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*

ARTICLE INFO

Article history:
Received November 19, 2016
Revised November 25, 2016
Accepted November 25, 2016

Keywords:
Overdispersion
Poisson
Zero inflated Poisson
Regression
Simulation data

ABSTRACT

Poisson distribution is one of discrete distribution that is often used in modeling of rare events. The data obtained in 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

Copyright © 2016 International Journal of Advances in Intelligent Informatics.
All rights reserved.

I. Introduction

Poisson distribution is one of the discrete distribution that is often used for modeling on rare occasions. The data obtained in the form of counts with non-negative integers. One form of analysis used to model count data is Poisson regression. Poisson regression analysis showed a relationship between the explanatory variables with the response variable that spread Poisson. Poisson regression has equidispersion assumptions, a condition in which the mean and variance of the response variable equal value. Deviation assumptions that often events on Poisson regression is overdispersion/ underdispersion. Overdispersion is the variance greater than the mean, while the value underdispersion is the variance smaller than mean value on response variable. Application about underdispersion on Poisson regression is rare eventring, it is because there is no low variance value of the response variable on real data [1].

Problem often encountered in Poisson regression was overdispersion. This condition is caused by the explanatory variable that can't be explained in the model, so it is possible the high variability of the response variable caused by other variables. Cause of overdispersion that often event in Poisson regression is zero probability that excess on the response variable. One result is the standard deviation of parameter estimate to be underestimate and the significance of the explanatory variables to be overstate, resulting invalid conclusions [2].

The handling model can be used to overcome overdipersion due to zero probability excess on the response variable Poisson regression such as models of hurdle Poisson regression, zero-inflated Poisson (ZIP) regression, and Semiparametric hurdle Poisson regression [3]. Handling model that will be used in this paper is the model of ZIP regression, because it is more convenient than models of hurdle Poisson regression and Semiparametric hurdle Poisson. Superiority of ZIP regression is very easily applied to several fields [4] such as agriculture, animal husbandry, biostatistics, and industry.

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 $\mathbf{y_i}$, $i=1,2,\ldots,n$ represents count of those rare occasions in the period with value of the parameter Poisson distribution lambda (λ_i). $\mathbf{y_i}$ is a Poisson random variable that spread by the mass function probability of the following

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

with assuming of Poisson regression is

$$E(Y_i) = Var(Y_i) = \lambda_i \tag{2}$$

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

$$Var\ (Y) = \tau\, E\ (Y)\ =\ \tau\, \lambda \tag{3}$$

with τ is ratio of dispersion. When there was overdispersion on Poisson regression then value of τ 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 x (p+1). Random variables y_i and i is a row vector of **X**, will be linked with the log link function.

$$\ln\left(\lambda_i\right) = \eta_i = \mathbf{Xb} \tag{4}$$

$$\lambda_i = \exp\left(\mathbf{Xb}\right) \tag{5}$$

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_i. The second step events on the probability that result of data count spread Poisson with parameter λ is denoted by (1- p_i) .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 $Y_i \sim ZIP(\lambda_i, p_i)$ have assumptions on the ZIP regression

$$E(Y_i) = (1-p_i)\lambda_i = \mu_i \tag{6}$$

and

$$Var(Y_i) = \mu_i + \left(\frac{p_i}{1-p_i}\right)\mu_i^2 \tag{7}$$

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 $y_1,\ldots,y_n$ 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 of divided into two components, namely discrete data models for λ 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

$$\ln(\lambda) = \mathbf{Xb}$$
$$\ln\left(\frac{p}{1-p}\right) = \mathbf{X}g \tag{8}$$

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 (λ) starting from λ= 0.6, 0.8, 1, 6, 8, 10, and 20, the zero probability (*p*) are *p*=0.1, 0.3, 0.5, and 0.7, and sample size (*n*) are *n*=100, 300, and 500. The data generated are useful to obtain parameter estimators of Poisson and ZIP regression. The coefficient of regression parameters were determined are θ_0=3, and θ_1=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 (μ,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 *λ*, *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 $x_i = (\ln(y_i) - \theta_0)/\theta_1$, 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).

