Modeling of infant mortality in west sulawesi using zero inflated poisson regression method

RAnnisa 1 EO Permatasari 2 and A T Rumiati 3
1Department of Statistics, Faculty of Mathematics, Computing, and Data Science, Sepuluh Nopember Institute of Technology, Teknik Mesin road 175, Keputih, Sukolilo, Surabaya, East Java
2 Department of Statistics, Faculty of Mathematics, Computing, and Data Science, Sepuluh Nopember Institute of Technology, Teknik Mesin road 175, Keputih, Sukolilo, Surabaya, East Java
3Department of Statistics, Faculty of Mathematics, Computing, and Data Science, Sepuluh Nopember Institute of Technology, Teknik Mesin road 175, Keputih, Sukolilo, Surabaya, East Java
e-mail: rocmatun.annisa@gmail.com, erma.oktania@gmail.com, and agnesrumiati@gmail.com

Abstract. The number of infant deaths, one of which can be reflected in the Infant Mortality Rate (IMR) is one indicator to determine the development of a country's development. The AKB target in the 2015-2019 National Medium-Term Development Plan is 24 deaths per 1,000 live births. Based on the SDKI 2017, the IMR value in West Sulawesi is 42 deaths per 1,000 live births. Factors suspected of affecting infant mortality in West Sulawesi are not only in the health but also in socio-economic fields. Data on the number of infant mortality in West Sulawesi is a data count that contains a value of 0. Therefore, the analysis used is Zero Inflated Poisson (ZIP) Regression to determine what factors influence infant mortality in West Sulawesi. The results of the ZIP regression analysis showed that the best model with the smallest AIC and AICc is 324,3407 and 326,7407 were models with a combination of the variables of the percentage of active maternal and child health services, percentage of society with good drinking water adequate access and percentage of society with proper sanitation access. The variable that has a significant effect on the log model is the percentage of active maternal and child health services, percentage of society with good drinking water adequate access and percentage of society with proper sanitation access. In the variable logit model that is influential is the percentage of active maternal and child health services.

1. Introduction
Healthy development is a part of country development that strived by the government. One of the indicators from health development is the Infant Mortality Rate (IMR). IMR is a number that shows the rate of infant deaths aged 0 to 1 year from every 1,000 live-births in a year or the infant’s death before one-year-old [1]. Based on data from the Indonesian Demographic and Health Survey (IDHS) at 2012, the national IMR reached 32 deaths per 1,000 live births. Therefore, one of the main national development goals in the National Midterm Development Plan (NMDP) at 2015-2019 the health and nutrition of the society with the target of IMR is 24 deaths per 1,000 live birth.
West Sulawesi is one of the provinces with the highest number of IMR in Indonesia. Based on the 2017 IDHS data, the IMR in West Sulawesi reached 42 deaths per 1,000 live births. This number is still far from the target set by the government in the NMDP 2015-2019. One hypothesized affecting factor of infant mortality is socio-economic factors. This can be seen from the still gap in infant mortality at the levels of educational, socio-economic, inter-urban and rural.

Previous researches that have been conducted on IMR are Handayani & Pratiwi (2016) [2]. This study was an analysis of labor habits in the Kaili Da’a tribe in Mamuju District, West Sulawesi. There were still many Kaili Da’a tribal people who give birth at home with the help of a family or a doctor. This made infant mortality in West Sulawesi increase. Other research that discusses the affecting factors of the maternal and infant mortality rate in Central Java using the Bivariate Generalized Poisson Regression method performed by Putri & Purhadi (2017) [3]. Data on the number of infant deaths in West Sulawesi does not meet the equidispersion assumption in the Poisson regression method. In the data on the number of Infant Mortality in Sulawesi was overdispersion, namely the variance value greater than the mean value. In the data on the number of infant mortality, some data are worth zero. In cases of overdispersion and zero data, the Zero Inflated Poisson (ZIP) regression model is recommended [2]. Therefore, this study analyzed the number of infant deaths in West Sulawesi by using the Zero Inflated Poisson regression method because of overdispersion cases. The response variable used is the number of infant mortality in West Sulawesi and the predictor variables are factors that are thought to affect the number of infant mortality in West Sulawesi.

2. Literature Review

2.1. Infant Mortality
Infant mortality is the death of individuals less than 1-year-old. Infant mortality is represented in Infant Mortality Rate (IMR). IMR is a number that shows the number of infant deaths aged 0 to 1 year from every 1,000 live births in a given year or the probability of a baby die before getting the age of one year [1]. Infant mortality is caused by many factors. Generally, infant mortality has two types of causes, that is endogenous causes and exogenous causes. Endogenous infant mortality is commonly referred to as neonatal death. Exogenous infant mortality or postneonatal infant mortality is infant mortality that occurs after the baby has an age of one month until the age of one year caused by factors related to the influence of the external environment [3].

2.2. Descriptive Statistics
The statistical method is the procedures used in collecting, presenting, analyzing, and interpreting data. The methods are divided into two, namely descriptive statistics and inferencing statistics. Descriptive statistics is a method related to the collection and presentation of data to provide information about the data and not concluding [4]. In this study, descriptive statistics will be used in the form of minimum values, maximum values, averages, and variances.

