Convergence of Mathematics and Statistics
Volume 01, Issue 1, March 2026, pages 01-12

DOI: https://doi.org/xx.xxxxx/xxxxxxxxx-xx
E-ISSN: xxxx-xxxxx

Open Access - Research Article

Overdispersion Data Analysis Using Gaussian Inverse Poisson Regression Model

Marsono [1]*, **M. Y. Matdoan** [2]

[1]Badan Pusat Statistik, Mamuju, Sulawesi Barat, Indonesia
[2]Program Studi Statistika, Fakultas Sains dan Teknologi, Universitas Pattimura, Ambon, Indonesia

Article Info

Article history:

Received: October 15th, 2025
Revised: November 25th, 2025
Accepted: December 18th, 2025
Published: January 18th, 2026

Published by:

ABSTRACT

One of the violations of assumptions in regression models in the calculation data is the case of overdispersion, where the value of variance is greater than the mean of the response variable. To overcome cases of overdispersion, a regression model is formed by mixing the Poisson distribution with several other distributions. Poisson Inverse Gaussian (PIG) is one of the regression models formed by mixed models designed for overdispersion data. The purpose of this study is to analyze the variables that affect the infant mortality rate in East Java Province in 2019. The model parameter estimator used is the Maximum Likelihood Estimator (MLE). From the minimum AIC value achieved, it can be seen that the Gaussian Inverse Poisson regression model is better than the Poisson regression model

Keywords: *Infant Mortality Rate, Overdispersion, Poisson Gaussian inverse, Poisson Model*

Corresponding Author: *marsono@bps.go..id*

1. Introduction

To model the relationship between one or more dependent variables or responses and one or more independent variables or predictors, classical linear regression or *Ordinary Least Square* (OLS) regression models are generally used[1]. In OLS regression there are several assumptions that must be met. In the reality in the field, not all types of data can meet these assumptions, for example in the case of counting data. Modelling of enumeration data using OLS regression cannot be done because it violates two assumptions required in OLS regression, namely error following a normal distribution (normality) and variance must be constant (homoscedasticity). Developments in data modelling gave rise to modelling for enumerated data with *Generalized Linear Models* (GLMs). GMLs are generalizations of the classical normal regression model or OLS regression of a rigorous range of assumptions and provide an analytical method for abnormal data [2][3]. Poisson regression is one of the family members of GLMs derived from the poisson distribution. The poisson distribution is a discrete distribution with a random variable value in the form of a positive integer so that it is a good choice for modeling enumeration data. The poisson distribution is determined by only one parameter that defines both *the mean* and the variance of the distribution, so in Poisson regression there is an assumption that must be fulfilled, namely that the *mean* and variance of the response variables must be the same (*equidispersion*)[4].

However, in reality, there is often a violation of this assumption where the variance is smaller than *the mean* (*underdispersion*) or the variance is greater than *the mean* (*overdispersion*). In most data *counts*, cases of overdispersion are sometimes found [5][6]. In practice, enumerated data often show considerable variance because many contain extra

zeros or a larger distribution of the values in the data or both [7][8]. Data with a large number of zeros can be caused by structural *zeros* or by zero samples (*sampling zeros*). Furthermore, Hu, Pavlicova and Nunes gave examples of the number of high-risk sexual behaviors in comparing the two. A zero value in the number of high-risk sexual behaviors can be caused by a person who does not have a partner so that they have never had sexual intercourse, which is called structural *zeros* or it can be because someone who already has a partner and has never had high-risk sexual intercourse, which is called zero sampling *zeros*[9].

The case of overdispersion if ignored can result in *underestimation* of standard errors, so that it can result in errors in decision-making in some hypothesis tests, for example, a predictor variable has a significant effect when in reality it has no significant effect [8]. In overcoming the case of overdispersion, several modelling was formed which are a combination of Poisson distributions with several distributions, both discrete and continuous (*mixed Poisson distribution*). *Mixed poisson distributions* are an alternative solution to overdispersion cases, but only a few distributions are often used in studies due to their complex calculations. One of them is the *Poisson Inverse Gaussian* (PIG) distribution which is a *mixed poisson distribution* with a random effect that has an Inverse Gaussian distribution. This distribution was first introduced by Holla in 1966 [10]. The PIG distribution itself is a form of *Siche (SI)* distribution with two parameters. SI is referred to as a better model than the negative binomial model, especially for data that is highly overdispersed and tends to be highly *skewed to the right*. However, the calculation is more complicated because it has three parameters on the probability density function. As a form of SI distribution, the PIG distribution is used in modelling counting data that is right-handed and has a slightly longer tail. The PIG distribution has a *closed form* of likelihood function and is easier to calculate, so many studies involving enumeration data use this model [11][12].

Willmot [12] demonstrated the potential of modelling with gaussian inverse poisson regression as an alternative to negative binomial regression on car insurance claims data. Six sets of car insurance claims data were presented with characteristics of nearly 80 percent of the data containing zero and resulted in the conclusion that modelling with PIG regression was a better model than the negative binomial regression model. Another study that uses the PIG regression model is a study by Shoukri, Asyali and Vandorp [13] which used the model on the number of mastitis cases in dairy cows in Ontario Canada and came to the conclusion that PIG regression is a better model than the negative binomial regression model. Furthermore, in some road safety studies such as accident data modelling and motor insurance research, the PIG regression model is often used as an alternative to the negative binomial regression model [13][14]. Modelling with PIG regression was also used in the study of Zha, Lord and Zou [15] in the case of the number of motorcycle accidents that occurred in two different places, namely in Texas and Washington. In the study, the data on the number of right-hand motorcycle accidents with a slightly long tail and 37 percent of the data contained a zero value. With smaller *Akaike Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC) values, it is shown that modelling with PIG regression is better for modelling the number of motorcycle accidents in the two places.

