Analysis of Bayesian Generalized Linear Models on the Number of Tuberculosis Patients in Indonesia with R

Femmy Diwidian[1, *], Anang Kurnia[2], Kusman Sadik[2]

{anangk@apps.ipb.ac.id}

[1] Mathematics Education Study Program , UIN Syarif Hidayatullah Jakarta , 15419, Indonesia
[2] Statistics Department, Bogor Agricultural University , Bogor , 16680, Indonesia

Abstract. *Generalized Linear Models (GLM)* is an extension of the linear regression model that aims to determine the causal relationship, the effect of independent variables on the dependent variable where the response variable is a member of the exponential family. In general, estimating parameters on GLM can be divided into two approaches, namely the frequentist method and the Bayesian GLM method. In this study, both approaches will be used to analyze the number of people suffering from tuberculosis in 34 provinces in Indonesia. The data used is based on 2018 Indonesia Health Profile Data and Information published by the Ministry of Health of the Republic of Indonesia in 2018. Based on the best model test criteria, this study provides results that the frequentist approach to GLM is better in matching the number of people suffering from tuberculosis in Indonesia compared to use Bayesian GLM.

Keywords: Generalized Linear Models (GLM), Frequentis Method, Bayesian GLM Method

1. Introduction

Generalized linear model (GLM) is an extension of the linear regression model assuming the predictor has a linear effect but does not assume a particular distribution of the response variable and is used when the response variable is a member of an exponential family [1]. GLM aims to determine the causal relationship, the effect of independent variables on the dependent variable. The superiority of GLM compared to ordinary linear regression lies in the distribution (curve shape) of dependent variables. Variable dependent on GLM is not socialized with normal distribution (symmetrical bell curve), but distributions that belong to an exponential family, namely; Binomial, Poisson, Negative Binomial, Normal, Gamma, Gaussian Inverse .

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290479

In GLM, an Inference test needs to be done, including the parameter test and model compatibility test. The parameter test is performed to determine whether the parameters in the GLM model are significant or not and the model compatibility test is to compare the model with the most likely model, which is a model in which many parameters are involved.

There are two approaches in modeling and estimating regression parameters, namely the Bayesian approach and the *frequentist* approach [2]. The Bayesian approach assumes that the data is fixed and that the model parameters are random, which in analyzing requires prior distribution of the regression parameters. While the frequentist approach assumes the opposite, random data and model parameters are fixed [3]. In this research an analysis of the two methods in modeling and estimating the parameters of the Tuberculosis data in Indonesia will be carried out .

Based on Indonesia Health Profile Data and Information published by the Ministry of Health of the Republic of Indonesia in 2018, Tuberculosis defined as an infectious disease caused by the bacterium *Mycobacterium tuberculosis* , the bacteria can invade various organs of the body, especially the lungs. Although only caused by bacteria but can be dangerous even to the point of death. Although vaccines and anti-tuberculosis drugs have been discovered, there is still an increase in the number of people with tuberculosis in Indonesia. This can be seen based on Indonesia Health Profile Data and Information 2018 published by the Ministry of Health of the Republic of Indonesia in 2018 where the number of tuberculosis patients increased from 360,770 to 511,873 cases[4]. According to Jumariana in her research in 2012 the risk factors that can cause Tuberculosis are genetic, malnutrition, vaccination, poverty and population density[5]. In addition, Tuberculosis is common in stressed populations, poor nutrition, overcrowding, inadequate house ventilation, inadequate health care. Based on the effects and effects caused by the attack of Tuberculosis , this research will look for factors that influence Tuberculosis by using GLM.

2. Materials

The data used secondary data taken based on 2018 Indonesia Health Profile Data and Information published by the Ministry of Health of the Republic of Indonesia in 2018 . Provinces as observation units so that the number of observations becomes 34 observation units. The variables used in this study are shown in **Table 1**.

