Count Data Modeling Using GAMLSS Approach and Its Application in Dengue Hemorrhagic Fever Cases in East Kalimantan Province, Indonesia

by M. Fathurahman

Submission date: 11-Dec-2021 06:45AM (UTC+1100)
Submission ID: 1726879429
File name: jsju_manuscript_proofreading.docx (43.44K)
Word count: 2396
Character count: 13136

Count Data Modeling Using GAMLSS Approach and Its Application in Dengue Hemorrhagic Fever Cases in East Kalimantan Province, Indonesia

Abstract

Generalized Additive Models for Location, Scale, and Shape (GAMLSS) is a robust approach used to model various types and characteristics of data. Therefore, this research aims to model the count data using GAMLSS approach through Poisson Regression (PR), Poisson Inverse Gaussian Regression (PIGR), and Negative Binomial Regression (NBR). PIGR and NBR are the best models compared to PR based on their application to modeling the number of dengue hemorrhagic fever (DHF) cases in East Kalimantan Province, Indonesia, in 2019. Furthermore, both models produced varying results on the factors with a significant effect on DHF. Only one factor of PIGR model, namely altitude, significantly affected these cases. Meanwhile, NBR model produced three factors that affected the number of dengue cases, such as altitude, population density, and health workers.

Keywords: data count, GAMLSS, PR, PIGR, NBR, DHF

Introduction

Count data is a type of statistical data often used in various research fields, such as social, economic, environmental, and health. The approach commonly used in its modeling process is Generalized Linear Models (GLM) or Generalized Additive Models (GAM). However, there are some weaknesses associated with using these two approaches. Firstly, they can only be used to model the relationship between explanatory variables and the location parameters of the response variable distribution. Secondly, they cannot model the relationship between explanatory variables with location parameters and the shape of the response distribution. These weaknesses can be overcome using an approach called Generalized Additive Models for Location, Scale, and Shape (GAMLSS).

GAMLSS is a development approach comprising GLM and GAM models that provide location parameters that describe only a limited aspect of the response variable distribution. This approach accommodates other parameters of the explanatory response variable distribution, such as scale and shape in the form of linear, nonlinear, parametric, nonparametric, and random effects functions. The response variable in GAMLSS follows a distribution that belongs to the exponential family with the addition of discrete, continuous, skewed, and kurtosis models. It can also be used to model count data, such as equidispersion, underdispersion, and overdispersion.

GAMLSS was used to model the number of Dengue Hemorrhagic Fever (DHF) cases in East Kalimantan Province, Indonesia, in 2019, using the count data type. DHF is a severe infectious disease caused by the dengue virus, contaminated through the bites of the *Aedes*

aegypti and *Aedes albopictus* mosquitoes. This disease disrupts the capillaries and blood-clotting systems, leading to bleeding and death while not treated properly. DHF is commonly found in tropical countries, such as Indonesia, and still a public health problem, specifically in East Kalimantan Province.

According to a 2020 publication by the Ministry of Health, the number of positive cases and deaths in East Kalimantan Province in 2019 was 6,723 and 44, respectively. This is in addition to Incidence Rate (IR) of 100,000 per population of 180.66, and Case Fatality Rate (CFR) of 0.65. %. Of the 34 provinces in Indonesia, East Kalimantan was ranked second for the highest IR value after North Kalimantan. IR value of DHF in this province was also very high and exceeded the overall value of 51.48 in Indonesia. This shows that the number of DHF cases in this province is very high.

The purpose of this research is to model the count data using GAMLSS approach through PR, PIGR, and NBR to determine the best model needed to analyze the number of DHF cases in East Kalimantan Province, Indonesia, in 2019. It also aims to determine the factors that significantly affect the number of DHF cases. Furthermore, this research is limited to PR, PIGR, and NBR models, which are modeled with GAMLSS approach as well as Rigby and Stasinopoulos (RS) algorithm.

Literature Review

GAMLSS

GAMLSS assumes that the response variable is Y_i for $i = 1,2,\dots,n$ with a probability distribution function $P(Y_i = y_i|\boldsymbol{\theta}^i)$ where $\boldsymbol{\theta}^i = [\theta_{i1} \quad \theta_{i2} \quad \cdots \quad \theta_{ip}]^T$. $\boldsymbol{\theta}^i$ is a vector of four distribution parameters, namely $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\upsilon}$, and $\boldsymbol{\tau}$ which are functions of explanatory variables. The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are referred to as location and scale, while $\boldsymbol{\upsilon}$ and $\boldsymbol{\tau}$ are known as skewness and kurtosis and included in the shape parameters.

