

Georgia Institute of Technology
ISYE 6420: Bayesian Statistics

Project – SAT Score Data by State

Name: Tianyu Yang
GTid: 903645962
Date: 2020/12/5

1. 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.

2. 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.

	A	B	C	D	E	F	G	H	I
1	State	Spend	StuTeaRat	Salary	PrcntTake	SATV	SATM	SATT	
2	Alabama	4.405	17.2	31.144	8	491	538	1029	
3	Alaska	8.963	17.6	47.951	47	445	489	934	
4	Arizona	4.778	19.3	32.175	27	448	496	944	
5	Arkansas	4.459	17.1	28.934	6	482	523	1005	
6	California	4.992	24	41.078	45	417	485	902	
7	Colorado	5.443	18.4	34.571	29	462	518	980	
8	Connecticut	8.817	14.4	50.045	81	431	477	908	
9	Delaware	7.03	16.6	39.076	68	429	468	897	
10	Florida	5.718	19.1	32.588	48	420	469	889	
11	Georgia	5.193	16.3	32.291	65	406	448	854	
12	Hawaii	6.078	17.9	38.518	57	407	482	889	
13	Idaho	4.21	19.1	29.783	15	468	511	979	
14	Illinois	6.136	17.3	39.431	13	488	560	1048	
15	Indiana	5.826	17.5	36.785	58	415	467	882	
16	Iowa	5.483	15.8	31.511	5	516	583	1099	
17	Kansas	5.817	15.1	34.652	9	503	557	1060	
18	Kentucky	5.217	17	32.257	11	477	522	999	
19	Louisiana	4.761	16.8	26.461	9	486	535	1021	
20	Maine	6.428	13.8	31.972	68	427	469	896	
21	Maryland	7.245	17	40.661	64	430	479	909	
22	Massachusetts	7.287	14.8	40.795	80	430	477	907	
23	Michigan	6.994	20.1	41.895	11	484	549	1033	
24	Minnesota	6	17.5	35.948	9	506	579	1085	
25	Mississippi	4.08	17.5	26.818	4	496	540	1036	
26	Missouri	5.383	15.5	31.189	9	495	550	1045	
27	Montana	5.692	16.3	28.785	21	473	536	1009	
28	Nebraska	5.935	14.5	30.922	9	494	556	1050	

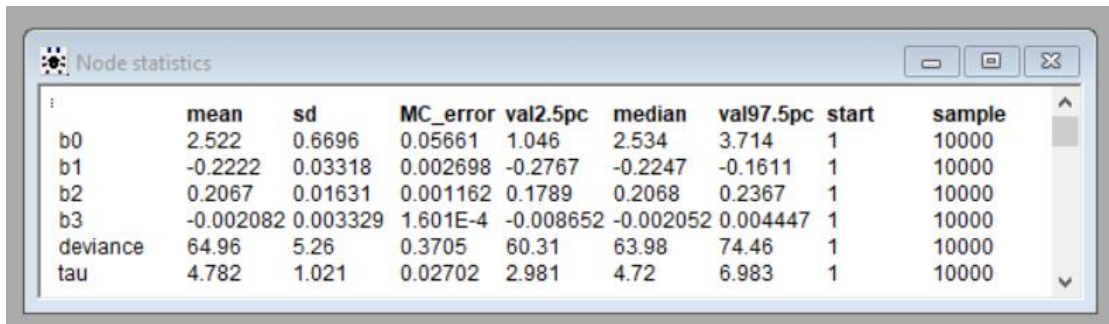
In this dataset, Spend is the spending on SAT. StuTeaRats is the average student teacher ratio for SAT. Salary is the average salary of the teachers. PrcntTake is the percentage of students taking the exam. In this project, I will only use these four columns to do the Bayesian regression analysis.

3. 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:



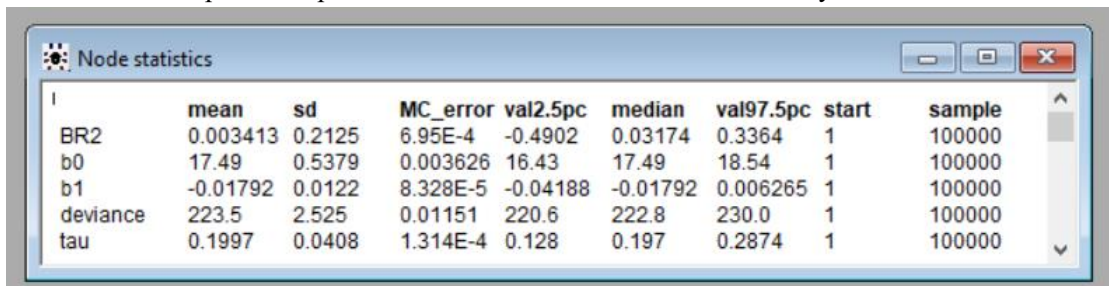
	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
b0	2.522	0.6696	0.05661	1.046	2.534	3.714	1	10000
b1	-0.2222	0.03318	0.002698	-0.2767	-0.2247	-0.1611	1	10000
b2	0.2067	0.01631	0.001162	0.1789	0.2068	0.2367	1	10000
b3	-0.002082	0.003329	1.601E-4	-0.008652	-0.002052	0.004447	1	10000
deviance	64.96	5.26	0.3705	60.31	63.98	74.46	1	10000
tau	4.782	1.021	0.02702	2.981	4.72	6.983	1	10000

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:



	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
BR2	0.003413	0.2125	6.95E-4	-0.4902	0.03174	0.3364	1	100000
b0	17.49	0.5379	0.003626	16.43	17.49	18.54	1	100000
b1	-0.01792	0.0122	8.328E-5	-0.04188	-0.01792	0.006265	1	100000
deviance	223.5	2.525	0.01151	220.6	222.8	230.0	1	100000
tau	0.1997	0.0408	1.314E-4	0.128	0.197	0.2874	1	100000

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.

4. 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.

5. Reference

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