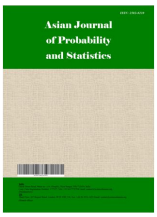

Asian Journal of Probability and Statistics

7(1): 71-86, 2020; Article no.AJPAS.55623
ISSN: 2582-0230

Parameter Estimation of Bayesian Multiple Regression Model with Informative Inverse Gamma Prior Distribution: Application to Malaria Symptom Dataset

Drinold Aluda Mbete[1*]

[1]*Department of Mathematics, Masinde Muliro University of Science and Technology, P.O. BOX 190-5000, Kakamega, Kenya.*

Author's contribution

The sole author designed, analyzed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/AJPAS/2020/v7i130174
Editor(s):
(1) Dr. S. M. Aqil Burney, University of Karachi, Pakistan.
Reviewers:
(1) Zimeras Stelios, University of the Aegean, Greece.
(2) Myron Hlynka, University of Windsor, Canada.
Complete Peer review History: http://www.sdiarticle4.com/review-history/55623

Received: 18 January 2020
Accepted: 23 March 2020
Published: 08 April 2020

Original Research Article

Abstract

Objectives: The study aims to develop a Bayesian multiple regression model with informative inverse gamma prior and fit the model to malaria symptom dataset.
Place and Duration of Study: The study was carried out in Masinde Muliro University of Science and Technology (MMUST). The study used 300 malaria related symptom dataset obtained from Health service records of different patients (students) between the time period of 1st January, 2015 to 20th December, 2015.
Methodology: Multiple linear regression model with Bayesian parameter estimation is used. The Normal prior distribution for θ parameter and inverse gamma prior distribution for the σ^2 parameter is derived. Gibbs sampler and Metropolis Hasting algorithm is used with Markov Chain Monte Carlo (MCMC) method to produce an iteration of about 102,491 with Burn-in of 2500 and thinning of 10 that resulting to effective sample size of 90000.
Results: The results shows that all the estimated posterior predictive p-values are between 0.05 and 0.95 indicating an adequate fit for the individual observation of the data in the model. The results also reveals that the data values and the average distance between the data values and the mean tend to be close to each other and the estimated coefficient of $\theta's$ approximately 95% draws fall within each of the corresponding highest posterior density intervals.

**Corresponding author: E-mail: drinoldmbete123@gmail.com;*

Conclusion: Though the Least Squares method is sufficient for estimating the coefficients of the regression parameters, the Bayesian estimates recorded comparatively very small standard errors making the Bayesian method more robust in analysing symptom dataset.

Keywords: Bayesian, regression, malaria; mcmc; symptoms; Plasmodium falciparum.

2010 Mathematics Subject Classification: 53C25, 83C05, 57N16.

1 Introduction

Malaria is an ancient disease that has been affecting people since the beginning of recorded time. It poses serious economic, social and health burdens in tropical and subtropical countries where it is predominantly found, ([1]). Malaria still remains a huge public health issue regardless of how many years of research has been conducted on how to combat it. According to WHO report [2], released in November 2017, the report shows that the number of malaria cases reported in the year 2016 was 216 million up from 211 million cases reported in 2015. The report also shows that malaria death estimates in 2016 stood at 445,000 compared to 446,000 deaths in 2015. The high burden of malaria cases in 2016 was in Africa at 90% with 91% cases of deaths reported in children. According to WHO report on malaria cases in Kenya, malaria is one of the leading causes of morbidity and fatality with about 3.5 million children at risk of developing severe malaria, out of which an estimated 34,000 children under five years die every year. The disease is also responsible for 30% of out-patient visits at health centres, economically, it is estimated that 170 million working hours are lost each year because of malaria illness, [3].

Symptoms are experienced deviations from an individual's perception of his or her normal healthy state of being, yet not necessarily an indicator of illness. A symptom can emerge from sensitivity to certain combinations of biological, social and environmental processes and vary in magnitude, severity, persistence and character. Symptoms can be subjectively reported or objectively observed. Depending on the disease, the scope and intensity, the duration of symptoms can vary over time.

The malaria symptoms can be grouped into two; symptoms for uncomplicated malaria (suspected malaria) and symptoms for complicated malaria (severe malaria). Malaria is considered uncomplicated when symptoms are present but there are no clinical or laboratory signs to indicate severity or vital organ dysfunction. The symptoms for uncomplicated malaria are non-specific i.e. they are self-reported symptoms that do not indicate a specific disease process, they are initial symptoms and include fever (high body temperature), chills, cough, headache, pains (joint, muscle, abdominal), muscle aches, loose stool, tiredness, nausea, high pressure, vomiting and diarrhoea. Infection with *Plasmodium falciparum* if not promptly treated can quickly progress to complicated malaria (severe malaria). The main symptoms for severe malaria include coma, severe breathing difficulties, low blood sugar, hallucination, prostration, immobility, confusion and incoherent speech, seizure, loss of consciousness, hyperparasitaemia, black quarter urine and low blood haemoglobin, [4].

2 Literature Review

Numerous mathematical models have been developed and carried out to gain insight into the transmission dynamics and control of malaria transmission in human population. Different approaches are helpful in guiding different stages of the disease through mathematical models that study transmission of malaria based on the reproduction number. A number of mathematical models have been developed and analysed to explain the dynamics of infectious diseases in humans. Many of

these models are described by systems of ordinary differential equations formulated under reasonable assumptions and parameters.

