**Group Member:**

Xiaodan Chen

Cheng Miao

Xiaoyu Wang

**Predict the onset of Diabetes based on Diagnostic Measures**

**Introduction**：The objective of the project is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements like pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, age. After data preprocessing, we will use MCMC, Logistic Regression, GLM, AUC to get prediction

**Proposed method(why should the method work):**

1. **Bayesian Logistic Regression：** GLM with rstanarm package, which we can efficiently apply well-prepared data to bayesian logistic model and get the posterior distributions of coefficients of predictors. In other words, we can get predictor posterior estimation distributions, which show all possible probabilities of a feature that can affect results.
2. **GLM:** A normal and standard generalized linear method (glm package in R) to do point estimation of coefficients of features.
3. **AUC:** This function computes the area under the receiver operating characteristic curve (AUROC) by pROC package and roc function. Through the value of AUC, we know how useful the model is.
4. **MCMC:** MCMC methods are used as a well defined tool for finding the posterior. Also, MCMC is the need for a likelihood function and for complex Probability model. MCMC starts from the sampling of any state, after sampling for a period of time, which approximates to the stationary distribution. And finally selects the sample after the approximation as the final sample. The target distribution is difficult to use in this case, so that we have eight binary variables X1, X2, X3, X4, X5, X6, X7, X8 are now sampled, and the relationship between the eight variables are really closed. If we sample directly, there is no front-to-back order between the eight variables. After the mcmc, until the sample no longer changes, it can be considered to approach the stationary distribution, take the sample at this time as the final sample. From this we get the posterior probability distribution.

**Experimental results and analysis:**

* **Preprocessing**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. All patients here are females at least 21 years old of Pima Indian heritage.

The dataset is already clean with no missing value, so we don’t need to clean the data. We removed those observation rows with ‘0’ in any of the variables. Then, we transform ‘Outcome’ into factor type and modify the data column names slightly for easier typing.

* **Logistic Regression**

A Bayesian version of logistic regression model can be estimated using the stan\_glm function. So we'll use a Student t prior with 7 degrees of freedom and a scale of 2.5.

We use the stan\_glm function is similar in syntax to [glm](http://www.rdocumentation.org/packages/stats/topics/glm) but rather than performing maximum likelihood estimation of generalized linear models, full Bayesian estimation is performed via MCMC. The Bayesian model adds priors on the coefficients of the GLM.

post <- stan\_glm(outcome ~ ., data = diabetes, family = binomial(link = "logit"),

prior = t\_prior, prior\_intercept = t\_prior, seed = 1)

Next, we use ggplot2 to plot the distribution of posteriors and find median coefficients in specific confidence interval(90%) for later prediction purpose. As for the prediction, we write the prediction function according to the definition of logistic regression. The results are probabilities between 0 and 1, then we set a threshold 0.5, if the result probability is more than 0.5, we set it to be class 1 while if it’s less than or equal to 0.5, we classify the result into class 0.

5% 95%

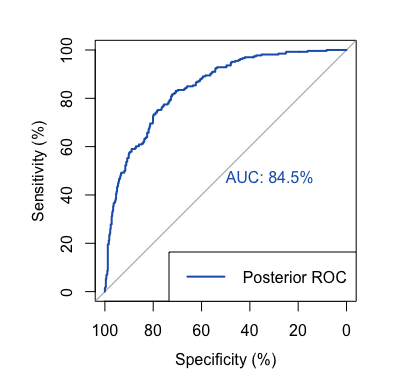
(Intercept) -1.04 -0.72

pregnancies 0.24 0.62 glucose 1.00 1.39 bp -0.44 -0.10 st -0.18 0.19

insulin -0.33 0.02 bmi 0.50 0.90 dpf 0.15 0.49 age -0.04 0.34

z = 0.43\*diabetes$pregnancies + 1.19\*diabetes$glucose - 0.27\*diabetes$bp -0.15\*diabetes$insulin + 0.69\*diabetes$bmi + 0.32\*diabetes$dpf + 0.15\*diabetes$age

y<- 1/(1+exp(-z))

For more easily interpretable predictive performance measures, we will compute posterior predictive probabilities and use them to compute ROC and AUC for Logistic Regression. we decided to use median coefficients to do prediction. The median coefficients are given above, we use pROC package to plot ROC and compute AUC as above on the left. AUC is about 84.5%, which is not bad.

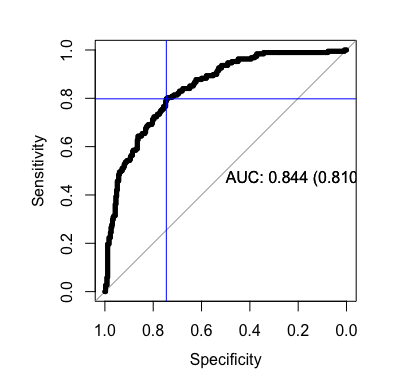
Until now, we used Logistic Regression get the prediction, we are interested other GLM method and how different with Bayesian one. Below is the application of GLM.

