ISSN No:-2456-2165

Overdispersion Data Modeling Cases of Filariasis in East Lampung Province, Indonesia

Herawati, N.[1*], Fitriyani, A.[2], Saidi, S.[3], Setiawan, E.[4]
[1,2,3,4]Department of Mathematics, University of Lampung, Bandar Lampung, Indonesia

Meliantari, D.[5]
[5]Department of Management, Dian Nusantara University, Jakarta, Indonesia

Abstract:- The Poisson regression model is a statistical model that can be utilized to analyze the relationship between independent variable X in the structure of continuous, discrete or mixed data and dependent variable Y in the form of discrete data. This model has several assumptions that must be met, one of which is that the variance value of the dependent variable must be equal to the average (equidispersion). If the variance value is larger than the mean, data overdispersion will occur. This study will compare the performance of Quasi-Poisson regression, Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB) regression models on overdispersion data, namely filariasis cases data in east Lampung province, Indonesia. The results showed that the ZIP method was the best for modeling data overdispersion compared to Quasi-Poisson regression and ZINB method. This is indicated by the smallest QAIC, QAICc and RMSE values of the ZIP model when compared to the Quasi-Poisson and ZINB regression models.

Keywords:- *Overdispertion; Quasi-Poisson; zero inflated Poisson; zero inflated negative binomial.*

I. INTRODUCTION

Linear regression analysis is related to the analysis of the effect of one or more explanatory variables on the response variable [1]. This regression method can be used when the dependent variable Y is continuous data.If the dependent variable Y is discrete and the independent variable X is in the form of continuous, discrete, or mixed data, one of the regression models that can be used to analyze the relationship between the two variables is the Poisson regression model. The Poisson regression model requires several assumptions that must be met, one of which is the equidispersion assumption. However, it is often found that the data has variances that are smaller or larger than the average. The situation where the variance of the data is smaller than the average data is called underdispersion. Meanwhile, if the variance of the data is greater than the average, it is called overdispersion. Overdispersion can occur because of several factors such as the presence of a large variance of the response variable which may be caused by other variables whose heterogeneity is not observed, the influence of other variables that cause the dependence of the probability of an event on the previous event, the existence of outliers, there are excess zeros in dependent variable [2-4]. To find out easily whether there is overdispersion or overdispersion in the data, it can be seen whether the possibility of binomial experimental results will produce failure or success events in the response variable is excessive or not. If there are too many failure events in the response variable, it is reasonable to expect that the data will be overly dispersed. Conversely, if there are too many successful events in the response variable, it is suspected that there is underdispersion in the data [5]. The statistical technique that can be used to test whether there is overdispersion in the data is to divide the residual deviation by the degrees of freedom [6].

$$\theta_1 = \frac{D^2}{db}, \quad D^2 = 2\sum_{i=1}^{n}\{y_i\, ln\left(\frac{y_i}{\mu_i}\right)\} \tag{1}$$

withdb=n-p-1; p is the number of parameters including constants, n is the number of observations and D^2 is the deviation value. If the residual deviation is greater than 1 then it can be said that the data is overdispersion.When overdispersion is neglected and standard Poisson regression is applied to these data, this can lead to underestimation of standard errors and p-values. As a consequence increasing the chance of an inflated Type I error.

Various approaches such as the negative binomial distribution and other Poisson mixtures to model overdispersion data but there are only a few models for under dispersion data. Research by [7-10] have found that there are several methods that can be used to estimate parameters in overdispersion data such as Quasi-Poisson regression, ZIP and Negative Binomial Regression and several other methods on different cases data [8-9], Another study briefly reviews the performance of Poisson, Generalized Poisson, ZIP, ZIGP and ZINB [11]. This study will compare the performance of Quasi-Poisson, Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB) regression models on filariasis cases data in east Lampung Province, Indonesia based on QAIC, $QAIC_C$ and RMSE values.