Mathematical modelling of malaria has a long history starting with the first models of malaria transmission dynamics by [5]. Ross introduced the first deterministic differential equation model of malaria by dividing the human population into susceptible and infected compartments, with the infected class returning to susceptible class again leading to the SIS structure. The Ross model outlines the basic features of malaria transmission and puts the main burden of transmission on mosquito-specific features thereby paving the way for mosquito-based malaria control programmes. The simple Ross model did not consider the latency period of the parasite in mosquitoes and their survival during the latent period. This resulted in MacDonald model which considered the latency period and introduced the Exposed class in the mosquitoes, [6]. In a natural extension to the Ross and McDonald's models, Anderson and May considered the 21 days latency period of the parasite in humans and introduced the Exposed class in human population in their model, [7]. Their model divided the host population into three compartments i.e susceptible, exposed and infected along with that in the mosquito population (susceptible, exposed and infected). Therefore, their model consisted of four differential equations describing the time evolution of both the exposed and infected classes for humans population and mosquito's population.

A study by [8] examined the relationship between malaria, environmental and socio-economic variables in Sudan using health production modified model where regression analysis method was used to analyse their model, their results showed a significant relationships between malaria, rainfall and water bodies while other variables such as Human Development Index, temperature, population density and percent of cultivated areas were not significant. [9] used robust Poisson regression model in his study to model the daily average number of cases in 10 districts of Ethiopia that was associated with rainfall, minimum temperature and maximum temperature as explanatory variable in a polynomial distributed lag model. In order to improve reliability and generalizability within similar climatic conditions, he grouped the districts into two climatic zones i.e hot zones and cold zones. The results showed that in hot zones, malaria was associated with rainfall and minimum temperature at relatively shorter lags whereas in cold districts, rainfall was associated with a delayed increase in malaria cases. The results also showed that in cold districts, minimum temperature was associated with malaria cases with a delayed effect while in hot districts, the effect of minimum temperature was non-significant at most lags, and much of its contribution was relatively immediate.

In many studies of medical treatment, symptoms are measured repeatedly over time in observation called longitudinal observation. Though we cannot observe directly latent variables, for instance, state of individual in case of infectious disease, we learn about it by measuring symptom(s). For the longitudinal models, two latent variables govern disease, one for the probability of experiencing a particular symptom and another for the severity of the experienced symptom. Thus the probability of a symptom and the severity of it depends on both latent variables and observed variables, [10]. Latent variable link observable data in the real world to symbolic data in the model. Bayesian statistics is often used for inferring latent variables, the common method used inferring latent variables in Bayesian statistics are; Hidden Markov Model (HMM), factor analysis, principal component analysis and Expectation Maximization (EM) algorithm, [10]. [11] developed an intra-individual consistency model using a logistic-type latent variable model. The latent variable in the model is also used to represent the propensity of symptoms and intensity of episodes as these could not be observed directly and need to be estimated through observation of symptom episodes in hypoglycaemia. Their model results showed that their was individual difference in symptom reporting and adults exhibit distinct intra-individual variability in symptom reporting. [12] extended the model developed by Zammit by allowing for different forms of symptom experiencing thresholds between groups variability when symptoms are classified in groups. Bayesian estimation was performed for all coefficients in the developed model without grouped symptoms and with grouped

symptoms. Their analysis showed that a multiplicative form of symptom propensity and episode intensity provides the most suitable symptom experiencing threshold. It also showed that groups of symptoms distinct propensity had significant impact on the consistency of symptom reporting especially on gender subjects.

3 Materials and Methods

3.1 The study area

The study was conducted in Masinde Muliro University of Science and Technology (MMUST) located in Kakamega Town, Kakamega County with an altitude of 1561m above the sea level with a student population of approximately 30000. The levels of malaria risk and transmission intensity in MMUST exhibit significant spatial and temporal variability related to variations in amount of rainfall, temperature, altitude, topography and human settlement pattern around the college. In this study area, malaria situation is typical of Sub-Saharan Africa making its transmission an all-year-round affair and seasonal variation. The MMUST Health service records shows that between 300-700 cases of malaria are reported each month which constitutes 75% of all out-patient cases reported in the health service. The peak period of malaria incidence occurs from April to August following the long rain season. The malaria cases reported in MMUST is either complicated (severe) malaria or uncomplicated malaria. For complicated malaria, the following symptoms are commonly displayed by students; hallucination, prostration, loss of consciousness, hyperparasitaemia, pallor, convulsions, low and high blood pressure, coma, convulsions, low and high pulse beat/min, anaemia and dark urine. For uncomplicated malaria, the following non-specific symptoms are commonly displayed by the students; headache, pains (joint, muscle, abdominal), loose stool, cough, fever, rigors, nausea and vomiting. For confirmatory test of malaria, blood slide (BS) for malaria parasite is carried out.

3.2 The model

In this section, malaria statistical model is developed and analyzed. The model subdivides the human population under study into classes depending on the disease status of the individual and observed symptoms. In this study, the explanatory variable are observable symptoms recorded by the health officer for each student who visited the health facility with malaria related symptoms and the response variable is the transition parameter represented by state of individual after observed symptoms i.e mild, moderate and severe. The observed symptoms are; X_1 - fever (high body temperature), X_2 - rigors, X_3 - convulsion, X_4 - sweating, X_5 - vomiting, X_6 - diarrhoea, X_7 - pallor, X_8 - cough and X_9 - prostration. The observed symptoms are then grouped and recorded in a scale based on the severity of the disease. For instance 0 implies no symptoms, 1 implies mild symptoms, 2 implies moderate and 3 implies severe symptoms. To aid our discussion, we first provide an overall modelling framework. Figure (3.2) provides a transition diagram on how an n^{th} individual susceptible to malaria evolves.

Figure 1 shows how an n^{th} individual can become infected at time t given that he/she was not infected at time $t-1$. The process in which an individual n becomes infected is denoted by Z_{nt}, thus when in state S at time $t-1$, the process is Z_{nt-1}. When an individual n transit from S to I, the process become Z_{nt}. Within state I, the process evolves with time to three different categories namely; mild (I_1), moderate (I_2) and severe (I_3).

