JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e-ISSN: 2614-8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 21, No. 1, September 2024, pp. 176-189
DOI: 10.20956/j.v21i1.34512

# The Robust Negative Binomial Regression Model on Under-five Mortality due to Pneumonia in the Province of East Java

**Anggun Yuliarum Qur'ani[1,2], Chandra Sari Widyaningrum[2], Sa'adatur Rohimiyah[3]**

*[1]Mathematics Study Programme, Udayana University, Indonesia*
*[1,2]Master of Mathematics Programme, Universitas Gadjah Mada, Indonesia*
*[3]Department of Nursing, Poltekkes Kemenkes Surabaya, Indonesia*
***Email**: [1]anggunyuliarum@gmail.com; [2]chandrasari1231@gmail.com;*
*[3]miasanmia3011@gmail.com*

**Abstract**

Robust Negative Binomial regression model (RNBR) is a modelling method to overcome a problem if there are outliers and overdispersion in the data. Outliers are data points that are significantly different from other data. Outliers have a significant effect on modelling to the resulting model. Furthermore, overdispersion is indicated by the presence of too large values of Pearson statistics. In this study, the RNBR model was used to determine the factors of the toddler immune variable at post neonatal age that significantly influenced the number of under-five deaths caused by pneumonia in East Java Province. Based on the modelling obtained, it shows that the RNBR model provides more robust results in handling outlier and overdispersion problems. This can be seen from the AIC value of the RNBR model is smaller than the AIC of the Poisson regression model. In addition, $s^2_{PR}$ and $s^2_{RBNR}$, which are measures of the influence of outliers on the model, decreased from 1 for the Poisson regression model to around 0.42 for the RNBR model.

**Keywords:** Robust Negative Binomial Regression (RNBR), Immunity, Toddler Mortality, Pneumonia

## 1. INTRODUCTION

Not all observed data are free from outliers. The presence of outliers in the data usually occurs due to recording/inputting errors, measurement scale errors, or indeed certain unusual events. This is very crucial in statistical data analysis which results in some assumptions on model errors not being met, including on model errors generated from linear regression analysis. Linear regression acts as a tool for analytical problem solving to determine the relationship between variables [25]. To handle the presence of outliers in the data, Robust regression offers a solution in accommodating outliers without having to eliminate them. In addition to the problem of outliers, there is data obtained from the counting process experiencing overdispersion, which is a situation

where the variance value of the data is greater than the mean value of the data. One of the causes of overdispersion is too many zeros (excess zeros) in the dependent variable [22]. One method to overcome overdispersion in regression models is Negative Binomial regression. If there are 2 problems, namely the presence of outliers and overdispersion, then Robust Negative Binomial Regression (RNBR) is one method that can provide a solution to the problem.

[1] modelled patient falls data using Negative Binomial regression with robust inference to evaluate the effectiveness of exercise interventions in reducing the number of falls among elderly people with Parkinson's disease. One of the results obtained from his research is the benefit of robust estimators for inference and as a diagnostic tool. [32] conducted a comparison of 3 methods, namely Poisson regression, Negative Binomial regression, and quasi-Poisson robust regression using the Bayesian approach. His research found that the quasi-Poisson robust regression model is more robust than other regression models.

The United Nations Children's Fund (UNICEF) report shows that pneumonia will be the infectious disease that contributes the most deaths to children under five years old in the world in 2021. The number reached 725,557 cases in 2021 [28]. Pneumonia is an acute infection of the lung tissue (alveoli) caused by bacteria, viruses or fungi. The typical symptoms of this disease are increased breathing frequency and shortness of breath due to sudden lung inflammation [17, 25]. Health education was conducted by [25] on Tuesday, 13 December 2022, in the Parung Serab Ciledug urban village area involving 20 participants. The health counselling that was conducted to parents showed an increase in knowledge about recognising the signs of pneumonia and handling emergencies at home. [4] conducted a study on risk factors associated with the incidence of pneumonia among under-fives in Kupang City using logistic regression. There was a correlation between nutritional status, immunisation status, and exclusive breastfeeding with the incidence of pneumonia. In addition, [11] used Geographically Weighted Negative Binomial Regression (GWNBR) to model pneumonia cases in East Java in 2021. His study resulted in 9 groups based on significant variables with the influencing factors in all districts/cities being population density and percentage of under-five health service coverage.

The number of pneumonia cases affecting children under five years of age in East Java Province in 2022 was 92,118. This was the highest number of cases among Malaria, Lung TB, and Leprosy [7]. And there were 145 cases of under-five deaths due to pneumonia in the same year [10]. Based on the explanation that has been described, researchers want to find out the factors of the under-five immune variable that significantly influence under-five deaths caused by pneumonia in East Java Province using the Robust Negative Binomial Regression model.

## 2. METHODS

### Robust Negative Binomial Regression (RNBR)

An outlier is a data point that is significantly different from the rest of the data [2]. Outliers have a significant effect on modelling and thus the resulting model. This is a problem that often arises in data analysis. These outliers usually occur due to recording/inputting errors, measurement scale errors, or indeed certain unusual events. Outliers can appear in both discrete and continuous data types. There is a special case in discrete data obtained from the counting process will obtain zero, and non-zero observation values. When many zero values are obtained from observations on the dependent variable, this is one of the causes of overdispersion, which is a state of data variance greater than the mean data [22].

