Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 15, Number 2 (2019), pp. 191-205

http://www.ripublication.com/gjpam.htm

Bayesian Analysis via Markov Chain Monte Carlo Algorithm on Logistic Regression Model

Emenyonu Sandra Chiaka [1*], Mohd Bakri Adam[2]

[1]Department of Mathematics and Statistics, College of Natural and Applied Sciences, Gregory University Uturu Abia State, Nigeria.

[2]Institute for Mathematical Research and Department of Mathematics, Faculty of Science Universiti Putra Malaysia, 43400 upm Malaysia

Abstract

This research introduces the Bayesian approach which is applied to the logistic regression coefficients via Markov Chain Monte Carlo algorithm for posterior distribution to be obtained. Coefficients of the variables of interest are first generated via maximum likelihood estimation and the significant variables are further identified. These coefficients of the significant variables were then estimated using the Bayesian logistic regression method with the incorporation of both non-informative flat prior and a non-flat prior distribution. These results were compared with those generated via the method of maximum likelihood. It was shown that both the Frequentist logistic regression and Bayesian logistic regression suggest that family history, waist circumference and body mass index are significant risk factors associated with the Type 2 diabetes mellitus, Bayesian logistic regression model with the non-informative flat prior distribution and frequentist logistic regression model yielded similar results, while the non-informative non-flat model showed a different result compared to the frequentist logistic regression model and also a significant decrease of the standard errors associated with the coefficients generated from the Bayesian analysis with the non-flat prior distribution its being shown. Consequently, making the coefficients in the model more stable. Thus, the non-flat prior yielded better model than the maximum likelihood estimate and the Bayesian with the non-informative flat prior.

Keywords: Binary logistic regression, bayesian logistic regression, coefficient estimate, posterior distribution, prior, maximum likelihood estimate.

1. INTRODUCTION

Logistic regression as a model has the advantage of combing a few but efficient independent variable which serves as one of the outstanding properties (Kerlinger & Pedhezar, 1973).

The dependent variable which is the type 2 diabetes assumes a value of 1 for the probability of occurrence of the disease and 0 for the probability of non-occurrence.

The logistic model is seen to be an approved method for statistical analysis in most areas of study in the past ten years and more (Lemeshow and Hosmer, 2000). On the other hand, (Pen et al., 2002) suggest that the logistic regression model is suitable for explaining and also for hypotheses testing about a categorical dependent variable and one or more categorical or continuous independent variables. Similarly, the logistic regression which is often known as a logit model, models the relationship between several independent variable and categorical outcome, see (Park, 2013). In addition, the frequentist logistic regression (FLR) makes use of the maximum likelihood estimate (MLE) in order to maximize the probability of obtaining the observed results via the fitted regression parameters. Thus, the FLR brings about point parameter estimates together with standard errors. The uncertainty related to the estimation of parameters or coefficients is measured by means of confidence interval based on the normality assumption. On the contrary, Bayesian logistic regression (BLR) method makes use of Markov Chain Monte Carlo (MCMC) method in other to obtain the posterior distribution of estimation based on a prior distribution and the likelihood. Thus findings suggest that using the iterative Markov Chain Monte Carlo simulation, BLR provides a rich set of results on parameter estimation. Several studies conclude that BLR performs better in posterior parameter estimation in general and the uncertainty estimation in particular than the ordinary logistic regression. Further reading can be sort from (Lau, 2006) and (Nicodemus, 2001). (Gilks et al., 1996) proposed that in Bayesian, the unknown coefficients $\boldsymbol{\beta}$ are obtained from posterior distribution, inferences are made based on moment, quantile and the highest density region shown in posterior outcome of the parameter $\boldsymbol{\pi}$. Further, the bias of maximum likelihood estimates is significant in relation to small samples and this weakness can be confronted by making use of the Bayesian logistic regression as an alternative method. The Bayesian approach is flexible and does not need to conform with challenging assumptions as proposed in the method of maximum likelihood or as in the frequentist approach. However, the use of Markov Chain Monte Carlo (MCMC) improves the flexibility of the Bayesian method and the advancement of the MCMC methods has made it feasible to fit several non-linear regression models, see (Acquah, 2013). This research aims at applying the BLR model to T2DM to determine the associated risk factors. Uncertainty associated in estimation of the parameters is expressed by means of the posterior distribution. The estimates for the coefficients are obtained by means of FLR, then BLR is also applied on the same variables for coefficient estimation, and the significance of every coefficient estimate is assessed by means of the posterior density generated from the Bayesian analysis. In the present study, factors influencing the occurrence of the disease were determined by applying the Bayesian logistic regression and assuming a non-informative flat and not-