Table 1. List of variables used

Variable	Information

Y	Number of Tuberculosis sufferers
X1	Population Density (/ km²)
X2	Poor Population Percentage
X3	Number of Puskesmas
X4	Percentage of Slum Households
X5	Percentage of Regencies / Cities that have PHBS
X6	Percentage of livable homes

3. Method

In this study GLM analysis will be carried out to determine the factors that influence Tuberculosis . GLM is a general form of the Linear Model that is modeled in form .

$$g(\mu_i) = X_i^T \beta \tag{1}$$

If a vector **y** has n components, which is the realization of a matrix response **Y** that each component is independent and Normal distribution or E(Y) = μ . As a transition from linear models to GLM , then elaborated form three components, namely: Random Component : a observational values of **Y** responses that are independent of certain distributions, Systematic Components: a linear combination of variable **X** with parameters denoted by η = Xβ and Link between random and Systematic / link function : a function that explains the expected value of the response variable (**Y**) which connects with explanatory variables through linear equations. Written by $\eta_i = g(.)$, function $g(.)$ is called the link function or link function .

There are two statistical inference approaches that will be carried out in this study. The first method is to use the Frequentist approach or often said with the classical approach. Deficiencies in the classical approach can occur because random sample information in past research is not taken into account. While the second method is to use you right Bayesian approach se like comparison frequentist approaches. Models of both approaches will be compared with the value of AIC so that it can be determined which of the best models of both approaches.

The procedures that will be carried out in this study are:

3.1 Evaluate the exponential family distribution in the data

Random variables have expectation $E(Y) = \mu$ and variance values $Var(Y) = \sigma^2$. Distribution of Y called an exponential family member if the opportunity distribution function can be written as:

$$f(y|\theta) = m(y)r(\theta)e^{s(\theta)t(y)} \tag{2}$$

which shows canonical parameters. According to Nelder et.al in general the equation can be written as:

$$f(y|\theta) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

Or

$$log[f(y|\theta)] = \ell(\theta; y, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \tag{3}$$

which shows the scale parameters.

The data uses a ratio scale where the chance of being exposed to Tuberculosis is small. Based on the data there are two possible distributions of the Y response, Poisson and negative Binomial. If the mean of the process is equal to the variance, the response Y will be Poisson distribution, whereas in reality, the assumption that the mean is equal to the variance (equidispersion) is rarely fulfilled. In general, the data found is discrete data with a variance that is greater than the mean value (overdispersion). Appropriate analysis in handling the overdispersion problem is to use a negative Binomial distribution. Therefore, it is necessary to check the presence or absence of overdispersion.

```
#Overdispersion test, The Sintax R used is as follows:

ovmod=glm(Y~.,data = dta,family = "poisson")

dispersiontest(ovmod)

mod1=glm.nb(Y~.,data=dta)

summary(mod1)
```

3.2 Modeling by *Frequentist* method

The frequentist approach was initially widely used in the field of Mathematics. But since the beginning of the 18th century, classical linear models began to be used in the field of Statistics to model astronomical problems. Gauss (1801) was able to predict the position of the appearance of Ceres asteroids using a linear model and find the Normal distribution. The simplest classical linear model is defined in equation (4).

$$\begin{gathered} y = f(x, \beta, \varepsilon) \\ y = \beta_0 + \beta x + \varepsilon \end{gathered} \tag{4}$$

where y is the dependent variable whose value depends on the independent variable x . β is an unknown model parameter, while it is a random variable that differs from the actual value of y with its estimated value. The random variable is assumed to follow the Normal (0, σ^2) distribution .

The Sintax R used is as follows for Frequentist method:

```
setwd("C:\\Users\\Femmy Diwidian\\Documents\\GLM\\tugas 3")

dta=readxl::read_excel("data tuberkulosis.xlsx",sheet = 2)

library(dplyr)

library(MASS)

library(AER)
```

```
dta=dta%>%dplyr::select(-Provinsi)
dta
```

3.3 Bayes Method Modeling

In using the Bayes method, the following assumptions and characteristics must be met, among others, the parameters of the model and the response variables are random, based on prior distribution, the distribution of model parameters is called prior distribution, any parameter of the prior distribution is called hyperparameter. While Bayesian methods focus on five important elements, namely: Information distribution priors (normal, Poisson, binomial, etc.), prior combined with the likelihood function, this combination function that generate distribution posterior, simutation derived from distribution posterior to produce empirical distributions for population parameters and draw conclusions from empirical distributions from the simulation of posterior distribution.

