Overdispersion study of poisson and zero-inflated poisson regression for some characteristics of the data on lamda, n, p

Lili Puspita Rahayu a,1,*, Kusman Sadik b,2, Indahwati b,3

a Ministry of Education and Culture of The Republic of Indonesia, Indonesia
b Department of Statistics, Bogor Agriculture University, Indonesia
1 lilipuspitarahayu@gmail.com *, 2 kusmansadik@gmail.com; 3 indah.stk@gmail.com
* corresponding author

I. Introduction

Poisson distribution is one of discrete distribution that is often used in modeling of rare events. The data obtained in the form of counts with non-negative integers. One of analysis that is used in modeling count data is Poisson regression. Deviation of assumption that often occurs in the Poisson regression is overdispersion. Cause of overdispersion is an excess zero probability on the response variable. Solving model that be used to overcome of overdispersion is zero-inflated Poisson (ZIP) regression. The research aimed to develop a study of overdispersion for Poisson and ZIP regression on some characteristics of the data. Overdispersion on some characteristics of the data that were studied in this research are simulated by combining the parameter of Poisson distribution (λ), zero probability (p), and sample size (n) on the response variable then comparing the Poisson and ZIP regression models. Overdispersion study on data simulation showed that the larger λ, n, and p, the better is the model of ZIP than Poisson regression. The results of this simulation are also strengthened by the exploration of Pearson residual in Poisson and ZIP regression.
In addition, estimate of the parameter on ZIP regression model can be interpretation easily, and can explain the reason of the mean smaller in the response variable.

Research that has been done before, starting the ZIP regression model developed as a solution overdispersion handling of Poisson regression models using simulation studies that Xie et al. [5] using a type II error in the simulation with combining the parameter of Poisson distribution, zero probability for sample size on the response variable. Numna [6] developed a Wald test for comparison of Poisson and ZIP regression models. Wald test development performed simulations with determination of zero probability on the response variable based on the value parameter of the Poisson distribution.

Development of the research that has been done previously, the researcher wanted to study develops the overdispersion on some characteristics of the data. Overdispersion which will be examined in this study are simulated by combining the value of the parameter of the Poisson distribution, zero probability, and sample size on the response variable. Furthermore, comparing Poisson and ZIP regression models based on exploration of the response variable, and the evaluation of the prediction models. Any simulation on characteristics data are expected to determine the cause of overdispersion on response variable.

II. Poission Regression

Hardin and Hilbe [7] stated that the Poisson regression model provides a standard framework for analysis of data count. Poisson regression is a form of general linear model. Let $y_i, i=1,2,...,n$ represents count of those rare occasions in the period with value of the parameter Poisson distribution lambda ($\lambda_i$), $y_i$ is a Poisson random variable that spread by the mass function probability of the following

$$f(y_i; \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, ...$$

with assuming of Poisson regression is

$$E(Y_i) = Var(Y_i) = \lambda_i$$

If Poisson regression is used for the condition overdispersion, then the result is not exact because the value of mean and variance Poisson regression contain dispersion

$$Var(Y) = \tau E(Y) = \tau \lambda$$

with $\tau$ is ratio of dispersion. When there was overdispersion on Poisson regression then value of $\tau$ is more than one and constant. Dispersion is a measure of variant of a group of data to the mean data. Small dispersion values showed a homogeneous variety in the data, while the big dispersion values indicate heterogeneity in data. Method to estimate parameters of Poisson regression coefficients are maximum likelihood method. Suppose $X$ is explanatory variable that size matrix $n \times (p+1)$. Random variables $y_i$ and $i$ is a row vector of $X$, will be linked with the log link function.

$$\ln(\lambda_i) = \eta_i = Xb$$

$$\lambda_i = \exp(Xb)$$

The model in (5) is a Poisson regression model with parameter $b$ is the coefficient estimate.

III. Zero Inflated Poisson Regression

Jansakul and Hinde [1] state that if the $Y_i$ are independent random variables that have a ZIP distribution, then the value of zero is assumed to arise from the same two steps. The first step
events on probability that only produces zero observations denoted by $p$. The second step events on the probability that result of data count spread Poisson with parameter $\lambda$ is denoted by $(1-p)$. In general, the zero value of the first step is called structural zeros, and the zero value of the second step is called sampling zeros. Variables $Y_i \sim ZIP(\lambda, p_i)$ have assumptions on the ZIP regression

$$E(Y_i) = (1 - p_i)\lambda_i = \mu_i$$

(6)

and

$$Var(Y_i) = \mu_i + \left(\frac{p_i}{1 - p_i}\right)\mu_i^2$$

(7)

variables $Y_i$ has overdispersion $p_i > 0$ if, then overdispersion will reduce to Poisson models when $p_i = 0$. Value $p_i > 0$ explains that there is an increase value of zero on the response variable.

