MODELING OF EXCESS ZERO COUNT DATA WITH ZERO INFLANTED POISSON REGRESSION

(Case Study of Filariasis Occurrence in East Java)

Dany Suhardiyanto
Department of Statistics, Faculty of Mathematics and Natural Science
Bogor Agricultural University, Indonesia

Muhammad Nur Aidi
Department of Statistics, Faculty of Mathematics and Natural Science
Bogor Agricultural University, Indonesia

Farit M Afendi
Department of Statistics, Faculty of Mathematics and Natural Science
Bogor Agricultural University, Indonesia

Abstract

Zero-inflated Poisson regression (ZIP) is used to model count data that has mostly an excess of zero counts (over 50%) and others follow Poisson distribution. The existence of the grouping of zeros (zero state) and Poisson distribution (Poisson state) data causes the parameter estimation process is done by expectation-maximization (EM) algorithm. ZIP regression is widely applied in the fields of health, epidemiology and also insurance. Filariasis or more commonly known as elephantiasis is an infectious disease caused by filarial worms and transmitted by various species of mosquitoes. In the year of 2000, Indonesia through the Ministry of Health agreed to take part in the world Filariasis elimination program, by launching the National Plan of Filariasis Elimination Acceleration Program in Indonesia from 2010 to 2014. The program was followed by breaking the chain of transmission by distributing mass drug of disease prevention (POMP) in endemic areas of Filariasis and also preventing and limiting defect cases caused by filariasis. In the province of East Java, the number

of clinical cases of chronic filariasis had been recorded until 2012 was 341 cases spreaded in 32 cities in 180 districts and 259 villages. Elephantiasis elimination program can be done more efficiently if the factors that influence it are known. The link of these factors with the number of filariasis sufferers can be analyzed by using regression method. The purpose of this study was to determine the factors that influence the occurrence of filariasis cases in East Java with Zero-Inflated Poisson regression. The result of this study shows there is a variable that gives significant impact on the number of filariasis cases, which is the percentage of families who have a healthy litter (X_3) both for state poisson models and zero inflation models.

Keywords: Zero-inflated Poisson regression, Excess Zero, Poisson regression, Filariasis

1. Introduction

The number of infrequent occurrences can be illustrated by the count data that follows the Poisson distribution. The modeling of count data of poisson distribution is often done by Poisson regression. However, there is assumption of similarities mean and variance of samples that must be fulfilled on the Poisson regression [1]. Error in this assumption can occur as a result of the diversity of the population or the great number of zeros (excess zero) in the data [2] and [3]. For that reason, the use of Poisson regression on the count data that contain lots of zeros (excess zero) can result in overdispersion cases. One of methods for analyzing the count data of Poisson distribution containing lots of zeros (Excess Zero) is the Zero-Inflated Poisson regression. Zero-Inflated Poisson (ZIP) regression was introduced by [4] to model the count data that mostly value zeros (more than 50%) and others follow poisson distribution. The estimation of ZIP regression parameter was conducted with Maximum Likelihood Estimation (MLE) model. The existence of the grouping of zeros (zero state) and Poisson distribution (Poisson state) data causes the parameter estimation process is done by Expectation-Maximization (EM) algorithm. ZIP Regression is applied in many areas of health, epidemiology and also insurance.

Filariasis or more commonly known as elephantiasis is an infectious disease caused by the filarial worms and transmitted by various species of mosquitoes. There are three species of worms that cause filariasis in Indonesia namely Wuchereria bancroft, Brugia timori and Brugia malayi. More than 70% of filariasis cases in Indonesia is caused by Brugia malayi. Filariasis can be transmitted by all types of mosquitoes. Filariasis clinical symptoms will appear

after a few bites of mosquitoes that have been filariasis-infected in a long period of time [5].

With the increasing of filariasis cases which occurs in many countries, especially 60% of cases occured in Southeast Asia, the World Health Organization (WHO) in 2000 declared "The Global Goal of Elimination of Lymphatic Filariasis as a Public Health Problem by the Year 2020." Indonesia through Ministry of Health agreed to take part in the world Filariasis elimination program, by launching a National Plan of Filariasis Elimination Acceleration Program in Indonesia from 2010 to 2014. The program was followed by breaking the chain of transmission by distributing mass drug of disease prevention (POMP) in endemic areas of Filariasis and also preventing and limiting defect cases caused by filariasis.

