JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 10, No. 1, January 2026, pp. 275-288

Modeling Zero-Inflated Poisson Invers Gaussian Regression Bayesian Approach

Berliana Jannah[1*], Ni Wayan Surya Wardhani[1], Eni Sumarminingsih[1]
[1]Department Statistics, Brawijaya University, Indonesia
berlianajannah@student.ub.ac.id

Article History:
Received : 15-08-2025
Revised : 26-10-2025
Accepted : 28-10-2025
Online : 17-01-2026

Keywords:
ZIPIGR;
Overdispersion;
Bayesian;
Excess Zero;
Gibbs Sampling.

ABSTRACT

Deaths due to dengue hemorrhagic fever (DHF) remains one of the most pressing public health issues in Indonesia, especially in urban areas such as Semarang City, which has a high population density and diverse environmental conditions that potentially increase the risk of transmission and death from DHF. This study aims to model the number of DHF in Semarang City using a Bayesian-based Zero-Inflated Poisson Inverse Gaussian Regression (ZIPIGR) approach. The research data was obtained from the Semarang City Health Office and the Central Statistics Agency (BPS) in 2024, with the response variable being the number of DHF deaths and five predictor variables. The data showed overdispersion and a high proportion of zeros (around 50%), indicating the presence of excess zeros in count data with a small sample size. The Bayesian ZIPIGR method was chosen because it can produce more stable parameter estimates than classical methods such as Maximum Likelihood Estimation (MLE), especially for data with complex likelihood functions, small sample sizes, and many zero values. Parameter estimation was performed using Gibbs Sampling simulation in the Markov Chain Monte Carlo (MCMC) framework. The results show that the Bayesian ZIPIGR model performs better than the MLE ZIPIGR model based on the Root Mean Square Error (RMSE) value. Factors that significantly influence DHF mortality are population density, slum area, and number of health workers. These results confirm that regional density and health worker capacity play an important role in increasing the risk of DHF mortality in urban areas. The developed model has been proven to be highly accurate in modeling count data with excess zero characteristics and makes an important contribution to health policy formulation. In practical terms, this model can be used to improve early warning systems and DHF control strategies in densely populated urban areas such as the city of Semarang.

https://doi.org/10.31764/jtam.v10i1.34068

A. INTRODUCTION

Poisson distribution is a discrete distribution with random variable values in the form of positive integers (Akinkunmi, 2019). Poisson regression assumes that the mean and variance of the response variable are equal, a condition known as equidispersion (Bektashi et al., 2022). However, the mean and variance of enumerated data are often not equal (Agresti, 2019), either the mean is greater than the variance (overdispersion) or the mean is less than the variance (underdispersion). In other words, the assumption of equidispersion is often violated. Enumerated data often show considerable variance because they contain many extra zeros or scatter that is larger than the values in the data or both (Aswi et al., 2022).

One approach to address overdispersion is by developing models that combine the Poisson distribution with various discrete or continuous distributions, known as mixed Poisson distributions (Lambert, 1992). While this method offers an alternative solution for overdispersion, only a few types are commonly applied in research because of the complexity of their calculations (Payne et al., 2017). Some mixed poisson distributions that have been developed are Zero-Inflated Poisson (ZIP), Generalized Poisson, Negative Binomial Poisson and Poisson Inverse Gaussian (PIG).

The Zero-Inflated Poisson (ZIP) model is a simple mixture model for discrete data with many zero values (Lambert, 1992). ZIP regression is able to control overdispersion in the Poisson distribution and zero value inflation so that the accuracy of parameter estimation can be guaranteed (Rahayu et al., 2016). In general, the ZIP regression model is still rarely used for count data that shows inflation due to zero values and overdispersion. Several studies related to Poisson regression problems and their applications from time to time always experience developments. Accoarding to the research of Amalia et al. (2021) found that Zero-Inflated Poisson (ZIP) can be used to analyze the number of excess zeros and overdispersion. Also Abdulhafedh (2023) conducted research comparing the Poisson regression, negative binomial regression, and zero-inflated Poisson (ZIP) models, with the result that ZIP is more effective for handling excess zeros in traffic accident data.

The Poisson Inverse Gaussian (PIG) distribution, introduced by Holla in 1966, is a form of mixed Poisson distribution where the random effect is modeled using an inverse Gaussian distribution ((Karlis & Xekalaki, 2005). Some research on Poisson Inverse Gaussian (PIG) regression has been conducted by Putri et al. (2020), the study found that the Poisson-Inverse Gaussian regression model provides a better fit than the Negative Binomial regression model for overdispersed count data, as indicated by its higher pseudo R-squared value in the horseshoe crabs case study. The research of Zha et al. (2016) used PIG regression modeling to analyze the number of motorcycle accidents that occurred in Texas and Washinton, by comparing the negative binomial regression model and the PIG regression model on the Akaike Information Criterion (AIC) value and the Bayesian Information Criterion (BIC) value.

