Journal of Mathematics Education and Application (JMEA)

Vol. 2, No 2, Juni 2023, pp. 96-107 DOI: https://doi.org/10.30596/jmea.v2i2.13591

Zero-Inflated Poisson Regression Testing In Handling Overdispersion On Poisson Regression

Mutia Sari[1*], Open Darnius[2]

[1] Student of Study Program of Mathematics, Universitas Sumatera Utara, Indonesia

[2] Lecturer of Study Program of Mathematics, Universitas Sumatera Utara, Indonesia

*Corresponding Author. E-mail: mutiasari256@gmail.com

Article Info

Article History

Received : 22 Februari 2023

Accepted : 10 Mei 2023

Published : 30 Juni 2023

Keywords:

Generalized Linear Model, Overdispersion, Poisson Regression, Zero-Inflated Poisson Regression.

ABSTRACT

The classical linear regression analysis is an analysis aimed at knowing the relationship between the response variables and the explanatory variables assuming the normal distribution data, but in the applied data is often not the case. Generalized Linear Model (GLM) was developed for data in the form of categorical and discrete distribution. In this study the data was raised which has a poisson distribution by as much as *n*, with average λ and the odds appearing zero *p*. Poisson regression is GLM for Poisson-distributed data assuming that $Var(X) = E(X)$, but asusumption is rare in applied data. For rare occurrences of a specified interval *X* variables are often zero-valued, thus causing overdispersion ($Var(X) > E(X)$). Lambert (1992) introduced a method for overcoming overdispersion in poisson regression i.e. the Zero-Inflated Poisson regression (ZIP). In this research conducted a ZIP regression test in overcoming overdispersion to see the opportunity limit p appears zero- valued as the value that causes overdispersion. Testing is done with RStudio ver. 1.1.463.0 software. Based on the simulated data obtained that Regression ZIP stopped overcoming overdis persion at the condition $n = 500$, $\lambda = 0.7$ with the odds $p = 0.2$ with a dispersion ratio of $\tau = 1.010$.

To cite this article:

INTRODUCTION

Regression analysis in statistics is one method that can be used to determine cause and effect relationships between variables. Classical regression analysis has the condition that the data is normally distributed. This analysis aims to determine the direction of the relationship between the explanatory variable and the response variable as well as to predict the value of the response variable if the value of the explanatory variable increases or decreases. In practice in the field, the data found often do not meet the assumptions required by classical linear regression. To overcome this, a Generalized Linear Model (GLM) was developed. GLM is used as an extension of the general regression model for response variables in the form of categorical data and discrete distribution.

The response variable used in this research is poisson distribution. The poisson distribution is a discrete probability distribution that expresses the probability of the number of events occurring in a given time period with a known average of events occurring in independent time. So for the formation of the regression model can be used Poisson regression which is one of the special cases of the Generalized Linear Model (GLM).

Poisson regression analysis shows the relationship between the explanatory variable and the response variable with a poisson spread. The characteristic of the response variable from the poisson distribution is that the mean and variance have the same value or equidispersion. However, in ap plied data, the response variable has a large diversity, in other words, deviations often occur in the form of overdispersion or underdispersion.

Overdispersion is the variance that is greater than the average value, while underdispersion is the variance that is smaller than the average value of the response variable. Overdispersion can occur due to heterogeneity in the response variable (A. Agresti, 2007) as well as due to excess zeros.

Lambert (1992) introduced zero-inflated poisson regression (ZIP regression) as a model for han dling the overdispersion problem in data with excess zeros. The advantage of ZIP regression is that it is very easy to apply in several fields such as agriculture, animal husbandry, biostatistics, and industry. In addition, the ZIP regression model is easily interpreted by parameter estimators, and can explain the reason for the overdispersion of the response variables (D. Lambert,1992).

In a previous study (Dewanti et al., 2016) examined the comparison of zero-inflated poisson (ZIP) and zero-inflated negative binomial (ZINB) regression analysis which can overcome overdispersion because it does not have the assumption of equidispersion as in poisson regression [3]. ZIP regression analysis has been able to control the value of zero, but has not fully controlled overdis persion, so in this study will be studied about the ability of ZIP regression in overcoming overdis persion in poisson regression.