Elephantiasis elimination program can be done more efficiently if the factors that influence it have been known. The link between these factors and the number of filariasis sufferers can be analyzed using regression methods. The relationship between filariasis sufferers in the Province of East Java as variable respon can be assumed to follow poisson distribution because occurrence of filariasis is a relatively infrequent event that resulted in the data of excess zero. Thus the purpose of this study was to determine the factors that influence the occurrence of filariasis cases in the Province of East Java with Zero Inflated Poisson regression.

2. REFERENCES

Zero-Inflation

The excess zero value at response variable (zero inflation) is often found in Poisson regression analysis both for discrete data or count data. If the value of zero has significant value in a research, then this data cannot be eliminated but must be included in the analysis process. In several studies, there are conditions in which too many zeros at response variable of more than 50 percent. According to [6], the large proportion of zeros in the data may result in the accuracy of inference. In addition, the Poisson regression is also no longer appropriate to model the actual data.

Zero-Inflated Poisson Regression (ZIP Regression)

ZIP regression model is one of the alternative methods for analyzing data with many zero values contained in the response variable. The great number of zero values in the data can result in a error on the assumption of mean similarity and variance in the Poisson distribution. For any observations on the response variables, which are independent Y1, Y2, Y3, ... Yn, and

$$Y_i \sim \begin{cases} 0 & , \text{ with probability } \pi_i \\ \text{Poisson}(\mu_i) & , \text{ with probability } (1-\pi_i) \end{cases}$$

Function of probability for Y_i is:

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

With parameter $\boldsymbol{\mu} = (\mu_1\ \mu_2 \dots\ \mu_n)^T$ dan $\boldsymbol{\pi} = (\pi_1\ \pi_2 \dots\ \pi_n)^T$ that meets equation:

$$\mu_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}}$$

$$\pi_i = \frac{1}{1+e^{\mathbf{x}_i^T \boldsymbol{\gamma}}} \quad \text{and} \quad (1-\pi_i) = \frac{1}{1+e^{\mathbf{x}_i^T \boldsymbol{\gamma}}}$$

ZIP regression model can be formulated as:

$$\ln \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}\ ; i = 1, \dots, n$$

$$\text{logit}\, \pi_i = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_k x_{ik}\ ; i = 1, \dots, n$$

Where,

β : Vector from regression parameter that will be assumed

γ : Vector from regression parameter that will be assumed

X : matrix of *n x* (*k+1*) containing explanatory variables related to probability at *Zero state* ($y_i = 0$) and mean at *poisson state* ($y_i > 0$).

Expectation value and variance of $\mathbf{Y}_i$ can be formulated as follows:

$$E(Y_i) = \mu_i(1-\pi_i)$$

$$Var(Y_i) = \mu_i(1-\pi_i)(1+\mu_i\pi_i)$$

Parameter estimation at ZIP regression models is conducted using Maximum Likelihood Estimation (MLE), where the probability density function of **Y**i is known [4]. The function of likelihood of the ZIP regression model formed is:

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \begin{cases} \prod_{i=1}^{n} \frac{e^{\mathbf{x}_i^T \boldsymbol{\gamma}} + e^{-e^{\mathbf{x}_i^T \boldsymbol{\beta}}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\gamma}}} & , \text{for } y_i = 0 \\ \prod_{i=1}^{n} \frac{1}{1+e^{\mathbf{x}_i^T \boldsymbol{\gamma}}} \frac{\left(\exp\left(-e^{\mathbf{x}_i^T \boldsymbol{\beta}} + (\mathbf{x}_i^T \boldsymbol{\beta}) y_i\right)\right)}{y_i!} & , \text{for } y_i > 0 \end{cases}$$

and the function of the ln likelihood is

$$\ln(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \begin{cases} \sum_{i=1}^{n} \ln\left(e^{\mathbf{x}_i^T \boldsymbol{\gamma}} + e^{-e^{\mathbf{x}_i^T \boldsymbol{\beta}}}\right) - \sum_{i=1}^{n} \ln\left(1 + e^{\mathbf{x}_i^T \boldsymbol{\gamma}}\right) & , \text{for } y_i = 0 \\ \sum_{i=1}^{n} \left(-e^{\mathbf{x}_i^T \boldsymbol{\beta}} + (\mathbf{x}_i^T \boldsymbol{\beta}) y_i\right) - \sum_{i=1}^{n} \ln\left(1 + e^{\mathbf{x}_i^T \boldsymbol{\gamma}}\right) - \sum_{i=1}^{n} y_i! & , \text{for } y_i > 0 \end{cases}$$

