International Journal of Electronics and Communications System
Volume 1, Issue 1, 25-32.
ISSN: 2798-2610
http://ejournal.radenintan.ac.id/index.php/IJECS/index

Modeling with generalized linear model on covid-19: Cases in Indonesia

Subian Saidi*
University of Lampung,
Lampung, INDONESIA

Netti Herawati
University of Lampung,
Lampung, INDONESIA

Khoirin Nisa
University of Lampung,
Lampung, INDONESIA

Article Info

Article history:

Received: March 16, 2021
Revised: May 25, 2021
Accepted: June 8, 2021

Keywords:

AIC,
Gamm,
Gaussian,
Generalized Linear Model,
Poisson.

Abstract

The ongoing Covid-19 outbreak has made scientists continue to research this Covid-19 case. Most of the research carried out is on the prediction and modeling of Covid-19 data. This study will also discuss Covid-19 data modeling. The model that is widely used is the linear model. However, if the classical assumption of normality is not met, a special method is needed. The method that can overcome this is the generalized linear model (GLM), with the assumption that the data is distributed in an exponential family. The distribution used in this study is the Gaussian, Poisson, and Gamma distribution. Where the three distributions will be compared to get the best model. The variables used in this study were the number of confirmed Covid-19 cases per day and the number of deaths due to Covid-19 per day. This study also aims to see how much influence the confirmation of Covid-19 has on the number of deaths due to Covid-19 per day. By using 3 types of exponential family distribution, the best result is the Gaussian distribution GLM. Selection of the best model using Akaike Information Criterion (AIC).

To cite this article: S. Saidi, N. Herawati, and K. Nisa, "Modeling with generalized linear model on covid-19: Cases in Indonesia," Int. J. Electron. Commun. Syst., vol. 1, no. 1, pp. 25–33, 2021

INTRODUCTION

In 2019-2020 there was an epidemic that attacked the world called covid-19 (coronavirus disease-19). The increase from day to day in the number of patients infected with the Covid-19 virus is already difficult to control. A clear and straightforward plan is needed from the government to tackle this problem [1]. Coronavirus disease 2019 (Covid-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus emerged in China in December 2019. Coronaviruses are a group of viruses from the subfamily Orthocronavirinae in the Coronaviridae family and the order Nidovirales. This group of viruses can cause disease in birds and mammals, including humans [2]. Since then, the virus has spread rapidly to various regions of the world. As of December 7, 2020, as recorded by WHO, there were 65.8 million confirmed cases of Covid-19 with 1.5 million of them died and Covid-19 has also spread to 220 countries or regions [3]. This Covid case has been declared a pandemic by WHO on March 11, 2020. In Indonesia alone, the number of Covid-19 cases is still increasing every day. As of December 10, 2020, there were 598,933 positive cases with 491,975 recoveries and 18,336 deaths [4].

The Covid-19 outbreak has brought changes to the economic structure of society [5]. The viruses have made many scientists and academics conduct various researches. However, the latter was an observational study without a formal control arm, and data on the time, as symptom onset were not reported. Using antibody-based therapy as early as possible after exposure to maximize their therapeutic effects is similar to the use of anti-hepatitis B virus or rabies immunoglobulin preparations [6]. Research that is often carried out is time series analysis and regression analysis on Covid-19 data. But in fact, in the regression analysis can be found variable y which is not normally distributed

• **Corresponding author:**
Subian Saidi, University of Lampung, Lampung, INDONESIA. ✉ subian.saidi@fmipa.unila.ac.id

[7]. If we continue to analyze with the assumption of normality, we will get poor analysis results. Therefore, a special analysis is needed to overcome this abnormality, one of which is GLM.

Research on GLM has been conducted also applied to Covid-19 data, as has been done by To et al (2021) on Covid-19 data in Canada and Rath (2020) on Covid-19 data in India[8], [9]. Ratt (2020) concluded that GLM is a good method for modeling and estimating Covid-19 data in India [9]. Therefore, the authors are interested in doing GLM modeling on Covid-19 data in Indonesia with the response variable being the mortality variable and the predictor variable being the confirmed status variable [10]–[18].

