BAREKENG: Journal of Mathematics and Its Applications
June 2025 Volume 19 Issue 2 Page 0721–0732
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss2pp0721-0732

ZERO INFLATED POISSON REGRESSION MODELS TO ANALYZE FACTORS THAT INFLUENCE THE NUMBER OF MEASLES CASE IN JAVA

Candra R.W.S. Weni Utomo[1*], Achmad Efendi[2], Ni Wayan S. Wardhani[3]

[1,2,3]Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya
Jl. Veteran, Malang, Jawa Timur, 65145, Indonesia

*Corresponding author's e-mail: ** **rezziningcandra@gmail.com**

Article History:

Received: 3rd June 2024
Revised: 22nd January 2025
Accepted: 3rd February 2025
Published: 1st April 2025

Keywords:

Excess Zero;
Measles;
Overdispersion;
Poisson Regression;
Zero Inflated Poisson Regression

ABSTRACT

Measles is an infectious disease that often occurs in children and is caused by the measles virus (morbillivirus) which can cause death. Thus, it is important to identify the factors that cause measles. The number of measles cases is used as response variable in the form discrete data so that Poisson Regression is commonly used. However, some assumptions are sometimes not met, such as overdispersion and excess zero so that can use Zero Inflated Poisson Regression to meet these assumptions. Because the model can overcome two common characteristics that are often found in count data, which are excess zero and overdispersion. The purpose of this study was to determine the factors that influence the number of measles cases in East Java. The data in the study used secondary data obtained from the Central Statistics Agency (BPS). The predictor variables used were the number of population, percentage of vaccination, percentage of poor people, and percentage of adequate sanitation. The results showed that the data is overdispersed because the variance is greater than the mean. There were four predictor variables, The p-value of the total population variable is <0.01, the percentage of vaccinations is 0.914, the percentage of poor people <0.01 and the percentage of proper sanitation is 0.014 so it can be concluded that the percentage of vaccinations has no effect on the number of measles cases and the other three variables affect the number of measles cases in East Java. The best model of affect the number of measles cases in East Java is Zero Inflated Poisson with AIC value 326.24. The ZIP model for measles case in East Java is $\widehat{\mu_i} = \exp(-11.91 + 2.9753X_1 - 0.08X_3 - 0.0223X_4)$..

How to cite this article:

C. R.W.S. W. Utomo, A. Efendi and N. W. S. Wardhani., "ZERO INFLATED POISSON REGRESSION MODELS TO ANALYZE FACTORS THAT INFLUENCE THE NUMBER OF MEASLES CASE IN JAVA," *BAREKENG: J. Math. & App.*, vol. 19, iss. 2, pp. 0721-0732, June, 2025.

1. INTRODUCTION

Measles is an infectious disease that often occurs in children and is caused by the measles virus (morbillivirus) which can lead to death [1]. According to the WHO, measles can be prevented by vaccination, measles vaccination can prevent 57 million deaths between 2000 and 2022 which can reduce the estimated deaths from measles from 761,000 in 2000 to 136,000 in 2022 [2]. However, according to Indonesia's health profile in 2022, the number of suspected measles case in Indonesia was 21,175 case, which occurred in all provinces in Indonesia [3]. Measles positive cases also spread in almost all provinces except Central Kalimantan and Maluku [4]. In Indonesia, the high incidence of measles is caused by several factors that are often referred to as measles risk factors. Some of the risk factors for measles include individual-related, demographic, economic, social, and health factors.

This research will focus on the number of measles cases in East Java. One of the main reasons for choosing East Java was due to the high prevalence of measles cases in certain areas while there were no cases in other areas, which can cause overdispersion in the data. Apart from being an area with a large number of measles problems in 2021, East Java is also an island with a dense population in Indonesia. The data used in this study was secondary data derived from the East Java Provincial Health Profile and the Central Statistics Agency (BPS). The data was only for 2021 due to incomplete data. Therefore, this study was conducted in 38 cities and districts in East Java.

The number of measles cases was a common count data using the Poisson regression method in its analysis. In Poisson regression, it is necessary to assume the existence of equidispersion conditions where the variance and mean values are the same, this rarely happens because of the emergence of Over dispersion which occurs due to several things, namely, the violation of Poisson conditions, namely the variance value is greater than the mean, the presence of excess zeros, and the presence of outliers in the data [5]. Therefore, Poisson regression is not suitable for modeling the data. There are several methods to analyze data with overdispersion problems in Poisson regression, including generalized Poisson regression, zero inflated Poisson regression, zero inflated negative binomial, and zero inflated Poisson regression.

