The Comparison Of Generalized Poisson Regression And Negative Binomial Reression Methods In Overcoming Overdispersion

Ayunanda Melliana, Yeni Setyorini, Haris Eko, Sistya Rosi, Purhadi

Abstract: Data on the number of cervical cancer cases are discrete data (count) which are usually analyzed with Poisson regression. The characteristics of the Poisson regression mean and variance must be the same, whereas in fact the count data is often becoming variance greater than the mean, which is often referred to over dispersion. To deal with the problem over dispersion, modelling can be done with Generalized Poisson Regression (GPR) and a Negative Binomial Regression because it does not require the mean value equal to the value of variance. Model GPR produces AIC value of 317.70. While the negative binomial regression models produced by AIC value 312.43. Then the best model is obtained from the negative binomial regression model because it produces the smallest AIC value.

Index Terms: cervical cancer; Generalized Poisson Regression; Negative Binomial Regression, AIC

1 PENDAHULUAN

POISSON regression is a nonlinear regression analysis of the Poisson distribution, where the analysis is highly suitable for use in analyzing discrete data (count) if the mean equal to the variance process. In fact, assuming a mean equal to the variance (equidispersion) rarely met while in general frequently encountered discrete data with variance greater than the mean (overdispersi) as occurs in the data the number of cervical cancer cases in East Java in this study. The implication of non-compliance is equidispersion of deviance value becomes very large models no longer appropriate Poisson regression to model the data. In addition, the models created will produce biased parameter estimates. The over dispersion down sides will also have consequences on the value of the standard error estimators for smaller which can further lead to inferential errors for the parameters. The appropriate anwalysis in this case is over dispersion of Generalized Poisson Regression (GPR) and negative binomial regression. The studies used in this study was to compare the best models of GPR and negative binomial regression, also determine the factors on the number of cervical cancer cases in East Java. Cervical cancer is cancer that occurs in the uterine cervix, the part of the female reproductive organs that are the entrance to the uterus located beteen the uterus and vagina. This research is carried out to obtain the best model by comparing the model obtained from analysis of GPR and Negative Binomial Regression. The result is also expected to provide additional information about any factors that significantly influence the occurrence of cervical cancer.

2 LITERATURE

2.1 POISSON REGRESSION

Poisson regression is a nonlinear regression analysis of the Poisson distribution, where the analysis is highly suitable for use in analyzing discrete data (count). Poisson regression model is a Generalized Linear Model (GLM) is the Poisson distribution is assumed response data [1], [2]. Poisson regression is said to contain over dispersion if the variance is greater than the value of the mean value. Over dispersion has the same impact with the assumption that if the offense discrete data occurred over dispersion but still used Poisson regression, the parameter estimates of the regression coefficients remain consistent but not efficient. This has an impact on the value of the standard error to under estimate, so that the conclusions become invalid. Over dispersion phenomenon [6] can be written as $Var(Y) > E(Y)$.

2.2 GENERALIZED POISSON REGRESSION (GPR)

The handling of equidispersion violations assumptions on the Poisson regression model is developed by using GPR. In the GPR models besides there are also parameter μ and θ as the dispersion parameter. GPR model is similar to the Poisson regression models, but assumed that the components are randomly distributed to general Poisson. In the analysis of GPR, if θ is equal to 0 then the model will be the model Poisson. If θ is more than 0 then GPR models represent data containing count over dispersion case and if θ is less than 0 represent data containing under dispersion count. The assessment of GPR parameter models is using MLE method. Log-likelihood function for the GPR model is

$$\ln L(\beta,\theta) = \sum_{i=1}^{n}\left\{y_i\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) - y_i \ln(1+\theta\exp(\mathbf{x}_i^T\boldsymbol{\beta})) + (y_i - 1)\ln(1+\theta y_i) + \Delta\right\}$$

with $\Delta = -\ln(y_i!) - \exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)(1+\theta y_i)(1+\theta\exp(\mathbf{x}_i^T\boldsymbol{\beta}))^{-1}$ (1)

To obtain parameter estimates β and θ the equation (1) is downgraded to β and θ using numerical methods, Newton-Raphson iteration.

