KnE Social Sciences

IAPA 2022
IAPA 2022 International Conference and International Indonesia Conference on Interdisiplinary Studies (IICIS)
Volume 2023

Knowledge E
enriching | engaging | empowering

Research Article

Model Comparison of Covid-19 in Lampung Indonesia

Tika Widayanti[1,4]*, Khoirin Nisa[2] and Tiyas Yulita[3]

[1]Departement of Geomatics Engineering, Institut Teknologi Sumatera, Indonesia
[2]Departement of Mathematics, University of Lampung, Indonesia
[3]Departement of Cyber Security, National Cyber and Crypto Institute, Indonesia
[4]Doctoral Program, Faculty of Science. University of Lampung, Indonesia

Abstract.
The COVID-19 has been pandemic in the world and has resulted in so many deaths due to being infected by the virus, therefore this study aims to determine a suitable model in estimating the number of deaths due to being infected by the COVID-19. This study focus in Lampung, Indonesia. The analysis of the death number was using three methods, there waspoisson regression, negative binomial regression (NBR), and generalized poisson regression (GPR). From the results, three predictor variables have significant effect to the model, there was positive cases of COVID-19, number of poor people, and life expectancy, while population density per km^2 has no significant effect. The best estimation model has smallest AIC and BIC values, and the poisson regression method is the best among other methods.

Keywords: Poisson Regression, NBR, GPR, COVID-19

Corresponding Author: Tika Widayanti; email: tikawidayanti@gt.itera.ac.id

Published 03 March 2023

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the IAPA 2022 Conference Committee.

1. Introduction

Poisson regression is a regression model in which the response variable is not normally distributed and is in the form of count data in an event. One of the assumptions underlying the Poisson regression is that the variance of the response variable must be equal to the mean. One of the development of Poisson regression is Generalized Poisson regression (GPR), which GPR is able to overcome if one of the assumptions in the Poisson regression is not fulfilled.

GPR is an extended method of Poisson regression which is able to overcome over / under dispersion, that is, if the variance value is equal to the average value, the dispersion parameter value is equal to zero, if the variance value is greater than the average value then the parameter value is greater from zero (overdispersion), but if the value of the variance is smaller than the average, the dispersion value is less than zero (underdispersion).

Research on the handling of over / under disperse using GPR, among others, Poisson analysis in clinical research [1], modeling count data [2], Covid in Nigeria [3]. In addition to GPR, there is another method that can determine the predictive model in the data count, namely negative binomial regression (NBR). This method is a combination of the poisson and gamma distribution, where the gamma distribution can overcome the overdispersion in the Poisson regression, which basically does not assume the existence of equilibrium.

From the description above, the count data analysis by comparing the Poisson regression method, GPR and NBR is considered very suitable if it is used to determine the best model for the number of deaths due to the COVID-19. The cases analyzed were those that occurred in Lampung Province, Indonesia with several variables that influenced, among others, the number of positives infected with the COVID-19, population density, the number of poor people, and the life expectancy in 15 districts / cities.

2. Materials and Method

2.1. Poisson Distribution

Poisson distribution is the probability of a rare occurrence in a given period of time. If the average number of success events is μ then the Poisson distribution that states the probability of success x in a given time interval is,

$$p(x;\mu) = \frac{e^{-\mu}\mu^{x}}{x!}; \qquad x = 0, 1, 2, \ldots; \qquad e = 2.7183$$

where μ is the mean and variance of the response variable y. While Poisson regression is a nonlinear regression used for computational data in which the response variable y follows the Poisson distribution, the Poisson regression can be expressed as follows [4],

$$\mu_i = exp\left(x_i^T \beta\right)$$

with

$$x_i = \begin{bmatrix} 1 & x_{1i} & x_{21} & \cdots & x_{ki} \end{bmatrix}^T$$

and,

$$\beta = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix}^T.$$

2.2. Generalized Linear Model

The generalized linear model (GLM) follows the regression model for univariate data, which is a very common exponential family distribution. Included in the exponential family are the normal distribution, binomial, poisson, geometry, negative binomial, exponential, gamma and inverse. If $y_i$is the response variable with i = 1, 2, ..., n, then GLM is defined as follows,

$$g\left(\mu_i\right) = g\left[E\left(y_i\right)\right] = x_i'\beta$$

where$x_i$is the regressor variable vector and β parameter vector or regression coefficient. Each GLM has three components, namely the distribution of the response variable (error structure), the linear predictor involving the regressor or covariate variable, and the link function that connects the linear predictor with the mean of the response variable. Suppose the response variables are normally distributed, the linear regression is,

and the link function is an identity link,

$$g\left(a\right) = a \; or$$

$$E\left(y\right) = \mu$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Thus, the general linear regression model is GLM dependent on the selection of the g-link function, and GLM can include nonlinear models [5].