Supposing $\boldsymbol{y}^T = [y_1 \quad y_2 \quad \cdots \quad y_n]$ denotes a vector of response variables with a size of $(n \times 1)$ and $g_k(\cdot), k = 1,2,3,4$ is a connecting function between distribution parameters and explanatory variables. GAMLSS model can be written as follows:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\gamma_{jk} \tag{1}$$

Supposing $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is an identity matrix with a size of $n \times n$ and $\gamma_{jk} = h_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ for all combinations of j and k in Equation (1), then GAMLSS model can be rewritten as follows:

$$\begin{aligned} g_k(\boldsymbol{\theta}_k) &= \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\boldsymbol{x}_{jk}) \\ g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\boldsymbol{x}_{j1}) \end{aligned} \tag{2}$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} h_{j2}(\boldsymbol{x}_{j2})$$

$$g_3(\boldsymbol{\upsilon}) = \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} h_{j3}(\boldsymbol{x}_{j3})$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} h_{j4}(\boldsymbol{x}_{j4})$$

where:

$\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\upsilon}, \boldsymbol{\tau}, \boldsymbol{\eta}_k$ = vector with length n

$\boldsymbol{\beta}_k^T = [\beta_{1k} \quad \beta_{2k} \quad \cdots \quad \beta_{J_k'k}]$ is the parameter vector

$\boldsymbol{X}_k$ = explanatory variable matrix of size $n \times J_k'$

h_{jk} = nonparametric smoothing function of the explanatory variables $\boldsymbol{x}_{jk}$, where $j = 1,2,\ldots,j_k$ is also a vector of length n and $k = 1,2,3,4$. The function h_{jk} is the unknown function of the explanatory variables X_{jk}, and $h_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ is a vector that evaluates the function h_{jk} on x_{jk}.

Inference to GAMLSS model includes parameter estimation and hypothesis testing. Parameter estimators can be obtained using the penalized likelihood method and numerical approach. These include RS algorithm, Cole and Green (CG) algorithm, and a mixture of RS and CG algorithms (RS-CG algorithm). However, this research only uses RS algorithm, while the parameter hypothesis testing was conducted with the likelihood ratio test and Wald procedures.

Distribution of Count Data on GAMLSS

Some distributions for modeling count data using GAMLSS approach are as follows:

1. Poisson distribution

This is a distribution of discrete random variables that express the number of successes of an experiment. According to Cameron & Trivedi (2013), it has the following characteristics:

- An event with a small probability that occurs in a population with a large number of members.
- Depends on a certain time interval.
- Events are included in the stochastic process.
- The recurrence of events follows the binomial distribution.

For example, assuming Y is a discrete random variable with Poisson distribution consisting of parameter μ, then the probability mass function is obtained as follows:

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}, y = 0,1,2,\ldots$$

Where μ is a location parameter that represents the probability of many successful events from Y. Poisson distribution has the same mean and variance, namely $E(Y) = \mu$ and $Var(Y) = \mu$. Meanwhile, the skewness is $S = E(Y-\mu)^3/\mu^3 = 1/\sqrt{\mu}$.

2. Poisson Inverse Gaussian Distribution

This is a mixed Poisson distribution with its shape depending on the random effect (υ). For example, $q(\upsilon)$ denotes the integral υ obtains the probability density function of υ and the marginal distribution for Y:

$$P(Y = y|\mu) = \int f(y|\mu, \upsilon)\, q(\upsilon) d\upsilon, \tag{3}$$

Where υ is assumed to have an inverse Gaussian distribution and a probability density function written as follows:

$$q(\upsilon) = (2\pi\tau\upsilon^3)^{-1/2} e^{-(\upsilon-1)^2/2\tau\upsilon}, \upsilon > 0 \tag{4}$$

Where $E(V) = 1$ and $\tau = Var(V)$. Based on Equations (3) and (4), PIG distribution is formed with the probability mass function as follows:

$$P(Y = y|\mu, \tau) = \left(\frac{2z}{\pi}\right)^{1/2} \frac{\mu^y e^{1/\tau} K_s(z)}{(z\tau)^y y!}, \tag{5}$$

Where $s = y - 1/2$, $z = \sqrt{1/\tau^2 + 2\mu/\tau}$, and $K_s(z) = K_{y-1/2}(1/\tau\sqrt{2\mu\tau + 1})$ is a modified Bassel function of the third kind.