Zero-Inflated Poisson Inverse Gaussian (ZIPIG) is a development of the Zero-Inflated Poisson (ZIP) and Poisson Inverse Gaussian (PIG) models (Hilbe, 2014). The ZIPIG model is very effective for handling data that experience overdispersion and have excess zero data properties. Chakraborty & Biswas (2024) applied the Zero Inflated Poisson Inverse Gaussian (ZIPIG) model to health data. The study found that the Zero-Inflated Poisson Inverse Gaussian (ZIPIG) regression model outperforms the Zero-Inflated Negative Binomial (ZINB) model in predicting dengue cases in Bangladesh, based on AIC and BIC criteria. According to Purhadi & Ermawati (2021) the Bivariate Zero-Inflated Poisson Inverse Gaussian Regression (BZIPIGR) model is effective for data experiencing overdispersion due to zero inflation in HIV and AIDS cases in Trenggalek and Ponorogo City. BZIPIGR is a development of ZIPIG, which uses two response variables commonly referred to as bivariate. Therefore, the ZIPIG model is recommended as a more suitable approach for modelling DHF incidence in similar over-dispersed and zero-inflated datasets.

Parameter estimation in the Zero-Inflated Poisson Inverse Gaussian (ZIPIG) model is generally done using Maximum Likelihood Estimation (MLE). MLE is chosen because it is

efficient and consistent on large sample sizes. According to Psutka & Psutka (2019), MLE works optimally when the sample is large, because the asymptotic assumptions underlying this method can be met. Azizan et al. (2020) revealed that the MLE approach produced low accurate and high bias estimates of the item parameters in small sample sizes regardless of the number of items. However, at small sample sizes, MLE tends to produce unstable parameter estimates and can result in bias, so caution is needed in its use in the context of limited samples. This limitation of MLE in handling small samples encourages the development and application of alternative methods, one of which is the Bayesian method. Utomo et al. (2025) found that Zero-Inflated Poisson (ZIP) with the Bayesian approarch is better than MLE approach as shown in the simulation study results on several small, medium and low sample sizes.

The number of deaths due to Dengue Fever (DHF) is one example of enumerated data in the health sector. According to Kemenkes RI (2022), Dengue Fever (DHF) is an infectious disease caused by the DHF virus carried by female Aedes aegypti & Aedes albopictus mosquitoes. This disease usually occurs in tropical and subtropical regions, where Southeast Asia is in the subtropical region. The Ministry of Health until the 15th week of 2024 stated that the total cases of morbidity cases due to DHF reported were 62,001 cases. The highest cases are in West Java (17,331 cases), Banten (5,877 cases), and Central Java (4,330 cases), while there are 475 cases of death due to DHF, with the highest deaths reported in West Java (158 cases), Central Java (105 cases), and East Java (37 cases). The high number of DHF deaths in Central Java needs to be a concern. DHF is still a problem that must be solved in these various regions because it is the province with the second most cases in Indonesia. Various treatments to prevent the spread of this virus have been carried out by the Central Java Provincial Health Office. However, the spread of this disease continues in this province.

Based on this description, the data related to the mortality rate due to DHF contains almost 50% zero values, so the researcher plans to develop a Zero-Inflated Poisson Inverse Gaussian (ZIPIG) analysis using the Bayesian parameter estimation approach to evaluating the best model between ZIPIG using Maximum Likelihood Estimation (MLE) and Bayesian on data deaths due to Dengue Hemorrhagic Fever (DHF) in Semarang City, as well as creating a model using Bayesian parameter estimators and determining the factors that significantly affect the number of deaths due to DHF in Semarang City.

B. METHODS

1. Data and Research Variable

This study is a quantitative study using secondary data obtained from the Semarang City Health Office and the Central Statistics Agency (BPS) in 2024, consisting of 16 subdistricts as observation units. The response variable in this study was the number of deaths due to DHF, while the predictor variables consisted of the number of DHF cases (X_1), population density (X_2), percentage of clean water sources (X_3), area of slums (X_4), and number of health workers (X_5). The analysis stages were carried out as follows, starting with (1) descriptive analysis to determine the characteristics of the data, followed by detecting overdispersion, multicollinearity, and zero inflation, then (2) formulating the Zero-Inflated Poisson Inverse Gaussian Regression (ZIPIGR) model, and (3) estimating the parameters using the MLE and Bayesian approaches. The Bayesian approach used the Gibbs Sampling method in Markov Chain

Monte Carlo (MCMC). The estimation process was carried out using R studio software, with 20,000 MCMC iterations, 5,000 initial iterations (burn-in) deleted, and a thinning interval of 10 to reduce autocorrelation between samples. Lastly, (4) MLE and Bayesian were compared using the smallest RMSE value.

In the context of dengue hemorrhagic fever (DHF) cases, the Poisson–Inverse Gaussian component describes the number of deaths due to DHF, while the zero-inflated component represents subdistricts that have no deaths, which may happen due to successful health interventions or random variation. Model performance was evaluated using Root Mean Square Error (RMSE) to compare the Bayesian ZIPIGR model and the Maximum Likelihood Estimation (MLE) ZIPIGR model, where a smaller RMSE value indicates better prediction accuracy. The application of this model provides an understanding of how factors such as population density, slum area size, access to clean drinking water, and the number of health workers affect deaths from DHF, thereby providing a basis for evidence-based health policies for the Semarang City Government in its efforts to control and prevent DHF.

2. Multicollinearity

According to Kyriazos & Poga (2023), The Variance Inflation Factor (VIF) indicates how much the regression coefficients are inflated as a result of multicollinearity. VIF measures how much the variance of regression coefficient estimates increases when multicollinearity occurs. A high VIF indicates high multicollinearity. If VIF = 1, the relationship between the predictor variables is mutually free (no multicollinearity occurs), A VIF value of 1 indicates no multicollinearity, while values greater than 1 reflect increasing levels of multicollinearity. Generally, a VIF exceeding 5 or 10 is considered high and signals serious multicollinearity issues.