Generalized Linear Model (GLM)

GLM is a general form of the Linear Model. In the classical linear model, Y is assumed to be normally distributed with (Y)=μ and variance^2. In GLM, the Y response variable can be distributed other than normal but is included in the exponential family (Exponential Terms). As a transition from the Linear model to the Generalized Linear model, the shape is described through three components [19].

Random Component, namely the observed values of the Y response that are mutually independent of any particular distribution Systematic Component, which is a linear combination of the X variable with parameter denoted by =Xβ[20]. The link between random and Systematic/ link function, which is a function that explains the expected value of the response variable (Y) which connects with the explanatory variables through the equation linear [21].

The three components will determine the model to be used in GLMs. The simplest link function isg(μ) = is called the identity link[22]. When GLM has the simplest connecting function, then the GLM is a linear model with continuous response. Another connecting function will relate nonlinearly to the predictor.

Exponential Distribution Term.

Regression analysis whose response belongs to one of the exponential families is called the Generalized Linear Model or better known as GLM (Generalized Linear Models).Regression model describing the response variable in the form of categorical with the predictor variable [23]. Either categorical or continuous Generalized Linear Models (GLM) extend the usual regression model to include response variables that are not normally distributed and the model functions for mean [24]. GLM is the distribution of responses that has various types, which are included in the Exponential Family. A random variable Y, included in the distribution belonging to the exponential term, if it has the form:

$$f_Y(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \qquad (1)$$

In some cases, the functions a, b, c, and d may contain another parameter called Nusian/disturbance. Some types of distribution that are often used in GLM can be described as follows:

Normal / Gaussian Distribution

The normal (or Gaussian) distribution was first described by Carl Friedrich Gauss in 1809 in the context of measurement errors in astronomy. Usually, normal distributions are compared by putting them on the same scale to obtain the standard normal distribution [25]. The form of the probability density function of the random variable Y which has a Normal / Gaussian distribution is

$$f(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right), \quad -\infty < y < \infty \qquad (2)$$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right), \quad -\infty < y < \infty \qquad (3)$$

With

$$b(\theta) = \frac{\theta}{\sigma^2}, \quad d(y) = \frac{y^2}{2\sigma^2}, \quad c(\theta) = -\frac{\sigma^2}{2\sigma^2 - \frac{1}{2}\log(2\pi\sigma^2)} \qquad (4)$$

Here is the Nusian parameter. So,

$E[Y] = \theta$ and $Var[Y] = \sigma^2$

Poisson distribution

The random variable Y which has a Poisson distribution has a probability density function:

$$f(y) = \frac{\theta^y e^{-\theta}}{y!}, y = 0, 1, 2, 3, \ldots = \exp[y \log\theta - \theta - \log y!] \quad (5)$$

In equation 5,

$$b(\theta) = \log\theta, \quad c(\theta) = -\theta, \quad d(y) = -\log y \quad (6)$$

Therefore $E[Y] = \theta$ and $Var\ [Y] = \theta$

Gamma Distribution

The random variable Y which has a Gamma distribution has a probability density function.

$$f(y) = \frac{\theta(y\theta)^{\phi-1}e^{-y\theta}}{\Gamma(\phi)}, y > 0, \quad (7)$$

$$f(y) = \exp[-y\theta + (\phi - 1)\log y + \phi\log\theta - \log\Gamma(\phi)] \quad (8)$$

With

$$b(\theta) = -\theta, \quad a(y) = y, \quad c(\theta) = \phi\log\theta - \log\Gamma(\phi), \quad d(y) = (\phi - 1)\log y \quad (9)$$

so,

$$E(Y) = \phi/\theta\ , \quad Var[Y] = \phi/\theta^2 \quad (10)$$

From hereσ is the Nusian parameter

Characteristics of Exponential Terms

The exponential distribution terms as discussed above. An exponential function is a function with the basic form f (x) = a^x , where a (a fixed base that is a real, positive number) is greater than zero and not equal to 1 [26]. Each of them has special characteristics as follows.