perfectly non- flat prior distributions for every unknown coefficient in the model. Although, several studies also used the BLR method with a non-informative flat prior distribution, there have not been many studies on the risk factors of type 2 diabetes mellitus using the Bayesian logistic regression method with a non-informative non-flat prior distribution. Therefore, we decided to incorporate this non-flat prior for the estimation of the parameters which to the best of our knowledge has not been used for the study of T2DM in Malaysia.

Several studies have employed the Binary logistic regression model to examine and analyse the impacts of the covariates on the binary outcome. For example, In a study, (aksu, 2006) reports the risk factors that are associated with Type 2 diabetes mellitus (T2DM) and determine the groups at risk for a public health program intervention in Nilufer district, Bursa, Turkey. The study comprised of 727 randomly selected patients of which 382 were women and 345 were men. Age, gender, education, family history of diabetes, hypertension, cigarette smoking, alcohol, occupational activity, physical exercise and body mass index were the predictor variables included in the model. A logistic regression analysis was performed which showed that out of all the predictors, only age, family history, hypertension and overweight (body mass index greater than 25kg/m^2) were statistically significant that is to say they were the risk factors which influenced T2DM in Nilufer.

However, a different study predicts cigarette smoking behaviour in high school student. (Adwere, 2011) evaluates the effect of a set of covariates in cigarette smoking behaviour of high school student by the use of logistic regression model for the analysis. In addition, the target outcome was current frequent cigarette use with five predictor variables such as race, frequency of cocaine use, initial cigarette smoking age, feeling sad or hopeless and physical inactive behaviour were considered. Consequently, all the predictors in the study were significant statistically and a conclusion was drawn, that all the predictors were associated with frequent cigarette use among high school students.

However, in recent years, (Majgi, 2012) in a study reveals the underlying risk factors causing diabetes mellitus in rural Puducherry India using cross-sectional data obtained from two villages. The body mass index, physical activities, family history of diabetes, smoking and rate of intake of alcohol were the variables considered.

Univariate analysis of the prevalence rate of selected risk factors was carried out and the significant variables were used in the analysis of the binary logistic regression. The significant variables for the univariate analysis are age, BMI, family history of diabetes, and the type of occupation.

These variables also make up the independent risk factors for diabetes in the binary logistic regression.

The analysis of the risk factors using logistic regression model, showed that higher Age, BMI, and occupation skill level were significantly contributing risk factors. The study also concluded that while maternal history of diabetes could be helpful in preventing diabetes, increase physical activity reduces the risk of diabetes mellitus.

Following the prevalence of osteoporosis and its associated factors among older men with Type 2 diabetes, (Chen, 2013) findings show that, no differences in terms of age, blood pressure, waist-to-hip ratio, body mass index, and testosterone levels observed. Consequently, the prevalence of low bone mineral density was significantly higher in Type 2 diabetes mellitus group compared to the control group and the risk of developing low bone mineral density (BMD) and fracture in T2DM subjects was increased by 46- and 26-fold respectively compared to control subjects.

In actual fact, logistic regression models have made a significant impact in different fields of study, as a result of the non-linearity of the logistic model. Conclusion is drawn by the maximum likelihood, although the method of maximum likelihood has its weaknesses which can be solved by the use of Bayesian which is a more flexible method. Following this, several fields of study have applied the Bayesian logistic regression. For example, (Mila, 2003) evaluate the extent of uncertainty associated with the estimation of coefficients generated from the logistic regression analysis of the prevalence of soybean sclerotinia stem rot. The study re-examines the Bayesian logistic regression of soybean sclerotinia stem rot (SSR) prevalence in the north-central region of the United States. Estimates from the posterior distribution for the coefficients were generated via the Gibbs sampler and both the informative and the non-informative prior distributions were considered. However, the informative and non-informative priors were examined and compared, and the predictor variables included were chosen on the basis of past logistic regression analysis. These predictor variables were average air temperature of July and August, total precipitation of July and August, an indicator variable for tillage effect and an indicator variable which represent state effect. In other words, based on the findings, which show that with the use of the non-informative prior distribution, the posterior estimates are similar with the coefficients estimate obtained from the logistic regression analysis, whereas the use of the informative prior has influence on the posterior distributions of the parameters, with result suggesting that the dataset may not have had enough information in order to yield estimates that can be trusted to influence some of the predictor variables on the prevalence of SSR. Hence, reliable estimate are essential for generating robust inferences and making sound predictions. On the other hand, (Sta Romana, 2007) investigate the relationship between type 2 diabetes mellitus (T2DM) and osteoporosis, a flat non- informative prior distribution was incorporated which expressed ignorance of the relationship between type 2 diabetes and osteoporosis and the coefficients of the variables of interest were generated. Findings show that, type 2 diabetes is found to be a protective factor for osteoporosis in this referred population of women. However, diabetes related factors like peripheral neuropathology can cause muscle imbalance, possibly hypoglycaemia that can bring about dizziness, nocturia, visual impairment affect fracture risk. Consequently, recommending that assessment, screening for osteoporosis and fracture risk reduction be performed among diabetic patients.

