: 10.1080/03610926.2012.665553.

[20] Lee J Bain and Max Engelhardt. *Introduction to Probability and Mathematical Statistics*. Duxbury Press Belmont, CA, 2000. https://openlibrary.org/books/OL7784515M/Introduction_to_Probability_and_Mathematical_Statistics_(Duxbury_Classic).

[21] Rahma Fitriani, Lidia Novita Chrisdiana, and Achmad Efendi. "Simulation on the Zero Inflated Negative Binomial (ZINB) to Model Overdispersed, Poisson Distributed Data". In: *IOP Conference Series: Materials Science and Engineering* 546.5 (2019). DOI: 10.1088/1757-899X/546/5/052025.

[22] Elizabeth H Payne, Mulugeta Gebregziabher, James W Hardin, Viswanathan Ramakrishnan, and Leonard E Egede. "An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data". In: *Communications in Statistics-Simulation and Computation* 47.6 (2018), pp. 1722–1738. DOI: 10.1080/03610918.2017.1323223.