Asian Journal of Management Entrepreneurship and Social Science
ISSN: 2808 7399
https://ajmesc.com/index.php/ajmesc Volume 04 Issue 01

Modeling The Number of Toddler Pneumonia Sufferers in DKI Jakarta using Negative Binomial Regression

Lucky Simarda[1], Dian Lestari[2*], Fevi Novkaniza[3], Arman Haqqi[4], Sindy Devila[5]

[1,2,3,4,5] Department of Mathematics , Faculty of Mathematics and Natural Science, Universitas Indonesia

*Coresponding author: dian.lestari@sci.ui.ac.id

Abstract

Acute lung tissue infection caused by various microorganisms, including fungi, viruses, and bacteria, is known as pneumonia. Pneumonia is the highest cause of child death worldwide. In Indonesia, pneumonia remains the leading cause of death among toddler (12-59 months old). By 2021, the national coverage of pneumonias among toddler was 34.8%, and the provinces with the highest coverage for toddler pneumonia were DKI Jakarta (53.0%), Banten (46.0%), and West Papua (45,7%). To find out the pattern of the relationship between the number of young people with pneumonia and the variables that affect it, a custom mathematical model is needed. The number of cases of toddler pneumonia in DKI Jakarta is a data count distributed by Poisson. Poisson regression is perfectly suitable for analyzing data that qualifies equidispersion. However, on the data, the number of toddler pneumonia cases in DKI Jakarta does not meet the equidispersion condition because the variance value is greater than the average or is called overdispersion. One of the methods developed to deal with overdispersion is negative binomial regression. The analysis showed that the average case of toddler pneumonia in Jakarta DKI was 454, Duren Sawit district recorded the highest case of 1329 cases and Sawah Besar district recorded the lowest case as 50 cases. The AIC criteria indicate that the Negative Binomial Regression model is a suitable model for modeling the number of cases of toddler pneumonia in Jakarta DKI with the smallest AIC value of 592,57. The best modeling results using the negative binomial regression method show two significant variables, they are the numbers of toddlers given exclusive breastfeeding and the numbers toddlers that were affected by covid-19.

Keywords: Count Data, Overdispersion, Poisson Regression

1. INTRODUCTION

Pneumonia remains a significant public health challenge worldwide, especially among toddlers residing in urban settings. In DKI Jakarta, Indonesia, the prevalence of pediatric pneumonia stands as a critical concern, necessitating a comprehensive understanding of its determinants and trends. By 2021, the nationwide coverage of toddler pneumonia was 34.8%, and the provinces with the highest coverage for toddler pneumonia are DKI Jakarta 53.0%, Banten 46.0%, and West Papua 45,7% (Directorate General Of Disease Prevention and Control, Indonesian Ministry of Health, 2021). Addressing this issue requires robust analytical approaches capable of untangling the intricate web of factors contributing to toddler pneumonia cases. This study aims to employ Negative Binomial Regression (NBR) as a powerful statistical tool to model and predict the incidence of toddler pneumonia in DKI Jakarta.

The complexities surrounding the epidemiology of toddler pneumonia are multifaceted and influenced by a myriad of socio-environmental factors. Previous research (Zhang Xian, et al, 2016) has underscored the significance of variables such as socio-economic conditions, housing quality, healthcare accessibility, air quality, and demographic attributes in shaping pneumonia outcomes among toddlers. However, synthesizing these variables within a unified analytical framework employing Negative Binomial Regression remains a novel pursuit in understanding toddler pneumonia dynamics in this specific urban context.

Based on research conducted by Irza, Anindya, and Resa (2020) on risk factors affecting pneumonia in children in Bandung City with analysis using negative binomial regression methods, significant risk factors are the percentage of malnutrition and population density

The choice of Negative Binomial Regression as the primary analytical method is predicated on its suitability for count data analysis and its ability to handle overdispersion commonly encountered in epidemiological datasets (Ismail & Jemain, 2007). Unlike conventional Poisson regression, NBR accommodates situations where the variance exceeds the mean, making it particularly relevant when modeling infrequent but impactful events like severe pneumonia cases among toddlers. By employing NBR, this study endeavors to uncover nuanced relationships between predictor variables and the reported incidence of pneumonia among this vulnerable demographic.

Furthermore, beyond elucidating individual factor contributions, the application of Negative Binomial Regression offers the potential for predictive modeling. By harnessing historical data encompassing diverse socio-environmental determinants, this research aspires to construct a predictive model capable of forecasting toddler pneumonia occurrences in DKI Jakarta. Such predictive insights hold immense promise for healthcare planning, resource allocation, and targeted intervention strategies aimed at mitigating the burden of toddler pneumonia in this urban setting.

This study answered two research questions which are: (i)How is the modeling of the number of cases of toddler pneumonia in DKI Jakarta with the method of Binomial Regression Negative? (ii) What variable explains the number of toddler suffering from pneumonia in DKI Jakarta? In line with the research questions, two research objectives were answered. The two research objectives are: (i) Modeling the number of pneumonia suferrers in toddler in DKI Jakarta with the method of Binomial Regression Negative. (ii)Identifying the variable that explains the amount of people suffering from toddler pneumonia in DKA Jakarta. Structure of this study includes introduction, literature review, research methodology, results, discussion and conclusion. Next section presents literature review.

2. LITERATURE REVIEW

2.1 Pneumonia

Acute infection of lung tissue caused by various microorganisms, including fungi, viruses and bacteria, is known as pneumonia. According to the World Health Organization, the lungs consist of small sacs called alveoli, which fill with air when a person breathes healthily. However, for pneumonia sufferers, the alveoli contain mucus or fluid which limits the flow of oxygen and makes breathing painful because the body's cells cannot function (WHO, 2022).