3. Overdispersion

In Poisson regression, one essential assumption is that the mean and variance of the response variable are equal (equidispersion). Overdispersion happens when the variance exceeds the mean, which can result from positive correlation or excessive variation in the response probabilities.

$$VT = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{\bar{y}} = (n-1)\frac{S^2}{\bar{y}} \tag{1}$$

This value is equal to the variance-to-ragam ratio, often referred to as the dispersion index, multiplied by n-1, where n is the sample size. If the value of the dispersion index is less than 1, it can be said that there is underdispersion, whereas overdispersion occurs when the dispersion index is more than 1 (Handarzeni, 2022).

4. Zero Inflated Poisson Invers Gaussian Regression (ZIPIGR)

Zero-Inflated Poisson Inverse Gaussian regression is a combined regression coding of Zero-Inflated distribution and Inverse Gaussian distribution. The ZIP model deals with excess zeros and the PIG model deals with overdispersion in the data. There are three parameters in the

ZIPIG model, namely the mean (μ), dispersion (τ), and zero inflated (p). Zero-Inflated Poisson Inverse Gaussian can be written as $Y_i \sim ZIPIG(\mu, \tau, p)$ with the following propability function.

$$P(Y = y \mid \mu, \tau, p) = \begin{cases} p + (1-p)P(Y = 0 \mid \mu, \tau), & \text{for } y = 0 \\ (1-p)P(Y = 0 \mid \mu, \tau), & \text{for } y = 1,2,3 \end{cases} \quad (2)$$

where for $y = 0$ written as follows:

$$P(Y = 0|\mu, \tau, p) = p + (1-p)\exp\left(\frac{1}{\tau}\right)\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}(2\mu\tau + 1)^{\frac{1}{4}}\left(\frac{\pi}{2\left(\frac{1}{\tau}\sqrt{2\mu\tau+1}\right)}\right)^{\frac{1}{2}}\exp\left(\frac{1}{\tau}\sqrt{2\mu\tau+1}\right) \quad (3)$$

where as for y=1,2, ... is written as follows:

$$(Y = y|\mu, \tau, p) = (1-p)\frac{\mu^y}{y!}\exp\left(\frac{1}{\tau}\right)\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}(2\mu\tau + 1)^{-\frac{\left(y-\frac{1}{2}\right)}{2}}\left(\frac{\pi}{2\left(\frac{1}{\tau}\sqrt{2\mu\tau+1}\right)}\right)^{y-\frac{1}{2}}\exp\left(\frac{1}{\tau}\sqrt{2\mu\tau+1}\right) \quad (4)$$

with $p = \frac{\exp(-\mu X_i^T \beta)}{1+\exp(-\mu X_i^T \beta)}$ and $1 - p = \frac{1}{1+\exp(-\mu X_i^T \beta)}$

Suppose a response variable $Y_i \sim ZIPIG(\mu, \tau, p)$ then the ZIPIG regression model can be written in two model components, namely the component for the Poisson state model (μ) and the zero-inflated model component, written:

for the model μ

$$\mu = \exp(\boldsymbol{X^T\beta}) \quad (5)$$

$$\ln(\mu) = \mathbf{X}^T\boldsymbol{\beta} \quad (6)$$

for zero inflated model is:

$$logit(p) = ln\frac{p}{1-p} = -\gamma\mathbf{X}_i^T\boldsymbol{\beta} \quad (7)$$

Distributing Equations (6) and (7) to Equations (3) and (4), the probability function for the ZIPIG regression model when $y = 0$ is obtained:

$$P(Y = 0|\boldsymbol{\beta}, \tau, p) = \frac{exp(-\mu X_i^T\beta)}{1 + exp(-\mu X_i^T\beta)} + \frac{1}{1 + exp(-\mu X_i^T\beta)}exp\left(\frac{1}{\tau}\right)\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}(2\,exp(\mathbf{X}^T\boldsymbol{\beta})\,\tau + 1)^{\frac{1}{4}} \left(\frac{\pi}{2\left(\frac{1}{\tau}\sqrt{2\exp(\mathbf{X}^T\boldsymbol{\beta})\,\tau+1}\right)}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{\tau}\sqrt{2\exp(\mathbf{X}^T\boldsymbol{\beta})\,\tau+1}\right) \quad (8)$$

for $y = 1,2, \ldots$ is written as follows:

$$P(Y=y|\boldsymbol{\beta},\tau,p)=\frac{1}{1+exp(-\mu X_i^T\beta)}\left(\frac{(exp(X_i^T\beta))^{y_i}exp\left(\frac{1}{\tau}\right)}{y_i!}\right)\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}(2\,exp(\mathbf{X}^T\boldsymbol{\beta})\,\tau+1)^{-\frac{\left(y_i-\frac{1}{2}\right)}{2}}$$
$$\left(\frac{\pi}{2\left(\frac{1}{\tau}\sqrt{2\exp(\mathbf{X}^T\boldsymbol{\beta})\,\tau+1+1}\right)}\right)^{y_i-\frac{1}{2}}\exp\left(-\frac{1}{\tau}\sqrt{2\exp(\mathbf{X}^T\boldsymbol{\beta})\,\tau+1+1}\right) \quad (9)$$

