Diabetes Prediction Using Bayesian Logistic Regression and Conventional Logistic Regression

Final Project for Bayesian Method for Data Science (DATS 6450)
Project Group 11: Li Du, Xiaochi Li, Yupeng Yang
Data Science, George Washington University

Abstract

Globally, diabetes has become one of the major risk factors to health. It involves problems with the hormone insulin, but its inception can be reflected by a series of other symptoms and indicators. This study aims at applying Bayesian method to construct logistic regression model to classify subjects with and without diagnosed diabetes from a group of predicting variables. Comparison between Bayesian and conventional logistic regression suggests marginal difference in their performance. Sensitivity study indicates that for the Pima dataset the size of the dataset and variables in the models are likely to play a more important role in the determination of posterior distributions of coefficients. This study also evaluated the influence of prior distribution of the coefficients to the posteriors.

INTRODUCTION

Diabetes is a number of diseases that involve problems with the hormone insulin[1]. Normally, the pancreas (an organ behind the stomach) releases insulin to help your body store and use the sugar and fat from the food you eat. Diabetes can occur when the pancreas produces very little or no insulin, or when the body does not respond appropriately to insulin. In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2015, diabetes was the direct cause of 1.6 million deaths and in 2012 high blood glucose was the cause of another 2.2 million deaths[2]. As yet, there is no cure. People with diabetes need to manage their disease to stay healthy.

Early detection and treatments of potential diabetes patients is necessary to reduce the healthy risk of having diabetes. To better understand the weights of certain variables and diagnostically predict whether a patient has diabetes based on these diagnostic measurements, our team found and analyzed the Pima Indians Diabetes Database[3] by using both of Bayesian and Conventional logistic regression models.

The database originally provided by National Institute of Diabetes and Digestive and Kidney Disease. It consists of several medical predictor variables and one target variable. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. The target variable, "Outcome", is consisted by 0 (no explicit diabetes) and 1 (explicit diabetes).

[1] Diabetes Health Center, https://www.webmd.com/diabetes/default.htm
[2] Diabetes World Health Organization, http://www.who.int/mediacentre/factsheets/fs312/en/
[3] Diabetes Database, https://www.kaggle.com/uciml/pima-indians-diabetes-database/data

METHODS & IDEAS

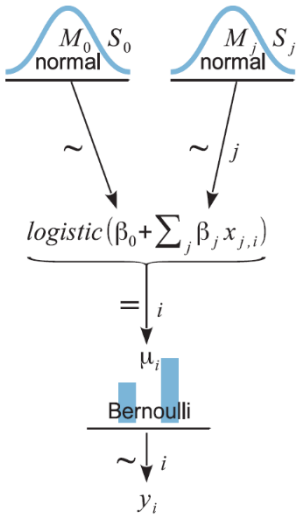
Since the dependent variable in our data is dichotomous, as the book recommended, logistic model should be the right way to analyze the data.[4]
Like the General Linear Model in frequentist school, Bayesian school also has their General Linear Model. The GLM can be written as follows:

$$\mu = f(lin(x), [parameters]) \text{ (1)}$$

$$y \sim pdf(\mu, [parameters]) \text{ (2)}$$

The predictors *x* are combined in the linear function $lin(x)$, and the function f in equation 1 is called the inverse link function, which is like the traditional GLM. However, the difference is that the data, y, are distributed around the central tendency μ according to the probability density function labeled "pdf". [5]

More specifically, in Bayesian logistic model, the linear combination of metric predictors is mapped to a probability value via the logistic function, and the predicted 0's and 1's are Bernoulli distributed around the probability. Restated formally and shown in the figure:[6]



$$\mu = logistic(\beta_0 + \sum \beta_j x_j) \text{ (3)}$$

$$y \sim Bernoulli(\mu) \text{ (4)}$$

$$logistic(x) = \frac{1}{1+e^{-x}} \text{ (5)}$$

We also standardized the x to reduce autocorrelation and improve the efficiency of the MCMC process.

$$logit(\mu) = \zeta_0 + \sum \zeta_j z_j = \zeta_0 + \sum \zeta_j \frac{x_j - \overline{x_j}}{s_{x_j}} = \zeta_0 - \sum \frac{\zeta_j}{s_{x_j}} \overline{x_j} + \sum \frac{\zeta_j}{s_{x_j}} x_j \text{ (6)}$$

We use the code from *Jags-Ydich-XmetMulti-Mlogistic.R* to implement the Bayesian logistic model in JAGS. This is the line of code to specify the connection between y and x:

*y[i] ~ dbern (ilogit (zbeta0 + sum (zbeta [1: Nx] * zx [i,1: Nx])))*

After we finish coding the model, JAGS will generate MCMC chain automatically and we can get the distribution of β from the result.