Zero inflated Poisson regression is used to model count data that has an excess of zero counts. Further, theory suggests that the excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently. Thus, the zero inflated Poisson model has two parts, a Poisson count model and the logit model for predicting excess zeros. Several studies that discuss various factors affecting the number of measles cases have been conducted such as those conducted by [6] using the generalized Poisson regression (GPR) approach, the results showed that the GPR model is better than the Poisson model, and the human development index and the percentage of working people in the labor force have a significant effect on the number of unemployment. Research conducted by [7] using the hurdle negative binomial regression approach, the results showed that in the first part with the logit model the variable that affected the discovery or not of measles was the variable of giving vitamin A. In the second part, the log model showed that each addition of one case of undernourished children under five will increase the number of measles cases by 1,0004 times. Research conducted by [8] using the zero inflated negative binomial regression approach, the results of the analysis show that in the ZINB regression, the factors that have a significant effect on the expected value of measles cases in West Java Province in 2020 are the percentage of vitamin A administration, the percentage of exclusive breastfeeding and the percentage of malnourished children under five. Meanwhile, factors that have a significant effect on the chance of measles in West Java Province in 2020 are the percentage of malnourished toddlers. This is used as a comparison in determining the factors that affect the number of measles cases in East Java Province in 2021. The novelty of this research is the application of zero inflated Poisson regression using maximum likelihood estimation and then comparing the two models with new data, namely measles data in East Java province in 2021.

2. RESEARCH METHODS

2.1 Materials and Data

The data used in this study consists of 38 districts and cities in East Java Province in 2021 and data published by the Central Statistics Agency (BPS) in 2021. The data used was secondary data from the East Java Provincial Health Profile in 2021. This study used district and city data in East Java Province. The response variable in this study was the number of measles cases (Y), while the predictor variables are population (X_1), percentage of vaccinations (X_2), percentage of poor people (X_3), and percentage of proper sanitation (X_4).

2.2 Non-Multicollinearity

Non-Multicollinearity Assumption is a test to determine whether the regression model has a correlation between the independent variables [9]. A regression model is said to be good if it is free from multicollinearity problems or there is no correlation between the independent variables. Multicollinearity occurs when there is a high correlation between several predictor variables of a multiple linear regression model. Multicollinearity can be detected through the variance inflation factor (VIF), Pearson correlation, or by looking at the eigenvalue and condition index (CI) [10]. In this study, multicollinearity detection using VIF was used. In regression with more than two predictor variables, VIF can be calculated using the formula:

$$VIF = \frac{1}{\left(1 - R_j^2\right)} \tag{1}$$

Where j is the predictor variable index and $R_j^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ the coefficient of determination in the regression model with j predictor variables and the other is the response variable in the model.

2.3 Overdispersion

Overdispersion is a condition that occurs when the variance is greater than the mean, this is a violation of the assumptions on the Poisson distribution [11]. The formula for detecting overdispersion is as follows [12].

$$\phi = \frac{\chi_p^2}{n - p - 1} \text{with } \chi_p^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{var(\hat{\mu}_i)} \tag{2}$$

Where y_i denotes the response variable of the i-th observation, n the number of observations, p the number of explanatory variables, $\hat{\mu}_i$ the mean estimate of the i-th response, and $var(\hat{\mu}_i)$ the variance estimate of the i-th response.

If the Pearson's Chi-squared test statistic divided by the independent degree is more than 1, then the data has over dispersion [13].

2.4 Homogeneity

Whether or not the assumptions are met can be tested with the Breusch Pagan Test using the following hypothesis.

$H_0: var(u|X) = E(u^2|X) = \sigma^2$

$E(u^2|X)$ represents the variance, if there is a violation of the assumption, then there is a relationship between u^2 the predictor variables. One form of relationship between u^2 and predictor variables that can be assumed is

$$u_i^2 = \delta_1 + \delta_2 X_{2i} + \cdots + \delta_k X_{ki} + v_i \tag{3}$$

If u^2 is on average independent of the predictor variable, then the hypothesis is equivalent to

$H_0: \delta_2 = \delta_3 = \cdots = \delta_k = 0$
H_1: There is at least one $\delta_j \neq 0, j = 2,3, \dots, k$

Using the Breusch-Pagan test statistic as follows.

$$B = \frac{1}{2} f^T \boldsymbol{Z}\left(\boldsymbol{Z}^T \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T f \tag{4}$$

Decision rules:

H_0 is rejected if $B > \chi^2_{(\alpha,p)}$ or p-value $< \alpha$ where p is the number of predictor variables.