2.3 NEGATIVE BINOMIAL REGRESSION (NBR)

In addition to GPR, over dispersion handling on Poisson regression can also be performed by using negative binomial models approach. In the negative binomial regression, if θ goes to zero then variable (Yi) goes to μi so will converge towards the negative binomial Poisson [3], [4]. Negative binomial regression models have the same form as the Poisson regression model Negative binomial regression parameter estimation performed by using MLE method. Log-likelihood equation for the negative binomial is.

$$\ln L(\theta,\beta) = \sum_{i=1}^{n}\left\{\left(\sum_{j=0}^{y_i-1}\left(j+\theta^{-1}\right)\right) - \ln y_i! - (y_i+\theta^{-1})\times\Delta\right\}$$

with $\Delta = -\ln(1+\exp(\mathbf{x}_i^T\boldsymbol{\beta}) + y_i \ln\theta + y_i \mathbf{x}_i^T\boldsymbol{\beta}$

2.4 AKAIKE INFORMATION CRITERION

AIC is one of the criteria in determining the best model [5] as follows.

$$AIC = -2\ln L(\tilde{\theta}) + 2k$$

Where $L(\hat{\theta})$ is *likelihoo valued*, and *k* the number of parameters. The best model is the model that has the smallest AIC value.

2.5 CERVICAL CANCER

Cervical cancer is cancer that occurs in the uterine cervix, the part of the female reproductive organs that are the entrance to the uterus located between the womb with a intercourse hole / vaginal [6]. The most common symptoms of cervical cancer are abnormal bleeding from the vagina or spots (blotches) vagina. This is especially true abnormal bleeding after sexual intercourse, but can also appear between 2 cycle menstrual bleeding, menorrhagia, or spotting / bleeding post menopause. If the bleeding lasts for a long time then the patient will complain of fatigue and weakness due to anemia was experiencing. Spotting is followed by a watery yellowish fishy smell can be a sign of malignancy [7]. Symptoms usually only appear when abnormal cells are transformed into malignant and infiltrate the surrounding tissue. Cervical cancer can spread to various organs, such as to the lymph nodes, vagina, bladder, rectum, endometrial (uterine lining), and ovary (ovarian). Each of them is providing different symptoms. Spread of cervical cancer in general circulation through the lymph nodes, blood circulation through rare [8].

3. RESULT AND CONCLUSION

3.1 ANALYSIS STAGE

To get the best modelling, some measures are used as follows: (1) Analyzing the correlation between the predictor variables to detect any cases of multicollinearity. (2) Getting the best model for the Poisson regression modelling of the number of cervical cancer cases. (4) Detecting the presence of over dispersion on the data by looking at the value of Pearson Chi-squares and Deviance divided by degrees freely. (5) Getting the best model for GPR on modelling the number of cervical cancer cases is to: a) Estimate the model parameters GPR, b) Test the significance of the model parameters simultaneously and partially GPR, c) Calculate the value of AIC of GPR models. (6) Getting the best model for the negative binomial regression modelling of the number of cases of cervical cancer case is to: a) Estimate the parameters of the negative binomial regression model, b) Test the significance of the negative binomial regression model parameters simultaneously and partially, c) Calculate the AIC values of the negative binomial regression model. (7) Comparing the AIC value from the GPR and the resulting negative binomial regression to obtain the best models. The best model is the model with the smallest AIC value.

3.2 MULTICOLINIERITY TEST

Before performing the analysis with the three methods that will be used is the Poisson regression, GPR, and negative binomial regression testing should be conducted multicollinearity of the data used. The Multicollinearity test is needed to be done as an initial assumption for parameter estimation. The criteria that can be used to detect one case of multicollinearity is by looking at the value of VIF. Multicollinearity occurs if the VIF value is greater than 10. VIF value of each predictor variable can be seen in Table 1.

Table 1. VIF value of 8 predictor variables

Variable	VIF
X_1	1,391
X_2	4,151
X_3	2,064
X_4	3,392
X_5	1,915
X_6	2,472
X_7	2,107
X_8	3,585

Tabel 1 shows that all predictor variables have VIF values <10. Thus all the variables can be included in the subsequent analysis modelling with Poisson regression, GPR, and Negative Binomial Regression.

3.3 THE MODELLING OF CERVICAL CANCER NUMBER CASES USING POISSON REGRESSION

The data rates of cervical cancer cases are a data count, where this type of data follows the Poisson distribution. The modelling with Poisson regression analysis is conducted to determine the factors that influence the number of cervical cancer. MLE method is used to obtain the estimation of the Poisson regression model parameters as shown in Table 2 and the resulting AIC value of 1942.2.