2.2 Risk Factors for Pneumonia

The high incidence of pneumonia cannot be separated from the risk factors for pneumonia. Risk factors that have been identified include nutritional status, low birth weight, lack of exclusive breastfeeding in the first six months of life, measles immunization and housing overcrowding (five or more people per room) (UNICEF/WHO, 2006). Based on the Disease Control and Environmental Health Profile (2012), risk factors that influence the incidence of pneumonia are poor nutrition, low levels of exclusive breastfeeding, indoor air pollution, residential density, low measles immunization coverage, and low birth weight.

During the COVID-19 pandemic, pneumonia deaths increased by more than 75%. Pneumonia cases are associated with COVID-19. Before the emergence of the COVID-19 pandemic, most pneumonia cases were caused by people's habits such as smoking and living an unhealthy lifestyle. However, since the emergence of the pandemic, pneumonia has become the most common complication of COVID-19 (Yasmin, 2020).

2.3 Poisson Distribution

The Poisson distribution is a distribution probability of several types of events random (Buonaccorsi and Skibiel, 2005). Distribution on the values in a variable random X where is the number of outcomes obtained in certain events for states the probability if the average the incident was known in time which are mutually independent. Distribution characteristics Poisson is the number of trials carried out in certain areas depending on the results of the experiments that occurred, probability of occurring briefly corresponds with the intervals carried out, and the probability that more than that can be ignored (Feng, 2008). A random variable Y has a Poisson distribution if it satisfies the probability function in equation (1)

$$f(y,\lambda) = \frac{e^{-\lambda}\lambda^{y}}{x!}, y = 0,1,2,\dots and\ \lambda > 0 \quad (1)$$
$$= 0, \quad y\ others$$

where λ is a parameter that shows the average number within a certain interval range. The Poisson distribution has the special property that the variance is equal to the mean.

2.4 Negative Binomial Distribution

The Negative Binomial Distribution has many ways of approaching it. According to Hogg and Craig (1995), there are 12 ways to approximate the negative binomial distribution, including those that can be approximated by the Bernoulli test series and the mixed Poisson Gamma distribution. The classic approach that is often used is a series of Bernoulli trials, namely the number of Bernoulli trials needed to achieve k successes, where each repetition is independent, and the probability of success for each trial is constant p, while a failed trial is 1-p. For example, the random variable X represents the number of trials required for k successes to occur. Then X has a negative Binomial distribution with a probability function as in equation (2)

$$f(x) = \binom{x-1}{k-1} p^{k}(1-p)^{x-k} \quad , x = k, k+1, k+2, \dots \quad (2)$$
$$= 0 \quad , \text{x others}$$

According to De Jong and Heller (2008), the negative binomial distribution can be determined for every positive value of k by using the gamma function as a substitute for factorial and its combination, namely:

$$f(y) = \frac{\Gamma(y+k)}{\Gamma(k)y!} p^{k}(1-p)^{y}, y = 0,1,2,3,\dots \quad (3)$$

2.5 Generalized Poisson Distribution

The Generalized Poisson distribution is a two-parameter discrete distribution. One parameter measures location and the other parameter measures dispersion. Define X as a discrete random variable and X as a non-negative integer. Let $P_x(\theta,\lambda)$ be the probability of X given the value x. Thus X is said to have a Generalized Poisson Distribution with parameters θ and λ if:

$$P_x(\theta,\lambda) = \begin{cases} \frac{\theta(\theta+x\lambda)^{x-1}e^{-\theta-x\lambda}}{x!} & x = 0,1,2,\dots \\ 0 & for\ x > m, when\ \lambda < 0 \end{cases} \tag{4}$$

Where $\theta > 0, max\ (-1, -\theta/m)\ \leq \lambda \leq 1, and\ (m \geq 4)$ is the largest positive integer with $\theta + m\lambda > 0\ when\ \lambda < 0$. The parameters θ and λ are independent with lower bounds on λ and $m \geq 4$ to ensure that there are at least five classes with non-zero probability when λ is negative. The parameter θ is an indicator of the intensity of the natural Poisson process, while the parameter λ can be considered as a measure of deviation from the Poisson process. The Generalized Poisson distribution model reduces to a Poisson probability model when $\lambda = 0$ (Consul, 1989).

The mean and variance for the Generalized Poisson distribution are:

$$E(X) = \frac{\theta}{1-\lambda} \tag{5}$$

$$Var(X) = \frac{\theta}{(1-\lambda)^3} \tag{6}$$

If λ=0, then the probability function of the Generalized Poisson distribution will have a Poisson distribution so that $E(X) = Var(X)$. If $\lambda > 0$, then $E(X) < Var(X)$ this indicates overdispersion. Meanwhile, if $\lambda < 0$, then $E(X) > Var(X)$, this indicates underdispersion.

2.6 Generalized Linear Model (GLM)

Generalized Linear Model (GLM) is an extension of linear regression which assumes that the predictor variables have a linear effect, but the response variable is not assumed to have a certain distribution. GLM is used when the distribution of the response variable belongs to the exponential family (Agresti, 2013).