The number of cases of infant deaths in an area is a form of enumerated data, so that in the modelling it can use Poisson regression. However, infant mortality data also have the potential to be over dispersed, so the modelling is not enough to use Poisson regression. In this study, modelling will be compared using Poisson regression and Gaussian Inverse Poisson regression to get the best model of the variables that affect the number of infant stills in East Java in 2019. The data used is secondary data from the health profile of East Java province in 2019.

2. Method Details

2.1. Fish Distribution

The Poisson distribution is a discrete probability function distribution that expresses the number of events that occur in a given period of time if the average of the events is known and in a time that is mutually free from the last event. Poisson distributions can also be used for the number of events at a given interval such as distance, area, or volume. The poisson distribution is a benchmark model for enumeration data [16]. The probability function for Poisson distributed data depends on a single parameter, i.e. the average (μ) with the probability distribution given by Equation (1).

$$f(y;\mu)=\frac{e^{-\mu}\mu^{y}}{y!};\text{for } y=0,1,2,\ldots \text{ and } \mu>0 \tag{1}$$

In Poisson distributions, the mean and variance are of the same value and can be written as Equation (2).

$$E(Y)=\text{var}(Y)=\mu \tag{2}$$

2.2. Poisson Regression

Poisson regression is a non-linear *regression model* that is often used to analysed discrete data. Poisson regression refers to the use of the Poisson distribution [17]. In Poisson regression there is an assumption, namely the Poisson distributed response variable (y) and there is no multicollinearity among the predictor variables (x) (Hinde in Astuti, 2013). Furthermore, the Poisson regression model with log linkage is given by equation (3).

$$y_i = \mu_i = \exp\left(\tilde{x}_i^T \tilde{\beta}\right) = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right) \tag{3}$$

where

y_i : Variable response to- $I = 1, 2, , \cdots n$.

μ_i : the average number of events that occurred in a given period of time.

x_i : variable predictor ke-i, $I = 1, 2, , \cdots n$.

β : parameters in the Poisson regression model.

2.3. Parameter Estimation Poisson Regression Model

Estimation of Poisson regression parameters was carried out using *the Maximum Likelihood* Estimation (MLE) method by maximizing the *likelihood* function of the Poisson regression model [18]. Given *the likelihood* function of the Poisson regression model as equation (4).

$$\begin{aligned} L\left(y \mid \tilde{\beta}\right) &= \prod_{i=1}^{n} \mu_i^{y_i} \frac{\exp\left(-\mu_i\right)}{y_i!} \\ &= \prod_{i=1}^{n} \exp\left(\tilde{x}_i^T \tilde{\beta}\right) \frac{\exp\left(-\tilde{x}_i^T \tilde{\beta}\right)}{y_i!} \end{aligned} \tag{4}$$

Next, the likelihood of equation (4) is obtained.

$$\begin{aligned} \ln L\left(y \mid \tilde{\beta}\right) &= \ln\left(\prod_{i=1}^{n} \mu_i^{y_i} \frac{\exp\left(-\mu_i\right)}{y_i!}\right) \\ &= \sum_{i=1}^{n} y_i \tilde{x}_i^T \tilde{\beta} - \sum_{i=1}^{n} \exp\left(\tilde{x}_i^T \tilde{\beta}\right) - \sum_{i=1}^{n} \ln y_i! \end{aligned} \tag{5}$$

Equation (5) is then derived from and $\tilde{\beta}$ equalized to zero as equation (6).

$$\frac{\partial L\left(y \mid \tilde{\beta}\right)}{\partial \tilde{\beta}} = \frac{\partial\left(\sum_{i=1}^{n} y_i \tilde{x}_i^T \tilde{\beta} - \sum_{i=1}^{n} \exp\left(\tilde{x}_i^T \tilde{\beta}\right) - \sum_{i=1}^{n} \ln y_i!\right)}{\partial \tilde{\beta}} 0 \tag{6}$$

The results of the parameter estimation produced in equation (6) are not *closeform* so it is necessary to perform numerical iterations using Newton-Raphson iteration. The goal of the numerical iteration method is to maximize the function of probability (Myers, 1990).

2.4. Inverse Gaussian Distribution

The Inverse Gaussian distribution is a continuous distribution with a density function similar to the gamma distribution but with greater emptiness and sharp tapering. The Gaussian inverse has two parameters and an opportunity density function that can be written as equations [19].

$$f\left(y\right) = \left(2\pi y^3 \sigma\right)^{-0.5} e^{-(y-\mu)^2 / 2y\mu^2\sigma^2}, \quad y > 0 \tag{7}$$

where $E\left(Y\right) = \mathrm{var}\left(Y\right) = \sigma^2 \mu^3$ and σ^2 is the dispersion parameter.

The gaussian inverse is used in conditions with extreme stiffness. The name inverse gaussian itself comes from the cumulative function which has an *inverse relationship* with the cumulative function (natural logarithm of the MGF function) normal distribution/gaussian distribution [20].

