Austrian Journal of Statistics
2026, Volume 55, Issue 3, 113–137.
http://www.ajs.or.at/

A Statistical Approach Zero Inflated Negative Binomial and Hurdle Negative Binomial Modelling for Toddler Deaths due to Pneumonia

A'yunin Sofro
Universitas Negeri Surabaya
East Java 60231, Indonesia

Mutia Eva Mustafidah
Universitas Negeri Surabaya
East Java 60231, Indonesia

Khusnia Nurul Khikmah
Universitas Palangka Raya
Central Kalimantan 74874, Indonesia

Abstract

The Generalized Linear Model (GLM) is an extension of the general regression model for response variables following an exponential family distribution, including normal, binomial, Poisson, negative binomial, exponential, and gamma. If the response variable is discrete and follows a Poisson distribution, then the Poisson regression model can be used for model formation. However, in its application, overdispersion often occurs, where the variance is greater than the mean. Overdispersion in Poisson regression can occur due to a large number of observations having zero values in the response variable (excess zeros). Data experiencing overdispersion and excess zeros are more suitable for using Zero Inflated Negative Binomial (ZINB) and Hurdle Negative Binomial (HNB) regressions.

In this study, the models were further developed into ZINB and HNB regression models with transformed variables to improve model performance. Real-world issues related to these methods can be encountered in mortality cases, where the data used pertain to the number of toddler deaths due to pneumonia in East Java in 2022.

The results of model selection using AIC show that the ZINB regression model with transformed variables is the best model in this study, with an AIC value of 58.63682. The results of the partial significance test of parameters in the ZINB regression model with transformed variables indicate that the percentage of vitamin A supplementation (x_5), and the percentage of exclusive breastfeeding (x_7) significantly influence the number of toddler deaths due to pneumonia.

Keywords: pneumonia, zero inflated negative binomial, hurdle negative binomial.

1. Introduction

The Generalized Linear Model (GLM) is an extension of the general regression model for response variables that follow exponential family distributions, including normal, binomial,

Poisson, negative binomial, exponential, and gamma distributions (Agresti 2015). When the response variable is discrete and follows a Poisson distribution, a Poisson regression model can be used to model the data (Cameron and Trivedi 2014). One important assumption in Poisson regression analysis is that the variance equals the mean, known as equidispersion. However, this assumption is often violated in some cases, leading to overdispersion, where the variance is greater than the mean. According to Hilbe (2011), when overdispersion occurs, Poisson regression is no longer suitable because it underestimates standard errors, leading to errors in determining the significance of regression parameters. One alternative to address overdispersion in Poisson regression is using the Negative Binomial regression (Saputro and Qudratullah 2021).

However, in some cases, count data not only experiences overdispersion but also excess zeros (Sharker, Balbuena, Marcoux, and Feng 2020). Excess zeros occur when the number of zero values in the response variable is greater than expected from the assumed distribution, which can lead to erroneous conclusions. According to Puig and Valero (2006), detecting excess zeros requires careful statistical testing because the presence of zero inflation can significantly impact model selection and parameter estimation. Blasco-Moreno, Pérez-Casany, Puig, Morante, and Baldó-Serra (2019) further emphasize that misidentifying excess zeros can lead to incorrect inferences, and they propose improved methodologies for assessing zero inflation in count data models.

In cases where both overdispersion and excess zeros are present, Negative Binomial regression is no longer suitable, and alternative regression models must be considered (Bilgic and Florkowski 2007). There are several regression models that can handle both overdispersion and excess zeros, including the Zero Inflated Negative Binomial (ZINB) and Hurdle Negative Binomial (HNB) regression models. ZINB is a regression model formed from a Poisson-Gamma mixture distribution. The ZINB regression model has two components: the Negative Binomial State and the Zero Inflation State. The first component, the Negative Binomial State, with probability $(1 - \pi_i)$, generates non-negative observations that follow a Negative Binomial distribution with mean μ_i. The second component, the Zero Inflation State, with probability (π_i), generates zero observations (Minami, Lennert-Cody, Gao, and Román-Verdesoto 2007). On the other hand, HNB is a regression model formed from a Poisson-Gamma mixture distribution. The Hurdle Negative Binomial model has two components: the Zero Hurdle State and the Truncated Negative Binomial State. The first component, the Zero Hurdle State, with probability (π_i), generates zero observations with a hurdle. The second component, the Truncated Negative Binomial State, with probability $(1 - \pi_i)$, generates non-negative observations after overcoming the hurdle, following a Negative Binomial distribution with mean μ_i (Ma, Yan, Wei, and Wang 2016).

The difference between zero-inflated and hurdle distributions lies in the concepts of structural zero and sampling zero. Structural zero refers to situations where the response value is always zero due to the nature of the data itself. For example, there may be conditions or areas where a particular disease cannot occur. On the other hand, sampling zero occurs when the response variable takes a zero value simply due to the absence of an event in the observed sample, but the event may still occur in other situations or populations. Zero-inflated models are appropriate when there are more zeros than expected from the assumed distribution, which includes both structural zero (i.e., events that cannot happen) and sampling zero (i.e., rare but possible events). Meanwhile, hurdle models are used when zeros occur solely due to sampling, that is, when count data has a hurdle that must be passed before positive values can be observed. After this hurdle is surpassed, the data follows a truncated distribution (e.g., Negative Binomial) for positive outcomes.

The ZINB and HNB models are chosen in this study because they are specifically designed to handle both overdispersion and excess zeros simultaneously. Overdispersion occurs when the variance of count data is greater than the mean, which is commonly found in real-world data. Poisson regression, which assumes equidispersion (mean equals variance), is unsuitable in this case, leading to biased parameter estimates. While Negative Binomial regression can handle

overdispersion, it is not effective in addressing excess zeros, i.e., when there are more zeros than expected from the underlying distribution (Yıldırım, Kaçıranlar, and Yıldırım 2022). The ZINB and HNB models are more appropriate in this situation because both models include components that specifically model the mechanisms of zero inflation or zero hurdle.

The ZINB model is particularly useful when there is a two-stage process in the formation of the data. In ZINB, the first stage models the excess zeros using a binary model, while the second stage models the count data (non-zero values) using a Negative Binomial distribution. This allows for a more accurate representation of data where most observations are zeros, but there are also processes that generate non-zero count values. Meanwhile, the HNB model is more useful when zeros are considered as a hurdle that must be crossed before the data follows a Negative Binomial distribution. The HNB model first models the occurrence of zeros and then models positive counts (after overcoming the hurdle) using the Negative Binomial distribution.

In the literature, there are alternative models that also handle excess zeros, such as the Zero Inflated Generalized Poisson, Zero Inflated Poisson Lindley, Zero-One-Inflated Poisson, or Zero-One-Inflated Poisson Lindley distributions. However, although these models can handle some cases of excess zeros, the ZINB and HNB models are preferred due to their flexibility in dealing with data that not only has high zero inflation but also significant overdispersion. The Zero Inflated Generalized Poisson, for example, while able to handle more complex distributions, is often less efficient in handling overdispersion compared to the Negative Binomial distribution used in ZINB and HNB. Likewise, the Zero-One-Inflated Poisson model, designed to handle data with structural and sampling zeros, is still less flexible in dealing with significant overdispersion. Furthermore, the Zero-One-Inflated Poisson Lindley distribution, though newer, has not yet proven superior in handling overdispersion, especially in cases with high levels of excess zeros.

Considering the complexity of the data and the goal of the study to obtain a model that is not only accurate in estimating the count data distribution but also capable of handling overdispersion and excess zeros optimally, ZINB and HNB are more appropriate choices compared to other distribution models.

This study applies both the ZINB and HNB models to analyze the factors influencing toddler deaths due to pneumonia in East Java. Data exhibiting overdispersion and excess zeros are better handled by these two models. Given the high rate of toddler deaths from pneumonia compared to other diseases such as diarrhea and dengue fever, as reported by the East Java Health Department, it is crucial to use the appropriate statistical methods to identify significant factors affecting toddler mortality and determine the most suitable model for this context.

In this study, the models were further developed into ZINB regression models with transformed variables and HNB regression models with transformed variables to improve model performance. Real-world issues related to these methods can be encountered in mortality cases, where the data used pertain to the number of toddler deaths due to pneumonia in East Java in 2022.

By employing both ZINB and HNB models, this research aims to fill gaps in previous studies on toddler deaths due to pneumonia, particularly related to challenges in handling data with zero inflation and overdispersion. The findings of this study are expected to contribute to more accurate risk assessment and the development of interventions to reduce toddler deaths from pneumonia in the region.

2. Mathematics model

2.1. Zero Inflated Negative Binomial regression

Zero Inflated Negative Binomial regression is a regression model derived from a Poisson-Gamma mixture distribution (Myers, Montgomery, Vining, and Robinson 2012). This distribution is one of the approaches within the Negative Binomial distribution used to address cases of overdispersion in Poisson regression. It accounts for two conditions: the Negative Binomial State and the Zero Inflation State. The first condition, the Negative Binomial State, occurs with a probability of $(1-\pi_i)$ and generates non-negative observations that follow a Negative Binomial distribution with a mean of μ_i. The second condition, the Zero Inflation State, occurs with a probability of (π_i) and generates observations with a value of zero (Saengthong, Bodhisuwan, and Thongteeraparp 2015).

The probability function of the ZINB regression model is as follows (Garay, Hashimoto, Ortega, and Lachos 2011):

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

Where $i = 1, 2, \ldots, n$, $0 \leq \pi_i \leq 1$, $\mu_i \geq 0$ and ϕ are dispersion parameters with $\phi > 0$, $\Gamma\left(y_i + \frac{1}{\phi}\right)$ is a gamma function of $y_i + \frac{1}{\phi}$ and $\Gamma\left(\frac{1}{\phi}\right)$ is the gamma function of $\frac{1}{\phi}$.

