p-ISSN 2338 – 3216
e-ISSN 2528 – 1070

JURNAL STATISTIKA
UNIVERSITAS MUHAMMADIYAH SEMARANG
11(1) 2023: 45-60

Jurnal Statistika
Universitas Muhammadiyah Semarang

MODELING COUNT DATA WITH OVER-DISPERSION USING GENERALIZED POISSON REGRESSION: A CASE STUDY OF LOW BIRTH WEIGHT IN INDONESIA

M. Fathurahman*
Department of Mathematics, Study Program of Statistics, Faculty of Mathematics and Natural Sciences, Mulawarman University, Indonesia

***e-mail**: *fathur@fmipa.unmul.ac.id*

Article Info:
Received: April 11, 2023
Accepted: May 8, 2023
Available Online: May 31, 2023

Keywords:
Count Data; Generalized Poisson Regression; Low Birth Weight; Over-dispersion; Poisson Regression.

Abstract: Poisson regression is commonly used in modeling count data in various research fields. An essential assumption must be met when using Poisson regression, which is that the count data of the response has the mean and variance must be equal, namely equip-dispersion. This assumption is often unmet because many data for the response that the variance is greater than the mean, called over-dispersion. If the Poisson regression model contains the over-dispersion, then will be produced an invalid model can under-estimate standard errors and misleading inference for regression parameters. Therefore, an approach is needed to overcome the over-dispersion problem in Poisson regression. The generalized Poisson regression can handle the over-dispersion in Poisson regression. This study aims to obtain the generalized Poisson regression model and the factors affecting the low birth weight in Indonesia in 2021. The result shows that the factors affecting the low birth weight in Indonesia based on the generalized Poisson regression model were: poverty rate, percentage of households with access to appropriate sanitation, percentage of pregnant women at risk of chronic energy deficiency receiving additional food, percentage of pregnant women who received blood-boosting tablets, and percentage of antenatal care.

1. INTRODUCTION

Poisson regression is widely used in modelling count data. Count data is one type of statistical data that shows the number of events over a particular time and can only be positive [1]. An essential assumption must be met in Poisson regression modelling; namely, the mean and variance of the response must be equal, called equid-dispersion [2]. This assumption is often unmet because, in many data in various research fields, the variance is greater than the mean, called over-dispersion. An invalid model can underestimate standard errors and misleading inferences for regression parameters [1]. Therefore, an approach is needed to overcome the over-dispersion problem in Poisson regression. The generalized Poisson regression is an alternative approach for handling it [2].

Several studies that model count data with over-dispersion using generalized Poisson regression have been proposed. The maximum likelihood and moment methods were used to estimate the generalized Poisson regression model parameters. In contrast, the significance test of the parameters was used by the likelihood ratio test method [3]. The restricted generalized

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

Poisson regression model was developed [4]. The generalized Poisson regression model was applied to model the infant mortality rate [5].

The generalized Poisson regression in this study was applied to model the factors affecting the low birth weight in Indonesia, in 2021. The modelling of low birth weight using the Poisson regression shows any over-dispersion problem. Low birth weight is a birth weight of fewer than 2500 grams. Low birth weight has always been a significant public health problem globally and is associated with various of short- and long-term consequences. Overall, 15 to 20 percent of all births worldwide are estimated to be low birth weight, representing more than 20 million births annually. WHO has committed to monitoring the progress of global change and supporting global targets to improve maternal, infant and child nutrition through six global nutrition targets by 2025. One of them is the third target which aims to achieve a 30 percent reduction in body weight and low birth weight by 2025. It means a target of a relative reduction of 3 percent per year between 2012 and 2025, namely a decrease from about 20 million to around 14 million babies with low birth weight [6].

A baby's weight at birth is the most crucial determinant of the chances of survival, growth, and development in the future. Mothers who continually maintain their health by consuming nutritious food and adopting a good lifestyle will give birth to healthy babies. In contrast, mothers who experience nutritional deficiencies have a risk of giving birth to babies with low body weight. The low birth weight reflects the health and nutrition situation and shows the level of survival and psychosocial development [7]. Babies with low birth weight have a higher risk of experiencing death, growth retardation, and development during childhood than babies who are not low birth weight [8]. Some of the factors that cause low birth weight are pregnant women experiencing chronic energy shortages, poor antenatal care, poverty, and poor sanitation [9], [10], [11].

This study aims to obtain the generalized Poisson regression model, the factors affecting modelling count data with overdispersion, and its application in low birth weight in Indonesia, in 2021. Following [3], [12], [13], [14], the Poisson regression and generalized Poisson regression models can be obtained by the maximum likelihood and Fisher-scoring methods. In contrast, the test of significant parameters of the Poisson regression and generalized Poisson regression models can be used by the likelihood ratio test and Wald test methods.

2. LITERATURE REVIEW

2.1. Low Birth Weight

Low birth weight is a baby born weighing less than 2,500 grams. Low birth weight consists of low birth weight (1,500-2,499 grams), very low birth weight (1,000-1,499 grams), and extremely low birth weight (< 1,000 grams). 60 to 80 percent of the infant mortality rate is due to low birth weight. Low birth weight has a greater risk of experiencing morbidity and mortality than babies born with normal weight. A gestation period of less than 37 weeks can cause complications in the baby due to the imperfect growth of the organs in the body. The lower the baby's weight, the more crucial it is to monitor its development in the weeks after birth. Low birth weight can be caused by two factors: premature birth and Intra Uterine Growth Restriction (IUGR), commonly called impaired fetal growth. Low birth weight can cause morbidity and even death [7].

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

2.2. Poisson Regression

Poisson regression is a nonlinear parametric regression model. The response of Poisson regression model (Y) follows the Poisson distribution, which has the probability mass function defined as [2]:

$$P(Y = y|\mu_1) = \frac{\exp(\mu_1)\,\mu_1^y}{y!}, y = 0,1,2, \ldots \tag{1}$$

where μ_1 is the parameter and $\mu_1 > 0$. $E(Y) = \mu_1$ and $(Y) = \mu_1$, respectively, symbolize the mean and variance of the Poisson distribution.

