Bayesian statistics modelling for SAT scoring

1. **Executive summary/abstract:** SAT Exam is the most important examination for college admissions in USA. Here we wound analyse the factors influencing the SAT score, starting from expenditure per pupil to student participation rate. We approached the problem with one Bayesian regression model along with noninformative priors and another along with informative priors. We found that both the models showed good convergence but the second model performed slightly better than the first one according to the model comparison method, namely deviance information criterion. The residuals were +-10% of the mean SAT scores and did not show any specific pattern. We found salary to be highest positive influencer for SAT score whereas number of test takers to be highest negative influencer. Current expenditure showed only 70% chance that it positively impacts SAT scores.
2. **Introduction:** Our objective is to identify external factors responsible for the students' performance in SAT examination. Here we are going to address such independent factors, their correlation with the SAT score. We will address the following issues for our analysis:
    - Whether higher current expenditure leads to better students' performance
    - SAT score variation due to change in pupil/teacher ratio
    - Whether salary increment of teachers causes increase in SAT scores
    - Influence of student participation rate on individual test scores
    - Most influential explanatory variable
    - Comparison of baseline model (with non-informative priors) and final model (with informative priors)
3. **Data:** The data we have chosen for our study is the popular dataset from Guber (1999) in his paper 'Getting What You Pay For: The Debate over Equity in Public School Expenditures'. It contains data that describes the per-pupil expenditure on education and the education outcomes (SAT scores) in various states from 1994–95. By modelling this data, we can find out the relationship between expenditure and education outcomes. The summary of the data is as follows:

Variable	Description
Expend	Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)
Ratio	Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994
Salary	Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)
Takers	Percentage of all eligible students taking the SAT, 1994-95
Verbal	Average verbal SAT score, 1994-95
Math	Average math SAT score, 1994-95
Total	Average total score on the SAT, 1994-95

Table 1: Variable description for SAT scores dataset

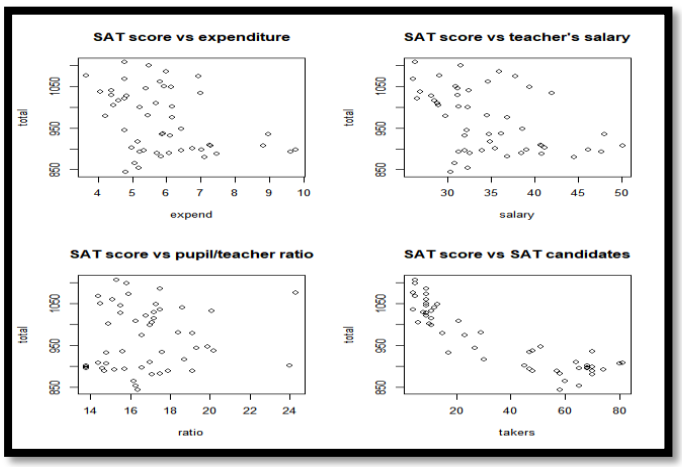

Figure 1: Scatter plots of SAT score against the parameters, current expenditure, pupil/teacher ratio, salary, examination takers

- As the expenditure on education increases, SAT score decreases.
- Pupit to teacher ratio does not show any strong impact on SAT score.
- Teacher salary increment causes slight deterioration in SAT scores.
- High participation rate largely decreases individual SAT score

4. **Model:** We will consider a model with non-informative priors(first model) and another model with informative priors(second model) and then compare them to select the best of them as our final model.

```
set.seed(183)
mod_string13 = " model {
        for (i in 1:length(y)) {
                y[i] ~ dnorm(mu[i], prec)
                mu[i] = alpha + b_Expend*Expend[i] + b_Ratio*Ratio[i]
                        + b_Salary*Salary[i] + b_Takers*Takers[i]

        }

        alpha ~ dnorm(0.0, 1.0/1.0e6)
        b_Expend ~ dnorm(0.0, 1.0/1.0e6)
        b_Ratio ~ dnorm(0.0, 1.0/1.0e6)
        b_Salary ~ dnorm(0.0, 1.0/1.0e6)
        b_Takers ~ dnorm(0.0, 1.0/1.0e6)

        prec ~ dgamma(5.0/2.0, 5.0*10.0/2.0)
        sigma2 = 1.0 / prec
        sigma = sqrt(sigma2)

} "
```

Figure 2(a): Baseline Model

```
set.seed(183)
mod_string11 = " model {
        for (i in 1:length(y)) {
                y[i] ~ dnorm(mu[i], prec)
                mu[i] = alpha + b_Expend*Expend[i] + b_Ratio*Ratio[i]
                + b_Salary*Salary[i] + b_Takers*Takers[i]

        }
        alpha ~ dnorm(9.0e2, 1.0/4.0e4)
        b_Expend ~ dnorm(0.0, 1.0/1.0e1)
        b_Ratio ~ dnorm(0.0, 1.0/1.0e1)
        b_Salary ~ dnorm(0.0, 1.0/1.0e1)
        b_Takers ~ dnorm(0.0, 1.0/1.0e1)

        prec ~ dgamma(5.0/2.0, 5.0*10.0/2.0)
        sigma2 = 1.0 / prec
        sigma = sqrt(sigma2)

} "
```