When $\pi_i = 0$, the random variable y_i has a Negative Binomial distribution with mean μ_i and dispersion parameter ϕ. It is further assumed that the parameters μ_i and π_i depend on the vector of predictor variables $\mathbf{x}_i$. So the Zero Inflated Negative Binomial regression model consisting of two models is expressed in the following equation:

1. Negative Binomial State Model

$$\begin{aligned} y_i|\mu_i &\sim NB(\mu_i) \\ \ln(\mu_i) &= \boldsymbol{x}_i^T\boldsymbol{\beta}. \end{aligned} \tag{2}$$

   Where $\mu_i \geq 0$, i=1, 2, ..., n, j=1, 2, ..., p, n is the number of observations and p is the number of predictor variables, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$.

2. Zero Inflation State Model

$$\text{logit}(\hat{\pi}_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\gamma}. \tag{3}$$

   Where $0 \leq \pi_i \leq 1$, i=1, 2, ..., n, j=1, 2, ..., p, n is the number of observations and p is the number of predictor variables, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)^T$.

2.2. Hurdle Negative Binomial regression

Hurdle Negative Binomial regression is a model derived from a mixture distribution of Poisson and Gamma (Myers *et al.* 2012). This distribution is one of the approaches within the Negative Binomial distribution used to address cases of overdispersion in Poisson regression. The regression consists of two components: the Zero Hurdle State and the Truncated Negative Binomial State. The first component, the Zero Hurdle State, with probability (π_i), generates zero-valued observations with a hurdle. The second component, the Truncated Negative Binomial State, with probability $(1-\pi_i)$, generates non-negative observations that have

passed through the hurdle (truncated), following a Negative Binomial distribution with mean μ_i (Saffari, Adnan, and Greene 2012).

The probability function of the HNB regression model is as follows (Greene 2008):

$$P(y_i) = \begin{cases} \pi_i & \text{for} \quad y_i = 0 \\ (1-\pi_i)\frac{\Gamma\left(y_i+\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right)y_i!}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{y_i}\frac{(1+\phi\mu_i)^{-\frac{1}{\phi}}}{1-(1+\phi\mu_i)^{-\frac{1}{\phi}}} & \text{for} \quad y_i > 0. \end{cases} \tag{4}$$

Where $i = 1, 2, \ldots, n$, $0 \leq \pi_i \leq 1$, $\mu_i \geq 0$ and ϕ are dispersion parameters with $\phi > 0$, $\Gamma\left(y_i + \frac{1}{\phi}\right)$ is the gamma function of $y_i + \frac{1}{\phi}$ and $\Gamma\left(\frac{1}{\phi}\right)$ is the gamma function of $\frac{1}{\phi}$.

When $\pi_i = 0$, the random variable y_i has a Negative Binomial distribution with mean μ_i and dispersion parameter ϕ. Next, it is assumed that the parameters μ_i and π_i depend on the vector of predictor variables $\boldsymbol{x}_i$. So the regression model Hurdle Negative Binomial can be expressed by two models as follows:

1. Zero Hurdle State Model

$$\text{logit}(\hat{\pi}_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\gamma}. \tag{5}$$

   Where $0 \leq \pi_i \leq 1$, i=1, 2, ..., n, j=1, 2, ..., p, n is the number of observations and p is the number of predictor variables, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)^T$.

2. Truncated Negative Binomial State Model

$$\begin{aligned} y_i|\mu_i &\sim NB(\mu_i) \\ \ln(\mu_i) &= \boldsymbol{x}_i^T\boldsymbol{\beta}. \end{aligned} \tag{6}$$

   Where $\mu_i \geq 0$, i=1, 2, ..., n, j=1, 2, ..., p, n is the number of observations and p is the number of predictor variables, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$.

3. Estimation of parameters

3.1. Parameter estimation in Zero Inflated Negative Binomial regression

The maximum likelihood estimation method can be used to estimate the parameters of the Zero Inflated Negative Binomial regression model.

Based on Equation (2), the obtained result is:

$$\ln(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

$$\mu_i = e^{\boldsymbol{x}_i^T\beta} \tag{7}$$

Meanwhile, based on Equation (3), the result obtained is:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\gamma}$$

$$\pi_i = \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}} \tag{8}$$

$$(1-\pi_i) = \frac{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}} - \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}} = \frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}. \tag{9}$$

Equations (7), (8), and (9) are substituted into Equation (1), resulting in the probability function of the ZINB regression model, as follows:

$$P(y_i) = \begin{cases} \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}} + \left(\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}\right)\left(\frac{1}{1+\phi e^{\boldsymbol{x}_i^T\beta}}\right)^{\frac{1}{\phi}} & \text{for} \quad y_i = 0 \\ \left(\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}\right)\frac{\Gamma\left(y_i+\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right)y_i!}\left(\frac{1}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\phi}}\left(\frac{\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{y_i} & \text{for} \quad y_i > 0. \end{cases} \tag{10}$$

Where i=1, 2, ..., n, j=1, 2, ..., p, n is the number of observations and p is the number of predictor variables, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, $\boldsymbol{\gamma}$= $(\gamma_1, \gamma_2, \ldots, \gamma_p)^T$ and ϕ is the dispersion parameter.

The likelihood function of the ZINB regression model can be written as follows:

$$L(\boldsymbol{\beta},\boldsymbol{\gamma}) = \begin{cases} \Pi_{i=1}^n \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}} + \frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}\left(\frac{1}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\phi}} & \text{for} \quad y_i = 0 \\ \Pi_{i=1}^n \frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}\frac{\Gamma\left(y_i+\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right)y_i!}\left(\frac{1}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\phi}}\left(\frac{\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{y_i} & \text{for} \quad y_i > 0. \end{cases} \tag{11}$$

The next step after obtaining the likelihood function of the ZINB regression model is to form the log-likelihood function based on Equation (11). The form of the log-likelihood function for the ZINB regression model can be written as follows:

$$\ln L(\boldsymbol{\beta},\boldsymbol{\gamma}) = \begin{cases} \sum_{i=1}^n \ln\left[e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}} + \left(\frac{1}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\phi}}\right] - \sum_{i=1}^n \ln\left[1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}\right] & \text{for} \quad y_i = 0 \\ \sum_{i=1}^n \ln\Gamma\left(y_i+\frac{1}{\phi}\right) + \frac{1}{\phi}\sum_{i=1}^n \ln\left(\frac{1}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right) + y_i\sum_{i=1}^n \ln\left(\frac{\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right) & \\ -\sum_{i=1}^n \ln\left(1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}\right) - \sum_{i=1}^n \ln(y_i!) - \sum_{i=1}^n \ln\Gamma\left(\frac{1}{\phi}\right) & \text{for} \quad y_i > 0. \end{cases} \tag{12}$$

The function ln likelihood in Equation (12) has 2 conditions that are combined, namely when $y_i = 0$ which represents the group zero Inflated state and $y_i > 0$ which represents the group negative binomial state. So that it makes the calculation difficult and it is not known which zero value comes from zero Inflated state and negative binomial state. To describe the condition of the variable y_i in detail, y_i will be redefined with a latent variable z_i to find out the zero value from zero Inflated state and negative binomial state which can be defined as follows (Hall 2000):

$$z_i = \begin{cases} 1 & \text{for } y_i = 0, \text{came from Zero Inflation State} \\ 0 & \text{for } y_i > 0, \text{came from Negative Binomial State.} \end{cases}$$

Then, determine the probability function of the latent variable z_i as follows:

$$P(z_i) = \begin{cases} \pi_i & \text{for} \quad z_i = 1 \\ 1-\pi_i & \text{for} \quad z_i = 0. \end{cases} \tag{13}$$

Based on Equation (13) with the condition of Zero Inflation State ($P(z_i = 1) = \pi_i$) and condition of Negative Binomial State ($P(z_i = 0) = 1 - \pi_i$). Then, the joint probability function between y_i and z_i which is as follows:

$$\begin{aligned} f(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y},\mathbf{z}) &= f(\boldsymbol{z})f(\boldsymbol{y}|\boldsymbol{z}) \\ &= f(\boldsymbol{z}|1,\pi_i)f(\boldsymbol{y}|\boldsymbol{z},\mu_i) \\ &= (\pi_i)^{z_i}(1-\pi_i)^{(1-z_i)}\left[\frac{\Gamma\left(y_i+\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right)y_i!}\left(\frac{1}{1+\phi\mu_i}\right)^{\frac{1}{\phi}}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{y_i}\right]^{(1-z_i)} \end{aligned} \tag{14}$$

Then Equations (7), (8), and (9) are substituted into Equation (14), yielding:

$$\begin{aligned} f(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y},\mathbf{z}) &= \left(\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}\right)^{z_i}\left(\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}\right)^{1-z_i}\left(\frac{\Gamma\left(y_i+\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right)y_i!}\left(\frac{1}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\phi}}\left(\frac{\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{y_i}\right)^{1-z_i} \\ &= \left(e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}\right)^{z_i}\left(\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}\right)\left(\frac{\Gamma\left(y_i+\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right)y_i!}\left(\frac{1}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\phi}}\left(\frac{\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{y_i}\right)^{1-z_i}. \end{aligned} \tag{15}$$

The new likelihood function from the joint probability function between y_i and z_i can be expressed as the following equation:

$$L(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y},\mathbf{z}) = \Pi_{i=1}^{n}\left(e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}\right)^{z_i}\left(\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}\right)\left(\frac{\Gamma\left(y_i+\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right)y_i!}\left(\frac{1}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{\frac{1}{\phi}}\left(\frac{\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right)^{y_i}\right)^{1-z_i}. \tag{16}$$

Thus, the ln likelihood function can be written as follows (Garay *et al.* 2011):

$$\begin{aligned} \ln L(\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y},\mathbf{z}) = &\sum_{i=1}^{n}(z_i)(\boldsymbol{x}_i^T\boldsymbol{\gamma}) - \sum_{i=1}^{n}\ln\left(1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}\right) + \left(\sum_{i=1}^{n}(1-z_i)\right)\sum_{i=1}^{n}\Gamma\left(y_i+\frac{1}{\phi}\right) - \sum_{i=1}^{n}\Gamma\left(\frac{1}{\phi}\right) \\ &- \sum_{i=1}^{n}\ln(y_i!) + y_i\sum_{i=1}^{n}\ln(\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}) - (y_i+\phi^{-1})\sum_{i=1}^{n}\ln\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right). \end{aligned} \tag{17}$$

Parameter estimation process $\boldsymbol{\beta},\boldsymbol{\gamma}$ is performed separately, so the complete likelihood function can be rewritten as Equations (18) and (19) as follows:

$$\begin{aligned} \ln L(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{z}) = &\left(\sum_{i=1}^{n}(1-z_i)\right)\sum_{i=1}^{n}\Gamma\left(y_i+\frac{1}{\phi}\right) - \sum_{i=1}^{n}\Gamma\left(\frac{1}{\phi}\right) - \sum_{i=1}^{n}\ln(y_i!) \\ &+ y_i\sum_{i=1}^{n}\ln(\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}) - (y_i+\phi^{-1})\sum_{i=1}^{n}\ln\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right) \end{aligned} \tag{18}$$

and

$$\ln L(\boldsymbol{\gamma}|\boldsymbol{y},\mathbf{z}) = \sum_{i=1}^{n}(z_i)(\mathbf{X}\boldsymbol{\gamma}) - \sum_{i=1}^{n}\ln\left(1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}\right). \tag{19}$$

1. To obtain parameter estimates $\boldsymbol{\beta}$ the ln likelihood function in Equation (18) is partially differentiated with respect to the parameters $\boldsymbol{\beta}$.

