Analysis of Wide Modified Rankin Score Dataset using Markov Chain Monte Carlo Simulation

Pranjal Kumar Pandey[1,*], Priya Dev[2], Akanksha Gupta[1], Abhishek Pathak[2], V.K. Shukla[3] and S.K. Upadhyay[1]

[1]*Department of Statistics, Banaras Hindu University, Varanasi - 221 005, India*

[2]*Department of Neurology, Institute of Medical Science, Banaras Hindu University, Varanasi - 221 005, India*

[3]*Department of General Surgery, Banaras Hindu University, Varanasi - 221 005, India*

Abstract: Brain hemorrhage and strokes are serious medical conditions that can have devastating effects on a person's overall well-being and are influenced by several factors. We often encounter such scenarios specially in medical field where a single variable is associated with several other features. Visualizing such datasets with a higher number of features poses a challenge due to their complexity. Additionally, the presence of a strong correlation structure among the features makes it hard to determine the impactful variables with the usual statistical procedure. The present paper deals with analysing real life wide Modified Rankin Score dataset within a Bayesian framework using a logistic regression model by employing Markov chain Monte Carlo simulation. Latterly, multiple covariates in the model are subject to testing against zero in order to simplify the model by utilizing a model comparison tool based on Bayes Information Criterion.

Keywords: Wide dataset, Logistic regression, Markov chain Monte Carlo, Covariates, Bayesian computation, Bayes information criterion.

1. INTRODUCTION

A brain hemorrhage, also known as intracranial hemorrhage, refers to bleeding within the brain tissue. This particular situation presents substantial health hazards and can result in grave outcomes, frequently resulting in permanent impairment or possibly even death. It may occur due to different factors, such as trauma, hypertension, aneurysms, or abnormalities in blood vessels. Brain hemorrhages can display various symptoms depending on the location and severity of the bleeding; however, recognizable signs are usually severe headaches, muscle weakness or paralysis, impaired speech or comprehension abilities, loss of awareness, and occurrences of seizures. In order to effectively manage and treat patients with this life-threatening condition, it is imperative for medical professionals to comprehend the underlying causes, symptoms, and treatments.

Individuals who have experienced a stroke often rely on the widely employed Modified Rankin Scale (mRS) to evaluate their degree of disability or dependence. Generally, there are a total of seven categories in the mRS, which range from 0 to 6 and represent distinct levels of disability. A lack of any symptoms whatsoever is represented by a score of 0, whereas a score of 6 signifies death [1]. This mRS may vary due to a large number of causes and studying the joint impact of these causes might be of interest to the medical practitioners.

Analysis of such problems which consist numerous covariates requires vast demand for data and advancement in data collection techniques might be a possible solution for this. Such datasets are characterized by a multitude of features or variables. Researchers frequently seek to establish a connection between these characteristics and particular patient outcomes. Large feature-rich datasets are hard to visualize effectively. When studying low-dimensional data, researchers can plot the response variable against each explanatory variable to ascertain which ones play a crucial role in predicting response, but the same becomes tough in the case of high-feature data. This is so because the single outcome is affected by numerous characteristics and the presence of strong correlation among them also increases its complexity. Regression models are found to be helpful in handling these situations. In problems involving regression with numerous predictors, the model is typically presumed to possess sparsity, thereby left with a few active predictors. When the usual statistical procedure struggles with high dimensional setting in procuring precise results, there has been a more recent proposal of employing Bayesian techniques [2, 3]. By fixing the parameter in their assumptions, frequentist methods tend to underestimate the variability of the parameter of interest. Contrarily, within Bayesian framework, all unspecified variables are subject to randomness and adhere to particular probability distributions. Needless to mention, the natural capability of these methods includes automatically quantifying uncertainty of the inference by means of the posterior distribution.

In statistics and data analysis, Bayesian methodology holds immense importance in addressing

*Address correspondence to this author at the Department of Statistics, Banaras Hindu University, Varanasi - 221 005, India;
E-mail: pranjal2802@bhu.ac.in

E-ISSN: 1929-6029/24

complex issues. It fundamentally applies Bayes theorem to update the understanding of a particular hypothesis when new information is accessible. The improved representation of our updated understanding is achieved by incorporating prior information and beliefs, thereby obtaining posterior distributions that better reflect the problem at hand [4]. Despite this fact, the posterior distributions typically do not have closed-form solutions and integrating the marginal term with respect to the large number of parameters is usually tough, which necessitates utilizing computational algorithms like Markov chain Monte Carlo (MCMC) for approximating the desired inferences [5]. This refined approach allows one to analyze an extensive array of complicated models while systematically accounting for uncertainty and variability in our inferences. We record states from the Markov chain that allows us to acquire a sample from the desired posterior distribution and, thereby, draw the sample-based posterior inferences.