II. POISSON REGRESSION MODEL

The Poisson discrete probability distribution is a probability distribution that states the occurrence of an event several times (y) within a certain time interval or space [1]. The parameter of the Poisson distribution is only one parameter, namely μ (lambda) which is the average number of events with probability function:

$$p(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!} \quad (2)$$

To model data that follows the Poisson discrete probability function, one can use the Poisson regression model which is the most commonly used statistical method for count data [3]. The Poisson Regression Model is one of the Generalized Linear Model (GLM) models that can be used to model count data and contingency tables. The Y variable is a value that follows the Poisson distribution. This model assumes that the logarithm of the expected value (average) can be modeled into a linear form by several unknown parameters. In the Poisson regression model, the covariates in the model are included by the inverse link function which describes the relationship of the linear predictor $x_i * \beta$ = ηi with the expected value of the response μ_i:

$$y = \mu_i + \varepsilon_i$$

$$\mu_i = exp^{(x_i * \beta)}$$

The estimation of the Poisson regression parameters, Maximum Likelihood Estimator (MLE) method is used to estimate the parameter β with conditional probability function:

$$P(y_i|x_{1i}, x_{2i}, x_{3i}, \dots, x_{pi}; \mu_i(x_{1i}, x_{2i}, x_{3i}, \dots, x_{pi})) = \frac{\left[e^{\left(\beta_0+\Sigma_{j=1}^{p}\beta_j X_{ji}\right)}\right]^{y_i} e^{\left[-e^{\left(\beta_0+\Sigma_{j=1}^{p}\beta_j X_{ji}\right)}\right]}}{y_i!}$$

The likelihood function is obtained by multiplying all conditional probability functions from y_i by x_i so that:

$$L(\beta) == \frac{\left[e^{\left(\Sigma_{i=1}^{n}\beta_0+\Sigma_{j=1}^{p}\beta_j X_{ji}\right)}\right]^{y_i} e^{\left[-e^{\left(\Sigma_{i=1}^{n}\beta_0+\Sigma_{j=1}^{p}\beta_j X_{ji}\right)}\right]}}{\prod_{i=1}^{n} y_i!}$$

ln likelihood function:

$$\ln L(\beta) = \ln\left(\left[e^{\left(\Sigma_{i=1}^{n}\beta_0+\Sigma_{j=1}^{p}\beta_j X_{ji}\right)}\right]^{y_i} + e^{\left[-e^{\left(\Sigma_{i=1}^{n}\beta_0+\Sigma_{j=1}^{p}\beta_j X_{ji}\right)}\right]} - \prod\nolimits_{i=1}^{n} y_i!\right)$$

$$= \left(e^{\left(\Sigma_{i=1}^{n} y_i(\beta_0+\Sigma_{j=1}^{p}\beta_j X_{ji})\right)}\right) - e^{\left(\Sigma_{i=1}^{n}\beta_0+\Sigma_{j=1}^{p}\beta_j X_{ji}\right)} - ln\prod\nolimits_{i=1}^{n} y_i!$$

The first derivative of $lnL(\beta)$to β_0:

$$\frac{\partial lnL(\beta)}{\partial(\beta_0)} = \sum_{i=1}^{n}\left\{y_i - e^{\beta_0+\Sigma_{j=1}^{p}\beta_j X_{ji}}\right\} = 0$$

Because it is known that the first derivative still contains parameters that cannot be solved by the linear method, it is maximized by the Newton Raphson numerical optimization method.

$$\hat{\beta}_{(i+1)} = \hat{\beta}_i - \left(\left(H(\beta)\right)^{-1}.U_i(\beta)\right)$$

When the data is overdispertion, Poisson regression is no longer a reliable statistical model. Other models such as Quasi-Poisson regression, zero inflated Poisson regression or zero inflated negative binomial regression are more reliable for overdispersion data[4].

A. *Quasi-Poisson Regression*

Quasi-Poisson regression is a Poisson regression model that is often used to model count data that suffers from overdispersion problems. This model is proven to be able to increase the value of the standard error in Poisson regression. Compared to Poisson regression, Quasi-Poisson regression can adjust the standard error value with the dispersion parameter value to produce a larger standard error. Let Y be a random variable such that$E(y_i) = \mu_i$and $var(y_i) = \mu_i k$.

To estimate the parameter β based on the chain theorem in Quasi-Poisson regression, use the following equation:

$$\frac{\partial lnL(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{\partial lnL(\beta)}{\partial \theta_i} \times \frac{\partial \theta_i}{\partial \mu_i} \times \frac{\partial \mu_i}{\partial \eta_i} \times \frac{\partial \eta_i}{\partial \beta_j} \right)$$

The maximum likelihood estimator βˆ can be estimated with the log likelihod function partially derived from θ_i, so the following equation is obtained:

$$\frac{\partial lnL(\beta)}{\partial \beta_j} = \frac{y_i - b'^{\theta_i}}{a(k)} = \frac{y_i - \mu_i}{a(k)}$$