2.5 Zero Inflated Poisson Regression

According to Lambert in [14], the ZIP regression method is a combined method of discrete-valued data with many zero values in the response variable. According to Jansakul & Hindie in [15], if Y_i is an independent random variable with ZIP distribution, the zero value of the observation is expected to occur in two ways corresponding to separate states. The first state is called the zero-state which occurs with probability ω and produces zero-valued observations, while the second state is called the Poisson state which occurs with probability $(1-\omega)$. According to [14] the ZIP distribution function is as follows

$$P(Y=y_i)=\begin{cases}\omega_i+(1-\omega_i)e^{-\mu_i}, & y_i=0\\ \dfrac{(1-\omega_i)e_i^{-\mu}\mu_i{}_i^{y}}{y_i!}, & y_i>0\end{cases} \tag{5}$$

With $Y_i \sim ZIP(\mu,\omega)$ and to model ω generally using a logit model, i.e.

$$\omega=\exp\frac{\boldsymbol{x}_i^T\boldsymbol{\gamma}}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\gamma})} \quad \mu=\exp\boldsymbol{x}_i^T\boldsymbol{\beta} \tag{6}$$

Where $\boldsymbol{x_i}$ is the i-th explanatory variable matrix $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is the regression parameter to be estimated. The relationship model of μ and ω is as follows

$$\omega\ln(\mu)=\boldsymbol{X}_i^T\boldsymbol{\beta} \text{ and } logit\ (\omega)=\ln\left(\frac{\omega}{1-\omega}\right)=\boldsymbol{X}_i^T\boldsymbol{\gamma} \tag{7}$$

where $\boldsymbol{X}$ is the matrix of explanatory variables, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $(p+1)\times 1$ and $(q+1)\times 1$ matrices of parameters to be estimated and ω is the probability at zero state. According to Jansakul & Hinde in [16], the mean and variance of ZIP regression are as follows [16].

$$E(Y_i)=\ \mu_i \text{ and } Var(Y_i)=\ \mu_i+\frac{\omega}{(1-\omega)}\mu_i^2 \tag{8}$$

From the equation above, it can be seen that the distribution of Y_i experiences overdispersion if $\omega>0$, because the variance value is greater than the average value.

2.6 Parameter Estimate

The estimation of ZIP regression parameters is done using the Maximum Likelihood method Estimation (MLE) method. The MLE method is one of the parameter estimation methods that can be used to estimate model parameters with known distributions.

Equation (6) is substituted into **Equation (5)** to obtain the following model.

$$P(Y=y_i)=\begin{cases}\exp\dfrac{x_i^T\gamma}{1+\exp(x_i^T\gamma)}+\left(1-\exp\dfrac{x_i^T\gamma}{1+\exp(x_i^T\gamma)}\right)e^{-\exp x_i^T\beta}, & y_i=0\\ \dfrac{\left(1-\exp\dfrac{x_i^T\gamma}{1+\exp(x_i^T\gamma)}\right)e_i^{-\exp x_i^T\beta}\exp x_i^T\beta_i^y}{y_i!}, & y_i>0\end{cases} \tag{9}$$

From **Equation (5)**, by multiplying all the probability functions of Y_i so that the ZIP likelihood function is obtained as follows.

$$L(\beta,\gamma|y_i)=\prod_{i=1}^n P(y_i;\beta,\gamma)$$

$$L(\beta,\gamma|y_i)=\begin{cases}\prod_{i=1}^n\left[\dfrac{\exp x_i^T\gamma+\exp\left(-e^{x_i^T\beta}\right)}{1+e^{x_i^T\gamma}}\right], & y_i=0\\ \prod_{i=1}^n\dfrac{\exp\left(-e^{x_i^T\beta}\right)\left(e^{x_i^T\beta}\right)^{y_i}}{\left(1+e^{x_i^T\gamma}\right)y_i!}, & y_i>0\end{cases} \tag{10}$$

Furthermore, from **Equation (6)**, the equation $ln\ likelihood$ is obtained:

For $y_i = 0$, function of $\ln L(\beta, \gamma | y_i)$ is obtained as follows.

$$\ln L(\beta, \gamma | y_i) = \sum_{i=1}^{n} \ln \left(\exp x_i^T \gamma + \exp \left(-e^{x_i^T \beta} \right) \right) - \sum_{i=1}^{n} \ln \left(1 + e^{x_i^T \gamma} \right) \tag{11}$$