Method to estimate parameters of ZIP regression coefficients is maximum likelihood method. Log-likelihood function for observations $y_1,...,y_n$ on ZIP regression model is used to simplify the calculation to get parameter estimate coefficient. Maximizing of log-likelihood function will give the same result as maximizing the likelihood function. ZIP regression models divided into two components, namely discrete data models for $\lambda$ and zero-inflation models for $p$. If the explanatory variables used in ln and logit model on the same value, then the ZIP regression model is

$$\ln(\lambda_i) = Xb$$

$$\ln\left(\frac{p_i}{1 - p_i}\right) = Xg$$

(8)

with $X$ is the matrix of explanatory variables, while $b$ and $g$ is the vector of parameter estimate of ZIP regression coefficients, each sized $(q+1)x1$ and $(r+1)x1$ in equation (8). Explanatory variables used in the model ln be the same or different from the explanatory variables used in the logit model. Maximum likelihood estimation for $b$ and $g$ are obtained by using the expectation maximization (EM) algorithm which provides a simple way, so that it can be applied to standard software to match the general linear model.

IV. Design of Simulation

The data used in this study is the simulation data. Simulation data was generated based on the characteristics of the data. Characteristics of the data in the form of lambda ($\lambda$) starting from $\lambda = 0.6, 0.8, 1, 6, 8, 10, \text{ and } 20$, the zero probability ($p$) are $p=0.1, 0.3, 0.5, \text{ and } 0.7$, and sample size ($n$) are $n=100, 300, \text{ and } 500$. The data generated are useful to obtain parameter estimators of Poisson and ZIP regression. The coefficient of regression parameters were determined are $\theta=3$, and $\theta=0.01$. Variables were determined to make the Poisson and ZIP regression models are explanatory variable ($X$), the response variable ($Y$).

Variable X consists of variable $X$ which is a normal random variable spreads ($\mu, 1$). Variable $X$ is assumed as a fixed variable. Generating variables $X$ and $Y$ on simulation study carried out by stages, namely:

1. Generating variable $Y$ based on the value of $\lambda$, $n$, $p$ have been determined.
2. Generating variable $X$ is with the first loop, are:
   - Separate variables $Y$ become $Y$ zero and $Y$ not zero.
   - Transform variable $X$ with formula $x_i = (\ln(y_i) - \theta)/\theta_i$, which is $y_i$ from variable $Y$ not zero.
• Initialize the result of transformation variable X as X not zero.

3. Generating variable X on variable Y not zero and Y zero with the second loop are

• If the variable Y is zero, then the variable x_i is gotten by sampling with replacement on variable X not zero.

• If the variable Y is not zero, then the variables x_i is generated from Normal distribution with mean of transformation result from 2(ii) and variance is 1 with sample size \( n=1 \).

Simulation data on the variables X and Y generated by the software program R ver.2.15.2 and will be repeated \( r=500 \) replications. There are 84 simulation conditions used in this study. The accuracy of parameter estimators in the Poisson and ZIP regression models can be seen from the relative absolute bias (RAB) and relative root mean square error (RRMSE) [8]. Furthermore, the accuracy of the estimate \( y \) in Poisson and ZIP regression models can be seen from Pearson residual (PR) and sum of absolute Pearson residual (SAPR) [9]. The equation of value RAB, RRMSE, PR, and SAPR are defined respectively in (9), (10), (11), and (12).

\[
RAB = \sum_{i=1}^{500} \left| \frac{\hat{\theta}_i - \theta}{\theta} \right| \times 100\% 
\]

(9)

\[
RRMSE = \sqrt{\frac{\sum_{i=1}^{500} (\hat{y}_i - \theta)^2}{n}} \times 100\%
\]

(10)

\[
PR_i = \frac{\sum_{i=1}^{500} (y_{ij} - \hat{y}_{ij})}{\text{Var}(Y)} \times 100\%
\]

(11)

\[
SAPR = \sum_{i=1}^{500} |PR_i|
\]

(12)

where \( r \) is the number of simulation replications, \( \hat{\theta}_i \) is the parameter estimators of Poisson and ZIP regression \( i^{th} \), \( \theta \) is the actual parameter. Then, \( y_{ij} \) is the response variable on repeat to-\( i \) and observations to-\( j \), and \( \hat{y}_{ij} \) is the estimate \( y \) of the Poisson and ZIP regression model on repeat to-\( i \) and observations to-\( j \), and \( \text{Var}(Y) \) is the estimate variance of Poisson and ZIP regression. Then, the smaller values of RAB, RRMSE, PR, and SAPR on regression model can be said to be getting better.