LITERATURE REVIEW

Poisson Regression

Poisson regression is a regression analysis that is usually used for data with responses in the form of discrete variables but not binary. In this case the data response is Poisson distribution with parameter λ. It is very important to note that this parameter λ is highly dependent on some particular unit or period of time, distance, area, volume, and so on. This distribution is then used to model an event whose existence is relatively rare or rare to occur in certain units. Poisson regression has the following assumptions [5]:

1. The response variable is discrete data.
2. The conditional distribution of the response variables follows the Poisson distribution.
3. The mean will be equal to the variance, E(Y) =Var(Y)

In the Poisson regression model, the connecting function used is the log link function because
the log function guarantees that the expected variable value of the response variable will be non-negative.

$$\eta_i = ln\ ln\ (\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i\ , i = 1,2,\cdots,n$$

The above equation can also be written as:

$$E(y_i) = \mu_i = exp\ exp\ (X_i\beta) + \varepsilon$$

Overdispersion

Poisson regression is said to contain overdispersion if the variance value is greater than the average value. Overdispersion has the same impact as the assumption violation if there is overdispersion in discrete data but Poisson regression is still used, the estimation of the regression coefficient parameters is consistent but not efficient. This has an impact on the standard error value which becomes an under estimate, so that the conclusion becomes invalid. The overdispersion phenomenon can be written as Var(Y) > E(Y). The relationship of the dispersion parameter (ϕ) with the variance and mean in the Poisson regression is:

$$\phi = \frac{Var(Y)}{\mu}$$

The calculation of the dispersion value using the Pearson Chi-Square is:

$$\phi = \frac{\chi^2}{df} = \frac{Pearson\ Chi - Square}{df}$$

Zero-Inflated Poisson (ZIP) Regression Model

The ZIP regression model is a simple mixed model for discrete data with many zero events. If it is an independent random variable with a ZIP distribution, then the zero value contained in the observation is thought to have occurred in two ways that correspond to separate states. The first state is called the zero state with probability and the second state is called the Poisson state with probability. Both states give the distribution of the two-component mixture. The probability function of the ZIP regression model is:

$$P(Y = y_i) = \{\pi_i + (1 - \pi_i)e^{-\lambda_i} \quad , y_i = 0 \ \frac{(1 - \pi_i)e^{-\lambda_i}\, {\lambda_i}^{y_i}}{y_i!} \quad , y_i > 0$$

In many applications there is little preliminary information about how relates π to λ. So to estimate the parameters use the link function:

$$log\ log\ (\lambda) = X\beta \ \ \text{and}\ (\pi) = -\tau X\beta$$

where the true value of the parameter is unknown, resulting in $\pi_i = (1 + \lambda_i^{\tau})^{-1}$

In generalized linear models equations, log(λ) and logit(π) are link functions or transformations that are generally used to linearize the Poisson mean and the Bernoulli probability of success. The ZIP model will then be written as ZIP (τ). The link function logit for parameter will be symmetric around the value 0.5. Two frequently used asymmetric link functions are log-log links, which are defined as:

$$log\ log\ (-\ log\ log\ (\pi)\ = \tau X\beta)\ \textit{ is equivalent to } \pi_i = exp\ exp\ (-\lambda_i^{\tau})$$

and a complementary log-log link which is defined as:

$$log\ log\ (-\ log\ log\ (1 - \pi)\ = -\tau X\beta)\ \text{ or } \pi_i = 1 - exp\ exp\ (-\lambda_i^{-\tau})\ .$$

Chi-Square Test on Poisson Regression and ZIP

The chi-square test is used to test the suitability of a set of data against a certain probability distribution. The chi-square test is used to test the suitability of the data set against the probability of the Poisson distribution and ZIP. In the chi-square test, the actual frequency in the category is compared with the theoretically expected frequency if the data follows the probability of the Poisson distribution and ZIP. The hypothesis of the chi-square test is

$$H_0 : p_l = p_l^0 \text{ and } H_0 : p_l \neq p_l^0$$

Then, the chi-square test statistic is the difference between the observed frequency and the theoretical frequency to the theoretical frequency of the Poisson and ZIP probability distributions.

$$\chi^2 = \sum_{l=0}^{m} \frac{(n_l - np_l)^2}{np_l}$$

Where pi is the probability mass function of the Poisson distribution and ZIP, ni is the observed frequency for each lth category, n is the sample size, and m is the number of categories. Thus the decision is to reject H_0 in α, is $\chi^2 > \chi^2_{\alpha,(m-p-1)}$. The rejection of H_0 on means that there is no match between the observed probability and the Poisson distribution probability or the response variable is not Poisson spread or ZIP is not spread.