   The first derivative of the ln likelihood function in Equation (18) with respect to $\boldsymbol{\beta}$ can be expressed as the following equation:

$$\frac{\partial \ln L\left(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{z}\right)}{\partial(\boldsymbol{\beta})} = \sum_{i=1}^{n}(1-z_i)\left(y_i\sum_{i=1}^{n}\boldsymbol{x}_i^T - (y_i+\phi^{-1})\sum_{i=1}^{n}\frac{\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\boldsymbol{x}_i^T}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right). \tag{20}$$

Then, the second derivative of the ln likelihood function in Equation (18) with respect to $\boldsymbol{\beta}$ can be expressed as the following equation:

$$\frac{\partial^2 \ln L\left(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{z}\right)}{\partial^2(\boldsymbol{\beta})} = -\sum_{i=1}^{n}(1-z_i)\left((y_i+\phi^{-1})\sum_{i=1}^{n}\frac{\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\left(\boldsymbol{x}_i^T\right)^2}{\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^2}\right). \tag{21}$$

Estimation for $\boldsymbol{\beta}$ direct estimation cannot be obtained directly because the results from the equation derived from the first derivative of the ln likelihood function with respect to each parameter do not have a closed form. To address this issue, numerical iteration methods are used to facilitate calculations, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method to obtain parameter estimates (Hilbe 2011). The steps of the BFGS numerical iteration method are as follows:

a. Determine the initial estimate of the parameter $\hat{\boldsymbol{\beta}}^{(0)}$ obtained using the OLS method and determine $\boldsymbol{H}^{(0)}$ as follows:

$$\hat{\boldsymbol{\beta}}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

With

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

and

$$\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T.$$

Then the matrix $\boldsymbol{H}^{(0)} = \boldsymbol{I}$

$$\boldsymbol{H}^{(0)} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & 0 & \ddots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

b. Forming a gradient vector **u**, which is a vector consisting of the first partial derivatives of the estimated parameters.

c. Forming the Hessian matrix $\boldsymbol{H}$, which is a matrix containing the partial derivatives of the two estimated parameters.

d. Insert the value of $\hat{\boldsymbol{\beta}}^{(0)}$ into the elements of the vector **u** and the matrix **H** to obtain the vector $\boldsymbol{u}\left(\hat{\boldsymbol{\beta}}^{(0)}\right)$ and the matrix $\boldsymbol{H}\left(\hat{\boldsymbol{\beta}}^{(0)}\right)$.

e. Perform a BFGS iteration *Quasi Newton* starting from $r = 0$ with the following equation:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} - a^{(r)}\left(\boldsymbol{H}^{(r)}\right)^{-1}\boldsymbol{u}^{(r)}.$$

With:

$\hat{\boldsymbol{\beta}}^{(r)}$ = a set of parameter estimators that converge at the rth iteration.

$a^{(r)}$=min $\ln L\left(\hat{\boldsymbol{\beta}}^{(r)} - a^{(r)}\left(\boldsymbol{H}^{(r)}\right)^{-1}\boldsymbol{u}^{(r)}\right)$.

f. Calculate $\boldsymbol{s}^{(r)} = \hat{\boldsymbol{\beta}}^{(r+1)} - \hat{\boldsymbol{\beta}}^{(r)}$ and $\boldsymbol{y}^{(r)} = \hat{\boldsymbol{u}}^{(r+1)} - \hat{\boldsymbol{u}}^{(r)}$ so that the BFGS update matrix is obtained as follows:

$$\boldsymbol{H}^{(r+1)} = \boldsymbol{H}^{(r)} - \frac{\boldsymbol{H}^{(r)}\boldsymbol{s}^{(r)}\left(\boldsymbol{s}^{(r)}\right)^{T}\boldsymbol{H}^{(r)}}{\left(\boldsymbol{s}^{(r)}\right)^{T}\boldsymbol{H}^{(r)}\boldsymbol{s}^{(r)}} + \frac{\boldsymbol{y}^{(r)}\left(\boldsymbol{y}^{(r)}\right)^{T}}{\left(\boldsymbol{s}^{(r)}\right)^{T}\boldsymbol{y}^{(r)}}.$$

g. The iteration process will stop if a parameter estimate has been obtained that is convergent by satisfying $\left(\left|\hat{\boldsymbol{\beta}}^{(r+1)} - \hat{\boldsymbol{\beta}}^{(r)}\right|\right) \leq \epsilon$ where ϵ is a very small value and has been applied previously, for example 10^{-4}.

2. Similar to the parameter maximization stage for $\boldsymbol{\beta}$, the maximization stage for parameter $\boldsymbol{\gamma}$ also utilizes numerical iteration methods such as the Newton Raphson method. To obtain parameter estimates $\boldsymbol{\gamma}$, the ln likelihood function from Equation (19) is partially differentiated with respect to the parameters $\boldsymbol{\gamma}$.

The first derivative of the ln likelihood function in Equation (19) with respect to ($\boldsymbol{\gamma}$) is expressed as follow:

$$\frac{\partial \ln L\left(\boldsymbol{\gamma}|\boldsymbol{y},\mathbf{z}\right)}{\partial(\boldsymbol{\gamma})} = \sum_{i=1}^{n}(z_i)\left(\boldsymbol{x}_i^T - \sum_{i=1}^{n}\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}\boldsymbol{x}_i^T\right). \tag{22}$$

Then, the second derivative to $\boldsymbol{\gamma}$ can be expressed as the following equation:

$$\frac{\partial^2 \ln L\left(\boldsymbol{\gamma}|\boldsymbol{y},\mathbf{z}\right))}{\partial^2(\boldsymbol{\gamma})} = -\sum_{i=1}^{n}\left(\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}\left(\boldsymbol{x}_i^T\right)^2}{\left(1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}\right)^2}\right). \tag{23}$$

Estimation for $\boldsymbol{\gamma}$ direct estimation cannot be obtained directly because the results from the equation derived from the first derivative of the ln likelihood function with respect to each parameter do not have a closed form. To address this issue, numerical iteration methods are used to facilitate calculations, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method to obtain parameter estimates (Hilbe 2011). The steps of the BFGS numerical iteration method are as follows:

a. Determine the initial estimate of the parameter $\hat{\boldsymbol{\gamma}}^{(0)}$ obtained using the OLS method and determine $\boldsymbol{H}^{(0)}$ as follows:

$$\hat{\boldsymbol{\gamma}}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

With

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

and

$$\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T.$$

Then the matrix $\boldsymbol{H}^{(0)} = \boldsymbol{I}$

$$\boldsymbol{H}^{(0)} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & 0 & \ddots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

b. Forming a gradient vector **u**, which is a vector consisting of the first partial derivatives of the estimated parameters.

c. Forming the Hessian matrix $\boldsymbol{H}$, which is a matrix containing the partial derivatives of the two estimated parameters.

d. Insert the value of $\hat{\boldsymbol{\gamma}}^{(0)}$ into the elements of the vector **u** and the matrix **H** to obtain the vector $\boldsymbol{u}\left(\hat{\boldsymbol{\gamma}}^{(0)}\right)$ and the matrix $\boldsymbol{H}\left(\hat{\boldsymbol{\gamma}}^{(0)}\right)$.

e. Perform a BFGS iteration *Quasi Newton* starting from $r = 0$ with the following equation:

$$\hat{\boldsymbol{\gamma}}^{(r+1)} = \hat{\boldsymbol{\gamma}}^{(r)} - a^{(r)} \left(\boldsymbol{H}^{(r)}\right)^{-1} \boldsymbol{u}^{(r)}.$$

With:

$\hat{\boldsymbol{\gamma}}^{(r)}$ = a set of parameter estimators that converge at the rth iteration.

$a^{(r)}$=min ln $L\left(\hat{\boldsymbol{\gamma}}^{(r)} - a^{(r)} \left(\boldsymbol{H}^{(r)}\right)^{-1} \boldsymbol{u}^{(r)}\right)$.

f. Calculate $\boldsymbol{s}^{(r)} = \hat{\boldsymbol{\gamma}}^{(r+1)} - \hat{\boldsymbol{\gamma}}^{(r)}$ and $\boldsymbol{y}^{(r)} = \hat{\boldsymbol{u}}^{(r+1)} - \hat{\boldsymbol{u}}^{(r)}$ so that the BFGS update matrix is obtained as follows:

$$\boldsymbol{H}^{(r+1)} = \boldsymbol{H}^{(r)} - \frac{\boldsymbol{H}^{(r)} \boldsymbol{s}^{(r)} \left(\boldsymbol{s}^{(r)}\right)^T \boldsymbol{H}^{(r)}}{\left(\boldsymbol{s}^{(r)}\right)^T \boldsymbol{H}^{(r)} \boldsymbol{s}^{(r)}} + \frac{\boldsymbol{y}^{(r)} \left(\boldsymbol{y}^{(r)}\right)^T}{\left(\boldsymbol{s}^{(r)}\right)^T \boldsymbol{y}^{(r)}}.$$

g. The iteration process will stop if a parameter estimate has been obtained that is convergent by satisfying $\left(\left|\hat{\boldsymbol{\gamma}}^{(r+1)} - \hat{\boldsymbol{\gamma}}^{(r)}\right|\right) \leq \epsilon$ where ϵ is a very small value and has been applied previously, for example 10^{-4}.