Reducing μ_i to$\theta_i$we get,

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{var(y_i)}{a(k)}$$

Then using a linear g function that connects the average value and response variables with predictor variables, is derived from β_j, so we get$\frac{\partial \eta_i}{\partial \beta_i} = x_{ij}$. In order to obtain the chain theorem equation as follows:

$$\frac{\partial lnL(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{(y_i - \mu_i)x_{ij}}{var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} \right)$$

The above equation is a score function for Quasi likelihood. To get a solution of this equation, the iteration method is used with the Iteratively Reweighted Least Square (IRLS) procedure.

$$\hat{\beta}^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} z^{(t)})$$

B. Zero Inflated Poisson Regression

If the count data has an excess of zero counts, Zero Inflated Poisson Regression (ZIP)is one of the statistical methods that can be used to model the data [3]. This model is a combination of zero-valued event distributions with probability distribution function:

$$P(Y = y_i) = \begin{cases} \omega_i + (1 - \omega_i)e^{-\mu_i}, y_i = 0 \\ \dfrac{(1 - \omega_i)e^{-\mu_i}\mu_i^{y_i}}{y_i!}, y_i > 0 \end{cases}$$

The μ connection model on the ZIP regression is:

$$\ln(\mu) = \boldsymbol{X}_i^T \boldsymbol{\beta} \text{and logit}(\omega) = \ln\left[\frac{\omega}{1 - \omega}\right] = \boldsymbol{X}_i^T \boldsymbol{\gamma}$$

with $E(Y_i) = (1 - \omega_i)\mu_iandVar(Y_i) = (1 - \omega_i)(\mu_i + \omega_i {\mu_i}^2)$.

In the estimation of the Poisson regression parameters, β is estimated using the Maximum Likelihood Estimator (MLE) method. Partial derivatives of the probability function are performed on the estimated parameters:

$\mu_i = \exp(x_i^T \beta)$and $\omega_i = \frac{e^{x_i^T \gamma}}{1 + e^{x_i^T \gamma}}$.

$$P(Y = y_i) = \begin{cases} \dfrac{\exp(x_i^T \gamma) + \exp\left(-e^{x_i^T \beta}\right)}{1 + x_i^T \gamma}, y_i = 0 \\ \left(\dfrac{1}{1 + x_i^T \gamma}\right) \dfrac{\left(\exp\left(-e^{x_i^T \beta}\right)\right)\left(-e^{x_i^T \beta}\right)^{y_i}}{y_i!}, \\ y_i > 0 \end{cases}$$

Next, the ln likelihood equation is made from the equation above:

$$\ln L(\beta,\gamma \mid y_i) = \begin{cases} \sum_{i=1}^{n} \ln\left(e^{x_i^T\gamma} + \exp\left(-e^{x_i^T\beta}\right)\right) - \sum_{i=1}^{n} \ln\left(1+e^{x_i^T\gamma}\right), & y_i = 0 \\ \sum_{i=1}^{n} \left((x_i^T\beta)y_i - e^{x_i^T\beta}\right) - \sum_{i=1}^{n} \ln\left(1+e^{x_i^T\gamma}\right) - \sum_{i=1}^{n} \ln y_i!, & y_i > 0 \end{cases}$$

The EM (Expectation Maximization) algorithm is used to maximize the ln likelihood function. The EM algorithm is an optimization method commonly used to maximize the likelihood function. Suppose Y is related to the indicator variable Z:

$$Z = \begin{cases} 1, y_i = 0 \\ 0, y_i > 0 \end{cases}$$

The steps for parameter estimation using the EM (Expectation Maximization) algorithm are:

Determination of the distribution of Z variables with

$(z_i=1)=\omega_i$ and $(z_i=0)=1-\omega_i$

Formation of a combined distribution between y_i and z_i is:

$$f(y_i, z_i|\omega_i, \mu_i) = (1-\omega_i)^{(1-z_i)}(\omega_i)^{z_i}\left(\frac{\exp(-\mu_i)\,{\mu_i}^{y_i}}{y_i!}\right)^{(1-z_i)}$$

Then got the normal log function:

$$\ln L(\boldsymbol{\beta},\boldsymbol{\gamma}|y,z) = \sum_{i=\frac{1}{n}}^{n}[z_i x_i^T\boldsymbol{\gamma} - \ln(1+\exp(x_i^T\gamma))] + \sum_{i=\frac{1}{n}}^{n}(1-z_i)(y_i x_i^T\boldsymbol{\beta} - \exp(x_i^T\boldsymbol{\beta}))] - \sum_{i=\frac{1}{n}}^{n}(1-z_i)\ln y_i!$$