$$\text{RAB} = \sum_{i=1}^{500} \left|\frac{\hat{\theta}_i - \theta}{\theta}\right| \text{x}100\% \tag{9}$$

$$RRMSE = \sqrt{\sum_{i=1}^{500}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right)^2} \times 100\% \tag{10}$$

$$PR_i = \frac{\sum_{j=1}^{n}(y_{ij} - \hat{y}_{ij})}{\sqrt{Var(Y)}} \tag{11}$$

then

$$SAPR = \sum_{i=1}^{500} |\text{PR}_i| \tag{12}$$

where r is the number of simulation replications, $\hat{\theta}_i$ is the parameter estimators of Poisson and ZIP regression ith, θ 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 Var (Y) is the estimate variance of Poisson and ZIP regression. Then, the smaller values of RAB, RRMSE, 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 λ, 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 againts λ, n, and p is tested indicates that the rise of the value of p effect on λ. Value of λ 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 λ that has small value still has p from Poisson distribution relatively large. This statement can be explained by the cumulative Poisson distribution table. Value λ is tested to 0.6, 0.8, 1, 6, and 8 still have zero probability of a Poisson distribution, while for the other λ is tested, namely 10 and 20 already have not zero probability of a Poisson distribution.

Exploration variable Y to λ, 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 α of 0.05 on simulation data at variable Y against a combination of λ, n, p are shown in Table 1. Scores test indicate that the larger λ, n, and p is tested, then the larger percentage of excess zero at the variable Y.

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

n	λ	p=0.1	p=0.3	p=0.5	p=0.7
100	0.6	6.5	23.4	45.6	88.9
	0.8	9.0	39.2	66.8	88.9
	1	10.0	55.0	86.2	92.8
	6	100.0	100.0	100.0	100.0
	8	100.0	100.0	100.0	100.0
	10	100.0	100.0	100.0	100.0
	20	100.0	100.0	100.0	100.0
300	0.6	13.2	55.0	89	96.2
	0.8	18.2	82.6	98.8	99.8
	1	24.4	96.4	100.0	100.0
	6	100.0	100.0	100.0	100.0
	8	100.0	100.0	100.0	100.0
	10	100.0	100.0	100.0	100.0
	20	100.0	100.0	100.0	100.0
500	0.6	17.6	77.0	98.0	99.6
	0.8	25.0	96.2	100	100.0
	1	39.0	100	100	100.0
	6	100.0	100.0	100.0	100.0
	8	100.0	100.0	100.0	100.0
	10	100.0	100.0	100.0	100.0
	20	100.0	100.0	100.0	100.0

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 (%)

n	*λ*	*p*=0.1		*p*=0.3		*p*=0.5		*p*=0.7	
		Poisson	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**
100	0.6	91.0	83.6	75.0	85.4	54.8	85.2	33.4	80.6
	0.8	89.2	80.2	62.6	82.0	32.0	81.6	18.0	85.4
	1	88.2	82.2	51.0	86.0	14.2	85.0	7.8	82.8
	6	0.0	84.4	0.0	83.4	0.0	82.8	0.0	82.0
	8	0.0	81.6	0.0	79.6	0.0	80.4	0.0	76.8
	10	0.0	79.4	0.0	78.6	0.0	78.2	0.0	75.0
	20	0.0	72.6	0.0	71.2	0.0	71.6	0.0	66.8
300	0.6	82.8	81.6	51.4	82.8	13.2	86.2	4.0	86.6
	0.8	82.4	85.4	25.2	86.6	1.8	86.8	0.2	83.2
	1	75.0	85.4	8.2	82.6	0.2	82.6	0.0	85.8
	6	0.0	85.8	0.0	84.6	0.0	86.4	0.0	84.4
	8	0.0	84.0	0.0	83.4	0.0	84.6	0.0	81.6
	10	0.0	82.6	0.0	83.0	0.0	80.6	0.0	79.4
	20	0.0	77.8	0.0	76.8	0.0	76.0	0.0	72.6
500	0.6	81.6	87.8	28.8	86.0	1.8	83.8	0.2	86.2
	0.8	76.2	81.8	7.6	85.8	0.2	86.2	0.0	87.0
	1	67.4	87.0	0.2	86.4	0.0	84.6	0.0	83.2
	6	0.0	86.0	0.0	86.6	0.0	85.8	0.0	86.4
	8	0.0	84.6	0.0	82.6	0.0	83.4	0.0	84.6
	10	0.0	81.8	0.0	82.4	0.0	81.4	0.0	80.6
	20	0.0	77.4	0.0	75.6	0.0	77.0	0.0	76.0

B. The testing overdipserion 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 excessi zero value, then ZIP regression became one of the completion of the Poisson regression. Overdipersion 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 b_0 and b_1. ZIP regression using k=4 is based on discrete models for λ and zero-inflation modesl for p are g_0 and g_1, then l_0 and l_1.