For $y_i > 0$, function of $\ln L(\beta, \gamma | y_i)$ is obtained as follows.

$$\ln L(\beta, \gamma | y_i) = \sum_{i=1}^{n} (\exp x_i^T \beta) y_i - e^{x_i^T \beta} - \sum_{i=1}^{n} \ln(1 + \exp x_i^T \gamma) - \sum_{i=1}^{n} \ln y_i! \tag{12}$$

The summation of the $ln\ likelihood$ functions in **Equation (11)** and **Equation (12)** will complicate the calculation because it is not known which zero value comes from the zero state and which zero value comes from the zero state. calculation because it is not known which zero value comes from the zero state and which comes from the Poisson state. Poisson state, the $ln\ likelihoood$ function cannot be solved using ordinary numerical methods. using ordinary numerical methods.

To maximize the $ln\ likelihood$ function in **Equation (11)** and **Equation (12)** then using the EM (Expectation Maximization) algorithm.

Suppose variable Y is related to indicator variable Z, namely.

$$Z = \begin{cases} 1, & \text{if } y_i \text{ comes from zero state} \\ 0, \text{if } y_i \text{ comes from poisson state} \end{cases}$$

Where Z is a random discrete variable in function distribution with data information missing value.

The problem is that if the value of the response variable y_i = 1,2,3, …, then the value of z_i = 0. Conversely, if the value of the response variable is y_i = 0, then z_i may be 0 or may be 1. Therefore, the value of z_i is considered missing. To overcome this problem, parameter estimation is performed using the EM algorithm.

2.7 Model Significance Testing

For testing significant model use test statistic G^2 which is maximum likelihood ratio test statistic with the following hypotheses.

$H_0: \beta_1 = \beta_2 = \cdots = \beta_j, j = 1,2, \ldots, 4$
H_1: there is at least one $\beta_j \neq 0, j = 1,2, \ldots, 4$

The test statistic used is the G^2 statistic:

$$G^2 = -2 \ln \left(\frac{L_0}{L_1} \right) \sim \chi_k^2 \tag{13}$$

Where L_0 is likelihood model without predictor variable, L_1 is likelihood with predictor variable. The test criteria for the G^2 statistic reject H_0 if test statistic $G^2 > \chi^2_{(a,k)}$, k is number of predictor variable.

2.8 Parameter Significance Testing

This test aims to see the significance of the regression coefficient parameters from each of the Zero Inflated Poisson regression parameters as follows:

1. For Poisson State

$H_0: \beta_j = 0$ (The j-th predictor variable has no significant effect on the response variable)
$H_0: \beta_j \neq 0$ (The j-th predictor variable has a significant effect on the response variable)

2. For Zero State

$H_0: \gamma_j = 0$ (The j-th predictor variable has no significant effect on the probability that the response variable is zero)

$H_0: \gamma_j \neq 0$ (The j-th predictor variable has a significant effect on the probability of the response variable being zero).

The test statistic used is the Wald test statistic:

$$W_j = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right) \text{or } W_j = \left(\frac{\hat{\gamma}_j}{SE(\hat{\gamma}_j)}\right) \quad (14)$$

The test criteria for the Wald statistic that follows a standard normal distribution is reject H_0 if $|W_j| > Z_{1-\frac{\alpha}{2}}$.

2.9 Best Model Selection

To choose the best model using Akaike's Information Criterion (AIC).

$$AIC = -2 \ln L(\hat{\beta}) + 2p \quad (15)$$

Where $L(\hat{\beta})$ is the likelihood value and p is the number of parameters. The better model to use is the one with the smallest AIC value.

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics of Data

Based on the data that has been obtained, the descriptive analysis of the data is as follows.

Table 1. Summary of Descriptive Analysis

Statistic	Variable				
	Y	X_1	X_2	X_3	X_4
Minimum	0	131216	48.23	4.09	51.64
Median	0.5	1025335	89.20	10.56	99.85
Mean	11.55	1056754.53	83.77	11.32	95.92
Maximum	97	2918543	100	23.76	100
Standard Deviation	24.88	677056.10	18.33	4.72	7.19