2.5. Overdispersion

In poisson regression there is an assumption that must be fulfilled, namely that the mean and variant of the response variable must be the same *(equidispersion)*. Overdispersion in poisson regression occurs when the variance of the response variable is greater than the mean. One way that can be done to detect the presence or absence of overdispersion in the response variable to be studied is with the statistics of *the chi-square Pearson t*-test show *n* by equation (8).

$$VT = \sum_{i=1}^{n} \frac{(y-\bar{y})^2}{\bar{y}} = (n-1)\frac{S^2}{\bar{y}} \tag{8}$$

The value VT in equation (8) is equal to the *variance-to-mean ratio*, which is often referred to as the disperse index, multiplied by $n-1$, which n is the sample size. If the value of the disperse index is less than 1, it can be said that underdispersion occurs, on the other hand, overdispersion occurs when the value of the disperse index is more than 1 [7]. The statistical test that can be used to detect overdispersion in a data set is the overdispersion test so in this test the hypothesis used is.

$H_0 : \text{var}(y) = \mu$

$H_1 : \text{var}(y) = \mu_i + a.g(.)$,

where $g(.)$ it is a certain function. In simple terms, if the value $a = 0$ can be said to be equidispersal, on the other hand, if $a > 0$ it can be said to be overdispersion. The value of the coefficient a can be estimated by *Ordinary Least Square* (OLS) regression with a statistical test used as equation (9)

$$T = \frac{1}{2}\sum_{i=1}^{n}\left(\left(Y_i - \bar{\mu}_i\right)^2 - Y_i\right) \tag{9}$$

The distribution that is asymptotic to T is the standard normal distribution under the zero hypothesis [5].

2.6. Gaussian Inverse Poisson distribution

The Gaussian Inverse Poisson distribution (PIG) is one of the *mixed* poisson distributions. The shape of *the mixed* poisson distribution depends on the distribution of the *random* effect (v). Suppose (g v) is a function of the probability density of and v the marginal distribution for Y obtained by the integral *vij*.

$$P(Y = y \mid \mu) = \int f(y \mid \mu, v)\, g(v)\, dv \tag{10}$$

where, $f(y \mid \mu, v) = e^{-v\mu}\frac{(\mu v)^y}{y!}$ and (g v) is given in equation (11).

For the PIG distribution, v in equation (10) is assumed to follow the inverse gaussian distribution and has an opportunity density function that can be written as equation (11).

$$g(v) = \left(2\pi v^3\right)^{-0.5} e^{-(v-1)^2/2\tau v}, \; v > 0 \tag{11}$$

where, $\tau = \text{var}(v), E(v) = 1$.

The PIG distribution is denoted by $PIG(\mu, \tau)$ that given by the equation (12).

$$P(y \mid \mu, \sigma) = \left(\frac{2z}{\pi}\right)^{\frac{1}{2}} \frac{\mu^y e^{\frac{1}{\tau}} K_s(z)}{(z\tau)^y\, y!} \tag{12}$$

where

$K_s(z) = K_{y-\frac{1}{2}}\left(\frac{1}{\tau}\sqrt{(2\mu\tau+1)}\right)$, $s = y-\frac{1}{2}$, and $z = \sqrt{\frac{1}{\tau^2}+\frac{2\mu}{\tau}}$.

$K_s(z)$ is the third type of modified Bessel function (Willmott, 1987).

The PIG distribution is determined by two parameters, namely the mean () as the location parameter and the dispersion parameter () as the shape parameter. $\mu\tau$

2.7. Inverse Gaussian Poisson regression

Inverse Gaussian Poisson regression is a non-linear *regression model* that is often used as an alternative to Poisson regression when the assumption of a variance value is greater than the mean value or what is called overdispersion. Given the Poisson Inverse Gaussian regression model as equation (13) follows (Purnamasari, 2016).

$$y_i = \mu_i = \exp\left(\tilde{x}_i^T\tilde{\beta}\right) = \exp\left(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}\right) \tag{13}$$

where, $y_i \sim PIG(\mu_i,\tau)$ and $\mu_i = \exp\left(\tilde{x}_i{}^T\tilde{\beta}\right)$ or $\ln(\mu_i) = \tilde{x}_i{}^T\tilde{\beta}$. With $\tilde{x}_i^T = \begin{bmatrix}1 & x_{i1} & x_{i2} & \cdots & x_{ip}\end{bmatrix}$ and $\tilde{\beta} = \begin{bmatrix}\beta_0 & \beta_1 & \beta_2 & \cdots & \beta_p\end{bmatrix}^T$, where $i = 1,2,\ldots,n$.

With the function of opportunity density as equation (14)

$$P(Y=y\mid x_i,\beta,\tau) = \left\{\frac{e^{x_i^T\beta y_i}e^{\frac{1}{\tau}}}{y_i!}\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}\left(2e^{x_i^T\beta}\tau+1\right)^{-\frac{\left(y_i-\frac{1}{2}\right)}{2}}K_{si}(z_i)\right\} \tag{14}$$

2.8. Estimation of PIG Regression Parameters

The estimation of PIG regression parameters is carried out using *the Maximum Likelihood* Estimation (MLE) method by maximizing the *likelihood* function of the PIG regression model. Given *the likelihood* function of the PIG regression model as equation (15).