1. The Gaussian distribution has the characteristic.
   a. Continuous scale with range -∞<y<∞
   b. Symmetrical
   c. Variance independent of the mean (constant variance)
2. Gamma distribution has the characteristics of
   a. Continuous scale with range 0<y<∞
   b. Not symmetrical
   c. The variance is quadratic with the mean =ϕμ^2
3. The Poisson distribution has the characteristic
   a. Discrete scale with range 0≤y<∞,y=0,1,2,⋯
   b. Not symmetrical
   c. The variance is linearly related to the mean =ϕμ
4. Bernoulli distribution (Binomial with n=1) has the characteristic
   a. Discrete scale with binary range y=0.1
   b. Symmetry depends on the value of p

Link Function

With the distribution of response data that does not always follow the Gaussian distribution, it means that the data range is also not always in the range of all real numbers, for example, the range of positive data is continuous, whole, or binary only. Meanwhile, a linear combination of predictors commonly referred to as linear predictors.$\eta = \sum_{ij=0}^{p} x_{ij}\beta_j$ is open to take any value of any real number. For any reason, we need a function that connects and simultaneously synchronizes the response with a linear predictor. This function is called the link function [27]. Thus, the link function simultaneously serves to maintain linearity so that the predictor remains linear and normality then the span between the linear predictors and y or y remains in sync. Among the link functions that can be used, there is a canonical link function, namely the relationship function that occurs when $b(\theta) = \eta = \sum_{j=0}^{p} \beta_j x_j$[28]

1. For the binomial distribution, for example, the functions that can be used are:
   a. the logit function, which is the canonical link function i.e.

$$\eta = \log\left(\frac{\mu}{1-\mu}\right) \quad (11)$$

   b. orbit function

$$\eta = \Phi^{-1}(\mu)$$

   Where Φ is the cumulative function of the Normal distribution, i.e.

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}z^2\right] dz \quad (12)$$

   c. log–log complementarity, i.e

$$\eta = \log[-\log(1-\mu)] \quad (13)$$

1. For a Gaussian distribution, the canonical *link* function is the identity $\mu_i = \eta_i$
2. For the Gamma distribution of reciprocal canonical link functions $\log\frac{1}{\mu_i} = \eta_i$ but the log link is also often used $\log(\mu_i) = \eta_i$
3. For a Poisson distribution, the canonical link function is $\log(\mu_i) = \eta_i$

Akaike Information Criterion (AIC)

In selecting the best model, this study uses the Akaike Information Criterion (AIC) [29]. The model has the smallest AIC value

$$AIC = -2\log L + 2P \quad (14)$$

Where: log L = the maximum value of the likelihood function of the Cox PH model regression,

P = number of independent variables in the Cox PH regression model[30].So, by using modeling with a generalized linear model, we aim to make a model on Covid-19 cases in Indonesia.

METHOD

The methods that will be used in this research are:

1. Collecting data from the covid19.go.id site
2. Selection of response variables and predictor variables (response variable is the number of patients dying from Covid-19 per day and the predictor variable is the number of confirmed Covid-19 statuses per day)
3. Descriptive statistics
4. Identify a suitable distribution
5. Linearity test
6. GLM modeling with 3 distributions
7. Selection of the best model
8. The estimated value of y guess

RESULTS AND DISCUSSION

Descriptive statistics

From the descriptive statistics table 1, it can be interpreted that the average confirmed Covid-19 patient is 2214 people and the average patient who died is 69 people. The least confirmed number of Covid-19 is 106 people in a day and the most confirmed number of Covid-19 is 6267 people in a day. Meanwhile, the number of patients who died due to Covid-19 was at least 7 people in a day and the number of patients who died from Covid-19 was at most 169 people in a day.

Table 1. Descriptive statistics

Variabel	Min	Quartil 1	Median	Mean	Quartil 3	Max
X	106	689	1853	2214	3737	6267
y	7,00	36,00	70,00	69,16	98,00	169,00

Identification of distribution

Before performing the generalized linear model analysis, it will first identify the distribution of the response variable y. To see the appropriate distribution, a goodness of fit test will be carried out. The goodness of fit test of several distributions is presented in table 2.