Changes to a generalized linear model show a form consisting of three parts, namely:

1. Random component, a GLM component that contains the response variable Y with independent observations ($y_1, \dots, y_n$) from a distribution in the exponential family.
2. Systematic component, GLM component that connects η_1 ,..., η_n to predictor variables through a linear model which is expressed as follows:
$$\eta_i = x_i^T \beta, i = 1,2, \dots, n \tag{7}$$
with $x_i^T = \begin{bmatrix}1 & x_{i1} \; x_{i2} \dots x_{ip}\end{bmatrix}^T$ and $\beta = [\beta_0 \quad \beta_1 \quad \beta_2 \dots \quad \beta_p]^T$ and n represents the number of observations.
3. The link between the random and systematic components (connecting function), namely a function that explains the expected value of the response variable Y which is connected to the predictor variables through a linear equation. Let $\mu_i = E(Y_i), i = 1, \dots, n.$. The model relates μ_i to η_i with $\eta_i = g(\mu_i)$. The identity link function can be expressed with the following equation:
$$g(\mu_i) = x_i^T \beta, i = 1,2, \dots, n \tag{8}$$

From the three components above, the connecting function will determine the appropriate model in GLM. If the GLM has the simplest link function ($g(\mu) = \mu$ or identity link), then the GLM is a linear model with a discrete response. In conclusion, GLM is a linear model for the transformed average of response variables that have an exponential family distribution (Agresti, 2013).

2.7 Maximum Likelihood Estimation

This research requires the Maximum Likelihood Estimation method to estimate parameters. To find out the estimated parameter values, a likelihood function is needed. Let $X_1, \dots, X_n$ is random samples that are mutually independent and have identical distribution with a probability density function $f(x; \theta), \theta \in \Omega$. Assume that θ is a scalar and its value is unknown. The likelihood function is given by an equation as in equation (7)
$$L(\theta) = L(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta), \theta \in \Omega \tag{9}$$
With $x = (x_1, \dots, x_n)'$. The log-likelihood function is defined as follows
$$l(\theta) = logL(\theta) = \sum_{i=1}^{n} \log f(x_i, \theta), \theta \in \Omega \tag{10}$$
Maximum Likelihood Estimation (MLE) is denoted as $\hat{\theta}. \hat{\theta} = \hat{\theta}(X)$ is the maximum likelihood estimator of θ if
$$\hat{\theta} = \max_{\theta \in \Omega} L(\theta; x) \tag{11}$$
(Hogg, McKean, & Craig, 2013)

In general, there is no final form of maximum likelihood estimator for Binomial Negative Regression. For this reason, a numerical method is used to obtain the estimated value of β. The iterative algorithm commonly used for Negative Binomial Regression is the Newton algorithm Raphson.

3. RESEARCH METHOD

The research carried out in this thesis is quantitative research by conducting literature studies, analysis of secondary data originating from the DKI Jakarta Provincial Health Service, the Central Agency for Statistics and Health Profiles of DKI Jakarta Province using the Negative Binomial Regression method with R software.

The research variables used in this study are divided into two, namely the response variable (Y) and the predictor variable (X). The response variable is the number of toddler pneumonia sufferers in DKI Jakarta in 2022.

Table 1. Variable of Research

Variable	Information	Data Scale
Y	Number of Pneumonia Cases in Toddlers	Ratio
X_1	Number of toddlers who are exclusively breastfed	Ratio
X_2	Number of Toddlers Who Get Vitamin A	Ratio
X_3	Number of Toddlers with Low Birth Weight (LBW)	Ratio
X_4	Total population	Ratio
X_5	Number of Clean Healthy Living Behaviors	Ratio
X_6	Number of Toddlers who received Complete Basic Immunization	Ratio
X_7	Number of Toddlers with Malnutrition	Ratio
X_8	Number of poor households	Ratio
X_9	Number of toddlers affected by Covid-19	Ratio

3.1 Multikollinearity

Multicollinearity occurs when there is a relationship between two predictor variables in the regression model which can be explained as a linear relationship between the predictor variables in the regression model. This can be detected through the correlation coefficient and Variance Inflation Factor (VIF) values. If the VIF value exceeds 10, then it indicates that there is a multicollinearity relationship in the regression model being used (D.C. Montgomerry, 2012). The VIF value can be formulated in equation (12).

$$VIF_j = \frac{1}{1-R_j^2}; j = 1,2, \dots, \text{k} \tag{12}$$

where R_j^2 is the coefficient of determination of the jth predictor variable regression model, where $0 \leq R_j^2 \leq 1$, so that $VIF_j \geq 1, \forall_{j.}$There is multicollinearity in the regression model when the VIF value > 10.

3.2 Poisson Regression Model

Poisson regression is a regression analysis method that is used to model data consisting of count data. The Poisson distribution has a mean parameter μ and a probability function in equation (13)

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!}\ ; y = 0, 1, 2, .. dan\ \mu > 0 \tag{13}$$

where, μ is the average number of events in a certain interval with *E(Y) = Var(Y) =* μ (A. Agresti, 2019). One of the characteristics of the Poisson distribution is when the mean is equal to the variance. However, in many cases, situations occur where the data variance is greater than the mean. This is known as overdispersion.

3.3 Overdispersion

Overdispersion is a condition that can occur when modeling uses the Poisson distribution. This matter occurs because the Poisson distribution has the same mean and variance, but in fact sometimes the variance is greater than the mean or vice versa, a situation like this is said with overdispersion (Fitrial & Fatikhurrizqi, 2021). To identify overdispersion, the method that can be used is to compare the ratio of variance to mean in an equation (14)

$$\theta = \frac{D(\widehat{\beta})}{df} = \frac{(-2\ln(\Lambda))}{n-k-1} = \frac{-2\ln\left(\frac{L(\widehat{\omega})}{L(\widehat{\Omega})}\right)}{n-k-1} \tag{14}$$

Overdispersion has an effect on the consistency of regression parameters, even though it is inefficient or subject to bias (A.C. Cameron and P.K. Triyedi, 1990). There are special conditions related to the dispersion parameter θ which influence the determination of the model being analyzed in equation (15).

$$\theta = \begin{cases} 0, model\ regresi\ poisson \\ > 0, model\ NBR\ overdispersi \\ < 0, model\ NBR\ underdispersi \end{cases} \tag{15}$$

Therefore, to deal with overdispersion, one option that can be used is to apply the Negative Binomial Regression method.

