**Georgia Institute of Technology**

**ISYE 6420: Bayesian Statistics**

**Project – SAT Score Data by State**

Name: Tianyu Yang

GTid: 903645962

Date: 2020/12/5

## Introduction

In this project, I will use the datasets from the Kaggle to finish the task. I will load the SAT score by States in the United States to doing the Bayesian Regression. Though we have many powerful Bayesian regression software tools, such as WinBUGS, PYMC, MATLAB, R or Python. In this project, I will use OpenBUGS to finish the tasks.

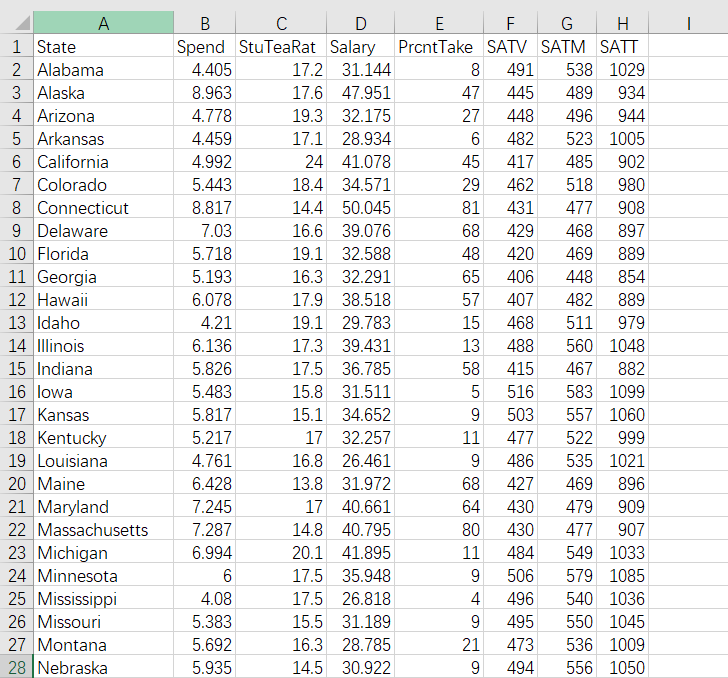
The dataset I use is a very popular dataset. It includes the data of SAT scores statistical data of states in the United States. In this dataset, I will use the Bayesian regression to get the relationship function of spending, student teacher ratio, salary and percentage of students taking SAT exam. It is really interesting to use the dataset from all the states in the United States to analyze this question because from the dataset we can know the factors of SAT examinations and by using the method of Bayesian Regression.

In this project, I will perform the Bayesian regression on this dataset and calculate the relationship of each of the factors in the SAT score dataset. I will use Bayesian Multiple Regression to run the model for this can be more flexible and handling complex models. What’s more, Bayesian model selection is superior (BIC/AIC). Also, Bayesian hierarchical model is easy to extend to many levels and can be much more accurate for the samples in small count. Also, Bayesian model can incorporate the prior information.

## Dataset

The datasets I used is from the Kaggle. It is the SAT Score Data by State. Here is the reference link for that datasets. <https://www.kaggle.com/billbasener/sat-score-data-by-state>.

In this dataset, it contains 8 columns and 50 rows. The columns include State, Spend, StuTeaRat, Salary, PrcntTake, SATV, SATM and SATT.



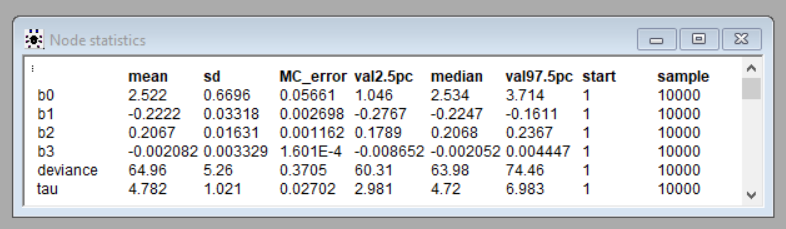
In this dataset, Spend is the spending on SAT. StuTeaRats is the average student teacher ratio for SAT. Salary is the average salary of the teachers. PrcntTake is the percentage of students taking the exam. In this project, I will only use these four columns to do the Bayesian regression analysis.

## Bayesian Regression Method

In this project, I will first let OpenBUGS get the likelihood function of the data and then choose a prior normal distribution for the parameters. Lastly, I will use the Bayes distribution theorem to get the posterior distribution of these parameters.

As for the codes, I write in OpenBUGS for this problem as the file project\_normal.odc which is attached in the submission.

We can get the relationship of Spend (SP), StuTeaRats(STR), Salary(SA) and PrcntTake(PT). In the solution, I use the prior in Normal distribution (0, 0.001) for four variables, which is b0, b1, b2 and b3. What’s more, I also use tau variable to use prior in the Gamma distribution of (0.001, 0.001). In the program, I set the samples count as 10000 and get the node statistics analysis as below:

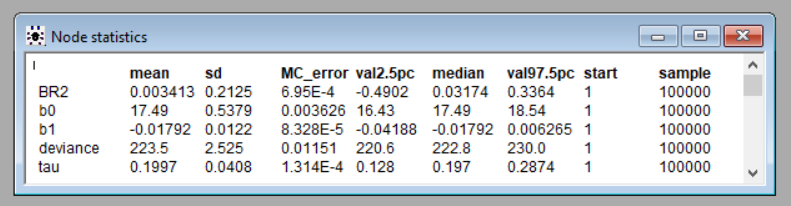


Therefore, the equation of the relationship for all parameters is:

SP = 2.522 + -0.2222 \* STR + 0.2067 \* SAL + -0.002082 \* PT

Besides, from the node statistics above, we can also get that the deviance is 64.96 and tau is 4.782.

What’s more, I also analyze the relationship between StuTeaRats and PrcntTake. I apply the simple linear regression to the dataset as the file project\_SLR.odc which is also attached in the submission. Also, from the code, I also get the correlation value for this simple linear regression model. Here I set the update sample count as 100000. The node statistics analysis is as below:



Therefore, the relationship equation should be:

STR = 17.49 + -0.01792 \* PT

Besides, from the node analysis, we can conclude the Correlation is 0.003413, deviance is 223.5 and tau is 0.1997.

## Conclusion

In conclusion, from the equation, we can get the relationship of Spend (SP), StuTeaRats(STR), Salary(SA) and PrcntTake(PT).

From the first equation we get above, we know that SP depends on STR, SA and PT. All of these three parameters have the relationship of SP. SP has negative correlation of STR and positive correlation of SAL. Though SP has negative correlation with PT but not too much.

What’s more, for the simple linear regression model of PTR and PT. The Bayesian correlation value is only 0.003413, which means that they are not dependent on each other. The relationship of these two variables are negative correlation.

## Reference

The SAT score by state is from the Kaggle <https://www.kaggle.com/billbasener/sat-score-data-by-state>.