The presence of overdispersion in many situations is clearly indicated by the presence of excessively large values of the Pearson statistic or goodness-of-fit deviations in the full model. Overdispersion in classical Poisson regression by calculating the sum of squares of the $n$ standardised error, $\sum_{i=1}^{n} z_i^2$, where $z_i = \frac{y_i - \hat{y}_i}{sd(\hat{y}_i)}$, and compared with the $\chi^2_{n-k}$ distribution,

so that the estimated overdispersion ratio

$$\text{the estimated overdispersion} = \frac{1}{n-k}\sum_{i=1}^{n} z_i^2 \tag{2.1}$$

is a summary of overdispersion in the data compared to the fitted model and an estimated value greater than 1 indicates an overdispersion model[12, 25]. Given observed values of the dependent variable $Y$ with $n$ observations $y_i$, and $n$ observed values of $p$ independent variables $X$, $x_{ik}$ with $i = 1,2,\dots,n$; $k = 1,2,\dots,p$. Assuming $Y \sim \text{Poisson}(\mu_i)$, then the probability mass function of $Y$

$$p(y_i;\mu_i) = P(Y = y_i;\mu_i) = \frac{e^{-\mu_i}\mu_i{}^{y_i}}{y_i!}; \quad y_i = 0,1,2\ldots \tag{2.2}$$

where mean and variance, $\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu_i$. The relationship of $y_i$ with $x_{ik}$ is as follows.

$$\mathrm{E}(Y = y_i) = \mu_i = e^{x_i{}^T\boldsymbol{\beta}} \tag{2.3}$$

where $\boldsymbol{x_i} = \{x_{i0}, x_{i1}, x_{i2}, \dots, x_{ip}\}^T$, and $\beta = \{\beta_0, \beta_1, \beta_2, \dots, \beta_p\}$. The models from equations (2.2) and (2.3) are known as Poisson regression models or log-linear models [20]. Poisson model parameters are usually estimated using the Maximum Likelihood Estimator (MLE). Define the likelihood function of (2.2) as follows [9, 20, 34]

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i{}^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-\left(e^{x_i{}^T\boldsymbol{\beta}}\right)}\left(e^{x_i{}^T\boldsymbol{\beta}}\right)^{y_i}}{y_i!} \tag{2.4}$$

So that the logarithmic form of both sides of equation (2.4) is obtained

$$l(\beta) = \sum_{i=1}^{n}(y_i \ln \mu_i - \mu_i - \ln y_i!) = \sum_{i=1}^{n}\left(y_i(x_i{}^T\boldsymbol{\beta}) - e^{x_i{}^T\boldsymbol{\beta}} - \ln y_i!\right) \tag{2.5}$$

by substituting $x_i{}^T\boldsymbol{\beta} = \beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}$ in equation (2.5) we get

$$l(\beta) = \sum_{i=1}^{n}\left\{y_i\left(\beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}\right) - e^{\beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}} - \ln y_i!\right\} \tag{2.6}$$

Then the MLE for $\beta_0, \beta_1, \dots, \beta_k$ are obtained from the first derivative and second derivative of equation (2.6) against $\beta$ which can be written in the following equation.

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta_0} &= \sum_{i=1}^{n}\left(y_i - e^{\beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}}\right) \\
\frac{\partial l(\beta)}{\partial \beta_1} &= \sum_{i=1}^{n}\left(y_i - e^{\beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}}\right) x_{i1} \\
&\dots \\
\frac{\partial l(\beta)}{\partial \beta_k} &= \sum_{i=1}^{n}\left(y_i - e^{\beta_0 + \sum_{k=1}^{p}\beta_k x_{ik}}\right) x_{ip}
\end{aligned} \tag{2.7}$$

$$\frac{\partial^2 l(\beta)}{\partial^2 \beta_k \beta_j} = -\sum_{i=1}^{n} e^{x_i{}^T\boldsymbol{\beta}} x_{ik} x_{ij} \tag{2.8}$$

Equations (2.7) and (2.8) are non-linear functions with respect to $\beta_0, \beta_1, \dots, \beta_k$, so to estimate the values of $\boldsymbol{\beta}$ it is necessary to use an iterative algorithm, one of which is the Newton-Raphson algorithm [20, 29].

Given a random variable $Y \sim NB(\mu_i)$, and an unknown but constant parameter $\phi$. Then the probability mass function $Y$ can be written with the following equation [19].

$$\mathrm{p}(y_i, \mu_i, \phi) = \frac{\Gamma(y_i+\phi\mu_i)\phi^{\phi\mu_i}}{y_i!\,\Gamma(\phi\mu_i)(1+\phi)^{y_i+\phi\mu_i}} \tag{2.9}$$

Suppose $\phi\mu_i = \frac{1}{\alpha}$, then equation (2.9) can be written as follows.

$$\begin{aligned}\mathrm{p}(y_i, \mu_i, \alpha) &= \frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)\left(\frac{1}{\alpha\mu_i}\right)^{\frac{1}{\alpha}}}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)\left(1+\frac{1}{\alpha\mu_i}\right)^{y_i+\frac{1}{\alpha}}} \\ &= \frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)\left(\frac{1}{\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(1+\frac{1}{\alpha\mu_i}\right)^{-\frac{1}{\alpha}}\left(1+\frac{1}{\alpha\mu_i}\right)^{-y_i}}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)} \\ &= \frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)}(1+\alpha\mu_i)^{-\frac{1}{\alpha}}\left(1+\frac{1}{\alpha\mu_i}\right)^{-y_i}\end{aligned} \tag{2.10}$$

where $\alpha > 0$ is the overdispersion parameter and is constant. And illustrating $x_{ik}$ can be substituted in the Negative Binomial distribution regression model based on the following relationship [1, 16, 19, 20].

$$\log \mu_i = \sum_{k=1}^{p} x_{ik}\beta_k = \boldsymbol{x}_i\boldsymbol{\beta} \tag{2.11}$$

Let $\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)} = \frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma(y_i+1)\,\Gamma\left(\frac{1}{\alpha}\right)} = \prod_{k=1}^{y_i}\left(y_i+\frac{1}{\alpha}+k\right) = \alpha^{-y_i}\prod_{k=1}^{y_i}(\alpha y_i+\alpha k+1)$. Model estimation is performed using the Maximum Likelihood Estimator (MLE), so that the log-likelihood function of equation (2.11) is obtained.