Table 2. Distribution test

Distribution	AD	P
Normal	1.715	<0.005
Lognormal	7.163	<0.005
Exponential	18.758	<0.003
2-Parameter Exponential	12.596	<0.010
Weibull	2.467	<0.010
3-Parameter Weibull	2.299	<0.005
Smallest Extreme Value	2.954	<0.010
Largest Extreme value	2.923	<0.010
Gamma	3.855	<0.005
Logistic	2.140	<0.005
Log-logistic	5.1743	<0.005

We could see from table 2, for all distributions have a p-value <0.05 so we reject H_0 and it can be concluded that the data does not follow one of these distributions.

Linearity Test

This linearity test, it will be done by looking at the scatter plot. With the help of the R program, the following graphic output is obtained:

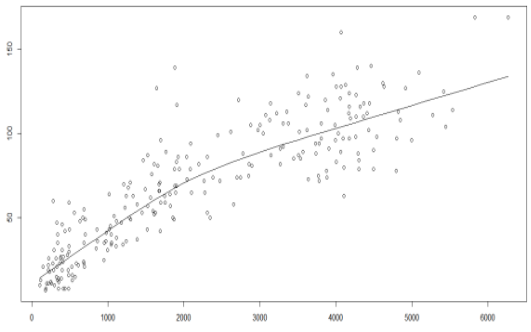

Figure 1. Linear plot of data

It can be seen from figure 1, that the data is relatively linear increasing. So it can be concluded that the data meet the assumption of linearity. Furthermore, the data will be analyzed with a generalized linear model.

Generalized Linear Model (GLM)

At this stage, GLM modeling will be carried out on three distributions of the exponential family, namely the Gaussian distribution (see table 3), the Poisson distribution (see table 4), and the Gamma distribution.

a. Gaussian

In GLM, this Gaussian distribution will use the link identity function. With the help of the R program, the following model is obtained:

$$y = 23{,}06763 + 0{,}02082\, x \qquad (11)$$

Table 3. Gaussian output

	Std. Error	t-value	Pr(>|t|)
β_0	2.031	11.36	<0,00000002
β_1	0.0007477	27.84	<0,00000002

From table 3, the column Pr(>|t|) can be seen that the variable x has a significant effect on y with a p-value <0.05

b. Poisson

In this Poisson distribution, GLM will use the link log function. With the help of the R program, the following model is obtained:

$$y = 3.489 + 0{,}0002892\, x \qquad (12)$$

Table 4. Poisson Output

	Std. Error	z-value	Pr(>|z|)
β_0	0.01614	216.18	<0,0000002
β_1	0.000004795	60.31	<0,0000002

From table 4, the Pr(>|t|) column can be seen that x has a significant effect on the y variable because the p-value <0.05

c. Gamma

In this Gamma distribution, GLM will use the link log function. With the help of the R program, the following model is obtained:

$$y = 3.3420523 + 0{,}0003462\, x$$

Table 5. Gamma Output

	Std. Error	t-value	Pr(>|t|)
β_0	0.0426629	78.34	<0,000001
β_1	0.0000157	22.05	<0,000001

From table 5, the column Pr(>|t|) can be seen that x has a significant effect on the value of y with p-value <0.05

Based on the modeling of the three distributions above, it is found that the x variable has a significant effect on the y variable. However, we need to compare the three models above to get the best model.

Selection of the Best Model

From the three distributions that have been used, the AIC value of each distribution will be sought. The best model is the model of the distribution that has the smallest AIC value. Below is a table of AIC values obtained with the help of the R program (see table 6).

Table 6. AIC .value

Model	Distribution	AIC
1	Gaussian	2127
2	Poisson	3246
3	Gamma	2239.9

From table 6, the smallest AIC value is in model 1, namely GLM with a Gaussian distribution with an AIC value of 2127, so the best model is model 1 GLM with a Gaussian distribution.