3.4 Negative Binomial Regression Model

Negative Binomial Regression is one of the solution methods used to overcome the problem of overdispersion. This model has a probability time function in equation (16)

$$f(y, \mu, \theta) = \frac{\Gamma(y+\theta^{-1})}{\Gamma(\theta^{-1})\Gamma(y+1)} \left(\frac{1}{1+\theta\mu}\right)^{\theta^{-1}} \left(\frac{\theta\mu}{1+\theta\mu}\right)^{y}, y = 0,1,2, ... \tag{16}$$

The Negative Binomial Regression Model is a combination of the Poisson and Gamma distributions. The negative binomial regression model has the following equation (17):

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) \tag{17}$$

The probability function of the Negative Binomial distribution is

$$f(y, \mu, \theta) = \exp\left(ln\left(\frac{\Gamma(y+\theta^{-1})}{\Gamma(\theta^{-1})\Gamma(y+1)}\right) + \frac{1}{\theta} ln\left(\frac{1}{1+\theta\mu}\right) + y\ ln\left(\frac{\theta\mu}{1+\theta\mu}\right)\right), y = 0,1,2, ... \tag{18}$$

The parameters in the negative binomial regression model are estimated using the Maximum Likelihood Estimation (MLE) method. To find the parameter values, Newton Rapshon iteration is used. The likelihood function of negative binomial regression can be described as follows:

$$L(\beta, \theta) = \prod_{j=1}^{n} \left(\frac{\Gamma(y_j+\theta^{-1})}{\Gamma(\theta^{-1})\Gamma(y_j+1)} \left(\frac{1}{1+\theta\mu_j}\right)^{\theta^{-1}} \left(\frac{\theta\mu_j}{1+\theta\mu_j}\right)^{y_j} \right) \tag{19}$$

The log-likelihood form is

$$\ln\left(L(\beta, \theta) = \sum_{j=1}^{n} \left(\ln\left(\frac{\Gamma(y_j+\theta^{-1})}{\Gamma(\theta^{-1})\Gamma(y_j+1)}\right) + \frac{1}{\theta} ln\left(\frac{1}{1+\theta\mu}\right) + y\ ln\left(\frac{\theta\mu}{1+\theta\mu}\right)\right)\right.$$

$$= \sum_{j=1}^{n} \left(\ln\left(\frac{\Gamma(y_j+\theta^{-1})}{\Gamma(\theta^{-1})\Gamma(y_j+1)} \right) + \ y_j ln\left(\frac{1}{1+\theta\mu} \right) - (y_j + \theta^{-1}) \ln\left(1+\theta\mu_j\right) \right) \quad (20)$$

The estimated values of β and θ are obtained by deriving the probability function ln to β and θ and equalizing them to zero. Then the Newton Raphson method is used to find a solution to the equation obtained.

3.5 Generalized Poisson Regression Model

Another GLM model that can be used to overcome overdispersion is the Generalized Poisson Regression (GPR) Model. The GPR model is the appropriate model to use when the data is in the form of count data to handle cases of overdispersion. GPR assumes that random variables follow a generalized Poisson distribution (Y. Setyorini, A. Melliana, H. Eko, S. Rosi and Purhadi, 2015). The parameters in the GPR model consist of μ which is the average parameter and the dispersion parameter θ. The GPR model is similar to Poisson regression in its distribution, as shown in equation (23)

$$f(y;\mu;\theta) = \left(\frac{\mu}{1+\theta\mu}\right)^{y} \left(\frac{1+\theta y}{y!}\right)^{y-1} \exp\left(\frac{-\mu(1+\theta y)}{1+\theta\mu}\right), y = 0,1,2,\dots; \mu > 0; -\infty < \theta < \infty \quad (21)$$

where, the mean and variance of the GPR model are adalah $E(Y) = \mu$ and $Var(Y) = \mu(1+\theta\mu)^2$ (A. Agresti, 2019). The GPR model is similar to the Poisson regression model, namely in equation (24)

$$\mu_i = \exp(x_i^T\beta) \quad (22)$$

Parameters in GPR are estimated using the MLE (Maximum Likelihood Estimation) method (Y. Setyorini, A. Melliana, H. Eko, S. Rosi and Purhadi, 2013). Parameter significance testing is divided into two types of testing, namely simultaneous testing and partial testing (Triyanto, Purhadi, B. W. Otok and S. W. Purnami, 2015). Testing the significance of parameters simultaneously uses the Maximum Likelihood Ratio Test (MLRT) method.

3.6 Model Feasibility Test

Test the suitability of model simultaneously by testing the deviance as follows:

$H_0: \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (all variables have no effect on the model)

$H_1: ther\ is\ at\ lest\ one\ \beta_j \neq 0, j = 0,1,2, \dots, p$ (there are variables that effect on the model)

The test statistic for feasibility of Negative Binomial Regression is Deviance. Deviance model is defined as twice the difference between the maximum log likelihood of the full model $L(\hat{\Omega})$ and the maximum log likelihood of the observed model L($\hat{\omega}$) or can be written:

$$D\hat{\beta} = -2ln\frac{L(\widehat{\omega})}{L(\widehat{\Omega})} = 2\left(\ln\left(\widehat{\Omega}\right) - \ln\left(\widehat{\omega}\right)\right) \quad (23)$$

Test criteria:

H_0 rejected at a specific α value if test statistics $D(\hat{\beta})$ more than $\chi^2_{(p,\alpha)}$, where p is the number of parameters of the model under the population minus the number of parameters under H_0.