Pearson Chi-Square Test on Poisson Regression and ZIP

The chi-square Pearson test is often used to measure the goodness of the Poisson and ZIP regression models. This test is carried out with the hypothesis that if the ratio produces a value of more than one, then the model experiences overdispersion on the alternative hypothesis (H_1). The hypothesis in the chi-squared Pearson test is

$$H_0: \tau = 1 \text{ and } H_1: \tau > 1$$

Pearson chi-square test statistic value can be defined as

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - E(Y)\right)^2}{Var(Y)}$$

with dispersion ratio

$$\tau = \frac{\chi^2}{n-k}$$

Under the condition that H_0 is true, the test statistic χ2 will approach the distribution with degrees of freedom (n−k), so the decision is to reject H_0 in α, if $\chi^2 > \chi^2_{\alpha,(n-k)}$,Identifying the diversity of data in the response variable (Y) to the Poisson regression and ZIP using the ratio criteria τ obtained from the statistical value of the chi-squared Pearson test to the degree of freedom from the Poisson regression and ZIP.

Wald's Test on Poisson Regression and ZIP

The Wald test is used to determine the explanatory variables that affect the response variable. Wald's test was applied to the Poisson and ZIP regression models. The hypothesis to test the significance of the Poisson and ZIP regression parameter coefficients, for example θ_i, is

$$H_0: \theta_i = 0 \text{ and } H_1: \theta_i \neq 0$$

The confidence interval in the Wald test for θ_i is$\widehat{\theta_i} \pm 1.96\, se\,(\widehat{\theta_i})$, with the test statistic used in the equation is

$$Z_i = \frac{\widehat{\theta_i}}{se\,(\widehat{\theta_i})}$$

Wald's test statistic on Z_i approaches the standard normal distribution when $\theta_i = \theta$. This situation is equivalent to Z^2 which is close to a chi-square distribution with degrees of freedom 1, so the Wald test statistic used is

$$W_i = \left[\frac{\widehat{\theta_i}}{se\,(\widehat{\theta_i})}\right]^2$$

where $(\widehat{\theta_i})$ is the parameter estimator coefficient θ_i and $se\,(\widehat{\theta_i})$ is the standard error estimator of the parameter estimator coefficient θ_i obtained from the variance estimator matrix (θ_i). The test statistic Wi will approach the distribution of χ2 with degrees of freedom 1 under the condition that H0 is true, so the decision is to reject H_0 in α, if $W_i > \chi^2_{\alpha,1}$. The rejection of H_0 on means that the i-th explanatory variable, for a certain i (i=1,2,...,k), has a significant effect on the response variable.

RESEARCH METHOD

This method the writer does through reading and taking data from books, articles or journals that support it to fulfill the theoretical basis in the analysis carried out. The data used in this research is simulation data. Simulation data is generated based on the characteristics of the data. Simulation data is useful for estimating the coefficients of the Poisson and ZIP regression parameters. The data will be simulated using the R programming language.

The simulation stages carried out in this study are:

1. Generating data response variable Y with Poisson distribution with λ = 0.7, 5, 10, 20 and the probability of appearing zero as p = 0.1, 0.2, 0.3,..., 0.9 much as n = 300, 400, 500.
2. Generating explanatory variables X which has a standard normal distribution.
3. Testing the fit of the data following the Poisson distribution using the chi-square test:
   (a) Record value Y along with the frequency.
   (b) Count the number of values Y.
   (c) Calculate the probability of each observation with the probability function Y:

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

   (d) Calculate the chi-square test statistic:

$$\chi^2 = \sum_{l=0}^{m} \frac{(n_l - np_l)^2}{np_l}$$

   (e) Do a test whether $\chi^2_{count} < \chi^2_{table}$
   (f) Calculate the percentage χ^2 that meets the test.
4. Testing the suitability of the data following the ZIP distribution using the chi-square test:
   (a) Record value Y along with the frequency.
   (b) Count the number of values Y .
   (c) Calculate the probability of each observation Y with the probability function:

$$P(Y = y_i) = \{\pi_i + (1-\pi_i)e^{-\lambda_i} \quad ,for\ y_i = 0 \ \frac{(1-\pi_i)e^{-\lambda_i}\,\lambda_i^{\,y_i}}{y_i!} \quad ,for\ y_i > 0$$

   (d) Calculate the chi-square test statistic:

$$\chi^2 = \sum_{l=0}^{m} \frac{(n_l - np_l)^2}{np_l}$$

   (e) Do a test whether $\chi^2_{count} < \chi^2_{table}$
   (f) Calculate the percentage χ^2 that meets the test.
5. Poisson regression model form.
6. Record the value of the coefficient on the poisson regression.
7. ZIP regression model form.
8. Record the coefficient values in the ZIP regression.
9. Calculate the dispersion ratio in poisson regression:
   (a) Calculate the poisson value:

$$\lambda_{pois} = exp\ exp\ (\beta_0 + \beta_1 X)$$

   (b) Calculate the chi-squared pearson test statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - E(Y)\right)^2}{Var(Y)}$$

   (c) Calculate the dispersion ratio:

$$\tau = \frac{\chi^2}{(n-k)}$$

   where (n−k) is degree of freedom

10.Calculate the dispersion ratio in poisson regression:

(a) Calculate value λ_i ZIP for discrete data:

$$\lambda_i = exp\ exp\ (\beta_0 + \beta_1 X)$$

(b) Calculate value p_i ZIP for zero-inflation:

$$p_i = \frac{exp\ exp\ (\beta_0 + \beta_1 X)}{1 + exp\ exp\ (\beta_0 + \beta_1 X)}$$

(c) Count $E(Y)$ and $Var(Y)$:

$$E(Y) = (1 - p_i)\lambda_i$$

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

(d) Calculate the chi-squared pearson test statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - E(Y)\right)^2}{Var(Y)}$$

(e) Calculate the dispersion ratio:

$$\tau = \frac{\chi^2}{(n - k)}$$

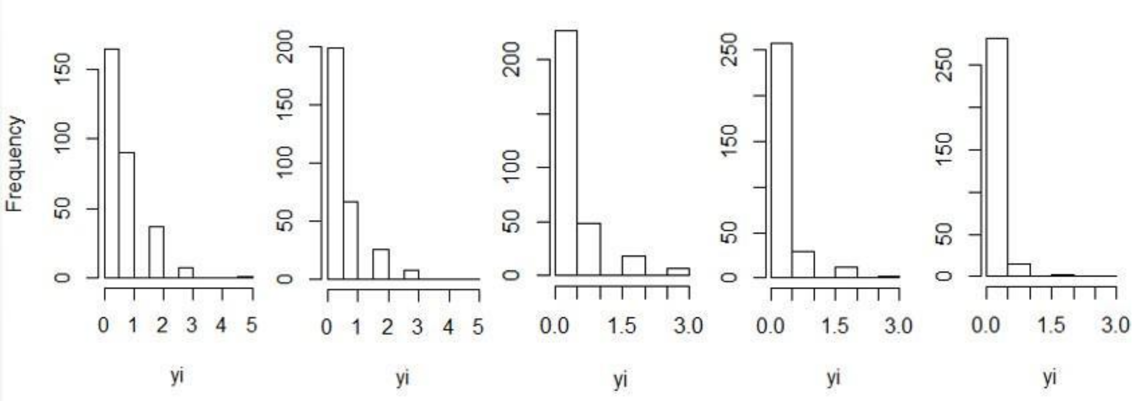
where (n−k) is degree of freedom

11. Perform the Wald test on Poisson and ZIP regression to see the relationship between the response variable and the explanatory variable.