$$L\left(y\mid\tilde{\beta},\tau\right) = \prod_{i=1}^{n}\left\{\frac{e^{\tilde{x}_i^T\tilde{\beta}y_i}e^{\frac{1}{\tau}}}{y_i!}\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}\left(2e^{\tilde{x}_i^T\tilde{\beta}}\tau+1\right)^{-\frac{\left(y_i-\frac{1}{2}\right)}{2}}K_{si}(z_i)\right\} \tag{15}$$

Next, the likelihood of the equation (15) is obtained.

$$\begin{aligned}\ln L\left(y\mid\tilde{\beta},\tau\right) &= \ln\left(\prod_{i=1}^{n}\frac{e^{\tilde{x}_i^T\tilde{\beta}y_i}e^{\frac{1}{\tau}}}{y_i!}\left(\frac{2}{\pi\tau}\right)^{\frac{1}{2}}\left(2e^{\tilde{x}_i^T\tilde{\beta}}\tau+1\right)^{-\frac{\left(y_i-\frac{1}{2}\right)}{2}}K_{si}(z_i)\right)\\ &= \sum_{i=1}^{n}y_i\tilde{x}_i^T\tilde{\beta}+\frac{n}{\tau}-\ln\left(\sum_{i=1}^{n}y_i!\right)+\frac{n}{2}\ln\left(\frac{2}{\pi}\right)-\frac{n}{2}\ln\tau\\ &\quad-\sum_{i=1}^{n}\left(\frac{2y_i-1}{4}\right)\ln\left(2\tilde{x}_i^T\tilde{\beta}+1\right)+\sum_{i=1}^{n}\ln K_{si}(z_i)\end{aligned} \tag{16}$$

Equation (16) is then derived against $\tilde{\beta}$ and τ then equalized to zero as equation (17).

$$\frac{\partial L\left(y\mid\tilde{\beta}\right)}{\partial\tilde{\beta}\partial\tau} = \frac{\partial\left(\sum_{i=1}^{n}y_i\tilde{x}_i^T\tilde{\beta}-\sum_{i=1}^{n}\exp\left(\tilde{x}_i^T\tilde{\beta}\right)-\sum_{i=1}^{n}\ln y_i!\right)}{\partial\tilde{\beta}\partial\tau}0 \tag{17}$$

The results of the parameter estimation produced in equation (17) are not *closeform* so it is necessary to perform numerical iteration using Newton-Raphson iteration. The purpose of the numerical iteration method is to maximize the probability function [8]. The Newton-Raphson algorithm can be done with the following steps.

1. Determine the initial estimate of the parameter $\hat{\tilde{\beta}}_{(0)}$ using *the Ordinary Least Square* (OLS) method as equation (18).

$$\hat{\tilde{\beta}}_{(0)} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}\tilde{y} \tag{18}$$

2. Define the gradient vector $\tilde{g}$ as equation (19).

$$\tilde{g}^T\left(\tilde{\beta}_{(m)}\right)_{(p+1)1} = \left(\frac{\partial L\left(y\mid\tilde{\beta}\right)}{\partial\beta_0}, \frac{\partial L\left(y\mid\tilde{\beta}\right)}{\partial\beta_1}, \ldots, \frac{\partial L\left(y\mid\tilde{\beta}\right)}{\partial\beta_p}\right)_{\tilde{\beta}=\tilde{\beta}_{(m)}} \tag{19}$$

3. Determine the Hessian matrix $\mathbf{H}$ as equation (20).

$$\mathbf{H}\left(\tilde{\beta}_{(m)}\right)_{(p+1)(p+1)} = \begin{bmatrix} \frac{\partial^2 L\left(y\mid\tilde{\beta}\right)}{\partial\beta_0^2} & \frac{\partial^2 L\left(y\mid\tilde{\beta}\right)}{\partial\beta_0\partial\beta_1} & \cdots & \frac{\partial^2 L\left(y\mid\tilde{\beta}\right)}{\partial\beta_0\partial\beta_p} \\ & \frac{\partial^2 L\left(y\mid\tilde{\beta}\right)}{\partial\beta_1^2} & \cdots & \frac{\partial^2 L\left(y\mid\tilde{\beta}\right)}{\partial\beta_1\partial\beta_p} \\ & \vdots & \ddots & \vdots \\ \text{Simetri} & & & \frac{\partial^2 L\left(y\mid\tilde{\beta}\right)}{\partial\beta_p^2} \end{bmatrix} \tag{20}$$

4. Enter values $\hat{\tilde{\beta}}_{(0)}$ into vector elements $\tilde{g}$ and matrix $\mathbf{H}$ so that obtained $\tilde{g}\hat{\tilde{\beta}}_{(0)}$ and matrix $\mathbf{H}\hat{\tilde{\beta}}_{(0)}$.
5. Starting from $m=0$ iteration on equations (21).

$$\tilde{\beta}_{(m+1)} = \tilde{\beta}_{(m)} - \mathbf{H}^{-1}\left(\tilde{\beta}_{(m)}\right)\tilde{g}\left(\tilde{\beta}_{(m)}\right) \tag{21}$$

6. The optimization of the converged parameter is obtained if $\left\|\tilde{\beta}_{(m+1)} - \tilde{\beta}_{(m)}\right\| \le \varepsilon$, if the convergent estimate has not been obtained, then step 5 is resumed until the iteration to $m = m+1$.

2.9. PIG Regression Model Parameter Testing

To determine the suitability of the model formed, it is necessary to test the parameters on the PIG regression model. Parameter testing is carried out with two events, namely simultaneous and partial parameter testing [13].