After testing using deviance and there are variables that explain the model, then a significance test is carried out on each variable independently (partially), by looking at whether there are significant independent variables in the model. This partial test uses the following hypothesis:

$H_0: \beta_j = 0$ (absence of significant influence of predictor variables on respons variables)

$H_1: \beta_j \neq 0, j = 0,1,2, \dots, p$ (the existence of a significant influence of predictor variable on response variable)

The test statistics for partial significance are as follow:

$$Z_{count} = \left(\frac{\widehat{\beta_J}}{se\left(\widehat{\beta_J}\right)}\right)^2, j = 0,1,2, \dots, k \quad (24)$$

Where $\widehat{\beta_J}$ is a conjecture of parameters β_j and $se\left(\widehat{\beta_J}\right)$ is a standard allgation of error from $\widehat{\beta_J}$.

Test criteria: H_0 rejected if test statistics $|Z_{count}|$more than $Z_{\frac{\alpha}{2}}$ then the jth parameter is significant to model.

3.7 Criteria for Selection of the Best Model

The best model is the one that is able to explain the relationship between predictor variables and response variables according to certain criteria. AIC or Akaike Information Criterion is one of the most commonly used criteria for selecting the best model in statistics(Bozdogan H, 2000).

In regression modeling, the AIC criterion is used to identify significant variables in the model. The AIC value is closely related to the deviance of the model. The model is considered better if the deviation value is smaller. The deviation value will decrease as the comparison between the likelihood function under H_0 and the likelihood function under the population increases. Therefore, the best model has the smallest AIC and deviance as explained in equation (25)

$$AIC = -2\ln L(\hat{\beta}) + 2k \tag{25}$$

where $L(\hat{\beta})$ represents the likelihood of each estimated model, while k indicates the number of parameters in the model.

4. RESULT

4.1 Descriptive Statistics

Descriptive statistics are used to understand the characteristics of the variables used in this research. Descriptive statistics on the number of pneumonia cases in toddlers are explained in table 2.

Table 2 Descriptive Statistics of Number of Toddlers Pneumonia Cases

Variable	Mean	Varians	Min	Maks
Y	453,90	109939,9	50	1329
X_1	4344	7799509	180	11236
X_2	1589	1877892	84	8252
X_3	144,4	2622,34	20	236
X_4	63644	1441874668	2362	15275
X_5	48075	997517687	2980	12689
X_6	41,52	1135,046	1	127
X_7	6828	9803666	395	1389
X_8	245,5	19488,77	2	735
X_9	166,98	10088,58	4	489

From the data listed in Table 4.2, it can be concluded that of the 44 sub-districts in DKI Jakarta Province, the average number of pneumonia cases in children under five in 2022 is around 454 cases. Duren Sawit District recorded the highest number of cases, reaching 1329 cases, while Sawah Besar District recorded the lowest number of cases, only 50 cases.

4.2 Multicolinearity Test

In building a regression model, the assumption that there are no cases of multicollinearity must be met. A condition where there is a strong relationship between two predictor variables in the regression model. This can be thought of as a linear relationship between the predictor variables in the model. If the VIF value is greater than 10 then multicollinearity occurs in the regression model. The VIF value of each predictor variable is presented in Table 3

Table 3 VIF Value of Predictor Variable

Variable	VIF
X_1	5,04
X_2	1,56
X_3	3,39
X_4	13,73
X_5	9,27
X_6	1,55
X_7	3,86
X_8	1,68
X_9	1,31

Based on Table 3, there is one variable that does not meet the non-multicollinearity assumption because the VIF value is greater than 10. So there is a predictor variable that is correlated with another predictor variable, namely the variable X4. Therefore, delete the variable X4.

Table 4 New VIF Value of Predictor Variable

Variable	VIF
X_1	4,64
X_2	1,47
X_3	3,38
X_5	3,89
X_6	1,55
X_7	3,11
X_8	1,52
X_9	1,31

Based on Table 4, because all predictor variables have VIF values <10, the multicollinearity problem has been resolved. It was concluded that the predictor variables were not correlated with each other, there was no multicollinearity. Therefore, all variables can be used in modeling.

4.3 Poisson Regression Model

Based on the multicollinearity case test based on table 4, the results show that there is no high correlation between all predictor variables. Therefore, there are eight predictor variables available for poisson regression modeling. The estimation results of the Poisson regression model are presented in Table 5

Table 5. Poissonl Regression Model Parameter Estimation

Parameter	Estimation	Z count	SE	P-*value*
$\beta 0$	5,342	179,731	0,02972	0,0000000*
β_1	0,0001347	25,860	0,000005208	0,0000000*
$\beta 2$	-0,0000171	-2,307	0,000007397	0,0021034*
$\beta 3$	-0,004424	-15,710	0,0002816	0,0000000*
β_5	0,00000161	3,484	0,0000004622	0,0004950*
$\beta 6$	-0,0001515	-0.596	0,0002543	0,5514920
$\beta 7$	0,00002512	6,865	0,000003659	0,0000000*
$\beta 8$	-0,00001598	-0,236	0,00006760	0,8131830
$\beta 9$	0,002996	37,707	0,00007944	0,0000000*
Devians	4842,552			
AIC	5198,5			