To compare the performance between Bayesian logistic model and traditional logistic model, we split the data into train set and test set, and use train set to train a traditional logistic model and a

---

[4] Table 15.3 from Kruschke, J.. Doing Bayesian data analysis, 2nd Edition: A Tutorial with R, JAGS and Stan. 2015 Academic Press/Elsevier. ISBN:9780124058880

[5] See section 15.4 "Formal Expression of the GLM" for more detailed discussion.

[6] See section 21.1 "Multiple Metric Predictors" for more detailed discussion.

Bayesian logistic model. And evaluate the performance by AUC and hit rate (accurate prediction rate).

To easily compare the performance of the models, we simply use the mode of the distribution of βs from the result of MCMC to build a logistic function.

$$\mathrm{y} = \mathrm{logistic}\left(mode(\beta_0) + \sum mode(\beta_j)x_j\right) \quad (7)$$

These parts are implemented in the *train_test_split, traditional_logistic, bayesian_logistic* function in the code *MCMC_logistic.R* and the high-level code is *Driver.R*

We also tried to find out the impact of different preprocess methods, like making the data more balanced or standardizing the data, on the performance of the models. This part is implemented as *preprocessor function* in *Preprocessor.R*.

Experimental Results and Analysis

Dataset and data preprocessing

The "Pima" dataset [7] is one of the most popular machine learning practice datasets. It includes nine variables with eight as numerical predictors and one target variable that is dichotomous (1 and 0). This dataset has 768 records each representing a subject that participated in a study and was examined for a series of demographical and biological parameters. However, due to the limitation of the study, data was not available for all the subjects and these missing values are simply marked as zeros in the raw data file.

Imputation of these values using aggregated metrics such mean or median of the same variable across the entire dataset has been a welcomed approached in many applications[8]. However, in order to rule out the uncertainties that are likely to be introduced by these "artificial" values, the primary part of the analyses in this project was based on a preprocessed dataset that only contains records with full data coverage (i.e. no missing values for each predictors). However, a sensitivity analysis was performed to evaluate the impact of these missing values and is presented in the Supplementary Information.

With appropriate preprocessing, the updated dataset for this analysis contains 392 observations and 9 variables. We also performed standardization to all predicting variables for this project. Other preprocessing practices were evaluated on a case-by-case basis and will be discussed in the sections below.

[7] Obtained from https://www.kaggle.com/uciml/pima-indians-diabetes-database/data

[8] Case Study on Pima Indian Diabetes. Available at https://www.kaggle.com/hinchou/case-study-on-pima-indian-diabetes

Logistic regression: conventional vs. Bayesian

As discussed in the previous section, JAGS was applied to generate MCMC chains that allow the estimation of regression parameters by their probability/frequency distribution. Because no prior information was assumed available for any of the predicting variables, priors for the coefficients and the intercept was assumed to follow a normal distribution centered at 0 with a standard deviation of 0.25 (dnorm(0, 0.25). However, a sensitivity study was performed to investigate the impact of different priors in the following sections.

Figure 1 shows the MCMC diagnostics of the standardized intercept (β_0) as an example. Nicely overlapped traces of chains and shrinkage factors that are close to 1 are good assurance of the representativeness of the samples from the posterior distribution. Given that Kruschke[9] suggested an effective sample size values of 10000 that is largely empirical, the ESS values in this analysis, ranging from 6531 to 12139, are demonstration of good accuracy for the posterior distribution.