(Mutshinda 2009) in a study, performs a Bayesian analysis using a Bayesian logistic regression model, for posterior estimates to be generated. A non-informative flat prior was incorporated. This method was done on a real world data from a biological assay

experiment. Further, the generated estimates from the posterior distribution were compared with the estimates obtained via the method of maximum likelihood. Finding has it that five replicates by dose level considered resulted in low precision estimation, due to the estimates of the standard errors being large. On the other hand, in a study on Bayesian logistic regression modelling via Markov Chain Monte Carlo (MCMC) algorithm, (Acquah, 2013), applies the Bayesian logistics regression for the estimation of parameters on economy data. A comparison was made between the classical logistic regression and the Bayesian logistic regression which suggests that higher per-capital income is associated with free trade of countries. The results show that there was a reduction in the standard errors associated with the parameters generated from the Bayesian analysis. As a result, causing greater stability to the parameters.

Here, both the frequentist and Bayesian logistic regression logistic regression methods where applied to the type 2 diabetes dataset.

MATERIALS AND METHODS

Permission was sought from clinical research centre Kuala Lumpur. The procedure was spearheaded by a family medical specialist who was invited to take part in the study. The main research group organised site feasibility study to recognise clinics that were eligible. Eligibility was based on personal willingness, readiness and agreement to be fully involved and be part of the research group. The research was based on a cluster randomised trial such that the clinics that were selected were done randomly because they met the inclusion criterion. The unit of randomization for the study was the primary health care clinics with males and females ≥28 years of age that were diagnosed with T2DM. Individuals with type 1 diabetes and severe hypertension Systolic blood pressure >180mmHg and Diastolic blood pressure >110 mmHg were excluded. A self-management booklet was shared to all the participants after the training was over and the necessary details were extracted from them. The variables collected during the study were as follows: Demography, social and biological variables and behavioural components.

Logistic regression model will be considered for the occurrence of type 2 diabetes as a discrete and binary response variable, and factors such as, age, sex, ethnicity, physical activity, family history of diabetes, hypertension, body mass index and waist circumference as explanatory variables. A statistical analysis was carried out to determine the effect of these factors with respect to type 2 diabetes occurrence. Suppose the Binary logistic regression model is given as:

$$\text{Logit}(\pi_i) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

$$\pi = P(y_i = 1 | x_1, \ldots, x_k). \qquad (1)$$

Then the estimates of the model can be of the form:

Logit $(\hat{\pi}_i) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. (2)

Where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ are estimates of the coefficient β and $\boldsymbol{x_i} = (x_i, x_2, \ldots, x_k)$ are the k independent variables, $\hat{\pi}_i$ is the estimate of the likelihood of type 2 diabetes occurrence.

Given the explanatory variables $x_i, x_2, \ldots, x_k$, π_i can be estimated as:

$$\hat{\pi}_i = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \quad (3)$$

However, Bayesian framework is the combination of the likelihood function and the prior distribution to yield the posterior distribution. Consequently, the response variable y_i follows a Bernoulli distribution with probability π and is given as:

$y_i \sim$ Bernoulli (π_i),

$$\hat{\pi}_i = \frac{\exp(\boldsymbol{x_i}\grave{}\,\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i}\grave{}\,\boldsymbol{\beta})}.$$

Where, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$, $\boldsymbol{y_i} = (y_1, y_2, \ldots, y_n)$ and $\boldsymbol{x_i} = (x_1, x_2, \ldots, x_k)$.

