A Study of ZIP and ZINB Regression Modeling for Count Data with Excess Zeros

R N Amalia, K Sadik*, and K A Notodiputro
Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor 16680, West Java, INDONESIA

*E-mail: kusmans@apps.ipb.ac.id

Abstract. Count data with excess number of zeros can cause overdispersion problems. Overdispersion is the presence of greater variability in a data set than would be expected. Overdispersion due to zero-inflated data can be handled either by Zero-inflated Poisson (ZIP) regression or Zero-inflated Negative Binomial (ZINB) regression. In this paper, the performances of different models have been compared based on their MSE, RMSE, bias, and AIC using simulation. The simulated data were generated using ZIP and ZINB distributions. The data were generated using a combination of observation size (n), mean (µ), and the proportion of zeros observation (ω) were imposed to facilitate comparison. To ensure the present of overdispersion, the score test has been applied to the generated data. As expected, the results showed that ZIP and ZINB regression performed better when compared to the Poisson regression. Moreover, in general the simulation results showed that ZINB regression showed better performance than ZIP and Poisson regressions. In this paper, ZINB regression was applied to analyze maternal mortality rate in East Java. The results showed that maternal mortality was significantly affected by the percentage of pregnant woman visiting the clinics for the first time as well as by the percentage of pregnant woman visiting clinics for the fourth time.

1. Introduction

Regression analysis is a parametric analysis technique that is used to model data and identify cause and effect relationships between several variables. Poisson regression is a type of nonlinear regression analysis that is able to model data with response variables in the form of counted data. Poisson regression can be modeled using Generalized Linear Model (GLM), which describes the relationships between explanatory variables and response variables in the form of linear equations. GLM consists of three main components: the random component, the systematic component, and the link function [1].

Poisson regression describes the relationship between a response variable which follows the Poisson distribution and its explanatory variables. Poisson distribution is characterized by the equal value of variance and expected value. A high number of observations with zero value on the response variable can cause overdispersion [2]. Overdispersion is a condition when the variance of the response variable is greater than the expected value. In the case of Poisson regression, overdispersion occurs when the response variable is greater than the mean.

The use of Poisson regression on overdispersed data can cause the standard error of the regression parameters to be underestimated, which in turn would result in invalid conclusions Muniswamy and Molla [4]. Overdispersion caused by excess zero-value observations can be handled by employing Zero-inflated Poisson (ZIP) regression. Results of a research done by Rahayu [5] showed that ZIP regression...
performed better than Poisson regression in modeling overdispersed data with excess zero values. Sepriawandani [7] compared Poisson regression model, Negative Binomial regression, and ZIP regression for overdispersion cases due to excess zeros. The results showed that ZIP regression performed better than Poisson regression and Negative Binomial regression.

Another zero-inflated regression model that can be used to handle overdispersion due to excess zeros is the Zero-inflated Negative Binomial (ZINB) regression model. A research done [9] showed that ZINB regression performed well in handling overdispersed data with a high dispersion parameter and proportion of zero-value observations. Therefore, in order to determine the best method in handling datasets with numerous zero observations in their response variables, this study conducted a simulation study to compare the performance of the model between Poisson regression, ZIP regression, and ZINB regression. The goodness-of-fit of the models are determined based on the value of Root Mean Square Error (RMSE), bias, and AIC.

2. Methodology
2.1. Materials
The materials used in this research are real data and simulated data. Simulated datasets with excess zeros in their response variables were employed in this research. These datasets were generated based on the following model:

\[ \ln(\mu_i) = \beta_0 + \beta_1 x_i \]

Table 1 shows various \( \beta_0 \) and \( \beta_1 \) values of each \( \mu \) parameter value.

| \( \mu \) | \( \beta_0 \) | \( \beta_1 \) |
|---------|---------|---------|
| 2       | 1.10    | -0.15   |
| 5       | 1.10    | 0.20    |
| 15      | 1.10    | 0.50    |
| 30      | 1.10    | 0.70    |

The response variable of each dataset were generated following the Zero-inflated Poisson and Zero-inflated Negative Binomial distribution with parameters \( \mu \) (mean) and \( \omega \) (proportion of zeros observation). The \( \mu \) parameter values were set at 2, 5, 15, and 30, while the \( \omega \) parameter values were set at 0.1, 0.3, 0.5, 0.7, and 0.9. The dispersion parameter (\( \tau \)) value of ZINB distribution was set as 2 (\( \tau = 2 \)). Explanatory variables were generated following the normal distribution, with parameter values set at \( \mu = 3 \) and \( \sigma = 1 \), assuming them as constant variables. Sample sizes were set at \( n = 30, 100, 200, \) and 500, representing small, moderate, and big sample sizes. Simulations were repeated 500 times.