3.2. Parameter estimation in Hurdle Negative Binomial regression

The maximum likelihood estimation (MLE) method can be used to estimate the parameters of the Hurdle Negative Binomial regression model.

Based on Equation (5), the result obtained is:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\gamma}$$

$$\pi_i = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}} \tag{24}$$

$$(1-\pi_i) = \frac{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}} - \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}} = \frac{1}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}. \tag{25}$$

Meanwhile, based on (6), the result obtained is:

$$\ln(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

$$\mu_i = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}. \tag{26}$$

Equations (24), (25), and (26) are substituted into Equation (4), resulting in the probability function of the HNB regression model, as follows:

$$P(y_i) = \begin{cases} \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}} & \text{for} \quad y_i = 0 \\ \frac{1}{1+e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}} \frac{\Gamma\left(y_i+\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right) y_i!} \left(\frac{\phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1+\phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}\right)^{y_i} \frac{\left(1+\phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}}{1-\left(1+\phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}} & \text{for} \quad y_i > 0. \end{cases} \tag{27}$$

Where i=1, 2, ..., n, j=1, 2, ..., p, n is the number of observations and p is the number of predictor variables, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)^T$ and ϕ is the dispersion parameter.

The likelihood function of the HNB regression model can be written as follows:

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \begin{cases} \Pi_{i=1}^n \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}} & \text{for} \quad y_i = 0 \\ \Pi_{i=1}^n \frac{1}{1+e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}} \frac{\Gamma\left(y_i+\frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right) y_i!} \left(\frac{\phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1+\phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}\right)^{y_i} \frac{\left(1+\phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}}{1-\left(1+\phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}} & \text{for} \quad y_i > 0. \end{cases} \tag{28}$$

The next step after obtaining the likelihood function of the HNB regression model is to construct the ln likelihood function based on Equation (28). The form of the ln likelihood function for the Hurdle Negative Binomial regression model can be written as the following equation:

$$\ln L(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \begin{cases} \sum_{i=1}^n \ln e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}} - \sum_{i=1}^n \ln\left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}\right) & \text{for} \quad y_i = 0 \\ - \sum_{i=1}^n \ln\left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}\right) + y_i \sum_{i=1}^n \left[\ln \phi + \ln e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right] - \left(\frac{1}{\phi} + y_i\right) & \\ \sum_{i=1}^n \left[\ln\left(1 + \phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)\right] + \sum_{i=1}^n \ln \Gamma\left(y_i + \frac{1}{\phi}\right) - \sum_{i=1}^n \ln\left(y_i!\right) - & \\ \sum_{i=1}^n \ln \Gamma\left(\frac{1}{\phi}\right) - \sum_{i=1}^n \left[\ln\left(1 - \left(1 + \phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}\right)\right] & \text{for} \quad y_i > 0. \end{cases} \tag{29}$$

The ln likelihood function in Equation (29) has two conditions combined when $y_i = 0$ representing the zero hurdle state, and when $y_i > 0$, representing the truncated negative binomial state. This makes calculations difficult and it cannot be determined which zero values originate from the zero inflation state and the negative binomial state (Famoye and Singh 2006). To describe the condition of the variable y_i in detail, y_i will be redefined with a latent variable z_i to determine the zero values originating from the zero hurdle state and the truncated negative binomial state, which can be defined as follows (Hall 2000):

$$z_i = \begin{cases} 1 & \text{for } y_i = 0, \text{came from Zero Hurdle State} \\ 0 & \text{for } y_i > 0, \text{came from Truncated Negative Binomial State.} \end{cases}$$

Then, determine the probability function of the latent variable z_i as follows:

$$P(z_i) = \begin{cases} \pi_i & \text{for } z_i = 1 \\ 1 - \pi_i & \text{for } z_i = 0. \end{cases} \tag{30}$$

Based on Equation (30) with the condition of Zero Hurdle State ($P(z_i = 1) = \pi_i$) and condition of Truncated Negative Binomial State ($P(z_i = 0) = 1 - \pi_i$). Then, the joint probability function between y_i and z_i which is as follows:

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}, \mathbf{z}) &= f(\boldsymbol{z}) f(\boldsymbol{y}|\boldsymbol{z}) \\ &= f(\boldsymbol{z}|1, \pi_i) f(\boldsymbol{y}|\boldsymbol{z}, \mu_i) \\ &= (\pi_i)^{z_i} (1-\pi_i)^{(1-z_i)} \left[\frac{\Gamma\left(y_i + \frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right) y_i!} \left(\frac{\phi e^{\boldsymbol{x}_i^T \beta}}{1 + \phi e^{\boldsymbol{x}_i^T \beta}} \right)^{y_i} \frac{\left(1 + \phi e^{\boldsymbol{x}_i^T \beta}\right)^{-\frac{1}{\phi}}}{1 - \left(1 + \phi e^{\boldsymbol{x}_i^T \beta}\right)^{-\frac{1}{\phi}}} \right]^{(1-z_i)} . \end{aligned} \tag{31}$$

Then Equations (24), (25), and (26) are substituted into Equation (31), yielding:

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}, \mathbf{z}) &= \left(\frac{e^{\boldsymbol{x}_i^T \gamma}}{1 + e^{\boldsymbol{x}_i^T \gamma}} \right)^{z_i} \left(\frac{1}{1 + e^{\boldsymbol{x}_i^T \gamma}} \right)^{1-z_i} \left(\frac{\Gamma\left(y_i + \frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right) y_i!} \frac{\phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + \phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}^{y_i} \frac{\left(1 + \phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}}{1 - \left(1 + \phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}} \right)^{1-z_i} \\ &= \left(e^{\boldsymbol{x}_i^T \gamma} \right)^{z_i} \left(\frac{1}{1 + e^{\boldsymbol{x}_i^T \gamma}} \right) \left(\frac{\Gamma\left(y_i + \frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right) y_i!} \left(\frac{\phi e^{\boldsymbol{x}_i^T \beta}}{1 + \phi e^{\boldsymbol{x}_i^T \beta}} \right)^{y_i} \frac{\left(1 + \phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}}{1 - \left(1 + \phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}} \right)^{1-z_i} . \end{aligned} \tag{32}$$

The new likelihood function from the joint probability function between y_i and z_i can be expressed as the following equation:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}, \mathbf{z}) = \Pi_{i=1}^{n} \left(e^{\boldsymbol{x}_i^T \gamma} \right)^{z_i} \frac{1}{1 + e^{\boldsymbol{x}_i^T \gamma}} \left(\frac{\Gamma\left(y_i + \frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right) y_i!} \left(\frac{\phi e^{\boldsymbol{x}_i^T \beta}}{1 + \phi e^{\boldsymbol{x}_i^T \beta}} \right)^{y_i} \frac{\left(1 + \phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}}{1 - \left(1 + \phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}} \right)^{1-z_i} . \tag{33}$$

Thus, the ln likelihood function can be written as follows (Garay *et al.* 2011):

$$\begin{aligned} \ln L(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}, \mathbf{z}) = &\sum_{i=1}^{n} (z_i)(\boldsymbol{x}_i^T \boldsymbol{\gamma}) - \sum_{i=1}^{n} \ln\left(1 + e^{\boldsymbol{x}_i^T \gamma}\right) + \left(\sum_{i=1}^{n} (1 - z_i) \right) \sum_{i=1}^{n} \Gamma\left(y_i + \frac{1}{\phi}\right) - \\ &\sum_{i=1}^{n} \Gamma\left(\frac{1}{\phi}\right) - \sum_{i=1}^{n} \ln y_i! - \left(\frac{1}{\phi} + y_i\right) \left[\sum_{i=1}^{n} \ln\left(1 + \phi e^{\boldsymbol{x}_i^T \beta}\right) \right] + \\ &y_i \sum_{i=1}^{n} \left(\ln \phi + \ln e^{\boldsymbol{x}_i^T \beta} \right) - \sum_{i=1}^{n} \ln\left(1 - \left(1 + \phi e^{\boldsymbol{x}_i^T \beta}\right)^{-\frac{1}{\phi}}\right) . \end{aligned} \tag{34}$$

Parameter estimation process of $\boldsymbol{\beta}, \boldsymbol{\gamma}$ is performed separately, so the complete likelihood function can be rewritten as Equations (35) and (36) as follows:

$$\begin{aligned} \ln L(\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{z}) = &\left(\sum_{i=1}^{n} (1 - z_i) \right) \sum_{i=1}^{n} \Gamma\left(y_i + \frac{1}{\phi}\right) - \sum_{i=1}^{n} \Gamma\left(\frac{1}{\phi}\right) - \sum_{i=1}^{n} \ln(y_i!) - \\ &\left(\frac{1}{\phi} + y_i\right) \left[\sum_{i=1}^{n} \ln\left(1 + \phi e^{\boldsymbol{x}_i^T \beta}\right) \right] + y_i \sum_{i=1}^{n} \left(\ln \phi + \ln e^{\boldsymbol{x}_i^T \beta} \right) - \\ &\sum_{i=1}^{n} \ln\left(1 - \left(1 + \phi e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}\right) . \end{aligned} \tag{35}$$

and

$$\ln L\left(\boldsymbol{\gamma}|\boldsymbol{y},\mathbf{z}\right)=\sum_{i=1}^{n}(z_i)(\boldsymbol{x}_i^T\boldsymbol{\gamma})-\sum_{i=1}^{n}\ln\left(1+e^{\boldsymbol{x}_i^T\gamma}\right). \tag{36}$$

1. To obtain parameter estimates $\boldsymbol{\beta}$ The ln likelihood function in Equation (35) is partially differentiated with respect to the parameters $\boldsymbol{\beta}$.