Estimated Value of y Guessing

After obtaining the best model, then the estimation of the value of y per day will be carried out. By using the Gaussian distribution GLM model, the estimated y value is presented in the appendix.

CONCLUSION

The conclusions obtained from the analysis are as follows the number of confirmed cases of Covid-19 in Indonesia has a significant influence on the number of deaths due to Covid-19 in Indonesia. After modeling with three distributions and then doing a comparison between the three models, the best model is obtained is a model with a Gaussian distribution with an AIC value of 2127.

REFERENCES

[1] I. Wahidah, R. Athallah, N. F. S. Hartono, M. C. A. Rafqie, and M. A. Septiadi, "Pandemik COVID-19: Analisis perencanaan pemerintah dan masyarakat dalam berbagai upaya pencegahan," *J. Manaj. dan Organ.*, vol. 11, no. 3, pp. 179–188, 2020, doi: 10.29244/jmo.v11i3.31695.

[2] N. R. Yunus and A. Rezki, "Kebijakan pemberlakuan lockdown sebagai antisipasi penyebaran corona virus Covid-19," *Jurmal Sos. Budaya Syar'i*, vol. 7, no. 3, pp. 227–238, 2020, doi: 10.15408/sjsbs.v7i3.15048.

[3] L. D. Rampal and B. S. M. Liew, "Coronavirus disease coronavirus disease (COVID-19) spreads," *Who*, vol. 75, no. 2, pp. 95–97, 2020.

[4] D. Prastiwi, "Update Corona 10 desember 2020: 598.933 positif, sembuh 491.975, meninggal 18.336," *liputan6.com*, 2020. https://www.liputan6.com/news/read/4430175/update-corona-10-desember-2020-598933-positif-sembuh-491975-meninggal-18336.

[5] S. Syafrida and R. Hartati, "Bersama melawan virus covid 19 di Indonesia," *SALAM J. Sos. dan Budaya Syar-i*, vol. 7, no. 6, pp. 495–508, 2020, doi: 10.15408/sjsbs.v7i6.15325.

[6] A. Gharbharan *et al.*, "Effects of potent neutralizing antibodies from convalescent plasma in patients hospitalized for severe SARS-CoV-2 infection," *Nat. Commun.*, vol. 12, no. 1, pp. 1–12, 2021, doi: 10.1038/s41467-021-23469-2.

[7] W. Trisunaryanti, T. Triyono, M. Mudasir, and A. Syoufian, "Multiple regression analysis of the influence of catalyst characters supported on γ-Al_2O_3 towards their hydrocracking conversion of asphaltene," *Indones. J. Chem.*, vol. 4, no. 1, pp. 6–11, 2010, doi: 10.22146/ijc.21868.

[8] T. To *et al.*, "UV, ozone, and COVID-19 transmission in Ontario, Canada using

generalised linear models," *Elsevier Public Heal. Emerg. Collect.*, 2021.

[9] S. Rath, A. Tripathy, and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 5, pp. 1467–1474, 2020, doi: 10.1016/j.dsx.2020.07.045.

[10] C. Bolancé and R. Vernic, "Multivariate count data generalized linear models: Three approaches based on the Sarmanov distribution," *Insur. Math. Econ.*, vol. 85, pp. 89–103, 2019, doi: 10.1016/j.insmatheco.2019.01.001.

[11] S. Kawano, H. Fujisawa, T. Takada, and T. Shiroishi, "Sparse principal component regression for generalized linear models," *Comput. Stat. Data Anal.*, vol. 124, pp. 180–196, 2018, doi: 10.1016/j.csda.2018.03.008.

[12] N. D. Nordin, M. S. D. Zan, and F. Abdullah, "Generalized linear model for enhancing the temperature measurement performance in Brillouin optical time domain analysis fiber sensor," *Opt. Fiber Technol.*, vol. 58, no. May, 2020, doi: 10.1016/j.yofte.2020.102298.

[13] R. E. Chandler, "Multisite, multivariate weather generation based on generalised linear models," *Environ. Model. Softw.*, vol. 134, no. September, p. 104867, 2020, doi: 10.1016/j.envsoft.2020.104867.