5. Bayesian Zero Inflated Poisson Inverse Gaussian Regression

The ZIPIG model has 2 joint distributions, so 2 priors are obtained for each of the ZIP and PIG models. According to Liu & Powers (2012), without prior knowledge of the distribution of the parameters, determining the prior distribution can use informative priors. The ZIPIG model parameters **α** and **β** are determined to be Normal (μ,σ^2) distributed so that the prior distribution can be written as:

$$f(\alpha,\beta)=\prod_{j=0}^{m}\left[\frac{1}{\sqrt{2\pi}\sigma_{aj}}e^{\left\{-\frac{\left(a_j-\mu_{aj}\right)^2}{2\sigma_{aj}^2}\right\}}\right]\times\prod_{j=0}^{l}\left[\frac{1}{\sqrt{2\pi}\sigma_{\beta j}}e^{\left\{-\frac{\left(\beta_j-\mu_{\beta j}\right)^2}{2\sigma_{\beta j}^2}\right\}}\right] \quad (10)$$

According to Lijoi et al. (2005) the PIG model $\alpha \geq 0$ and $\gamma > 0$, are denoted as $V \sim IG(\alpha,\gamma)$, so the prior distribution can be written as:

$$f(v)=\frac{\alpha}{\sqrt{2\pi}}v^{-\frac{3}{2}}exp\left[-\frac{1}{2}\left(\frac{\alpha^2}{v}+\gamma^2 v\right)+\gamma\alpha\right] \quad (11)$$

Therefore, the posterior distribution is the multiplication of the prior and likelihood distributions. Based on equations (10) and (11), the posterior distribution for the ZIPIG model is:

$$f(\alpha,\beta|y) \propto f(y|\alpha,\beta)f(\alpha,\beta)$$

$$\begin{aligned} f(\alpha,\beta|y) \propto & \prod_{\substack{i=1 \\ y_i=0}}^{n} \frac{1}{1+exp(-\mu \boldsymbol{X}_i^T \boldsymbol{\beta})} \left[exp(-\mu \boldsymbol{X}_i^T \boldsymbol{\beta}) + \exp\left(\frac{1}{\tau} - \frac{1}{\tau}\sqrt{(2(\exp(\mathbf{X}^T\boldsymbol{\beta}))\tau)+1}\right)\right] \\ & \times \prod_{\substack{i=1 \\ y_i>0}}^{n} \left(\frac{1}{1+exp(-\mu \boldsymbol{X}_i^T \boldsymbol{\beta})}\right)\left(\frac{(exp(-\gamma \boldsymbol{X}_i^T \boldsymbol{\beta}))^{y_i}\exp\left(\frac{1}{\tau}\right)}{y_i!}\right)\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}} [2\, exp(-\mu \boldsymbol{X}_i^T \boldsymbol{\beta})\,\tau \\ & + 1]^{-\left(\frac{y_i-\frac{1}{2}}{2}\right)} \left(\frac{\pi}{2\left(\frac{1}{\tau}\sqrt{(2(\exp(\mathbf{X}^T\boldsymbol{\beta}))\tau)+1}\right)}\right)^{y_i-\frac{1}{2}} \exp\left(-\frac{1}{\tau}\sqrt{(2(\exp(\mathbf{X}^T\boldsymbol{\beta}))\tau)+1}\right) \\ & \times \prod_{j=0}^{m}\left[\frac{1}{\sqrt{2\pi}\sigma_{aj}} e^{\left\{-\frac{(a_j-\mu_{aj})^2}{2\sigma_{aj}^2}\right\}}\right] \times \prod_{j=0}^{l}\left[\frac{1}{\sqrt{2\pi}\sigma_{\beta j}} e^{\left\{-\frac{(\beta_j-\mu_{\beta j})^2}{2\sigma_{\beta j}^2}\right\}}\right] \\ & \times \frac{\alpha}{\sqrt{2\pi}} v^{-\frac{3}{2}} exp\left[-\frac{1}{2}\left(\frac{\alpha^2}{v}+\gamma^2 v\right)+\gamma\alpha\right] \end{aligned} \tag{12}$$

6. Convergence Test

The MCMC convergence check is used to determine whether the generated samples are in accordance with the target distribution, namely the posterior distribution. MCMC convergence check can use trace plot, MC Error and autocorrelation. The formula for calculating MC error is:

$$MCE[G(\theta)] = \sqrt{\frac{1}{K(K-1)}\sum_{b=1}^{K}\left(\bar{G}(\theta)_b - G(\theta)\right)^2} \tag{13}$$

where, $\bar{G}(\theta)_b$ is the sample mean of each batch, $G(\theta)$ is the general sample mean, K is the number of batches.

7. Credible Interval

Accoarding to Hespanhol et al. (2019) states that testing the parameters of the Bayesian method uses the Credible Interval which uses the lower limit of the 2.5% percentile and the upper limit of the 97.5% percentile. The test is to determine the effect of each predictor variable on the response variable with the following hypothesis:

H_0: $\beta = 0$; there is no significant effect of the independent variable on the response variable

H_1: $\beta \neq 0$; there is a significant effect of the independent variable on the response variable

The decision criteria for rejecting or accepting $H_0$is based on whether or not a zero value appears in the Credible Interval of each parameter. If it contains zero value, then $H_0$is rejected.