In this article we have considered data on mRS of the individuals and the associated factors. For the purpose of analysis we have divided mRS of the individuals varying from 0-3 as good outcome and 4-6 as bad outcome thus converting the variable of interest into a dichotomous one. A logistic regression model, which is considered to be a good choice for data where dependent variable is dichotomous, is then used for analysis. The analysis of the model is performed completely in a Bayesian framework for the reasons mentioned above and finally MCMC techniques are employed to obtain the results.

The structure of the paper is mapped out in the following manner. The upcoming section supplies the necessary modelling formulation for implementing the proposed Bayesian approach, assuming a logistic regression model. The section also discusses briefly the Metropolis algorithm, which is an important and flexible MCMC technique. In Section 3, there is an exploration of variable selection with an overview of Bayes information criterion (BIC), utilized for determining significant variables within the model. In Section 4, a numerical illustration is given for a real life wide mRS dataset. Finally, the last section provides a succinct conclusion following the extensive list of the references.

2. BAYESIAN MODEL FORMULATION

To start with, let us assume a variable $Z = (z_1, z_2, \ldots, z_n)$ having n independent observations coming from a Bernoulli family taking binary values where 1 denotes the occurrence of the event and 0 signifying otherwise. Thus,

$$P(Z = z_i|\theta_i) \propto \theta_i^{z_i}(1-\theta_i)^{1-z_i},\ z_i = 0,1;\ 0 < \theta_i < 1;\ i = 1,2,\ldots,n \quad (1)$$

where $\theta_i = Pr(z_i = 1);\ i = 1,2,\ldots,n$, is the mean of Bernoulli distribution. Now, the variable Z is associated with several other features that are likely to affect it in one or other way. In scenarios like this, the logistic regression model seems to be an obvious choice where the parameters have a unique interpretation in terms of logarithm of odds ratios (see [6, 7]) a quantity which is of major interest to medical practitioners. By examining the relationship within a provided dataset, logistic regression enables the classification of data into distinct categories. Not only does it gauge the effectiveness of a predictor (based on the value of coefficient), but it also signifies its direction of association. When working with the logistic regression model, there exists a logit link function that establishes a connection between θ_i and explanatory features given by

$$log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 f_{i1} + \beta_2 f_{i2} + \ldots + \beta_k f_{ik}.$$

where $\beta_k{}'$s are the coefficients associated with the given features. On solving the above equation for θ_i and putting it in (1), we can write the likelihood function as

$$L = \prod_{i=1}^{n}\left(\frac{1}{1+exp\left(-\beta_0-\sum_{j=1}^{k}\beta_j f_{ij}\right)}\right)^{z_i}\left(\frac{exp\left(-\beta_0-\sum_{j=1}^{k}\beta_j f_{ij}\right)}{1+exp\left(-\beta_0-\sum_{j=1}^{k}\beta_j f_{ij}\right)}\right)^{1-z_i} \quad (2)$$

The next step involves specifying priors for the parameters of the regression model. In the case of non-availability of enough a priori information, one can consider non-informative or weakly informative priors (see, for example, [8, 9]). The main advantage of non-informative prior lies with the fact that the inferences are data dependent but, simultaneously, one shows confidence in Bayesian logic. This paper considers normal distributions with large variances as the prior distributions for the regression parameters. That is

$$\beta_j \sim N\left(\mu_j, \sigma_j^2\right), \quad (3)$$

where μ_j and σ_j are the hyperparameters associated with the prior of $\beta_j, j = 0,1,\ldots,k$. Now combining (2) and (3) using Bayes' theorem, one can get the joint posterior distribution up to proportionality given as

$$P\left(\bar{\beta}\middle|\bar{z},\bar{f},\bar{\mu},\bar{\sigma}\right) \propto$$

$$\prod_{i=1}^{n}\left[\left(\frac{1}{1+exp\left(-\beta_0-\sum_{j=1}^{k}\beta_j f_{ij}\right)}\right)^{z_i}\left(\frac{exp\left(-\beta_0-\sum_{j=1}^{k}\beta_j f_{ij}\right)}{1+exp\left(-\beta_0-\sum_{j=1}^{k}\beta_j f_{ij}\right)}\right)^{1-z_i}\right]$$

$$\prod_j\left[exp\left\{-\frac{(\beta_j-\mu_j)^2}{2\sigma_j^2}\right\}\right] \quad (4)$$