To demonstrate the performance of each method in real life, real data of maternal mortality were also employed in this research, obtained from the 2018 Publication of the East Java Province Health Profile published by the East Java Provincial Health Department. Variables used in this study are shown in Table 2.

| Variable | Definition |
|---------|------------|
| Y       | Number of maternal mortality in |
| X1      | Percentage of pregnant women receiving FE1 tablet |
| X2      | Percentage of pregnant women receiving FE3 tablet |
| X3      | The percentage of pregnant woman visiting clinics for the first time |
| X4      | The percentage of pregnant woman visiting clinics for the fourth time |

2.2. Data Analysis
In this section, the method of analysis data is divided into two parts. The first part explained the analysis of the simulated data and the second explained about analysis procedure of real data.
2.2.1. Analysis of Simulated Data

The analysis of the simulated data was carried out by the following steps:

1. Generating the $X$ variables
2. Generate the $Y$ variable
3. Performing score test on the response variable to identify overdispersion. The procedure of score test are as follows:

   Test hypotheses:
   
   $H_0: \omega = 0$ and $H_1: \omega > 0$

   Where $\omega$ is the probability of zero-valued response. The test statistics for score test are computed using:

   $S_\omega = \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2}$

   Where $n_0$ is the number of observations with zero-valued response, $n$ is the sample size, and $p_0 = e^{-\mu_0}$ with $\mu_0$ is the Poisson distribution parameter estimator given the $H_0$ condition. The null hypothesis is rejected when $S_\omega > \chi^2_{\alpha,1}$. Rejected null hypothesis implies overdispersion [5].

4. Performing data modelling using Poisson, ZIP, and ZINB regression.
5. Evaluating generated models

   The performances of different models were evaluated using the RMSE, bias, and AIC values. Models with lower RMSE, bias, and AIC values have better performance than those with higher values Rajitha and Sakthivel [6] and Sreelatha and Muniswamy [8]. Each value is computed as follows:

   - Root Mean Square Error (RMSE), based on the prediction results for each model, is calculated as:

     $RMSE = \sqrt{\frac{\sum_{i=1}^{n} (r_i)^2}{n}}$, dengan $r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(\hat{y})}}$

     Where $r_i$ is Pearson residual [3] value of the $i$-th observation.

   - Bias

     Bias computation can be done based on the predicted value and estimated parameters, or Absolute Relative Bias (ARB). Bias calculation follows equation below:

     $Bias = \frac{1}{n} \sum_{i=1}^{n} |\frac{y_i - \hat{y}_i}{\hat{y}_i}|$ dan $ARB = \left| \frac{\hat{\beta} - \beta}{\beta} \right|$

     Where $\hat{\beta}$ is the model parameter estimate and $\beta$ is the predetermined population parameter.

   - AIC, following the formula:

     $AIC = -2\ell + 2p$

     Where $\ell$ is the model’s log likelihood value and $p$ is the number of parameters used.

   The best model is determined by calculating the average value of RMSE, bias, and AIC for each model built from each combination of parameters and sample size.

2.2.2. Real Data

The real data analysis procedure is as follows:

1. Perform exploratory analysis on the response and explanatory variables
2. Conduct score test to detect overdispersion in the response variables
3. Detect multicollinearity in explanatory variables
4. Build models using Poisson regression, Zero-inflated Poisson regression (ZIP), and Zero-inflated Negative Binomial (ZINB) regression
5. Evaluate and interpret models
6. Perform cluster analysis
3. Results

3.1. Simulated Data

Table 3 shows the proportion of zero-valued observation in ZIP and ZINB distribution which was generated using every available parameter combination. It indicates that in the case of datasets with smaller mean value, the proportion of zero-valued observations are greater than the predetermined value. This was because in the case of Poisson and Binomial Negative distributions with a small middle value ($\mu = 2$), zero observations are generated in two stages: from the Poisson and Binomial Negative distribution and from predetermined $\omega$. In the case of datasets with greater the $\mu$ value, the proportion of zero-valued observations are the same as that which has been determined. This is because zero-valued observations are only generated from predefined values.

| n   | $\mu$ | 0.1  | 0.3  | 0.5  | 0.7  | 0.9  |
|-----|-------|------|------|------|------|------|
|     | ZIP   | ZINB | ZIP   | ZINB | ZIP   | ZINB | ZIP   | ZINB | ZIP   | ZINB |
| 2   | 0.23  | 0.35 | 0.40  | 0.50  | 0.57  | 0.64  | 0.75  | 0.79  | 0.92  | 0.93 |
| 5   | 0.10  | 0.16 | 0.30  | 0.35  | 0.50  | 0.53  | 0.70  | 0.72  | 0.90  | 0.91 |
| 15  | 0.10  | 0.11 | 0.30  | 0.31  | 0.50  | 0.51  | 0.70  | 0.70  | 0.90  | 0.90 |
| 30  | 0.10  | 0.10 | 0.30  | 0.30  | 0.50  | 0.50  | 0.70  | 0.70  | 0.90  | 0.90 |
| 100 | 0.23  | 0.34 | 0.40  | 0.48  | 0.57  | 0.63  | 0.74  | 0.78  | 0.92  | 0.93 |
| 5   | 0.20  | 0.17 | 0.30  | 0.35  | 0.50  | 0.54  | 0.70  | 0.72  | 0.90  | 0.91 |
| 15  | 0.20  | 0.12 | 0.30  | 0.32  | 0.50  | 0.51  | 0.70  | 0.71  | 0.90  | 0.90 |
| 30  | 0.20  | 0.11 | 0.30  | 0.31  | 0.50  | 0.51  | 0.70  | 0.70  | 0.90  | 0.90 |
| 200 | 0.23  | 0.33 | 0.40  | 0.48  | 0.57  | 0.63  | 0.74  | 0.78  | 0.92  | 0.93 |
| 5   | 0.10  | 0.17 | 0.30  | 0.35  | 0.50  | 0.54  | 0.70  | 0.72  | 0.90  | 0.91 |
| 15  | 0.10  | 0.12 | 0.30  | 0.32  | 0.50  | 0.51  | 0.70  | 0.71  | 0.90  | 0.90 |
| 30  | 0.10  | 0.11 | 0.30  | 0.31  | 0.50  | 0.51  | 0.70  | 0.70  | 0.90  | 0.90 |
| 500 | 0.10  | 0.11 | 0.30  | 0.31  | 0.50  | 0.51  | 0.70  | 0.70  | 0.90  | 0.90 |