3. Negative Binomial Distribution

Negative binomial distribution has two parameters, namely μ and σ, which denotes the parameter location and dispersion. The probability mass function of negative binomial distribution is formulated as follows:

$$P(Y = y|\mu, \sigma) = \frac{\Gamma((y+1)/\sigma)\alpha^y}{\Gamma(y+1)\Gamma(1/\sigma)} \left(\frac{(\mu\sigma)^y}{\mu\sigma + 1}\right)^{y+\frac{1}{\sigma}}, y = 0,1,2, \dots, \infty \tag{6}$$

Where $E(Y) = \mu$is the mean, $Var(Y) = \mu(1 + \alpha)$, and α is the dispersion parameter.

Best Model Selection

The criteria that can be used to obtain the best regression model based on GAMLSS approach are Global Deviance (GDEV), Akaike Information Criterion (AIC), and Schwarz Information Criterion (SBC), which are defined as follows:

$$\begin{aligned} \text{GDEV} &= -2\ell(\hat{\boldsymbol{\theta}}) \\ \text{AIC} &= -2\ell(\hat{\boldsymbol{\theta}}) + 2K \\ \text{SBC} &= -2\ell(\hat{\boldsymbol{\theta}}) + \log(n)\, K \end{aligned} \tag{7}$$

Where $\ell(\hat{\boldsymbol{\theta}})$ denotes the maximum log-likelihood of the model, K is the number of parameters in the model, and n is the sample size. The best models have the smallest GDEV, AIC, and SBC values.

Research Methodology

This research uses secondary data from the Central Statistics Agency and the Provincial Health Office of East Kalimantan, Indonesia. It uses one response variable (Y) and four explanatory variables (X_1,) X2, X3, X4, as shown in Table 1.

Table 1: Research variables

Symbol	Variable	Variable Type
X_1	An area	Continuous
X_2	Area altitude	Continuous
X_3	Population density	Continuous
X_4	Number of health workers	Discrete
Y	Number of DHF cases	Discrete

The steps of data analysis in this research are as follows:

1. Performing descriptive statistical analysis of research data.
2. Performing multicollinearity detection of explanatory variables.
3. Modeling the number of DHF cases using PR, PIGR, and NBR models based on GAMLSS approach.
4. Getting the best model needed for the number of DHF cases.
5. Getting the factors that influence DHF cases.
6. Drawing a conclusion.

Results and Discussion

The discussion starts with a descriptive analysis of research data, as shown in Table 2.

Table 2: Description of research data

Variable	Mean	SD	Max	Min
Y	672	621	1838	66
X_1	12734,7	11494,05	31051,7	163,1
X_2	54,1	65,04	174,63	5,98
X_3	380	579	1298	1
X_4	1381	903	2960	287

Description: SD = Standard deviation, Max = Maximum, Min = Minimum.

Table 2 shows that the average number of DHF cases at East Kalimantan Province in 2019 was 672. The highest and lowest numbers 1,838 and 6 cases, were found in Balikpapan City and Mahakam Ulu Regency, respectively. The data description on the number of DHF cases is shown in Figure 1. The average area of regencies/cities in East Kalimantan Province is 12,374.7 km^2, while the largest and smallest areas are found in East Kutai Regency (31,051.7 km^2) and Bontang City (163.1 km^2). The average height of 54.1 meters and height altitude of 174.63 masl are found in the regency/city in East Kalimantan Province and Mahakam Ulu Regency, respectively. Meanwhile, the regency/city with the lowest altitude in East Kutai Regency is at 31,051.7 km^2 with an average population density of 380 people/km^2. The highest population density is in Balikpapan City, with 1,298 people/km^2, and the lowest is in Mahakam Ulu Regency, as much as 1 person/km^2. Furthermore, the average number of health workers in

this province in 2019 was 1,381 people, with the highest and lowest found in Samarinda City (2,960 people) and Mahakam Ulu Regency (287 people), respectively.