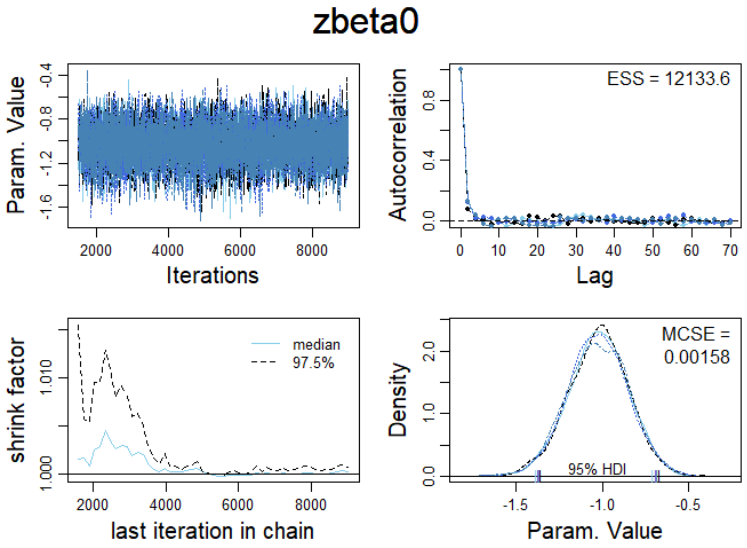


Figure 1. Illustration of MCMC diagnostics for the standardized intercept(β_0).

Bayesian logistic regression yielded very similar results compared to the conventional logistic regression. Visually, the 95% highest density intervals (HDI) of the posterior distributions (shown in Figure 2) encompass the values of corresponding coefficients from conventional logistic regression and the mode of these posterior distributions of the coefficients are in fact only marginally different from the coefficients generated by the logistic regression model.

To take one step further, we evaluated the two models using the test dataset that was generated by retrieving 30% of entire raw data while maintaining the relative number of records that belong to both classes (i.e. 1 and 0)[10]. Mode values from the 9 posterior distributions were used for the evaluation of the Bayesian logistic regression model.

---

[9] Kruschke, J.. Doing Bayesian data analysis, 2nd Edition: A Tutorial with R, JAGS and Stan. 2015 Academic Press/Elsevier. ISBN:9780124058880

[10] See section "Effect of unbalanced 1's and 0's" for more detailed discussion.

The performance demonstrated by confusion matrices of the two models are given in Table 1. The similar hit rates of the two models are consistent with the results demonstrated by Figure 2. Because MCMC chain are generated to represent the posterior distributions of the coefficients when the predicted variable follows certain probability density function (pdf), a Bernoulli distribution in the case of dichotomous variables, the mode of the posterior distributions may correspond to the maximum probability density of the predicted variables.

Table 1. Performance comparison of conventional logistic regression and Bayesian logistic regression

Logistic regression (hit rate=82%)

	[Predict] 0	[Predict] 1
[Actual] 0	74	17
[Actual] 1	4	22

Bayesian logistic regression (hit rate=83%)

	[Predict] 0	[Predict] 1
[Actual] 0	73	15
[Actual] 1	5	24

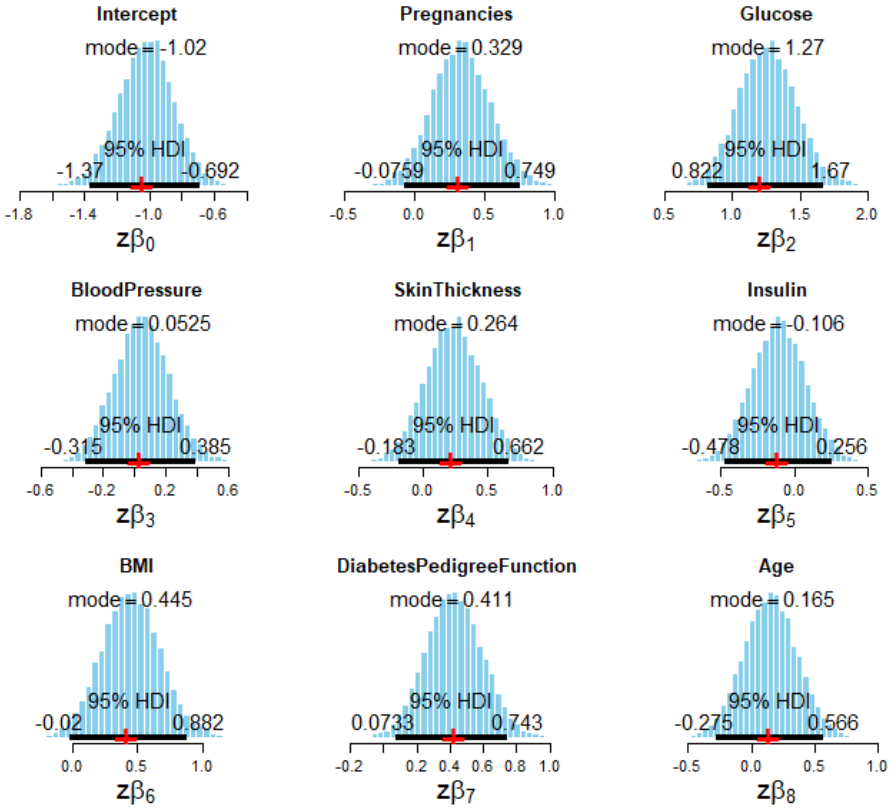


Figure 2. Posterior distributions of the regression coefficients and the intercept. Red crosses on the X-axis indicate corresponding values generated by the conventional logistic regression.