Table 4 shows the average ratio of deviance value to the degree of freedom of the Poisson regression on each non-zero response variable (Y>0) following the ZIP and ZINB distribution.

| n   | $\mu$ | 0.1  | 0.3  | 0.5  | 0.7  | 0.9  |
|-----|-------|------|------|------|------|------|
|     | ZIP   | ZINB | ZIP   | ZINB | ZIP   | ZINB | ZIP   | ZINB | ZIP   | ZINB |
| 2   | 0.9   | 8.9  | 1.0   | 8.5  | 1.0   | 8.4  | 0.9   | 7.7  | 0.4   | 3.0  |
| 5   | 0.9   | 3.7  | 0.9   | 3.4  | 0.9   | 3.4  | 0.8   | 3.1  | 0.3   | 1.1  |
| 15  | 0.9   | 1.7  | 0.9   | 1.7  | 0.9   | 1.6  | 0.8   | 1.4  | 0.4   | 0.6  |
| 30  | 1.0   | 18.7 | 1.4   | 17.3 | 1.4   | 16.7 | 1.1   | 16.2 | 0.4   | 7.4  |
| 2   | 1.0   | 3.7  | 1.0   | 3.5  | 1.0   | 3.5  | 0.9   | 3.3  | 0.8   | 2.9  |
| 5   | 1.0   | 1.9  | 1.0   | 1.9  | 1.0   | 1.8  | 0.9   | 1.8  | 0.8   | 1.5  |
| 15  | 1.0   | 8.3  | 1.0   | 7.6  | 1.0   | 7.5  | 1.0   | 6.8  | 0.9   | 6.8  |
| 30  | 1.2   | 15.4 | 1.2   | 14.3 | 1.3   | 14.1 | 1.5   | 11.3 | 1.2   | 12.0 |
| 2   | 1.0   | 1.9  | 1.0   | 1.9  | 1.0   | 1.9  | 1.0   | 1.9  | 0.9   | 1.7  |
| 5   | 1.0   | 3.7  | 1.0   | 3.6  | 1.0   | 3.6  | 1.0   | 3.5  | 0.9   | 3.2  |
| 15  | 1.1   | 8.2  | 1.0   | 8.0  | 1.1   | 8.0  | 1.1   | 7.7  | 0.9   | 6.9  |
| 30  | 1.4   | 15.5 | 1.3   | 15.0 | 1.3   | 14.6 | 1.5   | 14.3 | 1.1   | 12.4 |
| 2   | 1.0   | 2.0  | 1.0   | 2.0  | 1.0   | 2.0  | 1.0   | 1.9  | 1.0   | 1.9  |
| 5   | 1.0   | 3.7  | 1.0   | 3.7  | 1.0   | 3.7  | 1.0   | 3.7  | 1.0   | 3.4  |
Table 4 indicates that the deviance ratio (d) values in ZIP distributions were within the 0.8 – 1 range, which means that there were no cases of overdispersion in non-zero count data. ZINB distributions, on the other hand, show very diverse values of d. Large values in ZINB distributions indicated that there were cases of overdispersion in non-zero count data. Table 5 shows the power test of score tests under each combination of n, µ, and ω. The test results showed that the greater the µ and ω were, the higher the statistical power of the test would be.

Table 5 The Power test of score tests under each combination of n, µ, and ω

| n  | µ   | 0.1  | 0.3  | 0.5  | 0.7  | 0.9  |
|----|-----|------|------|------|------|------|
|    | ZIP | ZINB | ZIP  | ZINB | ZIP  | ZINB | ZIP  | ZINB | ZIP  | ZINB |
| 2  | 0.15| 1.00 | 0.69 | 1.00 | 0.93 | 1.00 | 0.97 | 1.00 | 0.81 | 1.00 |
| 30 | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 15 | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 30 | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2  | 0.43| 0.86 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 200| 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 15 | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 30 | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2  | 0.74| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 5  | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 15 | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 30 | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2  | 0.99| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 5  | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 15 | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 30 | 1.00| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 6 shows the average RMSE of models built from datasets with observation n = 200. For datasets whose response variable followed the ZIP distribution, the RMSE values of their Poisson regression models are greater than those of the ZIP regression and ZINB regression models. Greater µ and ω also implied higher Poisson regression models’ RMSE. In contrast, the RMSE values of ZIP regression and ZINB regression models tended to stay constant as the value of µ and ω increased. RMSE values of ZINB regression models were lower than that of other models when µ was large. RMSE values of ZIP regression and ZINB regression models were not far apart.