The distribution of $(y_i|\boldsymbol{x_i}\grave{}\,\boldsymbol{\beta}) = \pi^{y_i}(\pi^{1-y_i})$

For i =…, n, y_i is the number of successes and 1- y_i is the number of failures.

2.1 The likelihood function

The likelihood function is the probability density function of the data which is seen as a function of the parameter treating the observed data as fixed quantities.

For a given sample size n, the likelihood function is given as:

$L(Y|X\beta) = \prod_{i=1}^{n} (y_i|\boldsymbol{x_i}\grave{}\,\boldsymbol{\beta})$.

Recall that

$(y_i|\boldsymbol{x_i}\grave{}\,\boldsymbol{\beta}) = \pi^{y_i}(\pi^{1-y_i})$.

Where

$$\pi_i = \frac{\exp(\boldsymbol{x_i}\grave{}\,\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i}\grave{}\,\boldsymbol{\beta})}.$$

And

$1\text{-}\pi_i = \frac{1}{1+\exp(\boldsymbol{x_i\beta})}$.

Therefore, the likelihood function is of the form:

$$= \left(\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\right)^{y_i} \left(\frac{1}{1+\exp(x_i\beta)}\right)^{1-y_i} \qquad (4)$$

Hence, the likelihood function can be of the form:

$$= \exp\left(\sum_{i=1}^{n} \; y_i \, x_i\beta\right) \prod_{i=1}^{n} \left(\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\right) \qquad (5)$$

2.2 Prior distribution

After the model for our data has been selected, the specification of our prior distribution for the unknown model parameters is made. We assign a prior distribution to all the unknown parameters. Firstly we assume a non-informative flat prior with mean zero and a large variance to all the parameters. However, we also assume a prior distribution to all the unknown parameters with mean zero and small variance 1, this influences the posterior distribution. In Bayesian analysis, precision is used rather than the variance, a large variance is chosen for it to be considered as non-informative while a small variance makes the prior not to be perfectly flat. Our choice of large variance is 10000 (10^4). We assign a normal distribution as prior to each unknown parameters, and the normal distribution is of the form:

$$\text{P}(\beta_j) = \prod_{j=0}^{k} \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left\{-\frac{1}{2}\left(\frac{\beta_j-\mu_j}{\sigma_j}\right)^2\right\} \qquad (6)$$

Each β is assigned with mean zero and precision 0.0001, and it is expressed as

$\beta_j \sim$ N (0, 0.0001), j=0,…,k.

Where β_j includes all the coefficients having normal prior distributions with very large variance.

However, to have a prior that is not perfectly flat, using the normal prior distribution we give each unknown parameter a mean of zero and a variance of 1 with a known precision given as:

$\beta_j \sim$ N (0, 1), j=0,…,k.

Where β_j include all the coefficients having normal prior distributions with very small variance.

2.3 Posterior distribution

The posterior distribution of the coefficients β is obtained by multiplying the likelihood function in Equation (5) by the prior distribution in Equation (6). The posterior is given as

$$P(\beta|yx) \propto \prod_{i=1}^{n} L(y|x_{i\beta}^{`}) \times \prod_{j=0}^{k} P(\beta_j)$$

The above expression can be written as

$$p(\beta|yx) \propto \left\{ \exp\left(\sum_{i=1}^{n} y_i x_{i\beta}^{`} \right) \prod_{i=1}^{n} \left( \frac{\exp(x_{i\beta}^{`})}{1+\exp(x_{i\beta}^{`})} \right) \times \prod_{j=0}^{k} \left[\frac{1}{\sqrt{2\pi\sigma_j}} \exp\left(-\frac{1}{2} \frac{(\beta_j-\mu_j)^2}{\sigma_j} \right) \right] \right\} \quad (7)$$

In the above expression, showing the posterior distribution, the latter part of the equation is seen to be the normal prior distribution for the unknown beta parameters. On the other hand, the posterior distribution here has no standard form, in other words, the metropolis algorithm is used to solving and approximating the features of the marginal posterior density of every coefficient. In actual fact, to estimating the posterior distribution of the coefficients of the Bayesian logistic regression, the random walk metropolis algorithm will be used.