RESULTS AND DISCUSSION

VARIABLE *Y* EXPLORATION SIMULATION STUDY

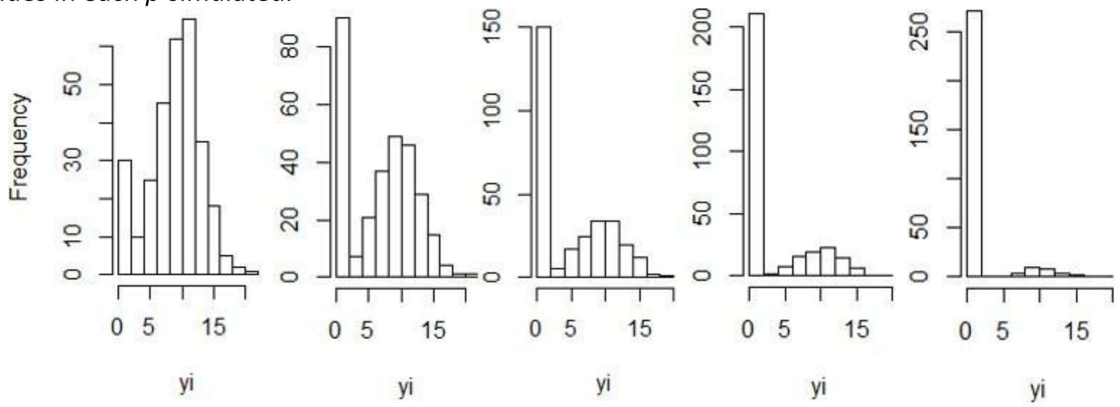
Variable *Y* is a response variable that contains data with a Poisson distribution. Data from the response variable is generated with several conditions. For λ = 0.7, 5, 10, 20 as many values n = 300, 400 and 500 with a probability of appearing a value of p = 0.1, 0.2, 0.3, ..., 0.9. Based on the simulation, it is shown that the value p has an effect on λ . Exploration of the variables *Y* will be shown through a histogram which aims to determine the condition of the Poisson distribution and ZIP on the variable *Y* .



$p = 0.1$ $p = 0.3$ $p = 0.5$ $p = 0.7$ $p = 0.9$

Figure 1. Histogram against the variable *Y* at *λ* = 0.7, *n* = 300 with *p* = 0.1, 0.3, 0.5, 0.7, 0.9

Histogram in Figure 4.1. shows that when the variable *Y* is valued *λ* = 0.7 with *p* = 0.1, 0.3, 0.5, 0.7 and 0.9, then the data indication is still Poisson spread. The histogram on *λ* = 0.7 shows that the mean of the variables is around the value of 0.7. When the variable *Y* has *p* = 0.1, then the mean changes to a value *λ* less than 0.7, namely *λ* = 0.643. In the condition of the *p* value from 0.1 to 0.9, there is a significant change in the value *λ* , meaning that the greater the value of *p*, the value *λ* goes to zero. Value *λ* = 0.7 and *p* = 0.7 and 0.9 on the variable *Y* indicates a zero excess chance. This condition corresponds to an increase in the frequency of zero values in each *p* simulated.



$p = 0.1$ $p = 0.3$ $p = 0.5$ $p = 0.7$ $p = 0.9$

Figure 2. Histogram against the variable *Y* at *λ* = 10, *n* = 300 with *p* = 0.1, 0.3, 0.5, 0.7, 0.9

Histogram in Figure 4.2. shows that the variables *Y* with *λ* = 10 in each *p* are simulated, it is indicated that the ZIP spread is indicated. The histogram on *λ* = 10 shows that the mean of the variable *Y* is around the value 10. The variable *Y* has many excess zeros in the event condition *λ* = 10, *p* = 0.5, 0.7, 0.9 so that the shape of the data distribution that occurs is the position of the zero value is separated from other values that are around the value of 10. The condition of the change in the p-value that is simulated *λ* = 10 in shows the same results as the condition of *λ* = 0.7. The results of the histogram indication will be tested using chi-squared, which shows that the conditions of the causes of overdispersion and the Poisson distribution on the variable *Y* due to excess zero or excessive zero values.