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 3. Dispersion ratio on Poisson and ZIP regression

n	λ	p=0.1		p=0.3		p=0.5		p=0.7	
		Poisson	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**
100	0.6	0.803	0.657	0.926	0.737	1.049*	0.808	1.188*	0.863
	0.8	0.749	0.595	0.911	0.686	1.069*	0.770	1.247*	0.838
	1	0.697	0.547	0.897	0.650	1.097*	0.741	1.286*	0.819
	6	0.628	0.383	1.847*	0.656	3.069*	0.769	4.291*	0.829
	8	0.815	0.443	2.447*	0.717	4.076*	0.818	5.695*	0.871
	10	1.022*	0.500	3.076*	0.764	5.102*	0.855	7.138*	0.901
	20	2.034*	0.675	6.135*	0.885	10.201*	0.943	14.246*	0.969
300	0.6	0.796	0.645	0.913	0.718	1.032*	0.791	1.156*	0.857
	0.8	0.733	0.581	0.897	0.669	1.060*	0.751	1.221*	0.825
	1	0.684	0.532	0.883	0.633	1.086*	0.723	1.286*	0.801
	6	0.622	0.377	1.827*	0.641	3.040*	0.749	4.246*	0.807
	8	0.808	0.437	2.422*	0.699	4.029*	0.796	5.631*	0.847
	10	1.013*	0.493	3.025*	0.745	5.036*	0.833	7.051*	0.877
	20	2.012*	0.665	6.038*	0.863	10.069*	0.917	14.092*	0.943
500	0.6	0.791	0.642	0.912*	0.715	1.032*	0.786	1.151*	0.853
	0.8	0.734	0.579	0.895*	0.666	1.055*	0.747	1.217*	0.821
	1	0.684	0.529	0.884*	0.630	1.082*	0.719	1.281*	0.797
	6	0.616	0.376	1.821*	0.637	3.030*	0.744	4.231*	0.803
	8	0.805	0.436	2.415*	0.696	4.023*	0.792	5.621*	0.843
	10	1.002*	0.491	3.019*	0.742	5.020*	0.828	7.034*	0.872
	20	2.006*	0.663	6.024*	0.858	10.050*	0.912	14.058*	0.938

*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.

Table 4. Percentage of overdispersion on Poisson and ZIP regression

n	λ	*p*=0.1		*p*=0.3		*p*=0.5		*p*=0.7	
		Poisson	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**
	0.6	0.0	0.0	0.0	0.0	7.2*	0.0	40.0*	0.0
	0.8	0.0	0.0	0.0	0.0	10.2*	0.0	60.0*	0.0
	1	0.0	0.0	0.0	0.0	13.6*	0.0	60.0*	0.0
100	6	0.0	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	8	0.0	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	10	1.0*	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	20	100.0*	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	0.6	0.0	0.0	0.0	0.0	9.2*	0.0	55.4*	0.0
	0.8	0.0	0.0	0.0	0.0	15.2*	0.0	77.2*	0.0
	1	0.0	0.0	0.0	0.0	23.8*	0.0	92.8*	0.0
300	6	0.0	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	8	0.0	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	10	2.6*	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	20	100.0*	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	0.6	0.0	0.0	0.0	0.0	9.6*	0.0	70.8*	0.0
	0.8	0.0	0.0	0.0	0.0	20.0*	0.0	94.0*	0.0
	1	0.0	0.0	0.0	0.0	36.0*	0.0	99.0*	0.0
500	6	0.0	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	8	0.0	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	10	1.2*	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0
	20	100.0*	0.0	100.0*	0.0	100.0*	0.0	100.0*	0.0

*Overdipersion

C. Evaluation of the estimation on Poisson and ZIP regression model

The goodness of fit on Poisson and ZIP regression model showed that the larger λ, *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 λ, *n*, and *p* are tested on estimate θ_1 in Table 5. The average of ARB on ZIP regression in Table 5 produces a value is smaller as enlargement λ, *n*, and *p* is tested against estimate θ_1. The value of ARB on Poisson regression has a minimum value that is contained in $\lambda = 6$. The average of ARB on estimate θ_1 indicates that Poisson regression is better than ZIP regression on the value of λ are 0.6, 0.8, and 1 in each of *p* and *n* is tested. Value of λ are 6, 8, 10, and 20 in each of the *p* and *n* is tested showed that ZIP regression is better than Poisson regression.