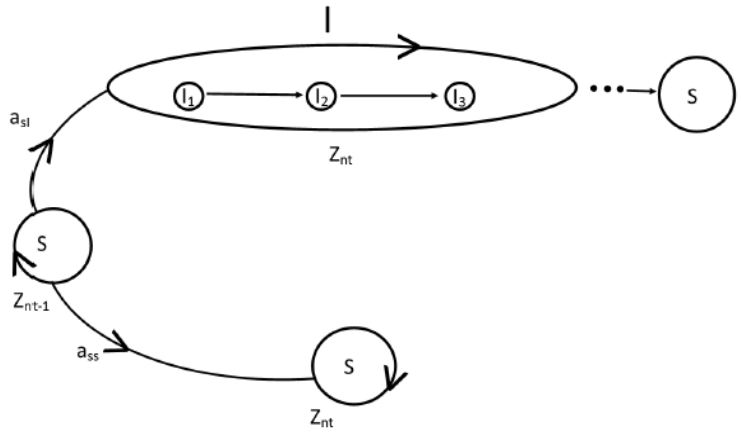

Fig. 1. Malaria transition diagram

3.3 Computation of transition probability

In this section, the transition probabilities shown in Figure (3.2) which are vital to our model are computed. The transition probabilities are:

$$\begin{aligned} a_{SS}^{(t)} &= P(Z_{nt} = S|Z_{nt-1} = S), \\ a_{SI}^{(t)} &= P(Z_{nt} = I|Z_{nt-1} = S), \end{aligned} \tag{3.1}$$

and

$$a_{II}^{(t)} = P(Z_{nt} = I|Z_{nt-1} = I)$$

To calculate the probabilities of $a's$, we require probit models. This model simply use the cumulative Gaussian normal distribution to calculate the probability of being in one state/category or not. For instance

$$a_{SI}^{(t)} = P(Z_{nt} = I|Z_{nt-1} = S) = \int_{-\infty}^{\psi_{nt}} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}t^2 dt = \phi(\psi_{nt}) \tag{3.2}$$

where ϕ is the cumulative standard normal distribution and the upper bound parameter ψ_{nt} is a transition parameter (i.e., $\psi_{nt} \in \Re$) which defines the transition of n^{th} student from a state of susceptibility at time $t-1$ to a state of illness at time t.

Similarly other probabilities of $a's$ can be computed in the same manner. For example

$$a_{SS}^{(t)} = P(Z_{nt} = S|Z_{nt-1} = S) = \int_{-\infty}^{\psi_{nt}^c} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}t^2 dt \tag{3.3}$$

where ψ_{nt}^c is the complement of ψ_{nt}. Rewriting the compliment, we have

$$a_{SS}^{(t)} = P(Z_{nt} = S|Z_{nt-1} = S) = \int_{\Re-\psi_{nt}}^{\Re} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}t^2 dt$$

which implies that

$$a_{SS}^{(t)} = P(Z_{nt} = S|Z_{nt-1} = S) = \int_{-\Re}^{\Re} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}t^2 dt - \int_{-\infty}^{\psi_{nt}} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}t^2 dt$$

upon simplification, we have

$$a_{SS}^{(t)} = P(Z_{nt} = S|Z_{nt-1} = S) = 1 - \int_{-\infty}^{\psi_{nt}} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}t^2 dt$$

Thus equation (3.3) becomes

$$a_{SS}^{(t)} = P(Z_{nt} = S|Z_{nt-1} = S) = \int_{-\infty}^{\psi_{nt}^c} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}t^2 dt = 1 - \phi(\psi_{nt}) \tag{3.4}$$

3.4 Computation of ψ_{nt}

The transition of n^{th} student from a state of susceptibility at time $t-1$ to a state of illness at time t is due to observable malaria related symptoms. Therefore we suspect the transition to be caused by the following relationship

$$\psi_{nt} = \delta_n + X_{n1}\theta_1 + X_{n2}\theta_2 + X_{n3}\theta_3 + X_{n4}\theta_4 + X_{n5}\theta_5 + X_{n6}\theta_6 + X_{n7}\theta_7 + X_{n8}\theta_8 + X_{n9}\theta_9 + \epsilon_n. \tag{3.5}$$

It is convenient to write the model in Equation (3.5) in matrix notation as

$$\Psi_{nt} = \delta + [X_{n1} + X_{n2} + ... + X_{np}] \begin{bmatrix} \theta_1 \\ \theta_2 \\ . \\ . \\ . \\ \theta_p \end{bmatrix} + \epsilon_i \tag{3.6}$$

Hence

$$\psi_{nt} = \delta + X\theta + \epsilon \tag{3.7}$$

where $\delta = (\xi_1, \xi_2, \xi_3)$ i.e., ξ_1 is the general trend for individual to be infected with malaria, ξ_2 is the climatic condition (Rainfall, relative humidity, temperature) accelerating the spread of malaria to an individual and ξ_3 is the environmental and socio-economic condition that affects an individual. θ is a $p \times 1$-vector of unknown regression parameters.

$\epsilon_n$is the unobserved random variable with expected value 0.
X is $(n \times p)$ vector of covariates where p is the number of symptoms ($p = 9$) and X is the observed symptoms i.e X_1 - fever (body temperature), X_2 - rigors, X_3 - sweating, X_4 - vomiting, X_5 - diarrhea, X_6 - weakness, X_7 - pallor, X_8 - cough and X_9 - prostration.