V. Results

Simulation study consisted of 84 cases of simulation which is a characteristic of data from the combination of \( \lambda, n, \) and \( p \). Simulations were performed to evaluate the results of estimating the parameters of the Poisson and ZIP regression using percentage of ARB, RRMSE, and average of SAPR. The value obtained from the simulation was repeated 500 times. The evaluation results of the simulation data to be clarified with the results of exploration and testing of the variable Y.

A. Exploration and testing at the variable Y

Characteristics of data simulation against \( \lambda, n, \) and \( p \) is tested indicates that the rise of the value of \( p \) effect on \( \lambda \). Value of \( \lambda = 0.6 \) with \( p \) that is tested 0.3, then the variables Y produces range \( p \) of 0.3 to 0.5. This is because the \( \lambda \) that has small value still has \( p \) from Poisson distribution relatively large. This statement can be explained by the cumulative Poisson distribution table. Value \( \lambda \) is tested to 0.6, 0.8, 1, 6, and 8 still have zero probability of a Poisson distribution, while for the other \( \lambda \) is tested, namely 10 and 20 already have not zero probability of a Poisson distribution.

Exploration variable Y to \( \lambda, n, \) and \( p \) is tested indication excess zero probability, so that the necessary tests on variables Y. The test is in the form of scores test and chi-square test were able to generalize the general conclusions on the results of exploration at variables Y. The condition of excess zero probability at the variable Y due to overdispersion. Flynn and Francis [10] state that when the score test results value of zero excess on a variable, then chances are the variables do not spread Poisson distribution, but has ZIP distribution.

The score test results with \( \alpha \) of 0.05 on simulation data at variable Y against a combination of \( \lambda, n, \) \( p \) are shown in Table 1. Scores test indicate that the larger \( \lambda, n, \) and \( p \) is tested, then the larger percentage of excess zero at the variable Y.

Lili Puspita Rahayu et. al. (Overdispersion study of poisson and zero-inflated poisson regression for ...)
Furthermore, the results of chi-square test with α of 0.05 for the Poisson and ZIP distribution against the combination of λ, n, p are shown in Table 2. The chi-square test for Poisson distribution shows that the larger λ, n, and p is tested, then percentage Poisson distribution will be smaller at the variable Y. The results of the score test and the chi-square test for Poisson distribution is inversely proportional to the larger λ, n, and p is tested. Chi-square test for ZIP distribution indicates that ZIP regression able to overcome overdispersion due to excess p at the variable Y. This condition is indicated by the larger value of λ, then the percentage Poisson distribution reaches 0%, while the percentage of ZIP distribution in the range of 60% to 80%.

Table 1. Percentage of score test on combination of λ, n, p (%)

<table>
<thead>
<tr>
<th>n</th>
<th>λ</th>
<th>p=0.1</th>
<th>p=0.3</th>
<th>p=0.5</th>
<th>p=0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>6.5</td>
<td>23.4</td>
<td>45.6</td>
<td>88.9</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>9.0</td>
<td>39.2</td>
<td>66.8</td>
<td>88.9</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10.0</td>
<td>55.0</td>
<td>86.2</td>
<td>92.8</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>6</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>8</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>10</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>20</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Furthermore, the results of chi-square test with α of 0.05 for the Poisson and ZIP distribution against the combination of λ, n, p are shown in Table 2. The chi-square test for Poisson distribution shows that the larger λ, n, and p is tested, then percentage Poisson distribution will be smaller at the variable Y. The results of the score test and the chi-square test for Poisson distribution is inversely proportional to the larger λ, n, and p is tested. Chi-square test for ZIP distribution indicates that ZIP regression able to overcome overdispersion due to excess p at the variable Y. This condition is indicated by the larger value of λ, then the percentage Poisson distribution reaches 0%, while the percentage of ZIP distribution in the range of 60% to 80%.

Table 2. Percentage of chi-square test on combination of λ, n, p (%)