On datasets whose response variables followed the ZINB distribution, the simulation results showed that the Poisson regression models’ RMSE were greater than those of the ZIP regression and ZINB regression models. The RMSE of the ZINB regression models tended to be stable despite the increase of µ and ω. Over all combinations of parameters, the ZINB regression models’ RMSE values were much smaller than those of the ZIP regression models.

Table 7 shows that greater µ and ω values tended to indicate greater AIC values of Poisson regression models. Conversely, greater µ and ω values tended to indicate smaller AIC values of the ZIP regression and ZINB regression models. As µ and ω increased, the AIC value of ZIP regression and ZINB regression models stayed smaller than those of the Poisson regression models. The AIC value of ZIP regression models tended to be smaller than that of ZINB regression models. However, when µ = 30, the AIC value of ZINB regression models were smaller than those of ZIP regression models at each increment in ω value.
Moreover, in the case of ZINB-distributed response variables, the Poisson regression models’ AIC values were greater than those of the ZIP regression and ZINB regression models. The ZINB regression models’ AIC values were smaller than other models. As $\mu$ increased, the difference between the AIC values of the ZIP regression models and those of the ZINB regression models was significant.

**Table 6** Average RMSE of models built from datasets with ZIP-distributed and ZINB-distributed response variable

| $\mu$ | $\omega$ | Response Variable Distribution | ZIP | ZINB | ZIP | ZINB |
|-------|---------|--------------------------------|-----|------|-----|------|
| 0.1   | 0.1     | Poisson                        | 1.09| 0.99 | 0.99| 1.45| 1.15| 1.00|
| 0.3   | 0.5     | Poisson                        | 1.25| 0.99 | 0.99| 1.57| 1.12| 1.00|
| 2     | 0.5     | Poisson                        | 1.40| 1.00 | 0.99| 1.69| 1.10| 1.00|
| 0.7   | 0.9     | Poisson                        | 1.53| 0.99 | 0.99| 1.78| 1.08| 0.99|
| 0.9   | 1.5     | Poisson                        | 1.60| 0.98 | 0.97| 1.85| 1.04| 0.98|
| 0.5   | 1.24    | Poisson                        | 1.24| 1.00 | 0.99| 2.08| 1.47| 1.00|
| 0.3   | 1.61    | Poisson                        | 1.61| 1.00 | 0.99| 2.32| 1.31| 1.00|
| 5     | 0.5     | Poisson                        | 1.92| 1.00 | 0.99| 2.54| 1.24| 1.00|
| 0.7   | 2.16    | Poisson                        | 2.16| 1.00 | 0.99| 2.74| 1.18| 0.99|
| 0.9   | 2.41    | Poisson                        | 2.41| 0.99 | 0.98| 2.85| 1.13| 0.99|
| 1.0   | 1.56    | Poisson                        | 1.56| 1.01 | 1.00| 3.17| 1.81| 1.00|
| 0.3   | 2.28    | Poisson                        | 2.28| 1.00 | 1.00| 3.64| 1.46| 1.00|
| 15    | 0.5     | Poisson                        | 2.82| 1.00 | 1.00| 4.01| 1.33| 1.00|
| 0.7   | 3.24    | Poisson                        | 3.24| 1.00 | 1.00| 4.32| 1.25| 0.99|
| 0.9   | 3.78    | Poisson                        | 3.78| 1.00 | 0.99| 4.51| 1.18| 0.99|
| 0.5   | 2.01    | Poisson                        | 2.01| 1.05 | 1.00| 4.43| 2.00| 1.00|
| 0.3   | 3.06    | Poisson                        | 3.06| 1.02 | 1.00| 5.11| 1.53| 1.00|
| 30    | 0.5     | Poisson                        | 3.85| 1.01 | 1.00| 5.61| 1.36| 1.00|
| 0.7   | 4.42    | Poisson                        | 4.42| 1.01 | 1.00| 6.09| 1.27| 1.00|
| 0.9   | 5.32    | Poisson                        | 5.32| 1.00 | 1.00| 6.34| 1.20| 0.99|