Metropolis Hastings Algorithm:

The Metropolis Hasting (MH) algorithm is an iterative algorithm that generates a Markov chain and allows the estimation of the posterior distribution.

The metropolis-Hastings algorithm does not need availability of full conditionals. Rather, it generates a sequence of samples from a probability distribution by using the full joint density function and proposal distribution. The basic MH algorithm can be described by the following steps:

1. Set a starting value θ_0
2. The value θ_i in the first step becomes the starting value θ_0
3. Draw a candidate parameter value θ^* from an arbitrary proposal density, g(.) which is uniform. The value being simulated is taken to be candidate because it has not been automatically accepted as a sample from the distribution of interest and this is based on the acceptance ratio.
4. Calculate the ratio at the candidate parameter value θ^* and the current value θ_i

$$\alpha = \left\lfloor 1, \frac{f(\theta^*)g(\theta_i|\theta^*}{f(\theta_i)g(\theta^*|\theta_i} \right\rfloor$$

5. The next value for θ_i is given as

$\theta_{i+1} = \{ \quad \theta^* \text{ With probability } \alpha \text{ , } \theta_i \text{ with probability 1- } \alpha. \quad \}$

6. Generate U from uniform (0,1).

7. Accept θ^* if U < α and go back to step 2, otherwise accept θ_i and return to step 2.

3. RESULTS AND DISCUSSION

The data used in the analysis consist of eight variables of which result show that five significantly contributed to the occurrence of type 2 diabetes Mellitus. Factors such as family history of diabetes, body mass index, and waist circumference were significant, whereas physical activity, ethnicity, and gender, hypertension and age showed no significance.

The estimates for all the predictor variables generated via the method of maximum likelihood are displayed in Table. 1 which consists of the variable, estimate, standard error, p-value, odds and the significance level for all variables. The extent of contribution exhibited by the variables in the model is due to their interaction and significance level.

Table 1: Analysis of Maximum Likelihood Estimate for all the variables in the model.

Variable	Estimate	Standard error	P-value	Odds ratio
Intercept	-3.899	1.063	<0.001	0.020
Age	0.011	0.010	0.247	1.011
Gender	-0.109	0.201	0.589	0.897
Hypertension	0.103	0.186	0.051	0.697
Physical activity	-0.102	0.182	0.592	0.903
Family history of diabetes	-0.360	0.127	0.008	3.149
Ethnicity	0.058	0.012	<0.584	1.108
Waist circumference	1.147	0.195	<0.001	1.060
Body mass index	-0.077	0.027	0.004	0.926

Table.2 shows the Coefficient estimates of via the method of maximum likelihood and the posterior distribution summaries of coefficient via Random walk metropolis algorithm for Type 2 diabetes occurrence with reference to the significant factors with non-informative (flat) prior distribution.

Table 2: Coefficient estimates of via the method of maximum likelihood and the posterior distribution summaries of coefficient via Random walk metropolis algorithm for type 2 diabetes occurrence with reference to the significant factors with non-informative (flat) prior distribution.

Variable	Estimate from FLR	Posterior mean	Posterior Standard deviation	Quantiles of posterior distribution 95% Credible interval	
				2.5%	97.5%
Intercept	-3.233	-3.257	0.811	-4.882	-1.685
Family history of diabetes	1.100	1.111	0.185	0.753	1.478
Waist Circumference	0.054	0.054	0.013	0.030	0.079
Body mass index	-0.069	-0.070	0.025	-0.119	-0.201

Considering the Bayesian logistic regression via the Random walk metropolis algorithm which was applied to type 2 diabetes data in other to draw up inferences about the effects of several risk factors contributing to the disease. Using the non-informative prior (flat), the means of the posterior distribution of every coefficient are similar to the coefficient estimates generated via the method of maximum likelihood. This is as a result of the Bayesian analysis making use of non-informative prior which basically uses the available information by the sample data.

The random walk metropolis algorithm allows the use of several values of the variances in order to achieve the required acceptance rate.

We tried several tuning parameters afterwards 25.3 was finally adopted which is close to the target value of 0.25 or 25 percent. With (roberts1997), suggesting that in Random walk metropolis algorithm, the optimal acceptance rate is around 25 percent.

The Random walk metropolis algorithm using the non-informative prior (flat) distribution assumed a normal density with zero mean and precision 0.0001 and also with zero (0) mean and precision one (1).