The above equation will be maximized using the EM algorithm, where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be estimated separately.

$$ln\, L(\beta,\gamma,y,z) = ln\, L(\beta,y,z) + ln\, L(\gamma,y,z) - \sum_{i=\frac{1}{n}}^{n}(1-z_i)\, ln\, y_i!$$

With

$$ln\, L(\gamma,y,z) = \sum_{i=\frac{1}{n}}^{n}[z_i x_i^T\gamma - \ln(1+exp(x_i^T\gamma))]$$

And

$$ln\, L\,(\beta,y,z) = \sum_{i=\frac{1}{n}}^{n}(1-z_i)(y_i x_i^T\beta - \exp(x_i^T\beta))]$$

The first stage in the EM Algorithm is the E-Step (expectation stage). Replace variables z_i dengan ${z_i}^{(k)}$ which is the expectation of z_i.

$$\ln\left(\boldsymbol{\gamma}^{(k)}, y, z^{(k)}\right) = \sum_{i=1}^{n}\left[z_i^{(k)} x_i^T\boldsymbol{\gamma}^{(k)} - \ln\left(1+\exp\left(x_i^T\boldsymbol{\gamma}^{(k)}\right)\right)\right]$$

$$\ln L\left(\boldsymbol{\beta}^{(k)}, y, z^{(k)}\right) = \sum_{i=1}^{n}\left(1-z_i^{(k)}\right)\left(y_i x_i^T\boldsymbol{\beta}^{(k)} - \exp\left(x_i^T\boldsymbol{\beta}^{(k)}\right)\right)$$

Each subsequent stage enters the maximization stage (M-step). Maximize β and γ in the equation by counting $\boldsymbol{\beta}^{(k+1)}$ and $\boldsymbol{\gamma}^{(k+1)}$ with Newton Raphson methode. Iterations are carried out until a convergent parameter estimator is obtained, namely at the time $|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}| < \varepsilon$ and $|\boldsymbol{\gamma}^{(k+1)} - \boldsymbol{\gamma}^{(k)}| < \varepsilon$.

C. Zero Inflated Negative Binomial Regression

As it is known that negative binomial regression can be used as an alternative method for modeling overdispersion data [4]. When there are excessive zeros in the count data, Zero Inflated Negative Binomial (ZINB) model can be used to overcome this problem. The probability distribution function of ZINB can be written as:

$$P(Y = y_i) = \begin{cases} p_i + (1 - p_i)\left(\frac{k}{\mu_i + k}\right)^k, y_i = 0 \\ (1 - p_i)\frac{\Gamma(k + y_i)}{\Gamma(y_i + 1)\Gamma(k)}\left(\frac{\mu_i}{\mu_i + k}\right)^{y_i}\left(\frac{k}{\mu_i + k}\right)^k, y_i > 0 \end{cases}$$

With mean and variance are $E(y_i) = (1 - p_i)\mu_i$ and $Var(y_i) = (1 - p_i)\mu_i(1 + \mu_i k^{-1} + p_i\mu_i)$ respectively.

The link function of μ in ZIP regression is:$\ln(\mu_i) = x_i^T\beta$and $logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = z_i^T\boldsymbol{\gamma},\ i = 1,2,\ldots n.$ To the estimate the parameters, partial derivatives of the probability function are performed. We get,

$$\mu_i = e^{x_i^T\beta}$$

$$p_i = \frac{e^{z_i^T\gamma}}{1 + e^{z_i^T\gamma}}$$

$$P(Y = y_i) = \begin{cases} \frac{1}{1 + e^{z_i^T\gamma}}\left[e^{x_i^T\beta} + \left(\frac{k}{e^{x_i^T\beta} + k}\right)^k\right], & y_i = 0 \\ \frac{1}{1 + e^{z_i^T\gamma}}\frac{\Gamma(\mathrm{k} + y_i)}{\Gamma(y_i + 1)\Gamma(\mathrm{k})}\left(\frac{e^{x_i^T\beta}}{e^{x_i^T\beta} + k}\right)^{y_i}\left(\frac{k}{e^{x_i^T\beta} + k}\right)^k, y_i > 0 \end{cases}$$