In this study, our focus is on the transition parameter ψ_{nt} which determines the state of n^{th} individual. The transition parameter ψ_{nt} is a continuous variable that depends on observed symptoms (vector of covariates) henceforth defined in Equation (3.7). Therefore equation (3.2) can now be written as

$$a_{SI}^{(t)} = \phi(\delta + X\theta + \epsilon) \tag{3.8}$$

3.5 Estimation of θ

In this study, δ_n is assumed to be known. Therefore Equation (3.7) can be written as

$$\psi_{nt} = X\theta + \epsilon \tag{3.9}$$

where the random error ϵ is assumed to be randomly distributed with mean 0 and unity variance i.e $\epsilon \sim N(0, 1)$. Assume also that the error term is the source of the randomness in the model which

implies that ψ_{nt} is also random i.e $\psi_{nt} \sim N(X\theta, 1)$.
Therefore the density function for the n^{th} observation is given as

$$f(\psi_{nt}|X,\theta) = (2\pi)^{\frac{-1}{2}} \exp -\frac{1}{2}(\psi_{nt} - X\theta)^{/}(\psi_{nt} - X\theta) \tag{3.10}$$

using the proportionality sign ($\propto$), the term that does not involve θ is not written. Therefore Equation (3.10) becomes

$$f(\psi_{nt}|X,\theta) \propto \exp -\frac{1}{2}(\psi_{nt} - X\theta)^{/}(\psi_{nt} - X\theta) \tag{3.11}$$

Suppose that $g(\theta)$ is assumed to be the conjugate prior of Equation (3.10) i.e $g(\theta) \sim N(0, \sigma^2)$. Then the distribution of $g(\theta)$ is given as

$$g(\theta) = (2\pi\sigma^2)^{\frac{-1}{2}} \exp -\frac{1}{2\sigma^2}(\theta)^{/}(\theta) \tag{3.12}$$

drop the term that does not involve a function of θ

$$g(\theta) \propto \exp -\frac{1}{2\sigma^2}(\theta)^{/}(\theta) \tag{3.13}$$

therefore, the posterior distribution of $P(\theta|\psi_{nt}, X)$ is obtained by multiplying Equation (3.11) and Equation (3.13)

$$P(\theta|\psi_{nt}, X) \propto \exp -\frac{1}{2}(\psi_{nt} - X\theta)^{/}(\psi_{nt} - X\theta) \exp -\frac{1}{2\sigma^2}(\theta)^{/}(\theta) \tag{3.14}$$

expand the term in the exponent

$$P(\theta|\psi_{nt}, X) \propto \exp -\frac{1}{2}\{(\psi_{nt}^{/}\psi_{nt} - \psi_{nt}^{/}X\theta - \theta^{/}X^{/}\psi_{nt} + \theta^{/}X^{/}X\theta + \frac{1}{\sigma^2}(\theta)^{/}(\theta)\}$$

drop the term that does not involve a function of θ

$$P(\theta|\psi_{nt}, X) \propto \exp -\frac{1}{2}\{(-\psi_{nt}^{/}X\theta - \theta^{/}X^{/}\psi_{nt} + \theta^{/}X^{/}X\theta + \frac{1}{\sigma^2}(\theta)^{/}(\theta)\} \tag{3.15}$$

rewrite Equation (3.15) as

$$P(\theta|\psi_{nt}, X) \propto \exp -\frac{1}{2}\{(-\psi_{nt}^{/}X\theta - \theta^{/}X^{/}\psi_{nt} + \theta^{/}(X^{/}X + \frac{1}{\sigma^2})\theta\} \tag{3.16}$$

upon simplification we get

$$P(\theta|\psi_{nt}, X) \propto \exp -\frac{1}{2}\{(\theta - (X^{/}X + \frac{1}{\sigma^2})^{-1}(X^{/}X + \frac{1}{\sigma^2})(\theta - (X^{/}X + \frac{1}{\sigma^2})^{-1}X^{/}\psi_{nt}\}$$

which implies that expected value of θ to be

$$E[\theta] = (X^{/}X + \frac{1}{\sigma^2})^{-1}X^{/}\psi_{nt} \tag{3.17}$$

thus, the Bayes estimate of $\hat{\theta}$ is

$$\hat{\theta} = (X^{/}X + \frac{1}{\sigma^2})^{-1}X^{/}\psi_{nt} \tag{3.18}$$

3.6 Estimation of σ^2

Since ψ_{nt} is assumed to be a random variable, let ψ_{nt} be randomly distributed with mean 0 and variance σ^2 i.e $\psi_{nt} \sim N(0, \sigma^2)$. Therefore the density function for the n^{th} observation is given as

$$f(\psi_{nt}|2\sigma^2) = (\frac{1}{\pi})^{\frac{1}{2}}(\frac{1}{2\sigma^2})^{\frac{1}{2}} \exp -\frac{1}{2\sigma^2}(\psi_{nt})^2 \tag{3.19}$$

drop the term that does not involve a function of σ^2

$$f(\psi_{nt}|\sigma^2) \propto (\frac{1}{2\sigma^2})^{\frac{1}{2}} \exp -\frac{1}{2\sigma^2}\psi_{nt}^2 \tag{3.20}$$

Let $\tau = (\frac{1}{2\sigma^2})$. Then Equation (3.20) becomes

$$f(\psi_{nt}|\sigma^2) \propto \tau^{\frac{1}{2}} \exp -\tau\psi_{nt}^2 \tag{3.21}$$

let $r \triangleq \frac{1}{2}$ then Equation (3.21) can be written as

$$f(\psi_{nt}|\sigma^2) \propto \tau^{r} \exp -\tau\psi_{nt}^2 \tag{3.22}$$

suppose the conjugate prior of τ is the gamma density $G(\alpha, \beta)$ i.e

$$p(\tau) = \frac{\beta}{\Gamma\alpha}\tau^{\alpha-1} \exp -\tau\beta \tag{3.23}$$

therefore the posterior distribution $P(\tau|\psi_{nt}$ is obtained by multiplying Equation (3.22 and Equation (3.23)i.e

$$p(\tau|\psi_{nt}) \propto \tau^{r} \exp -\tau\psi_{nt}^2 \times \frac{\beta}{\Gamma\alpha}\tau^{\alpha-1} \exp -\tau\beta \tag{3.24}$$

drop the term that does not involve a function of τ

$$p(\tau|\psi_{nt}) \propto \tau^{r} \exp -\tau\psi_{nt}^2 \times \tau^{\alpha-1} \exp -\tau\beta \tag{3.25}$$

multiply the like terms together

$$p(\tau|\psi_{nt}) \propto \tau^{r} + \alpha - 1 \exp -\tau(\psi_{nt}^2 + \beta) \tag{3.26}$$