Figure 2(b): Second model

- For the first model, we have considered normal distributions for all the four independent variables, namely, expend, ratio, salary and takers, and intercept with mean 0.0 and sd as 1000.
- For the second model, we have somewhat informative priors considering the likelihood in the way that how much the final SAT score depends on each of the explanatory variables considered.We assumed similar normal distributions for the four independent variables with mean 0.0 but with different vaiances for each of them since they impact our target variable SAT score differently as per the scatter plots we noticed earlier. We considered the intercept to be following a normal distribution with mean 900 and sd as 200.
- For both the models, we have run **10,000 samples** as **burn in**, so they got discarded. Next we ran 3 chains, each containing **50,000 samples.**

5. **Results:** As we have decided on our model, our first task would be to check the MCMC convergence diagnostics.
    - We can observe from the autocorrelation plots that the intercept i.e., alpha, coefficient for expenditure, pupil-teacher ratio, teacher's salary has high autocorrelation over 40 lags but student participation rate and variance of SAT scores show low autocorrelation.
    - Autocorrelation results have certainly improved in final model as compared to baseline model as evident in Figure 3.

- The trace plots in Figure 4 are roughly flat implying that the chains have converged more or less.

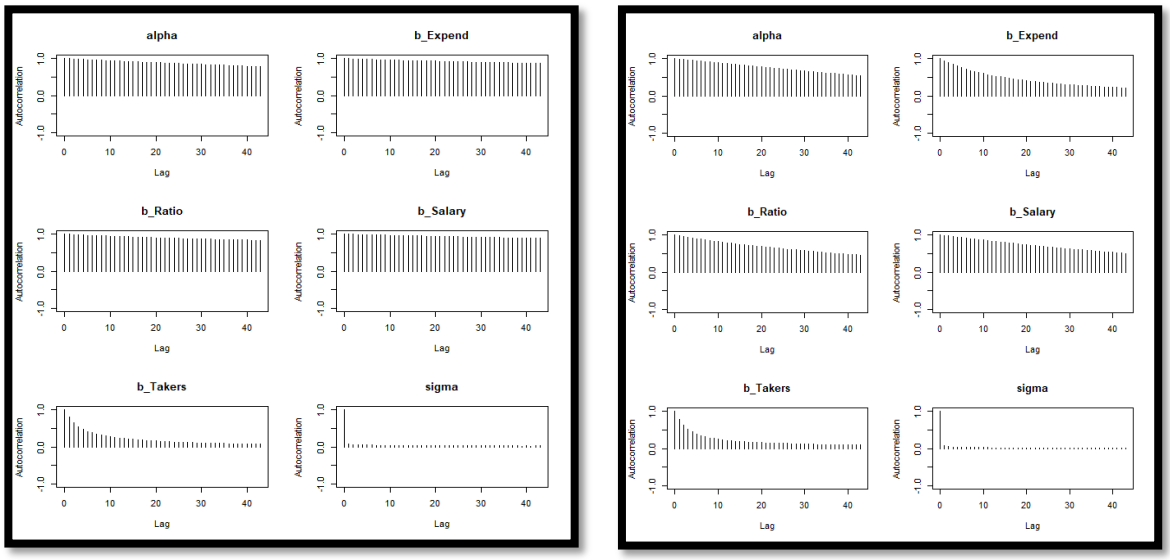

Figure 3: Autocorrelation plots of the parameters for the two model (left one for baseline model and right one for final model)

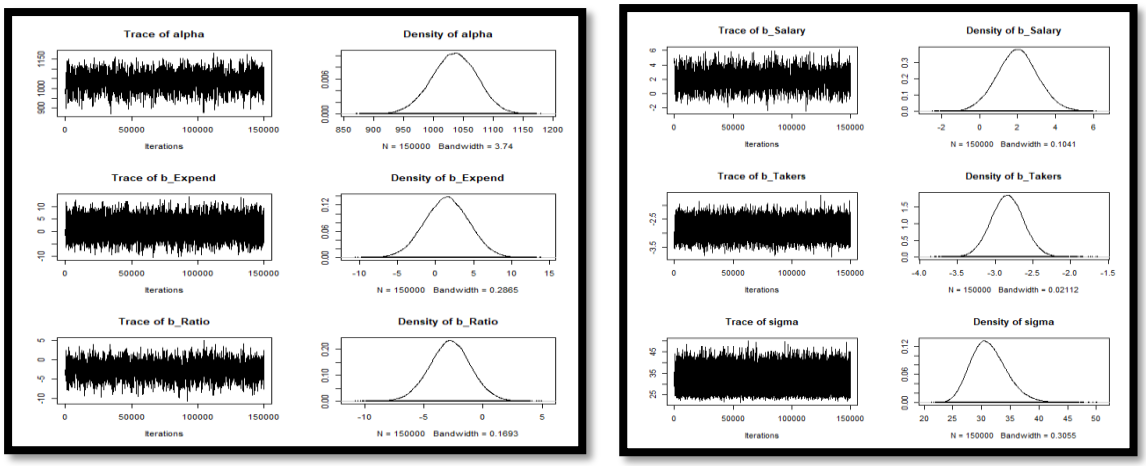

Figure 1: Trace plots of regression coefficients for convergence for final model)