The Sintax R used is as follows for Bayesian negative binomial

```
library(rstanarm)
library(dplyr)
dta=readxl::read_excel("data tuberkulosis.xlsx",sheet = 2)
dta=dta%>%dplyr::select(-Provinsi)
mod2=stan_glm.nb(Y~.,data=dta)
prior_summary(mod2)
summary(mod2)
plot(mod2)
```

3.4. Model Goodness Test

Test criteria for the best model used is *Akaike's Information Criteria* (AIC). *Akaike's Information Criterian (* AIC) is obtained from the equation:

$$AIC \; = -2 \log L_{fit} + \; 2k \tag{5}$$

The smallest AIC value indicates that the model is better compared to other models. This study uses the AIC (mod1) function for the OLS and AIC (mod2) approaches for the Bayesian approach in GLM.

4. Results and Discussion

4.1 Description

Based on the 2018 Indonesia Health Profile Data and Information published by the Ministry of Health of the Republic of Indonesia in 2018 where the Province was the observation unit so that the number of observations became 34 observation units The number of Tuberculosis sufferers in Indonesia is shown in **Figure 1** .

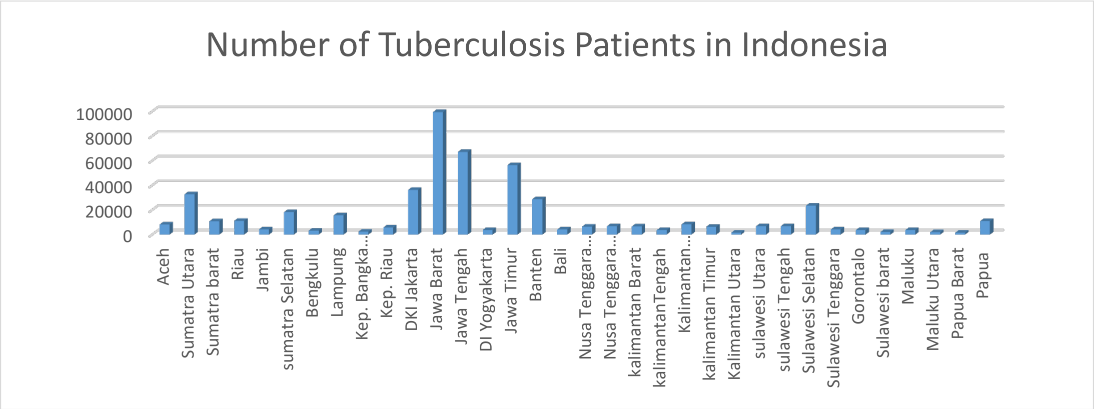

Fig 1. Number of Tuberculosis Patients in Indonesia

Figure 1 shows the large number of tuberculosis patients in Indonesia, where the highest number is in the province of West Java, while Central Java is ranked second followed by East Java, DKI Jakarta and Banten. If seen from the data above, almost 80% of the Java region has a high rate of tuberculosis compared to provinces outside of Java. This shows that the development of Tuberculosis on the island of Java is quite large which raises the question of what factors influence Tuberculosis in Indonesia

In this study the data structure Tuberculosis is made with N total in use as much of Indonesia's provinces to 34 provinces with independent variables as much as 6 namely, Y = number of patients with tuberculosis , X1 = k epadatan population (/ km^2), X2 = persetase p enduduk poor , X3 = number of puskesmas , X4 = Percentage of Slumy Households , X5 = Percentage of Regencies / Cities that have PHBS , and X 6 = Percentage of Livable Houses . The scatterplot in Figure 2 shows the correlation between X and Y