   The first derivative of the ln likelihood function in Equation (35) with respect to $\boldsymbol{\beta}$ can be expressed as the following equation:

$$\begin{aligned}\frac{\partial \ln L\left(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{z}\right)}{\partial(\boldsymbol{\beta})}=&\sum_{i=1}^{n}(1-z_i)y_i\sum_{i=1}^{n}\boldsymbol{x}_i^T-\left(\frac{1}{\phi}+y_i\right)\sum_{i=1}^{n}\frac{\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\boldsymbol{x}_i^T}{1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}-\\&\sum_{i=1}^{n}\frac{1}{\phi}\frac{\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^{-\frac{(\phi+1}{\phi}}\left(\phi\boldsymbol{x}_i^Te^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)}{1-\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^{-\frac{1}{\phi}}}.\end{aligned} \tag{37}$$

   Then, the second derivative of the ln likelihood function in Equation (35) with respect to $\boldsymbol{\beta}$ can be expressed as the following equation:

$$\begin{aligned}\frac{\partial^2 \ln L\left(\boldsymbol{\beta}|\boldsymbol{y},\mathbf{z}\right)}{\partial^2(\boldsymbol{\beta})}=&-\sum_{i=1}^{n}(1-z_i)\left(\left(\frac{1}{\phi}+y_i\right)\sum_{i=1}^{n}\frac{\phi e^{2\boldsymbol{x}_i^T\boldsymbol{\beta}}\left(\boldsymbol{x}_i^T\right)^2}{\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^2}\right)-\\&\sum_{i=1}^{n}\left(\frac{\phi\left(\boldsymbol{x}_i^T\right)^2e^{2\boldsymbol{x}_i^T\boldsymbol{\beta}}-(\phi+1)\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^{-\frac{(2\phi+1)}{\phi}}}{\phi^2\left(1-\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^{-\phi^{-1}}\right)^2}\right)-\\&\sum_{i=1}^{n}\left(\frac{\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^{-\frac{(\phi+1)}{\phi}}-\phi\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^{-\frac{2(\phi+1)}{\phi}}}{\phi^2\left(1-\left(1+\phi e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^{-\phi^{-1}}\right)^2}\right).\end{aligned} \tag{38}$$

   Estimation for $\boldsymbol{\beta}$ direct estimation cannot be obtained directly because the results from the equation derived from the first derivative of the ln likelihood function with respect to each parameter do not have a closed form. To address this issue, numerical iteration methods are used to facilitate calculations, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) iteration method to obtain parameter estimates (Hilbe 2011). The steps of the BFGS numerical iteration method are as follows:

   a. Determine the initial estimate of the parameter $\hat{\boldsymbol{\beta}}^{(0)}$ obtained using the OLS method and determine $\boldsymbol{H}^{(0)}$ as follows:

$$\hat{\boldsymbol{\beta}}^{(0)}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

   With

$$\boldsymbol{X}=\begin{bmatrix}1 & x_{11} & x_{12} & \dots & x_{1p}\\ 1 & x_{21} & x_{22} & \dots & x_{2p}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & x_{n2} & \dots & x_{np}\end{bmatrix}$$

and

$$\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T.$$

Then the matrix $\boldsymbol{H}^{(0)} = \boldsymbol{I}$

$$\boldsymbol{H}^{(0)} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & 0 & \ddots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

b. Forming a gradient vector **u**, which is a vector consisting of the first partial derivatives of the estimated parameters.

c. Forming the Hessian matrix $\boldsymbol{H}$, which is a matrix containing the partial derivatives of the two estimated parameters.

d. Insert the value of $\hat{\boldsymbol{\beta}}^{(0)}$ into the elements of the vector **u** and the matrix **H** to obtain the vector $\boldsymbol{u}\left(\hat{\boldsymbol{\beta}}^{(0)}\right)$ and the matrix $\boldsymbol{H}\left(\hat{\boldsymbol{\beta}}^{(0)}\right)$.

e. Perform a BFGS iteration *Quasi Newton* starting from $r = 0$ with the following equation:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} - a^{(r)} \left(\boldsymbol{H}^{(r)}\right)^{-1} \boldsymbol{u}^{(r)}.$$

With:

$\hat{\boldsymbol{\beta}}^{(r)}$ = a set of parameter estimators that converge at the rth iteration.

$a^{(r)}$=min ln $L\left(\hat{\boldsymbol{\beta}}^{(r)} - a^{(r)} \left(\boldsymbol{H}^{(r)}\right)^{-1} \boldsymbol{u}^{(r)}\right)$.

f. Calculate $\boldsymbol{s}^{(r)} = \hat{\boldsymbol{\beta}}^{(r+1)} - \hat{\boldsymbol{\beta}}^{(r)}$ and $\boldsymbol{y}^{(r)} = \hat{\boldsymbol{u}}^{(r+1)} - \hat{\boldsymbol{u}}^{(r)}$ so that the BFGS update matrix is obtained as follows:

$$\boldsymbol{H}^{(r+1)} = \boldsymbol{H}^{(r)} - \frac{\boldsymbol{H}^{(r)} \boldsymbol{s}^{(r)} \left(\boldsymbol{s}^{(r)}\right)^T \boldsymbol{H}^{(r)}}{\left(\boldsymbol{s}^{(r)}\right)^T \boldsymbol{H}^{(r)} \boldsymbol{s}^{(r)}} + \frac{\boldsymbol{y}^{(r)} \left(\boldsymbol{y}^{(r)}\right)^T}{\left(\boldsymbol{s}^{(r)}\right)^T \boldsymbol{y}^{(r)}}.$$

g. The iteration process will stop if a parameter estimate has been obtained that is convergent by satisfying $\left(\left|\hat{\boldsymbol{\beta}}^{(r+1)} - \hat{\boldsymbol{\beta}}^{(r)}\right|\right) \leq \epsilon$ where ϵ is a very small value and has been applied previously, for example 10^{-4}.

2. Similar to the parameter maximization stage for $\boldsymbol{\beta}$, the maximization stage for parameter $\boldsymbol{\gamma}$ also utilizes numerical iteration methods such as the Newton Raphson method. To obtain parameter estimates $\boldsymbol{\gamma}$, the ln likelihood function from Equation (36) is partially differentiated with respect to the parameters $\boldsymbol{\gamma}$.

The first derivative of the ln likelihood function in Equation (36) with respect to ($\boldsymbol{\gamma}$) is expressed as follow:

$$\frac{\partial \ln L\left(\boldsymbol{\gamma}|\boldsymbol{y}, \mathbf{z}\right)}{\partial(\boldsymbol{\gamma})} = \sum_{i=1}^{n} \left(z_i \boldsymbol{x}_i^T - \sum_{i=1}^{n} \frac{e^{\boldsymbol{x}_i^T \gamma} \boldsymbol{x}_i^T}{1 + e^{\boldsymbol{x}_i^T \gamma}} \right). \tag{39}$$

Then, the second derivative to $\boldsymbol{\gamma}$ can be expressed as the following equation:

$$\frac{\partial^2 \ln L\left(\boldsymbol{\gamma}|\boldsymbol{y},\mathbf{z}\right)}{\partial^2(\boldsymbol{\gamma})} = -\sum_{i=1}^{n}\left(\frac{e^{\boldsymbol{x}_i^T\gamma}\left(\boldsymbol{x}_i^T\right)^2}{\left(1+e^{\boldsymbol{x}_i^T\gamma}\right)^2}\right). \tag{40}$$

Estimation for $\boldsymbol{\gamma}$ direct estimation cannot be obtained directly because the results from the equation derived from the first derivative of the ln likelihood function with respect to each parameter do not have a closed form. To address this issue, numerical iteration methods are used to facilitate calculations, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method to obtain parameter estimates (Hilbe 2011). The steps of the BFGS numerical iteration method are as follows:

a. Determine the initial estimate of the parameter $\hat{\boldsymbol{\gamma}}^{(0)}$ obtained using the OLS method and determine $\boldsymbol{H}^{(0)}$ as follows:

$$\hat{\boldsymbol{\gamma}}^{(0)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

With

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

and

$$\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & \ldots & y_n \end{bmatrix}^T.$$

Then the matrix $\boldsymbol{H}^{(0)} = \boldsymbol{I}$

$$\boldsymbol{H}^{(0)} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & 0 & \ddots & \ldots \\ 0 & 0 & \ldots & 1 \end{bmatrix}.$$

b. Forming a gradient vector **u**, which is a vector consisting of the first partial derivatives of the estimated parameters.

c. Forming the Hessian matrix $\boldsymbol{H}$, which is a matrix containing the partial derivatives of the two estimated parameters.

d. Insert the value of $\hat{\boldsymbol{\gamma}}^{(0)}$ into the elements of the vector **u** and the matrix **H** to obtain the vector $\boldsymbol{u}\left(\hat{\boldsymbol{\gamma}}^{(0)}\right)$ and the matrix $\boldsymbol{H}\left(\hat{\boldsymbol{\gamma}}^{(0)}\right)$.

e. Perform a BFGS iteration *Quasi Newton* starting from $r = 0$ with the following equation:

$$\hat{\boldsymbol{\gamma}}^{(r+1)} = \hat{\boldsymbol{\gamma}}^{(r)} - a^{(r)}\left(\boldsymbol{H}^{(r)}\right)^{-1}\boldsymbol{u}^{(r)}.$$

With:

$\hat{\boldsymbol{\gamma}}^{(r)}$ = a set of parameter estimators that converge at the rth iteration.

$a^{(r)}$=min ln $L\left(\hat{\boldsymbol{\gamma}}^{(r)} - a^{(r)}\left(\boldsymbol{H}^{(r)}\right)^{-1}\boldsymbol{u}^{(r)}\right)$.

f. Calculate $\boldsymbol{s}^{(r)} = \hat{\boldsymbol{\gamma}}^{(r+1)} - \hat{\boldsymbol{\gamma}}^{(r)}$ and $\boldsymbol{y}^{(r)} = \hat{\boldsymbol{u}}^{(r+1)} - \hat{\boldsymbol{u}}^{(r)}$ so that the BFGS update matrix is obtained as follows:

$$\boldsymbol{H}^{(r+1)} = \boldsymbol{H}^{(r)} - \frac{\boldsymbol{H}^{(r)}\boldsymbol{s}^{(r)}\left(\boldsymbol{s}^{(r)}\right)^T\boldsymbol{H}^{(r)}}{\left(\boldsymbol{s}^{(r)}\right)^T\boldsymbol{H}^{(r)}\boldsymbol{s}^{(r)}} + \frac{\boldsymbol{y}^{(r)}\left(\boldsymbol{y}^{(r)}\right)^T}{\left(\boldsymbol{s}^{(r)}\right)^T\boldsymbol{y}^{(r)}}.$$

g. The iteration process will stop if a parameter estimate has been obtained that is convergent by satisfying $\left(\left|\hat{\boldsymbol{\gamma}}^{(r+1)} - \hat{\boldsymbol{\gamma}}^{(r)}\right|\right) \leq \epsilon$ where ϵ is a very small value and has been applied previously, for example 10^{-4}.