Suppose there are covariates namely $X_1, X_2, \ldots, X_k$, then the Poisson regression model can be written as follows [2]:

$$\zeta_1(\mathbf{x}) = \log(\mu_1) = \log[\exp(\mathbf{x}^T\boldsymbol{\theta}_1)] = \mathbf{x}^T\boldsymbol{\theta}_1 \tag{2}$$

where μ_1 is the mean of response. $\mathbf{x}^T = [X_0 \quad X_1 \quad X_2 \quad \cdots \quad X_k]$ is the vector of covariates with $X_0 = 1$. $\boldsymbol{\theta}_1 = [\theta_{10} \quad \theta_{11} \quad \theta_{12} \quad \cdots \quad \theta_{1k}]^T$ is the vector of regression parameters. k is the number of covariates. $\zeta_1(\mathbf{x})$ is the link function that depends on the covariates [15].

The Poisson regression model in Equation (2) can be obtained by estimating the model's parameter using the maximum likelihood method. The estimation begins with obtaining the likelihood and log-likelihood functions as follows:

$$\mathcal{L}(\boldsymbol{\theta}_1) = \prod_{i=1}^{n} P(Y_i = y_i|\mathbf{x}_i, \boldsymbol{\theta}_1) = \prod_{i=1}^{n}\left[\frac{\exp(-\exp(\mathbf{x}_i^T\boldsymbol{\theta}_1))\,(\exp(\mathbf{x}_i^T\boldsymbol{\theta}_1))^{y_i}}{y_i!}\right] \tag{3}$$

$$\ell(\boldsymbol{\theta}_1) = \log[\mathcal{L}(\boldsymbol{\theta}_1)] = \sum_{i=1}^{n}[y_i\mathbf{x}_i^T\boldsymbol{\theta}_1 - \exp(\mathbf{x}_i^T\boldsymbol{\theta}_1) - \log(y_i!)]. \tag{4}$$

It maximizes the log-likelihood function in Equation (4) by determining the first partial derivative of the log-likelihood function with respect to the estimated parameter and then equating it with zero,

$$\frac{\partial\ell(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1} = -\sum_{i=1}^{n}\left[\mathbf{x}_i^T\left(y_i - \exp\left(\mathbf{x}_i^T\widehat{\boldsymbol{\theta}}_1\right)\right)\right] = \mathbf{0}. \tag{5}$$

Based on Equation (5), the result of the first partial derivative of the log-likelihood function with respect to the estimated parameters produces an implicit function. Therefore, a numerical approach is needed to obtain the maximum likelihood estimator of the PR model parameters. One numerical approach is the Fisher-scoring method [12]. The Fisher-scoring algorithm for obtaining the maximum likelihood estimator of the Poisson regression model parameters is as follows:

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

1) Determine the initial value for $\widehat{\boldsymbol{\theta}}_1$, namely $\widehat{\boldsymbol{\theta}}_1^{(0)} = \left[\hat{\theta}_{10}^{(0)} \quad \hat{\theta}_{11}^{(0)} \quad \hat{\theta}_{12}^{(0)} \quad \cdots \quad \hat{\theta}_{1k}^{(0)}\right]^T$.
2) Determine the tolerance value, symbolized by δ for the iteration process stopping.
3) Start the iteration process using the following formula:
4)

$$\widehat{\boldsymbol{\theta}}_1^{(u+1)} = \widehat{\boldsymbol{\theta}}_1^{(u)} + \mathbf{I}^{-1}\left[\widehat{\boldsymbol{\theta}}_1^{(u)}\right]\mathbf{g}\left[\widehat{\boldsymbol{\theta}}_1^{(u)}\right], u = 0,1,2, \ldots \tag{6}$$

where $\mathbf{g}(\boldsymbol{\theta}_1)$ is the gradient vector, which has the elements in Equations (5). $\mathbf{I}(\boldsymbol{\theta}_1)$ is the information matrix and expressed as

$$\mathbf{I}(\boldsymbol{\theta}_1) = E\left[-\frac{\partial^2 \ell(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T}\right],$$

where the $\partial^2 \ell(\boldsymbol{\theta}_1)/\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T$ is the second partial derivative of the log-likelihood function with respect to the estimated parameters as follows:

$$\frac{\partial^2 \ell(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T} = \sum_{i=1}^{n} [\mathbf{x}_i^T \mathbf{x}_i \exp(\mathbf{x}_i^T \boldsymbol{\theta}_1)].$$

5) The iteration process stops at the u-th iteration when converged, namely $\left\|\widehat{\boldsymbol{\theta}}_1^{(u+1)} - \widehat{\boldsymbol{\theta}}_1^{(m)}\right\| \leq \delta$, where δ is the smallest positive number. The estimator values of the Poisson regression model parameters are obtained in the last iteration.

Following [13], the estimate for the variance-covariance matrix of $\boldsymbol{\theta}_1$ is $Cov\left(\widehat{\boldsymbol{\theta}}_1\right) = \left[\boldsymbol{I}\left(\widehat{\boldsymbol{\theta}}_1\right)\right]^{-1}$. The $Cov\left(\widehat{\boldsymbol{\theta}}_1\right)$ value can be used on the significance test of the Poisson regression model parameters below.