CHI-SQUARE TEST

The results of the chi-square test with *α* of 0.05 for the Poisson distribution and ZIP for the combination of *λ, n, p* are shown in Table 1. The chi-square test for the Poisson distribution shows that the larger *λ, n, p* simulated, the smaller the percentage of the Poisson distribution on the variable *Y* . The results of the chi-square test for the Poisson distribution are inversely proportional to the magnitude of the simulated *λ, n, p*. The chi-square test for the ZIP distribution shows that the ZIP regression is able to overcome the overdispersion caused by excess *p* in the variable *Y* . This condition is indicated by the greater the value of *λ* , the percentage of the Poisson distribution reaches 0% while the percentage of the ZIP distribution reaches the range of 60% to 80%.

Table 1. The percentage of the chi-square test against the combination *λ, n, p*

n	λ	0.1		0.2		0.3		0.4	
		Pois	Zip	Pois	Zip	Pois	Zip	Pois	Zip
300	0.7	76.4	85.6	20.8	87.2	0.4	86.6	0	86.6
	5	0	85.4	0	85.4	0	86.2	0	85.6
	10	0	82.6	0	82.2	0	83	0	80.4

	20	0	77.8	0	77	0	76.8	0	76
400	0.7	70.2	86	9	86.6	0	87.4	0	88.4
	5	0	86.4	0	87.2	0	86	0	85.4
	10	0	82.6	0	83.4	0	82	0	82.2
	20	0	76.8	0	77.2	0	77	0	77
500	0.7	70	84.8	5.2	86.6	0	87.2	0	86.6
	5	0	86.2	0	86	0	85.8	0	86.6
	10	0	81.8	0	81.8	0	82.4	0	83
	20	0	77.4	0	78.6	0	75.6	0	77.8

n	λ	0.5		0.6		0.7		0.8		0.9	
		Pois	Zip	Pois	Zip	Pois	Zip	Pois	Zip	Pois	Zip
300	0.7	0	83.8	0	84	0	83.8	0	84	0	82.4
	5	0	85.4	0	85.2	0	85.8	0	83.6	0	83
	10	0	80.6	0	79.8	0	79.4	0	77.6	0	75
	20	0	76	0	72.6	0	72.6	0	69.6	0	66.8
400	0.7	0	87.8	0	84.8	0	84.4	0	83.2	0	84.2
	5	0	86.8	0	85.6	0	85.2	0	85.8	0	83.6
	10	0	82.2	0	80.4	0	79.8	0	79.6	0	77.2
	20	0	78	0	77.4	0	72.6	0	70.8	0	68.2
500	0.7	0	87.8	0	88.2	0	83.8	0	83.6	0	84.4
	5	0	84.8	0	86.8	0	85.4	0	85.2	0	83.8
	10	0	81.4	0	82.2	0	80.6	0	79.8	0	78.2
	20	0	77	0	78	0	76	0	73	0	71.6

OVERDISPERSION OF POISSON REGRESSION AND ZIP

Exploration and testing of the variables showed that there was an indication of excess zero probability, so a chi-square test was carried out to determine whether the data had a Poisson distribution or ZIP. The simulation results for the combination of λ and p each n simulated show that when the variable Y has λ and p is larger, then there is overdispersion. Tests on the variables Y stated that the overdispersion condition affected the change in the Poisson distribution to the ZIP distribution. The chi-square test for the ZIP distribution shows that the ZIP regression is able to overcome the overdispersion caused by the excess zero value in the variable Y. Furthermore, Poisson regression and ZIP measured the goodness of the model based on overdispersion testing in each combination, λ, n and p which was simulated.

The overdispersion condition in each combination λ, n and p which is simulated in Poisson and ZIP regression can be known based on the ratio τ and the chi-squared Pearson test on α = 5%. The ratio shows the value of the chi-squared Pearson test statistic to the degree of freedom ($n{-}k$). The value of the degrees of freedom for the Poisson regression and ZIP is different, because the Poisson regression uses k = 2, which is the parameter estimator b_0 and b_1. ZIP regression uses k = 4 based on discrete models for λ and zero-inflation models for p i.e g_0 and g_1, and l_0 and l_1.

