Modelling of the number of malarias suffers in Indonesia using Bayesian generalized linear models

V M Santi¹*, A Kurnia² and K Sadik²

¹Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Negeri Jakarta, Indonesia
²Department of Statistics, Faculty of Mathematics and Natural Science, Institut Pertanian Bogor

*vera.indr4@gmail.com

Abstract. Generalized Linear Models (GLM) has been used for modelling various types of data where the distribution of response variables is an exponential family. Common examples include those for binomial and Poisson response data. The GLM regression model determines the structure of the explanatory variable or covariate information, where the link function specifically determines the relationship between the regression model and the expected value of the observation. Bayesian techniques can now be applied to complex modelling problems where they could not have been applied previously. This method is a simpler model than traditional frequentist techniques. Estimating the regression model parameters is done by using Bayesian GLM. In this paper, we study conducted modelling for the number of dengue sufferers in Indonesia using the Bayesian GLM approach with several prior distributions. There are 6 independent variables that have a significant effect on the regression model, that is population density, Gini ratio, proper sanitation access, healthy zoning, integrated control and total sanitation. Based on Akaike Information Criterion (AIC) and standard error, the Bayesian GLM estimation results for Cauchy and Normal prior distribution will converge to the same value as that obtained by GLM.

1. Introduction
Malaria is an acute infectious disease caused by protozoa of the genus Plasmodium. This parasite is transmitted by mosquito bites Anophelesbetina In humans, there are four species that cause malaria, namely P. falciparum, P. vivax, P. ovale, P. Malaria. The natural spread of the malaria parasite is caused by female Anopheles mosquitoes [1,2].

Areas with high malaria cases reported from Eastern Indonesia such as; Papua, West Papua, East Nusa Tenggara, Maluku and North Maluku Provinces. In other regions it is also reported to be still quite high, among others, in the provinces of Bengkulu, Bangka Belitung, Central Kalimantan, Lampung, and Central Sulawesi [3].

Generalized Linear Models (GLM) have been used to model various types of data where the distribution of response variables is an exponential family [4-7]. A common example included is the response variable with a binomial and Poisson distribution [8-11]. The GLM regression model determines the information structure of the explanatory variable or covariate, where the connection function specifically determines the relationship between the regression model and the expected value of the data. Simple linear models with normal spread errors are special cases of GLM. The Bayesian
The Bayesian approach in modeling and estimating regression parameters is a simpler model than traditional frequentist techniques [5,12].

The Bayesian approach to inferencing statistics is different from the frequentist technique which assumes that the data is fixed and the model parameters are random. Bayesian analysis of a set of data requires prior distribution information for the regression parameters used in the model [13,14]. Traditional frequentists assume that the model parameters are fixed and data is random.

The thing that needs to be considered in determining priors is that priors must be reduced rationally because priors can be strong or weak [12]. Weak priors are produced when we do not have much evidence or previous information which is the basis of priors. When priors are weak, the previous distribution will be wide which reflects many possible values and the function is likely to be more influential in creating a posterior distribution [12]. Conversely, strong priors arise when we have a lot of evidence or information that is the basis of priors. When priors are strong, the prior distribution will have a narrow interval, reflecting a smaller range of values and the function will likely be less influential in creating posterior.

The proposed of this paper is modeling the number of malaria sufferers in Indonesia with 6 independent variables such as population density per km$^2$, gini ratio, percentage of public health centers implementing integrated non-communicable disease control, percentage of households that have access to proper sanitation, percentage of districts/cities that organize regional arrangements health, and the percentage of villages that implement community based total sanitation [15]. Estimation of the regression model parameters is done using GLM and Bayesian GLM. Some prior distributions are used in estimating parameters of the Bayesian GLM model.

2. Material and methods

2.1. Material

The Bayesian approach to GLM in this paper applies to data on the number of malaria sufferers found in each province in Indonesia during 2017. Data was obtained from the 2017 Indonesian Health Profile published by the Indonesian Ministry of Health.

The variable response (Y) observed was the number of people who were positively affected by malaria who were found in each province in Indonesia. The explanatory variables observed are population density per km$^2$, gini ratio, percentage of community health centers that carry out integrated control of non-communicable diseases, percentage of households that have access to proper sanitation, percentage of regencies or cities that organize healthy regional arrangements, percentage of villages implementing community based on total sanitation.