<table>
<thead>
<tr>
<th>n</th>
<th>λ</th>
<th>p=0.1</th>
<th>p=0.3</th>
<th>p=0.5</th>
<th>p=0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>91.0</td>
<td>83.6</td>
<td>75.0</td>
<td>85.4</td>
<td>54.8</td>
</tr>
<tr>
<td>0.8</td>
<td>89.2</td>
<td>80.2</td>
<td>62.6</td>
<td>82.0</td>
<td>32.0</td>
</tr>
<tr>
<td>1</td>
<td>88.2</td>
<td>82.2</td>
<td>51.0</td>
<td>86.0</td>
<td>14.2</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
<td>0.0</td>
<td>84.4</td>
<td>0.0</td>
<td>83.4</td>
</tr>
<tr>
<td>8</td>
<td>0.0</td>
<td>81.6</td>
<td>0.0</td>
<td>79.6</td>
<td>0.0</td>
</tr>
<tr>
<td>10</td>
<td>0.0</td>
<td>79.4</td>
<td>0.0</td>
<td>78.6</td>
<td>0.0</td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
<td>72.6</td>
<td>0.0</td>
<td>71.2</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Furthermore, the results of chi-square test with α of 0.05 for the Poisson and ZIP distribution against the combination of λ, n, p are shown in Table 2. The chi-square test for Poisson distribution shows that the larger λ, n, and p is tested, then percentage Poisson distribution will be smaller at the variable Y. The results of the score test and the chi-square test for Poisson distribution is inversely proportional to the larger λ, n, and p is tested. Chi-square test for ZIP distribution indicates that ZIP regression able to overcome overdispersion due to excess p at the variable Y. This condition is indicated by the larger value of λ, then the percentage Poisson distribution reaches 0%, while the percentage of ZIP distribution in the range of 60% to 80%.

Table 2. Percentage of chi-square test on combination of λ, n, p (%)

<table>
<thead>
<tr>
<th>n</th>
<th>λ</th>
<th>p=0.1</th>
<th>p=0.3</th>
<th>p=0.5</th>
<th>p=0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>82.8</td>
<td>81.6</td>
<td>51.4</td>
<td>82.8</td>
<td>13.2</td>
</tr>
<tr>
<td>0.8</td>
<td>82.4</td>
<td>85.4</td>
<td>25.2</td>
<td>86.6</td>
<td>1.8</td>
</tr>
<tr>
<td>1</td>
<td>75.0</td>
<td>85.4</td>
<td>8.2</td>
<td>82.6</td>
<td>0.2</td>
</tr>
<tr>
<td>300</td>
<td>6</td>
<td>0.0</td>
<td>85.8</td>
<td>0.0</td>
<td>84.6</td>
</tr>
<tr>
<td>8</td>
<td>0.0</td>
<td>84.0</td>
<td>0.0</td>
<td>83.4</td>
<td>0.0</td>
</tr>
<tr>
<td>10</td>
<td>0.0</td>
<td>82.6</td>
<td>0.0</td>
<td>83.0</td>
<td>0.0</td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
<td>77.8</td>
<td>0.0</td>
<td>76.8</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Furthermore, the results of chi-square test with α of 0.05 for the Poisson and ZIP distribution against the combination of λ, n, p are shown in Table 2. The chi-square test for Poisson distribution shows that the larger λ, n, and p is tested, then percentage Poisson distribution will be smaller at the variable Y. The results of the score test and the chi-square test for Poisson distribution is inversely proportional to the larger λ, n, and p is tested. Chi-square test for ZIP distribution indicates that ZIP regression able to overcome overdispersion due to excess p at the variable Y. This condition is indicated by the larger value of λ, then the percentage Poisson distribution reaches 0%, while the percentage of ZIP distribution in the range of 60% to 80%.

Table 2. Percentage of chi-square test on combination of λ, n, p (%)

<table>
<thead>
<tr>
<th>n</th>
<th>λ</th>
<th>p=0.1</th>
<th>p=0.3</th>
<th>p=0.5</th>
<th>p=0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>81.6</td>
<td>87.8</td>
<td>28.8</td>
<td>86.0</td>
<td>1.8</td>
</tr>
<tr>
<td>0.8</td>
<td>76.2</td>
<td>81.8</td>
<td>7.6</td>
<td>85.8</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>67.4</td>
<td>87.0</td>
<td>0.2</td>
<td>86.4</td>
<td>0.0</td>
</tr>
<tr>
<td>500</td>
<td>6</td>
<td>0.0</td>
<td>86.0</td>
<td>0.0</td>
<td>86.6</td>
</tr>
<tr>
<td>8</td>
<td>0.0</td>
<td>84.6</td>
<td>0.0</td>
<td>83.4</td>
<td>0.0</td>
</tr>
<tr>
<td>10</td>
<td>0.0</td>
<td>81.8</td>
<td>0.0</td>
<td>82.4</td>
<td>0.0</td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
<td>77.4</td>
<td>0.0</td>
<td>75.6</td>
<td>0.0</td>
</tr>
</tbody>
</table>
B. The testing overdispersion on Poisson and ZIP regression

Results of exploration and testing of the simulation variables Y is a step that must be checked before performing Poisson regression analysis. When the variable Y has the characteristics of the data that led to overdispersion due to excess zero value, then ZIP regression became one of the completion of the Poisson regression. Overdispersion conditions on any combination of λ, n, and p are tested in Poisson and ZIP regression can be traced from the dispersion ratio (τ) and the Pearson chi-square test at 5% significance level.