In situations where numerous parameters are involved, the posterior becomes challenging to handle analytically and thus making use of MCMC technique becomes an important alternative. In this paper, the Metropolis algorithm is employed for obtaining samples from the posterior distribution specified up to proportionality in (4). A brief description of the same is provided below. With the given posterior, say, $p(\theta|\underline{x})$, one can build a Markov chain possessing an equilibrium distribution $p(\theta|\underline{x})$. Given a symmetric Markov kernel $q(\theta,\theta') = q(\theta',\theta)$, suppose θ is the current realized value of the chain in the state θ_ℓ, then one can propose θ' generated from $q(\theta,\theta')$ as the next realized value. The proposed value is, however, accepted with the probability

$$\alpha(\theta,\theta') = min\left\{\frac{p(\theta'|\underline{x})}{p(\theta|\underline{x})}, 1\right\}.$$

Otherwise, θ itself is retained as the next realized value. For the purpose of implementation, we have considered a multivariate normal kernel with an appropriately chosen mean and standard deviation c_s times Σ, where c_s is a scaling constant often taken between 0.5 and 1.0 in order to keep the acceptance probability reasonably high. One may refer, for example, Upadhyay *et al*., 2001 [10] for further details on the algorithm. The algorithm for the Metropolis chain is given below.

1. Choose a possible realization $\boldsymbol{\theta'}$ randomly from the symmetric proposal density $\boldsymbol{q}(\boldsymbol{\theta'}|\boldsymbol{\theta})$.

2. Determine the acceptance probability for $\boldsymbol{\theta'}$ by

$$\alpha(\theta,\theta') = min\left[\frac{p(\theta'|\underline{x})}{p(\theta|\underline{x})}, 1\right]$$

3. $\boldsymbol{\theta'}$ will be accepted with probability $\boldsymbol{\alpha}(\boldsymbol{\theta},\boldsymbol{\theta'})$ if

   - $p(\theta'|\underline{x}) \geq p(\theta|\underline{x})$
   - In case, $p(\theta'|\underline{x}) < p(\theta|\underline{x})$, we randomly generate a number from Uniform $[0,1]$ and accept θ' if the randomly generated quantity is less than $\frac{p(\theta'|\underline{x})}{p(\theta|\underline{x})}$.

4. If the proposed value is accepted, then $\boldsymbol{\theta_{\ell+1}} = \boldsymbol{\theta'}$ otherwise $\boldsymbol{\theta_{\ell+1}} = \boldsymbol{\theta}$ itself where $\boldsymbol{\theta_{\ell+1}}$ is the next realized state after $\boldsymbol{\theta_\ell}$.

3. VARIABLE SELECTION AND MODEL COMPARISON

The occurrence of a vast number of covariates influencing the response variable often arises in regression analysis, resulting in a bigger model dimension that becomes too cumbersome to manage. Occasionally, circumstances arise wherein certain covariates have a diminishing influence on the response variable, making it possible to eliminate them by testing their corresponding regression coefficients against zero. Dropping such covariates effectively decreases the dimensionality of the problem, leading to a simplified model. The procedure is known as variable selection. The variable selection, of course, simplifies a model but a further confirmatory assessment can be done by comparing the simplified model with the full model, possibly using a tool for model comparison. Obviously, the model comparison considers comparing a model with k covariates with a model having less than k covariates. If the result of model comparison supports a model having less than k covariates, one can comfortably drop the covariates from the full model and retain a model with less than k covariates for further inferences.

There are number of tools available for model comparison such as Akaike information criterion (AIC), Bayesian information criterion (BIC) and Deviance information criterion (DIC) etc. These tools mostly consider a trade-off between model complexity and its fit to the data and, accordingly, recommend a model. AIC is a frequentist criterion whereas both BIC and DIC use Bayesian approach for the same. As we are working in the Bayesian paradigm, this paper considers BIC initially proposed by Schwarz 1978 [11] for the purpose of model comparison. One can, of course, consider DIC as well but it is avoided here to make the approach simpler. The BIC can be defined as

$$BIC = -2 * \log L' + \kappa \log n$$

where κ denotes the number of parameters in the entertained model and L' is the maximized likelihood function evaluated at maximum likelihood (ML) estimates of the parameters, although the posterior modes are also recommended in the literature, especially when the priors are weak (see [12]). The two terms in the definition of BIC have their own significance in the sense that the first term corresponds to model fit whereas the second compensates for the model complexity, resulting in encouraging parsimony principle. The BIC criterion allows a model to be considered as the most appropriate that provides the least BIC.