2.3. Poisson Distribution
Poisson distribution is a distribution used in events that occur at a certain time interval or region. Poisson distribution is a distribution for events that have a small chance of occurrence. Poisson distribution is discrete variables [4].

The probability distribution function for Y’s Poisson random variable with λ’s parameter is shown in equation 1.

\[ P(Y_i = y_i) = \begin{cases} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!}, & \text{for } y_i = 0, 1, 2, \ldots \\ 0, & \text{for other } y_i \end{cases} \]

\[ E[Y_i] = \lambda_i, \quad \text{var}(Y_i) = \lambda_i \]
The Kolmogorov Smirnov test is used to determine whether a data is Poisson distributed or not. The testing hypothesis is follows:

\[ H_0 : \text{the data is Poisson distribution} \]
\[ H_1 : \text{the data is not Poisson distribution} \]

The statistic test in testing this hypothesis is the D statistic test. The D statistic test is the maximum value of the absolute value of the difference between the cumulative frequency distribution calculated from the data sample \( S_n(y) \) with the cumulative hypothesized distribution function Poisson \( F_0(y) \),

\[
D = \max \left| F_0(y) - S_n(y) \right|
\]  

The null hypothesis will be rejected if \( D > D_{(n,n)} \), where \( n \) is the sample size and \( D_{(n,n)} \) is the critical value D Kolmogorov Smirnov test with \( \alpha \) significance level or \( \alpha \) in the Kolmogorov Smirnov table.

2.4. Multicollinearity Assumptions

The multicollinearity (double collinearity) was found by Ragnar Frisch, which means that there is a perfect or definite linear relationship between some or all independent variables of the multiple regression model. Furthermore, the multicollinearity is the occurrence of high linear correlation between independent variables \( X_1, X_2, \ldots, X_k \) [5].

Multicollinearity cases can be detected with \( R^2 \) values. Multicollinearity occurs if \( R^2 \) is high in the model, but little or none of the regression parameters are significant if tested individually. Multicollinearity also occurs if the regression model obtains a regression coefficient with a different signs from \( Y \) and \( X_j \). For example, the correlation between \( Y \) is positive, but the regression coefficient is related to \( X_j \) is negative. According to Gujarati (2004) one way to identify the presence or absence of multicollinearity cases is through the value of Variance Inflating Factor (VIF). The VIF value shows how the variation of the parameter estimation results increases. VIF values of more than 10 indicate a case of multicollinearity. VIF value is formulated in equation 3.

\[
VIF_j = \frac{1}{1-R^2_j}
\]

With \( R^2_j \) is the coefficient of determination \( X_j \) with other predictor variables.

2.5. Overdispersion

Poisson regression is used for data with response variables that follow the Poisson distribution \( Y \sim \text{Poisson} \). The assumption is important to fulfill that the variance must be equal to the average or can be called equidispersion. In real data, these conditions often cannot be fulfilled, that is, often found data counts that have variance values greater than average or called overdispersion. However, if the value of variance is less than average, it is called underdispersion [6]. According to Hinde & Demetrio (2007) [9], there is an impossibility that equidispersion is not fulfilled in modeling, including the variety of observations (variety between individuals as components is not explained by the model), correlations between individual responses, groupings occur in populations and variables observed that was eliminated.

2.6. Zero Inflated Poisson (ZIP) Regression
Zero Inflated Poisson (ZIP) regression models are recommended for cases where the data with the response variable contains a zero value in a large proportion (zero inflation) [2]. ZIP regression model is an alternative method for analyzing data with many zero values contained in the response variable. The number of zero values in the data can lead to infringement from the similarity assumption of the mean and variance in the Poisson distribution. For each observation on the $y_1, y_2, ..., y_n$'s independent response variable and $Y$ follow the distribution is shown in equation 4.

$$Y_i = \begin{cases} 0, & \text{with the probability is } \pi_i \\ \text{Poisson}(\lambda_i), & \text{with probability is } (1 - \pi_i) \end{cases} \quad (4)$$

So, the probability function for $y_i$ is follow equation 5.

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1 - \pi_i) e^{-\lambda_i}, & \text{for } y_i = 0 \\ (1 - \pi_i) e^{-\lambda_i} \lambda_i^{y_i} / y_i!, & \text{for } y_i > 0 \end{cases} \quad (5)$$

with $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n)^T$'s parameters and $\pi = (\pi_1, \pi_2, ..., \pi_n)^T$'s parametersth that fulfill equation 6.

$$\lambda = \exp(x_i^T \beta) \quad (6)$$

with,

$$\log(\lambda) = \begin{bmatrix} \log \eta_1 \\ \log \eta_2 \\ \vdots \\ \log \eta_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \\ \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \\ \vdots \\ \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = X \beta \quad (7)$$

and

$$\logit(\pi) = \begin{bmatrix} \logit \phi_1 \\ \logit \phi_2 \\ \vdots \\ \logit \phi_n \end{bmatrix} = \begin{bmatrix} \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_k x_{ik} \\ \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_k x_{ik} \\ \vdots \\ \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_k x_{ik} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = X \gamma \quad (8)$$

where,
ZIP regression model can be written according to the equation 10.

\[ \log \lambda_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij}; i = 1, 2, \ldots, n; j = 1, 2, \ldots, k \]

\[ \log \pi_i = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_j x_{ij}; i = 1, 2, \ldots, n; j = 1, 2, \ldots, k \]

(10)

With,
\[ \beta \] : vector of regression parameters estimated is representing variables that affect Poisson state.
\[ \gamma \] : vector of regression parameters estimated is representing variables that affect zero state.
\[ X \] : matrix of variables that containing different sets of experimental factors related to the zero-state probability and Poisson mean in Poisson state.