*) significant with $\alpha = 5\%$

Based on Table 4.5, the deviation value obtained is 4842.9, which is greater than $\chi^2_{(0,05;8)}$ which was obtained at 15.507. It can be concluded that rejecting H_0 means that in cases of toddler pneumonia in DKI Jakarta there is at least one predictor variable that has a significant influence on the response variable. Test statistics for partial parameter tests $|Z_{count}|$ compared to $Z_{\frac{0,05}{2}} =$ 1,96. This table shows that the values of all variables $|Z_{count}|$ > 1,96 so reject H_0. This means that all variables have a significant influence on the Poisson regression. Therefore, the Poisson regression model equation formed based on table 4.5 is:

$$\hat{\mu} = exp(5,338 + 0,0001336X_1 - 0,00001809X_2 - 0,004392X_3 + 0,000001585X_5 + 0,00002488X_7 + 0,002987X_9) \qquad (26)$$

4.4 Overdispersion

Overdispersion can be identified using the residual deviation value and degrees of freedom. If the quotient of the residual deviation value with degrees of freedom is >1 then the model is said to be overdispersed. Therefore, it is necessary to carry out alternative analyzes to overcome cases of overdispersion, one of which is by using the Negative Binomial regression model. The residual deviation in the GLM Poisson regression model is 4842.9 with 37 degrees of freedom. The result for residual deviation and degrees of freedom is 130.89 > 1, so it can be concluded that there is a case of overdispersion in cases of toddler pneumonia in DKI Jakarta.

4.5 Negatif Binomial Regression Model

In the Poisson regression model, cases of overdispersion are overcome using the negative binomial regression method. Overdispersion will produce inefficient parameters. The results of Poisson regression modeling show that the ratio of deviation values to degrees of freedom is more than 1, which indicates that the number of cases of pneumonia under five in DKI Jakarta is overdispersion. Therefore, the negative binomial regression method is used.

Table 6. Negative Binomial Regression Model Parameter Estimation

Parameters	**Estimation**	**Z count**	**SE**	**P-*value***
$\beta 0$	4,966	18,505	0,02684	$0,0000000^*$
β_1	0,0001478	2,486	0,000005945	0,012936*
$\beta 2$	-0,00004008	-0,584	0,000006864	0,559285
$\beta 3$	-0,002096	-0,576	0,002773	0,449783
β_5	0,000001993	0,414	0,000004819	0,679207
$\beta 6$	0,00135	0,398	0,002848	0,690304
$\beta 7$	0,000008412	0,194	0,00004340	0,846326
$\beta 8$	-0,0001399	-0,205	0,0006815	0,837289
$\beta 9$	0,003400	3,865	0,0008798	$0,000111^*$
Devians	45,778			
AIC	602,8			

*) significant with $\alpha = 5\%$

Based on Table 6, the parameters $\beta_0, \beta_1, \beta_9$ have the value $|Z_{count}| > 1.96$ so rejecting H_0 then variables X_1 and X_9 are significant in the negative binomial regression model. Based on the significance of the variables in table 4.7, a negative binomial regression was carried out again without including predictor variables that were not significant. The results obtained from estimating the parameters of the negative binomial regression model with predictor variables X_1 and X_9 are as in table 7.

Table 7. Negative Binomial Regression Model Parameter Estimation (continued)

Parameter	Estimation	Z count	SE	P-*value*
$\beta 0$	4,777	26,899	0,1776	0,0000000*
β_1	0,0001517	5,196	0,00002919	0,000000203*
$\beta 9$	0,003260	4,016	0,0008118	0,0000592*
Devians	45,851			
AIC	592,57			

*) significant with $\alpha = 5\%$

According to table 7, the deviation value is 45.85, the significance level used is 5%, therefore, the value of $\chi^2_{(0,05;8)}$ obtained is 15.507, which indicates that the deviation value is greater than $\chi^2_{(0,05;8)}$. Thus, it can be concluded that rejecting H_0 means there is at least one predictor variable that has a significant influence. To test the parameters partially, we use the statistic $|Z_{count}|$ which is compared with $Z_{\frac{0.05}{2}} = 1.96$. Table 4.8 shows that all variables have a value of $|Z_{count}| > 1.96$ so reject H_0, which means all variables have a significant effect on the Negative Binomial regression. So the Negative Binomial Regression model equation formed based on table 7 is:

$$\hat{\mu} = \exp\,(4,777 + 0,0001517X_1 + 0,003260X_9) \qquad (27)$$

4.6 Generalized Poisson Regression Model

A GLM model that can also be an alternative to overcome cases of overdispersion is the Generalized Poisson Regression Model. The regression parameter (β) uses the Maximum Likelihood Estimation (MLE) method which is assisted by Newton Raphson iteration in the 15th iteration. Next, in the generalized Poisson regression model, the significance of parameters is tested simultaneously and partially. The estimation results of the GPR model are presented in table 8.