2.2. Methods

In this study, the Bayesian Generalized Linear Models (GLM) approach was used using several prior distributions, namely normal distribution, Cauchy, and t-student in modeling malaria cases in Indonesia. Secondary data in this study are derived from Ministry of Health data in 2017. The response variable which is the number of malaria sufferers in each province in Indonesia is assumed to have a Poison distribution. The first stage in this study was to do descriptive statistics to describe the distribution of the response variables. The second stages is modeling malaria data using six explanatory variables with Bayesian GLM for several prior distributions. The modeling results are compared with the GLM frequentist approach. AIC criteria and standard errors of parameter coefficients were used to see the goodness of fit model.

3. Result and discussion

An applied study of a set of data with Y response variables, namely the number of malaria sufferers in each province in Indonesia. Based on Table 1 below, the number of malaria sufferers in 34 provinces in Indonesia is presented in 2017.
Table 1. Number of malaria sufferers in Indonesia in 2017.

| No | Province             | Y  |
|----|----------------------|----|
| 1  | Aceh                 | 299|
| 2  | Sumatera Utara       | 2442|
| 3  | Sumatera Barat       | 512|
| 4  | Riau                 | 211|
| 5  | Jambi                | 187|
| 6  | Sumatera Selatan     | 911|
| 7  | Bengkulu             | 1031|
| 8  | Lampung              | 4297|
| 9  | Kep. Bangka Belitung | 95 |
| 10 | Kep. Riau            | 357|
| 11 | DKI Jakarta          | 111|
| 12 | Jawa Barat           | 328|
| 13 | Jawa Tengah          | 904|
| 14 | DI Yogyakarta        | 86 |
| 15 | Jawa Timur           | 107|
| 16 | Nusa Tenggara Barat | 765|
| 17 | Nusa Tenggara Timur | 30451|
| 18 | Kalimantan Barat     | 152|
| 19 | Kalimantan Tengah    | 760|
| 20 | Kalimantan Selatan   | 1135|
| 21 | Kalimantan Timur     | 1573|
| 22 | Sulawesi Utara       | 900 |
| 23 | Sulawesi Tengah      | 543 |
| 24 | Sulawesi Selatan     | 1201|
| 25 | Sulawesi Tenggara    | 596 |
| 26 | Gorotalo             | 46  |
| 27 | Sulawesi Barat       | 147 |
| 28 | Maluku               | 4019|
| 29 | Maluku Utara         | 957 |

The following is the histogram of the responses to the number of malaria sufferers in Indonesia as shown in Fig. 1 and the scatterplot of each independent variable as presented in Fig. 2 below. Based on Fig. 2, it can be seen that the independent variables used in this study are normal distribution and some have Poisson distribution.

Figure 1. Histogram of response variable of malaria cases.
Furthermore, the regression parameter estimation is performed using the Bayesian GLM approach with several prior distributions. The following results are obtained by the response variable which is Poisson distribution with its offset is the number of residents in each province.

Based on Table 3, it can be seen that in addition to using the prior t-student distribution it gives the same results as GLM based on the standard error of each coefficient parameters. With prior t-student distribution provides the estimation results where the standard error is slightly greater than the estimation with GLM. Bayesian GLM without the prior distribution also gives the exact same result as GLM because the "ARM" package in R will automatically select the prior distribution which is the best for estimating the model parameters, namely by using the empirical distribution as the prior distribution. One of the advantages of the Bayesian GLM perspective, even for all analyzes, is being able to make a good and credible interval of trust. In the Bayesian framework, the resulting intervals are truly believed to contain actual population parameters.

|     | GLM (Prior Cauchy) | GLM (Prior Normal) | GLM (Prior T-Student) | GLM (Tanpa Prior) |
|-----|-------------------|--------------------|-----------------------|------------------|
| Intercept | 6.01E-02 | 6.01E-02 | 0.0600656 | 0.0600659 | 6.01E-02 |
| X1 | 1.87E-05 | 1.87E-05 | 0.0000187 | 0.0000187 | 1.87E-05 |
| X2 | 1.42E-01 | 1.42E-01 | 0.14197 | 0.1419709 | 1.42E-01 |
| X3 | 3.23E-04 | 3.23E-04 | 0.000323 | 0.000323 | 3.23E-04 |
| X4 | 2.97E-04 | 2.97E-04 | 0.0002965 | 0.0002965 | 2.97E-04 |
| X5 | 9.47E-04 | 9.47E-04 | 0.000947 | 0.000947 | 9.47E-04 |
| X6 | 1.65E-04 | 1.65E-04 | 0.00165 | 0.001651 | 1.65E-04 |

Figure 2. Scatterplot from data of malaria in Indonesia.