The significance test on the Poisson regression model parameters aims to get the covariates affecting the response simultaneously and partially. The likelihood ratio test method is applied to the simultaneous test using the hypotheses:

$$\begin{aligned} &H_0: \theta_{11} = \theta_{12} = \cdots = \theta_{1k} = 0 \\ &H_1: \text{at least one of } \theta_{1j} \neq 0, j = 1,2, \ldots, k. \end{aligned} \tag{7}$$

The test statistic used to test the hypothesis in Equation (7) is Wilk's lambda statistic which can be obtained by the likelihood ratio test method, and is formulated as follows [13].

$$G_1^2 = -2 \log \Lambda_1 = 2\left[\ell\left(\widehat{\Omega}_1\right) - \ell(\widehat{\omega}_1)\right] \tag{8}$$

where $\ell(\widehat{\omega}_1)$ is the maximum value of the log-likelihood function for the set of model parameters under the null hypothesis (H_0) and $\ell\left(\widehat{\Omega}_1\right)$ is the maximum value of the log-likelihood function for the set of model parameters under the population are as follows:

$$\ell(\widehat{\omega}_1) = \sum_{i=1}^{n} \left[y_i \hat{\theta}_{10} - \exp\left(\hat{\theta}_{10}\right) - \log(y_i!)\right]$$

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

$$\ell(\widehat{\Omega}_1) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^T \widehat{\boldsymbol{\theta}}_1 - \exp\left(\mathbf{x}_i^T \widehat{\boldsymbol{\theta}}_1\right) - \log(y_i!) \right].$$

Wilk's lambda statistic in Equation (8) is asymptotically chi-square distributed [16]. Therefore, the rejected region of H_0 (the critical region) on the significance level (α) to test the hypothesis in Equation (7) is the null hypothesis rejected if the value of G_1^2 greater than the value of $\chi^2_{(\alpha,k_1)}$ or the null hypothesis is rejected when the p-value is less than the α value. k_1 is the degree of freedom being the difference between the number of model parameters under the population and the null hypothesis, namely $k_1 = (k+1) - 1 = k$ [13].

The parameter hypothesis testing carried out after the simultaneous test is a partial test. The hypothesis used for the partial test is:

$$\begin{aligned} &H_0: \theta_{1j} = 0 \\ &H_1: \theta_{1j} \neq 0, j = 1,2, \dots, k. \end{aligned} \tag{9}$$

The test statistic for testing the hypothesis in Equation (9) is Wald's statistic, formulated as follows [13].

$$W_1 = \frac{\hat{\theta}_{1j}}{\widehat{SE}\left(\hat{\theta}_{1j}\right)} \tag{10}$$

where $\hat{\theta}_{1j}$ is the estimated value of the maximum likelihood parameter of the Poisson regression model obtained by the Fisher-scoring method in Equation (6). $\widehat{SE}\left(\hat{\theta}_{1j}\right) = \left[\hat{V}\left(\hat{\theta}_{1j}\right)\right]^{1/2}$ is the maximum likelihood standard error estimate of the parameter of the Poisson regression model obtained from the main diagonal elements of the variance-covariance matrix, $Cov\left[\widehat{\boldsymbol{\theta}}_1\right] = \mathbf{I}^{-1}\left[\widehat{\boldsymbol{\theta}}_1\right]$ where $\mathbf{I}\left[\widehat{\boldsymbol{\theta}}_1\right]$ is the Fisher information matrix [13].

Wald's statistic in Equation (10) is asymptotically standard normal distributed [16] so that the critical region at the significance level to test the hypothesis in Equation (9) is the null hypothesis is rejected if the value of $|W_1|$ is greater than the value of $Z_{\alpha/2}$ or the null hypothesis is rejected if the p-value is less than the α value.

2.3. Over-dispersion

Over-dispersion is one of the most common problems in Poisson regression. The Poisson regression assumes the count data has the same variance value as its mean (equi-dispersion) [5]. Sometimes the count data contains over-dispersion, shown by the variance greater than the mean, namely $V(Y) > E(Y)$. Over-dispersion occurs due to unobserved sources of variability in the data or the effect of other variables that result in the probability of an event occurring depending on previous events. Over-dispersion can lead to underestimating the standard error, resulting in under-estimated parameters and the significance of the covariate effect being over-estimated. Over-dispersion in Poisson regression can be detected by the deviance divided by the degrees of freedom. If the value is greater than one, it is shown that there is over-dispersion [1].

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

2.4. Generalized Poisson Regression

Generalized Poisson regression is a development of the Poisson regression model. The generalized Poisson regression model can deal with under-dispersion and over-dispersion problems in Poisson regression [2]. The response (Y) of the generalized Poisson regression model has a generalized Poisson distribution with the probability mass function defined as follows [17]:

$$P(Y = y|\mu_2, \lambda) = \left[\frac{\mu_2}{1+\lambda\mu_2}\right]^y \left[\frac{(1+\lambda y)^{y-1}}{y!}\right] \exp\left[\frac{-\mu_2(1+\lambda y)}{1+\lambda\mu_2}\right], y = 0,1,2,\ldots \tag{11}$$

where μ_2 and λ are the parameters, for $\mu_2 > 0$ and $\lambda > 0$. $E(Y) = \mu_2$ and $V(Y) = \mu_2(1+\lambda\mu_2)^2$, respectively define the mean and variance of the generalized Poisson distribution.

Based on Equation (11), the generalized Poisson regression model can be written as follows [3]:

$$\zeta_2(\mathbf{x}) = \log(\mu_2) = \log[\exp(\mathbf{x}^T\boldsymbol{\theta_2})] = \mathbf{x}^T\boldsymbol{\theta_2} \tag{12}$$

where $\zeta_2(\mathbf{x})$ is the log link function that depends on the covariates. $\boldsymbol{\theta_2}$ is the parameter vector, and $\mathbf{x}^T$ is the vector of covariates, which are defined by $\boldsymbol{\theta_2} = [\theta_{20} \quad \theta_{21} \quad \theta_{22} \quad \cdots \quad \theta_{2k}]^T$ and $\mathbf{x}^T = [1 \quad X_1 \quad X_2 \quad \cdots \quad X_k]$, respectively.