8. Model Goodness Criteria

The criterion used to measure the goodness of the model after obtaining a model is the Root Mean Square Error (RMSE). RMSE is used based on the estimation error. The error shows how much the secondary data estimation results differ from the simulated data estimation values. This value is used to determine which model is the best. The RMSE formula is as follows.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_i^{(0)}\right)^2} \tag{14}$$

where, n is the number of simulations or observations, $\hat{\theta}_i$ is Estimation of parameters in the i-th simulation, and $\theta_i^{(0)}$ is the true value of the parameter in the i-th simulation.

C. RESULT AND DISCUSSION

1. Multicollinearity

The following are the results of testing the non-multicollinearity assumption using Rstudio software, as shown in Table 1.

Table 1. VIF Results of Each Model

Variables	VIF_j
The Number of DHF Cases (X_1)	5,981
Population Density (X_2),	1,574
Percentage Of Potable Water Sources (X_3),	1,134
Slum Area(X_4),	1,579
And The Number ff Health Workers (X_5)	6,055

Based on Table 1, it can be seen that the VIF value in all predictor variables < 10 so that it can be said that the number of DHF cases (X_1), percentage of population (X_2), percentage of potable water sources (X_3), slum area (X_4), number of health workers (X_5) in Semarang City are independent of each other.

2. Overdispersion and Excess Zero

In Poisson regression, the assumption is that the average is equal to the variance, if the variance is greater than the average, the data is overdispersed or the variance is smaller than the average, the data is underdispersed. Based on the overdispersion test, the Chi-Square value is 248.934 and the p-value <0.01 so it can be concluded that the data is overdispersed. Excess zero testing is done by calculating the ratio of the proportion of zeros in the data to the expectation of zero in the Poisson distribution. Excess zero data is generally characterized by a higher proportion of zero values compared to other values, with percentages ranging from 50% to 90% (Bimali et al., 2021). The DHF death data in Semarang City has a zero value of 62.5%, which is more than 50%, so the data is interpreted as zero-inflation.

3. Zero Inflation Poisson Invers Gaussian Regression (ZIPIG) with MLE

The results of the parameter estimates from the analysis of the data number of measles used MLE are as the following in Table 2.

Table 2. Results of ZIPIG Estimation

Parameter Estimators	$\widehat{\boldsymbol{\beta}}_j$	*P-value*	Decision
$\hat{\beta}_0$	10,512	0,532	Accept H_0
$\hat{\beta}_1$	0,138	0,438	Accept H_0
$\hat{\beta}_2$	-0,664	0,847	Accept H_0
$\hat{\beta}_3$	-0,037	0,843	Accept H_0
$\hat{\beta}_4$	-0,008	0,857	Accept H_0
$\hat{\beta}_5$	-0,518	0,605	Accept H_0
$\hat{\gamma}_0$	-11,650	0,419	Accept H_0
$\hat{\gamma}_1$	-0,113	0,330	Accept H_0
$\hat{\gamma}_2$	1,544	0,578	Accept H_0
$\hat{\gamma}_3$	0,085	0,452	Accept H_0
$\hat{\gamma}_4$	0,015	0,601	Accept H_0
$\hat{\gamma}_5$	4,150	0,578	Accept H_0
$\hat{\lambda}$	-37,316	<0,01*	Reject H_0

Based on Table 2, it can be seen that all parameters in ZIPIG regression using MLE have no significant effect on the number of DHF death cases.

4. Zero Inflation Poisson Invers Gaussian Regression (ZIPIG) with Bayesian Convergence Test with Trace Plot

The plot between the generated estimator and the iterations is called the trace plot. If the trace plot is random, then convergence is achieved. Iteration should be continued if convergence has not been achieved. Figure 1 shows the trace plot for the parameters β and γ.