The posterior summaries are shown in Table.2. Consists of three significant independent variables. A prior distribution was assigned to the variables and the posterior distributions were obtained. Summaries of the posterior estimates are shown in the above mentioned tables, with the inclusion of the credible interval and posterior standard deviation.

The first column shows the significant variables, the second contains estimate generated via the method of maximum likelihood, the third column contains the posterior means of the variables that is the posterior estimates of variables, and the fourth column contains the posterior standard deviation for each significant variable, which is a Bayesian equivalent of the standard error. The posterior standard deviation shows how good the mean estimates the samples. So the smaller the sample size, the smaller the posterior standard deviation and the closer the sample means to the population means. The smaller the posterior standard deviation and the closer the sample means to the population means. For the last column, this represents the 95% credible interval this is a Bayesian equivalent of the confidence interval. However, in Bayesian, the level or degree of belief is being reflected by the probability distribution. In other words, constructing the credible interval or credible region for our posterior estimates or posterior means, if the true value can be found in the credible interval then we can say that, given the observed data, there is a 95% probability that the true value falls within the credible interval. In other words, the credible interval values for the estimated coefficients of the predictor variable family history of diabetes, for the Random walk algorithm with a flat prior distribution, has a posterior mean of 1.111, there is 95% probability that the true value of the mean for FDM falls within the credible region of 0.753 and 1.478. Whereas, the waist circumference and the body mass index, having posterior means of 0.054 and -0.070 respectively we can say that there is 95% probability that the true value of the posterior means for waist circumference (WC) and body mass index (BMI) fall within the credible region of 0.030 and 0.079, -0.119 and -0.201.

Similarly, the Bayesian logistic regression via the Random walk metropolis algorithm incorporating a non-informative prior (not perfectly flat). The posterior distribution summaries of the parameters using the non-informative not perfectly flat prior are shown in Table 3. The use of this prior distribution influenced the posterior distribution of the intercept and the regression coefficients. The Table consists of the Variable, posterior mean, posterior standard deviation and the quantiles of the posterior distribution, which lies the credible interval of the variables. On the other hand, considering the posterior standard deviation and credible interval (which are the Bayesian equivalent of the confidence interval and standard error) for every coefficient assuming a non-informative not perfectly flat prior, shows that each standard deviation is smaller to that of the Bayesian analysis with the non-informative flat prior and the Frequentist analysis, implying that the smaller the standard deviation the better the model. In addition, based on the credible interval estimation for every coefficient in Table 3, the credible interval values for the estimated coefficients of the predictor variable family history of diabetes, for the Random walk algorithm with a non-flat prior distribution in has a posterior mean of 1.051, there is 95 probability that the true value of the mean for FDM falls within the credible region of 0.706 and 1.405.

Table 3: Posterior distribution summaries of the coefficients via Random walk metropolis algorithm for type 2 diabetes mellitus occurrence with reference to the significant factors with non-informative (not perfectly flat or non-flat) prior distribution.

Variable	Posterior mean	Posterior Standard deviation	Quantiles of posterior distribution 95% Credible interval	
			2.5%	97.5%
Intercept	-1.990	0.617	-3.212	-0.786
Family history	1.051	0.180	0.706	1.405
Waist Circumference	0.038	0.011	0.017	0.058
Body mass index	-0.058	0.024	-0.105	-0.113

Whereas, the waist circumference and the body mass index, having posterior means of 0.038 and -0.058 respectively we can say that there is 95% probability that the true value of the posterior means for waist circumference (WC) and body mass index (BMI) fall within the credible region of 0.017 and 0.058, -0.105 and -0.113. Since all the coefficients have shorter interval in the model with the non-flat prior compared to the MLE and the Bayesian logistic regression model with the non-informative flat prior, this is as a of the the interval bounds of the variable in that particular model having the shortest interval compared to the other models. That is to say that the model result that gives a shortest interval is considered to be reliable. In other words, the shorter the length of the interval for each coefficient, the better the model. In comparing the non-informative flat prior with the MLE method on T2DM data. The prior (flat) which is considered will overlap each other because a very large variance for the normal distribution brings about a very small precision which on the other hand yields results that are similar to those of the MLE. So making a comparison between the two methods, one can hardly say with confidence that the model is better than the other. However, with the use of another method, (that is a known variance that results in an informed or known precision) which will still be non- informative but not perfectly flat yielded a better results than the MLE and Bayesian with the flat prior.