2.7. ZIP Regression Parameter Estimate

ZIP Regression parameter estimator used the Maximum Likelihood Estimation (MLE) method. This method is usually used to estimate the parameters of a model whose density function is unknown. The first likelihood will be built from the ZIP function, by substituting the link function from the ZIP Regression model to its density function.

The likelihood function from ZIP regression model is:

\[ L(\gamma, \beta) = \prod_{i=1}^{n} \left( \frac{e^{x_i'\gamma}}{1 + e^{x_i'\gamma}} \right)^{y_i} \left( 1 + \frac{e^{x_i'\gamma}}{1 + e^{x_i'\gamma}} \right)^{1-y_i}, \text{for } y_i = 0 \]

\[ \prod_{i=1}^{n} \left( \frac{e^{x_i'\gamma}}{1 + e^{x_i'\gamma}} \right)^{y_i} \left( 1 + \frac{e^{x_i'\gamma}}{1 + e^{x_i'\gamma}} \right)^{1-y_i}, \text{for } y_i > 0 \]

(11)

And log likelihood function is:

\[ \log L(\gamma, \beta) = \sum_{i=1}^{n} \left( \log e^{x_i'\gamma} + \log \left( 1 + e^{x_i'\gamma} \right) \right), \text{for } y_i = 0 \]

\[ \sum_{i=1}^{n} \left( -e^{x_i'\gamma} + x_i'\gamma \right) y_i - \sum_{i=1}^{n} \log \left( 1 + e^{x_i'\gamma} \right) - \sum_{i=1}^{n} y_i, \text{for } y_i > 0 \]

(12)

Equation 12 is called incomplete likelihood, this is because the zero value in the first term is unknown which originates from the zero states and which originates from the Poisson state, so equation 12 is solved by redefining the variable with an indicator variable \( Z_i \), namely:

\[ Z_i = \begin{cases} 1, & \text{if } y_i \text{ from zero state} \\ 0, & \text{if } y_i \text{ from Poisson state} \end{cases} \]

(13)

If \( y_i > 0 \) then the value of \( Z_i = 0 \). However, if \( y_i = 0 \) so \( Z_i \) can be either 0 or 1. Therefore, is considered to be partially lost. These problems can be solved using the Expectation-Maximization (EM) algorithm. The EM algorithm is an alternative iterative method to maximize the likelihood function which contains incomplete data (missing). At each iteration, the EM algorithm has 2 stages: the expectation and maximization stages. The expectation stage is the expectation calculation stage of the log-likelihood function by taking into account incomplete data. The maximization stage is the
calculation phase to look for parameter estimators that maximize the log-likelihood function as a result of the expectation stage.

2.8. Test of ZIP Regression Parameters
Fitting of ZIP regression model can be used the Likelihood Ratio (LR) test [2].

1. Simultan Test
Simultan test is used to test \( \beta \) and \( \gamma \)'s parameters together. \( \beta \)'s parameter show Poisson state parameters and \( \gamma \)'s parameter show zero state parameters. The hypothesis used is:

- \( H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \)
- \( H_1: \) at least there is one \( \beta_j \neq 0, \ j = 1, 2, \ldots, k \)

and

- \( H_0: \gamma_1 = \gamma_2 = \ldots = \gamma_k = 0 \)
- \( H_1: \) at least there is one \( \gamma_j \neq 0, \ j = 1, 2, \ldots, k \)

With \( k + 1 \) is a number of parameters, \( \beta_j \) is parameters of log parameter and \( \gamma_j \) is parameters of logit parameters. The statistic test is the size of the likelihood ratio (deviants) formed to define a set of parameters under which the population \( \Omega = \{ \beta, \gamma \} \). The parameters under \( \omega = \{ \hat{\beta}_0, \hat{\gamma}_0 \} \). Then on the set parameters under the population, formed the likelihood function for the full model (saturated) involving all the predictor variables. While the population parameter under correct, set up the likelihood function for a model that does not involve the predictor variables. Both of the likelihood function \( L(\Omega) \) and \( L(\omega) \) are [7]:

\[
G^2 = -2 \ln \left( \frac{L(\hat{\omega})}{L(\hat{\Omega})} \right) = -2 \ln \left( \prod_{j=1}^{n} \left[ \frac{\exp(-e^{x_j}) (e^{x_j})^{y_j} \gamma_j!}{\left( 1 + e^{x_j} \right) (e^{x_j} \gamma_j)^{y_j}} \right] \right)
\]

\[
G^2 \text{ 's statistics is } \chi^2_{(df)} \text{ distribution so that the significance level of } \alpha \text{ reject } H_0 \text{ if the value of } G^2 > \chi^2_{(df, \alpha)} \text{, where } df \text{ is the number of parameters under population reduced the number of parameters under } H_0.
\]