* **Simple GLM**

Based on the same data preprocessing, we simply change ‘Outcome’ as a factor and classify the results as ‘Pos’ and ‘Neg’ that is easier for using caret package. The first step is to create train and test set by data splitting function:   
trainIndex <- createDataPartition(dt$Outcome, p = .7, list = FALSE, times = 1)

In this case we may turn to k-fold cross validation, the most popular flavor of which being 10-fold cross validation.

fitControl = trainControl(method = "repeatedcv", number = 10, repeats = 20, summaryFunction = twoClassSummary, classProbs = TRUE, savePredictions = T)

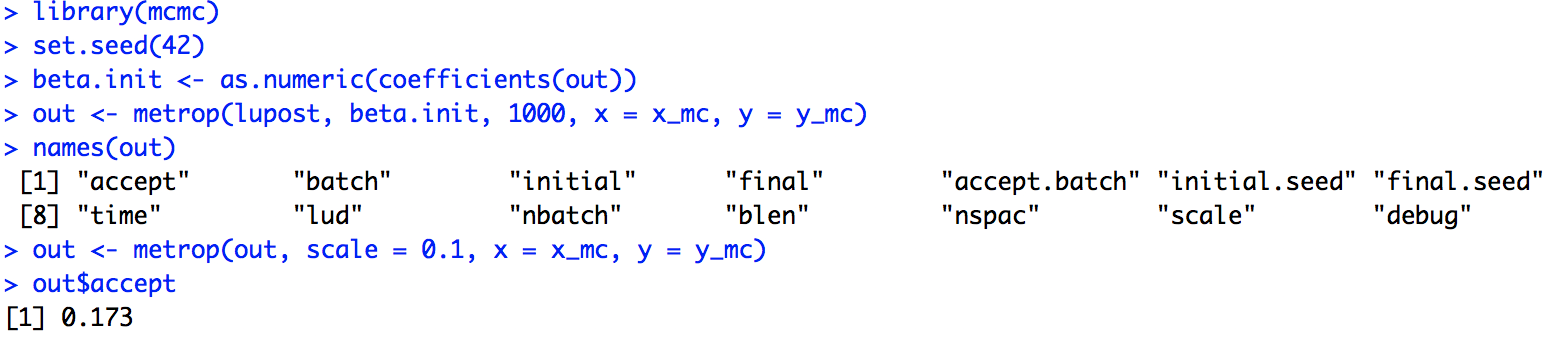
In k-fold cross validation, the data set is split randomly into k partitions, We then fit our model to a data set consisting of k-1 of the original k parts, and use the remaining portion for validation. That is we estimate the out of sample error using the portion of data left out of the fitting procedure. We repeat this k times and our estimate for the out of sample error is the the average over the k validation runs.