Ratio of τ shows the value of a statistical result Pearson chi-square test of the degrees of freedom (n-k). The value of degrees of freedom Poisson and ZIP regression is different, because the Poisson regression using k=2, are the parameter estimate b0 and b1. ZIP regression using k=4 is based on discrete models for λ and zero-inflation models for p are g0 and g1, then l0 and l1.

The average ratio of τ obtained from 500 times of repetition of the combination of λ, n, and p is tested in Table 3. The overdispersion indicated by the ratio τ is greater than one. The ratio τ on Poisson regression will be compared with ZIP regression. The ratio τ most at risk are at p=0.7 in Poisson regression with the value λ is the larger. The ratio τ of the Poisson regression showed that the larger λ and p, then the ratio τ more than one in every n is tested. Poisson regression suffered overdispersion by the larger λ, p, and n is tested. The ratio τ of ZIP regression has a value of less than one in every λ, p, and n is tested, so ZIP regression able to overcome overdispersion caused excess zero probability at the variable Y.

<table>
<thead>
<tr>
<th>n</th>
<th>λ</th>
<th>p=0.1</th>
<th>p=0.3</th>
<th>p=0.5</th>
<th>p=0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Poisson</td>
<td>ZIP</td>
<td>Poisson</td>
<td>ZIP</td>
<td>Poisson</td>
</tr>
<tr>
<td>0.6</td>
<td>0.083</td>
<td>0.657</td>
<td>0.926</td>
<td>0.737</td>
<td>1.049</td>
</tr>
<tr>
<td>0.8</td>
<td>0.749</td>
<td>0.595</td>
<td>0.911</td>
<td>0.686</td>
<td>1.069</td>
</tr>
<tr>
<td>1</td>
<td>0.697</td>
<td>0.547</td>
<td>0.897</td>
<td>0.650</td>
<td>1.097</td>
</tr>
<tr>
<td>6</td>
<td>0.628</td>
<td>0.383</td>
<td>1.847</td>
<td>0.656</td>
<td>3.069</td>
</tr>
<tr>
<td>8</td>
<td>0.815</td>
<td>0.443</td>
<td>2.447</td>
<td>0.717</td>
<td>4.076</td>
</tr>
<tr>
<td>10</td>
<td>1.022</td>
<td>0.500</td>
<td>3.076</td>
<td>0.764</td>
<td>5.102</td>
</tr>
<tr>
<td>20</td>
<td>2.034</td>
<td>0.675</td>
<td>6.135</td>
<td>0.885</td>
<td>10.201</td>
</tr>
<tr>
<td>0.6</td>
<td>0.796</td>
<td>0.645</td>
<td>0.913</td>
<td>0.718</td>
<td>1.032</td>
</tr>
<tr>
<td>0.8</td>
<td>0.733</td>
<td>0.581</td>
<td>0.897</td>
<td>0.669</td>
<td>1.060</td>
</tr>
<tr>
<td>1</td>
<td>0.684</td>
<td>0.532</td>
<td>0.883</td>
<td>0.633</td>
<td>1.086</td>
</tr>
<tr>
<td>6</td>
<td>0.622</td>
<td>0.377</td>
<td>1.827</td>
<td>0.641</td>
<td>3.040</td>
</tr>
<tr>
<td>8</td>
<td>0.808</td>
<td>0.437</td>
<td>2.422</td>
<td>0.699</td>
<td>4.029</td>
</tr>
<tr>
<td>10</td>
<td>1.013</td>
<td>0.493</td>
<td>3.025</td>
<td>0.745</td>
<td>5.036</td>
</tr>
<tr>
<td>20</td>
<td>2.012</td>
<td>0.665</td>
<td>6.038</td>
<td>0.863</td>
<td>10.069</td>
</tr>
<tr>
<td>0.6</td>
<td>0.791</td>
<td>0.642</td>
<td>0.912</td>
<td>0.715</td>
<td>1.032</td>
</tr>
<tr>
<td>0.8</td>
<td>0.734</td>
<td>0.579</td>
<td>0.895</td>
<td>0.666</td>
<td>1.055</td>
</tr>
<tr>
<td>1</td>
<td>0.684</td>
<td>0.529</td>
<td>0.884</td>
<td>0.630</td>
<td>1.082</td>
</tr>
<tr>
<td>6</td>
<td>0.616</td>
<td>0.376</td>
<td>1.821</td>
<td>0.637</td>
<td>3.030</td>
</tr>
<tr>
<td>8</td>
<td>0.805</td>
<td>0.436</td>
<td>2.415</td>
<td>0.696</td>
<td>4.023</td>
</tr>
<tr>
<td>10</td>
<td>1.002</td>
<td>0.491</td>
<td>3.019</td>
<td>0.742</td>
<td>5.020</td>
</tr>
<tr>
<td>20</td>
<td>2.006</td>
<td>0.663</td>
<td>6.024</td>
<td>0.858</td>
<td>10.050</td>
</tr>
</tbody>
</table>