2. Individual Test
In testing individual parameters there are two tests are testing for the log model parameters and testing for the logit model parameters. The hypotheses testing the log model parameters for Poisson states is:

- \( H_0: \beta_j = 0 \)
- \( H_1: \beta_j \neq 0, \ j = 1, 2, \ldots, k \)

Statistic test:

\[
Z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}
\]

And hypothesis for testing of logit model for zero state is:
H₀: γᵢ = 0
H₁: γᵢ ≠ 0, j = 1, 2, ..., k

Statistic test:

\[ Z = \frac{\hat{\gamma}_j}{SE(\hat{\gamma}_j)} \]  \hspace{1cm} (16)

With \( z_i \) is an-i indicator variables, \( X_j \) is an-j predictor variables, \( Y_i \) is an-i response variables, \( \hat{\gamma} \) and \( \hat{\beta} \) is an estimates parameter. The rejecting area for log and logit model is reject \( H_0 \) if \( Z > Z_{\alpha/2} \) or \( Z < -Z_{\alpha/2} \).

2.9. Selection of The Best Model

The best model in Zero Inflated Poisson regression is obtained by selecting the best model using the criteria of Akaike's Information Criterion (AIC) and Akaike's Information Corrected Criterion (AICc). Usually, the AIC value is in line with the deviance value of the model. The smaller the deviance value, the smaller the error rate generated by the model so that the model obtained becomes more precise. Therefore, the best model is the model with the smallest AIC value. AIC values are formulated in equation 17 [8].

\[ AIC = -2 \times (\text{log likelihood}) + 2k \]  \hspace{1cm} (17)

with,

\( k \) : the number of parameters in the model.

The formulation of AICc is showed in equation 18.

\[ AIC_c = -2 \times (\text{log likelihood}) + 2k \left(1 - \frac{n}{n-k-1}\right) \]  \hspace{1cm} (18)

with,

\( n \) : the number of observation.

The best model have small AICc value.

3. Research Method

3.1. Data Source

The data used in this research are secondary data obtained from the Heath Profiles of each district in West Sulawesi in 2016 is issued by the West Sulawesi Health Office. The unit of observation is 69 sub-districts in West Sulawesi Province.

3.2. Research Variables

Research variables used in this research are response and predictor variable. The variables in this research are shown on table 1.

| Variables                                      | Scale       |
|------------------------------------------------|-------------|
| Y                                              | Number of infant deaths | Ratio     |
| X₁                                             | Percentage of Active Maternal and Child Health Services | Ratio |
| X₂                                             | Percentage of Household with Health and Clean Behavior | Ratio |
| X₃                                             | Percentage of Universal Child Immunization (UCI) Village | Ratio |
| X₄                                             | Percentage of maternity mothers who assisted by medical workers | Ratio |
| X₅                                             | Percentage of society with good drinking water adequate access | Ratio |
| X₆                                             | Percentage of society with proper sanitation access | Ratio |
3.3. Analytical Steps

The analytical steps are taken as follow:
1. Study the literature and formulate problems.
2. Collecting data on the number of infant deaths and factors thought to affect the number of infant mortality in West Sulawesi.
3. Test the Poisson distribution on the response variable.
4. Perform overdispersion testing on the variable number of infant mortality.
5. Perform descriptive statistical analysis.
6. Detect multicollinearity with VIF based on equation 3.
7. Perform analysis with the Zero Inflated Poisson regression method based on equation 5.
8. Choosing the best model with AIC and AICC based on equations 17 and 18.
9. Interpreting the model formed.
10. Draw conclusions and suggestions.

4. Analysis and Discussion

4.1. Exploration of the Number of Infant Mortality and Factors Suspected to Affect Infant Mortality in West Sulawesi

Infant mortality is the death of individuals aged 0 to 1 year. Table 2 shows the highest and lowest number of infant mortality in West Sulawesi 2016.

Table 2. The Highest and Lowest Number of Infant Mortality

| Subdistrict | The number of infant mortality |
|-------------|-------------------------------|
| Pasakayu, Banggae | 15 |
| Matangnga, Balanipa, Luyo, Mapili, Tapoyo, Pangale, Balabalakang, Kalumpang, Bonehau, Sampaga, Papalang, Mambi, Bambang, Rantim, Tabang, Nosu | 0 |

Based on Table 2, the regions that have the highest number of infant mortality is 15 deaths in 2016. The regions are Pasakayu subdistricts and Banggae subdistricts. Meanwhile, there are 16 regions have not infant mortality. The regions are Matangnga, Balanipa, Luyo, Mapili, Tapoyo, Pangale, Balabalakang, Kalumpang, Bonehau, Sampaga, Papalang, Mambi, Bambang, Rantim, Tabang, Nosu.