Table 5. Average of ARB againts estimate θ_1 on Poisson and ZIP regression

n	λ	*p*=0.1		*p*=0.3		*p*=0.5		*p*=0.7	
		Poisson	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**
	0.6	19.476*	75.753	21.241*	103.494	22.586*	143.374	26.006*	166.700
	0.8	15.442*	51.298	16.928*	73.799	19.136*	105.497	21.128*	160.675
	1	13.883*	38.542	14.307*	57.739	17.575*	83.119	19.588*	122.616
100	6	6.878	3.073*	11.644	4.503*	14.651	5.066*	18.614	5.626*
	8	7.786	1.229*	12.766	1.511*	17.554	1.762*	21.789	1.837*
	10	8.224	0.465*	14.140	0.619*	20.226	0.648*	23.569	0.861*
	20	11.106	0.425*	20.382	0.452*	24.256	0.544*	31.381	0.736*
	0.6	10.389*	70.421	11.202*	94.119	11.825*	124.301	13.310*	171.498
	0.8	8.973*	50.053	9.221*	70.143	10.868*	95.410	10.905*	129.607
	1	7.996*	37.781	8.925*	56.098	9.808*	77.995	10.602*	105.616
300	6	3.792	3.306*	6.851	4.656*	8.561	5.055*	10.816	5.435*
	8	4.291	1.263*	7.945	1.566*	9.604	1.692*	10.920	1.770*
	10	4.671	0.400*	8.039	0.477*	10.186	0.540*	13.146	0.561*
	20	6.389	0.286*	11.708	0.291*	15.122	0.326*	18.366	0.426*
	0.6	8.188*	70.067	8.862*	92.178	9.113*	121.813	10.264*	161.514
	0.8	6.664*	49.296	7.110*	69.294	7.991*	94.169	8.798*	125.793
	1	5.481*	37.741	6.650*	55.680	6.961*	77.259	8.568*	103.199
500	6	2.801	3.226*	5.182	4.654*	6.397	5.121*	7.980	5.436*
	8	3.482	1.255*	5.601	1.590*	7.814	1.694*	8.682	1.743*
	10	3.754	0.422*	6.353	0.499*	8.395	0.517*	10.205	0.550*
	20	4.805	0.251*	8.932	0.272*	11.032	0.278*	14.338	0.341*

* Average of ARB is smaller

The result of RRMSE is similar with the ARB on estimate θ_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 againts estimate θ_1 on Poisson and ZIP regression

n	λ	p=0.1		p=0.3		p=0.5		p=0.7	
		Poisson	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**
100	0.6	23.931*	78.167	27.325*	107.113	28.948*	148.998	30.586*	177.728
	0.8	19.443*	52.643	21.295*	75.578	24.684*	108.939	27.273*	169.880
	1	13.883*	38.542	18.047*	58.795	22.019*	85.038	25.117*	129.679
	6	8.413	3.512*	14.546	4.950*	18.534	5.566*	24.056	6.420*
	8	9.725	1.681*	15.946	1.856*	21.576	2.240*	27.152	2.528*
	10	10.166	0.683*	17.949	0.913*	25.113	0.915*	29.879	1.214*
	20	14.009	0.535*	25.870	0.582*	31.394	0.693*	40.008	0.940*
300	0.6	13.088*	70.968	14.109*	94.960	15.152*	125.887	16.502*	176.730
	0.8	11.101*	50.462	11.411*	70.635	13.491*	96.326	13.943*	131.993
	1	9.856*	38.109	11.226*	56.421	12.552*	78.553	13.306*	107.173
	6	4.805	3.489*	8.597	4.815*	10.538	5.221*	13.605	5.680*
	8	5.474	1.420*	9.805	1.694*	12.220	1.820*	13.896	1.991*
	10	5.984	0.541*	10.201	0.597*	12.824	0.655*	16.695	0.718*
	20	7.940	0.354*	14.687	0.362*	18.903	0.411*	23.665	0.533*
500	0.6	10.065*	70.389	11.074*	92.632	11.480*	122.711	13.031*	163.979
	0.8	8.433*	49.534	8.944*	69.564	10.004*	94.700	11.058*	127.234
	1	6.853*	37.922	8.413*	55.887	8.962*	77.586	10.618*	104.121
	6	3.559	3.321*	6.535	4.730*	8.271	5.212*	9.985	5.592*
	8	4.297	1.355*	7.044	1.671*	9.584	1.780*	10.844	1.878*
	10	4.694	0.518*	7.994	0.583*	10.545	0.608*	12.877	0.659*
	20	6.023	0.306*	11.293	0.332*	13.830	0.341*	18.083	0.422*

*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 λ and zero-inflation model for *p*. In Table 7 shows that the larger λ, *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 λ, *n*, and *p* is tested.