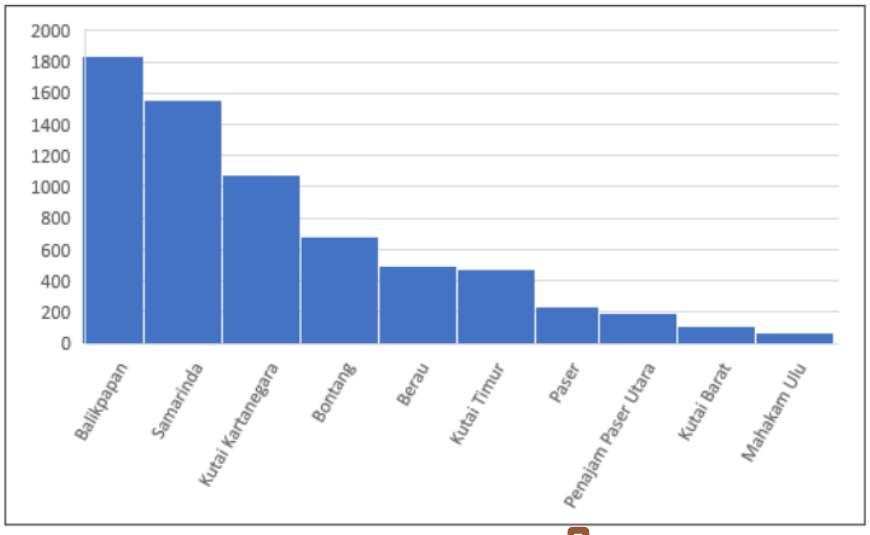

Figure 1: The description of the number of DHF cases in East Kalimantan Province, Indonesia, in 2019

This is followed by testing the prerequisites for data analysis of PR, PIGR, and NBR models, namely the detection of collinearity between explanatory variables. Multicollinearity detection uses the value of Variance Inflation Factor (VIF) with an explanatory variable comprising VIF value of more than 10, as shown in Table 3.

Table 3: VIF value of explanatory variables

Explanatory Variables	VIF Value
X_1	2,8981
X_2	1,3499
X_3	5,1523
X_4	2,9927

Table 3 shows that all explanatory variables have VIF value of less than 10. These results indicate that there is no multicollinearity in the explanatory variables. Therefore, it is important to use PR, PIGR, and NBR models.

The results of modeling the number of DHF cases in East Kalimantan Province, Indonesia, in 2019, using PR, PIGR, and NBR models with GAMLSS approach are shown in Table 4.

Table 4: Estimation and testing of PR model parameters

Model	Par	Est	Se	Z	p-value

PR	β_0	4,932	$6{,}621\times10^{-2}$	74,50	$8{,}26\times10^{-9}$*
	β_1	$3{,}237\times10^{-5}$	$2{,}672\times10^{-6}$	12,11	$6{,}77\times10^{-5}$*
	β_2	$-5{,}257\times10^{-3}$	$4{,}235\times10^{-4}$	-12,41	$6{,}01\times10^{-5}$*
	β_3	$9{,}746\times10^{-4}$	$6{,}659\times10^{-5}$	14,64	$2{,}69\times10^{-5}$*
	β_4	$4{,}934\times10^{-4}$	$2{,}523\times10^{-5}$	19,55	$6{,}45\times10^{-6}$*
PIGR	β_0	5,008	0,274	18,278	$5{,}27\times10^{-5}$*
	β_1	$2{,}68\times10^{-5}$	$1{,}78\times10^{-5}$	1,506	0,2066
	β_2	$-6{,}724\times10^{-3}$	$1{,}973\times10^{-3}$	-3,408	0,0271*
	β_3	$9{,}685\times10^{-4}$	$5{,}856\times10^{-4}$	1,654	0,1735
	β_4	$5{,}185\times10^{-4}$	$3{,}125\times10^{-4}$	1,659	0,1724
NBR	β_0	4,998	$2{,}330\times10^{-1}$	21,449	$2{,}79\times10^{-5}$*
	β_1	$2{,}757\times10^{-5}$	$1{,}512\times10^{-5}$	1,824	0,1423
	β_2	$-6{,}791\times10^{-3}$	$1{,}683\times10^{-3}$	-4,034	0,0157*
	β_3	$9{,}609\times10^{-4}$	$2{,}702\times10^{-4}$	3,556	0,0237*
	β_4	$2{,}240\times10^{-4}$	$3{,}493\times10^{-5}$	15,002	0,0001*

Description: Par = Parameter, Est = Estimate, Se = Standard error.
* Significant at the level of significance, $\alpha = 0{,}05$.