$$\begin{aligned}l(y_i, \mu_i, \alpha) &= \log\prod_{i=1}^{n}\left(\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{y_i!\,\Gamma\left(\frac{1}{\alpha}\right)}(1+\alpha\mu_i)^{-\frac{1}{\alpha}}\left(1+\frac{1}{\alpha\mu_i}\right)^{-y_i}\right) \\ &= \sum_{i=1}^{n}\left(-\log y_i! + \log\left(\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)}\right) - \frac{1}{\alpha}\log(1+\alpha\mu_i) - y_i\log\left(1+\frac{1}{\alpha\mu_i}\right)\right) \\ &= \sum_{i=1}^{n}\left(-\log y_i! + \log\left(\frac{\Gamma\left(y_i+\frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)}\right) - \frac{1}{\alpha}\log(1+\alpha\mu_i) + y_i\log\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)\right) \\ &= \sum_{i=1}^{n}\left(-\log y_i! + \mathrm{dlg}\left(\mathrm{y_i}, \frac{1}{\alpha}\right) - \frac{1}{\alpha}\log(1+\alpha\mu_i) + y_i\log(\alpha\mu_i) - \right. \\ &\qquad \left. y_i\log(1+\alpha\mu_i)\right)\end{aligned} \tag{2.12}$$

where $\mathrm{dlg}\left(\mathrm{y_i}, \frac{1}{\alpha}\right) = \log\Gamma\left(y_i+\frac{1}{\alpha}\right) - \log\Gamma\left(\frac{1}{\alpha}\right)$. Next, from equation (2.12), the first derivative of $\boldsymbol{\beta}$ and $\alpha$ are sought, thus obtained:

$$\Psi_\beta(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \alpha) = \frac{\partial l(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \alpha)}{\partial \beta_k} = \frac{\partial l(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \alpha)}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_k} = \sum_{i=1}^{n}\left(\frac{y_i-\mu_i}{\mu_i(1+\alpha\mu_i)}\right)\frac{\partial \mu_i}{\partial \beta_k}\boldsymbol{x}_i \tag{2.13}$$

$$\Psi_\alpha(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \alpha) = \frac{\partial l(y_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \alpha)}{\partial \alpha} = \sum_{i=1}^{n}\left\{\mathrm{ddg}\left(\mathrm{y_i}, \frac{1}{\alpha}\right) - \frac{1}{\alpha^2}\log(1+\alpha\mu_i) + \frac{\mu_i\left(y_i+\frac{1}{\alpha}\right)}{1+\alpha\mu_i}\right\} \tag{2.14}$$

And for the second derivative of equation (2.12) against $\boldsymbol{\beta}$ and $\alpha$ are

$$\frac{\partial^2 l(y_i,\boldsymbol{x}_i,\boldsymbol{\beta},\alpha)}{\partial\beta_j\beta_k} = \sum_{i=1}^{n}\left(\frac{1+\alpha y_i}{(1+\alpha\mu_i)^2}\right)\mu_i x_{ij}x_{ik} \tag{2.15}$$

$$\frac{\partial^2 l(y_i,\boldsymbol{x}_i,\boldsymbol{\beta},\alpha)}{\partial\alpha^2} = \sum_{i=1}^{n}\left\{\mathrm{dtg}\left(\mathrm{y_i},\frac{1}{\alpha}\right) - \frac{2}{\alpha^3}\log(1+\alpha\mu_i) + \frac{2\mu_i}{\alpha^2(1+\alpha\mu_i)} + \frac{{\mu_i}^2\left(y_i+\frac{1}{\alpha}\right)}{(1+\alpha\mu_i)^2}\right\} \tag{2.16}$$

$$\frac{\partial^2 l(y_i,\boldsymbol{x}_i,\boldsymbol{\beta},\alpha)}{\partial\beta_k\partial\alpha} = -\sum_{i=1}^{n}\left(\frac{y_i-\mu_i}{(1+\alpha\mu_i)^2}\right)\boldsymbol{x}_i \tag{2.17}$$

Where the ddg and dtg functions are di-gamma derived functions and tri-gamma derived functions which in detail can be seen in the following equations [16, 20].

$$\begin{aligned}\mathrm{ddg}\left(\mathrm{y_i},\frac{1}{\alpha}\right) &= \frac{\partial \mathrm{dlg}\left(\mathrm{y_i},\frac{1}{\alpha}\right)}{\partial\alpha}\\ &= \sum_{k=1}^{y_i}\frac{y_i-k}{\alpha y_i-\alpha k+1}\end{aligned} \tag{2.18}$$

$$\begin{aligned}\mathrm{dtg}\left(\mathrm{y_i},\frac{1}{\alpha}\right) &= \frac{\partial^2 \mathrm{dgg}\left(\mathrm{y_i},\frac{1}{\alpha}\right)}{\partial\alpha^2}\\ &= \sum_{k=1}^{y_i}\frac{-(y_i-k)^2}{(\alpha y_i-\alpha k+1)^2}\end{aligned} \tag{2.19}$$

Just like estimating the Poisson regression model parameters, estimating the Negative Binomial regression model parameters $\boldsymbol{\beta}$ and $\alpha$ in equations (2.13) - (2.17) using Fisher's algorithm.

Robust regression analysis was developed as an improvement to least squares estimation in the presence of outliers, and to provide information on which observations are valid, and whether they should be discarded. The main objective of robust regression analysis is to fit a model that represents the information in most of the data [6]. One of the popular methods used in Robust regression analysis is the M-Estimator. Given $\varepsilon_i$ s the error of the $i$ −th data, which is the difference between the value of the $i$ −th observation and its estimated value. Using the least squares method (OLS) by minimising the value of $\varepsilon_i$ using different error functions, namely

$$\min_{\hat{\boldsymbol{\beta}}}\sum_{i=1}^{n}\rho(\varepsilon_i) \tag{2.20}$$

where $\rho$ m is a positive definite function, and single-valued at zero. Deriving equation (2.20) on $\boldsymbol{\beta}_k = (\beta_0,\beta_1,\beta_2,\dots,\beta_p)$ are obtained.

$$\sum_{i=1}^{n}\frac{\partial\rho(\varepsilon_i)}{\partial\varepsilon_i}\frac{\partial\varepsilon_i}{\partial\beta_k} = \mathbf{0} \tag{2.21}$$

Let $\varphi(\varepsilon_i) = \frac{\partial\rho(\varepsilon_i)}{\partial\varepsilon_i}$ be the derivative function of $\rho$, and $\boldsymbol{x}_i = (x_{i1},x_{i2},\dots,x_{ip})$, also $\mathbf{0} = (0,0,\dots,0)$. Therefore, equation (2.21) can also be written as

$$\sum_{i=1}^{n}\varphi(\varepsilon_i)\frac{\partial\varepsilon_i}{\partial\beta_k} = \mathbf{0} \tag{2.22}$$