Based on **Table 1**, shows that the lowest number of measles incidences is 0, meaning that there are no measles cases in an area, while the highest measles incidence is 97, which occurred in Surabaya City. The average number of measles cases that occurred in East Java in 2021 was 11.55. The population variable shows that Surabaya City with a total of 2.918,543 people followed by Malang Regency as much as 2,637,160 people. Meanwhile, the city/district with the lowest population level is Mojokerto City with 131,216 people. Surabaya City as a metropolitan city, has a much larger population compared to other cities/districts, which may be due to urbanization, economic activity centers, and better public facilities. The percentage of vaccination variable shows that in some areas the highest percentage is 100%. This indicates that vaccination coverage in most of these areas is close to or has even reached the expected target, while the lowest percentage is 48.23. These low vaccination coverage areas represent about 10.5% of the total districts/municipalities in East Java, indicating challenges in vaccination distribution in certain areas. The percentage of poor people variable shows that the lowest value is 4.09% while the highest is 23.76%. The average percentage of poor people in East Java is 11.32. The percentage of proper sanitation variable shows that the lowest value is 51.64% while the highest is 100%. The average percentage of proper sanitation in East Java is 95.92.

3.2 Non-Multicollinearity

Multicollinearity check is used to determine whether there is a correlation between two or more predictor variables. Multicollinearity is checked through the VIF values in **Table 2** below which shows the VIF score of each predictor variable.

Table 2. VIF Value for Predictor Variable

Variable	VIF Value
Population (X_1)	1.17
Percentage of Vaccination (X_2)	1.35
Percentage of Poor People (X_3)	1.22
Percentage of Proper Sanitation (X_4)	1.44

Based on the results in Table 2, VIF on the variable population (X_1), variable percentage of vaccination (X_2), variable percentage of poor people (X_3), and variable percentage of proper sanitation (X_4) all VIF values are smaller than 10 for all predictor variables. The small VIF value in this study indicates that there is no correlation between the predictor variables. Therefore, it can be concluded that there is no indication of multicollinearity in the predictor variables.

3.3 Overdispersion

Overdispersion this is greater than 1 so the data is overdispersed in the Poisson regression model. Was checked by calculating the Pearson's chi-squared value divided by the degrees of freedom. The result shows a value of 29.26 this is greater than 1 so the data is overdispersed in the Poisson regression model.

3.4 Homogeneity

The homogeneity test was conducted using the Breusch-Pagan test. The following is the hypothesis of the Breusch-Pagan test.

H_0: ${\sigma_1}^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$ (variety is homogeneous)
H_1: ${\sigma_i}^2 \neq \sigma^2, i = 1,2, \dots, n$ (variety is not homogeneous)

Based on the results of the BP analysis of 8.551 and $\chi^2_{0,05} = 9.488$ because $8.551 < 9.488$, it can be concluded that the data is homogeneous, or the variance between variables is the same.

3.5 Poisson Regression

Poisson regression model for measles cases in East Java is presented in Table 3.

Table 3. Zero Inflated Poisson Parameter Estimation

Variable	Parameter	Estimation	p-value
Intercept	β_0	−13.551	$< \mathbf{0.01}$
X_1	β_1	2.558	$< \mathbf{0.05}$
X_2	β_2	0.011	$\mathbf{0.013}$
X_3	β_3	−0.160	$< \mathbf{0.05}$
X_4	β_4	0.010	0.383
AIC		760.51	
$\boldsymbol{G^2}$		673.9	

Note: the p-value in **bold** is a significant predictors variable.

Simultaneous test with the following hypotheses.

H_0: $\beta_1 = \beta_2 = \cdots = \beta_j, j = 1,2, \dots, 4$
H_1: there is at least one $\beta_j \neq 0, j = 1,2, \dots, 4$

Based on Table 3, it can be concluded that H_0 is rejected because $G^2 = 673.9$ and value of $\chi^2_{(0,05;4)} = 9.488$ where $673.9 > 9.488$ which means that there is at least one predictor variable that affects the number of measles cases.

Partial test for parameters is carried out with the following hypotheses.

H_0: $\beta_i = 0$
H_1: $\beta_i \neq 0, i = 1,2,3,4$

Based on the Table 3, it can be seen that the parameter coefficient β_1 is 2.558 with a p-value < 0.05, which means that there is a real influence between the variable population (X_1) and the number of measles cases (Y). The parameter coefficient β_2 is 0.011 with a p-value $= 0.013$, this also means that there is a real influence between the variable percentage of vaccinations (X_2) on the number of measles cases (Y). The coefficient of β_3 is -0.160 with a p-value < 0.05, this means that there is an influence between the variable percentage of poor people (X_3) on the number of measles cases (Y). While the parameter coefficient β_4 is 0.010 with a p-value $= 0.383$, this means that there is no real influence between the variable percentage of proper sanitation (X_4) on the number of measles cases (Y).