1. Simultaneous parameter testing
   The hypotheses used in simultaneous parameter testing are given as follows.
   $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$
   H_1 : at least one $\beta_l \neq 0$ with $j = 1, 2, \ldots, p$
   The test statistics used are Devian test statistics or commonly referred to as test statistics G formulated by equations (22).

$$\begin{aligned} G &= -2\ln\left(\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right) \\ &= 2\left(\ln L(\hat{\omega}) - \ln L(\hat{\Omega})\right) \end{aligned} \tag{22}$$

   where

$$L\left(\hat{\Omega}\right)=\prod_{i=1}^{n}\left(\frac{\left(e^{\tilde{x}_i^T\tilde{\beta}}\right)^{y_i}e^{\frac{1}{\hat{\tau}}}}{y_i!}\left(\frac{2}{\pi\hat{\tau}}\right)^{\frac{1}{2}}\left(2\left(e^{\tilde{x}_i^T\tilde{\beta}}\right)\hat{\tau}+1\right)^{-\left(\frac{y_i-\frac{1}{2}}{2}\right)}K_{S_i}\left(z_i\right)\right)$$

$$\ln L\left(\hat{\Omega}\right)=\sum_{i=1}^{n}y_i\tilde{x}_i^T\tilde{\beta}+\frac{n}{\hat{\tau}}-\ln\left(\sum_{i=1}^{n}y_i!\right)+\frac{n}{2}\ln\left(\frac{n}{\pi}\right)-\frac{n}{2}\ln\hat{\tau}-\sum_{i=1}^{n}\frac{2y_i-1}{4}\ln\left(2\tilde{x}_i^T\tilde{\beta}\hat{\tau}+1\right)+\sum_{i=1}^{n}\ln K_{S_i}\left(z_i\right)$$

$$L\left(\hat{\omega}\right)=\prod_{i=1}^{n}\left(\frac{\left(e^{\beta_0}\right)^{y_i}e^{\frac{1}{\hat{\tau}_\omega}}}{y_i!}\left(\frac{2}{\pi\hat{\tau}_\omega}\right)^{\frac{1}{2}}\left(2\left(e^{\beta_0}\right)\hat{\tau}_\omega+1\right)^{-\left(\frac{y_i-\frac{1}{2}}{2}\right)}K_{S_i}\left(z_i\right)\right)$$

$$\ln L\left(\hat{\omega}\right)=\sum_{i=1}^{n}y_i\beta_0+\frac{n}{\hat{\tau}_\omega}-\ln\left(\sum_{i=1}^{n}y_i!\right)+\frac{n}{2}\ln\left(\frac{n}{\pi}\right)-\frac{n}{2}\ln\hat{\tau}_\omega-\sum_{i=1}^{n}\frac{2y_i-1}{4}\ln\left(2\beta_0\hat{\tau}_\omega+1\right)+\sum_{i=1}^{n}\ln K_{S_i}\left(z_i\right)$$

This test statistic G follows the *Chi–Square distribution*, because the test G compares the observed frequency $L\left(\hat{\omega}\right)$ with the expected frequency $L\left(\hat{\Omega}\right)$ [16]. Therefore, it is necessary to approach the *Chi-Square* distribution if it is close to infinity, with a free degreenq. $H_o$will be rejected at the level of significance α if the value $G \geq \chi^2_{(q;\alpha)}$ or $p-value < \alpha$, with the conclusion that there is at least one coefficient . $\beta_j \neq 0$

2. Partial parameter testing

Hypothesis testing is partially performed for both parameters, i.e. and with the hypotheses used in the partial parameter testing given as follows.$\beta_j\tau$

$H_0:\beta_j=0$ and $H_1:\beta_j\neq 0 \quad j=1,2,....,p$.

$H_0:\tau=0$ and $H_1:\tau\neq 0$.

The test statistics used are t-tests formulated by equations (23) for parameters and parameters.$\beta_j\tau$

$$t_{hitung}=\frac{\hat{\beta}_j}{se\left(\hat{\beta}_j\right)} \text{ and } t_{hitung}=\frac{\tau}{se\left(\tau\right)} \tag{23}$$

where $se(\hat{\beta}_j)$ is the *standard error* of $\hat{\beta}_j$ with, $se\left(\hat{\beta}_j\right)=\sqrt{\text{var}\left(\hat{\beta}_j\right)}$ and where $se(\hat{\tau})$ is the *standard error* of $\hat{\tau}$ with, $se\left(\hat{\tau}\right)=\sqrt{\text{var}\left(\hat{\tau}\right)}$. An area of rejection H_0 is $\left|t_{hitung}\right| > t_{\frac{\alpha}{2},(n-p-1)}$ or minus H_0 if $p-value$ it is less than the level of significance.

2.10. Data Source

The data used in this study is secondary data from the health profile of East Java province in 2019. The number of samples used was 38 districts/cities in East Java province. With 5 predictor variables used. The following is given a definition of each response variable and predictor.

Table 1. Response Variables and Predictors

Variabel	Information
y	Total Infant Deaths
x_1	Percentage of Visits by Pregnant Women with K1
x_2	Percentage of Visits for Pregnant Women with K4
x_3	Percentage of Births by Health Workers
x_4	Percentage of Pregnant Women Getting Blood Supplement Tablets (TTD)
x_5	Percentage of Obstetric Complications Treated

2.11. Analysis Steps

The steps to determine the factors that affect the number of infant deaths in East Java Province in 2019 with the Poisson Inverse Gaussian regression model are as follows.