Further, the inclusion of information about parameter values into the analysis through the choice of non- perfectly non-flat prior had an influence on the model. Owing to the fact that a known variance was used resulting to a known precision. On the other hand, when the standard deviation is small, the sample mean is close to each of the sample point, thereby making the result reliable.

Convergence can be checked visually. For instance, the unimodal shape of the kernel density can indicate convergence, this is to say, when there is no convergence, the density plots shows no unimodality. In addition, the 75000 represents the chain being run for every coefficient, that is 75000 iterations for each of the chains as displayed in

Fig. 1 Convergence was checked also by the use of the density plots and the trace plots as shown in the figures mentioned above. The plots in Fig 1 show the history of trace plots and the density distribution for Random walk algorithm of the corresponding posterior coefficient estimates for the variable of interest.

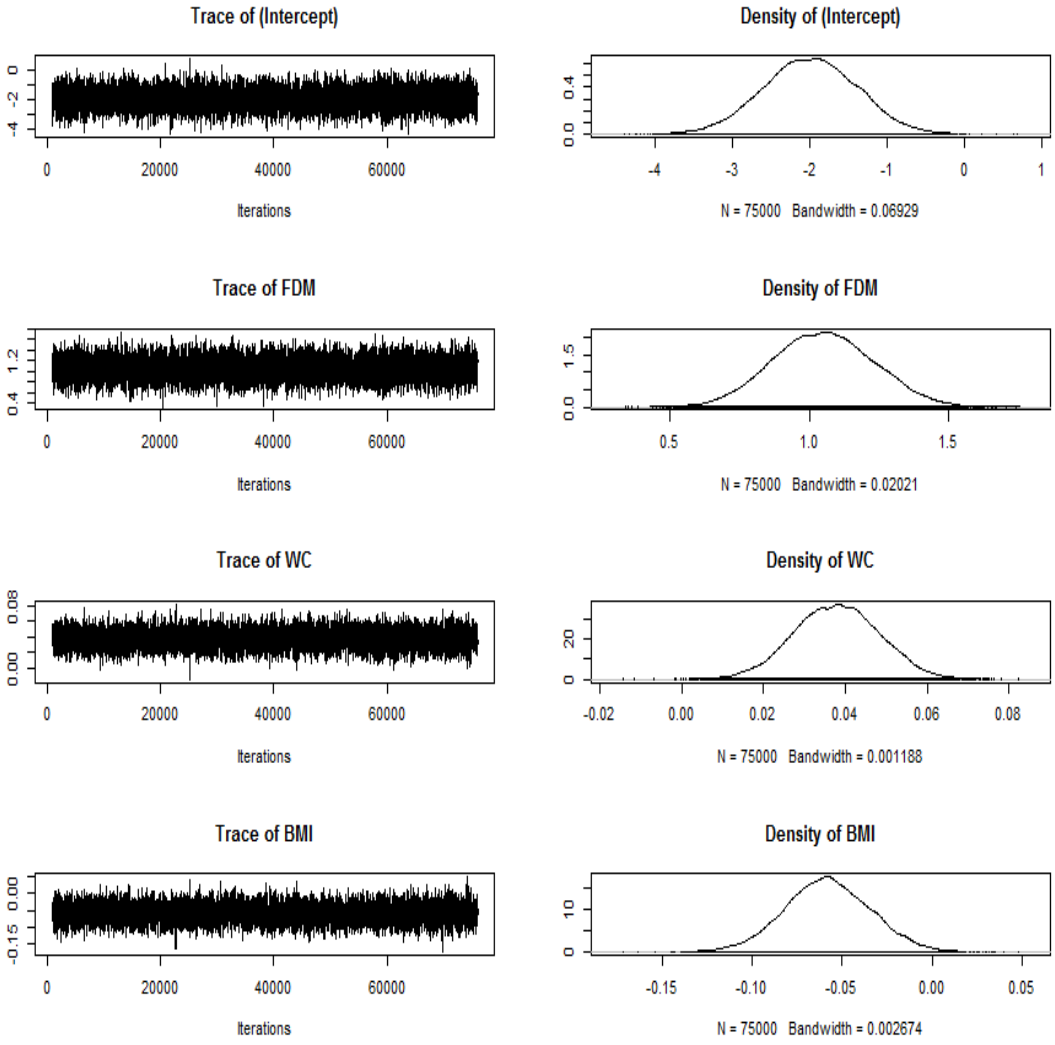

Figure1: Trace plots and density distribution of the corresponding posterior estimates of the Intercept, family history of diabetes (FDM) and waist circumference (WC) and body mass index (BMI) via Random walk metropolis algorithm.