Characteristics of the factors thought affect to infant mortality in West Sulawesi in 2016 are shown in Table 3.

Table 3. Descriptive statistics Suspected Factors Influencing Infant Mortality in West Sulawesi in 2016

| Variable | Mean | Variance | Min | Max |
|----------|------|----------|-----|-----|
| X₁       | 60.75| 1578.08  | 0.00| 100.00 |
| X₂       | 67.96| 865.40   | 0.00| 160.75 |
| X₃       | 63.18| 861.97   | 0.00| 125.00 |
| X₄       | 77.60| 160.19   | 43.79| 103.16 |
| X₅       | 38.89| 1366.44  | 0.00| 120.39 |
| X₆       | 49.52| 1468.95  | 0.00| 137.37 |

Based on Table 3, the highest average is found in the variable of the percentage of maternity mothers who assisted by medical workers ($X₄$) is 77.60%, the meaning that 77.60% of mothers in each district in West Sulawesi had delivered labor assisted by health workers. The variable has the largest variance is 1578.08 so it can be interpreted that the diversity of the active maternal and child health services variable is very high.
4.2. Poisson Distribution Testing for The Number of Infant Mortality in West Sulawesi

Testing the Poisson distribution on data on the number of infant mortality in West Sulawesi using the Kolmogorov Smirnov test shown in Table 4.

Table 4. Poisson Distribution Testing for Response Variables

| Variable                  | Kolmogorov Smirnov |
|---------------------------|--------------------|
| The number of infant mortality | 0.24181            |

Based on the test results obtained D value of 0.24181. If the value is compared with a value of $D_{(0.05;69)}$ or 0.1637, a decision to reject $H_0$ will be obtained. The conclusion that can be drawn is data on the number of infant deaths in West Sulawesi in 2016 without the Poisson distribution.

However, in testing using easy fit software, the results show that the data has the Poisson distribution in second. Table 5 shows the test results using the help of easy fit software.

Table 5. Testing of Data Distribution Using Easy Fit Software

| Variable  | Kolmogorov Smirnov |
|-----------|--------------------|
| Statistic test | Rank             |
| Geometric  | 0.22848            | 1                  |
| Poisson    | 0.24181            | 2                  |

Table 5 shows that the data on the number of infant mortality in West Sulawesi in 2016 was Poisson ranked second.

4.3. Overdispersion Testing on Number of Infant Mortality in West Sulawesi

Overdispersion is a case that a variable has a variance greater than the average. The mean and variance variables for the number of infant deaths is shown in Table 6.

Table 6. Overdispersion Testing on Number of Infant Mortality in West Sulawesi

| Variable                  | Mean   | Variance |
|---------------------------|--------|----------|
| Number of Infant Mortality| 3.377  | 13.915   |

Based on Table 6 it can be seen that the average of the variable number of infant mortality is 3.377 and the variance is 13.915. The variance is greater than the average or 13.915 greater than 3.377 so it can be concluded that the variable number of infant mortality occurred overdispersion.

In order to make this assumption more valid, overdispersion testing is compared by the value of the deviance per degrees of freedom. If the result more than 1, it is called overdispersion.

Table 7. Overdispersion Testing on data Number of Infant Mortality in West Sulawesi

| Variable                  | Deviants | Deviants / Degrees of Freedom |
|---------------------------|----------|-------------------------------|
| Number of Infant Mortality| 262.83   | 3.856                         |

Table 7 shows the value of deviance per degrees of freedom is 3.856 so it can be concluded that data on the number of infant mortality in West Sulawesi in 2016 occurred overdispersion.

4.4. Assumptions of Multicollinearity Testing on Predictor Variables Suspected of Influencing Infant Mortality in West Sulawesi

There are two ways to determine the multicollinearity case of the predictor variables were used, with the correlation coefficient between the predictor variables and the value of Variance Inflation Factor (VIF) of each predictor. The following is the correlation coefficient between the predictor variables suspected to affect the number of infant mortality in West Sulawesi shown by Table 8.
The correlation coefficient between the variables $X_1, X_2, X_3, X_4, X_5$, and $X_6$ shown in Table 10 is not more than ±0.95, so it can be said that the predictor variables used do not have a correlation or multicollinearity does not occur. In addition to using correlation values, checking multicollinearity assumptions can use VIF values. Following are the VIF values on the factors thought to influence the number of infant mortality in West Sulawesi in 2016 shown in Table 9.

Table 9. VIF of Factors that Thought to Influence of Infant Mortality in West Sulawesi in 2016

| Variabel | VIF |
|----------|-----|
| $X_1$    | 1.441 |
| $X_2$    | 1.243 |
| $X_3$    | 1.143 |
| $X_4$    | 1.111 |
| $X_5$    | 1.306 |
| $X_6$    | 1.154 |

Table 9 shows the VIF on each of the predictor variables. There are no VIF for all predictor variables that exceed 10, so it can be concluded that the predictor variables or factors that are suspected to influence the number of infant deaths in West Sulawesi in 2016 do not correlate with the predictor variables or multicollinearity cases do not occur.