The regression coefficients considered as the mean of the posterior distribution values are as follows:

Parameters	Mean	SD
alpha	1046.019	52.35458
b_Expend	4.13102	10.40014
b_Ratio	-3.67778	3.206845
b_Salary	1.721441	2.334885
b_Takers	-2.907	0.224763
sigma	31.53568	3.217505

Parameters	Mean	SD
alpha	1034.452	38.26456
b_Expend	1.479787	2.931581
b_Ratio	-2.77393	1.763144
b_Salary	1.993279	1.078099
b_Takers	-2.83762	0.217326
sigma	31.32188	3.181241

Table 2: Posterior distribution mean values for the coefficients (left one for baseline model and right one for final model)

From the above tables, we can say that the intercept and coefficients take higher values i.e., the parameters influence the total SAT score more intensely according to the baseline model compared to the final model.

Now let us compare the model using well-known DIC values.

Baseline model	Final model
Mean deviance: 492	Mean deviance: 490.8
penalty 6.927	penalty 4.665
Penalized deviance: 498.9	Penalized deviance: 495.4

Lower DIC values indicate better models. Therefore, our second model is better than the baseline model. Let us evaluate the residuals and plot them for clear insights.

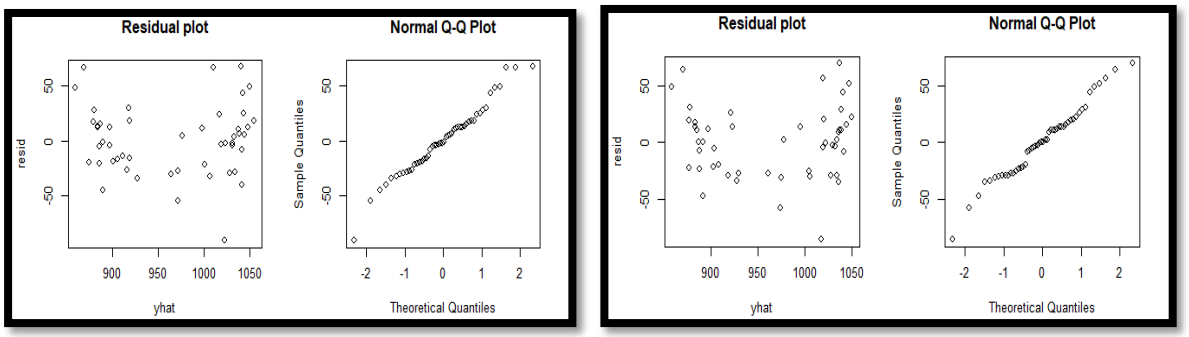

Figure 2: Residual plot and qqnorm plot for baseline model (left) and final model (right)

The residual plot shows that the residuals i.e., the difference between the responses and the predicted values is +-50 as compared to average SAT score of 950. The normal qq plot is almost straight indicating that the residuals are normally distributed.

6. **Conclusions:** We have developed two models for our problem statement, one with noninformative priors and the other with informative priors and thereafter, have observed that the second model performed better than the first as per DIC. Therefore, we have chosen the former as our final model.

$$y_i \sim N(\mu, \sigma^2)$$
$$\mu_i = \alpha + \beta_{Expend} * Expend_i + \beta_{Ratio} * Ratio_i + \beta_{Salary} * Salary_i + \beta_{Takers} * Takers_i$$
$$\alpha \sim (900, 4.0e4)$$
$$\beta_{Expend} \sim (0.0, 10.0)$$
$$\beta_{Ratio} \sim (0.0, 10.0)$$
$$\beta_{Salary} \sim (0.0, 10.0)$$
$$\beta_{Takers} \sim (0.0, 10.0)$$

Now, we will the questions we started our discussion with based on this model.

- We have observed from the posterior distribution that there is only 69% chance that there is a positive correlation between current expenditure and total SAT score
- There is a 94% chance that increase in pupil/teacher ratio will enhance individual student SAT score implying that more students in a class will benefit them for SAT.
- There is a 96% chance of a positive correlation between teacher's salary and total SAT score. Increment in salary will give rise to quality teaching, thus preparing students for better performance in examination.
- There is a 100% chance that greater student participation rate in SAT examination will degrade individual student SAT score.
- Among the explanatory variables considered, salary has highest positive correlation and number of test takers has highest negative correlation against SAT scores.
- Top 5 countries in the dataset showing highest residuals are - North Dakota, New Hampshire, Iowa, Massachusetts, Utah and South Dakota.