Table 2. Dispersion ratio to Poisson Regression and ZIP

n	λ	0.1		0.2		0.3		0.4	
		Pois	Zip	Pois	Zip	Pois	Zip	Pois	Zip
300	0.7	0.658	0.600	0.723	0.611	0.789	0.643	0.789	0.589
	5	0.959	0.658	1.257	0.650	1.602	0.656	1.879	0.639
	10	1.233	0.612	1.869	0.616	2.488	0.612	3.064	0.605
	20	1.811	0.603	3.010	0.606	4.186	0.601	5.390	0.601
400	0.7	0.856	0.803	0.929	0.804	0.991	0.804	1.079	0.814
	5	1.287	0.877	1.706	0.867	2.089	0.854	2.475	0.848
	10	1.640	0.820	2.431	0.802	3.287	0.809	4.148	0.814
	20	2.412	0.806	3.999	0.799	5.628	0.805	7.209	0.807
500	0.7	1.049	0.976	1.149	1.010	1.169	0.980	1.317	0.996
	5	1.566	1.066	2.116	1.079	2.599	1.052	3.146	1.064
	10	2.023	1.003	3.082	1.016	4.085	1.016	5.107	1.006
	20	3.007	0.999	5.033	1.007	7.006	1.006	9.034	1.005

n	λ	0.5		0.6		0.7		0.8		0.9	
		Pois	Zip	Pois	Zip	Pois	Zip	Pois	Zip	Pois	Zip
300	0.7	0.832	0.595	0.853	0.581	0.850	0.580	0.840	0.597	0.833	0.603
	5	2.212	0.645	2.470	0.639	2.731	0.633	3.038	0.614	3.363	0.590
	10	3.698	0.610	4.359	0.612	4.959	0.602	5.846	0.605	6.434	0.585
	20	6.656	0.605	7.864	0.606	9.005	0.599	10.642	0.603	12.311	0.598
400	0.7	1.192	0.805	1.123	0.791	1.182	0.776	1.189	0.774	1.161	0.814

	5	2.947	0.859	3.313	0.848	3.651	0.841	3.929	0.847	4.544	0.837
	10	4.930	0.812	5.711	0.808	6.641	0.817	7.496	0.803	8.676	0.789
	20	8.855	0.803	10.450	0.806	12.073	0.808	13.937	0.798	15.730	0.800
500	0.7	1.404	1.009	1.562	1.026	1.504	0.985	1.543	0.977	1.413	0.992
	5	3.591	1.047	4.178	1.066	4.677	1.058	5.049	1.051	5.405	1.007
	10	6.164	1.011	7.211	1.015	8.197	1.013	9.320	1.000	10.770	1.004
	20	11.066	1.009	13.048	1.006	15.089	1.009	17.314	1.002	19.736	1.002

Based on Table 2 it is known that the limit of the ZIP regression ability in overcoming overdispersion is in conditions *n* = 500*, λ* = 0.7 at probability *p* = 0.2 with a dispersion ratio *τ* = 1.010. In Table 2 it is shown that the ZIP regression stopped overcoming overdispersion at *p* = 0.2, but at *p* = 0.3 the dispersion ratio value *τ* = 0.980, based on the chi-square Pearson test in this condition there was no overdispersion. So that when *p* = 0.2 the simulation is carried out randomly without using the set.seed function 100 times the simulation.

Based on the simulation results of 100 runs for these conditions, the percentage of experiencing overdispersion is 59% so that it can be concluded that zero-inflated Poisson regression overcomes overdispersion under conditions of *n* = 500, *λ* = 0.7 at probability *p* = 0.2.

POISSON REGRESSION MODEL

Poisson regression analysis was used to evaluate the relationship between variable *X* and variable *Y* with a Poisson distribution. Poisson regression model estimation on variable X with Y when condition *n* = 100, *λ* = 20, *p* = 0.7 is as follows:

$$\lambda_i = exp\ exp\ (1.8209 - 0.1627X_i)\ .$$

The interpretation of the Poisson regression model for the *X* variable is significant to the *Y* variable, that is, every increase in the *X* variable will cause a decrease in the *Y* variable.