$$= \sum_{\substack{i=1 \\ y_i=0}}^{n} \ln\left(e^{\mathbf{x}_i^T\mathbf{Y}} + e^{-e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right) - \sum_{i=1}^{n} \ln\left(1 + e^{\mathbf{x}_i^T\mathbf{Y}}\right) + \sum_{\substack{i=1 \\ y_{i>0}}}^{n} \left(-e^{\mathbf{x}_i^T\boldsymbol{\beta}} + (\mathbf{x}_i^T\boldsymbol{\beta})y_i\right) - \sum_{\substack{i=1 \\ y_{i>0}}}^{n} y_i!$$

The above equation is referred to as incomplete likelihood. This is because the zero value on the first syllable is not known which ones coming from the zero state and which ones from the Poisson state, so it is resolved by redefining $\mathbf{Y}_i$ variables with indicator variable $\mathbf{Z}_i$ that:

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

If yi > 0, then the value of Zi = 0. However, if yi = 0, then it can be 0 or 1. This problem can be solved by using EM (Expectation-Maximization) algorithm. EM algorithm is one alternative iterative method for maximizing the function of likelihood that contains incomplete (missing) data. In addition, the EM algorithm is also used in the data that contain the latent variables [7].

At each iteration, the EM algorithm consists of two stages: Expectation stage and Maximization stage. Expectation stage is the stage of expectation calculation of the function of In likelihood with regard to the incomplete data. Whereas, maximization stage is the calculation stage to find parameter estimators that maximize the function of ln likelihood as the result from previous expectation stage.

To estimate the parameters of Z_i, that have been defined, the following steps are done:

1. Calculate the expectation value of Z_i from , which is:

$$E\left(Z_i|y_i, \gamma^{(m)}, \beta^{(m)}\right) = Z_i^{(m)}$$

$$Z_i^{(m)} = P\left(Z_I = 1|y_i, \gamma^{(m)}, \beta^{(m)}\right)$$

$$= \begin{cases} P\left(Z_I = 1|y_i, \gamma^{(m)}, \beta^{(m)}\right) & , \text{for } y_i = 0 \\ 0 & , \text{for } y_i > 0 \end{cases}$$

$$= \begin{cases} \dfrac{1}{1 + \exp(-\mathbf{x}_i^T\boldsymbol{\gamma} - \exp(\mathbf{x}_i^T\boldsymbol{\beta}))} & , \text{for } y_i = 0 \\ 0 & , \text{for } y_i > 0 \end{cases}$$

2. Maximation of $\boldsymbol{\beta}$ is done with Newton-Raphson iterative method to obtain $\boldsymbol{\beta}^{(m+1)}$. With gradient vector (**g**) and Hessian matrix (**H**) as follows:

$$\mathbf{g}^{T^{(m)}} = \mathbf{X}^T\mathbf{S}^{(m)}(\mathbf{y} - \boldsymbol{\mu})$$

And

$$\mathbf{H}^{(m)} = -\mathbf{X}^T\mathbf{S}^{(m)}\mathbf{T}\mathbf{X}$$

$\mathbf{S}^{(m)}$is diagonal matrix with $\left(1 - Z_i^{(m)}\right)$ as the main diagonal element of $\mathbf{T}$ and diagonal matrix with $\boldsymbol{\mu}$ as its main diagonal.

3. Maximation of $\boldsymbol{\gamma}$ where for each value of $y_i > 0$ value of $Z_i^{(m)} = 0$ so that it becomes:

   $\ln L\left(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{Z}^{(m)}\right) = \sum_{y_i=0} Z_i^{(m)}\mathbf{x}_i^T\boldsymbol{\gamma} - \sum_{y_i=0} Z_i^{(m)} \ln\left(1 + e^{\mathbf{x}_i^T\boldsymbol{\gamma}}\right) -$
   $\sum_{y_i=0}\left(1 - Z_i^{(m)}\right) \ln\left(1 + e^{\mathbf{x}_i^T\boldsymbol{\gamma}}\right)$

