Final Project

Entong Li

12/9/2019

Executive Summary:

Research goals:

When students apply to undergraduate or graduate school, they need to provide the information about their parents' educational background, and their race group. It seems like there are some relationships between academic performance and those two factors. However, how these two factors will affect the test score, which is one of the good way to represent their performance, of the students is unclear. Thus, this report will main explore how parents' educational background, race group will help to predict the exam score, or the test preparation course will have more effects on predicting the score.

Main findings and Implications:

Although the better `parental level of education` and different group of `race ethinicity` could make the studnets have higher score in the tests. Most of the time, the `parental level of education` is not as important as people think in predicting whether a student could preform well in school, while `race ethnicity` has more weight in the preformance of students. Also, if a student want to have a good grade or preformance in school, taking the test preparation course is more efficient than having a good background, as people could not change their background and families but they can improve their ability by learning.

Description of Data:

Dataset:

The dataset used to explored is from `Royce Kimmons`. It collects the scores and other informations of 1000 studnets from public school. It could be accessed here: http://roycekimmons.com/tools/generated_data/exams.

Variables:

1. (Response):

- math score (22 - 100)
- reading score (29 - 100)
- writing score (25 - 100)
- total score (87 - 299)

2. (Explanatory):

- parental level of education (from level 1 to 6):
  - some high school, high school, some college, associate's degree, bachelor's degree, master's degree.
- race ethnicity (5 groups): group A to group E.
- test preparation course: completed or none.
- gender: Female or Male

Bayesian Statistical Analyses of Data:

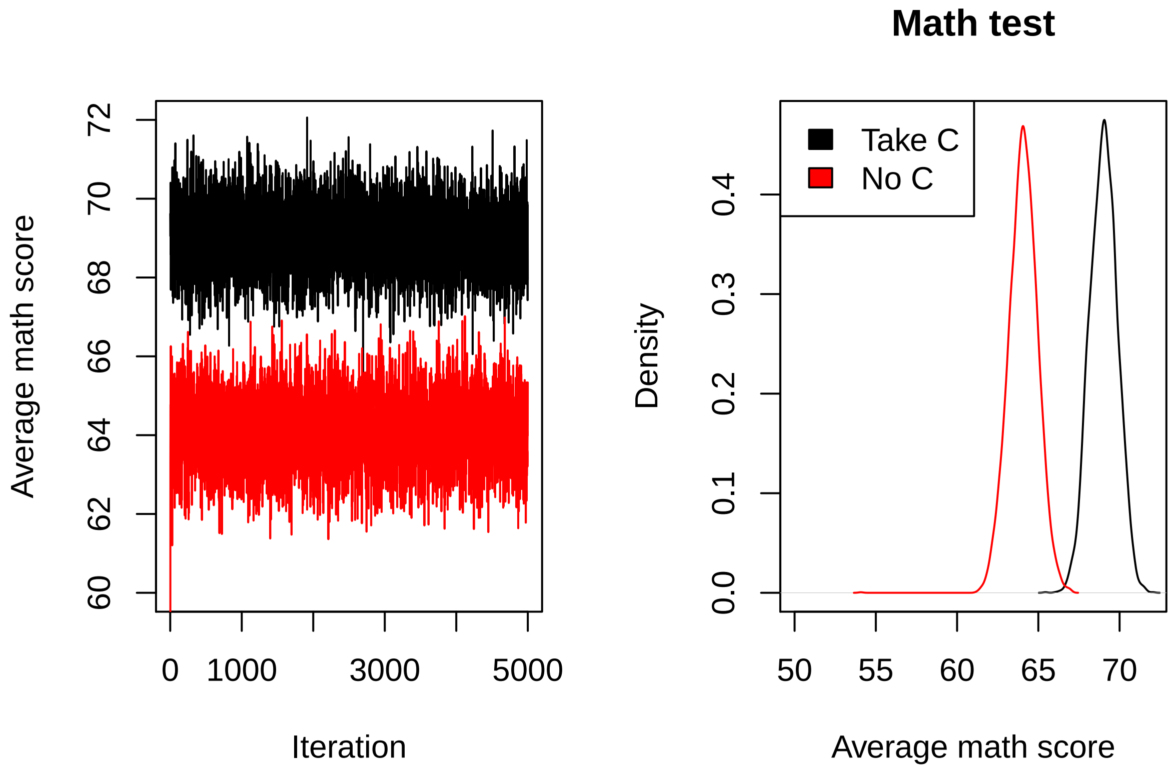
The the density of the test score is pretty similar to the normal distribution, so the conjugate prior about y (response variable) will be nornal distributed. The analyses can be mainly concluded by two steps: first check whether the test preparation course will be an efficient predictor by building multivariate normal models; second step is using `parental level of education`, `race ethnicity`, and `test preparation course` if necessary, to build an linear regression with test score, then use the Bayesian estimation to get the estimated value of slope of each explanatory variables.