ZIP REGRESSION MODEL

ZIP regression analysis is an analysis that evaluates the relationship between variable *X* and variable *Y* that spreads ZIP. The ZIP distribution is caused by the increasing value of zero in the Poisson distribution. The ZIP regression model consists of two model components, namely the discrete data model for and the zero-inflation model for *p*. The ZIP regression model for the variable *X* with *Y* when the conditions are *n* = 100*, λ* = 20*, p* = 0.7 is as follows:

Discrete data Model for *λ:*

$$\lambda_i = exp\ exp\ (3.05452 - 0.01945X_i)$$

Zero-Inflation Model for *p:*

$$p_i = \frac{exp(0.8807 + 0.1975X_i)}{1 + exp(0.8807 + 0.1975X_i)}$$

So the y estimator in the ZIP regression is

$$\widehat{y_i} = (1 - p_i)\lambda_i$$

The interpretation of the discrete model for in the ZIP regression which is significant for the *Y* variable, that is, every increase in the *X* variable, it will cause a decrease in the number of occurrences of the *Y* variable by $e^{-0.01945} = 0.981 \approx 1$. Interpretation of the zero-inflation model for *p* at ZIP regression is significant for variable *Y* , that is, for every increase in variable *X* , the risk of occurrence of *y* increases $e^{0.1975} = 1.218$times.

WALD TEST

Wald's test statistic will approach the distribution with degrees of freedom 1, so the decision is to accept H_0 at α of 0.05, if $W_i < \chi^2_{\alpha.1}$ with a value of $\chi^2_{\alpha.1} = 23.68479$. The rejection of H_0 on the X variable to the Y variable means that the X variable does not have a significant effect on Y with α of 0.05. Acceptance of H_0 occurs in the Poisson and ZIP regressions.

CONCLUSION

Based on the results of calculations and analysis obtained the study of overdispersion of the sim ulated data from the simulated combinations *λ, n, p* shows that the larger the value *λ, n* and *p* the chi-square test, the smaller the percentage of the Poisson distribution. The simulation results show that the ZIP regression stops overcoming overdispersion under conditions *n* = 500*, λ* = 0.7 with *p* = 0.2. The Wald test performed on the simulation data showed that there was no effect between the variables *Y* with the Poisson distribution and the variables with the standard normal distribution.

REFERENCES

Agresti, *An Introduction to Categorical Data Analysis second edition*. New Jersey: Jon Wiley and Sons, 2007.

D. Lambert, "Zero-inflated poisson regression, with an application to defects in manufactur ing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

N. P. P. Dewanti, M. Susilawati, and I. G. A. M. Srinadi, "Perbandingan regresi zero inflated poisson (zip) dan regresi zero inflated negative binomial (zinb) pada data overdispersion (studi kasus: Angka kematian ibu di provinsi bali)," *E-Jurnal Matematika*, vol. 5, no. 4, pp. 133–138, 2016.

D. Downing and J. Clark, *Statistics the easy way*. Univ of California Press, 1997.

M. Pateta, *Fitting Poisson Regression Models Using the Genmod Procedure*, USA: SAS Institute Inc, 2005.

P. McCullagh and J. A. Nelder, *Generalized linear models. Chapman and Hall*. London, UK: Chapman and Hall, 1989.

V. Ricci, "Fitting distributions with R, "*Contributed Documentation available on CRAN*, vol. 96, pp. 1–24, 2005.

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

M. H. Degroot and M. J. Schervish, *Probability and Statistics Fourth Edition*. Boston: Pearson Education In, 2012.

I. M. Nur, et al, “Penerapan Generalized Poisson Regression I untuk Mengatasi Overdispersi pada Regresi Poisson (Studi Kasus: Pemodelan Jumlah Kasus Kanker Serviks di Provinsi Kalimantan Timur)”, *Jurnal Eksponensial,* vol. 7, no. 1, pp. 59-77, 2016.

A. Lestari, et al, “Pemodelan Regresi Zero-Inflated Poisson (Aplikasi pada Data Pekerja Seks Komersial di Klinik Reproduksi Putat Surabaya)”, *Phytagoras,* vol. 5, no. 2, pp. 57-72, 2009.

L. P. Rahayu, “Kajian Overdispersi pada Regreasi Poisson dan *Zero-Inflated* Poisson untuk Beberapa Karakteristik Data [tesis]”. Bogor: Institut Pertanian Bogor, 2014.

D. Karlis and I. Ntzoufras, “Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R,” *Journal of Statistical Software*, vol. 14, no. 10, pp. 1–36, 2005.