4.5. Zero Inflated Poisson (ZIP) Regression on the Number of Infant Mortality

The Zero Inflated Poisson Regression model is a model used to analyze count data that experience overdispersion and contain zero values in large proportions. The estimation results of the ZIP Regression model parameters are shown in Table 10.

Table 10. ZIP Regression Parameter Estimation Results

| Parameter | Estimation | SE   | Z    | p-value |
|-----------|------------|------|------|---------|
| $\beta_0$ | 1.1430     | 0.8156 | 1.402 | 0.1611  |
| $\beta_1$ | 0.0067     | 0.0025 | 2.661 | 0.0078  |
| $\beta_2$ | -0.0034    | 0.0037 | -0.952 | 0.3411  |
| $\beta_3$ | -0.0017    | 0.0028 | -0.597 | 0.5502  |
| $\beta_4$ | -0.0015    | 0.0075 | -0.199 | 0.8422  |
| $\beta_5$ | 0.0174     | 0.0025 | 7.067 | 1.59x10^{-12} |
| $\beta_6$ | -0.0121    | 0.0025 | -4.837 | 1.32x10^{-6} |
| $\gamma_0$ | 6.6776 | 3.634 | 1.838 | 0.0661 |
| $\gamma_1$ | -0.0512 | 0.0194 | -2.634 | 0.0084 |
| $\gamma_2$ | 0.0161 | 0.0201 | 0.802 | 0.4227 |
| $\gamma_3$ | 0.0065 | 0.0207 | 0.313 | 0.7541 |
| $\gamma_4$ | -0.0755 | 0.0409 | -1.846 | 0.0649 |


An initial analysis is testing the suitability of the model to determine whether the model is appropriate. The significance level used is $\alpha = 0.05$, the rejection area is $\chi^2_{0.05;14}$ of 24.996. So, the decision that can be taken is to reject $H_0$ because the value of $G^2$ is 303 more than $\chi^2_{0.05;14}$, which is 24.996. The conclusion is that the ZIP regression model is appropriate. After knowing that the model is suitable, the next step is partial testing. Partial testing is used to find out which parameters have a significant effect on the model.

With a significance level $\alpha = 0.05$, a decision can be taken if Z score is more than $Z_{(0.025)}$ or Z score is more than 1.96 and Z is less than $-Z_{(0.025)}$ or Z is less than -1.96. Based on Table 12, the value of Z in the log model that is more than the critical region $-Z_{(0.025)}$ is $\beta_2, \beta_3$ and $\beta_5$ parameter with a value of -0.952; -0.597 and -0.199. The decision that can be taken is failing to reject $H_0$, so the conclusion is the variable percentage of Household with Health and Clean Behavior ($X_1$), percentage of UCI Village ($X_3$) and percentage of maternity mothers who assisted by medical workers ($X_4$) do not have a significant effect on the log model. Z score in the parameters $\beta_1$ and $\beta_5$ are 2.661 and 7.067, more than the $Z_{(0.025)}$ score is 1.96. The decision that can be taken is to reject $H_0$ and the conclusion is $X_1$’s variable or percentage of active maternal and child health services and $X_4$’s variable or percentage of society with good drinking water adequate access has a significant effect on the model. While the Z score less than the critical area $-Z_{(0.025)}$ is parameter $\beta_6$, which is -4.837 less than -1.96, it is found that the decision to reject $H_0$, The conclusion that can be drawn is the percentage of society with proper sanitation access to proper sanitation has a significant effect.

The decision to reject $H_0$ can be taken if Z is more than $Z_{(0.025)}$ or less than $-Z_{(0.025)}$ which is 1.96 or -1.96. The Z score in Table 12 that is less than $Z_{(0.025)}$ or 1.96 is the parameters $\gamma_2$ and $\gamma_3$ namely 0.802 and 0.313, so the decision is to fail to reject $H_0$. The conclusion is that the variables $X_2$ and $X_3$ or the percentage of Household with Health and Clean Behavior and the percentage of UCI village have no significant effect. Z score greater than $-Z_{(0.025)}$ or -1.96 are parameters $\gamma_4, \gamma_5$ and $\gamma_6$ with a value of -1.846; -0.021 and -1.383. So the decision that can be taken is failing to reject $H_0$ and the conclusion is the variables $X_4, X_5$ and $X_6$ or the percentage of maternity mothers who assisted by medical workers, the percentage of society with good drinking water adequate access and percentage of society with proper sanitation access has no significant effect. Based on Table 12, the value of Z in the logit model parameter $\gamma_1$ is -2.222 less than $-Z_{(0.025)}$ which is -1.96. So, the decision is to reject $H_0$. The conclusion obtained is that in the logit model, the percentage of active maternal and child health services or $X_1$ significant effect.

### 4.6. Selection of the Best Model on Data on Number of Infant Mortality and Factors Suspected of Influencing

The selection of the best model in Zero Inflated Poisson regression uses the criteria of Akaike's Information Criterion (AIC) and Akaike's Information Corrected Criterion (AICc). The combined results of all variables obtained the smallest AIC and AICC values shown in Table 11.
Variable Combination $X_1, X_5, X_6$ show the lowest AIC and AICc values compared to the combination of other variables. The smaller the value of AIC and AICc, the better the model. So, it can be concluded that the best model is a model with a combination of variables $X_1, X_5$ and $X_6$ or a combination of percentage of active maternal and child health services, percentage of society with good drinking water adequate access and percentage of society with proper sanitation access.