   Suppose $y_1$until $y_{n_)}$ is 0, or can be formulated as $y_1, y_2, \ldots, y_{n_)}$. Then, being defined that diagonal matrix $\mathbf{V}^{(m)}$ with main diagonal element:

   $$\mathbf{V}_*^T = \left(1 - Z_1^{(m)}, 1 - Z_2^{(m)}, \ldots, 1 - Z_{n+1}^{(m)}, 1 - Z_{n+n_0}^{(m)}\right)$$

   also defined:

   $$\mathbf{y}_*^T = \left(y_1, y_2, \ldots, y_n, y_{n+1}, \ldots, y_{n+n_)}\right)$$
   $$\mathbf{X}_*^T = (\mathbf{1}, \mathrm{X}_1^T, \mathrm{X}_2^T, \ldots, \mathrm{X}_k^T)$$
   $$\boldsymbol{\pi}_*^T = \left(\pi_1, \pi_2, \ldots, \pi_n, \pi_{n+1}, \ldots, \pi_{n+n_)}\right)$$

   With $\mathbf{X}_*^T$ of $(k+1) \times (n+n_0)$ then can be formulated in the form of:

   $\ln L\left(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{Z}^{(m)}\right) = \sum_{i=0}^{n+n_0} y_{*i} v_i^{(m)} \mathbf{X}_{*i}^T\boldsymbol{\gamma} - \sum_{i=1}^{n+n_0} v_i^{(m)} \ln\left(1 + e^{\mathbf{x}_i^T\boldsymbol{\gamma}}\right)$

   Gradient vector (**g**) and Hessian matrix (**H**) can be formulated as follows:

   $\mathbf{g}^T = \mathbf{X}_*^T\mathbf{V}^{(m)}(\mathbf{y}_* - \boldsymbol{\pi}_*)$

   And

   $\mathbf{H}^{(m)} = -\mathbf{X}^T\mathbf{V}^{(m)}\mathbf{Q}_*\mathbf{X}_*$

   Where $\mathbf{Q}_*$is diagonal matrix with its main diagonal element of $\pi_i(1 - \pi_i)$. Steps to obtain $\gamma^{(m+1)}$ is identical with Newton-Raphson iterative method, like it is done at the step to maximize, with $\mathbf{y}_*$ as response of $\mathbf{X}_*$, as variable matrix, and $\mathbf{V}^{(m)}$ as weighing matrix.
4. Change $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ into $\widehat{\boldsymbol{\beta}}^{(m+1)}$and $\hat{\boldsymbol{\gamma}}^{(m+1)}$, then redo the first step (expectation stage).
5. Expectation and Maximization stage is carried out continuously until obtaining the convergent parameter estimator.

Testing ZIP Regression Model Parameter

Testing parameters in ZIP regression models is done using Maximum Likelihood Ratio Test (MLRT). Each of tests and hypotheses and their statistics of likelihood ratio used will be described below [4].

1. Simultaneous Testing (Parameter of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$)

   Parameters tested in this simultaneous test covers all parameters of $\boldsymbol{\beta}$ and

$\boldsymbol{\gamma}$ together. The hypotheses used is:
H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = \gamma_1 = \gamma_2 = \cdots = \gamma_k = 0$
H_1: there is one for minimum or $\beta_j \neq 0$ or $\gamma_j \neq 0$, $j = 1,2,\ldots,k$.

$$G = -2\ln\left[\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right]$$

$$= -2\ln\left[\frac{\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{\hat{\gamma}_0}}\right)\left(e^{\hat{\gamma}_0}\right)^{z_i}\left(\frac{e^{-e^{\hat{\beta}_0}}\left(e^{\hat{\beta}_0}\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}{\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}}\right)\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}\right)^{z_i}\left(\frac{e^{-e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}}\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}\right]$$

Statistics of G distributed $\chi^2_{(df)}$ so that at significant level of α rejected if the value of $G > \chi^2_{(df,\alpha)}$, where df is the number of parameters under population minus the number of parameters under H_0.