Tabel 8. Generalized Poisson Regression Model Parameter Estimation

Parameter	Estimation	Z count	SE	P-*value*
$\beta 0$ $\beta 0$	5,238 1,567	19,358 28,251	0,2706 0,05547	$0{,}0000000^*$ 0.0000000*
β_1	0,0001158	2,417	0,00004790	0,015632*
$\beta 2$	-0,00004688	-0,725	0,000006466	0,468444
$\beta 3$	-0,002697	-1,083	0,002490	0,278714
β_5	0,000001744	0,420	0,000004152	0,674441
$\beta 6$	0,0005725	-0.239	0,002398	0,811325
$\beta 7$	0,00002582	0,755	0,00003420	0,453010
$\beta 8$	0,0002644	0,444	0,0005996	0,656820
$\beta 9$	0,002573	3,484	0,0007386	$0{,}000494^*$
Devians	347,29			
AIC	603,004			

*) significant with $\alpha = 5\%$

Based on table 8, the deviation value is 347.29 > $\chi^2_{(0,05;8)}$ of 15.507, which means reject H_0 so that there is at least one significant variable in the Generalized Poisson regression model. Then, the partial test results in the parameters $\beta_{0,}\beta_1$, and β_9 having the value $|Z_{count}| > Z_{\frac{0.05}{2}} =$ 1.96, which means that if H_0 is rejected, the parameters $\beta_{0,}\beta_1$, and β_9 are significant for the GPR model. Of the eight predictor variables, there are two significant predictor variables, namely X_1 and X_9 because they have a $p - value \ < \ \alpha = 0.05$. Next, the GPR model parameter estimation was carried out again by not including predictor variables that were not significant. The results obtained by estimating the parameters of the GPR model with predictor variables $X_1 \ and \ X_9$ are as follows:

Tabel 9. Generalized Poisson Regression Model Parameter Estimation (continued)

Parameter	Estimation	Z count	SE	P-*value*
$\beta 0$ $\beta 0$	5,131 1,579	27,874 28,565	0,1841 0,055276	$0{,}0000000^*$ $0{,}0000000^*$
β_1	0,0001195	5,164	0,00002313	0,000000203*
$\beta 9$	0,002206	3,20	0,0006789	$0{,}0000592^*$

Devians	349,98			
AIC	600,30			

*) significant with $\alpha = 5\%$

According to table 9, the deviation value is 349.98, and the significance level is 5%. Therefore, the value of $\chi^2_{(0,05;8)}$is 15.507, which indicates that the deviation value is greater than $\chi^2_{(0,05;8)}$. Thus, it can be concluded that rejecting H_0 means that there is at least one predictor variable that has a significant influence on the response variable. To test the parameters partially, statistic $|Z_{count}|$ used by comparing with $Z_{\frac{0,05}{2}} = 1,96$. Based on Table 4.10, it shows that all variables have a value of $|Z_{count}|$> 1.96 so reject H_0, which means all predictor variables have a significant effect on the Generalized Poisson regression. So the GPR model equation formed based on table 9 is:

$$\hat{\mu} = \exp\,(5{,}131 + 1{,}579 + 0{,}0001195X_1 + 0{,}002206X_9) \qquad (28)$$

4.7 AIC

The final step in modeling the number of toddler pneumonia sufferers in DKI Jakarta is to determine the best model using the Akaike Information Criterion (AIC) value.

Table 10 AIC Value

Method	AIC Values
Regresi Binomial Negatif Regresi Generalized Poisson	592,57 600,30

Table 10. shows that the AIC values produced by the two models are different and the AIC value of the Negative Binomial Regression model has the smallest value, namely 592.57. So the Negative Binomial Regression model is suitable for modeling risk factors that influence toddlers pneumonia in DKI Jakarta.

5. DISCUSSION

Based on the analysis results, the NBR model is the best model compared to the GPR model because it has the smallest AIC value. The best model obtained is as follows.

$$\hat{\mu} = \exp\,(5{,}131 + 1{,}579 + 0{,}0001195X_1 + 0{,}002206X_9)$$

Based on the model obtained, it can be seen that of the 2 predictor variable coefficients, only 1 has a logical relationship to the number of cases of pneumonia under five in DKI Jakarta.

This variable is the number of toddlers affected by Covid-19 (X_9). This variable logically has a positive relationship with the number of toddlers' pneumonia and in the model it has a positive coefficient. Meanwhile, the variable number of toddlers who are exclusively breastfed (X_1) should have a negative relationship with the number of toddlers' pneumonia but has a positive coefficient.

Discrepancies in the direction of the relationship in the model can be caused by dependencies between predictor variables or because of the data collection process. The data used is the sum or total data for one year (2022) even though the development of the size of each variable in one year is not always constant so the final number does not necessarily represent the conditions in each month.

The number of toddlers who are given exclusive breast milk and the number of toddlers affected by Covid-19 has been statistically proven to influence the number of cases of toddler pneumonia, so local governments and related institutions need to pay attention to these risk factors to maintain and increase the coverage of exclusive breastfeeding for toddlers, and carry out socialization/counseling about clean living behavior, using masks, maintaining distance so that Covid-19 cases continue to decline. Thus, the risk of toddler pneumonia can be reduced to reduce the toddler mortality rate due to pneumonia, which is still very high in DKI Jakarta. A low infant/child mortality rate is an indicator of community welfare.

Suggestions for further research are to pay more attention to collinearity between predictor variables so that the interpretation of the resulting model is in accordance with the logic of thinking in accordance with the principles of related science. Future research is also expected to be able to compare the resulting models with various criteria for model goodness, not just AIC, apart from that it is also recommended to pay attention to spatial aspects considering that pneumonia is an infectious disease.

6. CONCLUSION

The results of the analysis show that the average number of pneumonia cases in children in DKI Jakarta is 454, with Duren Sawit District recording the highest cases at 1329 cases and Sawah Besar District recording the lowest cases at 50 cases. The results show that the eight influencing variables generally have quite low significance, and there are no cases of multicollinearity. The AIC criteria show that the Negative Binomial Regression model is a suitable model for modeling the number of cases of pneumonia in children in DKI Jakarta with the smallest AIC value, namely 592.57. The best modeling results using the Negative Binomial Regression method show two significant variables, namely the number of toddlers who are given exclusive breast milk and the number of toddlers who are affected by Covid-

19. Every unit increase in the number of toddlers with Covid-19 will increase the number of toddlers suffering from pneumonia by 0.326% times the initial response variable.