**Table 7** Average AIC of models built from datasets with ZIP-distributed and ZINB-distributed response variable

| $\mu$ | $\omega$ | Response Variable Distribution | ZIP | ZINB | ZIP | ZINB |
|-------|---------|--------------------------------|-----|------|-----|------|
| 0.1   | 0.5     | Poisson                        | 682.43| 678.71| 679.57| 782.38| 737.33| 719.17|
| 0.3   | 0.7     | Poisson                        | 665.68| 629.22| 631.30| 741.64| 653.33| 638.75|
| 2     | 0.7     | Poisson                        | 608.07| 533.23| 534.18| 661.98| 534.05| 524.24|
| 0.9   | 0.9     | Poisson                        | 486.68| 387.13| 387.83| 513.68| 375.21| 370.44|
| 0.5   | 1.0     | Poisson                        | 242.44| 171.38| 171.57| 261.48| 164.21| 163.97|
| 0.3   | 1.0     | Poisson                        | 1012.60| 933.93| 936.65| 1432.83| 1244.78| 1071.50|
| 5     | 0.7     | Poisson                        | 1161.96| 867.46| 871.26| 1502.94| 1079.13| 949.40|
| 0.5   | 0.9     | Poisson                        | 1192.87| 724.65| 725.87| 1437.61| 863.56| 771.22|
| 0.3   | 0.7     | Poisson                        | 1032.76| 513.29| 513.95| 1196.98| 588.52| 535.85|
| 15    | 0.9     | Poisson                        | 598.91| 224.43| 224.33| 636.70| 240.98| 227.45|
| 0.1   | 1.0     | Poisson                        | 1514.05| 1125.83| 1128.00| 2786.20| 2192.58| 1390.38|
| 0.3   | 1.0     | Poisson                        | 2136.01| 1017.45| 1020.98| 3242.23| 1812.86| 1208.87|
| 7     | 0.5     | Poisson                        | 2462.69| 834.31| 835.07| 3310.64| 1400.03| 967.03|
| 0.9   | 1.0     | Poisson                        | 2260.56| 581.35| 581.48| 2845.64| 898.26| 655.17|
| 0.1   | 1.0     | Poisson                        | 1481.81| 246.76| 246.61| 1586.96| 341.14| 271.27|
| 30    | 0.3     | Poisson                        | 2179.35| 1301.73| 1293.85| 4863.36| 3553.64| 1602.15|
| 0.5   | 0.5     | Poisson                        | 3558.49| 1148.27| 1144.33| 5997.59| 2820.39| 1376.64|
| 0.7   | 0.5     | Poisson                        | 4383.65| 930.07| 927.20| 6230.23| 2114.36| 1086.39|
0.7  4110.80  646.84  642.59  5520.40  1319.37  729.02
0.9  2904.61  264.95  266.19  3136.33  463.70  295.90

The ARB of models built from datasets with ZIP-distributed response variable are showed by Table 8. The estimated parameter value of ZIP regression and ZINB regression models were determined based on the count data model. ARB was determined by comparing the estimated value of the model parameters to the predetermined parameter values.

Average $\beta_1$ ARB of Poisson regression models tended to enlarge as $\mu$ and $\omega$ increased. When $\mu = 2$, $\beta_1$ ARB of Poisson regression models were lower than the $\beta_1$ ARB of ZIP and ZINB regression models. However, as $\mu$ increased, $\beta_1$ ARB of Poisson regression models were greater than the $\beta_1$ ARB of ZIP and ZINB regression models. This means that in the case of overdispersion, Poisson did not perform well in estimating the $\beta_1$. Meanwhile, when the values of $\mu$ and $\omega$ were large, average $\beta_0$ ARB of Poisson regression models were lower than those of ZIP and ZINB regression models. Conversely, when the values of $\mu$ and $\omega$ were small, average $\beta_0$ ARB of Poisson regression models were higher than those of ZIP and ZINB regression models.

| $\mu$ | $\omega$ | $\beta_0$ | $\beta_1$ |
|-------|----------|-----------|-----------|
| Poisson | ZIP | ZINB | Poisson | ZIP | ZINB |
| 0.1 | 0.37 | 0.36 | 0.32 | 0.69 | 0.91 | 0.76 |
| 0.3 | 0.45 | 0.37 | 0.29 | 0.43 | 0.93 | 0.67 |
| 2 | 0.5 | 0.79 | 0.40 | 0.35 | 0.53 | 1.01 | 0.84 |
| 0.7 | 1.16 | 0.40 | 0.35 | 0.54 | 1.00 | 0.83 |
| 0.9 | 2.29 | 0.59 | 0.55 | 0.96 | 1.41 | 1.22 |
| 0.1 | 0.54 | 0.52 | 0.52 | 1.20 | 0.98 | 0.99 |
| 0.3 | 0.46 | 0.52 | 0.52 | 1.48 | 0.98 | 0.99 |
| 5 | 0.5 | 0.16 | 0.51 | 0.51 | 1.43 | 0.96 | 0.97 |
| 0.7 | 0.29 | 0.50 | 0.50 | 1.55 | 0.98 | 0.98 |
| 0.9 | 1.33 | 0.55 | 0.55 | 1.40 | 1.01 | 1.01 |
| 0.1 | 1.35 | 1.33 | 1.33 | 1.05 | 0.97 | 0.97 |
| 0.3 | 1.31 | 1.37 | 1.37 | 1.18 | 0.99 | 0.99 |
| 15 | 0.5 | 0.97 | 1.35 | 1.35 | 1.16 | 0.98 | 0.98 |
| 0.7 | 0.51 | 1.30 | 1.30 | 1.18 | 0.96 | 0.96 |
| 0.9 | 0.45 | 1.42 | 1.42 | 1.14 | 0.98 | 0.98 |
| 0.1 | 1.94 | 1.92 | 1.92 | 1.03 | 0.97 | 0.97 |
| 0.3 | 1.89 | 1.95 | 1.95 | 1.12 | 0.98 | 0.98 |
| 30 | 0.5 | 1.56 | 1.95 | 1.95 | 1.10 | 0.97 | 0.97 |
| 0.7 | 1.08 | 1.87 | 1.87 | 1.12 | 0.95 | 0.96 |
| 0.9 | 0.17 | 2.04 | 2.04 | 1.09 | 0.97 | 0.97 |