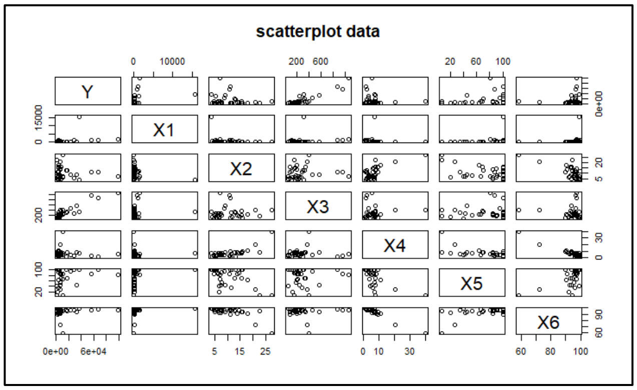

Fig 2 . Scatterplot correlation between X and Y

Based on the Scatterplot in Figure 2, it appears that there is a fairly strong positive correlation between Y and X3, and there is also a negative correlation between X4 and X6 and X4 and X2

Data to be used is the amount it has to do testing whether there is o verdispersi can be determined so that the most appropriate resolution techniques to the data. If there is an overdispersion of data, then using a negative binomial distribution becomes more effective . Results of testing Overdispersion test on R indicates that z = 2.8057 with p-value = 0.00251 , an alternative hypothesis: true dispersion is greater than 1 , resulting in a dispersion = 2356.358. B erdasarkan overdispersi test are overd ispesi thus fulfilled the assumption that the data distribution is negative binomial

4.2 Negative binomial frequentist modeling

The following results are the results of GLM with a negative Binomial distribution that will be used in estimating regression parameters.

Table.2 Deviance Residuals

Min	1Q	Median	3Q	Max
-1.8215	-0.9721	-0.2512	0.3129	2.6102

In Table 2. it can be seen that the minimum value of residuals is -1.8215 while the maximum value is 2.6102, with the middle value at -0.2512, it can be predicted that the distribution of residuals is more congested on the left. For more details can be seen in the image below

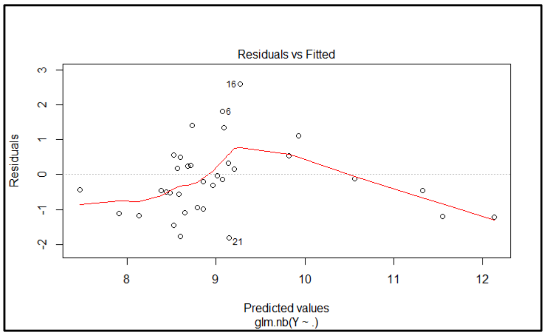

Fig 2 . Residual and Fitted Number of Tuberculosis Patients in Indonesia

Table.3 coefficient values of the model

	Estimate	**Std. Error**	**z value**	**Pr (> | z |)**
(Intercept)	8.140e+00	4.370e+00	1.863	0.06250.
X1	5.950e-05	3.568e-05	1.668	0.09539.
X2	-4.061e-02	2.094e-02	-1.939	0.05245.
X3	3.689e-03	3.361e-04	10.975	<2e-16 ***
X4	2.289e-02	5.070e-02	0.452	0.65162
X5	9.062e-03	3.411e-03	2. 657	0.00788 **
X6	-6.333e-03	4.237e-02	-0.149	0.88117

Table 3. the output model using the frequentis t approach to the data shows that the independent variables that influence the real level of 0.001 are the variables X3 and X5, namely the number of Puskesmas and the percentage of districts / cities that have PHBS. So it can be said that Tuberculosis is influenced by the number of Puskesmas and the percentage of regencies / cities that have PHBS at the real level of 0.001. whereas at the real level 0.1 there was an increase in the influential variables, namely intercept, variable X1 (Population Density (/ km^2)) and variable X2 (Number of Poor Population). So that at a significant level of 0.1 there are 5 variables that affect tuberculosis, namely intercept, X1 (Population Density (/ km^2)), X 2 (Number of Poor Residents), X3 (Number of Puskesmas) and X5) .