1. Conduct an examination of multicollinearity cases using the VIF criteria.
2. Perform modeling which includes parameter estimation and testing of significant parameters for Poisson regression models and Poisson Inverse Gaussian regression.
3. Perform an overdispersion test
4. Modeling using Poisson Inverse Gaussian regression which includes parameter estimation and hypothesis testing simultaneously and partially.
5. Compare AIC values to find the best model.
6. Interpret the obtained Poisson Inverse Gaussian regression model.
7. Draw conclusions from the results of the analysis.

3. Results and Discussion

Poisson Inverse Gaussian regression (PIG) is a regression that can be applied to data *count* that is overdispersed. In this study, PIG regression modelling was applied to data on the number of infant deaths in East Java province in 2019. Before performing PIG regression modelling, it is necessary to perform a multicollinearity test. The multicollinearity test aims to find out whether the predictor variables are linearly related to each other (collinearity). The criteria used to check the collinearity between the predictor variables is by looking at the *value of the Variance Inflation Factors* (VIF) in the predictor variables. The following are the VIF values for each predictor variable.

Table 2. VIF value.

VIF Value				
x_1	x_2	x_3	x_4	x_5
2.619	2.526	2.910	1.601	1.925

The VIF value of each predictor variable in Table 2 shows that the value is less than 10, so that there is no collinearity between the predictor variables. Because there is no collinearity between predictor variables, predictor variables can be included in the model.

3.1. Estimation of Poisson Regression Model Parameters and Inverse Gaussian Poisson Regression

In this study, data on the number of infant deaths in East Java province in 2019 will be modeled using two methods, namely Poisson regression as a comparison when cases of overdispersion are ignored and PIG regression for cases of overdispersion. Based on the results of the data analysis that has been carried out, the following are the estimated results for each model shown by Table 3 and Table 4.

Table 3. Parameter Estimation in a Poisson Regression Model

Parameter	Parameter Values	Standard *Error*	Value t	*P-Value*	Results
$\hat{\beta}_0$	2.267	0.456	4.575	**0.000**	Signifikan
$\hat{\beta}_1$	0.019	0.007	2.696	**0.007**	Signifikan
$\hat{\beta}_2$	-0.061	0.005	-12.541	**0.000**	Signifikan
$\hat{\beta}_3$	0.048	0.007	6.714	**0.000**	Signifikan
$\hat{\beta}_4$	0.007	0.003	1.995	**0.046**	Signifikan
$\hat{\beta}_5$	0.004	0.002	2.549	**0.011**	Signifikan

Table 4. Parameter Estimation in the PIG Regression Model

Parameter	Parameter Values	Standard *Error*	Value t	*P-Value*	Results
$\hat{\beta}_0$	2.451	3.014	0.813	0.422	Insignificant
$\hat{\beta}_1$	0.049	0.042	1.164	0.253	Insignificant
$\hat{\beta}_2$	-0.078	0.029	-2.704	**0.011**	Signifikan
$\hat{\beta}_3$	0.045	0.043	1.030	0.311	Insignificant
$\hat{\beta}_4$	-0.002	0.017	-0.103	0.919	Insignificant
$\hat{\beta}_5$	-0.001	0.009	-0.109	0.914	Insignificant
$\hat{\tau}$	-0.741	0.281	-2.636	**0.013**	Signifikan

Based on the results of parameter estimation obtained in Table 3 and Table 4, the results of the model estimation for the poisson regression model and PIG regression are obtained as equations (24) and (25).

$$\hat{y} = \exp\left(2.267 + 0.019x_1 - 0.061x_2 + 0.048x_3 + 0.007x_4 + 0.004x_5\right) \tag{24}$$

$$\hat{y} = \exp\left(2.451 + 0.049x_1 - 0.078x_2 + 0.045x_3 - 0.002x_4 - 0.001x_5\right) \tag{25}$$

The AIC value generated for the poisson regression model was 1050 while the AIC value generated for the PIG regression model was 401.68. The best model obtained when viewed from the smallest AIC value is the PIG regression model, but in the PIG regression model the significant parameters are only 2 parameters. Meanwhile, in the poisson regression model, all parameters in the model are significant. In this study, the model to be chosen is the model with the smallest AIC value, namely the PIG regression model.

3.2. Overdispersion Case Examination

An examination of overdispersion cases was carried out to see whether the data on the number of infant deaths in East Java province in 2019 was overdispersed or not. The following are given hypotheses and test statistics used.

$H_0 : \text{var}(y) = \mu$

$H_1 : \text{var}(y) = \mu + \alpha.g(.)$.

Table 5. Overdispersion Test

Value α	Value z	*p-value*	VT
35.73	2.53	0.006	1395.751

By using *the AER* package in the R *software*, a value $\alpha = 35,73$ with a *p-value* of 0.006 is obtained less than the 5% significance level so that it H_0 is rejected. It can be concluded that there is an overdispersion in the data on the number of infant deaths in East Java province in 2019. In another way, it is to look at the VT value obtained based on equation (8) which is greater than 1. So that the right method to model the number of infant deaths in East Java province in 2019 is Poisson Inverse Gaussian regression (PIG).

3.3. PIG Regression Model Parameter Estimation

Furthermore, parameter estimation will be carried out using the *backward* method for the PIG regression model to obtain the best model results. The following are some possible results of the model in PIG regression shown by Table 6.