3.6 Zero Inflated Poisson Regression

Poisson regression is used as an alternative when overdispersion conditions occur in the Poisson regression model. Because overdispersion occurs in the Poisson model, the zero inflated Poisson model is used to analyze the data. Table 4, shows the parameter estimation results of the zero inflated Poisson model.

Table 4. Zero Inflated Poisson Parameter Estimation

Variable	Parameter	Estimation	p-value
Intercept	$\boldsymbol{\beta_0}$	-11.910	< 0.01
X_1	$\boldsymbol{\beta_1}$	2.9753	< 0.01
X_2	$\boldsymbol{\beta_2}$	-0.0007	0.9137
X_3	$\boldsymbol{\beta_3}$	-0.0800	< 0.01
X_4	$\boldsymbol{\beta_4}$	-0.0022	0,0135
Intercept	$\boldsymbol{\gamma_0}$	-4.203	0.651
X_1	$\boldsymbol{\gamma_1}$	1.061	0.385
X_2	$\boldsymbol{\gamma_2}$	0.003	0.877
X_3	$\boldsymbol{\gamma_3}$	0.057	0.498
X_4	$\boldsymbol{\gamma_4}$	-0.032	0.539
AIC		326.24	
$\boldsymbol{G^2}$		897.5	

Simultaneous test with the following hypotheses.

$H_0: \beta_1 = \beta_2 = \cdots = \beta_j, j = 1,2,3,4$
H_1: there is at least one $\beta_j \neq 0, j = 1,2,3,4$

Based on Table 4, it can be concluded that H_0 is rejected because $G^2 = 897.5$ and value of $\chi^2_{(0,05;4)} = 9.488$ where $897.5 > 9.488$ which means that there is at least one predictor variable that affects the number of measles cases.

Partial test for parameters is carried out with the following hypotheses.

$H_0: \beta_i = 0$
$H_1: \beta_i \neq 0, i = 1,2,3,4$

Based on Table 4, model Poisson state variable X_1, X_3, and X_4 has p-value < 0.05. Therefore, variable population size (X_1), percentage poor people (X_2), and variable percentage proper sanitation (X_4) affect the number of measles cases and significant. while in the zero state model there are no variables that have a significant effect on the number of measles cases. The insignificant zero-state variable indicates that it does not affect the additional zero mechanism in the ZIP model, but indicate that the zero-state component is not significantly affected by the predictor variables. When the zero-state is insignificant, the main attention in interpretation can be directed more towards the count-state component, which models the count distribution (including zeros derived from Poisson processes).

3.7 Best Model Selection

The best model is determined using the AIC value of each model. The better model to use is the one with the smallest AIC value **Table 5** shows AIC values for every models.

Table 5. AIC Values for every Models

Model	AIC
Poisson Regression	760.51
Zero Inflated Poisson Regression	326.24

Based on **Table 5**, the smallest AIC value is the AIC value for the Zero Inflated Poisson Regression model which is 326.24. Therefore, it can be concluded that the best model in this study is Zero Inflated Poisson Regression.

3.8 Interpretation

Parameter estimates for the Poisson state model and the zero state model are as follows.

$$\log(\mu_i) = -11.91 + (2.9753)X_1 - (0.0006593)X_2 - (0.08)X_3 - (0.0223)X_4$$

$$\log it(\omega_i) = -4.203 + 1.061X_1 - 0.003X_2 + 0.057X_3 - 0.032X_4$$

Based on testing the significance of the parameters of the Zero Inflated Poisson regression model, only significant parameters can be interpreted, namely: X_1 (the population), X_3 (the percentage of poor people), and X_4 (the percentage of proper sanitation).

For β_1 equal to 2.9753, it holds that every increase in population (X_1) by 1 person assuming the value of other variables is held constant, the average number of measles (Y) tends to increase by $\mu_i = \exp(2.9753) = 19.6 \approx 20$ cases of measles in East Java. This indicates that an increase in population is directly related to an increase in the number of measles cases. Therefore, as the population increases, the probability of more measles cases is also higher. For β_3 equal to -0,08 states that every increase in percentage poor people (X_3) by 1% assuming the values of other variables is held constant, the average number of measles (Y) tends to decrease by $\mu_i = \exp(-0.08) = 1.1 \approx 1$ case of measles in East Java. This suggests that despite an increase in the number of poor people, a higher poverty rate is more or less associated with a decrease in the number of measles cases. This could indicate that other factors (such as decreased mobility or better disease control in poor communities) may reduce measles case numbers. For β_4 equal to -0,0223 state that every increase in percentage proper sanitation (X_4) by 1% assuming the values of other vaariables is held constant, the average number of measles (Y) tends to decrease by $\mu_i = \exp(-0.0223) = 1{,}02 \approx 1$ case of measles in East Java. The improved sanitation contributed to a decrease in the number of measles cases, demonstrating the importance of health and sanitation infrastructure in controlling the spread of the disease.