4.3 Bayesian Modeling on negative binomial distribution

The following results are the results of Bayesian glm with a negative Binomial distribution that will be used in estimating the regression and prior parameters used in the data .

Table.4 Priors for 'mod2' models

Intercept **(after predictors centered)**	**Coefficients**	**Auxiliary** **(reciprocal_dispersion)**

~ normal (location = 0, scale = 10)	~ normal (location = [0,0,0, ...], scale = [2.5,2.5,2.5, ...]) ** adjusted scale = [0.00093,0.43833,0.01022,...]	~ exponential (rate = 1)

By using the syntax booth glm.nb in Y data that spread negative binomials in the sampling algorithm using priors in table.5 to 4000 posterior sample size, 34 observations (provincy) and 7 predictors, the following model estimates are generated.

Table 5. Summary of Bayesian model

	mean	**sd**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
(Intercept)	8.0	5.3	-2.0	4.4	7.8	11.5	19.2
X1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X2	0.0	0.0	-0.1	-0.1	0.0	0.0	0.0
X3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X4	0.0	0.1	-0.1	0.0	0.0	0.1	0.1
X5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X6	0.0	0.1	-0.1	0.0	0.0	0.0	0.1
reciprocal_dispersion	3.3	0.8	1.8	2.7	3.2	3.7	5.1
mean_PPD	21184.2	7837.1	11119.6	15982.5	19520.6	24548.2	41317.9
log-posterior	-344.6	2.3	-350.0	-345.9	-344.3	-343.0	-341.3

Table 6. diagnostic of Bayesian model

	mcse	**Rhat**	**n_eff**
(Intercept)	0.1	1.0	1720
X1	0.0	1.0	1961
X2	0.0	1.0	3110
X3	0.0	1.0	3531
X4	0.0	1.0	1586
X5	0.0	1.0	2871
X6	0.0	1.0	1748
reciprocal_dispersion	0.0	1.0	2937
mean_PPD	148.2	1.0	2798
log-posterior	0.1	1.0	1396

For more clearly mod2 can be illustrated in **Figure 3** below

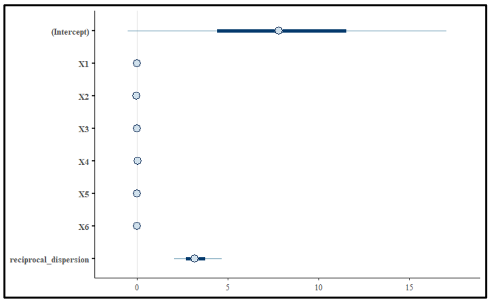

Fig 3. Plot model with a Bayesian approach to GLM

If you see a model based on the results of Bayesian glm, there are no predictors that affect Y, so it can be said that of the 7 predictors, none of them affect Tuberculosis.

4.4 Model Goodness Test

The selection of the best model from the several models presented, was chosen based on the criteria used. The choice of criteria as a comparison model to obtain the best model in this study is used Akaike's Information Criteria (AIC), AIC is a criterion that balances the model's goodness of fit based on the likelihood value with the many parameters of the model. Thus, the best model is the model with the smallest Akaike's Information Criterian (AIC) value. The negative binomial Frequentist approach assures the AIC value of 672.08 while the negayive Bayesian binomial approach to GLM AIC value generated is 674.3.

Based on the best model AIC test criteria, this study provides result that approach better frequentis on GLM in fitted the number of people suffering from Tuberculosis in Indonesia compared to using Bayesian GLM.

Reference.

[1] Nelder, J. A., and Wedderburn, (1972). :Generalized Linear Model". Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3.

[2] Mamouridis, V., 2011, Additive Mixed Models applied to the study of red shrimp landings: comparison between frequentist and Bayesian perspectives

[3] Lopiano, et.al. 2013. Generalized Linear Models in R Frequentist and Bayesian Methods. Starkweather, Jon. 2011. Bayesian Generalized Linear Models in R.