Overdispersion

The percentage of overdispersion have the similar result with the ratio of τ. Pearson chi-square test that significant shows overdispersion. Value of overdispersion contained in Table 4 show that in every λ, p, and n is tested to get overdispersion value reaching 0% for ZIP regression. These results indicate that ZIP regression is better to handle overdispersion caused excess zero probability on the variable Y. The percentage of overdispersion on Poisson regression showed that the larger λ and p, then the greater overdispersion in each n is tested that indicated by the value reached 100%. These results are consistent with the score test and the chi-square test that indicates the larger λ, n, and p is tested, then the greater zero probability that appears and the less spread Poisson at the variable Y.
The goodness of fit on Poisson and ZIP regression model showed that the larger \( \lambda, n, \) and \( p \) is tested, then ZIP regression is better than Poisson regression. Furthermore, the comparison of Poisson and ZIP regression models is taken based on the evaluation parameter estimate and \( y \). The average of ARB for each combination of \( \lambda, n, \) and \( p \) are tested on estimate \( \theta_1 \) in Table 5. The average of ARB on ZIP regression in Table 5 produces a value is smaller as enlargement \( \lambda, n, \) and \( p \) is tested against estimate \( \theta_1 \). The value of ARB on Poisson regression has a minimum value that is contained in \( \lambda = 6 \). The average of ARB on estimate \( \theta_1 \) indicates that Poisson regression is better than ZIP regression on the value of \( \lambda \) are 0.6, 0.8, and 1 in each of \( p \) is tested. Value of \( \lambda \) are 6, 8, 10, and 20 in each of the \( p \) is tested showed that ZIP regression is better than Poisson regression.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \lambda )</th>
<th>( p=0.1 )</th>
<th>( p=0.3 )</th>
<th>( p=0.5 )</th>
<th>( p=0.7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>19.47*</td>
<td>75.753</td>
<td>21.241*</td>
<td>103.494</td>
<td>22.568*</td>
</tr>
<tr>
<td>0.8</td>
<td>15.442*</td>
<td>51.298</td>
<td>16.928*</td>
<td>73.799</td>
<td>19.136*</td>
</tr>
<tr>
<td>1</td>
<td>13.883*</td>
<td>38.540</td>
<td>14.307*</td>
<td>57.739</td>
<td>17.575*</td>
</tr>
<tr>
<td>8</td>
<td>7.786</td>
<td>1.229*</td>
<td>12.766</td>
<td>1.511*</td>
<td>17.554</td>
</tr>
<tr>
<td>10</td>
<td>8.224</td>
<td>0.465*</td>
<td>14.140</td>
<td>0.619*</td>
<td>20.226</td>
</tr>
<tr>
<td>20</td>
<td>11.106</td>
<td>0.425*</td>
<td>20.382</td>
<td>0.452*</td>
<td>24.256</td>
</tr>
<tr>
<td>0.6</td>
<td>10.389*</td>
<td>70.421</td>
<td>11.202*</td>
<td>94.119</td>
<td>11.825*</td>
</tr>
<tr>
<td>0.8</td>
<td>8.973*</td>
<td>50.053</td>
<td>9.221*</td>
<td>70.143</td>
<td>10.868*</td>
</tr>
<tr>
<td>1</td>
<td>7.996*</td>
<td>37.781</td>
<td>8.925*</td>
<td>56.098</td>
<td>9.808*</td>
</tr>
<tr>
<td>8</td>
<td>4.291</td>
<td>1.263*</td>
<td>7.945</td>
<td>1.566*</td>
<td>9.604</td>
</tr>
<tr>
<td>10</td>
<td>4.671</td>
<td>0.400*</td>
<td>8.039</td>
<td>0.477*</td>
<td>10.186</td>
</tr>
<tr>
<td>20</td>
<td>6.389</td>
<td>0.286*</td>
<td>11.708</td>
<td>0.291*</td>
<td>15.122</td>
</tr>
<tr>
<td>0.6</td>
<td>8.188*</td>
<td>70.067</td>
<td>8.862*</td>
<td>92.178</td>
<td>9.113*</td>
</tr>
<tr>
<td>0.8</td>
<td>6.664*</td>
<td>49.296</td>
<td>7.110*</td>
<td>69.294</td>
<td>7.991*</td>
</tr>
<tr>
<td>1</td>
<td>5.481*</td>
<td>37.741</td>
<td>6.650*</td>
<td>55.680</td>
<td>6.961*</td>
</tr>
<tr>
<td>6</td>
<td>2.801</td>
<td>3.226*</td>
<td>5.182</td>
<td>4.654*</td>
<td>6.397</td>
</tr>
<tr>
<td>8</td>
<td>3.482</td>
<td>1.255*</td>
<td>5.601</td>
<td>1.590*</td>
<td>7.814</td>
</tr>
<tr>
<td>10</td>
<td>3.754</td>
<td>0.422*</td>
<td>6.353</td>
<td>0.499*</td>
<td>8.395</td>
</tr>
<tr>
<td>20</td>
<td>4.805</td>
<td>0.251*</td>
<td>8.932</td>
<td>0.272*</td>
<td>11.032</td>
</tr>
</tbody>
</table>