4. NUMERICAL ILLUSTRATION

The present section provides a numerical illustration of the model proposed in Section 2 based on a real dataset pertaining to the mRS score collected from individual patients at Sir Sunderlal Hospital, Banaras Hindu University. The dataset consists of 285 observations of a binary variable mRS score where a

Table 1: Names of the Associated Explanatory Variables

Name of the explanatory variable	Interpretation
Age	Containing information about Age of the individuals
Heartattack-DAA	Individuals diagnosed with Heart-attack at the time of admission
Diabetes-DAA	Subjects that are diagnosed with Diabetes at the time of admission
Cholesterol	Cholesterol level
Atrialfibrillation-DAA	Individuals diagnosed with Atrial fibrillation at the time of admission
Family stroke history	Information about incidence of stroke to individual's family members
Location	Stroke occurred in left or right part of the brain
Midline shift	Information about the midline shift occurred in brain
Alcohol	Alcohol consumption
TLC	Total Leukocyte count
Creatinine	Creatinine level of individuals
GCS	Glasgow Coma Scale

value 0 indicates a good mRS score and a value 1 indicates a poor mRS score. It may be noted that a poor mRS score is indicative of going against the health of patients. Besides, the data set consists of 12 other explanatory variables that may affect the main variable mRS score in some or other way. Table **1** provides the names of these explanatory variables for a ready reference. It may be noted that several other explanatory variables could have also been observed, but the paper considers only those which are of prime importance according to the experts.

Dev *et al*., 2022a and Dev *et al*., 2022b [13, 14] also considered this dataset and analyzed the same using frequentist approach. They, however, considered observing the impact of a single variable named antibiotic on happening of the event of types of stroke and relied on some simple tools without focussing much on the regression setup. The objective here is different as well as more elaborative in the sense that the paper provides a complete Bayes analysis considering a number of explanatory variables in a logistic regression framework with the ultimate aim of seeing if the explanatory variables have a major role in affecting the mRS score.

Next, the Bayesian model formulation given in Section 2 can be applied on the present dataset presuming that the subscript i varies from 1 to 285 and the subscript j varies from 1 to 12 in the posterior (4) corresponding to considered logistic regression model. As discussed, the normal prior in each case was considered with mean 0 and a presumably large variance 30. The Metropolis algorithm as discussed in Section 2 was then applied using multivariate normal kernel with ML estimates of the parameters and the corresponding Hessian-based approximation as the initial values. The scaling constant c_s was assessed to be 0.6 to get a reasonably good acceptance probability. A single long run of the chain was progressed and the convergence based on ergodic averages was monitored through R graphics. It was found at about $40K$ iterations before the final sample of size 1000 was chosen at a constant gap of 10 to make serial correlation among the generating outcomes reasonably small. Some of the posterior based inferences based on these finally generated samples are shown in Table **2**. These inferences are shown in the form of estimated posterior mean, standard deviation and the highest posterior density interval with coverage probability 0.95 (0.95 HI) in each case.

It can be seen from Table **2** that the positive values of estimated regression coefficients indicate the increase in mRS score, a finding that might not be appropriate for the overall health of patients. As such, the diagnosed heart-attack, diabetes and atrial fibrillation at the time of admission, individual's family having history of stroke, location of stroke, midline shift in brain, regular consumption of alcohol and creatinine levels are significantly affecting mRS score.

Additionally, it can also be observed that some of the covariates such as age, cholesterol, TLC and GCS have very narrow 0.95 HI covering a zero and posing negligible impact (values of the associated regression coefficients are very close to zero) on the outcome variable. If one considers the associated coefficients as zero, the model will be considerably simplified and the resulting inferences are likely to be easier. To examine the issue of whether the reduced model is really advantageous, it is proposed to compare the full model with the reduced model using BIC. The results are shown in Table **3**. It can be seen that the BIC value for the reduced model is smaller than the corresponding

Table 2: Posterior Estimates of Significant Parameters

Coefficients	Posterior mean	Posterior SD	0.95 HI
$\beta_{Intercept}$	-16.22	6.69	(-30.84, -3.54)
β_{Age}	0.05	0.04	(-0.01,0.07)
$\beta_{Heartattack-DAA}$	1.01	1.54	(-2.0,4.06)
$\beta_{Diabetes-DAA}$	1.92	0.91	(-0.21,3.75)
$\beta_{Cholesterol}$	-0.01	0.02	(-0.025,0.03)
$\beta_{Atrialfibrillation-DAA}$	0.99	1.63	(-1.96,4.36)
$\beta_{Familystroke}$	1.31	1.28	(-1.18,3.84)
$\beta_{Location}$	1.10	0.84	(-0.50,3.02)
$\beta_{Midlineshift}$	1.25	0.94	(-0.58,3.09)
$\beta_{Alcohol}$	2.77	1.78	(-0.48,4.85)
β_{TLC}	-0.02	0.07	(-0.09,0.10)
$\beta_{Creatinine\ \ level}$	3.02	3.00	(-0.06,10.93)
β_{GCS}	-0.03	0.05	(-0.15,0.08)