This research is in line with research conducted by **[17]** with the results of his research show that measles immunization status results obtained p-value $0.02 < 0.05$, the influence maternal knowledge on the incidence of measles p-value $0.00 < 0.05$, and the effect of housing density on the incidence of measles house p-value $0.00 < 0.05$. Conclusion that was an effect of immunization status, maternal knowledge and density of residential home with a case of measles in the health centers Tejakula I. The research by **[18]** the results show that immunization status, nutritional status, mother's knowledge, and occupancy density. The variable with the highest risk factor value for measles incidence in Indonesia is the occupancy density, followed by the mother's knowledge, immunization status, and nutritional status.

4. CONCLUSIONS

Based on the results of the analysis that has been done, it can be concluded that the data on the number of measles cases in East Java is overdispersed so that the Zero Inflated Poisson (ZIP) regression model can be used to overcome this problem. The ZIP model can model additional zeros through the zero inflation process and model other count data using the Poisson distribution. There are three variables that are influential and significant, namely the population variable (X_1), the percentage of poor people (X_3), and the percentage of proper sanitation (X_4). The ZIP model for measles cases in East Java is as follows.

$$\widehat{\mu_i} = \exp(-11.91 + 2.9753X_1 - 0.08X_3 - 0.0223X_4)$$

The AIC value in the ZIP regression model is 326.24, this shows that the AIC value of ZIP is smaller than the Poisson regression model, therefore the appropriate model to determine the factors that affect the number of measles patients in East Java is the Zero Inflated Poisson Regression model.

ACKNOWLEDGEMENT

With my gratitude to Allah SWT for grace and guidance. Alhamdulillah, the author can complete this research on time. The author dedicates this research to two great people in life, namely Mom and Dad, thank you for always praying endlessly. Thank you for myself for struggling, and do not forget to also thank the author for the supervisor, boyfriend and all friends involved who always provide help, prayers and support to the author. The author is very grateful to have all of you in this life.

REFERENCES

[1] K. R, “PENYAKIT CAMPAK,” *perpustakaankemenkes.go.id*, Online, Agustus 2022.

[2] A. A. Minta *et al.*, “PROGRESS TOWARD MEASLES ELIMINATION — WORLDWIDE, 2000–2022,” *MMWR Morb. Mortal. Wkly. Rep.*, vol. 72, no. 46, pp. 1262–1268, Nov. 2023, doi: 10.15585/mmwr.mm7246a3.

[3] S. Chacko *et al.*, “PROGRESS TOWARD MEASLES AND RUBELLA ELIMINATION — INDONESIA, 2013–2022,” *MMWR Morb. Mortal. Wkly. Rep.*, vol. 72, no. 42, pp. 1134–1139, Oct. 2023, doi: 10.15585/mmwr.mm7242a2.

[4] A. Riantina, Najma, and R. J. Sitorus, “ANALYSIS OF RISK FACTORS AFFECTING THE INCIDENT OF MEASLES IN INDONESIA: LITERATURE REVIEW,” *J. Nurs. Public Health*, vol. 12, no. 1, Apr. 2024.

[5] A. Rahayu, “MODEL-MODEL REGRESI UNTUK MENGATASI MASALAH OVERDIPERSI PADA REGRESI POISSON,” *J. Peqguruang Conf. Ser.*, vol. 2, no. 1, p. 1, Jun. 2021, doi: 10.35329/jp.v2i1.1866.

[6] H. Ihsan, W. Sanusi, and R. Ulfadwiyanti, “MODEL GENERALIZED POISSON REGRESSION (GPR) DAN PENERAPANNYA PADA ANGKA PENGANGGURAN BAGI PENDUDUK USIA KERJA DI PROVINSI SULAWESI SELATAN,” *J. Math. Comput. Stat.*, vol. 3, no. 2, p. 109, Feb. 2021, doi: 10.35580/jmathcos.v3i2.19190.