Determine whether the course could help the test:

In order to maker sure that whether the `test preparation courses` is a efficient predictor to predict the test score, use the multivatiate model to see whether a student has taken the course will have higher average test score than the one has not. In this step, bivariate normal distribution will be applied to model the data.

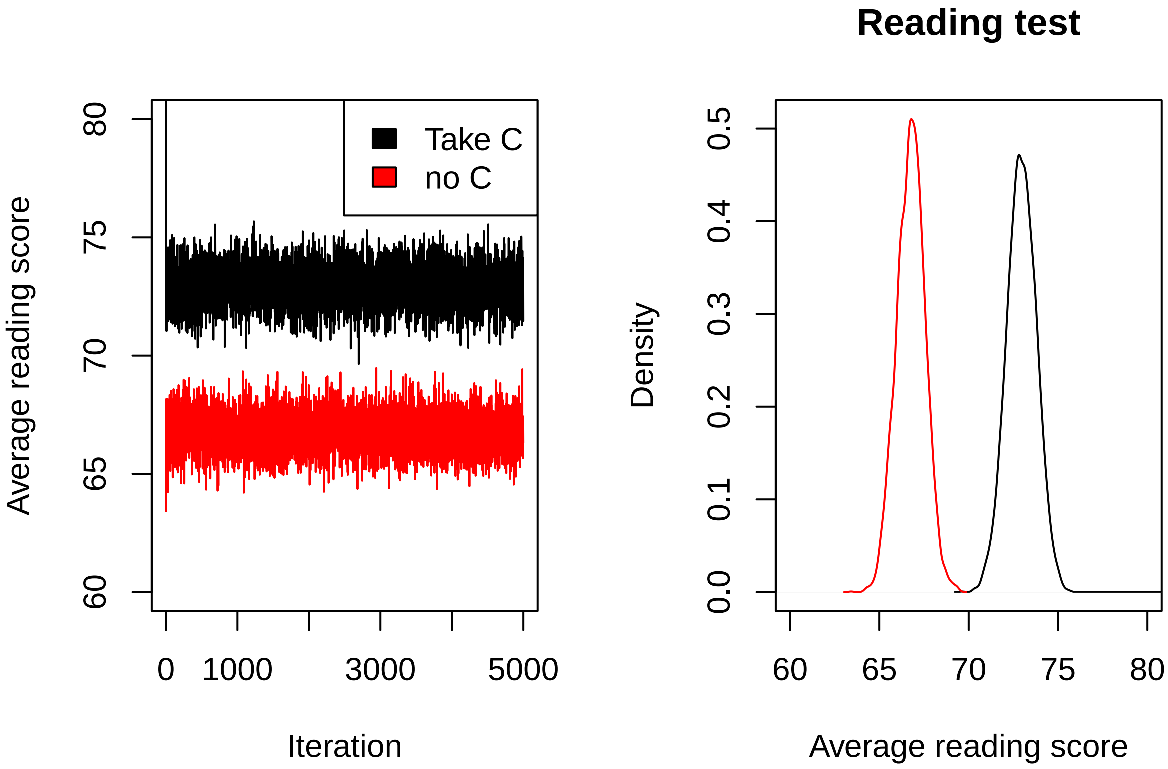
There are total four groups of tests: math, reading, writing, and total. When determining the mean (θ) and variance (Σ) of semi-conjugate prior distribution, the sample mean of data and sample covariance matrix will be used here, and the $\nu_0 = 4$, as each group will have two variables: score of student taken the course, score of student has not taken the course. And in order to save the space, **take test preparation course** will be denoted as **Take C**, and **do not take test preparation course** will be denoted as **No C**.