The generalized Poisson regression model in Equation (12) can be obtained by estimating the model parameters using the maximum likelihood method [7]. The initial step is forming the likelihood and log-likelihood functions. Suppose $\boldsymbol{\theta} = [\lambda \quad \boldsymbol{\theta}_2^T]^T$, then the likelihood and log-likelihood functions are formulated as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} P(Y_i = y_i|\mathbf{x}_i, \lambda, \boldsymbol{\theta_2}) \\ &= \prod_{i=1}^{n} \left\{\left[\frac{\mu_{2i}}{1+\lambda\mu_{2i}}\right]^{y_i} \left[\frac{(1+\lambda y_i)^{y_i-1}}{y_i!}\right] \exp\left[\frac{-\mu_{2i}(1+\lambda y_i)}{1+\lambda\mu_{2i}}\right]\right\},\end{aligned} \tag{13}$$

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \log[\mathcal{L}(\boldsymbol{\theta})] \\ &= \sum_{i=1}^{n} \log\left\{\left[\frac{\mu_{2i}}{1+\lambda\mu_{2i}}\right]^{y_i} \left[\frac{(1+\lambda y_i)^{y_i-1}}{y_i!}\right] \exp\left[\frac{-\mu_{2i}(1+\lambda y_i)}{1+\lambda\mu_{2i}}\right]\right\} \\ &= \sum_{i=1}^{n} \{y_i[\log(\exp(\mathbf{x}_i^T\boldsymbol{\theta_2})) - \log(1+\lambda\exp(\mathbf{x}_i^T\boldsymbol{\theta_2}))] + (y_i-1)\log(1+\lambda y_i) \\ &\qquad - \log(y_i!) - \frac{\exp(\mathbf{x}_i^T\boldsymbol{\theta_2})\,(1+\lambda y_i)}{1+\lambda\exp(\mathbf{x}_i^T\boldsymbol{\theta_2})}\Big\}.\end{aligned} \tag{14}$$

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

The next step is maximizing the log-likelihood function in Equation (14) by determining the first partial derivative of the log-likelihood function for the estimated parameters is then equated to zero,

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2} = \sum_{i=1}^{n} \left\{ y_i \mathbf{x}_i^T - \frac{\lambda y_i \mathbf{x}_i^T \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2)}{1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2)} + \left[\frac{\mathbf{x}_i^T \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2)\,(1 + \lambda y_i)}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))} \right] \times \left[\frac{\lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2)}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))} - 1 \right] \right\} = \mathbf{0}, \tag{15}$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \lambda} = \sum_{i=1}^{n} \left\{ \frac{y_i (y_i - 1)}{1 + \lambda y_i} + \frac{\exp(\mathbf{x}_i^T \boldsymbol{\theta}_2)}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))} \left[\frac{(1 + \lambda y_i) \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2)}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))} - 2 y_i \right] \right\} = \mathbf{0}. \tag{16}$$

The maximum likelihood parameter estimator of the generalized Poisson regression model in Equations (15) and (16) is an implicit function. Therefore, the maximum likelihood estimator cannot be obtained explicitly and requires a numerical approach. As in the Poisson regression model, a numerical approach with the Fisher-scoring method is used to obtain the maximum likelihood parameter estimator of the generalized Poisson regression model. The Fisher-scoring algorithm used is as follows:

1) Determine the initial value for $\widehat{\boldsymbol{\theta}}^{(0)}$.
2) Determine the value of gradient vector, $\mathbf{g}\left[\widehat{\boldsymbol{\theta}}^{(u)}\right]$.
3) Determine the value of Fisher information matrix inverse, $\mathbf{I}^{-1}\left[\widehat{\boldsymbol{\theta}}^{(u)}\right]$.
4) Carry out the Fisher-scoring iteration process using the following formula:

$$\widehat{\boldsymbol{\theta}}^{(u+1)} = \widehat{\boldsymbol{\theta}}^{(u)} + \mathbf{I}^{-1}\left[\widehat{\boldsymbol{\theta}}^{(u)}\right] \mathbf{g}\left[\widehat{\boldsymbol{\theta}}^{(u)}\right], u = 0,1,2, \ldots \tag{17}$$

where $\mathbf{g}(\boldsymbol{\theta})$ is the gradient vector, which has the elements in Equations (15) and (16). $\mathbf{I}(\boldsymbol{\theta})$ is the information matrix and defined as

$$\mathbf{I}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right],$$

where the $\partial^2 \ell(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$ is the second partial derivative of the log-likelihood function with respect to the estimated parameters as follows:

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T} = \sum_{i=1}^{n} \left\{ \left[\frac{\lambda y_i \mathbf{x}_i^T \mathbf{x}_i \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2)}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))} \right] \left[\frac{\lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2)}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))} - 1 \right] + \left[\frac{\mathbf{x}_i^T \mathbf{x}_i \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2)\,(1 + \lambda y_i)}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))} \right] \left[\frac{\lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2) - 1}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))} \right] \times \left[\frac{1}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))} \right] \right\}, \tag{18}$$

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \lambda^2} = \sum_{i=1}^{n} \left\{ -\frac{y_i^2(y_i - 1)}{(1 + \lambda y_i)^2} + \left[\frac{\exp(2\mathbf{x}_i^T \boldsymbol{\theta}_2)}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))^2} \right] \times \left[y_i - \frac{2(1 + \lambda y_i)\exp(\mathbf{x}_i^T \boldsymbol{\theta}_2)}{(1 + \lambda \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2))} \right] \right\}. \quad (19)$$

5) The iteration process stops when convergent conditions are met, namely $\left\| \widehat{\boldsymbol{\theta}}^{(u+1)} - \widehat{\boldsymbol{\theta}}^{(u)} \right\| \leq \delta$, where δ is the smallest positive number. The maximum likelihood parameter estimator of the generalized Poisson regression model was obtained from the last iteration.