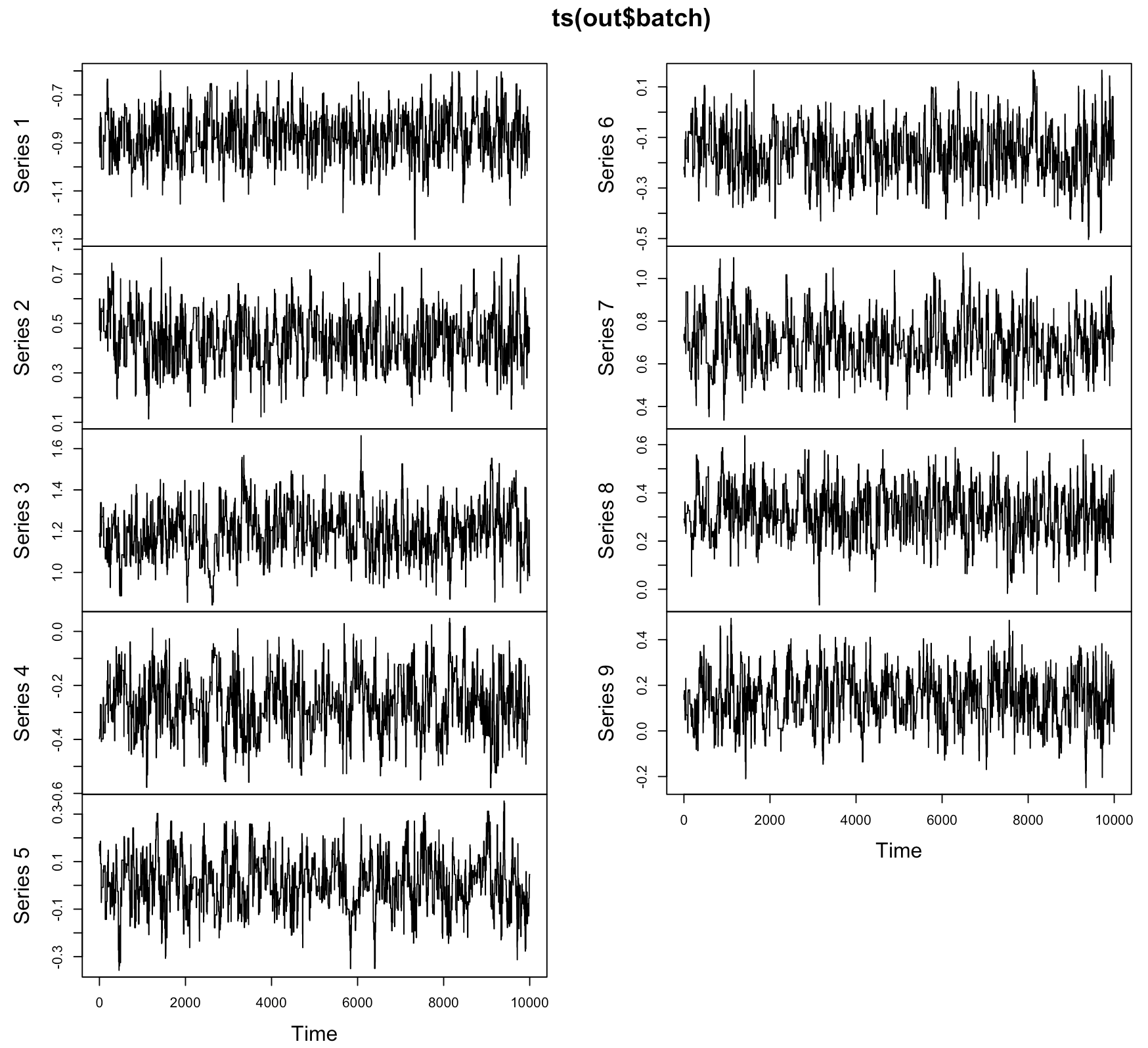
Then, we use train function and training set to predict by GLM:

model = train(Outcome ~., data = train, method = 'glm', preProcess = c('center', 'scale'), trControl = fitControl, metric = 'ROC')