2.3. Generalized Poisson Regression

Generalized Poisson Regression (GPR) is a suitable model for data on the amount of over-spread or under-spread, overdispersion is a greater-than-average variance, while under-dispersion is less-than-average variance. So there are two parameters in GPR, namely μ as average and θ as dispertion parameter. Suppose y is the response variable, then GPR is as follows,

$$f\left(y;\mu;\theta\right) = \left(\frac{\mu}{1+\theta\mu}\right)^{y} \frac{\left(1+\theta y\right)^{y-1}}{y!} exp\left(\frac{-\mu(1+\theta y)}{1+\theta\mu}\right), \qquad y = 0, 1, 2, \ldots .n$$

for the mean and its variants, E (y) = μ and$var\left(y\right) = \left(1+\theta\mu\right)^2$. If θ is equal to 0 then the GPR model will be a normal Poisson model, if θ> 0 then the GPR model will be

excessive , whereas if θ <0 then the GPR model indicates that there is less diffusion, the GPR model has the same shape as the Poisson model [6], i.e

$$\mu_i = exp\left(x_i^T \beta\right),\ i = 1, 2,\ \ldots,\ nx_i = \begin{bmatrix} 1 & x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$$

$$\beta = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix} (12)$$

The function assumes that *g* is the mean of the response variable until the linear predictor is,

$$g\left(\mu_i\right) = \eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \cdots + \beta_k x_k = x_i' \beta. (13)$$

the g function is called the link function, while the relationship between the mean and linear predictors is

$$\mu_i = g^{-1}\left(\eta_i\right) = g^{-1}(x_i' \beta) (14)$$

the commonly used link function in Poisson distribution is identity link, i.e.,

$$g\left(\mu_i\right) = \mu_i = x_i' \beta (15)$$

where $E\left(y_i\right) = \mu_i = x'\beta$, if $\mu_i = g^{-1}\left(x_i'\beta\right) = (x_i'\beta)$. In addition, there are also link functions used in Poisson distribution, namely log links,

$$g\left(\mu_i\right) = ln\left(\mu_i\right) = x_i{}' \beta (16)$$

for this function, the relationship between the response of the mean variable and its linear predictor is,

$$\mu_i = g^{-1}\left(x_i'\beta\right) = e^{x_i{}'\beta}. (17)$$

The log link function is excellent for Poisson regression, this is because it can predict all values for the response variable. Estimating the parameters in Poisson regression using maximum probability, for example a random sample and observation of the reaction variables y and x as predictor variables, then the probability function is

$$\mathscr{L}\left(\beta; y\right) = \prod_{i=1}^{n} f_i\left(y_i\right) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i{}^{y_i}}{y_i!} = \frac{\prod_{i=1}^{n} \mu_i^{y_i} exp\left(-\sum_{i=1}^{n} \mu_i\right)}{\prod_{i=1}^{n} y_i!} (18)$$

where$\mu_i = g^{-1}\left(x_i{}'\beta\right)$.The function of the selected link is then maximized with the log-likelihood

$$ln\mathscr{L}\left(\beta; y\right) = \sum_{i=1}^{n} y_i ln\left(\mu_i\right) - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} ln(y_i!) \ (19)$$

the appropriate estimates for parameter **b** in the Poisson regression model are,

$$\hat{y}_i = \mu_i = g^{-1}\left(x_i' b\right) (20)$$

if using an identity link, the prediction equation becomes,

$$\hat{y}_i = \mu_i = g^{-1}\left(x_i' b\right) = x_i' b (21)$$

therefore,

$$\hat{y}_i = \mu_i = g^{-1}\left(x_i' b\right) = exp(x_i' b) (22)$$

conclusions on models and parameters follow several approaches using logistical regression. To determine the merits of the model, Pearson chi-square deviance model and statistics were used [5].