.

REFERENCES

Agresti, A. 2007. Categorical Data Analisys. Second Edition. New York:John Wiley and Sons,Inc.

Agresti, A. 2015. Foundation of Linear and Generalized Linear Models. Hokoben, New Jersey:John Wiley & Sons.

Anwar, A., & Dharmayanti I. (2014). Pneumonia in Children Under Five in Indonesia. National Journal of Public Health, 8(8), 369-365

Arisandi, A., Herdiani, E. T., & Sahriman, S. (2018). Application of Generalized Poisson Regression in Overcoming Overdispersion in Dengue Hemorrhagic Fever Data.18(2).

Banaszak, I. W., & Bręborowicz, A. (2013). Pneumonia in Children. Retrieved February 7, 2023, from http//cdn.intechopen.com-/pdfs-wm/42153.pdf

Bozdogan, H. 2000. Akaike's Information Criterion and Recent Development in Inforation Complexity, Mathematical Psychology,44,62-91

Brown, R., & McDaid, J. (2003). Factors Affecting Retirement Mortality. North American Actuarial Journal, 7, 24-43.

Buonaccorsi and Skibiel, (2005). Buonaccorsi, V. & Skibel, A. (2005). A "Striking" demonstration of Poisson distribution. Teaching Statistics, 27(1)

Cameron, C., & Trivedi. 1998. Regression Analysis of Count Data. Cambridge University Press.

Consul, P.C. (1989) Generalized Poisson Distributions: Properties and Applications. Marcel Dekker, New York

Darnah. (2011). Overcoming Overdispersion in Poisson Regression Models with Generalized Poisson Regression I Handling Overdispersion on Poisson Regression Models with Generalized Poisson Regression I. Exponential Journal, 2(2).

D.C. Montgomerry, (2012). Design and Analysis of Experiments. 8th Edition, John Wiley & Sons, New York.

Disease Control and Environmental Health Profile. (2012). Risk Factors that Influence the Incidence of Pneumonia. Jakarta : Ministry Health of the Republic of Indonesia.

Feng, et al. (2008). The Poisson-Dirichlet Distribution and Related Topics: Models and Asymptotic. Bandung: Ertama.

Fitrial, N., & Fatikhurrizqi, A. (2021). Modeling The Number of Covid-19 Cases In Indonesia Using The Poisson Regression and Negative Binomial Regression Approach. National Seminar on Official Statistics, 2020(1), 65-72. https://doi.org/10.34123/semnasoffstat.v2020i1.465.

Hocking, R. R. (1996). Methods and Applications of Linier Models: Regression and the Analysis of Variance. New York: John Wiley & Sons. Inc

Hogg, R. V., & Craig, A. T. (1995). Introduction to Mathematical Statistics (7th Ed.), Upper Saddle River. New Jersey: Prentice Hall.

Hilbe, J. M. 2011. Negative Binomial Regression. Cambridge University Press.

Indonesian Ministry of Health. (2021). Health Profile Indonesia 2021. Jakarta : Ministry Health of the Republic of Indonesia

Ismail, N & Jemain, A. 2007. Handling Overdispersion with Negative Binomial Negative and Generalized Poisson Regression Models. Casualty Actuarial Society Forum, 106.

Irza, Anindya, and Resa (2020). Identifying Factors That Influence Pneumonia Cases in Toddlers in Bandung City Using Negative Binomial Regression. Proceedings of the National Statistics Seminar, 9, 41. https://doi.org/10.1234/pns.v9i.57

Jamilatuzzahro, Caraka, R.E., & Herliansyah, R. (2018). Generalized Linear Model Application in R. Yogyakarta: Innosain.

Lindsey, J. K.. (1997). Applying Generalized Linear Models. New York: Springer.M. Fathurahman, "Selection of the best regression model using Akaike's information criterion and Schwarz information criterion," J. Inform. Mulawarman, vols. 4, no. 3, 2009.

Maya Santi, V., & Wiyono, A. (2021). Modeling the Number of Malaria Cases in Indonesia Using the Generalized Linear Model. Journal of Statistics and Its Applications, 5(1).

McCullagh, P., & Nelder, J.A. (1989). Generalized Linear Models. Monographs on Statistics and Applied Probability. New York: Chapman and Hall.

Saidi, S., Herawati, N., & Nisa, K. (2021). Modeling with generalized linear models on covid-19: Cases in Indonesia. International Journal of Electronics and Communications Systems, 1(1), 25–33

Triyanto, Purhadi, B. W. Otok and S. W. Purnami.(2015). Parameter estimation of geographically weigthed multivariate Poisson regression .Applied Mathematical Sciences. 9(82). 4081-4093

Yasmin, et al (2020). Caregiver's perception of COVID-19 vaccination, and intention to vaccinate their children against the disease: a questionnaire based qualitative study. Annals of Medicine & Surgery.

Y. Setyorini, A. Melliana, H. Eko, S. Rosi and Purhadi. (2015) The Comparison Of Generalized Poisson Regression And Negative Binomial Reression Methods In Overcoming Overdispersion. International Journal of Scientific & Technology.

Zhang Xian, et al. (2016). Pneumonia and influenza hospitalizations among children under 5 years of age in Suzhou, China, 2005-2011. Influenza and Other Respiratory Viruses, ISIRV.

WHO. 2020. Pneumonia. http://www.who.int/mediacentre. [Accessed January 25, 2022]