Define the weighting function $w(\varepsilon_i) = \frac{\varphi(\varepsilon_i)}{\varepsilon_i}$, then equation (2.22) becomes

$$\sum_{i=1}^{n}w(\varepsilon_i)\varepsilon_i\frac{\partial\varepsilon_i}{\partial\beta_k} = \mathbf{0} \tag{2.23}$$

And to obtain the solution to equation (2.23), the iteratively re-weighted least squares (IRWLS) method is used which leads to a rather fast calculation and stable calculation [1, 6] as follows.

$$\min_{\hat{\boldsymbol{\beta}}}\sum_{i=1}^{n}w\left({\varepsilon_i}^{(k-1)}\right){\varepsilon_i}^2 \tag{2.24}$$

where $k$ is the number of iterations used so that the value converges. Thus, the Robust Negative Binomial Regression (RNBR) model estimation using weighted MLE for the $\boldsymbol{\beta}$ parameter from equation (2.13) is as follows.

$$\sum_{i=1}^{n}\left[\boldsymbol{w}^{*}(y_i,\mu_i)\Psi_{\beta}(y_i,\boldsymbol{x}_i,\boldsymbol{\beta},\alpha)w(\boldsymbol{x}_i)-a_i(\boldsymbol{\beta},w^{*},w)\right]=\mathbf{0} \tag{2.25}$$

where $\boldsymbol{w}^{*}(y_i,\mu_i)\in[0,1]$ is the independent variable weighting function written as $\frac{\varphi(\varepsilon_i)}{\varepsilon_i}$, $w(\boldsymbol{x}_i)$ is the function $w(\boldsymbol{x}_i)$ is the weight to limit the impact of $\boldsymbol{x}_i$ that may occur, and $a_i(\boldsymbol{\beta},w^{*},w)=E\left(\boldsymbol{w}^{*}(y_i,\mu_i)\Psi_{\beta}(y_i,\boldsymbol{x}_i,\boldsymbol{\beta},\alpha)w(\boldsymbol{x}_i)\right)$ is a correction that ensures consistency in the model. And the MLE of the weighted RNBR model for parameter $\alpha$ in (2.14) is as follows.

$$\sum_{i=1}^{n}\left[\frac{\varphi(\varepsilon_i)}{\varepsilon_i}\Psi_{\alpha}(y_i,\boldsymbol{x}_i,\boldsymbol{\beta},\alpha)w(\boldsymbol{x}_i)-b_i(\boldsymbol{\alpha})\right]=\mathbf{0} \tag{2.26}$$

where $\frac{\varphi(\varepsilon_i)}{\varepsilon_i}$ s is a weight analogous to the weight on the $\widehat{\boldsymbol{\beta}}$, and $b_i(\boldsymbol{\alpha})=E\left(\frac{\varphi(\varepsilon_i)}{\varepsilon_i}\Psi_{\alpha}(y_i,\boldsymbol{x}_i,\boldsymbol{\beta},\alpha)w(\boldsymbol{x}_i)\right)$ is a correction that ensures consistency in the model[1]. RNBR parameter estimation uses the GEM algorithm [5].

Regression model diagnostics are used to evaluate the regression model formed. This is by conducting an assumption test on the model error from the regression modelling results is a consideration of the goodness of the modelling results. There are 3 assumptions that usually need to be met in regression modelling, namely the assumption of homoscedasticity of model errors, the assumption of autocorrelation of model errors, and the assumption of multicollinearity between independent variables. If any of the assumptions are not met, then there are several treatments before stating that the regression model is not suitable for use in data analysis. Unlike the regression model, the robust model accommodates the presence of outliers in the model. Due to the presence of outliers, the test assumption of normality of model errors is not met. The first Robust model assumption is the assumption of homoscedasticity of model errors which is not met. To determine this assumption, model errors are tested using the Breusch-Pagan test with the Lagrange-Multiplyer (LM) test statistic [8] as follows.

$$LM=\frac{1}{2}f'Z(Z'Z)^{-1}Z'f \tag{2.27}$$

which tests the null hypothesis that $\sigma_{\varepsilon_1}^2=\sigma_{\varepsilon_2}^2=\cdots=\sigma_{\varepsilon_i}^2=0$. The value of the LM test statistic converges in the distribution ${\chi^2}_{p-1;\alpha}$. If the $p-$value $>\alpha$, then the null hypothesis is rejected. This means that the assumption of homoscedasticity is not met. In other words, the assumption of heteroscedasticity is met. There are several ways to handle homoskedasticity assumptions that are not met, one of which is spatial modelling or Robust modelling. In this study, it is more suitable to use robust modelling rather than spatial modelling. Because there is no spatial effect on this data modelling and it is more inclined to indicate outliers in the data. The second assumption in the Robust model is the assumption of autocorrelation between model errors. The existence of autocorrelation between model errors can be determined using the following Durbin-Watson (DB) test statistic [21]

$$d=\frac{\sum_{i=2}^{n}(e_i-e_{i-1})^2}{\sum_{i=1}^{n}e_i^2} \tag{2.28}$$

which tests $H_0$: $\rho=0$, and $H_1$: $\rho>0$. If the value of $0<d<4$ which means close to the value of 0 0 is an indication of negative autocorrelation, and approaching the value of 4 is an indication of positive autocorrelation. If $p-$value $<\alpha$, then the alternative hypothesis is accepted, which means

that there is autocorrelation between model errors. So this is one indication that the Robust model is suitable for modelling the data analysed. The third assumption that makes the Robust model one of the alternatives used is the presence of multicollinearity between independent variables. The existence of this problem can use the Variance Inflation Factors (VIF) test with test statistics written in the following equation.

$$VIF_k = \frac{1}{1-R_k^2} \tag{2.29}$$

where $R_k^2$ is the $R^2$ of the $k-$th independent variable. if the value of $VIF_k > 5$, then there is multicollinearity between the independent variables [21].