2.4. Negative Binomial Regression

Negative binomial is the result of a combination of the poisson and gamma distribution, where the Poisson model has heterogeneity and gamma with a mean of 1.The function of the negative binomial distribution is [7],

$$f\left(y; \mu, \alpha\right) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i + 1)\Gamma\left(1/\alpha\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha}\left(1 - \frac{1}{1+\alpha\mu_i}\right)^{y_i} (23)$$

Whereas for the form of the parameter coefficient of the negative binomial log-likelihood is,

$$\mathscr{L}\left(\beta_j; y; \alpha\right) = \sum_{i=1}^{n} i ln\left(\frac{\alpha\, exp(x_i'\beta)}{1+\alpha\, exp(x_i'\beta)}\right) - \frac{1}{\alpha} ln(1 + \alpha exp\left(x_i'\beta\right)\ ) (24)$$

$$+ ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - ln\Gamma\left(y_i + 1\right) - ln\Gamma\left(\frac{1}{\alpha}\right) (25)$$

2.5. AIC and BIC

Two criteria regarding a better model are indicated by the value of Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC), if the value of both has a lower value then the model is the best model. AIC was developed by HorotsugaAkaike in 1974 which has two forms of AIC statistics,

$$AIC = -\frac{2(\mathscr{L} - k)}{n} (26)$$

and

$$AIC = -2\mathscr{L} + 2k = -2(\mathscr{L} - k) (27)$$

Where $\mathscr{L}$ is the log-likelihood model and *k* the number of predictor variables including the intercept. Whereas BIC was developed by AndrianRaftery in 1986, BIC is based on the deviance value and is defined as follows,

$$BIC = -2\mathscr{L} + k ln(n)\ (28)$$

Where *k* is the number of predictor variables [7].

2.4. Multikolinearity Test

Variance Information Factors (VIF) are the values used to indicate the existence of between predictor variables, VIF can be calculated as part of the regression analysis process. VIF is formulated as follows,

$$VIF = \frac{1}{1 - R_j^2} (29)$$

With the interpretation that if the VIF value is equal to 1 then there is no correlation, if the VIF value is between 1 and 5 then the correlation is moderate, and if the VIF value is greater than 5 then there is a high correlation [8].

2.5. COVID-19

A virus that is currently endemic that can cause respiratory syndrome is referred to as COVID-19, or COVID-19 for short, this epidemic was first discovered in the city of Wuhan in China. On December 29, 2019, a group of pneumonia was reported, then the epidemic spread to become a worldwide pandemic in March 2020. The COVID-19 actually existed before. COVID-19 comes from Latin which means crown, because its shape when under an electronic microscope looks like a crown.

Common symptoms experienced when infected with COVID-19 are fever, fatigue and tuberculosis. Some patients may experience aches and pains, nasal congestion, runny nose, sore throat or diarrhea. These symptoms are mild and occur gradually, but some infected people do not show any symptoms. A large number of people (about 80%) recover from the disease without the need for special treatment. About 1 in every 6 people infected becomes seriously ill and has difficulty breathing.

People with medical problems such as diabetes, high blood pressure and heart disease and older are more susceptible to serious illness that requires medical treatment. Coronavirus spreads from person to person through small droplets from the nose or mouth that are spread when a person coughs, speaks or exhales. These drops then fall on objects touched by others. The person then touches the eyes, nose or mouth [9]. However, there are also studies showing that the virus can spread in free air [10].

3. Results and Discussion

In this research there were several variables, including response variables (total mortality due to COVID-19), and 5 predictor variables, namely x_1 (positive case of COVID-19 infection), x_2(population density per km^2), x_3 (number of poor people in thousands of people), and x_4 (life expectancy). The data used are cumulative data from the initial confirmation of COVID-19 infection until March 20, 2020 in Lampung Province, Indonesia. The data was obtained from the Department of Health and Lampung Central Bureau for Statistics.

Testing for the presence or absence of multicollinearity is to determine the value of the Variance Inflation Factor (VIF) for each variable, if the value is less than 5 then there is no multicollinear

Table 1: VIF Value.