For math test:



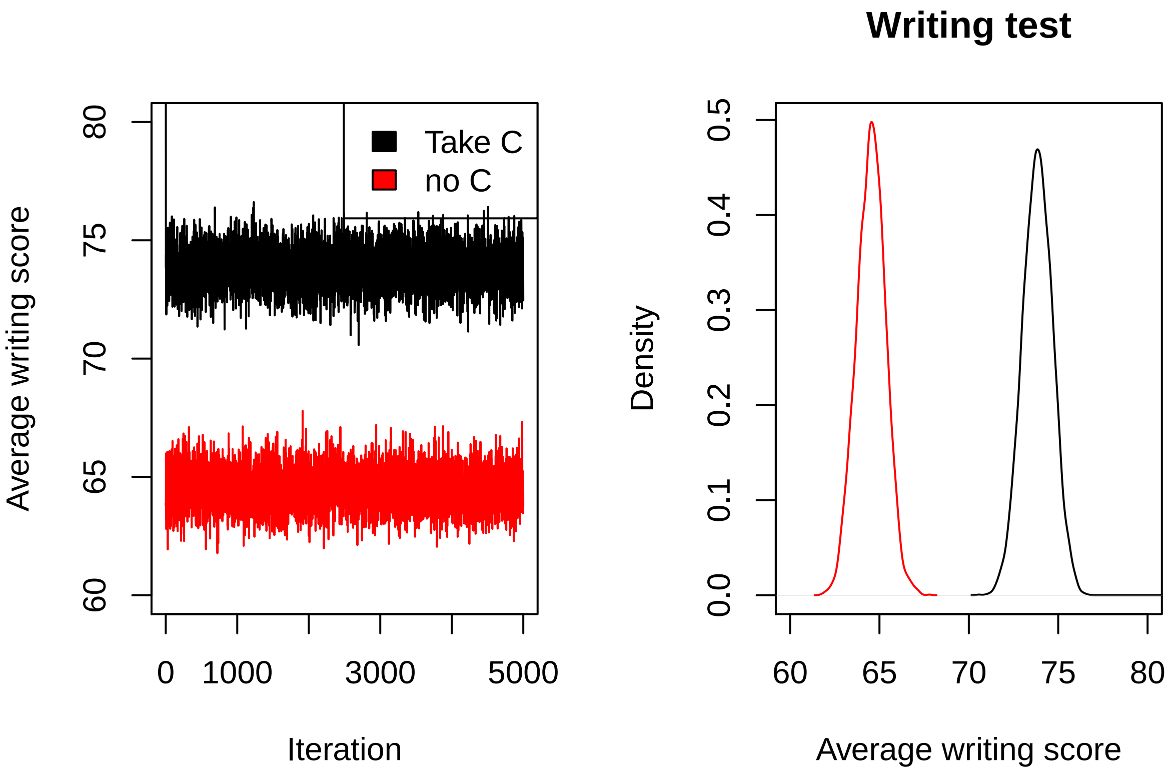
From the graphs above, it's clear that for math test, students who have taken the course are always have higher score than the students has not. The students taken the course will have average math score around 69, while the student didn't take the course will have average math score around 64. So the test preparation course will be a strongly effective and efficient for predicting the math score.

For reading test:



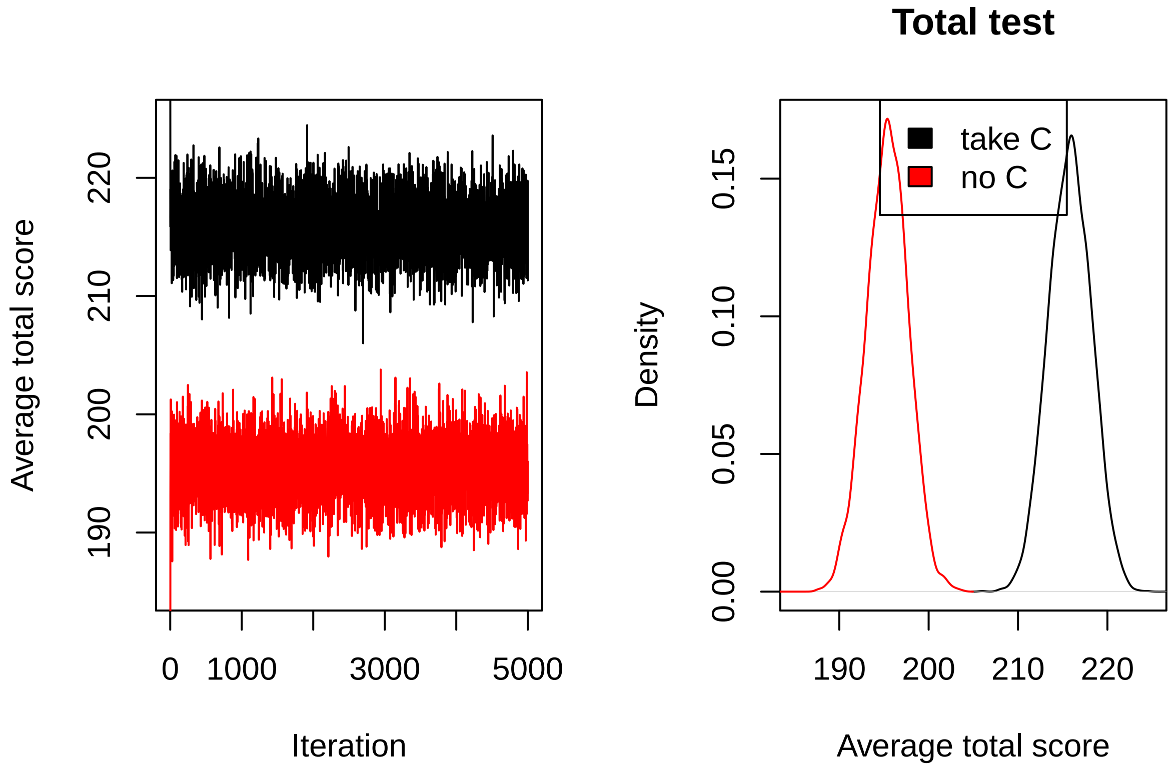
For the reading test, the gap of average reading score between students taken the course and students didn't is larger than the gap of average of math score, the average scores from the `Take C` group will always have higher value than that from `No C` group. Students taken the course will have average reading score around 73, while the students didn't take the course will have average reading score around 67. So the test preparation course will also be a strongly effective and efficient for predicting the reading score.