CONCLUSION

In this study, the type 2 diabetes and its associated risk factors were addressed by the use of Bayesian logistic regression (BLR) model via the random walk metropolis algorithm. The Bayesian method incorporated a non-informative flat prior distribution and also another prior, which is still non informative but not perfectly flat. These

models allowed us to analyse the uncertainty associated with the parameter estimation. Comparison between the frequentist logistic regression and Bayesian logistic regression models revealed a similarity in the model results owing to the use of non-informative flat prior distribution. Therefore, our study shows that the use of non -informative but not perfectly flat yielded better results than the MLE and Bayesian with the flat prior.

Competing interests

The authors declare that no competing interests exist.

Acknowledgement

We wish to thank the staff from Kuala Lumpur clinical research centre for providing the data used in this research.

REFERENCES

[1] Acquah, H. D.-G, 2013. Bayesian logistic regression modelling via markov chain monte carlo algorithm. Journal of Social and Development Sciences., 4: 193-197.

[2] Adwere-Boamah, J, 2011. Multiple logistic regression analysis of cigarette use among high school students. Journal of Case Studies in Education., 1: 1-19.

[3] Aksu, H., Pala, K. and Aksu, H, 2006. Prevalence and associated risk factors of type 2 diabetes mellitus in nilufer district, bursa, turkey. International Journal of Diabetes and Metabolism 14: 98-102.

[4] Chen, H. –L., Deng, L. –L. and Li, J. –F, 2013. Prevalence of osteoporosis and its associated factors among older men with type 2 diabetes. International Journal of Endocrinology. pages 1-9.

[5] Gilks, W.R., S. Richardson, and D.J. Spiegelhalter, 1996. Introducing Markov Chain Monte Carlo. Pages 1-19. Springer.

[6] Kerlinger, F. and Pedhazur, E., 1973. Multiple regression in behavioural research. Holt, Rinehart and Winston, New York.

[7] Lau, E., Leung, P., Kwok, T., Woo, J., Lynn, H., Orwell, E., Cummings, S., and Cauley, J., 2006. The determinants of bone mineral density in Chinese men-results from mr.os (hong kong), the first cohort study on osteoporosis in Asian men. Osteoporosis International., 17(2): 297-303.

[8] Lemeshow, S. and Hosmer, D., 2000. Applied logistic regression. Wiley Series in Probability and Statististics. Wiley-Interscience; 2 Sub edition

[9] Mafauzy, M., (2006). Diabetes mellitus in Malaysia. Medical Journal of Malaysia., 61(4): 397-398.

[10] Mafauzy, M., Z. Hussein and S. Chan, 2011. The status of diabetes control in Malaysia: results of diabcare., (2008). Med .J. Malaysia., 66(3): 175-181.

[11] Majgi, S., (2012). Risk factors of diabetes mellitus in rural puducherry. Online Journal of Health and Applied Sciences, 11: 1-7.

[12] Mila, A.L., Yang, X. B. and Carriquiry, A. L., 2003. Bayesian logistic regression of soyabean sclerotinia stem rot prevalence in the US north-central region: accounting for uncertainty in parameter estimation., 93: 758-764.

[13] Mutshinda, C. M., 2009. Markov chain monte carlo-based bayesian analysis of binary response regression, with illustration in dose-response assessment. Modern Applied Science ., 3 p19.

[14] Nicodemus, K.K. and A.R. Folsom, 2001. Type 1 and type 2 diabetes and incident hip fractures in postmenopausal women. Diabetes Care., 24(7): 1192-1197.

[15] Park, H., (2013). An introduction logistic regression: from basic concepts to interpretation with particular attention to nursing domain. Journal of Korean Academy of Nursing., 43: 154-164.

[16] Peng, C . Y. J., Lee, K. L and Ingersoll, G. M., (2002). An introduction to logistic regression analysis and reporting. The Journal of Educational Research ., 96: 3-14.

[17] Sta Roman, M. and Li-Yu, J. T., (2007). Investigation of the relationship between type 2 diabetes and osteoporosis using bayesian inference. Journal of Clinical Densitometry., 10: 386-390.