Variable	x_1	x_2	x_3	x_4
VIF	10,242647	13,509245	1,405732	2,903073

From Table 1, it can be seen that the VIF values for x_1 and x_2 are more than 5, so that x_2 is shown to be multicollinear, so x_2 is omitted for further analysis.

Table 2: VIF Value (x_1,$x_3$and x_4).

Variable	x_1	x_3	x_4
VIF	2,334939	1,003252	2,339727

Once x_2 is removed and the VIF value is recalculated, the VIF value is obtained as shown in Table 2, it can be seen that all VIF values are below 5, this indicates that there is no multicholinierity, so further analysis can be carried out.

The results of calculations using the three methodspoisson regression, generalized poisson regression and negative binomial regression are shown in Table 3.

4. Conclusions and recommendations

Based on the result of analysis usepoisson regression, negative binomial, and GPR, there was found that the predictor variables x_1 (positive case of COVID-19 infection), x_3(number of poor people in thousands of people) and x_4 (life expectancy) has significant effect on the response variable. The best method for estimating the number of deaths due to the COVID-19 in the Indonesian, Lampung Province is the poison regression which has the smallest AIC and BIC values.

Table 3: Parameter Estimate, AIC and BIC the Models.

Poisson Regression				
Coefficients:	Estimate	Std. Error	z value	
(Intercept)	-109.84316	30.05423	-3.655	
x1		0.05893	0.00298	19.777
x3		-0.09167	0.03751 -	2.444
x4		1.61697	0.44758	3.613
AIC: 114.99				
BIC: 117,8255				
Generalized Poisson Regression				
Coefficients:	Estimate	Std. Error	z value	
(Intercept)	-9.05E+00	3.14E+00	-2.881	
x1		5.00E-04	2.65E-05	18.887
x3		6.53E-03	8.95E-04	7.29
x4		1.65E-01	4.53E-02	3.635
AIC: 123.77				
BIC: 126,6051				
Negative Binomial Regression				
Coefficients:	Estimate	Std. Error	z value	
(Intercept)	-9.50E+00	4.43E+00	-2.145	
x1		5.26E-04	6.99E-05	7.517
x3		6.09E-03	1.76E-03	3.469
x4		1.71E-01	6.42E-02	2.67
AIC: 119.09				
BIC: 122,6281				

References

[1] Kianifard, F. And Gallo, P. P. (1995) Poisson Regression Analysis in Clinical Research. *Journal of Biopharmaceutical Statistics*. Vol 5 (1) : 115-129.

[2] Maxwell, O., Mayowa, B., Chinedu, I. U. and Peace, A. (2018). Modeling Count Data; A Generalized Linear Model Framework. *American Journal of Mathematics and Statistics*. Vol 6 (6) : 179 – 183

[3] Adams, S. O., Bamanga, M. A., Olanrewaju, S. O., Yahaya, H. U. and Akano, R. O., (2020). ModelingCOVID-19 Cases in Nigeria Using Some Selected Count data Regression Models. *International Journal of Healthcare and Medical Sciences*. Vol 6 (4) : 64-73.

[4] Agresti, A. (2002). *Categorical Data Analysis second Edition.* John Wiley & Sons. New York.

[5] Montgomery, dkk. (2010). *Generalized Linear Models*. John Wiley & Sons. New Jersey.

[6] Famoye, F., Wulu., J. T., Singh, K. P. (2004). On the Generalized Poisson Regression Model with an Application to Accident Data. *Journal of Data Science*. Vol. 2 : 287-295.

[7] Hilbe, J. M. (2011). *Negative Binomial Regression*. Second Edition. Cambridge University Press. New York.

[8] Daoud, J. I. (2017). Multicollinearity and Regression Analysis. *Journal of Physics : Conf. Series* Vol. 949 (012009).

[9] WHO. (2020). https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[10] Morawska, L. and Cao, J. (2020). Airborne Transmission of SARS_CoV-2: The Word should Face the Reality. *Environmenta International.*Vol 139 (105730).

[11] Statistics of Lampung Province. (2021). https://lampung.bps.go.id/subject/30/kesehatan.htm.