If the maximum likelihood parameter estimator of the generalized Poisson regression model has been obtained, then parameter hypothesis testing can be carried out. This test consists of a simultaneous test and a partial test. The simultaneous test is used to determine the effect of the covariates on the response simultaneously. In contrast, the partial test is used to determine the effect of each covariate on the response individually. The hypothesis for the simultaneous test is:

$$H_0: \theta_{21} = \theta_{22} = \cdots = \theta_{2k} = 0$$
$$H_1: \text{at least one of } \theta_{2j} \neq 0, j = 1,2, \ldots, k. \quad (20)$$

The test statistic used to test the hypothesis in Equation (20) is Wilk's lambda statistic (G_2^2) which can be obtained by the likelihood ratio test method and is formulated as follows [2]:

$$G_2^2 = -2 \log \Lambda_2 = 2\left[\ell\left(\widehat{\Omega}_2\right) - \ell(\widehat{\omega}_2)\right] \quad (21)$$

where $\ell(\widehat{\omega}_2)$ and $\ell\left(\widehat{\Omega}_2\right)$ are the values of maximum log-likelihood function under the null hypothesis and population, respectively. The $\ell(\widehat{\omega}_2)$ and $\ell(\widehat{\omega}_2)$ are obtained by:

$$\ell(\widehat{\omega}_2) = \sum_{i=1}^{n} \left[y_i \hat{\theta}_{20} - \exp\left(\hat{\theta}_{20}\right) - \log(y_i!) \right]$$
$$\ell\left(\widehat{\Omega}_2\right) = \sum_{i=1}^{n} \left[y_i \widehat{\boldsymbol{\theta}}_2^T \mathbf{x}_i - \exp(\mathbf{x}_i^T \boldsymbol{\theta}_2) - \log(y_i!) \right].$$

Wilk's lambda (G_2^2) statistic in Equation (21) is asymptotically chi-square distributed [16]. Therefore, the critical region of the null hypothesis in Equation (20) on the significance level (α) is rejected when the G_2^2 statistic value is greater than the $\chi^2_{(\alpha,k_2)}$ value (i.e., $G_2^2 > \chi^2_{(\alpha,k_2)}$) or the p-value is less than α, where k_2 is the degrees of freedom, which is $k_2 = (k + 2) - 2 = k$.

The next test is the partial test. The Wald test method is used for this test that has the hypothesis is:

$$H_0: \theta_{2j} = 0$$
$$H_1: \theta_{2j} \neq 0, j = 1,2, \ldots, k. \quad (22)$$

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

The statistical test for testing the hypotheses in Equation (22) is Wald statistic, and formulated by

$$W_2 = \frac{\hat{\theta}_{2j}}{\widehat{SE}\left(\hat{\theta}_{2j}\right)} \tag{23}$$

where $\hat{\theta}_{2j}$ is the estimated value of the maximum likelihood parameter of the generalized Poisson regression model obtained by the Fisher-scoring method in Equation (17). $\widehat{SE}\left(\hat{\theta}_{2j}\right) = \left[\hat{V}\left(\hat{\theta}_{2j}\right)\right]^{1/2}$ is the standard error estimated value of the maximum likelihood parameter of the generalized Poisson regression model obtained from the main diagonal elements of the variance-covariance matrix, $Cov\left[\widehat{\boldsymbol{\theta}}_2\right] = \mathbf{I}^{-1}\left[\widehat{\boldsymbol{\theta}}_2\right]$ where $\mathbf{I}\left[\widehat{\boldsymbol{\theta}}_2\right]$ is the Fisher information matrix [13].

The Wald statistic (W_2) in Equation (23) is asymptotically standard normal distributed [16] so that the critical region at the significance level (α) to test the hypothesis in Equation (22) is the null hypothesis is rejected when the value of W_2 is greater than the value of $Z_{\alpha/2}$ (i.e., $|W_2| > Z_{\alpha/2}$) or the null hypothesis is rejected when the p-value is less than the α value.

3. METHODOLOGY

3.1. Data Sources

The data in this study is secondary data obtained from the Ministry of Health of the Republic of Indonesia [18] and the Central Statistics Agency of the Republic of Indonesia [19]. This research unit is all provinces in Indonesia in 2021, namely 34 provinces.

3.2. Research Variables

The research variables used in this study contain the response (Y) and the covariates (X_j), for $j = 1,2,\ldots,8$, which are presented in Table 1.

Table 1. Research Variables

Variables	Variables Description	Variables Type
Y	Low birth weight in Indonesia	Discrete
X_1	Poverty rate	Continuous
X_2	Percentage of households occupying livable houses	Continuous
X_3	Percentage of food processing places that meet the requirements according to the standard	Continuous
X_4	Percentage of households that have access to safe drinking water	Continuous
X_5	Percentage of households that have access to proper sanitation	Continuous
X_6	Percentage of pregnant women at risk of chronic energy deficiency receiving additional food	Continuous
X_7	Percentage of pregnant women who received blood-boosting tablets	Continuous
X_8	Percentage of antenatal care	Continuous

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

3.3. Data Analysis Techniques

The techniques of data analysis in this study are as follows:

1. Analyzing the statistical description of the research variables.
2. Detecting the multicollinearity of covariates.
3. Modeling the low birth weight in Indonesia using Poisson regression.
4. Detecting over-dispersion.
5. Modeling the low birth weight in Indonesia using generalized Poisson regression.
6. Getting the factors that affect the low birth weight in Indonesia.
7. Interpreting the generalized Poisson regression model of low birth weight in Indonesia.
8. Getting the conclusions.

4. RESULTS AND DISCUSSION

4.1. Statistical Descriptive Analysis

Analyzing and modeling the low birth weight in Indonesia using generalized Poisson regression begins with the descriptive statistical analysis of research variables. The results are shown in Table 2.

Table 2. Statistical Descriptive Results of Research Variables

Variables	Mean	Standard Deviation	Maximum	Minimum
Y	3,286	4,775	22,574	177
X_1	10.76	5.40	26.86	4.53
X_2	60.14	12.16	85.15	27.60
X_3	51.59	13.88	81.10	16.50
X_4	86.68	8.46	99.86	64,92
X_5	80.97	9.93	97.12	40,81
X_6	90.00	12.57	100	42.20
X_7	79.46	11.43	92.6	37.20
X_8	81.31	16.45	100	13.00

Table 1 shows that Indonesia's average low birth weight in 2021 was 3,286, with a standard deviation of 4,775. The highest and lowest, 22,574 and 177, were found in West Java Province and North Sulawesi Province, respectively. One of the reasons for the high low birth weight in West Java Province compared to North Sulawesi Province is the larger population in West Java Province. The visualization of the distribution of low birth weight in Indonesia in 2021 is presented in Figure 1.