4. Research methods

4.1. Research data

The type of research conducted is quantitative research. The data for this study were obtained from the Health Profile of East Java Province published by the East Java Health Office and East Java in Figures published by the Central Statistics Agency of East Java. The dataset includes observations from all 38 districts and cities in East Java Province.

In this study, the response variable (y) is the number of toddler deaths due to pneumonia. The predictor variables (x) include the percentage of health service coverage for toddlers (x_1), the percentage of complete basic immunization coverage (x_2), the percentage of healthy homes (x_3), and the number of community health centers (x_4), the percentage of vitamin A supplementation (x_5), the percentage of malnourished toddlers (x_6), the percentage of exclusive breastfeeding (x_7), the percentage of low birth weight toddlers (x_8).

4.2. Research design

The research design is structured to achieve the study's objectives. The initial step involves examining overdispersion to determine whether the data exhibits this issue. One potential cause of overdispersion is an excess of zero values in the response variable (excess zeros). Therefore, an examination of excess zeros is also conducted. If the data exhibits both overdispersion and excess zeros, Poisson regression becomes unsuitable, necessitating alternative regression methods. Suitable methods include Zero Inflated Negative Binomial regression and Hurdle Negative Binomial regression.

In line with the research objectives, modeling is performed using ZINB and HNB regression on the number of toddler deaths due to pneumonia. Parameters in the ZINB and HNB regression models are estimated using the Maximum Likelihood Estimation method, and a stepwise procedure is applied to identify the best model. Furthermore, the research was extended by developing ZINB regression models with transformed variables and HNB regression models with transformed variables to improve model performance and better capture nonlinear relationships among predictors. The best model, selected based on the smallest Akaike Information Criterion (AIC) value, will then be used for prediction. The next step is to test the significance of the parameters in the regression models, conducted simultaneously using the G-test and partially using the Wald test. Finally, comprehensive residual diagnostics are performed to verify that the assumptions underlying the regression analyses are adequately met, thereby reinforcing the validity and reliability of the study's findings.

5. Results

Descriptive statistics are used to summarize the relationships between research variables and translate the data into more easily understandable and interpretable information. The de-

scriptive statistics of the variables in the data used in this study include the maximum value, minimum value, mean value, variance, and standard deviation, which are used to describe the variables in the data. Below are the results of the descriptive statistics for the research variables.

Table 1: Descriptive statistics of predictor variables

Variable	Mean	Min	Maks	Std.dev	Varians
y	1,053	0,00	8,00	1,874	3,511
x_1	80,99	5,51	100	27,665	765,354
x_2	61,98	34,72	81,09	12,129	147,117
x_3	56,43	13,22	88,65	20,763	431,121
x_4	25,5	3,00	63,00	12,83	164,5
x_5	90,82	70,30	99,30	6,932	48,052
x_6	1,161	0,00	5,60	1,076	1,157
x_7	34,21	11,16	54,83	9,476	89,798
x_8	11,734	5,530	21,610	5,143	26,45

Table 1 presents the descriptive statistics of the predictor variables used in the analysis. The results indicate significant variation among the variables, with some showing high dispersion levels. The response variable (y) has a relatively low mean, while the predictor variables (x) show a diverse range of values.

As the next step, an overdispersion assessment will be conducted to determine whether the variance of the response variable exceeds its mean, which could impact the suitability of the Poisson regression model.

A common issue in Poisson regression is that the variance of the dependent variable is greater than its mean (overdispersion), which can lead to underestimation. To address this, overdispersion is evaluated using the deviance and Pearson Chi-Square statistics to determine its presence in the data. From the calculations, the deviance value is 64.13162 with 29 degrees of freedom, and the Pearson Chi-Square value is 72.61627 with 29 degrees of freedom. The dispersion ratio, calculated by dividing the deviance by the degrees of freedom, yields a result of 2.211435. Similarly, dividing the Pearson Chi-Square value by the degrees of freedom produces a ratio of 2.504009. Since both ratios are greater than one ($2.211435 > 1$ and $2.504009 > 1$), the null hypothesis (H_0) is rejected.

This indicates that the data on toddler deaths due to pneumonia in East Java in 2022 exhibits overdispersion. Consequently, the data does not meet the Poisson regression assumption of equidispersion. One of the factors contributing to overdispersion is the high proportion of zero values in the response variable, known as excess zeros. This condition occurs when the number of zero values in the response variable is significantly higher than other values (with a zero percentage exceeding 50%). Therefore, the next step is to examine the presence of excess zeros in the response variable. Figure 1 will illustrate whether the data used exhibits an excess zeros issue, as shown below.

Based on the Figure 1 above, it shows that in this study, 22 instances of the response variable have a value of zero, constituting 57.89%. Therefore, it can be concluded that the data on toddler deaths due to pneumonia in East Java in 2022 suffers from excess zeros because the proportion of zero values exceeds 50%.

The issues of overdispersion and excess zeros render Poisson regression unsuitable, as they can result in standard errors of parameter estimates being underestimated, leading to conclusions that do not accurately reflect the data. Consequently, alternative regression models, such as Zero-Inflated Negative Binomial regression and Hurdle Negative Binomial regression, are more appropriate for modeling the number of toddler deaths due to pneumonia.

The explanatory variables proposed in this study were modelled against the response variables where the modelling used the ZINB regression model. The best model was selected

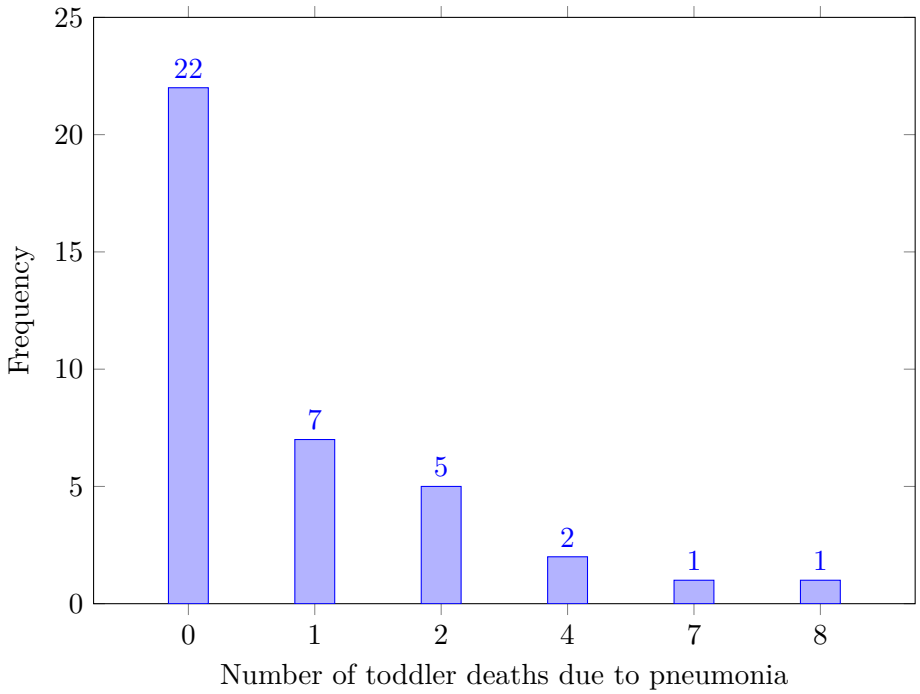

Figure 1: Frequency histogram of number of toddler deaths due to pneumonia

using the backward stepwise method. The best modelling results with the ZINB regression model backward stepwise. The parameters of the ZINB regression model, used to model the number of toddler deaths due to pneumonia, were estimated using the Maximum Likelihood Estimation method.

The results of these parameter estimations are presented in Table 2.

Table 2: Regression parameter estimation results

Variable	Estimation	Std. Error	Z-value	P-value
ln function				
β_0	0.1758	0.2469	0.712	0.47658
β_1	-0.6633	0.2305	-2.878	0.00401
β_2	0.2251	0.2622	0.858	0.39077
β_3	-0.5071	0.2061	-2.461	0.01385
β_4	-0.7647	0.3413	-2.240	0.02507
β_6	1.4548	0.3678	3.956	$7.64e$-05
β_7	-0.4631	0.1909	-2.426	0.01525
β_8	0.6047	0.1916	3.156	0.00160
logit function				
γ_0	-68.770	162.791	-0.422	0.673
γ_1	-10.267	57.762	-0.178	0.859
γ_2	-19.391	50.598	-0.383	0.702
γ_3	-1.948	43.833	-0.044	0.965
γ_4	-73.324	171.115	-0.429	0.668
γ_6	77.847	181.547	0.429	0.668
γ_7	-26.462	69.523	-0.381	0.703
γ_8	30.471	84.990	0.359	0.720
significance level	$\alpha = 0.05$			
		G test=43.229		

The equation for the Zero Inflated Negative Binomial regression model formed based on Table 2 as follows:

1. Negative Binomial State Model

$$\begin{aligned}\ln(\mu_i) =& 0.1758 - 0.6633x_{i1} + 0.2251x_{i2} - 0.5071x_{i3} - 0.7647x_{i4} + \\ & 1.4548x_{i6} - 0.4631x_{i7} + 0.6047x_{i8}.\end{aligned}$$

2. Zero inflated State Model

$$\begin{aligned}\text{logit}(\hat{\pi}_i) =& -68.770 - 10.267x_{i1} - 19.391x_{i2} - 1.948x_{i3} - 73.324x_{i4} + \\ & 77.847x_{i6} - 26.462x_{i7} + 30.471x_{i8}.\end{aligned}$$

The modeled data was developed by adding several scenarios through the application of transformations to the independent variables using square root and logarithmic functions. Subsequently, the transformed variables were used in the modeling process using the ZINB regression model with transformed variables to obtain more optimal results.