Next, the ln likelihood equation is made from the equation above:

$$\ln L(\theta \mid y_i) = \begin{cases} -\sum_{i=1}^{n}\ln\left(1 + e^{z_i^T\gamma}\right) + \sum_{i=1}^{n}\ln\left[e^{z_i^T\gamma} + \left(\frac{k}{e^{x_i^T\beta} + k}\right)^k\right], \\ -\sum_{i=1}^{n}\ln\left(1 + e^{z_i^T\gamma}\right) + \sum_{i=1}^{n}\ln[\Gamma(k + y_i)] - \sum_{i=1}^{n}\ln[\Gamma(y_i + 1)] - \\ \sum_{i=1}^{n}\ln[\Gamma(k)] + \sum_{i=1}^{n} y_i\ln\left(\frac{e^{x_i^T\beta}}{e^{x_i^T\beta} + k}\right) + \sum_{i=1}^{n} k\ln\left(\frac{k}{e^{x_i^T\beta} + k}\right), \\ y_i > 0 \end{cases}$$

The EM (Expectation Maximization) algorithm is applied to maximize the ln likelihood function. Suppose Y is related to the indicator variable W:

$$w_i = \begin{cases} 1, y_i = 0 \\ 0, y_i > 0 \end{cases}$$

$$ln\, L(\theta, y_i, w_i) = \sum_{i=1}^{n}\{w_i\, z_i^T\gamma - ln(1 + z_i^T\gamma) + (1 - w_i)\ln[g(y_i; \beta, \kappa)]\}$$

The EM algorithm is carried out which is usually called the E-Step (expectation stage). At this stage, (s) is used to symbolize iteration.

ISSN No:-2456-2165

$$\widehat{w}_i^{(s)} = \begin{cases} \left[1 + e^{-z_i^T \hat{\gamma}^{(s)}} \left(\frac{\hat{k}^{(s)}}{e^T \widehat{\boldsymbol{\beta}}^{(s)} + \hat{k}^{(s)}}\right)^{\hat{k}^{(s)}}\right]^{-1}, y_i = 0 \\ 0, \text{jika } y_i > 0 \end{cases}$$

And

$\left(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}^{(s)}\right) = \ln L\left(\left(\boldsymbol{\theta} \mid \boldsymbol{y}_{i'}, \boldsymbol{w}_i\right) \mid \boldsymbol{y}_i, \widehat{\boldsymbol{\theta}}^{(s)}\right) = \sum_{i=1}^{n} Q_{1i}\left(\boldsymbol{\gamma} \mid \widehat{\boldsymbol{\theta}}^{(s)}\right) + \sum_{i=1}^{n} Q_{2i}\left(\boldsymbol{\beta}, \boldsymbol{k} \mid \widehat{\boldsymbol{\theta}}^{(s)}\right)$. When $Q_{1i}\left(\boldsymbol{\gamma} \mid \widehat{\boldsymbol{\theta}}^{(s)}\right) = \widehat{w}_i^{(s)} \boldsymbol{z}_i^T \boldsymbol{\gamma} - \ln\left(1 + e^{z_i^T \gamma}\right)$

With

$$Q_{2i}\left(\boldsymbol{\beta}, \boldsymbol{k} \mid \widehat{\boldsymbol{\theta}}^{(s)}\right) = \left(1 - \widehat{w}_i^{(s)}\right) \ln\left[\frac{\Gamma(k + y_i)}{\Gamma(k)\Gamma(1 + y_i)} \left(\frac{e^{x[/\beta}}{e^{x^{\top}\beta} + k}\right)^{y_i} \left(\frac{k}{e^{e^{\top}[\beta} + k}\right)^{k}\right]$$

Each subsequent stage enters the maximization stage (M-step). Maximize β and γ in the equation by counting $\boldsymbol{\beta}^{(\boldsymbol{k+1})}$ and $\boldsymbol{\gamma}^{(\boldsymbol{k+1})}$ with Newton Raphson methode. Iterations are carried out until a convergent parameter estimator is obtained, which is $|\beta^{(k+1)} - \beta^{(k)}| < \varepsilon$ and $|\gamma^{(k+1)} - \gamma^{(k)}| < \varepsilon$.