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

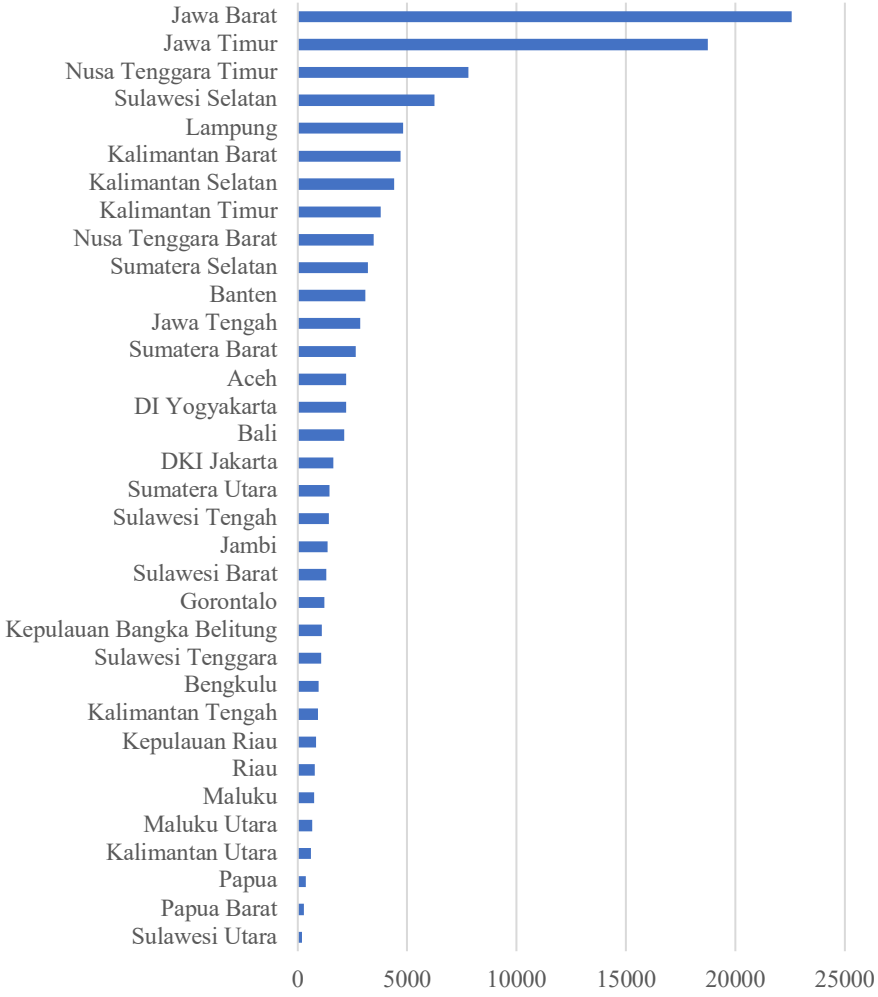

Fig 1. Distribution of Low Birth Weight in Indonesia in 2021

4.2. Detecting Multicollinearity

Multicollinearity is a problem in the generalized Poisson regression modeling; namely, the covariates are correlated to each other. This study's multicollinearity detection uses the Variance Inflation Factor (VIF) [20]. The generalized Poisson regression model has a multicollinearity problem when the VIF value of covariates is greater than 10. The VIF value of all covariates given in Table 3 shows that all covariates have a VIF value of less than 10. Therefore, there is no multicollinearity, and all of covariates can model low birth weight using the generalized Poisson regression model.

Table 3. The VIF Values of Covariates

Covariates	VIF Values
X_1	2.2160
X_2	1.7668
X_3	1.1652
X_4	1.9131

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

Covariates	VIF Values
X_5	2.3129
X_6	1.8983
X_7	2.5916
X_8	2.7612

4.3. Modeling Low Birth Weight Using Poisson Regression

The modeling of low birth weight in Indonesia in 2021 using Poisson regression begins with estimating and significance testing of the Poisson regression model parameters, were displayed in Table 4.

Table 4. Parameter Estimates, Standard Error, and Statistical Test Values of the Partial Test for the Poisson Regression Model

Parameter	Estimate	Standard Error	W_1	p-value
θ_{10}	1.7986	0.0746	24.1099	$< 2 \times 10^{-16*}$
θ_{11}	0.0524	0.0008	65.5	$< 2 \times 10^{-16*}$
θ_{12}	0.0067	0.0005	13.4	$< 2 \times 10^{-16*}$
θ_{13}	-0.0011	0.0002	-5.5	$7.54 \times 10^{-7*}$
θ_{14}	0.0433	0.0005	86.6	$< 2 \times 10^{-16*}$
θ_{15}	-0.0573	0.0006	-95.5	$< 2 \times 10^{-16*}$
θ_{16}	-0.0266	0.0004	-66.5	$< 2 \times 10^{-16*}$
θ_{17}	0.0585	0.0005	117	$< 2 \times 10^{-16*}$
θ_{18}	0.0446	0.0005	89.2	$< 2 \times 10^{-16*}$

*) Indicates significance at the significance level, $\alpha = 0.1$.

Based on Table 4, the Poisson regression model was obtained, and can be written as follows:

$$\hat{\zeta}_1(\mathbf{x}) = 1.7986 + 0.0524X_1 + 0.0067X_2 - 0.0011X_3 + 0.0433X_4 - 0.0573X_5 - 0.0266X_6 + 0.0585X_7 + 0.0446X_8. \quad (24)$$

The simultaneous influence of the hypothesis was carried out using Wilk's lambda statistic in Equation (8). The hypothesis was formulated as follows:

$H_0: \theta_{11} = \theta_{12} = \cdots = \theta_{18} = 0$

H_1: at least one of $\theta_{1j} \neq 0, j = 1,2, \ldots,8$.