2. Simultaneous Testing (Parameter of $\boldsymbol{\beta}$)
Parameters tested in this simultaneous test covers all parameters of $\boldsymbol{\beta}$ together. The hypotheses used is:
H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$
H_1: there is one for minimum or $\beta_j \neq 0$, $j = 1,2,\ldots,k$.

$$G = -2\ln\left[\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right]$$

$$= -2\ln\left[\frac{\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}}\right)\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}\right)^{z_i}\left(\frac{e^{-e^{\hat{\beta}_0}}\left(e^{\hat{\beta}_0}\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}{\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}}\right)\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}\right)^{z_i}\left(\frac{e^{-e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}}\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}\right]$$

Where G distributed $\chi^2_{(k)}$ so that rejected if the value of $G > \chi^2_{(k,\alpha)}$.

3. Simultaneous Testing (Parameter of $\boldsymbol{\gamma}$)
Parameters tested in this simultaneous test covers all parameters of $\boldsymbol{\gamma}$ together. The hypotheses used is:
H_0: $\gamma_1 = \gamma_2 = \cdots = \gamma_k = 0$
H_1: there is one for minimum or $\gamma_j \neq 0$, $j = 1,2,\ldots,k$.

$$G = -2\ln\left[\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right]$$

$$= -2\ln\left[\frac{\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{\hat{\gamma}_0}}\right)\left(e^{\hat{\gamma}_0}\right)^{z_i}\left(\frac{e^{-e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}}\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}{\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}}\right)\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}\right)^{z_i}\left(\frac{e^{-e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}}\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}\right]$$

Where *G* distributed $\chi^2_{(k)}$ so that rejected if the value of $G > \chi^2_{(k,\alpha)}$.

4. Partial Testing (parameter of $\boldsymbol{\beta}$)

   Parameters tested in this test covers all parameters partially. The hypotheses used is:

   H_0: $\beta_j = 0$

   H_1: $\beta_j \neq 0,\ j = 1,2,\ldots,k.$

$$G = -2\ln\left[\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right]$$

$$= -2\ln\left[\frac{\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}}\right)\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}\right)^{z_i}\left(\frac{e^{-e^{\hat{\beta}^*}}\left(e^{\hat{\beta}^*}\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}{\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}}\right)\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}\right)^{z_i}\left(\frac{e^{-e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}}\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}\right]$$

Where *G* distributed $\chi^2_{(1)}$ so that rejected if the value of $G > \chi^2_{(1,\alpha)}$.

5. Partial Testing (parameter $\boldsymbol{\gamma}$)

   Parameters tested in this test covers all parameters partially. The hypotheses used is:

   H_0: $\gamma_j = 0$

   H_1: $\gamma_j \neq 0,\ j = 1,2,\ldots,k.$

$$G = -2\ln\left[\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right]$$

$$= -2\ln\left[\frac{\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}^*}}\right)\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}^*}\right)^{z_i}\left(\frac{e^{-e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}}\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}{\prod_{i=1}^{n}\left[\left(\frac{1}{1+e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}}\right)\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\gamma}}}\right)^{z_i}\left(\frac{e^{-e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}}\left(e^{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}\right)^{y_i}}{y_i!}\right)^{1-z_i}\right]}\right]$$

Where G distributed $\chi^2_{(1)}$ so that rejected if the value of $G > \chi^2_{(1,\alpha)}$.

3. Data

The data used are secondary data on the occurrence of elephantiasis (filariasis), from the Health Profile of Province of East Java 2012. The units of observation in this study covered 38 districts/cities in East Java which covered 29 districts and 9 cities. The explanation of each variable is presented in Table 1.

Table 1 Details of Variables Used in the study

Name of Variables	Status	Explanation
The number of occurrence of Filariasis (Y)	Response Variable	The number of *Filariasis* cases in each districts/cities in East Java
BPHBS household (X1)	Explanatory Variable	Percentage of households with clean and healthy living behavior in each district / city
Healthy toilet (X2)	Explanatory Variable	Percentage of households that have healthy toilet in each district / city
Healthy trash (X3)	Explanatory Variable	Percentage of households that have healthy trash in each district / city
Healthy waste water management (X4)	Explanatory Variable	Percentage of households that have healthy waste water management in each district / city

Analysis Method

Data analysis was conducted by the following steps.

a. Conducting descriptive analysis on the variables of the study.
b. Examining the proportion of zero value on the response variable.
c. Modeling Zero Inflated Poisson (ZIP) regression
d. Testing the significance of the regression model parameters.
e. Interpreting Zero Inflated Poisson (ZIP) regression model formed.
f. Determining the level of goodness of Zero Inflated Poisson (ZIP) regression model formed.