Therefore

$$p(\tau|\psi_{nt}) \propto G(r + \alpha - 1, \psi_{nt}^2 + \beta) \tag{3.27}$$

which implies that expected value to be

$$E[\tau] = \frac{r + \alpha - 1}{\psi_{nt}^2 + \beta} \tag{3.28}$$

and variance to be

$$Var[\tau] = \frac{r + \alpha - 1}{\psi_{nt}^2 + \beta} \tag{3.29}$$

using the defination of τ

$$\sigma^2 = \frac{1}{2\tau}$$

therefore the Bayes estimate of σ^2 is

$$\sigma^2 = \frac{1}{2\tau} = \frac{\psi_{nt}^2 + \beta}{2r + 2\alpha - 2} \qquad provided \quad 2r + 2\alpha > 2 \tag{3.30}$$

This corresponds to the limiting case when $\alpha = 0$ and $\beta = 0$, hence the Bayes estimator of $\hat{\sigma^2}$ is

$$\hat{\sigma^2} = \frac{\psi_{nt}^2}{r - 1} \qquad provided \quad r > 1 \tag{3.31}$$

4 Results and Discussion

4.1 Description of data

For this study, unstructured test data was used for analysis, the symptoms were identified for each individual student through doctor's interrogation and student narration. The following symptoms were identified; fever, rigors, convulsion, prostration, vomiting, cough, diarrhoea, pallor and sweating. The symptoms were then grouped and recorded on an ordinal scale of 0 to 3 with 0 being no symptoms and 3 being maximum symptoms. Based on the malaria cases within the institution, this study used several independent variables ($X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$ and X_9) and one dependent variable (Y). Here are the variables:
(X_1=fever
(X_2=rigors
(X_3=convulsion
(X_4=sweating
(X_5=vomiting
(X_6=diarrhoea
(X_7=pallor
(X_8=cough
(X_9=prostration
Y= transition parameter

These description is summarized using descriptive statistics shown in Table 1.

Table 1. Descriptive statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
X_1	300	1.806667	.901108	0	3
X_2	300	1.716667	1.194912	0	3
X_3	300	1.743333	1.250066	0	3
X_4	300	0.79	.4079888	0	1
X_5	300	1.903333	1.047687	0	3
X_6	300	1.23	1.077483	0	3
X_7	300	1.8	1.075728	0	3
X_8	300	1.533333	1.214002	0	3
X_9	300	0.34	0.4951943	0	3
Y	300	1.923333	1.217537	0	3

Table 1 shows data values and average distance between the data values. Generally, the results shows that the mean tends to be very close to each other. For instance fever (mean=1.806667, SD=0.901108) implies that 18.1% of the observed symptoms with non missing values displayed from the data were fever with a variation of 0.901108. Standard deviation results shows how close the values are from the mean value, for instance, the value of 1.12 (cough) implies that the individual response on average was a little over 0.3 point away from the mean.

Before carrying out the analysis, the study sought to find out whether there was multicollinearity in the dataset. Multicollinearity was tested using variance inflation factor (VIF) and results are shown in Table 2. For instance, a high VIF i.e A VIF > 10 or a $\frac{1}{VIF} < 0.1$ indicates high collinearity between the associated independent variables with the other variables in the model. The results also shows that the VIF values less than 10 imply that the data values are free of multicollinearity

and inference obtained using data values are reliable. For instance, a VIF of 1.74 indicates that the variance of estimated coefficient is 1.74 times higher due to correlation between the independent variables. Added-variable plots (avplots) shown in Figure 2 are also used to depict the relationship between state of individual, observed symptoms and adjusting effects of other symptoms in order to uncover the observation exerting a disproportionate influence on the regression (outlier). High leverage observation displayed in Figure 2 shows the symptom that influences the coefficient values and this symptom can be seen to be horizontally distant from the rest of the other data values.

Table 2. Variance Inflation factor

Variable	VIF	$\frac{1}{VIF}$
X_1	3.74	0.267446
X_2	2.67	0.374987
X_3	2.06	0.484668
X_4	1.58	0.632941
X_5	1.51	0.661853
X_6	1.07	0.932220
X_7	1.04	0.960790
X_8	1.02	0.978364
X_9	1.01	0.993272
Mean VIF	1.74	

Before carrying out analysis using regression model, the first step is to carry out correlation test to determine whether there is a relationship between dependent variable with all independent variables i.e
$H_0 : \rho = 0$ (There is no correlation between two variables)
$H_1 : \rho \neq 0$ (There is correlation between two variables)

The correlation test results between independent variables with dependent variable are shown in Table 3, the results From shows that the independent variables are correlated with the dependent variable. This is indicated by p-value results which are less than α, where α value is 0.05. This implies that the correlation value is not equal to zero i.e the variables are correlated to each other. Therefore there exist a relationship between independent variables with dependent variable and that the data values can be analysed using regression model.