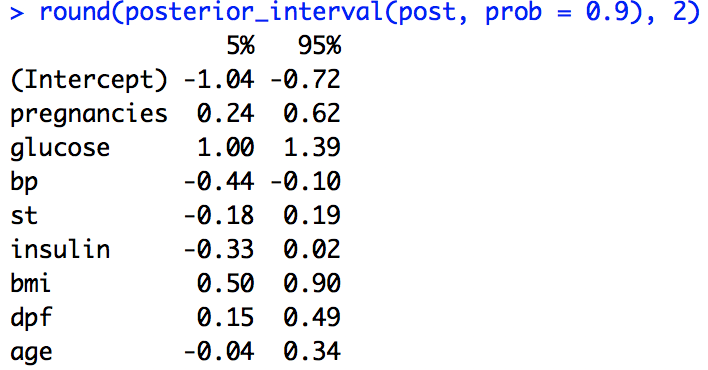
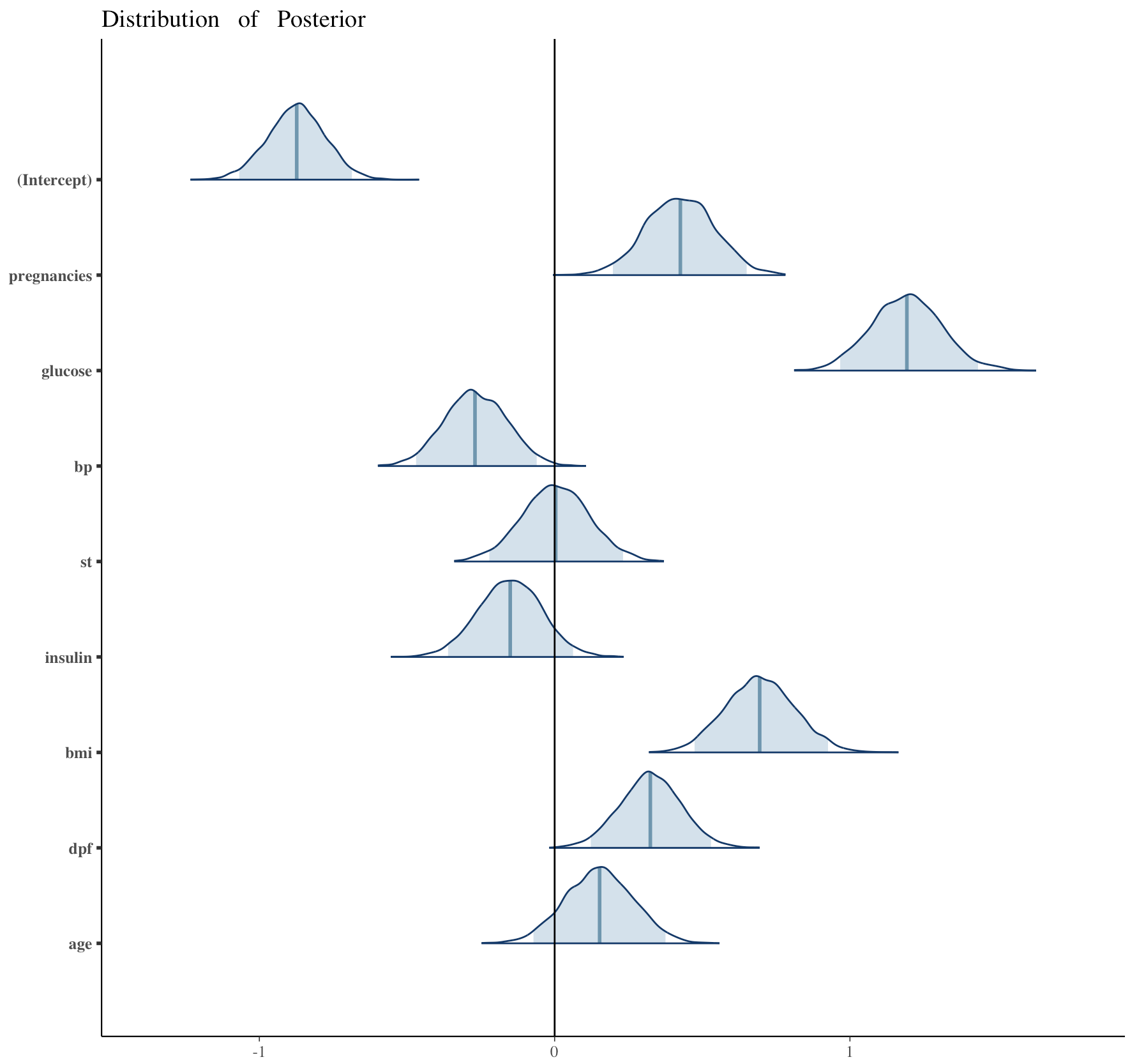
The final model of our prediction system gives us the value of each coefficient of the variables used in the system. Accordingly, compute AUC for simple GLM as shown above on the right with AUC is about 84.4%. Two models’ AUC are pretty close or even the same, 84.5% and 84.4%.

* **MCMC**

First, according to the Bayesian method of the problem, a non-standardized posterior distribution function is constructed. Here is the construction of x, y and beta, and build function to called “Logarithmic posterior distribution probability density”. Then run the MCMC initially to determine the appropriate scale. Next, determine the appropriate simulation batch and batch length. An lupost function that evaluates the log unnormalized density of the desired stationary distribution of the Markov chain. The output is in the component out$batch returned by the metrop function. in order to get a higher acceptance rate (out$accept), It is generally accepted that an acceptance rate of about 20% is right.

In the time series trace plots, we want to try to avoid flat bits in one direction. In this case, it looks like there was a about 1000 iterations, after which the MCMC sampler seems to mix well.

**“stan\_glm”** returns the posterior distribution for the 8 parameters. This figure can intuitively see the posterior probability distribution of the eight parameters. The corresponding interval posterior median estimates using 'coef' function and to get a sense for the uncertainty in our estimated we can use the posterior\_interval function to get Bayesian uncertainty intervals.



**CONCLUSION：**

The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. Diabetes mellitus is a chronic disease and a major public health challenge worldwide. Using Bayesian methods to aid people to predict diabetes has gain major popularity. The project was proposed to predict the persons whether diabetic or not.

There are some limitations of this study. Firstly, considering the diabetes dataset, there might be other risk factors that the data collections did not consider. According to, other important factors include gestational diabetes, family history, metabolic syndrome, smoking, inactive lifestyles, certain dietary patterns etc. MCMC is very useful for calculating posterior probability distributions, but our data volume is still very small. MCMC can be used for high-dimensional data samples, which will have more research value. The proper prediction model would need more data gathering to make it more accurate. This can be achieved by collecting diabetes datasets from multiple sources, generating a model from each dataset. Secondly, in this study we only use GLM to predict diabetes. In order to find a best prediction model, other machine learning methods such as Neural Network will be tested to compare the predicting results.