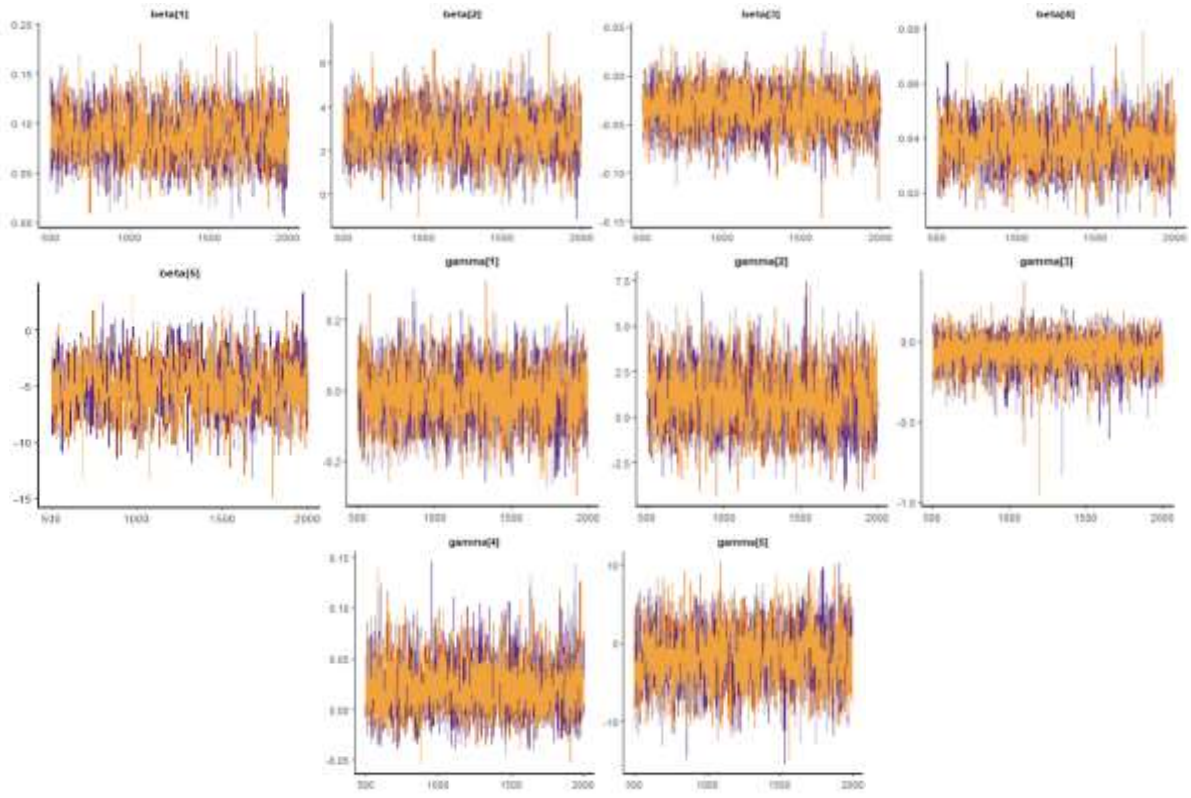

Figure 1. The Results of Trace Plot

Figure 1 shows that the trace plot is random or does not contain a trend when 2000 iterations are performed. It can be concluded that the parameters β and γ are convergent.

5. Convergence Test with MC Error

In addition to plotting, convergence checking can also be done by comparing the MC Error with 5% standard deviation of each parameter. Convergence is said to be achieved if the MC Error value is less than 5% standard deviation. MC Error on each parameter of the ZIPIG Bayesian Regression model in Table 3 as follows:

Table 3. MC Error for each Parameter of ZIPIG Bayesian Models

Parameter Estimators	Standard Deviation	5% Standard Deviation	MC Error	Decision
$\hat{\beta}_1$	0,02	0,001	0,00052	Convergent
$\hat{\beta}_2$	1,04	0,052	0,02381	Convergent
$\hat{\beta}_3$	0,02	0,001	0,00033	Convergent
$\hat{\beta}_4$	0,01	0,0005	0,00017	Convergent
$\hat{\beta}_5$	2,21	0,1105	0,05085	Convergent
$\hat{\gamma}_1$	0,08	0,004	0,00141	Convergent
$\hat{\gamma}_2$	1,61	0,081	0,03484	Convergent
$\hat{\gamma}_3$	0,10	0,005	0,00143	Convergent
$\hat{\gamma}_4$	0,02	0,001	0,00039	Convergent
$\hat{\gamma}_5$	3,53	0,176	0,07875	Convergent
$\hat{\lambda}$	13,72	0,686	0,18670	Convergent

Based on Table 3, the MC Error value for each parameter is less than 5% standard deviation. So it can be concluded that each parameter has converged or the generated sample comes from the desired posterior distribution.

6. Factors Influencing DHF Deaths

Credible intervals are used to test parameters based on the following hypothesis.

$H_0: \hat{\beta}_j = 0, H_1: \hat{\beta}_j \neq 0$,

$H_0: \hat{\gamma}_j = 0, H_1: \hat{\gamma}_j \neq 0$,

$H_0: \hat{\lambda} = 0, H_1: \hat{\lambda} \neq 0$

Where $\hat{\beta}_j, \hat{\gamma}_j$ and λ are ZIPIG model parameters. Credible intervals for each parameter are presented in table 4 below:

Table 4. Credible Interval

Parameter Estimators	Value of Parameter Estimators	Persentil 2,5%	Persentil 97,5%	Decision
$\hat{\beta}_1$	0,088	0,042	0,135	Accept H_0
$\hat{\beta}_2$	2,774	0,754	4,855	Reject H_0
$\hat{\beta}_3$	-0,035	-0,077	0,003	Accept H_0
$\hat{\beta}_4$	0,037	0,022	0,054	Reject H_0
$\hat{\beta}_5$	-4,888	-9,327	-0,593	Reject H_0
$\hat{\gamma}_1$	-0,013	-0,164	0,137	Accept H_0

Parameter Estimators	Value of Parameter Estimators	Persentil 2,5%	Persentil 97,5%	Decision
$\hat{\gamma}_2$	0,989	-1,995	4,168	Accept H_0
$\hat{\gamma}_3$	-0,065	-0,291	0,107	Accept H_0
$\hat{\gamma}_4$	0,022	-0,020	0,077	Accept H_0
$\hat{\gamma}_5$	-1,538	-8,689	5,169	Accept H_0
$\hat{\lambda}$	19,688	2,626	54,109	Reject H_0

Parameters are significant if the credible interval does not contain zeros in the 2.5% to 97.5% percentile interval. Based on Table 4.7, it can be seen that the factors that influence DHF are X_2= Population Density, X_4= Slum Area, and X_5= Number of Health Workers.

7. Interpretation

The ZIPIG regression model formed using the Bayesian method is as follows.

$$\text{logit}\left(\hat{\beta}\right) = 2{,}774X_2 + 0{,}037X_4 - 4{,}888X_5$$

a. If the population density increases by 1(person/km^2), the probability of DHF death increases by $\frac{1}{\exp(2{,}774)} = 0{,}0625$ or by 6.25%, assuming the values of slum area and number of other health workers are constant.
b. If the area of slums increases by 1 (ha), the probability of DHF deaths increases by $\frac{1}{\exp(0{,}037)} = 0{,}964$ or by 96.37% assuming the values of population density and the number of health workers are considered constant.
c. If the number of health workers increases by 1 person, the probability of DHF deaths decreases by $\frac{1}{\exp(-4{,}888)} = 0{,}00757$ or by 0.75%, assuming the values of population density and slum area are constant.