Table 3. Correlation test

Variable	correlation coefficient	p-value
X_1	0.1419	0.0139
X_2	0.6448	0.0000
X_3	0.8726	0.0000
X_4	-0.0191	0.0424
X_5	0.5605	0.0000
X_6	0.0874	0.0309
X_7	0.8437	0.0000
X_8	0.5776	0.0000
X_9	0.1710	0.0030

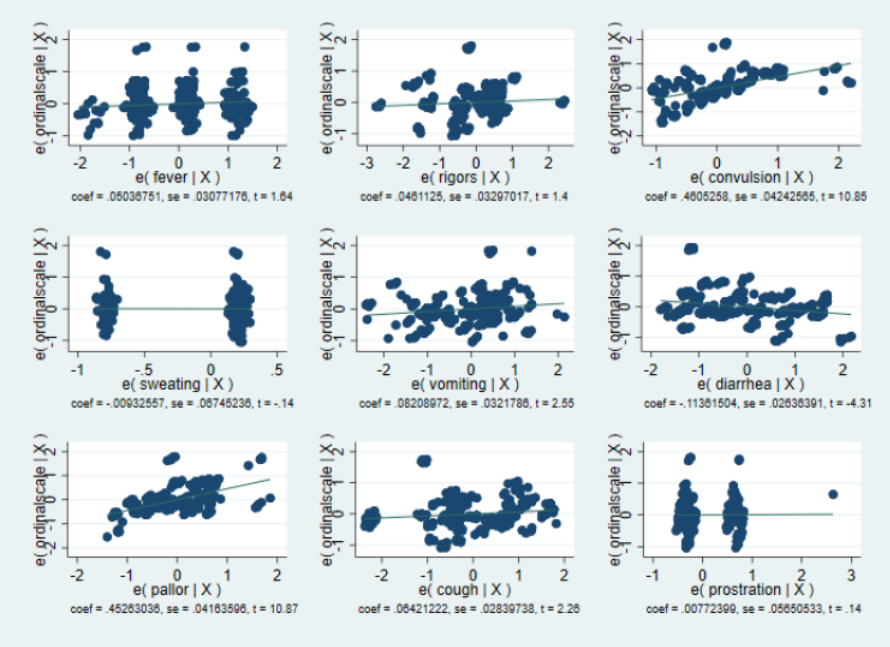

Fig. 2. avplots plot for symptom dataset

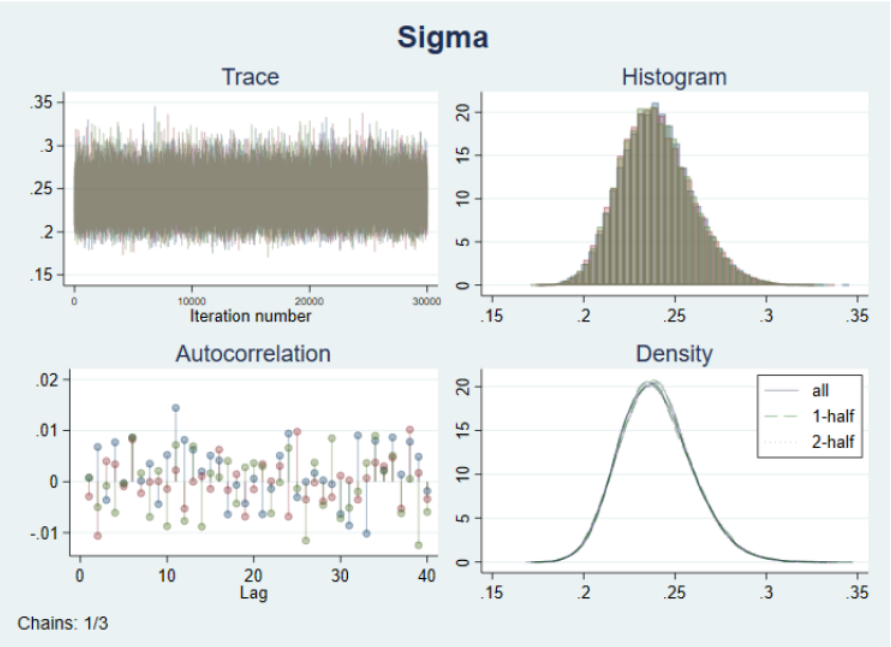

Fig. 3. Bayesgraph Diagnostics

Table 4. Linear regression model using OLS method

Source	SS	df	MS	Number of obs	=	300
				F(9, 290)	=	186.74
Model	378.009585	9	42.001065	Prob > F	=	0.0000
Residual	65.2270818	290	0.224920972	R-squared	=	0.8528
				Adj R-squared	=	0.8483
Total	443.236667	299	1.48239688	Root MSE	=	.47426
Y	Coef.	Std. Err.	t	P> $\|t\|$	[95% CI]	
X_1	.0503675	.0307718	1.64	0.003	-.0101968	.1109318
X_2	.0461125	.0329702	1.40	0.163	-.0187787	.111003
X_3	.4605258	.0424257	10.85	0.000	.3770246	.544027
X_4	-.0093256	.0674524	-0.14	0.890	-.1420838	.1234327
X_5	.0820897	.0321786	2.55	0.011	.0187565	.1454229
X_6	-.113615	.0263639	-4.31	0.000	-.1655039	-.0617262
X_7	.4526304	.041636	10.87	0.891	.3706834	.5345773
X_8	.0642122	.0283974	2.26	0.024	.0083211	.1201033
X_9	.007724	.0565053	0.14	0.000	-.1034886	.1189365
intercept	.0253763	.1020134	0.25	0.804	-.1754043	.2261569

From the estimate value of the parameter coefficients in Table 4, the regression model results shows that shows that 7 variables have positive influence and 2 variables have negative influence. The results also show that Convulsion, prostration, vomiting, diarrhoea, fever and pallor variables have significantly contributes to the transition of individual from one state to the other state. This is based on the result ($F(9.290) = 186.74, p < .005$). Based also on the result of $R^2(R^2 = 0.8528)$ value, the five predictor variables are able to explain 85.3% of the variance in the model. The regression results also shows that for each one-point increase in the transition outcome, sweating and diarrhoea decreases by -0.0093256 and -0.113615 respectively. The Standard Error results(SE) indicates the reliability mean whereby small SE results indicates that the sample mean is more accurate reflection of the actual population mean. The regression result indicate that 7 variables of the 9 independent variables affect the transition of an individual from one state to the other state. Therefore, the model in Equation (3.9) can be written as;

$$Y = 0.0254 + 0.0.0504X_1 + 0.0461X_2 + 0.04605X_3 - 0.0093X_4 + 0.00821X_5 - 0.1136X_6 + 0.4526X_7 + 0.0642X_8 + 0.0077X_9$$