In the meantime, conventional logistic regression tries to maximize the likelihood for its predictions based on the input. Therefore, when no strong priors are defined for Bayesian logistic regression and training data that are sufficiently large in size, these two models may show good

agreement. While the influence of outliers to the relative performances were not investigated in this project, it is also likely that the absence of major outliers in the training and testing datasets contribute to the similar performance.

Another interesting observation is that a few coefficients are associated with large p-values suggesting weak predictive power for these variables (such as blood pressure, skin thickness etc.), whereas posterior distributions for all 9 variables display similar spans of HDIs. It is also possible that the logistic regression is sensitive to the input dataset. Although we realize that regularization may be reasonable solutions to identifying such variables, this may be beyond the scope of this project and would be considered as future work.

Effect of unbalanced 1's and 0's

The original dataset has unequal number of records in the 1 and 0 class. Parameter estimates become more uncertain (i.e. larger HDIs) when there are more records in one class than the other. While small margin would probably not cause serious deterioration of modeling accuracy, this preprocessed dataset has 262 0's, which is more than twice as many of 1's (130). Therefore, we further processed the raw data to create a new dataset that have equal number of 1's and 0's with a total number of 260 observations. To eliminate the confounding factor of sample size, we created another dataset by randomly sampling 260 observations from the preprocessed dataset for Bayesian logistic regression modeling. Figure 3 demonstrate the distribution of 0's and 1's in the unbalanced dataset and balanced dataset.

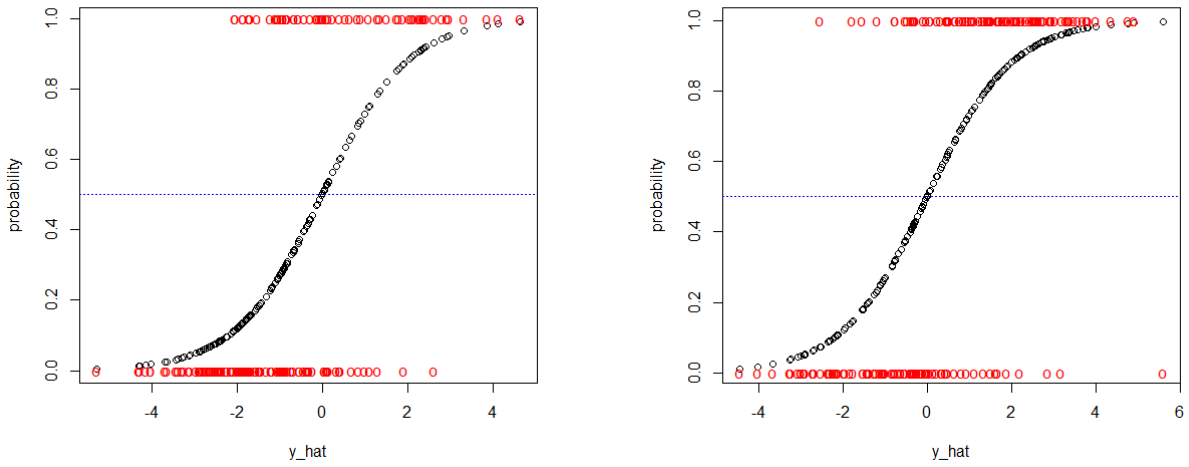


Figure 3. Demonstration of unequal number of 0's and 1's. Left panel: unbalanced (0: 169, 1: 91). Right panel: balanced (130 each class).

Interestingly, the uncertainties that were assumed to be caused by unequal number of records in the two classes were not reduced by balancing the two classes (see Figure 4). The spans of the HDIs for the posterior distributions are similar or even larger compared to unbalanced case. One likely reason would be the reduced sample size introduced more uncertainties. In the meantime,

it is also probable that including some of the variables with weak predictive power complicated the situation. Note that large variations were observed for variables such as blood pressure, skin thickness, insulin between the two cases, suggesting that these variables might be unstable in the logistic regression models.

As was mentioned above, logistic regression model with regularization may solve this issue by placing large penalties on the weight of these variables. We will consider including this analysis in our future work. However, these results suggest that sample size and the variables in the modes play a bigger role in the overall uncertainties of both the Bayesian and conventional logistic regression.