Based on the parameter estimation results of the ZIPIG model using the MLE method (Table 2), all variables showed no significant effect on the number of DHF deaths. However, through the Bayesian approach (Table 4), it was found that three variables namely population density, slum area, and number of health workers had a statistically significant effect. This shows the advantage of the Bayesian approach in detecting hidden patterns in data with small size and high proportion of zeros, and in producing more stable and informative parameter estimates.

8. Best Selection Models

To determine the goodness of the model between one method and another, Root Mean Square Error (RMSE) is used. The method that has the smallest RMSE is the best method. Here is the RMSE of each method, as shown in Table 5.

Table 5. RMSE Results of Each Model

Method	RMSE
ZIPIG MLE	28,132
ZIPIG Bayesian	24,259

Based on Table 5, it can be seen that the RMSE value in ZIPIG using MLE is greater than ZIPIG using Bayesian. This means that the best model in modelling the number of deaths due to DHF in Semarang is ZIPIG using Bayesian. The RMSE tends to increase at a higher proportion of zeros in the MLE method, while in the Bayesian method, the effect of the proportion of zeros is smaller than that of the MLE. This indicates that Bayesian is more robust to changes in the proportion of zeros and is superior in situations with small sample sizes or high proportions of zeros.

These results is similar to Utomo et al. (2025) study, in which the Bayesian method proved to be superior to maximum likelihood estimation in terms of RMSE values. In addition, (Al-Sharoot & Al-Badry, 2024; Shukla et al., 2017; Xu et al., 2025) also calculated the relative efficiency of estimators for the dataset and found that Bayesian estimators were better than MLE in all cases, particularly when dealing with small sample sizes or 10-20 samples.

D. CONCLUSION AND SUGGESTIONS

Based on the research results, it can be concluded that the Zero-Inflated Poisson Inverse Gaussian Regression (ZIPIGR) model with the Bayesian approach produces a lower RMSE value than the MLE approach, making it more accurate in modelling data with a high proportion of zeros and a small sample size. Data on the number of DHF deaths in Semarang City in 2024 has overdispersion and 62.5% of the data is zero, which indicates that the ZIPIG model is very appropriate to use. Based on the credible interval, there are three predictor variables that significantly affect the number of DHF, namely: Population density (positive), Area of slums (positive), and Number of health workers (negative). A limitation of this study is that it does not include spatial components, even though DHF cases are likely to have a distribution pattern influenced by location and proximity between regions. Therefore, further research is recommended to integrate spatial analysis, such as Bayesian Spatial ZIPIG, in order to improve the accuracy of predictions and interpretation of results. Furthermore, the government is advised to increase health workers and strengthen environmental control in densely populated areas and slums to reduce Dengue Hemorrhagic Fever (DHF) deaths.

ACKNOWLEDGEMENT

The author sincerely acknowledges to the supervisors and reviewers for their valuable contributions and constructive feedback, which greatly enhanced the quality of this research.

REFERENCES

Abdulhafedh, A. (2023). Incorporating Zero-Inflated Poisson (ZIP) Regression Model in Crash Frequency Analysis. *International Journal of Novel Research Interdisciplinary Studies*, *10*(1), 6–18. https://doi.org/10.5281/zenodo.7632596

Agresti, A. (2019). *An Introduction to Categorical Data Analysis* (3rd ed.). John Wiley & Sons, Inc. http://www.wiley.com/go/wsps

Akinkunmi, M. (2019). *Introduction To Statistics Using R* (1st ed.). Springer International Publisher. https://id.scribd.com/document/797441439/Introduction-to-Statistics-Using-r

Al-Sharoot, M. H., & Al-Badry, A. O. (2024). Estimating the survival function of three parameters Lindley distribution for patients with COVID-19 by using MLE and Bayesian estimates. *Warith Scientific Journal*, *6*(8), 1017–1030.

https://iasj.rdd.edu.iq/journals/uploads/2025/01/25/b16f1b3ec9f6162d7854421a793eb6b7.pdf

Amalia, R. N., Sadik, K., & Notodiputro, K. A. (2021). A Study of ZIP and ZINB Regression Modeling for Count Data with Excess Zeros. *Journal of Physics: Conference Series*, *1863*(1), 1–12. https://doi.org/10.1088/1742-6596/1863/1/012022

Aswi, A., Astuti, S. A., & Sudarmin, S. (2022). Evaluating the Performance of Zero-Inflated and Hurdle Poisson Models for Modeling Overdispersion in Count Data. *Inferensi*, *5*(1), 17–22. https://doi.org/10.12962/j27213862.v5i1.12422

Azizan, N. H., Mahmud, Z., Rambli, A., & Hafizah Azizan, N. (2020). Rasch Rating Scale Item Estimates using Maximum Likelihood Approach: Effects of Sample Size on the Accuracy and Bias of the Estimates. *International Journal of Advanced Science and Technology*, *29*(4s), 2526–2531. https://www.researchgate.net/publication/343152554