Figure 3. Standard errors of each variable from GLM and Bayesian GLM.
4. Conclusion
Bayesian GLM is an alternative method in estimating the regression model parameters. If the sample size is larger, the Bayesian GLM estimation results will converge to the same value as that obtained by GLM whose base is frequentist. Based on AIC and standard error of coefficient parameters, the results of estimating parameters for malaria cases in Indonesia with prior Cauchy, t-student and Normal distributions almost give the same results. The six independent variables are population density, gini ratio, access to proper sanitation, integrated control of non-communicable diseases, healthy regional arrangements and community-based total sanitation that have a significant influence on malaria’s disease in Indonesia.

Acknowledgement
Thank you to Mr. Anang Kurnia and Mr. Kusman Sadik for helping, giving input, guiding and directing the author in completing this paper.

References
[1] R Puasa, A Asrul and A Kader 2018 Identifikasi Plasmodium Malaria Didesa Beringin Jaya Kecamatan Oba Tengah Kota Tidore Kepulauan, Jurnal Riset Kesehatan
[2] Soedarto 2009 Penyakit Menular di Indonesia, Sagung Seto, Surabaya
[3] P P Ditjen and P L Kemenkes 2014 Pedoman Teknis Pemeriksaan Parasit Malaria, Jakarta
[4] A Agresti 2015 Foundation of Linear and Generalized Linear Models, New Jersey: Wiley
[5] S Das and K D Dey 2007 On Bayesian Analysis of Generalized Linear Models: A New Perspective, Statistical and Applied Mathematical Sciences Institute.
[6] A J Dobson and Barnett 2008 An Introduction to Generalized Linear Models, Third Edition, London and New York: Chapman and Hall
[7] Lopiano et al, 2013 Generalized Linear Models in R Frequentist and Bayesian Methods, https://www.samsi.info/wp-content/uploads/2016/03/Lopiano_glms_may2013.pdf
[8] A J Dobson 2002 An Introduction to Generalized Linear Models, Second Edition, London and New York: Chapman and Hall
[9] P McCullagh and J A Nelder 1989 Generalized Linear Models, Second edition, London: Chapman and Hall
[10] J A Nelder and R W M Wedderburn 1972 Generalized Linear Models, Journal of The Royal Statistical Society, Series A (General) 135(3)
[11] V M Santi 2017 Pengembangan Model Regresi Logistik Multinomial untuk Klasifikasi Politik pada Pemilihan Umum, Jurnal Statistika dan Aplikasinya, 2(1)
[12] J Starkweather 2011 Bayesian Generalized Linear Models in R, University oh North Texas.
[13] M-H Chen and J G Ibrahim 2002 Conjugate Prior for Generalized Linear Models, Statistica Sinica 13
[14] R Sultan and S P Ahmad 2012 Posterior Estimates of Poisson Distribution Using R Software, Journal of Modern Applied Statistical Method, 11
[15] S Untung et al. 2018 Profil Kesehatan Indonesia Tahun 2017, Jakarta: Kementerian Kesehatan Republik Indonesia.
Attachment. Data of Malaria Case in Indonesia in 2017.