III. METHODS

This research was conducted using filariasis cases data in east Lampung province, Indonesia in 2013 obtained from the Lampung Health Service website (n = 33). The dependent variable (Y) used is the number of cases of filariasis in each district and the independent variables are clean and healthy behavior (X_1), healthy homes (X_2) and sanitation (X_3). The first step taken in data analysis is to see the existence of multicollinearity through the VIF value.Then to evaluate the presence of overdispersion data and extra zero countsPearson's chi square and Deviance values of θ_1 are used. Further analysis was carried out by modeling the data with Quasi-Poisson regression, ZIP, and ZINB. To confirm whether a set of independent variables is simultaneously significant for the model or not G test was used with formula $G = -2\ln[L_0 - L_1] \sim \chi^2_{(\alpha,p)}$. In addition, to confirm whether each independent variable in the model is significant or not, the Wald test, $W_j = \frac{\hat{\beta}_j^2}{(SE(\hat{\beta}_j^2))^2}$, was used. The suitability of the model was evaluated based on the R square value of each model. Selection of the best model was obtained by comparing the Quasi Akaike Information Criteria (QAIC, QAICc) [12]. In addition, the root mean square error (RMSE) as an error-based model selection guidance that shows how much the difference between the estimated results and the value to be estimated was also applied. All these values were employed to determine which model is the best. The model that has the smallest QAIC, QAICcand RMSE was selected as the best model.

$$QAIC = -2\left(\frac{\log(L)}{k}\right) + 2p$$

$$QAIC_C = QAIC + \frac{2p(p+1)}{n - p - 1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{Y}_i)^2}{n}}$$

IV. RESULTS AND DISCUSSION

To begin our data analysis, we examined the presence of multicollinearity between independent variables in the filariasis filariasis cases data in east Lampung province, Indonesia using VIF. Correlation between independent variables was not found in data based on VIF values as shown in Table 1.

Table 1: VIF Value

Variable	VIF
X_1	1.161621
X_2	1.297452
X_3	1.156203

It can be seen in Table 1 that the VIF value of each independent variables <10. It indicates that there is no multicollinearity between independent variables. So, there is no multicollinearity for the explanatory variables for in the filariasis case data in East Lampung. Therefore, all independent variables in the data can be included in the regression model.

Next, the presence of zero inflation in the data was checked.It was found that the dependent variable in filariasis cases data in east Lampung province, Indonesia had zero counts of more than 60% as seen in Table 2. This means that the data is experiencing zero inflation.

Table 2: Analysis Zero Inflation

Value	Total	Precentage (%)
0	20	60
1	6	18
2	4	12
3	2	7
5	1	3
Total	33	100

It shows in Table 2 that the percentage of zero counts in dependent variable is 60%, this value is greater than the condition for zero inflation, which is 50%. This means that

filariasis cases data in east Lampung province, Indonesia is experiencing zero inflation. As it is known that the presence of zeros in the response variable in discrete data is common. However, this does not mean that zero count data is not important in statistical modeling. In most cases in discrete data modeling, zero count data provide significance meaning that must be included in the analysis.

Then a test was carried out for the existence of overdispersion in filariasis cases in east Lampung province, Indonesia by looking at the deviance value divided by the degrees of freedom (df). The results are shown in Table 3.

Table 3: Analysis of Overdispersion Data

Criteria	**Value**	**df**	**Deviance/df**
Deviance	47.57	29	1.64034

Theoretically, if the Deviance/df value is greater than one, then the data is overdispersion. As shown in Table 3, the deviation/df of filariasis cases in East Lampung, Indonesia is more than one. This denotes overdispersion of data.

After examining all possible problems such as multicollinearity, zero inflation and overdispersion, it is found that there is no multicollinearity between the independent variables but there are zero inflation and overdispersion. This is an indication that it is no longer possible to analyze filariasis cases data in east Lampung province, Indonesia using the Poisson regression model. The data should be analyzed using another suitable regression model for zero inflation and overdispersion data. The first is to analyze the data with Quasi-Poisson regression model, ZIP and ZINB. Table 4 shows the results of the parameter estimation analysis using the three models.