Tabel 2. Poisson Regression Model Parameter Estimation

Parameter	Estimation	Standard Error	z	p-value
β_0	-27,8895	0,000348	-80217	<,0001
β_1	0,003123	0,000702	4,45	<,0001
β_2	0,06991	0,004963	14,08	<,0001
β_3	0,5484	0,01991	27,54	<,0001
β_4	-0,1788	0,009921	-18,02	<,0001
β_5	0,1644	0,007918	20,77	<,0001
β_6	0,1604	0,01266	12,67	<,0001
β_7	-0,3012	0,01192	-25,27	<,0001
β_8	0,0001	0,000012	8,51	<,0001

Table 2 shows that the 15% significance level can be seen that the p-value of all the parameters is smaller than 0.15. so that the parameters β_1, β_2, β_3, β_4, β_5, β_6, β_7, and β_8 and significant effect on the model. Then Poisson regression models generated by.

$$\hat{\mu} = \exp(-27{,}8895 + 0{,}0031X_1 + 0{,}06991X_2 + 0{,}5484X_3 - 0{,}1788X_4 + \Delta)$$

With $\Delta = 0{,}1644X_5 + 0{,}1604X_6 - 0{,}3012X_7 + 0{,}001X_8)$

Variabel prediktor yang berpengaruh terhadap jumlah kasus kanker Predictor variables that influence the number of cervical cancer cases in East Java is the number of health facilities (X_1), the percentage of the female population aged ≤ 17 years of first marriage (X_2), the percentage of the female population (X_3), the percentage of poor (X_4), the percentage of

the population women who use contraceptives (X_5), the percentage of women with number of children born more than 4 (X_6), the percentage of women aged ≥ 35 years (X_7), and the average expenditure on food consumption of meat (X_8).

3.4 Overdispersion Test

In the Poisson regression analysis using, there is assumptions contained equidispersion that the mean value of the variance must be met. However, this assumption is rarely fulfilled so that it appears the case over dispersion. For detecting the over dispersion, can be seen from the value of deviance / db or Pearson / db. If the value of deviance / or Pearson db / db is greater than 1, it can be said to be a case of over dispersion whereas if it is less than 1 then there under dispersion.

Table 3. Value of Deviance and Pearson Poisson Regression Model

Criterion	value	df	value/df
Deviance	1791,256	28	63,9734
Pearson Chi-Square	2682,0817	28	95,7886

Table 3 shows that the value of deviance / db and Pearson chi-square/db greater than 1 so it can be concluded on the Poisson regression models the number of cervical cancer cases in East Java occurred overdispersion. Overdispersion led to models created will produce biased parameter estimates. Over dispersion presence will also have consequences on the value of the standard error estimator for the smaller (underestimate) which subsequently can lead to errors in inference for the parameters. To overcome this, the modelling is done using the Generalized Poisson regression (GPR) and negative binomial regression, where both methods can accommodate the dispersion parameter.

3.5 The Modelling of Number of Cervical Cancer Cases Using Generalized Poisson Regression (GPR)

GPR is one model that can be used to overcome over dispersion on Poisson regression. Before making models, a step that must be done is the estimation of parameters and testing parameters simultaneously and partially. GPR assessment of the model parameters shown in Table 4 and the smallest AIC value is obtained from a combination of six variables to the model GPR of 317.7.

Table 4. The Parameter Estimation of GPR Model

Parameter	Estimation	Standard Error	Z	P-Value
β_0	-1249,09	0,0294	-42484	<,0001
β_1	2,3986	0,5072	4,73	<,0001
β_2	0,6871	1,7987	0,38	0,7046
β_3	11,8043	3,2124	3,67	0,0008
β_5	3,3261	2,135	1,56	0,1278
β_6	4,6565	1,6974	2,74	0,0093
β_8	0,01719	0,002038	8,43	<,0001
θ	0,4117	0,07723	5,33	<,0001

From Table 4 with a significance level of 15%, it can be seen that the p-value is smaller for around 0.15 and the parameter are β_1, β_3, β_5, β_6, and β_8.

$$\hat{\mu} = \exp(-1249{,}09 + 2{,}3986X_1 + 11{,}8043X_3 + 3{,}3261X_5 + \Delta)$$

with $\Delta = 4{,}6565X_6 + 0{,}01719X_8$

Predictor variables that influence the number of cervical cancer cases in East Java is based on GPR models which are generated based on the number of health facilities (X_1), the percentage of the female population (X_3), the percentage of women who use contraceptives (X_5), the percentage of women with number of children born more of 4 (X_6), the average expenditure for meat consumption (X_8).

3.6 The Modelling of Number of Cervical Cancer Cases Using Negative Binomial Regression (NBR)

Another way to handle over dispersion on the Poisson regression model aside of using GPR, it can also uses a negative binomial regression model. The model with the smallest AIC value is the best model. In Table 5, it is shown the possibility of a negative binomial regression model with the smallest AIC value for each combination of variables ranging from one to six combinations of predictor variables using a significance level of 15%.