[14] J. George, J. Letha, and P. G. Jairaj, "Daily rainfall prediction using generalized linear bivariate model – a case study," *Procedia Technol.*, vol. 24, pp. 31–38, 2016, doi: 10.1016/j.protcy.2016.05.006.

[15] A. Habahbeh, S. O. Fadiya, and M. Akkaya, "Factors influencing SMEs CloudERP adoption: A test with generalized linear model and artificial neural network," *Data Br.*, vol. 20, pp. 969–977, 2018, doi: 10.1016/j.dib.2018.07.012.

[16] A. Ubaidilah, A. Kurnia, and K. Sadik, "Generalized multilevel linear model dengan pendekatan bayesian untuk pemodelan data pengeluaran perkapita rumah tangga," *J. Apl. Stat. dan Komputasi Stat.*, vol. 9, no. 1, 2017.

[17] Z. Ummah, Suliyanto, and Sediono, "Estimasi model linier tergeneralisasi gaussian berdasarkan maximum likelihood estimator dengan menggunakan algoritma fisher scoring," *J. Mat.*, vol. 1, no. 1, pp. 110–120, 2012.

[18] A. Calabrese, J. W. Schumacher, D. M. Schneider, L. Paninski, and S. M. N. Woolley, "A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds," *PLoS One*, vol. 6, no. 1, 2011, doi: 10.1371/journal.pone.0016104.

[19] P. McCullagh and J. A. Nelder, "Glmbook.Pdf," *Chapman and Hall*. pp. 1–511, 1989.

[20] F. J. Carvalho, D. G. de Santana, and L. B. de Araújo, "Why analyze germination experiments using generalized linear models?," *J. Seed Sci.*, vol. 40, no. 3, pp. 281–287, 2018, doi: 10.1590/2317-1545v40n3185259.

[21] G. Cerda, C. Pérez, J. I. Navarro, M. Aguilar, J. A. Casas, and E. Aragón, "Explanatory model of emotional-cognitive variables in school mathematics performance: a longitudinal study in primary school," *Front. Psychol.*, vol. 6, no. September, pp. 1–10, 2015, doi: 10.3389/fpsyg.2015.01363.

[22] J. E. Cote, "Sociological perspectives on identity formation: The culture-identity link and identity capital," *J. Adolesc.*, vol. 19, no. 5, pp. 417–28, 1996, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9245295.

[23] A. Wibowo and M. R. Ridha, "Comparison of logistic regression model and MARS using multicollinearity data simulation," *JTAM | J. Teor. dan Apl. Mat.*, vol. 4, no. 1, p. 39, 2020, doi: 10.31764/jtam.v4i1.1801.

[24] N. S. U, D. Ispriyanti, and T. Widiharih, "Online di : http://ejournal-

s1.undip.ac.id/index.php/gaussian aplikasi model regresi poisson tergeneralisasi pada kasus angka kematian bayi di jawa tengah tahun 2007," vol. 2, pp. 361–368, 2013.

[25] R. G. Brereton, "The normal distribution," *J. Chemom.*, vol. 28, no. 11, pp. 789–792, 2014, doi: 10.1002/cem.2655.

[26] S. A. Mousel, "The exponential function expository paper shawn," *MAT Expo. Pap.*, 2006.

[27] C. Czado and A. E. Raftery, "Choosing the link function and accounting for link uncertainty in generalized linear models using Bayes factors," *Stat. Pap.*, vol. 47, no. 3, pp. 419–442, 2006, doi: 10.1007/s00362-006-0296-9.

[28] piet de jong and G. z. Heller, *www.GFX.0fees.net*. 2014.

[29] M. Fathurahman, "Pemilihan model regresi terbaik menggunakan metode akaike's information criterion dan schwarz information criterion," *J. Inform. Mulawarman*, vol. 4, no. 3, 2009.

[30] M. Fathurahman, "Pemilihan model regresi terbaik menggunakan akaike's information criterion," *J. EKSPONENSIAL*, vol. 1, no. 2, pp. 26–33, 2010.