4. Results And Discussion

Overview of filariasis in East Java

Elephantiasis is an infectious tropical disease transmitted by mosquitoes,

similar to malaria, leprosy and dengue fever. Based on geographical location, East Java lies in $7,12^{o}$ – $8,48^{o}$ South latitude (LS) and 111^{o} – $114,40^{o}$ East longitude (BT). East Java consists of 38 regencies / cities covering 29 districts and 9 cities.

Response variable used in this study was the number of filariasis cases in Province of East Java in 2012. The greatest number of filariasis cases was 4 cases occurred in Lamongan, while the fewest number of filariasis cases was 0 case (no case of filariasis) occurred in 25 Districts / Cities among others Tulungagung Districts, Blitar Districts, Kediri Districts, Lumajang Districts, Jember Districts, Bondowoso Districts, Probolinggo Districts, Pasuruan Districts, Mojokerto Districts, Nganjuk Districts, Madiun Districts, Magetan Districts, Ngawi Districts, Tuban Districts, Gresik Districts, Sampang Districts, Sumenep Districts, Kediri City, Blitar City, Probolinggo City, Pasuruan City, Mojokerto City, Madiun City, Surabaya City and Batu City. The percentage of zero value in response variable has the greatest percentage which was 65.79%. This became the focus of study.

Modeling the filariasis cases that used ZIP regression model used four explanatory variables, i.e. the percentage of BPHBS household (X1), the percentage of households who have healthy toilet (X2), the percentage of households who have healthy trash (X3) and the percentage of households who have healthy wastewater management (X4). Descriptive overview of each explanatory variable is presented in Table 2.

Table 2 Descriptive Analysis of Explanatory Variables

Variable	Mean	STDEV	Minimum	Maximum
X_1	43.722	14.788	8.500	65.740
X_2	77.286	16.042	25.320	97.460
X_3	60.220	25.619	0	88.620
X_4	58.974	26.100	0	100.000

Examination of Zero-Inflation of Response Variable

Examination of zero inflation was done by calculating the percentage of zero observations at response variable. The examination result of zero inflation on the response variable is presented in Table 3.

Table 3 Examination Result of *Zero Inflation* at Response Variable

The number of Filariasis Cases	Frequency of The Number of Filariasis Cases	Percentage	Cumulative Percentage
0	25	65.79	25
1	7	18.42	32
2	4	10.53	36
3	1	2.63	37

4	1	2.63	38

Examination Result of *Zero Inflation* at Response Variable in Table 3 shows that there is *zero inflation* in response variable because the percentage of zero value observations was more than 50 percent, which was 65.79 percent.

Modeling Zero-Inflated Poisson (ZIP) Regression

The Zero-Inflated Poisson (ZIP) regression model is a regression model that can be used to model data with response variables that have Poisson distribution that contains many zero-value observations at response variable and occurs overdispersion. This model is the development of poisson regression model for data with lots of zero-value observations at the response variable (*zero inflation*). ZIP regression model was applied to the filariasis cases in the Province of East Java. The modeling of Filariasis occurrence used ZIP regression model with four explanatory variables, i.e. the percentage of BPHBS household (X1), the percentage of households who have healthy toilet (X2), percentage of households who have healthy trash (X3), and the percentage of households who have healthy wastewater management (X4). To determine the level of significance of the parameter estimation results at ZIP regression model, significance testing was conducted simultaneously and partially. According to Hosmer and Lemeshow (2000), the testing of the significance of parameter estimation results at ZIP regression models simultaneously utilizes statistical G test statistics and the significance testing partially utilizes *t* test statistics. The results of the estimation of ZIP regression model parameter on filariasis cases and the value of G test statistic and *t* test statistics are presented in detail in Table 6.