Table 11. The smallest AIC and AICc Values in a Variable Combination

| Variable Combination | AIC       | AICc      |
|----------------------|-----------|-----------|
| $X_1, X_5, X_6$      | 324.3407  | 326.7407  |

Based on Table 11, the values of AIC and AICc on the combination of variables $X_1, X_5$ and $X_6$ is 324.3407 and 326.7407, show the lowest AIC and AICc values compared to the combination of other variables. The smaller the value of AIC and AICc, the better the model. So, it can be concluded that the best model is a model with a combination of variables $X_1, X_5$ and $X_6$ or a combination of percentage of active maternal and child health services, percentage of society with good drinking water adequate access and percentage of society with proper sanitation access.

Then a simultaneous test and a partial test are performed to find out which variables have a significant effect on each model. Table 12 shows the results of parameter estimation on the combination of variables $X_1, X_5$ and $X_6$.

Table 12. ZIP Regression Best Model Parameter Estimation Results on variables $X_1, X_5$ and $X_6$

| Parameter | Estimation | SE   | Z    | p-value |
|-----------|------------|------|------|---------|
| $\beta_0$ | 0.7344     | 0.2782 | 2.640 | 0.00829 |
| $\beta_1$ | 0.0061     | 0.0024 | 2.491 | 0.01273 |
| $\beta_5$ | 0.0174     | 0.0025 | 6.954 | 3.54x10^{-12} |
| $\beta_6$ | -0.0124    | 0.0025 | -5.067 | 4.03x10^{-6} |
| $\gamma_0$ | 1.4107     | 1.1193 | 1.260 | 0.2075 |
| $\gamma_1$ | -0.0431    | 0.0200 | -2.155 | 0.0312 |
| $\gamma_5$ | 0.0024     | 0.0237 | 0.099 | 0.9209 |
| $\gamma_6$ | -0.0384    | 0.0335 | -1.148 | 0.2510 |

Furthermore, a simultaneous model test is performed to determine whether the best model is suitable and can be used. With a significance level of $\alpha = 0.05$, the rejection area used is $\chi^2_{(0.05;8)}$ which is 15.507. The decision that can be taken is to reject $H_0$ because the $G^2$ value of 308.4 is more than $\chi^2_{(0.05;8)}$ that of 15.507. The conclusion that can be drawn is that the best ZIP Regression model is appropriate and can be used. Next, a partial test is performed to find out which parameters have a significant effect on the model.

The decision to reject $H_0$ can be taken if the value of $Z$ is more than $Z_{(0.025)}$ which is 1.96 or $Z < -Z_{(0.025)}$ is -1.96. In Table 14 shows the score of $Z$’s in parameter $\beta_6$ that is equal to -5.067 whose score is less than $-Z_{(0.025)}$ or -1.96. The conclusion that can be taken is to reject $H_0$ so that the conclusion is variable $X_6$ or the percentage of society with proper sanitation access has a significant effect. The parameters $\beta_1$ and $\beta_5$ have $Z$ score of 2.491 and 6.954. Because the score of $Z$’s is more than 1.96, it can be decided to start with reject $H_0$. So the conclusion is the percentage of active maternal and child health services and percentage of society with good drinking water adequate access have a significant effect on the log model.

The decision to reject $H_0$ can be taken if the value of $Z$ is more than $Z_{(0.025)}$ which is 1.96 or $Z$ is less than $-Z_{(0.025)}$ which is -1.96. Based on Table 14 it can be seen that the parameter $\gamma_1$ has a $Z$ score
of 0.099. This value is less than the value $Z_{0.025}$ or 1.96, so the decision is a failure to reject $H_0$ and the conclusion is the variable percentage of society with good drinking water adequate access ($X_4$) has no significant effect. Parameter $\gamma_6$ has a value of $Z$ is -1.148 whose value is more than the value of -1.96. The decision was the failure to reject $H_0$ and the conclusion that could be drawn was that the percentage of society with proper sanitation access ($X_6$) had no significant effect on the logit model. The parameter $\gamma_1$ has a Z value of -2.155 which is less than $-Z_{0.025}$ is -1.96 so the decision is reject $H_0$. The conclusion that can be drawn is $X_1$ or percentage of active maternal and child health services has a significant effect on the logit model. 