The assumptions test result on the regression model showed that residual is not normally distributed, not independent and not identical. Therefore the IID Normal assumptions on the Ordinary least square(OLS) regression model is not met. The VIF test for the model shows that their is multicollinearity in the model. Based on this results, the multiple linear regression model using the OLS parameter estimation method is not a suitable method for parameter estimation. Therefore in this study, multiple linear regression model with Bayesian parameter estimation is used to find parameter estimates as it treats all the model parameters as random quantities and enable one to make probability statements about the likely values of parameters and assign probabilities of interest. The prior distribution used in this study is the Normal distribution for θ parameter and the inverse gamma distribution for the σ^2 parameter. Gibbs sampler and Metropolis Hasting algorithm is used with Markov Chain Monte Carlo (MCMC) method to produce iteration of about 102,491 with Burn-in of 2500 and thinning of 10 resulting to effective sample size of 90000 for inference so as to eliminate potential problems due to autocorrelation. The convergence of the Markov chain Monte Carlo chains was monitored using trace plots, autocorrelation function (ACF), and Gelman-Rubin diagnostics. The Metropolis Hasting (MH) algorithm achieve an overall rate AR of 34% and an average efficiency of about 99% as shown in Table 5. Therefore the Bayesian normal regression model in Equation (3.14) was fitted to data so as to obtain the parameter estimates. Based on Equation (3.18) and Equation (3.31), the estimate of mean vector $\hat{\theta}$ and variance $\hat{\sigma^2}$ were computed from the same data as follows respectively. The results obtained is shown in Equation (4.1) and

Equation (4.2).

$$\hat{\theta} = \begin{bmatrix} 0.0259496 \\ 0.0502708 \\ 0.0460683 \\ 0.4605055 \\ -0.0096916 \\ 0.0818986 \\ -0.1135517 \\ 0.4526221 \\ 0.0643508 \\ 0.0077789 \end{bmatrix} \tag{4.1}$$

similarly , based on Equation (3.31) estimate of variance was computed and result obtained as follow;

$$\hat{\sigma^2} = \begin{bmatrix} 0.23956 \end{bmatrix} \tag{4.2}$$

The model of linear regression analysis with Bayesian approach is

$Y = 0.0254 + 0.0.0504X_1 + 0.0461X_2 + 0.04605X_3 - 0.0093X_4 + 0.00821X_5 - 0.1136X_6 + 0.4526X_7 + 0.0642X_8 + 0.0077X_9$

Table 5. Bayesian simulation results

Name		value
Number of chains Per MCMC	=	3
Iterations	=	102,491
Burn-in	=	2,500
Sample size	=	10, 000
Number of observation	=	300
Average acceptance rate	=	0.3394
Average efficiency:minimum	=	.9927
Average efficiency:maximum	=	.9993
Maximum Gelman-Rubin Rc	=	1.176
Average marginal likelihood	=	-231.54658

From Table 6, it can be seen that estimated coefficient $\hat{\theta}$ are almost the same for the Ordinary Least Square (OLS) and the Bayesian model but the estimate of Bayesian model are smaller. The estimated coefficient of $\theta's$ approximately 95% draws fall within each of the corresponding highest posterior density intervals (HPD). Table 7 shows the efficiency summaries, The closer ESS (effective sample size) are to the MCMC sample size, the less correlated the MCMC sample is and more precise our estimates of parameters are. Values below 1% of the MCMC sample size indicate a problem in efficiency. From the results, the efficiency estimates are more than 99% indicating a good ESS for parameter estimates. Table 7 also shows the correlation times which are viewed as estimates of autocorrelation lags in the MCMC samples, for instance, the correlation times of the coefficients is 1 indicating a perfect correlation among MCMC sample size.

For convergence of MCMC, graphical diagnostic plot are plotted for the coefficient as shown in Figure 3. The displayed diagnostic include a trace plot, an autocorrelation plot, histogram and

Kernel density estimate overlaid with densities estimated using the first and second halves of the MCMC sample. A graphical summary for the variance parameter does not show any obvious problems. The trace plot reveals a good coverage of the domain of the marginal distribution, while the histogram and kernel density plots resemble the shape of an expected inverse-gamma distribution.

Table 6. Linear regression model using Bayesian approach

Y	Mean	Std. Dev.	MCSE	Median	HPD[95%	Cred.Interval]
X_1	.056306	.0317274	.000106	.0503697	-.0110428	.1130387
X_2	.0460683	.0339402	.000113	.0459884	-.0193016	.1131818
X_3	.4605055	.0437595	.000146	.4605534	.3747775	.5460086
X_4	-.0096916	.069681	.000232	-.0094585	-.1461358	.1256172
X_5	.0818986	.0331668	.000111	.0820614	.0190056	.1488435
X_6	-.1135517	.0271493	.00009	-.1134825	-.1663481	-.0601344
X_7	.4526221	.0432297	.000144	.4525424	.3671538	.5370451
X_8	.0643508	.0292297	.000097	.0643512	.0052016	.1200136
X_9	.0077789	.0584017	.000195	.0078181	-.1069035	.1222403
intercept	.0259496	.1051948	.000351	.0260526	-.1858322	.2276062
σ^2	.23956	.0198361	.000066	.2384632	.2019022	.278897

Table 7. Efficiency Summaries

	Number of chains	=	3
	MCMC sample size	=	90,000
	Efficiency:	min =	.9927
		avg =	.9993
		max =	1
Y	ESS	Corr. time	Efficiency
X_1	90000.00	1.00	1.0000
X_1	90000.00	1.00	1.0000
X_1	90000.00	1.00	1.0000
X_1	90000.00	1.00	1.0000
X_1	89339.73	1.01	0.9927
X_1	90000.00	1.00	1.0000
X_1	90000.00	1.00	1.0000
X_1	90000.00	1.00	1.0000
X_1	90000.00	1.00	1.0000
intercept	90000.00	1.00	1.0000
σ^2	90000.00	1.00	1.0000

To determine whether the model fits well and predictions of future observations. Samples are drawn from the posterior predictive distribution of Y. $\_ysim1$ is specified using bayespredict so as to simulate the outcome values and use a random-number seed for reproducibility as shown in Table 8. The first column contains posterior means, MCMC estimates of the expected outcome observations with respect to the posterior predictive distribution.