\(^{*}\)Average of ARB is smaller

The result of RMSE is similar with the ARB on estimate \( \theta_1 \) in Table 6. MSE contains two components, namely variance estimate (accuracy) and the bias (accuracy) [11]. The estimation with
good character of MSE is that can controls the variance and bias. The big value of RRMSE show the big variance estimate, so the risk to the estimation results, the accuracy estimation is lower.

Table 6. Average of RRMSE againsts estimate $\theta_1$ on Poisson and ZIP regression

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\lambda$</th>
<th>$p=0.1$</th>
<th>$p=0.3$</th>
<th>$p=0.5$</th>
<th>$p=0.7$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Poisson</td>
<td>ZIP</td>
<td>Poisson</td>
<td>ZIP</td>
<td>Poisson</td>
</tr>
<tr>
<td>0.6</td>
<td>23.931°</td>
<td>78.167°</td>
<td>27.325°</td>
<td>107.113°</td>
<td>28.948°</td>
</tr>
<tr>
<td>0.8</td>
<td>19.443°</td>
<td>52.643°</td>
<td>21.295°</td>
<td>75.578°</td>
<td>24.684°</td>
</tr>
<tr>
<td>1</td>
<td>13.883°</td>
<td>38.542°</td>
<td>18.047°</td>
<td>58.795°</td>
<td>22.019°</td>
</tr>
<tr>
<td>8</td>
<td>9.725°</td>
<td>1.681°</td>
<td>15.946°</td>
<td>1.856°</td>
<td>21.576°</td>
</tr>
<tr>
<td>10</td>
<td>10.166°</td>
<td>0.683°</td>
<td>17.949°</td>
<td>0.913°</td>
<td>25.113°</td>
</tr>
<tr>
<td>20</td>
<td>14.009°</td>
<td>0.535°</td>
<td>25.870°</td>
<td>0.582°</td>
<td>31.394°</td>
</tr>
</tbody>
</table>

*Average of RRMSE is smaller

The average of SAPR on the estimate $y$ is gotten by Poisson and ZIP regression models. The estimate $y$ in ZIP regression model using two models are discrete model for $\lambda$ and zero-inflation model for $p$. In Table 7 shows that the larger $\lambda$, $n$, and $p$ is tested, then the larger the value of SAPR. ZIP Regression have the average of SAPR that is smaller than Poisson regression in every $\lambda$, $n$, and $p$ is tested.