Table 4: Quasi-Poisson Regression Parameter Estimate using Quasi-Poisson Regression, ZIP and ZINB

Model	Parameter	Estimated Parameter	SE	*p-value*
Quasi-Poisson	β_0	0.7163	4.6330	0.8268
	β_1	-0.0125	0.0066	0.0204
	β_2	-0.0324	0.0683	0.5014
	β_3	0.02802	0.0428	0.3550
ZIP	β_0	1.0104	3.3610	0.7637
	β_1	-0.0019	0.0057	0.7365
	β_2	-0.0726	0.0522	0.1639
	β_3	0.0639	0.0311	0.0401
	γ_0	-1.6707	9.3879	0.8588
	γ_1	0.0230	0.0137	0.0943
	γ_2	-0.0996	0.1533	0.5159
	γ_3	0.0974	0.0717	0.1748
ZINB	β_0	1.0099	3.3611	0.7638
	β_1	-0.0019	0.0057	0.7365
	β_2	-0.0726	0.0522	0.1640
	β_3	0.0639	0.0311	0.0401
	γ_0	-1.6798	9.3880	0.8580
	γ_1	0.0230	0.0137	0.0942
	γ_2	-0.0994	0.1533	0.5164
	γ_3	0.0974	0.0718	0.1749

From Table 4, we get theQuasi-Poisson regression model is $\hat{\mu} = \exp(0.716382 - 0.012583X_1 - 0.032465X_2 + 0.02802X_3)$. The ZIP regression model is:

$$\ln(\mu) = 1.010491 - 0.001944X_1 - 0.072692X_2 + 0.063980X_3$$

$\text{logit}(\omega) = -1.67070 + 0.02306X_1 - 0.09962X_2 - 0.09740X_3$and the ZINB regression model is

$$\ln(\mu) = 1.009936 - 0.001944X_1 - 0.072682X_2 + 0.063979X_3$$

$$\text{logit}(p_i) = -1.67981 + 0.02307X_1 - 0.099492X_2 - 0.09741X_3$$

To inquire the significance of the independent variables on the response variables simultaneously ($H_0: \beta_j = 0$vs $H_1: \beta_j \neq 0$, j = 1, 2, 3) in the Quasi-Poisson, ZIP and ZINB regression models, G-test was undertaken. Reject H_0 if G-test $> \chi^2_{(0.05,3)}$ =7.815. Then to test the significance of the independent variables on the response variables individually ($H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$, j = 1, 2, 3) in the Quasi-Poisson regression model the ZIP and ZINB Wald test was applied and H_0 was rejected if$Wald\ test > \chi^2_{(0.05,1)} =$ 3.84. The test results are presented in Table 5.

Table 5: Quasi-Poisson Regression Wald Test

Model	Parameter	G-test	Wald test	Decision
Quasi-Poisson	$\beta 1$	Reject H_0	4.39810	Reject H_0
	$\beta 2$		0.22556	Accept H_0
	$\beta 3$		0.42688	Accept H_0
ZIP	$\beta 1$	Reject H_0	0.11323	Accept H_0
	$\beta 2$		1.93768	Accept H_0
	$\beta 3$		4.21512	Reject H_0
	$\gamma 1$		2.80040	Accept H_0
	$\gamma 2$		0.42206	Accept H_0
	$\gamma 3$		1.84124	Accept H_0
ZINB	$\beta 1$	Reject H_0	0.11323	Accept H_0
	$\beta 2$		1.93692	Accept H_0
	$\beta 3$		4.21471	Reject H_0
	$\gamma 1$		2.79876	Accept H_0
	$\gamma 2$		0.42107	Accept H_0
	$\gamma 3$		1.84059	Accept H_0

From Table 5, it can be seen that for the Quasi-Poisson Regression model the independent variables simultaneously affect the independent variables. The same results apply to the ZIP and ZINB models. This means that clean and healthy behavior (X_1), healthy homes (X_2) and sanitation (X_3) simultaneously or at least one of the independent variables affect the number of cases of filariasis in Lampung Province, Indonesia. In addition, the results of the Wald test as shown in Table 5 for the Quasi-Poisson regression model just clean and healthy living behavior (X_1) which affects filariasis cases in east Lampung province, Indonesia. Whereas for ZIP and ZINB only sanitation (X_3) has an effect on filariasis cases in East Lampung Province, Indonesia.

Based on the results of the G-test and Wald test above, it is necessary to further probe which regression model is the best in describing filariasis cases in east Lampung province, Indonesia. We tested the best model based on the R-square, QAIC, $\boldsymbol{QAIC_C}$ and RMSE values. The test results can be seen in Table 6 below.