The next step is to know whether the independent variables simultaneously have a significant effect on the dependent variable. In data with independent variables that are the sum of binary events, the $G$ test [3] is used to perform the simultaneous test, namely

$$G^2 = -2(L_0 - L_1) \sim \chi^2_{(\alpha,m-1))} \tag{2.30}$$

where $L_0$ is the log-likelihood for the simple model, $L_1$ is the log-likelihood function for the complete model, $m$ is the number of dependent and independent variables. The G-test tests the hypotheses $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ and $H_1: \beta_k \neq 0$. If the $p$ value $< \alpha$, then the independent variables simultaneously have a significant effect on the dependent variable. After testing the parameters simultaneously, the next stage is to conduct a partial parameter test which aims to determine which independent variables have a significant effect on the dependent variable. In this study, the Wald test [3] is used to test the parameters partially with the test statistics can be written in the following equation.

$$W = \left(\frac{\hat{\beta}_q}{se(\hat{\beta}_q)}\right)^2 \sim \chi^2_{(\alpha,1)} \tag{2.31}$$

with hypothesis $H_0: \beta_q = 0$ and $H_1: \beta_q \neq 0$, where $q = \{0,1,\dots,k\}$ are parameter coefficients for intercept and independent variables. If the $p$ value $< \alpha$, then the independent variable partially has a significant effect on the dependent variable. The last step is to see which model is better by looking at the value of Akaike's Information Criterion (AIC) which can be written [31] by

$$AIC = 2k - 2\ln(L) \tag{2.32}$$

where $L$ is the maximum value of the likelihood function, and $k$ is the number of parameters estimated in the model.

**Data Type and Source**

This study uses secondary data consisting of 3 independent variables and 1 dependent variable. Some of the under-five immune variables that affect the incidence of pneumonia used in this study were obtained from [4, 13, 30, 33], and the size value of each dependent and independent variable was obtained from [10].

The detailed description of the variables involved in the analysis can be seen in Table 1.

**Table 1** Research Variables

| Variables | Description | Unit |
|---|---|---|
| $Y$ | Number of post neonatal deaths (age 29 days-11 months) | Toddlers |
| $X_1$ | Malnourished toddlers | Percent |
| $X_2$ | Complete primary immunisation | Percent |
| $X_3$ | IMD (Early Breastfeeding Initiation) | Percent |

**Data Analysis Procedure**

The following procedures need to be carried out in the analysis using the Robust Negative Binomial Regression method.

1. Perform descriptive statistical analysis and check for outliers in the data.
2. Perform Poisson regression modelling by estimating the parameter $\beta$ using equation (2.7) with Fisher's algorithm.
3. Check for overdispersion when modelling Poisson regression using equation (2.1). If there is overdispersion and outliers, Robust Negative Binomial Regression (RNBR) modelling can be performed by performing step (4).
4. Perform parameter estimation of the RNBR $(\alpha, \beta)$ model in Equations (2.13) and (2.14) with the Generalised Expectation-Maximisation (GEM) algorithm.
5. Perform simultaneous parameter tests for Poisson model parameters and RNBN model parameters using the $G$ test in equation (2.31).
6. Perform partial parameter tests for each model using the Wald test statistic in equation (2.32).
7. Checking the error assumption of the model and independent variabels assumption formed by using equation (2.27)-(2.29).
8. Choosing the best model between Poisson and RNBN regression is more appropriate in this data analysis using the AIC statistic in equation (2.33).
9. In this study, the analysis used R software, package "lmtest", function "bptest, dwtest" in [15] to check the assumptions of heteroskedastity and autocorrelation of model errors; package "nortest", function "ad.test" in [14] to check the assumption of normality of model errors; package "car", function "vif" to check the assumption of multicollinearity; package "performance", function "check_overdispersion" in [18] for overdispersion; function "glm" to estimate Poisson regression parameters; package "robmixglm", function "robmixglm" in [5] to estimate RNBR parameters.
10. Interpretation of results and discussion.

## 3. RESULTS AND DISCUSSION

Outliers are one of the problems that usually exist in data, so in the data analysis process, it must be handled first or use data analysis that allows outliers in the data. In this study, we will analyse data on under-five deaths due to pneumonia in East Java in 2022. The incidence of pneumonia was the highest among Malaria, Lung TB, and Leprosy in East Java at the time.

There are several references to pneumonia as the main cause of under-five mortality. According to [25], there are three pathways that form the framework of pneumonia: host, agent, and environment. In this case, the host is the treatment performed on toddlers, the agent is the cause of airborne pneumonia, and the environment consists of several variables that measure environmental health. In this study, only the treatment of under-five children was taken. There are several variables that are strongly suspected to be variables of the immune system in toddlers, especially in the post-neonatal age (29 days-11 months), that influence the outbreak of pneumonia,

namely malnutrition, complete primary immunisation in the post-neonatal age, and early breastfeeding (IMD).

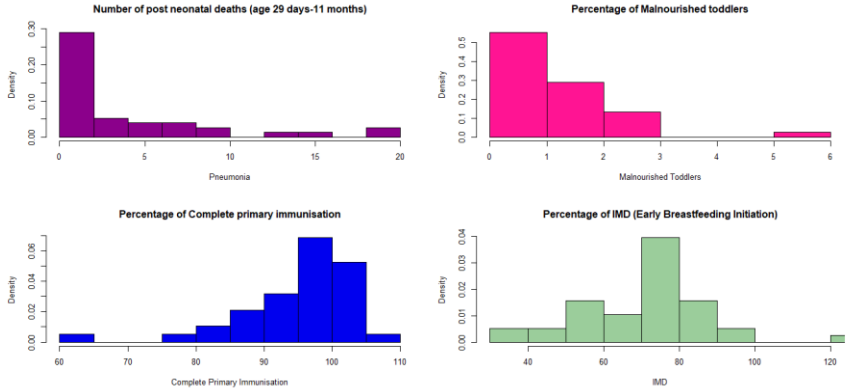


**Figure 1 Histogram of 4 research variables**

Looking at Figure 1, it can be seen that almost the data from all variables have skewed. This is an indication that there are outliers in the data and only the IMD variable is not indicated as an outlier in the data. For more detail, it can be seen in the multivariate plot shown in Figure 2.