The parameter estimation results of the ZINB regression model with transformed variables are presented in Table 3.

Table 3: Regression parameter estimation results

Variable	Estimation	Std. Error	Z-value	P-value
ln function				
β_0	-0.6947	0.9701	-0.716	0.474
β_1	0.3271	2.3082	0.142	0.887
β_3	0.8671	0.6283	1.380	0.168
β_5	-1.0895	0.2638	-4.130	3.62e-05
β_7	-0.5224	0.2330	-2.242	0.025
logit function				
γ_0	-16.73	97.09	-0.172	0.863
γ_2	-29.12	147.88	-0.197	0.844
γ_4	-20.72	110.79	-0.187	0.852
γ_8	-20.71	102.62	-0.202	0.840
significance level	$\alpha = 0.05$			
		G test=24.341		

The equation for the ZINB regression model with transformed variables, formed based on Table 3, is as follows:

1. Negative Binomial State Model

$$\ln(\mu_i) = -0.6947 + 0.3271\ln(x_{i1} + 1) + 0.8671\sqrt{x_{i3}} - 1.0895x_{i5} - 0.5224x_{i7}$$

2. Zero inflated State Model

$$\text{logit}(\hat{\pi}_i) = -16.73 - 29.12x_{i2} - 20.72x_{i4} - 20.71x_{i8}$$

Based on Table 3, the results of parameter significance testing for the ZINB regression model with transformed variables using the G-test yielded a value of 24.341. The testing criterion is to reject H_0 if $G > \chi^2_{\alpha;p}$, with $p = 7$ and a significance level of $\alpha = 0.05$. Since $24.341 > 14.067$, H_0 is rejected, indicating that at least one explanatory variable significantly affects the response variable. Furthermore, partial significance testing using the Wald test, obtained by dividing the estimated coefficient by its standard error (equivalent to the Z-value), was

also performed. The decision rule is to reject H_0 if $|W_j| > Z_{\frac{\alpha}{2}}$ or if the p-value $< \alpha = 0.05$, implying that the predictor variable has a significant influence on the response variable. Based on Table 3, the predictor variables that meet the testing criteria and significantly influence the number of toddler deaths due to pneumonia are the percentage of vitamin A supplementation (x_5) and the percentage of exclusive breastfeeding (x_7).

The count component shows that both x_5 and x_7 have negative, statistically significant effects ($p < 0.05$), indicating that increases in these variables are associated with a decrease in the expected number of toddler deaths. Meanwhile, the transformed variables $\ln(x_{i1} + 1)$ and $\sqrt{x_3}$ have positive but insignificant effects. In the zero-inflation part, the variables x_2, x_4, and x_8 have large negative but statistically insignificant coefficients, suggesting that they do not significantly explain the structural zeros in the data. Compared with the ZINB regression model selected by backward stepwise selection, the ZINB regression model with transformed variables achieves better estimation stability and model fit. This indicates that applying logarithmic and square-root transformations to selected predictors improves the model's ability to explain variations in toddler deaths due to pneumonia.

Apart from Zero-Inflated Negative Binomial regression, this research also uses Hurdle Negative Binomial regression to model the number of toddler deaths due to pneumonia, with parameters estimated using Maximum Likelihood Estimation. The best model was selected using the backward stepwise method. The results of these parameter estimations are presented in Table 4

Table 4: Regression parameter estimation results

Parameter	Estimation	Std. Error	Z-value	P-value
ln function				
β_0	0.3794	0.3448	1.100	0.271210
β_1	-0.3111	0.2327	-1.337	0.181325
β_2	-0.8636	0.2594	-3.329	0.000871
β_5	-0.7410	0.2301	-3.221	0.001277
β_7	-0.7605	0.2900	-2.623	0.008727
logit function				
γ_0	-0.38932	0.36747	-1.059	0.2894
γ_1	-0.68082	0.41242	-1.651	0.0988
γ_2	0.89387	0.46230	1.934	0.0532
γ_5	-0.23144	0.39280	-0.589	0.5557
γ_7	0.08034	0.40067	0.201	0.8411
significance level	$\alpha = 0.05$			
		G test=22.334		

The equation for the Hurdle Negative Binomial regression model formed based on Table 4 as follows:

1. Zero Hurdle State Model

$$\text{logit}(\hat{\pi}_i) = 0.3794 - 0.3111x_{i1} - 0.8636x_{i2} - 0.7410x_{i5} - 0.7605x_{i7}$$

2. Truncated Negative Binomial State Model

$$\ln(\mu_i) = -0.38932 - 0.68082x_{i1} + 0.89387x_{i2} - 0.23144x_{i5} + 0.08034x_{i7}$$

After obtaining the best model using the backward stepwise method, the modeling was further developed by applying transformations to the independent variables using square root and logarithmic functions. The transformed variables were then used in the modeling process using the HNB regression model with transformed variables to obtain more optimal results, and the parameter estimation results are presented in Table 5.

Table 5: Regression parameter estimation results

Variable	Estimation	Std. Error	Z-value	P-value
ln function				
β_0	-8.8728	3.7866	-2.343	0.01912
β_1	18.9359	6.7725	2.796	0.00517
β_3	-0.8905	1.0557	-0.844	0.39894
β_6	-1.1702	1.0216	-1.145	0.25200
β_7	-1.8852	0.8544	-2.206	0.02735
logit function				
γ_0	-1.09971	0.58776	-1.871	0.0613
γ_2	1.47598	0.78035	1.891	0.0586
γ_4	0.09751	0.43428	0.225	0.8223
γ_8	1.04296	0.58072	1.796	0.0725
significance level	$\alpha = 0.05$			
		G test=22.079		

The equation for the HNB regression model with transformed variables, formed based on Table 5, is as follows:

1. Zero Hurdle State Model

$$\ln(\mu_i) = -8.8728 + 18.9359\ln(x_{i1}+1) - 0.8905\sqrt{x_{i3}} - 1.1702x_{i6} - 1.8852x_{i7}$$

2. Truncated Negative Binomial State Model

$$\text{logit}(\hat{\pi}_i) = -1.09971 + 1.47598x_{i2} + 0.09751x_{i4} + 1.04296x_{i8}$$

Based on Table 5, the results of parameter significance testing for the HNB regression model with transformed variables using the G-test yielded a value of 22.079. The testing criterion is to reject H_0 if $G > \chi^2_{\alpha;p}$, with $p = 7$ and a significance level of $\alpha = 0.05$. Since $22.079 > 14.067$, H_0 is rejected, indicating that at least one explanatory variable significantly affects the response variable.

Furthermore, partial significance testing using the Wald test, obtained by dividing the estimated coefficient by its standard error (equivalent to the Z-value), was also performed. The decision rule is to reject H_0 if $|W_j| > Z_{\alpha/2}$ or if the p-value $< \alpha = 0.05$, implying that the predictor variable has a significant influence on the response variable.

The count component shows that both $\ln(x_{i1}+1)$ and x_7 have statistically significant effects ($p < 0.05$), where $\ln(x_{i1}+1)$ has a positive effect and x_7 has an adverse effect, indicating that an increase in health service coverage for toddlers is associated with an increase in the expected number of toddler deaths. In contrast, an increase in exclusive breastfeeding is associated with a decrease in the expected number of toddler deaths. Meanwhile, the transformed variable $\sqrt{x_3}$ and variable x_6 have adverse but statistically insignificant effects.

In the zero-hurdle part, the variables x_2, x_4, and x_8 have positive but statistically insignificant coefficients, suggesting that they do not significantly explain the excess zeros in the data.

Based on Table 5, the predictor variables that meet the testing criteria and significantly influence the number of toddler deaths due to pneumonia are the percentage of health service coverage for toddlers (x_1) and the percentage of exclusive breastfeeding (x_7).

Compared to the HNB regression model selected using the backward stepwise method, the HNB regression model with transformed variables provides better model performance in terms of parameter stability and model fit (log-likelihood = −20.441). This indicates that applying logarithmic and square-root transformations to selected predictors improves the model's explanatory power in describing variations in toddler deaths due to pneumonia.

After obtaining the Zero Inflated Negative Binomial, ZINB regression model with transformed variables, Hurdle Negative Binomial, and HNB regression model with transformed variables, the selection of the best model will be conducted using the Akaike Information Criterion (AIC) values. The best model is the one with the smallest AIC value. The best model is selected by comparing models that are more appropriate for the research data.

In this research, the models used are the Zero Inflated Negative Binomial, ZINB regression model with transformed variables, the Hurdle Negative Binomial, HNB regression model with transformed variables, so that the two models will be compared based on the AIC value of the ZINB regression model and the AIC value of the HNB regression model, to find out which method model is most suitable for researching the number of toddler deaths due to pneumonia in East Java in 2022. AIC values shown in Table 6.

Table 6: AIC value of ZINB and HNB

Regression Model	AIC Value
Zero Inflated Negative Binomial (ZINB)	95.94633
Hurdle Negative Binomial (HNB)	104.8086
ZINB regression model with transformed variables	58.63682
HNB regression model with transformed variables	60.88117

Based on the AIC value obtained in Table 6, the ZINB regression model with transformed variables has a smaller AIC value. The AIC value is in line with the deviation value of the model. The smaller the deviation value, the smaller the error rate produced by the model, so that the model obtained is more precise. Therefore, the best model is the model with the smallest AIC value, namely the ZINB regression model with transformed variables with an AIC value of 58.63682. So it can be concluded that the ZINB regression model can model data on the number of toddler deaths due to pneumonia in East Java.

After obtaining the best model in this research, namely the ZINB regression model with transformed variables. Then interpret the results of the predictor variables that have a significant influence on the predictor variables, namely the coefficient value x_5 is $3.62e-05$, so it can be interpreted that if there is an increase in the percentage of vitamin A supplementation by 1% it can lead to a decrease in the number of toddler deaths due to pneumonia by $e^{(3.62e-05)} = 1.0000362$ times. The coefficient value x_7 is 0.025, so it can be interpreted that if there is an increase in the percentage of exclusive breastfeeding by 1%, it can cause a decrease in the number of cases of death of toddlers due to pneumonia by $e^{(0.025)} = 1.025315$ times.