Table 7. Average of SAPR againts estimate *y* on Poisson and ZIP regression

n	λ	p=0.1		p=0.3		p=0.5		p=0.7	
		Poisson	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**	**Poisson**	**ZIP**
100	0.6	0.166	0.155*	0.171	0.155*	0.160	0.148*	0.145	0.096*
	0.8	0.180	0.153*	0.191	0.154*	0.184	0.154*	0.163	0.135*
	1	0.154	0.139*	0.185	0.158*	0.190	0.155*	0.175	0.141*
	6	0.081	0.069*	0.241	0.145*	0.346	0.170*	0.366	0.163*
	8	0.106	0.078*	0.281	0.151*	0.401	0.176*	0.431	0.168*
	10	0.123	0.084*	0.330	0.158*	0.462	0.182*	0.502	0.171*
	20	0.166	0.097*	0.459	0.169*	0.645	0.191*	0.709	0.177*
300	0.6	0.504	0.466*	0.523	0.470*	0.503	0.444*	0.425	0.374*
	0.8	0.516	0.455*	0.563	0.472*	0.559	0.455*	0.496	0.393*
	1	0.476	0.437*	0.571	0.475*	0.583	0.465	0.519	0.402*
	6	0.272	0.216*	0.700	0.432*	1.005	0.513*	1.090	0.488*
	8	0.324	0.236*	0.816	0.453*	1.162	0.529*	1.274	0.500*
	10	0.371	0.252*	0.962	0.473*	1.334	0.544*	1.447	0.512*
	20	0.492	0.289*	1.311	0.506*	1.881	0.571*	2.043	0.530*
500	0.6	0.861	0.779*	0.875	0.784*	0.837	0.744*	0.725	0.634*
	0.8	0.839	0.747*	0.898	0.783*	0.909	0.762*	0.829	0.657*
	1	0.832	0.715*	0.931	0.784*	0.947	0.780*	0.863	0.678*
	6	0.459	0.361*	1.187	0.722*	1.673	0.855*	1.811	0.815*
	8	0.513	0.389*	1.379	0.757*	1.973	0.884*	2.119	0.836*
	10	0.586	0.418*	1.576	0.786*	2.225	0.908*	2.407	0.852*
	20	0.864	0.484*	2.233	0.845*	3.166	0.950*	3.416	0.884*

* Average of SAPR is smaller

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 θ_1 in Poisson and ZIP regression indicates that the larger λ, *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 λ, *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 λ, *n, p* that is tested indicates that the greater λ, *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 θ_1, and the average of SAPR on estimate *y*, along with the growing λ, *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

[1] N. Jansakul and J. P. Hinde, "Score tests for zero-inflated Poisson models," *Comput. Stat. Data Anal.*, vol. 40, no. 1, pp. 75–96, 2002.

[2] N. Ismail and A. A. Jemain, "Handling overdispersion with negative binomial and generalized Poisson regression models," in *Casualty Actuarial Society Forum*, 2007, pp. 103–158.

[3] M. Ridout, C. G. B. Demétrio, and J. Hinde, "Models for count data with many zeros," in *Proceedings of the XIXth international biometric conference*, 1998, vol. 19, pp. 179–192.

[4] D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

[5] M. Xie, B. He, and T. N. Goh, "Zero-inflated Poisson model in statistical process control," *Comput. Stat. Data Anal.*, vol. 38, no. 2, pp. 191–201, 2001.

[6] S. Numna, "Analysis of extra zero counts using zero-inflated Poisson models," Prince of Songkla University, 2009.

[7] J. W. Hardin, J. M. Hilbe, and J. Hilbe, *Generalized linear models and extensions*. Stata press, 2007.

[8] R. Savic and M. Lavielle, "Performance in population models for count data, part II: a new SAEM algorithm," *J. Pharmacokinet. Pharmacodyn.*, vol. 36, no. 4, pp. 367–379, 2009.

[9] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*, vol. 53. Cambridge university press, 2013.

[10] M. Flynn and L. A. Francis, "More flexible GLMs zero-inflated models and hybrid models," *Casualty Actuar. Soc*, vol. 2009, pp. 148–224, 2009.

[11] G. Casella and R. L. Berger, *Statistical inference*, vol. 2. Duxbury Pacific Grove, CA, 2002.