Table 6. Estimation of Probability Parameters in the PIG Regression Model

Model	Variables in the Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\tau}$
M_1	x_1, x_2, x_3, x_4, x_5	2.451	0.049	-0.078	0.045	-0.002	-0.001	-0.741
M_2	x_2, x_3, x_4, x_5	400.208	-	-0.083	0.076	0.000	0.003	-0.704
M_3	x_1, x_3, x_4, x_5	0.676	0.051	-	-0.008	-0.022	0.013	-0.514

Model	Variables in the Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\tau}$
M_4	x_1, x_2, x_4, x_5	2.903	0.073	-0.066	-	0.000	-0.000	-0.697
M_5	x_1, x_2, x_3, x_5	2.774	0.044	-0.084	0.050	-	-0.001	-0.733
M_6	x_1, x_2, x_3, x_4	2.475	0.047	-0.076	0.044	-0.002	-	-0.741
M_7	x_1, x_2, x_3	2.747	0.042	-0.081	0.049	-	-	-0.735
M_8	x_1, x_2	2.965	0.073	-0.065	-	-	-	-0.697

Based on the results of parameter estimation in Table 6, PIG regression model estimates were obtained for several of them.

$M_1 = \hat{y} = \exp(2.451 + 0.049x_1 - 0.078x_2 + 0.045x_3 - 0.002x_4 - 0.001x_5)$,

$M_2 = \hat{y} = \exp(400.208 + -0.083x_2 + 0.076x_3 - 0.000x_4 - 0.003x_5)$,

$M_3 = \hat{y} = \exp(0.676 + 0.051x_1 - 0.008x_3 - 0.022x_4 - 0.013x_5)$,

$M_4 = \hat{y} = \exp(2.903 + 0.073x_1 - 0.066x_2 + 0.000x_4 - 0.000x_5)$,

$M_5 = \hat{y} = \exp(2.774 + 0.044x_1 - 0.084x_2 + 0.050x_3 - 0.001x_5)$,

$M_6 = \hat{y} = \exp(2.475 + 0.047x_1 - 0.076x_2 + 0.044x_3 - 0.002x_4)$,

$M_7 = \hat{y} = \exp(2.747 + 0.042x_1 - 0.081x_2 + 0.049x_3)$,

$M_8 = \hat{y} = \exp(2.965 + 0.073x_1 - 0.065x_2)$.

3.4. Simultaneous Parameter Testing

Simultaneous parameter testing is performed for each of the possible models in PIG regression. The following is given the test statistical value G as equation (22) shown by Table 7, it can be seen that the test statistical value G for each model, be it a model M_1, M_2 until the model M_8 produces a test statistical value G that is greater than the value χ^2_{tabel}, so that a rejection decision can be taken H_0 Because H_0, then it can be concluded that all parameters are simultaneously significant for all possible models in PIG regression.

Table 7. Simultaneous Parameter Testing

Model	Variables in the Model	Statistical Value G	Free Degrees	Value χ^2_{Tabel}	Results
M_1	x_1, x_2, x_3, x_4, x_5	389.67	32	46.194	Less H_0
M_2	x_2, x_3, x_4, x_5	390.85	33	47.400	Less H_0
M_3	x_1, x_3, x_4, x_5	396.87	33	47.400	Less H_0
M_4	x_1, x_2, x_4, x_5	390.86	33	47.400	Less H_0
M_5	x_1, x_2, x_3, x_5	389.68	33	47.400	Less H_0
M_6	x_1, x_2, x_3, x_4	389.68	33	47.400	Less H_0
M_7	x_1, x_2, x_3	389.70	34	48.602	Less H_0
M_8	x_1, x_2	390.86	35	49.802	Less H_0

3.5. Partial Parameter Testing and Best Model Selection

To see the influence of each parameter on the PIG regression model, a partial parameter test was performed. Partial parameter testing is only given for PIG regression models that have the smallest AIC values. The smallest AIC value for each of the possible models in the PIG regression is given by the following Table 8.

Table 8. AIC Values for Each Model

Model	Variables in the Model	AIC Value
M_1	x_1, x_2, x_3, x_4, x_5	403.67
M_2	x_2, x_3, x_4, x_5	402.85
M_3	x_1, x_3, x_4, x_5	408.87
M_4	x_1, x_2, x_4, x_5	402.86
M_5	x_1, x_2, x_3, x_5	401.68
M_6	x_1, x_2, x_3, x_4	401.68
M_7	x_1, x_2, x_3	399.70
M_8	x_1, x_2	**398.86**

In this study, the best model chosen was the model with the smallest AIC value. Based on Table 8, the model with the smallest AIC value is the one M_8 with an AIC value, which is 398.86, so the best model chosen is the M_8 model. Furthermore, the testing parameters are partially $H_0 : \tau = 0$ to M_8 be provided by Table 9 with the test hypothesis given as follows.

Hypotheses for model parameters.

$H_0 : \beta_l = 0$

$H_1 : \beta_l \neq 0$ with $l = 1, 2, ..., k$.

Hypotheses for dispersion parameters.

$H_1 : \tau \neq 0$.