[7] Tasya Azzahra, “PEMODELAN REGRESI HURDLE NEGATIVE BINOMIAL PADA JUMLAH KASUS DIFTERI JAWA BARAT 2020,” *J. Ris. Stat.*, pp. 125–130, Dec. 2023, doi: 10.29313/jrs.v3i2.3014.

[8] I. Amarita and N. Hajarisman, “PENERAPAN MODEL REGRESI ZERO INFLATED NEGATIVE BINOMIAL PADA KASUS CAMPAK DI PROVINSI JAWA BARAT TAHUN 2020,” *Bdg. Conf. Ser. Stat.*, vol. 3, no. 2, pp. 737–744, Aug. 2023, doi: 10.29313/bcss.v3i2.9311.

[9] M. Sriningsih, D. Hatidja, and J. D. Prang, “PENANGANAN MULTIKOLINEARITAS DENGAN MENGGUNAKAN ANALISIS REGRESI KOMPONEN UTAMA PADA KASUS IMPOR BERAS DI PROVINSI SULUT,” *J. Ilm. SAINS*, vol. vol.18, no. 1, no. 1, p. 18, Jul. 2018, doi: 10.35799/jis.18.1.2018.19396.

[10] A. A. R. Fernandes and Solimun, *ANALISIS REGRESI DALAM PENDEKATAN FLEKSIBEL (ilustrasi dengan Paket Program R)*. 2021, UB Press.

[11] A. Rahayu, “MODEL-MODEL REGRESI UNTUK MENGATASI MASALAH OVERDIPERSI PADA REGRESI POISSON,” *J. Peqguruang Conf. Ser.*, vol. 2, no. 1, p. 1, Jun. 2021, doi: 10.35329/jp.v2i1.1866.

[12] V. Eminita, A. Kurnia, and K. Sadik, “PENANGANAN MULTIKOLINEARITAS DENGAN MENGGUNAKAN ANALISIS REGRESI KOMPONEN UTAMA PADA KASUS IMPOR BERAS DI PROVINSI SULUT,” *FIBONACCI J. Pendidik. Mat. Dan Mat.*, vol. 5, no. 1, p. 71, Jul. 2019, doi: 10.24853/fbc.5.1.71-80.

[13] D. P. Prami Meitriani, K. G. Sukarsa, and I. P. E. N. Kencana, “PENERAPAN REGRESI QUASI-LIKELIHOOD PADA DATA CACAH (COUNT DATA) YANG MENGALAMI OVERDISPERSI DALAM REGRESI POISSON,” *E-J. Mat.*, vol. 2, no. 2, p. 37, May 2013, doi: 10.24843/MTK.2013. v02.i02.p036.

[14] N. S. N. Ihsan and N. Hajarisman, “PENERAPAN MODEL REGRESI ZERO-INFLATED POISSON PADA KASUS KEMATIAN BAYI DI KOTA BANDUNG TAHUN 2020,” *Bdg. Conf. Ser. Stat.*, vol. 3, no. 2, Jul. 2023, doi: 10.29313/bcss.v3i2.8042.

[15] A. D. Aprilia, “REGRESI ZERO INFLATED POISSON UNTUK PEMODELAN ANGKA POSITIF PENYAKIT MALARIA DI JAWA TIMUR,” *J. Ilm. Mat.*, 2023.

[16] K. Tawiah, S. Iddi, and A. Lotsi, “ON ZERO-INFLATED HIERARCHICAL POISSON MODELS WITH APPLICATION TO MATERNAL MORTALITY DATA,” *Int. J. Math. Math. Sci.*, vol. 2020, pp. 1–8, Dec. 2020, doi: 10.1155/2020/1407320.

[17] N. Giarsawan, I. W. S. Asmara, and A. E. Yulianti, “FAKTOR-FAKTOR YANG MEMPENGARUHI KEJADIAN CAMPAK DI WILAYAH PUSKESMAS TEJAKULA I KECAMATAN TEJAKULA KABUPATEN BULELENG TAHUN 2012,” *J. Esehatan Lingkung.*, vol. 4, no.2, pp. 140–145, Nov. 2014.

[18] F. H. Ramadhani, R. Azizah, J. Jalaludin, S. Martini, and L. Sulistyorini, "META-ANALYSIS AND SYSTEMATIC REVIEW: RISK FACTORS OF MEASLES INCIDENCE IN INDONESIA (2012 – 2021)," *J. Kesehat. Masy.*, vol. 19, no. 1, pp. 138–148, Jul. 2023, doi: 10.15294/kemas.v19i1.43060.