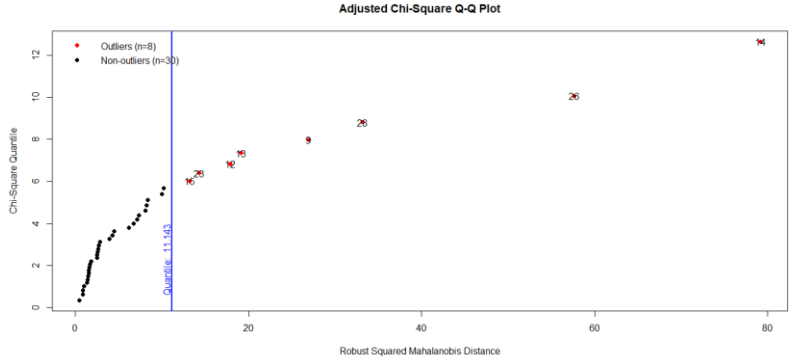


**Figure 2 Multivariate QQ plot of the study variables to detect outliers**

The black dots in Figure 2 are values that are not considered outliers, and the red dots in the figure are considered outliers in the data. There are 8 points that are considered outliers in the data to be analyzed. Since many values are considered outliers in the data, this indicates that multiple linear regression is not suitable as a method to model this data. Therefore, robust regression is a method of dealing with the outlier problem by allowing the inclusion of outliers in the analysis. In addition, the dependent variable used is data obtained from the counting process, so Poisson regression is an appropriate modelling method in this study. However, looking back at Figure 1, it can be seen that the pneumonia variable, which is the data obtained from the counting process, has a zero value, which dominates the pneumonia mortality cases. This means that in 2022 there were no pneumonia deaths in several districts/cities in East Java province. This indicates that there is overdispersion in the data when there are too many zeros (excess zeros) in the dependent variable. This problem can be handled by using Negative Binomial regression.

In this study, the dependent variable used is count data, so the first modelling uses Poisson regression. From Poisson regression modelling, parameter estimates $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are obtained, where $p$ being the number of independent variables used in data analysis. From the results of the analysis, the three variables have a significant influence on the number of deaths of toddlers of post neonatal age in East Java Province in 2022. In addition to the parameter estimation results using

Fisher's algorithm, there is information that the standard error value $s_{PR}^2$ for the Poisson regression model is 1. This $s_{PR}^2$ size indicates the influence of outliers on the model. For more details of the Poisson regression model parameter estimation results can be seen in Table 2.

**Table 2** Table 2 Estimation of parameters $\beta_k, k = 0,1,2,\ldots,p$ using Poisson Regression with 5% significance level

| Variables ($\hat{\beta}_k$) | Estimated Values | p-values | Decision | Interpretation |
|---|---|---|---|---|
| Intercept ($\hat{\beta}_0$) | 7.850339 | 5.12e-16 | Reject $H_0$ | Variables that have a significant effect on the number of under-five deaths in post neonatal age due to pneumonia |
| Malnourished toddlers ($\hat{\beta}_1$) | 0.262452 | 7.33e-05 | Reject $H_0$ | |
| Complete primary immunisation ($\hat{\beta}_2$) | -0.062010 | 6.13e-13 | Reject $H_0$ | |
| IMD (Early Breastfeeding Initiation) ($\hat{\beta}_3$) | -0.015449 | 0.0018 | Reject $H_0$ | |

Source: Estimation of Poisson Regression model parameters using R Software

where $i = 1,2,\ldots,38$. After we got the parameter estimation results, we will test whether each parameter for each independent variable has a significant effect simultaneously on the dependent variable for each model using the $G$ test. The $G$ test results for the Poisson model give the $G$ value of $5.594$ with $p$ −value of $0.061 > \alpha$. Its giving the result to accept the null hypothesis. Thus, for the Poisson model, the independent variables simultaneously **do not** have a significant effect on the dependent variable. So if we can conclude from Table 2 and the results of the G test, using the Poisson model the independent variables only partially have a significant effect on the dependent variable, but not simultaneously.

From Table 2, a Poisson model can be formed as in equation (3.1) and the information obtained is that an increase in the percentage of malnourished children under five has an effect on the increase in the number of under-five deaths in the post neonatal age due to pneumonia in East Java Province. In addition, an increase in the percentage of under-fives who received complete basic immunisation and early breastfeeding initiation can reduce the number of under-five deaths in East Java Province. From the estimation results that have been done, a Poisson model can be formed based on equation (2.3) and Table 2 as follows.

$$\mu_{i(RP)} = \exp(7.850339 + 0.262452 * \text{Malnourished toddlers} - 0.062010 * \text{Complete primary immunisation} - 0.015449 * \text{IMD}) \quad (3.1)$$

The next step of analysis to see the suitability of the modelling used in the analysis, especially in Poisson regression modelling is whether overdispersion is detected in the data. For this reason, the next stage of analysis is to see whether there is overdispersion in the data in Poisson regression modelling, which can be seen in Table 3.

**Table 3** Testing for overdispersion in Poisson regression with a 5% significance level

| Dispersion Ratio | *Pearson's* $\chi^2_{n-k}$ | $p$-values | Keputusan | Interpretation |
|---|---|---|---|---|
| 5.565 | 172.511 | < 0.001 | Reject $H_0$ | Overdispersion detected |

Source: Testing overdispersion in Poisson Regression using R Software

Due to the overdispersion in Poisson regression modelling seen in Table 3, it can be said that this modelling is not suitable to be used as an analysis method on this data. And because there are outlier and overdispersion problems that have been detected in the tests in Figure 2 and Table 3, the Robust Negative Binomial Regression (RNBR) model is desired to overcome these 2 problems. Therefore, the next analysis in this study is to estimate the parameters of RNBR modelling. For more detailed RNBR parameter estimation results can be seen in Table 4.