So that the best ZIP regression model is

$$\log(\lambda_i) = 0.7344 + 0.0061 x_{1i} + 0.0174 x_{5i} - 0.0124 x_{6i}$$

and

$$\logit(\pi_i) = 1.411 - 0.0431 x_{1i} + 0.0024 x_{5i} - 0.0384 x_{6i}$$

The log model shows a model that influences Poisson state with variable percentage of active maternal and child health services, percentage of society with good drinking water adequate access and percentage of society with proper sanitation access. As the percentage of active maternal and child health services in one unit increases, the chance of infant mortality will increase by $e^{0.0061}$ or 1.006 times. If the percentage of society with good drinking water adequate access increases by one, the chance of infant mortality will increase by $e^{0.0174}$ or 1.0176 times. If the percentage of society with proper sanitation access increases by one unit, it will reduce the chance of infant mortality by $e^{-0.0124}$ or 0.9877 times. The logit model shows a model that affects on zero state where the variable that has a significant effect is the percentage of active maternal and child health services variables. The risk of infant mortality from a zero state that has an active maternal and child health service is $e^{0.0431}$ or 0.9578 fold compared to the absence of infant mortality from Poisson state. The discrepancy of signs in the log and logit model that is formed is due to the absence of a linear relationship between the variable number of infant mortality and the predictor variable that has a significant effect. The following is a graph showing that there is no linear relationship between the response variables and their predictors.

**Figure 1** The Graph of ln (y) with X_1  

**Figure 2** The Graph of ln (y) with X_5  

**Figure 3** The Graph of ln (y) with X_6
Figures 1, 2 and 3 show the relationship between the variable ln (y) with the three predictor variables that have a significant effect indicating there is no linear relationship. That is because the distribution of data represented by blue dots does not form certain patterns. So the mark on the parameter estimation is not by the existing theory.

5. Conclusions and Recommendations

A. Conclusions

Based on the analysis and discussions that carried out in chapter 4, the conclusions are:

1. The characteristics of factors that are thought to affect the number of infant deaths in West Sulawesi are the average percentage of active Maternal and Child Health Services of 60.75%, the variable percentage of household with Health and Clean Behavior has an average of 67.47%, the percentage of UCI villages is 63.16%, the average percentage of maternity mothers who assisted by medical workers is 77.60%, the percentage of the society that has access to adequate drinking water is 38.89% and the percentage of the society that have access to proper sanitation is 49.52%. Two districts have the highest number of infant deaths are Banggai and Pasangkayu districts with 15 infant deaths. The other 16 districts that did not occur in infant mortality during 2016 are Tapoyo, Pangale, Balabalakang, Kalumpang, Bonehau, Sampaga, Papalang, Mambi, Bambang, Rantim, Tabang, Nosu, Matangnga, Balanipa, Luyo and Mapili.

2. The results of ZIP regression with the best method are the combination of variables $X_1$, $X_5$ and $X_6$ or the percentage of active Maternal and Child Health Services and the percentage of maternity mothers who assisted by medical workers. In the log or Poisson state model, it is found that the variables that have a significant effect are the percentage of active Maternal and Child Health Services, the percentage of society with adequate access of drinking water and the percentage of society with proper sanitation. In the logit or zero state model, it is found that the variable which has a significant effect is the percentage of active Maternal and Child Health Services.

B. Recommendations

Based on the conclusions, the recommendations for further research can be formulated as follows:

1. We recommend that the West Sulawesi provincial health office add a predictor variable in terms of socio-economic aspects, because this variable is difficult to obtain even though this variable also affects in the number of infant deaths.

2. We recommend that the West Sulawesi provincial health office focus more on improving active maternal and child health services and improving the quality of childbirth by medical workers.

References

[1] Badan Pusat Statistik, "Sistem Informasi Rujukan Statistik," 12 Februari 2019. [Online]. Available: https://bps.go.id/index.php?r=indikator/view&id=79.

[2] S. Handayani and N. L. Pratiwi, "Health Seekong Behaviour pada Persalinan Suku Kaili Da'a di Desa Wulai Kabupaten Mumuju Utara Sulawesi Barat," Buletin Penelitian Sistem Kesehatan, vol. 19, pp. 91-98, 2016.

[3] M. P. Putri and P., "Analisis Faktor-faktor yang Berpengaruh Terhadap Jumlah Kematian Bayi di Provinsi Jawa Tengah," Journal Sains dan Seni ITS, vol. 06, pp. 108-114, 2017.

[4] D. Lambert, "Zero-Inflated Poisson Regression, With an Application to Defect in Manufacturing," American Statistical Association and the American Society for Quality Control, vol. 34, pp. 1-14, 1992.

[5] K. K. Indonesia, "Profil Kesehatan Indonesia 2014," Kementrian Kesehatan Republik Indonesia, Jakarta, 2015.

[6] R. E. Walpole, R. H. Myres, S. L. Myres and K. Ye, Probability & Statistics for Engineers & Scientists 8th Edition, United States of America: Prentice Hall, 2007.
[7] Setiawan and D. E. Kusrini, Ekonometrika, Yogyakarta: C.V Andi Offset, 2010.
[8] J. M. Hilbe, Negative Binomial Regression, New York: Cambridge University Press, 2011.
[9] J. Hinde and C. B. Demetrio, Overdispersi: Model and Elimination, Exeter: Departement of Laver Building, 2007.
[10] D. B. Hall and J. Shen, "Robust Estimation for Zero Inflated Poisson Regression," Scandividia Journal of Statistics, 2009.
[11] H. Bozdogan, Akaike's Information Criterion and Recent Developments in Information Complexity. Journal of Mathematical Psychology, vol. 44, 62-91, 2000.