The Wilk's lambda statistic value was 80,105.23, and the $\chi^2_{(\alpha,k_1)}$ value was 13.3616 with a p-value was less than 0.001 (i.e., $p < 0.001$). Therefore, the null hypothesis was rejected and it can be concluded that the poverty rate, percentage of households occupying livable houses, percentage of food processing places that meet the requirements according to the standard, percentage of households that have access to safe drinking, percentage of households that have access to proper sanitation, percentage of pregnant women at risk of chronic energy deficiency receiving additional food, percentage of pregnant women who received blood-

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

boosting tablets, and percentage of antenatal care were simultaneously significantly influencing the low birth weight in Indonesia.

The partial test was used to obtain covariates that significantly influencing the low birth weight in Indonesia. This test was employed by the Wald statistic in Equation (10), which has the hypotheses as follows:

$H_0: \theta_{1j} = 0$
$H_1: \theta_{1j} \neq 0, j = 1,2, \ldots,8.$

Based on Table 4, the Wald statistic value for all parameters ($|W_1|$) was more than the value of $Z_{\alpha/2}$, and the p-value for all parameters was less than the α value. Therefore, the null hypothesis was rejected, and the conclusion was poverty rate, percentage of households occupying livable houses, percentage of food processing places that meet the requirements according to the standard, percentage of households that have access to safe drinking, percentage of households that have access to proper sanitation, percentage of pregnant women at risk of chronic energy deficiency receiving additional food, percentage of pregnant women who received blood-boosting tablets, and percentage of antenatal care were partially significantly influencing the low birth weight in Indonesia.

4.4. Detecting Over-dispersion

Overdispersion detection is done by comparing the variance value of the response to the average value and the deviation value of the Poisson regression model divided by the degree of freedom. If the variance value of the response is more than the average value, then overdispersion occurs. Meanwhile, if the value of the deviation divided by the value of degrees of freedom is more than 1, then overdispersion occurs. Based on the descriptive statistical analysis result in Table 2, the variance value of low birth weight was 22,797,386 and the average value was 3,286. Since the variance value is more than the average value, overdispersion occurs. Based on the results of Poisson regression modeling, the deviance value was 58,108.76, and the degrees of freedom value was 25. The deviance divided by the degrees of freedom were more than 1. These results indicate overdispersion and show that there is overdispersion in Poisson regression.

4.5. Modeling Low Birth Weight Using Generalized Poisson Regression

Since there is a problem of overdispersion in Poisson regression, the modeling of low birth weight in Indonesia in 2021 needs to be adequately modeled using Poisson regression. Therefore, generalized Poisson regression is one of the appropriate models to model. The result of modeling low birth weight in Indonesia in 2021 using generalized Poisson regression is presented in Table 5.

Table 5. Parameter Estimates, Standard Error, and Statistical Test Values of the Partial Test for the Generalized Poisson Regression Model

Parameter	Estimate	Standard Error	W_2	p-value
θ_{20}	4.4497	1.9240	2.3127	0.0207*
θ_{21}	0.0494	0.0278	1.7770	0.0758*
θ_{22}	-0.0014	0.0116	-0.1207	0.9066
θ_{23}	0.0012	0.0077	0.1558	0.8743

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

Parameter	Estimate	Standard Error	W_2	p-value
θ_{24}	0.0246	0.0163	1.5092	0.1324
θ_{25}	-0.0340	0.0162	-2.0988	0.0355*
θ_{26}	-0.0193	0.0116	-1.6638	0.0948*
θ_{27}	0.0374	0.0149	2.5101	0.0124*
θ_{28}	0.0291	0.0110	2.6455	0.0080*

*) Indicates significance at the significance level, $\alpha = 0.1$.

The generalized Poisson regression can be obtained based on the parameter estimates results in Table 5, and it was expressed as follows:

$$\hat{\zeta}_2(\mathbf{x}) = 4.4497 + 0.0494X_1 - 0.0014X_2 + 0.0012X_3 + 0.0246X_4 \\ -0.0340X_5 - 0.0193X_6 + 0.0374X_7 + 0.0291X_8. \quad (25)$$

The simultaneous influence of the hypothesis was carried out using Wilk's lambda statistic in Equation (21). The hypothesis was formulated as follows:

$H_0: \theta_{21} = \theta_{22} = \cdots = \theta_{28} = 0$

$H_1:$ at least one of $\theta_{2j} \neq 0, j = 1,2, \ldots,8.$

Wilk's lambda statistic value was 15.0237, and the $\chi^2_{(\alpha,k_2)}$ value was 13.3616 with a p-value of 0.0587. Therefore, the null hypothesis was rejected, and the conclusion was poverty rate, percentage of households occupying livable houses, percentage of food processing places that meet the requirements according to the standard, percentage of households that have access to safe drinking, percentage of households that have access to proper sanitation, percentage of pregnant women at risk of chronic energy deficiency receiving additional food, percentage of pregnant women who received blood-boosting tablets, and percentage of antenatal care were simultaneously significantly influencing the low birth weight in Indonesia.

The partial test was used to obtain covariates significantly influencing the low birth weight in Indonesia. The Wald statistic in Equation (23) was applied in this test, which has the hypothesis as follows:

$H_0: \theta_{2j} = 0$

$H_1: \theta_{2j} \neq 0, j = 1,2, \ldots,8.$

The Wald statistic value ($|W_2|$) of the parameters of θ_{21}, θ_{25}, θ_{26}, θ_{27}, and θ_{28} in Table 5 was greater than the value of $Z_{\alpha/2}$ with the p-value was less than the α value. Therefore, the null hypothesis was rejected, and the conclusion was poverty rate, percentage of households that have access to proper sanitation, percentage of pregnant women at risk of chronic energy deficiency receiving additional food, percentage of pregnant women who received blood-boosting tablets, and percentage of antenatal care were partially significantly influencing the low birth weight in Indonesia.