Table 5. Possible Negative Binomial Regression Model of Variable Combination

Y with Xi	AIC	Significant Parameter
X_7	310,06	β_0, β_7
X_4 X_7	310,41	β_0, β_7
X_3 X_5 X_7	311,39	β_5, β_7
X_2 X_4 X_6 X_8	312,43	β_2, β_6, β_8
X_1 X_2 X_4 X_6 X_8	314,29	β_2, β_6, β_8
X_1 X_2 X_3 X_4 X_6 X_8	316,41	β_6

Tabel 5 shows the model that has the most significant parameter and the smallest AIC value in each combination of predictor variables. The combination is a combination of four predictor variables that have a more significant parameter with the smallest AIC value than the combination of the five, and six predictor variables that is equal to 312.43. Thus the combination of the four predictor variables will be analyzed further to obtain a model of BNR. Parameter estimation results of negative binomial regression models shown in Table 6.

Table 6. Parameter Estimation of Negative Binomial Regression Model

Parameter	Estimation	SE	Z	P-*Value*
β_0	-0,0726568	23,068417	-0,031	0,9749
β_2	0,0740198	0,0420631	1,760	0,0785
β_4	-0,1201200	0,0882950	-1,360	0,1737
β_6	0,1603880	0,0998950	1,606	0,1084
β_8	0,0001464	0,0001012	1,447	0,1479
Θ	0,2633	0,0636		

Based on Table 6 with a significance level of 15%, it can be seen that the p-values of all parameters are smaller than 0.15 except for β0 dan β4 parameter. The parameter of θ is 0.2633 or greater than 0 that indicates overdispersion. So the negative binomial regression model was generated as follows.

$$\hat{\mu} = \exp(-0{,}0726 + 0{,}074X_2 + 0{,}1604X_6 + 0{,}0001X_8)$$

Predictor variables that influence the number of cervical cancer cases in East Java by a negative binomial regression model is the percentage of the female population aged ≤ 17 years of first marriage (X_2), the percentage of women with number of children born more than 4 (X_6), and the average expenditure on food consumption of meat (X_8).

3.7 CONCLUSION

Poisson regression model comparisons, GPR, and negative binomial regression was conducted to determine a better model used in modelling the number of cases of cervical cancer each district / city in East Java. The criteria for selection of the best model used is AIC. Best model is the model that has the smallest AIC value.

Table 7. The Choosing of best model

Model		Significant Variable	AIC
GPR		X_1 X_3 X_5 X_6 X_8	317,7
Regresi Negatif	Binomial	X_2 X_6 X_8	312,43

Based on AIC values in Table 7, the smallest AIC value is a negative binomial regression model. Then the best model for the number of cases of cervical cancer is obtained from the negative binomial regression model. This suggests that the negative binomial regression model is more appropriate in the case overdispersion poisson regression. With the resulting model is as follows:

$$\hat{\mu} = \exp(-0{,}0726 + 0{,}074X_2 + 0{,}1604X_6 + 0{,}0001X_8)$$

The models can be explained that every 1% increase in the population of women who first married age ≤ 17 years it will increase the number of cervical cancer cases in East Java by exp (0.06991) ≈ 1. Each 1% increase in the number of women with children who are born more than 4 then it will increase the number of cervical cancer cases in East Java by exp (0.1604) ≈ 1. Each increase in 1 rupiah average expenditure on food consumption of meat per month will reduce the number of cervical cancer cases in East Java by exp (0.0001) ≈ 1.

ACKNOWLEDGMENT

The author wishes to thank profusely to all those who have helped in this study. The author also thank to the East Java Provincial Health Office who have assisted in the provision of variable data response, the Central Bureau of Statistics who has assisted in providing data predictor variables.

REFFERENCES

[1] Hardin JW dan Hilbe JM. (2007). Generalized Linier Models and Extensions. Texas: A Strata Perss Publication.

[2] McCullagh P dan Nelder JA. 1983. Generalized Linier Models. London: Chapmann and Hall.

[3] Greene, W., Functional Forms For The Negative Binomial Model For Count Data. Foundations and Trends in Econometrics. Working Paper, Department of Economics, Stern School of Business, New York University, 2008: 585-590.

[4] Cameron, A.C. dan Trivedi, P.K. (1998). Regression Analysis of Count Data. Cambridge: Cambridge University Press.

[5] Bozdogan, H. (2000). Akaike's Information Criterion and Recent Developments in Information Complexity, Mathematical Psychology, 44, 62-91.

[6] Ferlay J. G (2002). Cancer Incidence, Mortality and Prevalence Worldwide. Lyon: IARC CancerBase, 2004.

[7] Mansjoer, A, Triyanti, K, dan Savitri, R. (2008). Kapita Selekta Kedokteran. Edisi ketiga. Fakultas Kedokteran Universitas Indonesia :Media Aesculapis.

[8] Prawirohardjo, S. (2008). Ilmu Kandungan. Second Edition. Jakarta : PT Bina Pustaka.