Table 6 Result of Estimation of ZIP Regression Model Parameter

Parameter	Estimation	SE	*t* Calculated Value	(Pr >| *t*|)
$\hat{\beta}_0$	-2.680785	1.320942	-2.03	0.0424*
$\hat{\beta}_1$	0.021780	0.022177	0.98	0.3261
$\hat{\beta}_2$	-0.012550	0.013079	-0.96	0.3373
$\hat{\beta}_3$	0.049433	0.022414	2.21	0.0274*
$\hat{\beta}_4$	-0.005950	0.019436	-0.31	0.7595
$\hat{\gamma}_0$	-16.334441	7.813373	-2.09	0.0366*
$\hat{\gamma}_1$	0.117029	0.077640	1.51	0.1317
$\hat{\gamma}_2$	0.068710	0.059112	1.16	0.2451
$\hat{\gamma}_3$	0.149569	0.075519	1.98	0.0476*

Parameter	Estimation	SE	*t* Calculated Value	(Pr >| *t*|)
$\hat{\gamma}_4$	-0.077171	0.059775	-1.29	0.1967
G Test Statistics = 61,76				

*) Significant with significance level of 5 percent

The test result of significance of estimation of ZIP regression model parameter simultaneously with significance level of 5 percent based on the G test statistics. The value of G test statistics was 61.76. The value of G test statistics was greater than $\chi^2_{(0,05;10)} = 18{,}307$. This shows that simultaneously at the explanatory variables of X1, X2, X3, and X4 gave significant impact on the response variable. For the result of significance of estimation of ZIP regression model parameter partially with the significance level of 5 percent was based on the *t* test statistics. Based on Table 5, there are two explanatory variables at the estimation of poisson state model parameter and one explanatory variable at the estimation of zero inflation model parameter having *t* calculated value greater than or equal to *t* (α / 2; 37) = 2,00 and having p value less than α (0.05). This shows that the explanatory variables that significantly affected partially at state poisson and zero inflation models were the intercept and the percentage of households who have healthy trash (X3).

The equation of Zero Inflated Poisson (ZIP) regression model formed is as follows.

a. *poisson state* model for $\hat{\mu}$

$$\hat{\mu} = \exp(-2{,}681 + 0{,}0217X_1 - 0{,}0125X_2 + 0{,}0494X_3 - 0{,}0059X_4)$$

b. *zero inflation* model for $\hat{\pi}$

$$\hat{\pi} = \frac{\exp(-16{,}3344 + 0{,}117X_1 + 0{,}068X_2 + 0{,}149X_3 - 0{,}077X_4)}{1 + \exp(-16{,}3344 + 0{,}117X_1 + 0{,}068X_2 + 0{,}149X_3 - 0{,}077X_4)}$$

Based on the poisson state and zero inflation models formed, there were signs of regression coefficient in contrast to the theory that the percentage of BPHBS households (X1), the percentage of households who have healthy toilet (X2) and the percentage of households who have healthy trash (X3). The existence of the regression coefficient that has a contrary sign to the theory of probability was caused by the shape of the data pattern of the explanatory variables that have positive correlation with the response variables. This different sign was also assumed because the data used was only 38 observations so that the models created were less able to describe what had been expected.

5. Conclusion

1. The Modeling of count data with Zero Excess can be done by using Zero Inflated regression at the response variable, which is the occurrence of filariasis because the percentage of zero-value observations is more than 50 percent, which is 65.79 percent.
2. Based on Zero Inflated Poisson (ZIP) regression model formed, explanatory variables having significant impact on the number of Filariasis occurrence are intercept and the percentage of households who have healthy trash (X3) both for the poisson state and the zero inflation models.

References

[1] Khosghoftaar, T. M., Gao, K., dan Szabo, R. M. (2004). "Comparing Software Fault Predictions of Pure and Zero-Inflated Poisson Regression Models". International Journal of System Science, Vol. 36, No. 11

[2] Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., dan Kirchner, U. (1999). "The Zero-Inflated Poisson Model and The Decayed, Missing, and Filled Teeth Index in Dental Epidemiology". Journal of Royal Statistical Society A, Vol.162, Part.2.

[3] Mouatassim, Y. & Ezzahld, E. H. (2012). "Poisson Regression and Zero-Inflated Poisson Regression : Application to Private Health Insurance Data", European Actuary Journal, Vol. 2

[4] Lambert, D. (1992). Zero Inflated Poisson Regression, With an Application to Defect in Manufacturing. *Technometric, 34*(1).

[5] Department of Health Republic Of Indonesia 2008. *Profil Kesehatan Indonesia. Center of data and information* Department of Health Republic Of Indonesia

[6] Famoye, F., & Singh, K. P. 2006. Zero Inflated Poisson Regression Model with an Applications Domestic Violence to Accident Data. Journal of Data Science, 117-130.

[7] Dempster, A. P., Laird, N. M., dan Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of Royal Statistical Society B, Vol. 39, No.1.