Based on Table 9, the simulation results show that the average ARB of $\beta_0$ and $\beta_1$ in Poisson regression models are greater than in ZIP and ZINB regression models. Out of all three, ZINB displayed the lowest ARB value. Table 10 shows that the absolute bias of ZINB and ZIP regression models were much lower than those of the Poisson regression models. The absolute bias of Poisson regression models tended to increase as the values of $\mu$ and $\omega$ increased. In datasets with ZIP-distributed response variable, the absolute bias difference between ZIP and ZINB regression models was not significant. However, in datasets with ZINB-distributed response variables, the absolute bias difference between ZIP and ZINB regression models was considerably large.
Table 9 Average ARB of models built from datasets with ZINB-distributed response variable

| µ   | ω  |          |       |          |       |          |       |
|-----|----|----------|-------|----------|-------|----------|-------|
|     |    | ARB      |       | ARB      |       | ARB      |       |
|     |    | Poisson  | ZIP   | ZINB     | Poisson| ZIP      | ZINB  |
| 0.1 | 0.02 | 0.11     | 0.02  | 0.29     | 0.17  | 0.00     |       |
| 0.3 | 0.03 | 0.14     | 0.03  | 0.75     | 0.07  | 0.09     |       |
| 0.5 | 0.34 | 0.11     | 0.01  | 0.75     | 0.14  | 0.01     |       |
| 0.7 | 0.81 | 0.08     | 0.07  | 0.77     | 0.18  | 0.07     |       |
| 0.9 | 1.90 | 0.16     | 0.02  | 0.56     | 0.14  | 0.32     |       |
| 1.0 | 0.02 | 0.13     | 0.04  | 0.22     | 0.13  | 0.08     |       |
| 0.3 | 0.08 | 0.12     | 0.01  | 0.47     | 0.10  | 0.03     |       |
| 0.5 | 0.41 | 0.13     | 0.02  | 0.42     | 0.13  | 0.06     |       |
| 0.7 | 0.82 | 0.11     | 0.00  | 0.53     | 0.10  | 0.02     |       |
| 0.9 | 1.94 | 0.10     | 0.03  | 0.37     | 0.12  | 0.02     |       |
| 1.0 | 0.01 | 0.06     | 0.02  | 0.07     | 0.03  | 0.01     |       |
| 0.3 | 0.09 | 0.08     | 0.02  | 0.18     | 0.05  | 0.02     |       |
| 0.5 | 0.47 | 0.05     | 0.01  | 0.12     | 0.02  | 0.00     |       |
| 0.7 | 0.89 | 0.05     | 0.00  | 0.17     | 0.03  | 0.01     |       |
| 0.9 | 1.99 | 0.08     | 0.00  | 0.11     | 0.07  | 0.02     |       |
| 1.0 | 0.02 | 0.03     | 0.00  | 0.04     | 0.01  | 0.00     |       |
| 0.3 | 0.14 | 0.04     | 0.01  | 0.10     | 0.01  | 0.00     |       |
| 0.5 | 0.49 | 0.01     | 0.01  | 0.08     | 0.01  | 0.00     |       |
| 0.7 | 0.93 | 0.02     | 0.00  | 0.10     | 0.01  | 0.00     |       |
| 0.9 | 2.19 | 0.00     | 0.06  | 0.01     | 0.01  | 0.01     |       |

Table 10 Absolute bias of models built from datasets with ZIP-distributed and ZINB-distributed response variable