To access the goodness of fit of the model, the results of the replicated outcome samples are compared with observed outcome using posterior predictive p-values as shown in Table 9. The posterior predictive p-values are typically computed for functions of the data. The results shows that all the estimated posterior predictive p-values are between 0.05 and 0.95 and thus indicate adequate fit for the individual observation.

Table 8. Posterior summary statistics

				Number of chains	=	3
				MCMC sample size	=	90,000
	Mean	Std. Dev.	MCSE	Median	[95%	Equal-tailed Cred. Interval]
_ysim1_1	1.770065	.4952522	.001651	1.771355	.7998321	2.739653
_ysim1_2	.5299451	.4970351	.001657	.5290812	-.443362	1.50753
_ysim1_3	.4200158	.494928	.00165	.4206923	-.5531654	1.388663
_ysim1_4	3.085513	.5006306	.001669	3.084935	2.102757	4.068597
_ysim1_5	3.231541	.495178	.001651	3.234449	2.261567	4.205788
_ysim1_6	.1144848	.4959197	.001653	.1128442	-.8576428	1.082826
_ysim1_7	2.056877	.503976	.00168	2.056064	1.073278	3.048767
_ysim1_8	3.406604	.4959012	.001653	3.405427	2.436004	4.383048
_ysim1_9	-.0786547	.4944269	.001648	-.0785352	-1.05821	.8909879
_ysim1_10	3.327874	.498201	.001661	3.329099	2.34821	4.302601
_ysim1_11	3.152681	.4960874	.001654	3.153086	2.179684	4.118937

Table 9. posterior predictive summary statistics

		Number of chains	=	3
		MCMC sample size	=	90,000
T	Mean	Std. Dev.	E(T_{obs})	P(T>= T_{obs})
_ysim1_1	1.770065	.4952522	2	.3203333
_ysim1_2	.5299451	.4970351	0	.8569778
_ysim1_3	.4200158	.494928	1	.1205778
_ysim1_4	3.085513	.5006306	3	.5673889
_ysim1_5	3.231541	.495178	3	.6800556
_ysim1_6	.1144848	.4959197	0	.5918
_ysim1_7	2.056877	.503976	2	.5442222
_ysim1_8	3.406604	.4959012	3	.7950889
_ysim1_9	-.0786547	.4944269	0	.4373778
_ysim1_10	3.327874	.498201	3	.7463778
_ysim1_11	3.152681	.4960874	3	.6222778

5 Conclusions

From Table 4 and Table 6, it can be seen that estimated coefficients of θ are almost the same for the Least Squares model and the Bayesian model though the Bayesian estimates recorded comparatively very small errors making the Bayesian method more robust. The study reveals that, though the Least Squares method is sufficient for estimating the coefficients of the regression parameters, the Bayesian estimates recorded comparatively very small standard errors making the Bayesian method more robust. The use of additional information provided by the assumption of univariate normal prior distribution of the θs accounted for the smaller standard errors of the Bayesian estimates. For this reason, we recommend the use of Bayesian approach for predicting and estimating parameters of interest in the context of symptom(s) dataset.

Acknowledgment

The author is grateful to the referees for their careful reading, constructive criticisms, comments and suggestions, which have helped us to improve this work significantly.

Competing Interests

The author has declared that no competing interest exist.

References

[1] Mandal S, Sarkar RR, Sinha S. Mathematical models of malaria–a review. Malaria Journal. 2011;10:202.

[2] WHO. Malaria Fact Sheet 2017 Report; 2017.

[3] WHO. World Malaria Report; 2012.

[4] Dondorp AM, Day NP. The treatment of severe malaria. Trans R Soc Trop Med Hyg. 2007;101:633-634.

[5] Ross R. The prevention of malaria. Murray, London; 1911;2.

[6] Macdonald G. The analysis of Equilibrium in malaria. Trop. Dis. Bull. 1952;49:813-829.

[7] Anderson RM, May RM. Infectious diseases of humans: Dynamics and control. London: Oxford University Press; 1991.

[8] Martens P, Kovats RS, Nijhof S, De Vries P, Livermore MTJ, Bradley DJ, Cox J, McMichael AJ. Climate change and future populations at risk of malaria. Global Environmental Change. 1999;9:S89-S107.

[9] Teklehaimanot HD, Lipsitch M, Teklehaimanot A, Schwartz J. Weather-based prediction of plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I. Patterns of lagged weather effects reflect biological mechanisms. Malaria Journal. 2004;3:41.

[10] Xu J, Zeger S. Joint analysis of longitudinal data comprising repeated measures and time to events. Journal of the Royal Statistical Society Series C. Applied Statistics. 2001;50;375-87.

[11] Zammit, Nicola N,George Streftaris, Gavin J. Gibson, Ian J. Deary, Brian M. Frier. Modelling the consistency of hypoglycaemic symptoms: High variability in diabetes. Diabetes Technology and Therapeutics. 2011;13(5):571-578.

[12] Syahida Zulkafli, George Streftaris, Gavin J. Gibson, Nicola N. Zammitt. Bayesian modelling of the consistency of symptoms reported during hypoglycemia for individual patients. Malaysian Journal of Mathematical Sciences. 2016;10(S):27-39.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle4.com/review-history/55623