Finally, the interpretation of the generalized Poisson regression model in Equation (25), especially for the significant covariates are as follows:

1) If the poverty rate (X_1) increases by 1%, the average low birth weight will increase by exp(0.0494) or 1.0506 times, where the other covariates are fixed.

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

2) If the percentage of households with access to proper sanitation (X_5) increases by 1%, then the average low birth weight will decrease by exp(-0.0340) or 0.9666 times, where the other covariates are fixed.
3) If the percentage of pregnant women at risk of chronic energy deficiency receiving additional food (X_6) increases by 1%, then the average of low birth weight will decrease by exp(-0.0193) or 0.9809 times, where the other covariates are fixed.
4) If the percentage of pregnant women who received blood-boosting tablets (X_7) by 1%, then the average of low birth weight will increase by exp(0.0374) or 1.0381 times, where the other covariates are fixed.
5) If the percentage of antenatal care (X_8) increases by 1%, then the average low birth weight will increase by exp(0.0291) or 1.0295 times, where the other covariates are fixed.

5. CONCLUSION

Generalized Poisson regression is an accurate regression technique for modeling and handling count data with overdispersion in Poisson regression. The generalized Poisson regression was developed from the generalized linear models. The maximum likelihood and Fisher-scoring methods were used to estimate the generalized Poisson regression model parameters, whereas the likelihood ratio test and Wald test methods can be employed to test the significance of parameters. The generalized Poisson regression model was applied to modeling low birth weight in Indonesia in 2021. The factor affecting the low birth weight in Indonesia based on the generalized Poisson regression model were: poverty rate, percentage of households with access to appropriate sanitation, percentage of pregnant women at risk of chronic energy deficiency receiving additional food, percentage of pregnant women who received blood-boosting tablets, and percentage of antenatal care. However, this study still needs to be continued by using a spatial regression approach for future research, such as geographically weighted generalized Poisson regression, because there is any spatial heterogeneity in the generalized Poisson regression model.

REFERENCES

[1] M. Fathurahman, I. Purnamasari and S. Prangga, "Negative binomial regression analys on dengue hemorrhagic fever cases in East Kalimantan Province," *AIP Conferen Proceedings,* vol. 2668, no. 070002, pp. 1-6, 2022.

[2] J. M. Hilbe, Modeling Count Data, New York: Cambridge University Press, 2014.

[3] P. C. Consul and F. Famoye, "Generalized Poisson regression model," *Commun. Sta Theor. Method.,* vol. 21, no. 1, pp. 89-109, 1992.

[4] F. Famoye, "Restricted generalized Poisson regression," *Commun. Stat. Theor. Metho* vol. 22, no. 5, pp. 1335-1354, 1993.

[5] A. Prahutama and Sudarno, "Modelling infant mortality rate in Central Java, Indones use generalized Poisson regression method," *J. Phys.: Conf. Ser.,* vol. 1025, no. 01210 pp. 1-9, 2018.

[6] World Health Organization, Global nutrition targets 2025: low birth weight, Genev WHO, 2014.

[7] I. Hartiningrum and N. Fitriyah, "Bayi berat lahir rendah (BBLR) di Provinsi Jawa Tim tahun 2012-2016," *J. Biometrika dan Kependud,* vol. 7, no. 2, pp. 97-104, 2018.

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]

[8] K. Rajashree, "Study on the factors associated with low birth weight among newborn delivered in a Tetiary-Care Hospital, Shimoga, Karnataka," *Int. J. Med. Sci. Publi Health,* vol. 4, no. 9, pp. 1287-1290, 2015.

[9] The Ministry of Health of the Republic of Indonesia, Guidelines for Managing Chron Energy Deficiency (KEK) in Pregnant Women, Jakarta: The Ministry of Health of tl Republic of Indonesia, 2015.

[10] R. Sutan, M. Mazlina, N. M. Aimi and M. T. Azmi, "Determinant of low birth infants: matched case control study," *Open J. Prev. Med,* vol. 4, no. 3, pp. 91-99, 2014.

[11] K. P. N. Perera and K. Manzur, "Socio-economic and nutritional determinants of low bir weight in India," *N. Am. J. Med. Sci,* vol. 6, no. 7, pp. 302-308, 2014.

[12] M. Fathurahman, "Pemodelan indeks pembangunan kesehatan masyarak kabupaten/kota di Pulau Kalimantan menggunakan pendekatan regresi probit," *J. Varia* vol. 2, no. 2, pp. 47-54, 2019.

[13] M. Fathurahman, "Regresi binomial negatif untuk memodelkan kematian bayi Kalimantan Timur," *J. Eksponensial,* vol. 13, no. 1, pp. 79-86, 2022.

[14] M. Fathurahman, "Inverse Gaussian regression modeling and its application in neonat mortality cases in Indonesia," *BAREKENG: J. Math. Appl.,* vol. 16, no. 4, pp. 1197-120 2022.

[15] A. Agresti, Foundations of Linear and Generalized Linear Models, New Jersey: Joh Wiley & Sons, 2015.

[16] Y. Pawitan, All Likelihood: Statistical Modelling and Inference Using Likelihood, 1st ec Oxford: Clarendon Press, 2001.

[17] A. C. Cameron and T. P. K, Regression Analysis of Count Data, 2nd ed., New Yor Cambridge University Press, 2013.

[18] The Ministry of Health of the Republic of Indonesia, The Indonesian Profile of Heal 2021, Jakarta: The Ministry of Health of the Republic of Indonesia, 2022.

[19] The Central Statistics Agency of the Republic of Indonesia, Indonesian Statistics, Jakart The Central Statistics Agency of the Republic of Indonesia, 2022.

[20] D. N. Gujarati and D. C. Porter, Basic Econometrics, 5th ed., New York: McGrav Hill/Irwin, 2009.

[DOI: 10.14710/JSUNIMUS.11.1.2023.45-60]