Table 9. Partial Parameter Testing M_8

Parameter	Parameter Values	Standard *Error*	Value t	*P-Value*	Results
$\hat{\beta}_0$	2.965	3.065	0.967	0.340	Failed to Reject *H0*
$\hat{\beta}_1$	0.073	0.028	2.555	**0.015**	Subtract *H0*
$\hat{\beta}_2$	-0.065	0.019	-3.517	**0.001**	Subtract *H0*
$\hat{\tau}$	-0.697	0.282	-2.469	**0.019**	Subtract *H0*

By looking at the *p-value* in Table 9 above, it can be seen that the *p-value* for the parameter $\hat{\beta}_1$ and $\hat{\beta}_2$ less than the significance value of 5% = 0.05 which means that variable x_1 and x_2 significant affect the number of infant deaths in East Java Province in 2019. Likewise, it is significant that the data on the number of infant deaths in East Java province in 2019 contains overdispersion, so the appropriate method for modelling the case is PIG regression.$\hat{\tau}$

3.6. Model Interpretation

The interpretation of the PIG regression model was carried out by looking at each coefficient in the model that has the smallest AIC value, namely the M8 model given by the following equation (26).

$$M_8 = \hat{y} = \exp(2.965 + 0.073x_1 - 0.065x_2) \quad (26)$$

Based on the model in equation (27) above, it is possible to interpret each variable by calculating $\exp(\hat{\beta}_j)$, $j = 1, 2$ the value of. For the variable Percentage of Visits of Pregnant Women with $K_1(x_1)$ a significant value $\exp(\hat{\beta}_1) = \exp(0.073) = 1.076$ that for an increase of one percent of the variable of the Percentage of Visits of Pregnant Women will $K_1(x_1)$ increase the number of infant deaths in East Java province in 2019. This is inversely proportional to the actual theory that the risk of infant mortality should decrease if the percentage of visits by pregnant women is $K_1(x_1)$ higher. However, for the variable Percentage of Visits of Pregnant Women with $K_4(x_2)$ a value $\exp(\hat{\beta}_1) = \exp(0.065) = 1.067$ and a coefficient of negative value, which means that every one percent increase in the Percentage of Visits of Pregnant Women will $K_4(x_2)$ reduce the number of infant deaths in East Java Province in 2019.

4. Conclusion

Based on the analysis of the data that has been carried out, it can be concluded that the PIG regression model provides better model results than the Poisson regression model in overcoming data that have cases of overdispersion by looking at the smallest AIC value. Furthermore, of the several possible PIG regression models that are formed, the best model obtained is the one with the smallest AIC value, i.e. the M_8 With the predictor variables in the M8 model are the Percentage of Visits of Pregnant Women with $K_1(x_1)$ and the Percentage of Visits of Pregnant Women with $K_4(x_2)$. Based on the model M_8, it can be concluded that the risk of infant mortality in East Java province in 2019 will increase if pregnant women's visits are only carried out, K_1 while if pregnant women's visits are also carried out K_4 , it will reduce the risk of infant deaths in East Java province in 2019.

References

1. Yadav, B. (2021). "Can Generalized Poisson model replace any other count regression model?" *Journal of Statistical Computation and Simulation*.
2. Lee, W. (2022). "Revisiting the analysis pipeline for overdispersed Poisson regression." *Statistical Modelling*.
3. Perreault, S. et al. (2024). "Case-crossover designs and overdispersion with application in air pollution epidemiology." *Environmental Health Perspectives*.
4. Burger, D. A. et al. (2025). "A robust contaminated discrete Weibull regression model for outlier-prone count data." *arXiv preprint*.
5. Otto, A. F. et al. (2025). "Modeling Bounded Count Environmental Data Using a Contaminated Beta-Binomial Regression Model." *arXiv preprint*.
6. Cooper, A. et al. (2025). "Dominating Hyperplane Regularization for Variable Selection in Multivariate Count Regression." *arXiv preprint*.
7. Berk, R., & MacDonald, J. M. (2008). "Overdispersion and Poisson Regression." *Journal of Quantitative Criminology*, 24(3), 269–284.
8. Hilbe, J. M. (2007). *Negative Binomial Regression* (1st ed.). Cambridge University Press.
9. Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data* (2nd ed.). Cambridge University Press.
10. Souza, R. F. (2023). "overdisp: Overdispersion in Count Data Multiple Regression Analysis." *CRAN R Package*. Link
11. Payne, E. H. et al. (2015). "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.
12. Smith, D., & Faddy, M. (2016). "Mean and Variance Modeling of Under and Overdispersed Count Data." *Journal of Statistical Software*, 69(6), 1–23.
13. Campbell, H. (2021). "The consequences of checking for zero-inflation and overdispersion in count data models." *Methods in Ecology and Evolution*, 12(3), 401–412.
14. Hilbe, J. M. (2011). *Negative Binomial Regression* (1st ed.). Cambridge University Press.
15. Ver Hoef, J. M., & Boveng, P. L. (2007). "Quasi-Poisson vs. Negative Binomial Regression: How should we model overdispersed count data?" *Ecology*, 88(11), 2766–2772.
16. Souza, R. F. (2023). "overdisp: Overdispersion in Count Data Multiple Regression Analysis." *CRAN R Package*.
17. Berk, R., & MacDonald, J. M. (2008). "Overdispersion and Poisson Regression." *Journal of Quantitative Criminology*, 24(3), 269–284.
18. Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.
19. Ver Hoef, J. M., & Boveng, P. L. (2007). "Quasi-Poisson vs. Negative Binomial Regression: How should we model overdispersed count data?" *Ecology*, 88(11), 2766–2772.
20. Souza, R. F. (2023). "overdisp: Overdispersion in Count Data Multiple Regression Analysis." *CRAN R Package*.