**Table 4** Estimation of parameters $\alpha$ and $\beta_k, k = 0,1,2, \dots, p$ using RNBR with a significance level of 5%

| Variables ($\hat{\beta}_k$) | Estimated Values | p-values | Decision | Interpretation |
|---|---|---|---|---|
| Intercept ($\hat{\beta}_0$) | 28.37181 | 0.000608 | Reject $H_0$ | Variables that have a significant effect on the number of under-five deaths in post neonatal age due to pneumonia |
| Malnourished toddlers ($\hat{\beta}_1$) | 1.25621 | 4.05e-05 | Reject $H_0$ | |
| Complete primary immunisation ($\hat{\beta}_2$) | -0.24935 | 0.004478 | Reject $H_0$ | |
| IMD (Early Breastfeeding Initiation) ($\hat{\beta}_3$) | -0.04714 | 0.017865 | Reject $H_0$ | |
| Overdispersion parameter ($\hat{\alpha}$) | 0.67803 | | | |

Source: Estimation of RNBR model parameters using R Software

For the RNBR model, we get the $G$ value of $30.159$ with $p$ −value of $0.0000002825 < \alpha$. Thus, the decision to reject the null hypothesis is obtained, which means that for the RNBR model, the independent variables simultaneously have a significant effect on the dependent variable. Table 4 and $G$ test for RNBR model, it is obtained that for the RNBR model, the independent variables have a significant effect simultaneously and partially on the dependent variable. And the estimation results that have been carried out, the RNBR model can be formed based on equation (2.11) and Table 4 as follows.

$$\mu_{i(RNBR)} = \exp(28.37181 + 1.25621 * \text{Malnourished toddlers} - 0.24935 * \text{Complete primary immunisation} - 0.04714 * \text{IMD}) \quad (3.2)$$

From the RNBR model formed in equation (3.2), information was obtained that most of the significant influences were in addition to the 3 factors that became the research variables. As in the Poisson regression model, malnourished toddlers have an influence in increasing the under-five mortality rate due to pneumonia in East Java Province. And for the treatment of complete basic immunization and early breastfeeding initiation for under-fives, it has a significant effect in reducing under-five mortality due to pneumonia in East Java Province with $s^2_{RBNR}$ of 0.41905 and the RNBR model estimates that there are 0.67803% outliers in the model.

The next step of analysis is the diagnostic of the two models that have been formed. There are 3 assumptions that will be tested on model errors and variable assumptions between independent variables for diagnostics on the model that has been formed. Detailed test results can be seen in Table 5.

**Table 5** Model diagnostics of Poisson Regression and RNBR model errors

| Model | Homoskedatisity Test | Non-Autocorrelation Test | Non-Multicollinearity Test |
|---|---|---|---|
| Poisson Regression | ✗ | ✓ | ✓ |
| RNBR | ✗ | ✓ | ✓ |

Source: Model diagnostics using R Software

✓ : assumptions fulfilled
✗ : assumptions not fulfilled

In Table 5, to get suitable results in Poisson regression modelling, it is necessary to meet 3 assumptions in the regression. To overcome the assumptions of homoskedasticity of errors in the Poisson regression model, the presence of outliers, and overdispersion, RNBR is more suitable in modelling this data. In addition, spatial effects do not have a significant influence on the results of modelling this data. Furthermore, to get the best model of the two models can be seen in the smallest AIC value. The AIC values for both models can be seen in Table 6.

**Table 6** Goodness of fit of Poisson Regression and RNBR models

| Model | AIC |
|---|---|
| Poisson Regression | 266.73 |
| RNBR | 227.9292 |

Source: Goodness of fit model using R Software

Comparison of the AIC values of the Poisson regression model and the RNBR model in Table 6 shows that the AIC value of the RNBR model is smaller than the AIC value of the Poisson model. This means that the RNBR model provides better results in modelling this data. In other words, from the overall analysis results, the RNBR model is better and more suitable for use in this study.

## 4. CONCLUSION

The Robust Negative Binomial Regression (RNBR) model provides more robust results in dealing with outlier and overdispersion problems. This can be seen from the AIC value of the RNBR model is smaller than the AIC of the Poisson regression model. In addition, it can be seen that the value of $s_{PR}^2$ and $s_{RBNR}^2$ which is a measure of the influence of outliers on the model gives a decrease in the value of 1 for the Poisson regression model to around 0.42 for the RNBR model. furthermore, only the RNBR modelling provides results that the independent variables have a simultaneous significant effect on the dependent variable. From the parameter estimation that has been done, the RBNR model is formed as follows.

$$\mu_{i(RNBR)} = \exp(28{,}37181 + 1{,}25621 * \text{Malnourished toddlers} - 0{,}24935 * \text{Complete primary immunisation} - 0{,}04714 * \text{IMD}) \quad (5.1)$$

From both modelling, there are 3 strong factors in terms of immunity that are the main influence on the number of under-five deaths at post neonatal age in East Java Province in 2022, namely the percentage of under-fives with malnutrition, the percentage of under-fives who get complete basic immunization treatment and the presence of early breastfeeding initiation (IMD).

The percentage of malnourished toddlers has an influence in adding to the number of under-five deaths in East Java Province. And 2 other factors, can reduce the number of under-five deaths in East Java Province in 2022.