| No. | PROVINCE                  | Y      | X1  | X2   | X3   | X4  | X5  | X6   | Total Population |
|-----|---------------------------|--------|-----|------|------|-----|-----|------|------------------|
| 1   | Aceh                      | 299    | 89.54 | 0.33 | 40.95 | 63.38 | 6   | 33.45 | 5189466          |
| 2   | Sumatera Utara            | 2442   | 195.42 | 0.34  | 43.59 | 73   | 17  | 23.18 | 14262147         |
| 3   | Sumatera Barat            | 512    | 126.66 | 0.31  | 60.98 | 52.77 | 19  | 45.42 | 5321489          |
| 4   | Riau                      | 211    | 76.51  | 0.33  | 28    | 70.04 | 11  | 63.58 | 6657911          |
| 5   | Jambi                     | 187    | 70.22  | 0.33  | 72.93 | 64.2  | 11  | 42   | 3515017          |
| 6   | Sumatera Selatan          | 911    | 90.26  | 0.37  | 59.25 | 66.36 | 14  | 51.93 | 8266983          |
| 7   | Bengkulu                  | 1031   | 97.11  | 0.35  | 39.78 | 42.71 | 8   | 50.3  | 1934269         |
| 8   | Lampung                   | 4297   | 239.42 | 0.33  | 70.3  | 52.89 | 10  | 47.31 | 8289577          |
| 9   | Kep. Bangka Belitung      | 95     | 87.12  | 0.28  | 100   | 83.56 | 7   | 93.61 | 1430865          |
| 10  | Kep. Riau                 | 357    | 253.93 | 0.36  | 46.67 | 86.33 | 5   | 44.23 | 2082694         |
| 11  | DKI Jakarta               | 111    | 15623.61 | 0.41   | 67.35 | 91.13 | 6   | 43.45 | 10374235         |
| 12  | Jawa Barat               | 328    | 1357.85 | 0.39  | 41.8  | 64.4 | 27  | 42.79 | 48037827         |
| 13  | Jawa Tengah              | 904    | 1044.43 | 0.37  | 59.79 | 71.84 | 35  | 70.84 | 34257865         |
| 14  | DI Yogyakarta            | 86     | 1200.76 | 0.44  | 88.43 | 89.4  | 5   | 98.86 | 3762167          |
| 15  | Jawa Timur               | 107    | 822.03  | 0.42  | 86.29 | 68.83 | 38  | 71.63 | 39292972         |
| 16  | Banten                   | 42     | 1288.24 | 0.38  | 68.22 | 71.68 | 6   | 77.43 | 12448160         |
| 17  | Bali                     | 33     | 734.69  | 0.38  | 38.14 | 90.51 | 9   | 70.53 | 4246528          |
| 18  | Nusa Tenggara Barat      | 765    | 266.83  | 0.38  | 62.42 | 69.25 | 10  | 97.01 | 4955578          |
| 19  | Nusa Tenggara Timur      | 30451  | 108.53  | 0.36  | 29.37 | 45.31 | 7   | 72.53 | 5287302          |
| 20  | Kalimantan Barat         | 152    | 33.48  | 0.33  | 53.36 | 49.65 | 8   | 31.36 | 4932499          |
| 21  | Kalimantan Tengah        | 760    | 16.97   | 0.33  | 57.51 | 45.46 | 2   | 59.2  | 2605274          |
| 22  | Kalimantan Selatan       | 1135   | 106.33  | 0.35  | 40.17 | 58.09 | 10  | 54.93 | 4119794          |
| 23  | Kalimantan Timur         | 1573   | 27.7    | 0.33  | 27.48 | 72.83 | 9   | 30.83 | 3575449          |
| 24  | Kalimantan Utara         | 65     | 9.16    | 0.31  | 25    | 66.59 | 4   | 21.37 | 691058           |
| 25  | Sulawesi Utara           | 900    | 177.67  | 0.39  | 40.98 | 71.93 | 14  | 15.99 | 2461028          |
| 26  | Sulawesi Tengah          | 543    | 47.97   | 0.35  | 40.53 | 61.12 | 7   | 39.07 | 2966325          |
| 27  | Sulawesi Selatan         | 1201   | 186.02  | 0.43  | 49.33 | 76.73 | 24  | 67.48 | 8690294          |
| 28  | Sulawesi Tenggara        | 596    | 68.36   | 0.4   | 32.43 | 69.52 | 9   | 36.13 | 2602389          |
| 29  | Gorontalo                | 46     | 103.77  | 0.41  | 32.61 | 58.75 | 6   | 48.15 | 1168190          |
| 30  | Sulawesi Barat           | 147    | 79.28   | 0.34  | 21.36 | 59.48 | 4   | 69.75 | 1330961          |
| 31  | Maluku                   | 4019   | 37.19   | 0.32  | 15.54 | 63.29 | 3   | 15.41 | 1744654          |
| 32  | Maluku Utara             | 957    | 37.81   | 0.33  | 21.26 | 66.18 | 3   | 21.19 | 1209342          |
| 33  | Papua Barat              | 13706  | 8.89    | 0.39  | 20.98 | 65.3 | 0   | 17.91 | 915361           |
| 34  | Papua                    | 192648 | 10.23   | 0.4   | 3.39  | 33.06 | 1   | 4.85  | 3265202          |

Source: Indonesian Health Profile Publication Book 2017 Ministry of Health of Indonesia.

Information:
Y = Number of malaria positive patients
X1 = Population density
X2 = GINI RATIO
X3 = Percentage of Public Health Centers implementing Integrated Non-Communicable Disease Control
X4 = Percentage of Households that have access to proper sanitation
X5 = Percentage of Districts / Cities that organize regional arrangements health
X6 = Percentage of villages that implement Community Based Total Sanitation