| µ   | ω  |          |       |          |       |          |       |
|-----|----|----------|-------|----------|-------|----------|-------|
|     |    | Response Variable Distribution |       |          |       |          |       |
|     |    | ZIP      | ZINB  | Poisson  | ZIP   | ZINB     |       |
| 0.1 | 0.88 | 0.81     | 0.80  | 1.12     | 0.89  | 0.77     |       |
| 0.3 | 1.04 | 0.83     | 0.83  | 1.20     | 0.86  | 0.77     |       |
| 0.5 | 1.13 | 0.81     | 0.80  | 1.24     | 0.81  | 0.73     |       |
| 0.7 | 1.13 | 0.74     | 0.73  | 1.17     | 0.71  | 0.65     |       |
| 0.9 | 0.79 | 0.48     | 0.48  | 0.81     | 0.46  | 0.43     |       |
| 1.0 | 0.98 | 0.79     | 0.78  | 1.60     | 1.13  | 0.77     |       |
| 0.3 | 1.37 | 0.85     | 0.84  | 1.81     | 1.02  | 0.78     |       |
| 0.5 | 1.69 | 0.88     | 0.88  | 1.93     | 0.94  | 0.76     |       |
| 0.7 | 1.77 | 0.82     | 0.81  | 1.87     | 0.82  | 0.69     |       |
| 0.9 | 1.33 | 0.55     | 0.54  | 1.33     | 0.53  | 0.47     |       |
| 1.0 | 1.18 | 0.77     | 0.75  | 2.39     | 1.40  | 0.77     |       |
| 0.3 | 1.96 | 0.87     | 0.86  | 2.78     | 1.15  | 0.78     |       |
| 0.5 | 2.63 | 0.94     | 0.93  | 3.03     | 1.03  | 0.77     |       |
| 0.7 | 2.80 | 0.87     | 0.87  | 2.96     | 0.88  | 0.70     |       |
| 0.9 | 2.19 | 0.58     | 0.58  | 2.12     | 0.57  | 0.48     |       |
| 0.1 | 1.46 | 0.77     | 0.74  | 3.25     | 1.55  | 0.77     |       |
| 0.3 | 2.67 | 0.89     | 0.87  | 3.82     | 1.20  | 0.79     |       |
| 0.5 | 3.67 | 0.96     | 0.95  | 4.16     | 1.06  | 0.78     |       |
| 0.7 | 3.88 | 0.89     | 0.88  | 4.10     | 0.90  | 0.71     |       |
| 0.9 | 3.12 | 0.59     | 0.59  | 2.95     | 0.58  | 0.49     |       |
Generally speaking, the results of the simulation study are presented in Table 11 which shows a summary of the performance of each model based on the distribution of the data. When the response variable followed ZIP distribution, both ZIP and ZINB regression models showed good performance in modeling the data. Meanwhile, when the response variable followed ZINB distribution, the ZINB regression model was more appropriate to be used in modeling the data.

Table 11 Summary of the performance of each model based on the distribution of the data

| Simulated model | Model Prediction | ZIP Regression | ZINB Regression |
|-----------------|-----------------|----------------|----------------|
| ZIP             | Good            | Good           |                |
| ZINB            | Bad             | Good           |                |

3.2. Real Data Analysis

Figure 1 shows the histogram of the number of maternal mortality in East Java in 2018. It can be seen that the distribution of the number of maternal mortality was skewed to the right. Based on the boxplot in Figure 2, it appears that there are 2 observations identified as outliers.

The average number of maternal mortality in 2018 was 3.26. Jember Regency had the highest number of maternal mortality (12) in that year, followed by Banyuwangi and Bondowoso, both with 8 cases of maternal mortality. Zero-valued observations indicated no maternal mortality in the area. The proportion of zero-valued observations was 0.20.

3.2.1. Data Modeling

Overdispersion was detected using the score test. The results showed that p-value < α = 0.05, rejecting the null hypothesis. This means that the response variable was overdispersed.

Table 12 shows the goodness-of-fit metrics score of each model. It can be seen that the AIC of the ZIP regression model was lower than that of other models, but the difference between ZIP and ZINB regression models was not statistically significant. The RMSE and absolute bias value of ZINB regression model, however, were relatively lower than those of other models. Therefore, in this case, ZINB regression was chosen to model the data.

Table 13 shows the results of the ZINB regression modeling. ZINB regression carried out modeling in two stages. The first stage is modeling the count data, producing a count data model. According to the test results, explanatory variables X3 and X4 of the count data model influenced the number of
maternal mortality at a significance level of $\alpha = 0.1$. In the zero-inflation model, on the other hand, maternal mortality number was not influenced by any of the explanatory variables.

**Table 12** Goodness-of-fit score of each model

|       | POISSON | ZIP       | ZINB       |
|-------|---------|-----------|------------|
| AIC   | 177.474 | 168.982   | 169.8489   |
| MSE   | 1.609   | 1.181     | 0.997      |
| RMSE  | 1.268   | 1.087     | 0.999      |
| ABS.Bias | 1.026   | 0.875     | 0.805      |

**Table 13** Results of ZINB regression model parameter test with four explanatory variables

| Variables | Parameter estimation | se | p-value |
|-----------|----------------------|----|---------|
| Intersep  | -6.649               | 4.391 | 0.129 |
| X1        | 0.016                | 0.049 | 0.746 |
| X2        | 0.023                | 0.034 | 0.500 |
| X3        | 0.115                | 0.027 | 0.094 |
| X4        | -0.77                | 0.068 | 0.004 |

3.2.2. *Area Clustering*

Further analysis towards maternal mortality number was done by employing cluster analysis. Cities/regencies within the same cluster are considered to have similar characteristics. Complete-linkage hierarchical clustering method was employed in this study. It was chosen due to the relatively small number of observations (38). Distance was measured using Euclidean distance. Figure 3 shows the dendogram of clustering results. The red-colored line acted as the cutoff line, resulting in three optimal clusters.