## REFERENCES

[1] Aeberhard, W.H., Cantoni, E. and Heritier, S., 2014. Robust inference in the negative binomial regression model with an application to falls data. *Biometrics*, Vol. 70, No. 4, 920–931.
[2] Aggarwal, C.C., 2017. *Outlier Analysis*. Springer International Publishing, Cham.
[3] Agresti, A., 2013. Categorical Data Analysis (3rd ed.). Wiley Interscience, New Jersey. Belum ditulis
[4] Banhae, Y.K., Abanit, Y.M. and Namuwali, D., 2023. Faktor Risiko yang Berhubungan dengan Kejadian Pneumonia pada Balita di Kota Kupang. *Jurnal Ilmiah Permas: Jurnal Ilmiah STIKES Kendal*, Vol. 13, No. 3, 1099–1106.
[5] Beath, K., 2022. Robust Generalized Linear Models (GLM) using Mixtures. *CRAN R*, Package, (May 2022).
[6] Bhar, L., 2007. *Robust Regression*. Course Online, Indian Agricultural Statistics Research Institute (I.C.A.R.), New Dehli.
[7] BPS Jatim., 2023. Jumlah Jenis Penyakit Malaria, TB Paru, Pneumonia, Kusta Menurut Kabupaten/Kota di Provinsi Jawa Timur Tahun 2022. *Badan Pusat Statistik Provinsi Jawa Timur*, Surabaya.
[8] Breushch, T.S. and Pagan, A.R., 1979. A Simple Test For Heteroscedasticity And Random Coefficient Variation. *Econometrica*, Vol. 47, No. 5, 1287–1294.
[9] Croux, C., 2003. *The Poisson Regression Model*. Online Course, KU Leuven, Netherland.
[10] Dinas Kesehatan Jatim., 2023. *Profil Kesehatan Provinsi Jawa Timur Tahun 2022*. Dinas Kesehatan Provinsi Jawa Timur, Surabaya.
[11] Facrotul, N., 2023. *Pemetaan Kasus Pneumonia Balita Di Jawa Timur Berdasarkan Hasil Pemodelan Dengan Geographically Weighted Negative Binomial Regression (GWNBR)*. Tesis Diploma, Institut Teknologi Sepuluh November, Surabaya.
[12] Gelman, A. and Hills, J., 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.
[13] Girma, F., Ayana, M., Abdissa, B., Aboma, M., Ketema, D., Kolola, T. and Wake, S.K., 2023. Determinants of under-five pneumonia among children visited in nine public health Hospitals in Ethiopia. *Clinical Epidemiology and Global Health*, Vol. 24, 101441, 1-6.
[14] Gross, J. and Ligges, U., 2015. Tests for Normality. *CRAN R*, Packages, (Jul. 2015).
[15] Hothorn, T., Zeileis, A., Farebrother, R.W., Cummins, C., Millo, G. and Mitchell, D., 2022. Testing Linear Regression Models. *CRAN R*, Packages, (Mar. 2022).
[16] Jansakul, N. and Hinde, J.P., 2004. Linear mean-variance negative binomial models for analysis of orange tissue-culture data. *Songklanakarin Journal of Science and Technology*, Vol. 26, No. 5, 683-696.
[17] Kemenkes RI., 2021. *Profil Kesehatan Indonesia Tahun 2021*. Kementerian Kesehatan Republik Indonesia, Jakarta.
[18] Lüdecke, D., Makowski, D., Ben-Shachar, M.S., Patil, I., Waggoner, P., Wiernik, B.M., Arel-Bundock, V., Thériault, R., Jullum, M. and Bacher, E., 2023. Assessment of Regression Models Performance. *CRAN R*, Package, (Oct. 2023).
[19] McCullagh, P. and Nelder, J.A., 1989. *Generalized Linier Models*. Chapman And Hall, New York.

[20] Molla, D.T. & Muniswamy, B., 2012. Power of Tests for Overdispersion Parameter in Negative Binomial Regression Model. *IOSR Journal of Mathematics (IOSRJM)*, Vol. 1, No. 4, 29–36.
[21] Montgomery, D.C., Peck, E.A. and Vining, G.G., 2001. *Introduction to Linear Regression Analysis*. John Wiley & Sons, New York.
[22] Nugroho, P.A. and Danardono, D., 2016. *Pemodelan Data Cacah Excess Zero Menggunakan Model Berbasis Poisson Dan Binomial Negatif*. Tesis, Universitas Gadjah Mada, Yogyakarta.
[23] Nuraini, H. & Febriana, F., 2023. Peningkatan Pengetahuan Mengenali Tanda Kegawatan Pneumonia dan Penanganannya pada Anak Setelah dilakukan Penyuluhan Kesehatan. *SIGDIMAS : Publikasi Kegiatan Pengabdian Masyarakat*, Vol. 1, No. 1, 35–40.
[24] Owusu, S., 2018. Analysis Of The Effects Of Overdispersion In Population Dynamics. A Research Thesis, Pan African University, Cameroun.
[25] Qur'ani, A.Y., 2023. Pemodelan Principal Component Regression Analysis dari Faktor Penanganan Stunting saat Pandemi Covid-19 di Indonesia. *Ulil Albab*, Vol. 2, No. 8, 3922–3931.
[26] Romeu, J.L., 2003. Anderson-Darling: A Goodness of Fit Test for Small Samples Assumptions. *START : Selected Topics in Assurance Related Technologies*, Vol. 10, No. 5, 1–6.
[27] Rustiyanto, E., 2012. *Faktor Risiko Kejadian Pneumonia Pada Balita (Studi Kasus Di Puskesmas Umbulharjo II Kota Yogyakarta)*. Tesis, Universitas Diponegoro, Semarang.
[28] Santika, E.F., 2023. Pneumonia Jadi Penyebab Terbesar Kematian Balita di Dunia 2021. *Databoks*, Layanan Konsumen & Kesehatan. https://databoks.katadata.co.id/datapublish/2023/12/04/pneumonia-jadi-penyebab-terbesar-kematian-balita-di-dunia-2021. [9 Februari 2024]
[29] Setyawan, Y., Suryowati, K. and Octaviana, D., 2022. Application of Negative Binomial Regression Analysis to Overcome the Overdispersion of Poisson Regression Model for Malnutrition Cases in Indonesia. *Parameter: Journal of Statistics*., Vol. 2, No. 2, 1–9.
[30] Solomon, Y., Kofole, Z., Fantaye, T. and Ejigu, S., 2022. Prevalence of pneumonia and its determinant factors among under-five children in Gamo Zone, Southern Ethiopia, 2021. *Frontiers in Pediatrics*, Vol. 10, 1017386, 1-8.
[31] Symonds, M.R.E. and Moussalli, A., 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, Vol. 65, No. 1, 13–21.
[32] Tan, S.Z., 2021. *The Robustness of Count Models in the Presence of Measurement Error and Process Error*. Thesis, University Of Helsinki, Finland.
[33] Utami, P.F., Rusgiyono, A. and Ispriyanti, D., 2021. Pemodelan Semiparametric Geographically Weighted Regression Pada Kasus Pneumonia Balita Provinsi Jawa Tengah. *Jurnal Gaussian*, Vol. 10, No. 2, 250–258.
[34] Walpole, R.E., Myers, R.H., Myers, S.L. and Ye, K., 2017. *Probability & statistics for engineers & scientists: MyStatLab update*. Pearson, Boston.