The next step is to perform a residual diagnostic using the Shapiro-Wilk test. For the ZINB regression model with transformed variables, the test yields $W = 0.89421$ and p-value is 0.01165, while for the HNB regression model with transformed variables, it yields $W = 0.90442$ and p-value is 0.01968.

Since the W values models are less than 1 and the p-values are smaller than 0.05 (leading to the rejection of H_0), it can be concluded that the residuals in both the ZINB regression model with transformed variables and the HNB regression model with transformed variables do not follow a normal distribution. To further support this conclusion, a Q–Q plot will be added to visually assess the normality of the residuals, as shown in the following figure.

Figure 2 presents the Q–Q plots of the residuals for the ZINB and HNB regression models with transformed variables. In both plots, the points deviate noticeably from the reference line, particularly at the upper and lower tails, indicating that the residuals are not normally

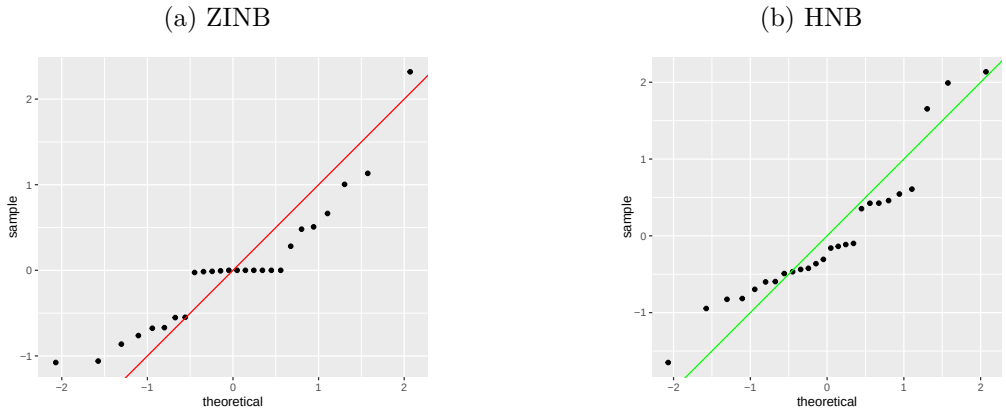

Figure 2: Q–Q plots of the residuals for the ZINB (left) and HNB (right) regression models with transformed variables

distributed. These visual patterns are consistent with the results of the Shapiro–Wilk tests, where the W values were less than 1 and the p-values were smaller than 0.05, leading to the rejection of the null hypothesis of normality. Therefore, it can be concluded that the residuals in both the ZINB and HNB regression models with transformed variables do not follow a normal distribution.

6. Discussion

Based on the results, it is evident that the best model is the ZINB regression model with transformed variables. Factors that significantly influence the number of toddler deaths from pneumonia are the percentage of vitamin A supplementation (x_5) and the percentage of exclusive breastfeeding (x_7).

The percentage of vitamin A supplementation (x_5) significantly affects pneumonia cases in toddlers, as supported by the study conducted by Saputri and Purhadi (2022). An increase in vitamin A supplementation coverage indicates that more toddlers receive essential nutrients that strengthen their immune systems. Vitamin A plays a crucial role in maintaining respiratory health and enhancing the body's ability to fight infections, including pneumonia. As the coverage of vitamin A supplementation increases, the risk of pneumonia in toddlers decreases, leading to lower morbidity and mortality rates.

The percentage of exclusive breastfeeding (x_7) plays a significant role in reducing pneumonia cases, as highlighted by research from Hartati, Nurhaeni, and Gayatri (2012). Increased rates of exclusive breastfeeding help toddlers build stronger immune systems, thereby reducing the risk of infections and, consequently, mortality. Breast milk contains essential components such as hormones, antibodies, nutrients, and antioxidants that support a toddler's development and immune system.

The findings of this study provide valuable reference material for future research. The significant factors identified as influencing toddler deaths due to pneumonia can be used as a basis for preventing and reducing mortality in East Java. Based on this analysis, local governments can take strategic actions to improve factors that reduce pneumonia cases and control those that exacerbate them.

7. Conclusion

ZINB regression model with transformed variables is the best model in this research, because it has the smallest AIC value, namely 58.63682. The ZINB regression model with transformed

variables can be formulated as follows:

1. Negative Binomial State Model

$$\ln(\mu_i) = -0.6947 + 0.3271 \ln(x_{i1} + 1) + 0.8671\sqrt{x_{i3}} - 1.0895x_{i5} - 0.5224x_{i7}$$

2. Zero inflated State Model

$$\text{logit}(\hat{\pi}_i) = -16.73 - 29.12x_{i2} - 20.72x_{i4} - 20.71x_{i8}$$

Based on the G test, it was found that the ZINB model was appropriate to use for this data. Meanwhile, in the Wald test, it was found that factors that had a significance influence the number of toddler deaths due to pneumonia are the percentage of vitamin A supplementation (x_5) and the percentage of exclusive breastfeeding (x_7).

References

Agresti A (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons. ISBN 978-1-118-73003-4.

Bilgic A, Florkowski WJ (2007). "Application of a Hurdle Negative Binomial Count Data Model to Demand for Bass Fishing in the Southeastern United States." *Journal of Environmental Management*, **83**(4), 478–490. doi:10.1016/j.jenvman.2006.10.009.

Blasco-Moreno A, Pérez-Casany FJ, Puig M, Morante X, Baldó-Serra M (2019). "What Does a Zero Mean? Understanding False, Random and Structural Zeros in Ecology." *Methods in Ecology and Evolution*, **10**(7), 949–959. doi:10.1111/2041-210x.13185.

Cameron AC, Trivedi PK (2014). *Regression Analysis of Count Data*. Cambridge University Press. ISBN 9781139013567. doi:10.1017/CBO9781139013567.

Famoye F, Singh KP (2006). "Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data." *Journal of Data Science*, **4**(1), 117–130. doi:10.6339/JDS.2006.04(1).257.

Garay AM, Hashimoto EM, Ortega EMM, Lachos VH (2011). "On Estimation and Influence Diagnostics for Zero-Inflated Negative Binomial Regression Models." *Computational Statistics & Data Analysis*, **55**(3), 1304–1318. doi:10.1016/j.csda.2010.09.019.

Greene W (2008). "Functional Forms for the Negative Binomial Model for Count Data." *Economics Letters*, **99**(3), 585–590. doi:10.1016/j.econlet.2007.10.015.

Hall DB (2000). "Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study." *Biometrics*, **56**(4), 1030–1039. doi:10.1111/j.0006-341x.2000.01030.x.

Hartati S, Nurhaeni N, Gayatri D (2012). "Faktor Risiko Terjadinya Pneumonia pada Anak Balita." *Jurnal Keperawatan Indonesia*, **15**(1), 13–20. doi:10.7454/jki.v15i1.42.

Hilbe JM (2011). *Negative Binomial Regression*. Cambridge University Press. ISBN 9780511973420. doi:10.1017/CBO9780511973420.

Ma L, Yan X, Wei C, Wang J (2016). "Modeling the Equivalent Property Damage Only Crash Rate for Road Segments using the Hurdle Regression Framework." *Analytic Methods in Accident Research*, **11**, 48–61. doi:10.1016/j.amar.2016.07.001.

Minami M, Lennert-Cody CE, Gao W, Román-Verdesoto M (2007). "Modeling Shark Bycatch: The Zero-Inflated Negative Binomial Regression Model with Smoothing." *Fisheries Research*, **84**(2), 210–221. doi:10.1016/j.fishres.2006.10.019.

Myers RH, Montgomery DC, Vining GG, Robinson TJ (2012). *Generalized Linear Models: With Applications in Engineering and the Sciences*. John Wiley & Sons. ISBN 978-0-470-45463-3.

Puig P, Valero J (2006). "Count Data Distributions: Some Characterizations with Applications." *Journal of the American Statistical Association*, **101**(473), 332–340. doi:10.1198/016214505000000718.

Saengthong P, Bodhisuwan W, Thongteeraparp A (2015). "The Zero Inflated Negative Binomial-Crack Distribution: Some Properties and Parameter Estimation." *Songklanakarin Journal of Science and Technology*, **37**(6), 701–711.

Saffari SE, Adnan R, Greene W (2012). "Hurdle Negative Binomial Regression Model with Right Censored Count Data." *SORT: Statistics and Operations Research Transactions*, **36**(2), 181–194.

Saputri VA, Purhadi P (2022). "Pemodelan Faktor-Faktor yang Mempengaruhi Kasus Pneumonia pada Balita di Provinsi Jawa Barat dengan Metode Geographically Weighted Generalized Poisson Regression." *Inferensi*, **5**(2), 91. doi:10.12962/j27213862.v5i2.12619.

Saputro MIA, Qudratullah MF (2021). "Estimation of Zero-Inflated Negative Binomial Regression Parameters using the Maximum Likelihood Method." In *Proceeding International Conference on Science and Engineering*, volume 4, pp. 240–254.

Sharker S, Balbuena L, Marcoux G, Feng CX (2020). "Modeling Socio-Demographic and Clinical Factors Influencing Psychiatric Inpatient Service Use: A Comparison of Models for Zero-Inflated and Overdispersed Count Data." *BMC Medical Research Methodology*, **20**, 232. doi:10.1186/s12874-020-01112-w.

Yıldırım G, Kaçıranlar S, Yıldırım H (2022). "Poisson and Negative Binomial Regression Models for Zero-Inflated Data: An Experimental Study." *Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics*, **71**(2), 601–615. doi:10.31801/cfsuasmas.988880.

Affiliation:

A'yunin Sofro
Department of Actuarial Science
Faculty of Mathematics and Natural Sciences
Universitas Negeri Surabaya
East Java 60231, Indonesia
E-mail: ayuninsofro@unesa.ac.id

Austrian Journal of Statistics
published by the Austrian Society of Statistics
http://www.ajs.or.at/
http://www.osg.or.at/
Volume 55, Issue 3
doi:10.17713/ajs.v55.i3.1907
Submitted: 2024-03-15
Accepted: 2026-06-01