Table 4 shows that all PR model parameters are significant at a significance level of 0.05. Furthermore, for PIGR model, only two parameters are significant, namely β_0 and β_2. Meanwhile, the significant parameters in NBR model are four, namely $\beta_0, \beta_2, \beta_3$, and β_4.

Furthermore, this research selects the best model needed to determine the number of DHF cases in East Kalimantan Province in 2019. The results obtained are shown in Table 5.

Table 5: Selection of the best model with GDEV, AIC, and SBC values

Model	GDEV	AIC	SBC
PR	416,7973	426,7973	428,3102
PIGR	122,0417*	134,0417*	135,8572*
NBR	122,2912	134,2912	136,1067

* Best Model

Based on Table 5, the model that has the smallest GDEV, AIC, and SBC values is PIGR. These results indicate that the best model for analyzing the number of DHF cases in East Kalimantan Province in 2019 is PIGR model. Meanwhile, the values of GDEV, AIC, and SBC for NBR model are relatively the same or close to PIGR. Therefore, it is necessary to consider the selection of NBR model as an alternative to analyzing the number of DHF cases in East Kalimantan Province in 2019, specifically NBR model with many significant parameters compared to PIGR.

PIGR model is obtained as follows:

$$\log(\hat{\mu}) = 5{,}008 + 2{,}68\times10^{-5}X_1 - 6{,}724\times10^{-3}X_2 + 9{,}685\times10^{-4}X_3 + 5{,}185\times10^{-4}X_4.$$

The significant explanatory variable in PIGR model above is the altitude of the region(X_2), which is used to interpret regencies/cities with a tendency of reducing the number of dengue cases by 1.0067 times.

Conclusion

GAMLSS is a flexible approach adaptable to various characteristics of data or distributions. This approach is able to accommodate other parameters from the distribution of response variables related to explanatory variables. These examples are scale and shape parameters in the form of linear, nonlinear, parametric, nonparametric functions, and random effects. The response variable in GAMLSS not only follows a distribution that belongs to the exponential family, rather it also includes discrete and continuous distributions with highly skewed and kurtosis. Therefore, GAMLSS approach can be used to model count data, such as equidispersion, underdispersion, and overdispersion.

Based on the application of GAMLSS through PR, PIGR, and NBR models in analyzing the number of DHF cases in East Kalimantan Province, Indonesia, in 2019, PIGR was obtained as the best. However, NBR model can be used as an alternative because it has the same GDEV, AIC, and SBC values as PIGR. NBR produces more significant parameters than PIGR. The area's height is the only factor that significantly affects the number of DHF cases in East Kalimantan Province in 2019 based on PIGR model. Meanwhile, NBR produces three factors with a significant effect, namely the area's altitude, population density, and the number of health workers. PIGR and NBR models produce a regional altitude factor that significantly affects the number of DHF cases in East Kalimantan Province, Indonesia, in 2019.

Count Data Modeling Using GAMLSS Approach and Its Application in Dengue Hemorrhagic Fever Cases in East Kalimantan Province, Indonesia

ORIGINALITY REPORT

6%	3%	5%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Klugman, . "Discrete Distributions and Processes", Wiley Series in Probability and Statistics, 2012. Publication	1%
2	Getiye Dejenu Kibret, Tadesse Awoke Ayele, Adino Tesfahun. "Incidence and Predictors of Sever Adverse Drug Reactions Among Patients on Antiretroviral Therapy at Debre Markos Referral Hospital, Northwest Ethiopia", Research Square, 2019 Publication	1%
3	Mickaël Campo, Benoît Louvet, Sofiene Harabi. "Dimensions of social identification with the team as predictors of the coach-created training climate in rugby: A group-actor partner interdependence modelling perspective", Psychology of Sport and Exercise, 2022 Publication	1%

4	www.nature.com Internet Source	1 %
5	J P M Kolanus. "Functional Properties and Chemical Composition of Dried Surimi Mackerel (Scomberomorus Sp) With Different Cryoprotectants and Drying Methods", Journal of Physics: Conference Series, 2020 Publication	1 %
6	louisdl.louislibraries.org Internet Source	1 %
7	Adi Wijaya, Surya Darma, Dio Caisar Darma. "Spatial Interaction Between Regions: Study of the East Kalimantan Province, Indonesia", International Journal of Sustainable Development and Planning, 2020 Publication	1 %
8	archive.org Internet Source	1 %
9	etd.repository.ugm.ac.id Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		