[4] Indonesia Health Profile Data and Information published by the Ministry of Health of the Republic of Indonesia in 2018

[5] Jumariana, Faktor-Faktor yang Berhubungan dengan Kejadian Tuberculosis Paru di wilayah Kerja Puskesmas Maccini Sawah Kota Makasar 2012, UIN Alauddin, Makasar 2012

[6] Agresti, A., 2015 Foundation of Linear and *Generalized Linear Models* , New Jersey: Wiley

[7] Bakker, Ryan. Bayesian Methods: Review of Generalized Linear Models. https://spia.uga.edu/faculty_pages/rbakker/bayes/Day2.applied.bayes.pdf

[8]Das, Sourish and Dey, K. *Dipak 2007, on Bayesian Analysis of Generalized Linear Models: A New Perspective,* Statistical and Applied Mathematical Science Institute.

[9]Dobson, Annete, J., and Barnett, 2008, *An Introduction to Generalized Linear Models,* Third edition, London and New York: Chapman and Hall.

[10]Starkweather, Jon. 2011. Bayesian Generalized Linear Models in R. https://it.unt.edu/sites/default/files/bayesglm_jds_jan2011.pdf

[11]Wakefield, Jon. 2011. Module 4: Bayesian Methods. http://faculty.washington.edu/jonno/SISG-2011/lectures/JW-Lec3_2.pdf

[12]Zahro, Jamilatuz. 2018. Application of Generalized Linear Model in R.ISBN :978-602-6542-46-5

Attachment 1

Data of tuberculosis

	Y	X1	X2	X3	X4	X5	X6
1	8145	91.13	15.68	348	8.06	69.57	93.38
2	32651	197.52	8.94	581	7.71	45.45	94.03
3	10754	128.11	6.55	275	6.58	84.21	95.38
4	11135	78.31	7.21	216	4.06	66.67	97.49
5	4297	71.32	7.85	195	4.04	90.91	95.99
6	18114	91.39	12.82	332	7.93	70.59	94.54
7	3283	98.56	15.41	180	6.66	100.00	93.70
8	15570	241.76	13.01	302	2.21	100.00	96.03
9	2353	88.89	4.77	64	2.28	100.00	99.10
10	5693	260.50	5.83	83	2.53	100.00	98.44
11	36241	15764.26	3.55	321	7.07	100.00	99.36
12	99398	1376.11	7.25	1069	5.92	81.48	97.09
13	67063	1051.53	11.19	881	1.75	100.00	97.40
14	3776	1213.75	11.81	121	1.13	100.00	99.46
15	56445	826.38	10.85	967	2.59	86.84	96.95
16	28797	1313.24	5.25	242	3.36	100.00	96.93
17	4114	742.58	3.91	120	1.79	100.00	99.03
18	6437	269.95	14.63	166	6.43	90.00	97.06
19	6746	110.26	21.03	381	20.65	27.27	72.06
20	6570	33.95	7.37	244	5.87	42.86	92.84
21	3548	17.32	5.10	200	7.05	100.00	93.73
22	8209	107.96	4.65	233	4.48	84.62	96.64
23	6387	28.27	6.06	183	3.82	90.00	98.20
24	1549	9.49	6.86	56	4.47	40.00	99.16
25	6786	179.36	7.59	193	7.24	53.33	95.59
26	6901	48.68	13.69	202	8.51	84.62	91.10
27	23427	187.77	8.87	458	3.67	95.83	96.60
28	4093	69.71	11.32	284	6.09	35.29	96.20
29	3521	105.31	15.83	93	9.76	100.00	93.85
30	2097	80.75	11.22	94	9.67	100.00	89.88

31	3656	37.81	17.85	208	11.05	63.64	90.55
32	1883	38.54	6.62	134	7.39	20.00	92.18
33	1421	9.41	22.66	159	8.10	7.69	94.55
34	10813	10.41	27.43	408	40.01	6.90	58.23