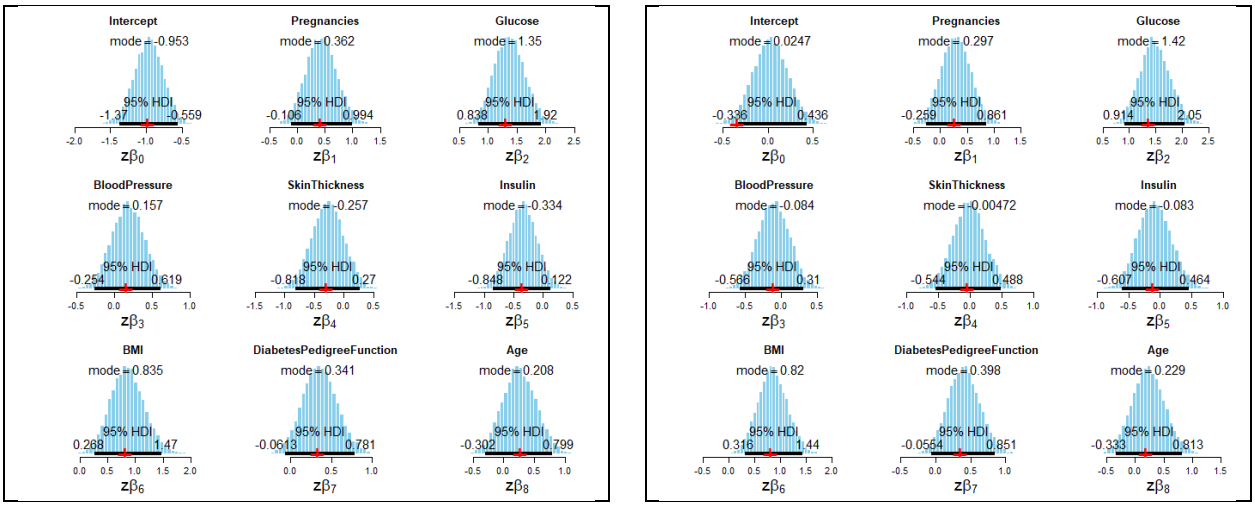


Figure 4. Posterior distributions of the regression coefficients and the intercept. Left panel shows the results without the number of 0's and 1's balanced. Right panel shows the ones with the number of 0's and 1's balanced. Red crosses on the X-axis indicate corresponding values generated by the conventional logistic regression.

Effect of different priors

One of the key advantages for Bayesian methods are associated to its capability of incorporating the impact of prior belief into the posterior probability distribution. For this project, because we do not have reliable priors for these coefficients, a "vague prior" that is defined as a normal distribution, dnorm (0, 0.25), was applied to all the coefficients. In order to understand how the priors would affect the posterior distributions, we analyzed two scenarios: 1) run the MCMC with only 50 randomly selected samples and obtain the posteriors for all the coefficients. Assuming the priors all follow normal distributions, we extracted mean and one quarter of the HDI width and use them to define the prior for a MCMC run with the all the samples. The results are shown in Figure 5.

Changing the priors based on the ones derived from a preliminary MCMC run did not significantly alter the posterior distribution of the coefficients. However, one should note that the

preliminary run has priors that are equal to the control case (dnrom(0, 0.25)) and this may intrinsically complicated the situation. In contrast, a sigma value that is large enough and generally considered non-informative resulted in larger differences compared to the control case. As the posteriors are determined by both the priors and the likelihood, this may imply a relatively weak likelihood built upon a dataset that is either not larger enough in size or a model that is not sufficiently tuned (e.g. excluding weak variables).