![Dendogram of clustering results](image)

**Figure 3** Dendogram of clustering results

3.2.3. *Cluster Interpretation*

Generated clusters were interpreted by looking at the centroid of each cluster. Table 14 shows the information regarding generated clusters. According to the characteristics of generated clusters, researchers categorized the areas (cities/regencies) into 3 conditions: good, bad, and moderate. Areas in cluster 1 were of the bad condition, areas in cluster 2 were of the moderate condition, and areas in cluster 3 were of the good condition.

**Table 14** Clustering results

| Cluster | Centroid | Number of Observation |
|---------|----------|-----------------------|
|         | Estimated | AKIH FE1 FE3 K1 K4   |


Cluster 1 | 6.00 | 98.91 | 87.32 | 99.95 | 85.55 | 7
Cluster 2 | 2.83 | 98.13 | 93.61 | 99.02 | 94.39 | 23
Cluster 3 | 2.37 | 89.01 | 80.82 | 93.42 | 84.43 | 8

Figure 4 shows the map of cities/regencies in East Java based on the categories generated in the clustering stage. Red-colored areas were areas with high number of maternal mortality, thus requiring a high level of attention from the government. Restraining the number of maternal mortality can be done by improving health facilities, providing maternal needs (i.e. FE1 and FE3 tablets), and encouraging expecting mothers to commit on Q1 and Q4 visits. Yellow-colored areas were areas with relatively lower maternal mortality number compared to other areas, but also with lower number of FE1 and FE3 tablets recipient and Q4 visits. The government, therefore, should improve overall health facilities in these areas, starting by raising the distribution of FE1 and FE3 tablets. Finally, blue-colored areas were areas in good condition.

![Figure 4 Category-based map of East Java](image)

4. Conclusion

Our modeling study, which was conducted using simulated data, showed that Poisson regression, ZIP regression, and ZINB regression have similar performance in modeling non-overdispersed data. In cases without overdispersion, it is recommended to use Poisson regression. In the case of datasets with overdispersion, ZINB and ZIP regressions performed better than Poisson regression. The modeling results showed that when the response variable followed ZINB distribution (where non-zero count data are overdispersed), ZINB regression performed better than ZIP regression. On the other hand, when the response variable followed ZIP distribution (where non-zero count data are not overdispersed), ZINB and ZIP regression were equally good in modeling the data. However, when the µ value is high, ZINB regression was preferred in modeling the data.

Maternal mortality number was modelled using ZINB regression. Significant exploratory variables in the count data model were the percentage of pregnant woman visiting clinics for the first time (X3) and the percentage of pregnant woman visiting clinics for the fourth time (X4). No explanatory variable was significant in the zero-inflation model. Cluster analysis of maternal mortality number yielded in three categories. Areas were sorted into one of the three (good, moderate, and bad) categories based on the estimated maternal mortality and the explanatory variables used in the model.

References

[1] Agresti A 2014 Foundation of Linear and Generalized Linear Models. New Jersey (US): John Wileyand Sons, Inc.
[2] Liu X, Winter B, Tang L, Zhang B, Zhang Z and Zhang H 2017 Simulating comparisons of computing algorithms fitting zero-inflated Poisson models for zero abundant counts. *Journal of Statistical Computation and Simulations*. doi: 10.1080/00949655.2017.1327590.

[3] McCullagh P and Nelder JA 1989 *Generalized Linear Models, Second Edition*. London (GB): Chapman & Hall.

[4] Muniswamy B and Molla DT 2012 Power of tests for overdispersion parameter in Negative Binomial regression model. *Journal of Mathematics*. 4(1):29-36.

[5] Rahayu LP 2014 Kajian Overdispersi pada Regresi Poisson dan Zero-Inflated Poisson untuk Beberapa Karakteristik Data [thesis]. Bogor (ID): IPB University.

[6] Rajitha CS, Sakthivel KM. 2019 Model selection for count data with excess number of zero counts. *American Journal of Applied Mathematics and Statistics*. 7(1):43-51.

[7] Sepriawandani 2019 Kajian Perbandingan Regresi Binomial Negatif dan Zero-inflated Poisson dalam Mengatasi Overdispersi [thesis]. Bogor (ID): Institut Pertanian Bogor.

[8] Sreelatha CH and Muniswamy B. 2018. A Wald Test for Overdispersion in Zero Inflated Poisson Regression Model. *International Journal of Mathematical*. 9(6): 201-212

[9] Yang S, Puggioni G, Harlow LL and Redding CA 2017 A comparison of different methods of zero inflated data analysis and an application in health surveys. *Journal of Modern Applied Statistical Methods*. 16(1):518–543.

[10] Yesilova A, Kaya Y, Kaki B, Kasap I. 2010. Analysis of Plant Protection Studies with Excess Zeros Using Zero-Inflated and Negative Binomial Hurdle Models. *Gazi University Journal of Science*. 23(2):131-136.