value for the full model and, therefore, one can propose the reduced model for developing the desired inferences. The fact is further strengthened by the fact that the explanatory variables age, cholesterol, TLC and GCS have negligible impact on the outcome variable.

Table 3: BIC Values for the Full and Simplified Model

Model	BIC value
Full Model	335.82
Reduced Model (excluding age, cholesterol, TLC and GCS)	302.62

5. CONCLUSION

In medical studies, a number of explanatory variables are often observed. A few among these have no or almost negligible impact on the outcome variable. The present paper analyses one such dataset on mRS scores of neurological patients using a logistic regression model. The complete Bayes analysis is provided using the Metropolis algorithm and the analysis shows that the entire development is not only routine but also capable of providing almost every inferential aspect that one desires. The results exhibit that some of the variables have almost negligible effect on the mRS score and, accordingly, a reduced model can be proposed, leaving these explanatory variables. Finally, the results based on BIC values suggest that the reduced model can undoubtedly be used for the desired inferential developments.

REFERENCES

[1] Broderick JP, Adeoye O, Elm J. Evolution of the modified rankin scale and its use in future stroke trials. Stroke 2017; 8(7): 2007-2012. https://doi.org/10.1161/STROKEAHA.117.017866

[2] Mallick H, Yi N. Bayesian methods for high dimensional linear models. Journal of Biometrics & Biostatistics 2013; 1: 001-005. https://doi.org/10.1098/rsta.2009.0159

[3] Johnstone IM, Titterington DM. Statistical challenges of high-dimensional data. Philosophical Transactions of the Royal Society A: Mathematical. Physical and Engineering Sciences 2009; 367(1906): 4237-4253.

[4] Joyce J. Bayes' Theorem. In Zalta EN, Editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Fall edition 2021.

[5] Chen M-H, Shao Q-M, Ibrahim JG. Monte Carlo Methods in Bayesian Computation. Springer Science & Business Media 2012.

[6] Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika 1979; 66(3): 403-411. https://doi.org/10.1093/biomet/66.3.403

[7] LaValley MP. Logistic regression. Circulation 2008; 117(18): 2395-2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658

[8] Syversveen AR. Noninformative Bayesian priors. interpretation and problems with construction and applications. Preprint Statistics 1998; 3(3): 1-11.

[9] Gupta A, Upadhyay S. On the use of a logistic regression model in the gene-environment problem: A Bayesian approach. American Journal of Mathematical and Management Sciences 2019; 38(4): 363-372. https://doi.org/10.1080/01966324.2019.1570406

[10] Upadhyay S, Vasishta N, Smith A. Bayes inference in life testing and reliability via Markov chain Monte Carlo simulation. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 2001; 63(1): 15-40.

[11] Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics 1978; 6(2): 461-464. https://doi.org/10.1214/aos/1176344136

[12] Sharma R, Srivastava R, Upadhyay SK. A hierarchical Bayes analysis and comparison of ph Weibull and ph Exponential

models for one-shot device testing experiment. International Journal of Reliability, Quality and Safety Engineering 2021; 28(05): 2150036. https://doi.org/10.1142/S0218539321500364

[13] Dev P, Singh VK, Kumar A, Chaurasia RN, Kumar A, Mishra VN, Joshi D, Pathak A, *et al*. Raised blood urea nitrogen-creatinine ratio as a predictor of mortality at 30 days in spontaneous intracerebral hemorrhage: An experience from a tertiary care center. Neurology India 2022a; 70(4): 1562-1567. https://doi.org/10.4103/0028-3886.355134

[14] Dev P, Singh VK, Kumar A, Chaurasia RN, Singh NA, Gautam P, Dhimani NR, Mishra VN, Joshi D, Pathak A. Use of ceftriaxone as a predictor of good outcome in stroke patients: A retrospective chart review. Annals of Neurosciences 2022b; 29(2-3): 116-120. https://doi.org/10.1177/09727531221086736

Received on 26-11-2023 Accepted on 19-12-2023 Published on 18-01-2024

https://doi.org/10.6000/1929-6029.2024.13.02