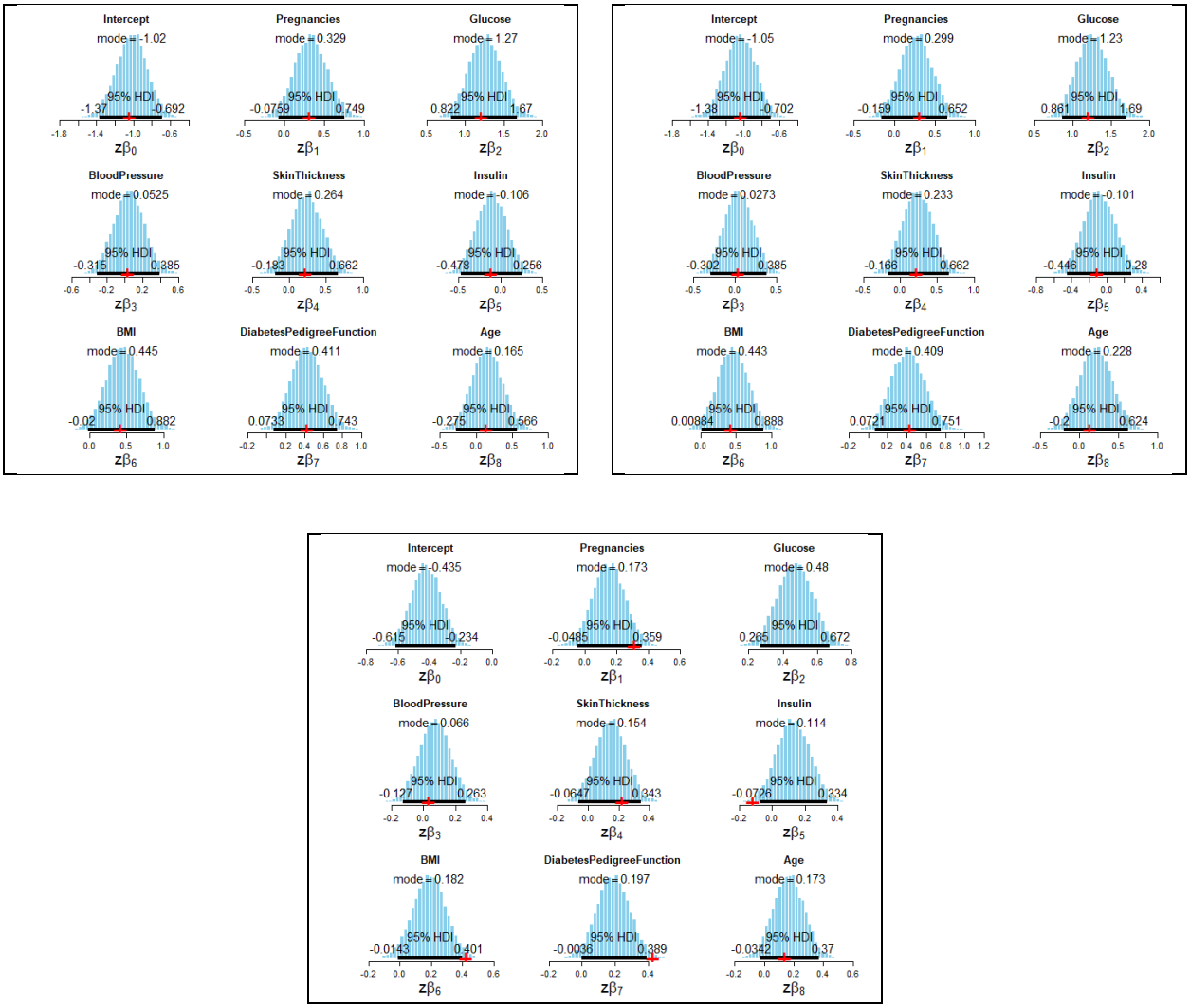


Figure 5. Illustration of the impact of different prior distributions. Top-left panel shows an MCMC run with all coefficients set to dnorm(0, 0.25); Bottom panel shows an MCMC run with all coefficients set to dnorm(0, 50). Top-right panel shows the posteriors from an MCMC run with all coefficients set to a normal distribution derived from running MCMC with 50 randomly selected samples.

CONCLUSION

This project focused on building Bayesian logistic regression modes using JAGS and evaluating the impacts from input data and priors to the posteriors. The Bayesian logistic regression model were evaluated and compared against the conventional logistic regression model as well. In all the cases evaluated in this project, Bayesian logistic regression model and conventional logistic regression model only display marginally differences in their performance gauged by hit rates. For the models built using the Pima dataset, unbalanced number of records in the two classes is not a dictating factor to the posterior distributions of the coefficients, whereas samples size and selection of variables play a more significant role. The evaluation of various priors also suggests that the posterior is synergy of prior and likelihood that is determined by the dataset. The fact that altering priors dramatically had a significant effect on posterior indicate that Pima data may not provide strong enough likelihood.

APPENDIX

The full diagnostic information for the JAGS analysis performed in "*Logistic regression: conventional vs. Bayesian"* are attached in a separated file "Appendix_Diagnostics(Bayesian_vs_logistic).pdf"