Bektashi, X., Rexhepi, S., & Limani–Bektashi, N. (2022). Dispersion of Count Data: A Case Study of Poisson Distribution and Its Limitations. *Asian Journal of Probability and Statistics*, *19*(2), 18–28. https://doi.org/10.9734/ajpas/2022/v19i230464

Bimali, M., Ounpraseuth, S. T., & Williams, D. K. (2021). Impact of magnitude of zero inflation of covariates on statistical inference and model selection. *Journal of Statistics Applications and Probability*, *10*(2), 287–292. https://doi.org/10.18576/jsap/100201

Chakraborty, S., & Biswas, S. C. (2024). Modelling Zero-Inflated Over Dispersed Dengue Data via Zero-Inflated Poisson Inverse Gaussian Regression Model: A Case Study of Bangladesh. *Acta Scientifica Malaysia*, *8*(1), 11–14. https://doi.org/10.26480/asm.01.2024.11.14

Handarzeni, S. A. (2022). Modeling of Tuberculosis Cases in Sumatra Region using Poisson Inverse Gaussian Regression. *JSDS: Journal of Statistics and Data Science*, *1*(2), 36–43. https://doi.org/10.2991/assehr.k.201010.007

Hespanhol, L., Vallio, C. S., Costa, L. M., & Saragiotto, B. T. (2019). Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian Journal of Physical Therapy*, *23*(4), 290–301. https://doi.org/10.1016/j.bjpt.2018.12.006

Hilbe, J. M. (2014). *Modeling Count Data* (1st ed.). Cambridge University Press. www.cambridge.org/9781107611252

Karlis, D., & Xekalaki, E. (2005). Mixed Poisson Distributions. *International Statistical Review*, *73*(1), 35–58. http://www.jstor.org/stable/25472639

Kemenkes RI. (2022). *Membuka Lembaran Baru Untuk Hidup Sejahtera*. Laporan Tahunan 2022 Demam Berdarah Dengue.

Kyriazos, T., & Poga, M. (2023). Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions. *Open Journal of Statistics*, *13*(3), 404–424. https://doi.org/10.4236/ojs.2023.133020

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14. https://doi.org/10.2307/1269547

Lijoi, A., Mena, R. H., & Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, *100*(472), 1278–1291. https://doi.org/10.1198/016214505000000132

Liu, H., & Powers, D. A. (2012). Bayesian Inference for zero inflated poisson regression models. *Journal of Statistics : Advance in Theory and Aplications*, *7*(2), 155–188. https://www.researchgate.net/publication/265778813_Bayesian_inference_for_zero-inflated_Poisson_regression_models

Payne, E. H., Hardin, J. W., Egede, L. E., Ramakrishnan, V., Selassie, A., & Gebregziabher, M. (2017). Approaches for dealing with various sources of overdispersion in modeling count data: Scale adjustment versus modeling. *Statistical Methods in Medical Research*, *26*(4), 1802–1823. https://doi.org/10.1177/0962280215588569

Psutka, J. V., & Psutka, J. (2019). Sample size for maximum-likelihood estimates of Gaussian model depending on dimensionality of pattern space. *Pattern Recognition*, *91*(2), 25–33. https://doi.org/10.1016/j.patcog.2019.01.046

Purhadi, & Ermawati. (2021). Bivariate Zero-Inflated Poisson Inverse Gaussian Regression Model and Its Application. *International Journal on Advanced Science Engineering Information Technology*, *11*(6), 2407–2415. https://doi.org/10.18517/ijaseit.11.6.14217

Putri, G. N., Nurrohmah, S., & Fithriani, I. (2020). Comparing Poisson-Inverse Gaussian Model and Negative Binomial Model on case study: Horseshoe crabs data. *Basic and Applied Sciences Interdisciplinary Conference 2017*, *1442*(1). https://doi.org/10.1088/1742-6596/1442/1/012028

Rahayu, L. P., Sadik, K., & Indahwati. (2016). Overdispersion study of poisson and zero-inflated poisson regression for some characteristics of the data on lamda, n, p. *International Journal of Advances in Intelligent Informatics*, *2*(3), 140–148. https://doi.org/10.26555/ijain.v2i3.73

Shukla, G., Kumar, V., Shukla, S., & Giri Goswami, M. (2017). Comparison between Bayesian and Maximum Likelihood Estimation of Scale Parameter in Generalized Gamma Type Distribution with Known Shape Parameters under Different Loss Functions. *International Journal on Emerging Technologies (Special Issue NCETST-2017)*, *8*(1), 288–294. www.researchtrend.net

Weni Utomo, C. R., Efendi, A., & Wardhani, N. W. S. (2025). Simulation Study of Bayesian Zero Inflated Poisson Regression. *CAUCHY: Jurnal Matematika Murni Dan Aplikasi*, *10*(1), 213–223. https://doi.org/10.18860/ca.v10i1.30207

Xu, A., Wang, B., Zhu, D., Pang, J., & Lian, X. (2025). Bayesian Reliability Assessment of Permanent Magnet Brake Under Small Sample Size. *IEEE Transactions on Reliability*, *74*(1), 2107–2117. https://doi.org/10.1109/TR.2024.3381072

Zha, L., Lord, D., & Zou, Y. (2016). The Poisson inverse Gaussian (PIG) generalized linear regression model for analyzing motor vehicle crash data. *Journal of Transportation Safety and Security*, *8*(1), 18–35. https://doi.org/10.1080/19439962.2014.977502