Table 7. Average of SAPR againsts estimate $y$ on Poisson and ZIP regression

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\lambda$</th>
<th>$p=0.1$</th>
<th>$p=0.3$</th>
<th>$p=0.5$</th>
<th>$p=0.7$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Poisson</td>
<td>ZIP</td>
<td>Poisson</td>
<td>ZIP</td>
<td>Poisson</td>
</tr>
<tr>
<td>0.6</td>
<td>0.166°</td>
<td>0.155°</td>
<td>0.171°</td>
<td>0.155°</td>
<td>0.160°</td>
</tr>
<tr>
<td>0.8</td>
<td>0.180°</td>
<td>0.153°</td>
<td>0.191°</td>
<td>0.154°</td>
<td>0.184°</td>
</tr>
<tr>
<td>1</td>
<td>0.154°</td>
<td>0.139°</td>
<td>0.185°</td>
<td>0.158°</td>
<td>0.190°</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
<td>0.081°</td>
<td>0.069°</td>
<td>0.241°</td>
<td>0.145°</td>
</tr>
<tr>
<td>8</td>
<td>0.106°</td>
<td>0.078°</td>
<td>0.281°</td>
<td>0.151°</td>
<td>0.401°</td>
</tr>
<tr>
<td>10</td>
<td>0.123°</td>
<td>0.084°</td>
<td>0.330°</td>
<td>0.158°</td>
<td>0.462°</td>
</tr>
<tr>
<td>20</td>
<td>0.166°</td>
<td>0.097°</td>
<td>0.459°</td>
<td>0.169°</td>
<td>0.645°</td>
</tr>
<tr>
<td>300</td>
<td>6</td>
<td>0.504°</td>
<td>0.466°</td>
<td>0.523°</td>
<td>0.470°</td>
</tr>
<tr>
<td>8</td>
<td>0.516°</td>
<td>0.455°</td>
<td>0.563°</td>
<td>0.472°</td>
<td>0.559°</td>
</tr>
<tr>
<td>1</td>
<td>0.476°</td>
<td>0.437°</td>
<td>0.571°</td>
<td>0.475°</td>
<td>0.583°</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
<td>0.272°</td>
<td>0.216°</td>
<td>0.700°</td>
<td>0.432°</td>
</tr>
<tr>
<td>8</td>
<td>0.324°</td>
<td>0.236°</td>
<td>0.816°</td>
<td>0.453°</td>
<td>1.162°</td>
</tr>
<tr>
<td>10</td>
<td>0.371°</td>
<td>0.252°</td>
<td>0.962°</td>
<td>0.473°</td>
<td>1.334°</td>
</tr>
<tr>
<td>20</td>
<td>0.492°</td>
<td>0.289°</td>
<td>1.311°</td>
<td>0.506°</td>
<td>1.881°</td>
</tr>
<tr>
<td>500</td>
<td>6</td>
<td>0.861°</td>
<td>0.779°</td>
<td>0.875°</td>
<td>0.784°</td>
</tr>
<tr>
<td>8</td>
<td>0.839°</td>
<td>0.747°</td>
<td>0.898°</td>
<td>0.783°</td>
<td>0.909°</td>
</tr>
<tr>
<td>1</td>
<td>0.832°</td>
<td>0.715°</td>
<td>0.931°</td>
<td>0.784°</td>
<td>0.947°</td>
</tr>
<tr>
<td>500</td>
<td>6</td>
<td>0.459°</td>
<td>0.361°</td>
<td>1.187°</td>
<td>0.722°</td>
</tr>
<tr>
<td>8</td>
<td>0.513°</td>
<td>0.389°</td>
<td>1.379°</td>
<td>0.757°</td>
<td>1.973°</td>
</tr>
<tr>
<td>10</td>
<td>0.586°</td>
<td>0.418°</td>
<td>1.576°</td>
<td>0.786°</td>
<td>2.225°</td>
</tr>
<tr>
<td>20</td>
<td>0.864°</td>
<td>0.484°</td>
<td>2.233°</td>
<td>0.845°</td>
<td>3.166°</td>
</tr>
</tbody>
</table>

*Average of SAPR is smaller

Lili Puspita Rahayu et. al. (Overdispersion study of poisson and zero-inflated poisson regression for ...
A good estimator should have a small bias and variance. Poisson and ZIP regression compared based on the ability to control bias and variance estimate to get the high accuracy estimation. Bias and variance of Poisson and ZIP regression parameter estimate shown on average of ARB and RRMSE. Then, to determine the predictability of Poisson and ZIP regression models shown on average of SAPR on the value estimate $y$. Value of ARB and RRMSE on estimate $\theta_1$ in Poisson and ZIP regression indicates that the larger $\lambda$, $n$, and $p$ is tested, then ZIP regression is better than Poisson regression. The average of SAPR estimate $y$ from the Poisson and ZIP regression model showed that ZIP regression is better than Poisson regression in every $\lambda$, $n$, and $p$ is tested. ZIP regression is better than Poisson regression in solving the overdispersion due to the many of excess zero value at the variable Y. The estimation evaluation results of Poisson and ZIP regression model in accordance with the results of the testing overdispersion, exploration and testing of the variable Y.

VI. Conclusion

Overdispersion study of the simulation data from combination of $\lambda$, $n$, $p$ that is tested indicates that the greater $\lambda$, $n$, and $p$, then the scores test get excess the zero probability are getting bigger and chi-squared test get percentage of Poisson distribution is getting smaller. The comparison showed that ZIP regression is better than Poisson regression based on the percentage of overdispersion and ratio of dispersion, the value of ARB and RRMSE on estimate $\theta_1$, and the average of SAPR on estimate $y$, along with the growing $\lambda$, $n$, and $p$ that is tested.

VII. Open Problem

The design of simulation performed in this study started from generate variable Y and the variable X is based on the determination of parameters, so there is the relationship between the two variables. In a subsequent study to do different phases of design simulation and there is a residual effect.

References