Table 6: R-square, QAIC, QAICc, and RMSE

Model	R Square	QAIC	$\boldsymbol{QAIC_C}$	RMSE
Quasi-Poisson	0.1231	73.4045	75.6269	1.62588
ZIP	0.1378	57.3644	63.3644	1.13568
ZINB	0.1355	59.3657	67.1918	1.13569

Based on the results of the analysis of the R-square value it is known that the ZIP model gives the largest R-square value of 0.1378, although the difference in value with Quasi-Poisson (0.1231) and ZINB (0.1355) is very small. . This means that the independent variables in the ZIP model are able to explain the dependent variable by 13.78%, while the rest is explained by other external factors. Furthermore, if seen in Table 6, the ZIP model gives the lowest QAIC and QAICc values, namely 57.3644 and 63.3644 respectively. Furthermore, if seen in Table 6, the ZIP model gives the lowest QAIC and QAICc values, namely 57.3644 and 63.3644 respectively. Meanwhile, the Quasi-Poisson and ZINB models provide slightly higher QAIC and QAICc values than ZIP. Likewise for the RMSE value on ZIP. The RMSE value for ZIP (1.13568) is the smallest compared to the RMSE Quasi-Poisson and ZINB values, namely 1.62588 and 1.13569 for Quasi-Poisson and ZINB respectively.

From the results of comparing the R-square, QAIC, QAICc and RMSE values in Quasi-Poisson, ZIP and ZINB regression models, it can be stated that the best regression model that can explain the overdispersion of filariasis case data in the province of East Lampung Indonesia is ZIP regression model which is

$$\ln(\mu) = 1.0104 - 0.0019X_1 - 0.0726X_2 + 0.0639X_3$$

And

$$\text{logit}\omega) = -1.6707 + 0.0230X_1 - 0.0996X_2 - 0.0974X_3$$

ISSN No:-2456-2165

V. CONCLUSION

When over dispersion occurs, the Poisson regression model is no longer proper to use because the resulting model will produce invalid conclusions. With overdispersion in the data, Quasi-Poisson regression models, ZIP and ZINB are some of the models that can be applied to model data overdispersion. From the results of the study by comparing Quasi-Poisson, ZIP and ZINB regression models to filarias is case data in East Lampung Province, Indonesia, it was found that the R-square value of the Quasi-Poisson, ZIP and ZINB regression models did not have much difference. This proves that the application of the Quasi-Poisson, ZIP and ZINB regression models to filarias is case data in East Lampung Province, Indonesia gives quite good results. This can also confirm that these three models can be used to model data overdispersion. This can also confirm that these three models can be used to model data overdispersion. However, of the three models when regarded from the QAIC, QAICc and RMSE values, the ZIP regression model give the smallest values. This states that the best model that can describe filariasis case data in East Lampung Province of Indonesia based on the smallest QAIC, QAICc and RMSE values is the ZIP regression model.

REFERENCES

[1.] Frome E.L, Kutner, M.K, & Beauchamp, J.J. 1973. Regression analysis of Poisson distributed data. *Journal of the American Statistical Association* 68: 935-40.

[2.] McCullagh, P. &Nelder, J.A. 1983. *Generalized Linear Models*. Chapman and Hall, London.

[3.] Cameron, C.A. & Trivedi, P.K. 1998. *Regression Analysis of Count Data*. Cambridge University Pr, London.

[4.] Berk,R. &MacDonald.2008. *OverdispersionandPoisson Regression*.Springer, Philadelpia.

[5.] Jansakul, N., &Hinde, J. P. 2008. Score tests for extra-zero models in zero-inflated negative binomial models. *Communications in statistics-simulation and computation*. 38(1). 92-108.

[6.] Dean, C. B. 1992. Testing for overdispersion in Poisson and binomial regression models.*Journal of the American Statistical Association* 87: 451–457.

[7.] Yang, Z., J. W. Hardin, C. L. Addy, and Q. H. Vuong. 2007. Testing approaches foroverdispersion in Poisson regression versus the generalized Poisson model. *BiometricalJournal* 49: 565–584.

[8.] VerHoef, J. M., &Boveng, P. L. 2007. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?. *Ecology*. 88(11): 2766-2772.

[9.] Ma, L., Yan, X., &Qiao, W. (2014). A quasi-Poisson approach on modeling accident hazard index for urban road segments. *Discrete dynamics in nature and society*. 2014: 1-8.

[10.] Lambert, D. 1992. Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics*. 34(1): 1-14.

[11.] Wagh, Y. S. &Kamalja, K. K. (2018). Zero-inflated models and estimation in zero-inflated Poisson distribution. *Communications in Statistics-Simulation and Computation*. 47(8): 2248-2265.

[12.] Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*. 33(2): 261-304