For writing test:



For the writing score, the average writing score from students taken the course is absolutely higher than that from the students didn't take the course. The difference between these two groups for writing test is more significant than the previous two tests, as the densities of these two groups do not have any overlabed. So, for the writing test, students taken the test preparation course will absolutely have higher score.

For total score:



For total score, the difference of average score between students taken the course and students didn't is very significant. It's very clear that students take the course will have higher average total score than the students didn't take the course. Most of the students take the course will have average total score around 215, while the students didn't take the course will have average total score around 195.

Overall, `test preparation course` is an effecitive and efficient predictor to predict the scores of an student.

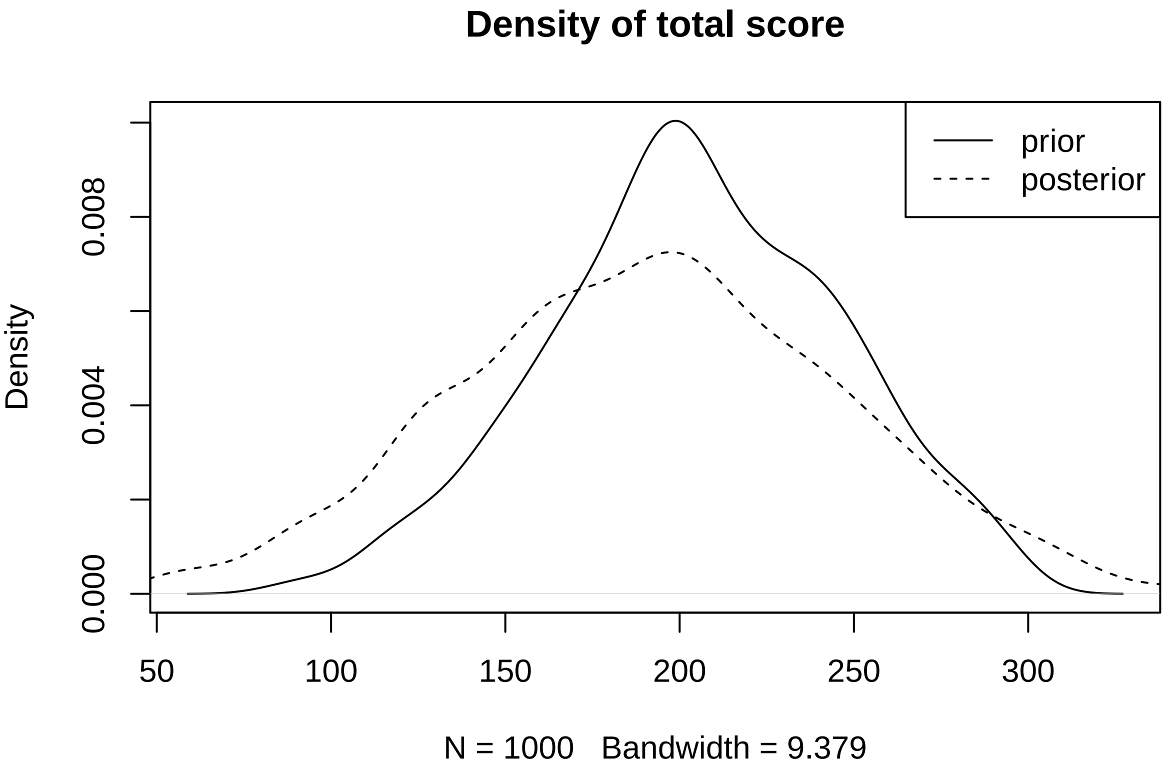
Bayesian Linear Regression:

In order to find out the relationship between score and parental level of education, race ethnictity, and test preparation course, using the regression model will be one of the useful and efficient way. Here,$Y = (y_1, y_2, ..., y_n)$ for $i = 1, 2, 3, ..., n$, and y_i will be represented the total test score,or math score, or reading score, or writing score, of student i, X will be a matrix of $x_{i,j}$. $x_{i,j}$ is the information of student i, where $x_{i,1}$ is parental level of education, $x_{i,2}$ is race ethnicity, and $x_{i,3}$ is test preparation course. The linear model will be $Y \sim N(X\beta, \sigma^2)$.

Y is total score:

Table 1: Table 1: Total Score

	posterior mean	2.5%	97.5%
test preparation course	35.71297	28.11235	43.06796
parental level of education	22.41850	20.28582	24.54042
race ethnicity	35.14234	33.10980	37.36387

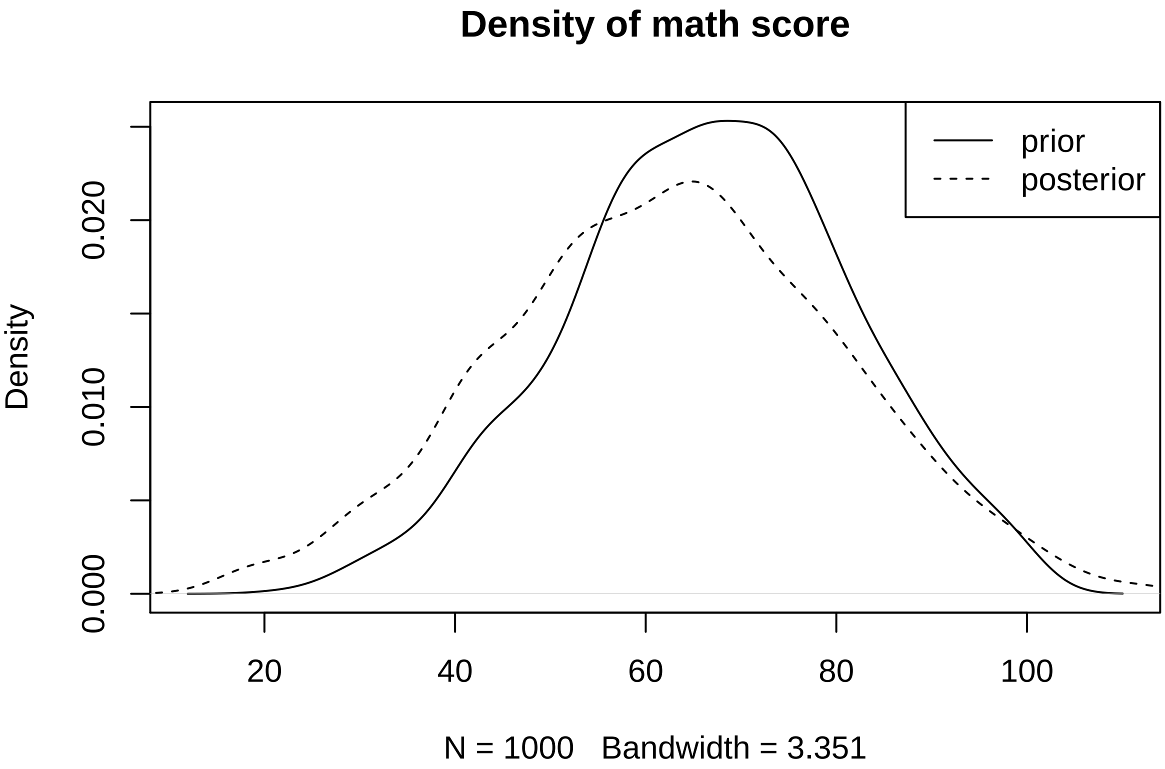


According to the Table 1, the estimated slope of each predictor is positive, meaning that these three variables have positive relationship with the response variable: total score. For total score, the `parental level of education` has the smallest slope, so it affects the total score least comparing to the other two variables. And the `test preparation course` has the largest value of slope, meaning that it affects the total score mostly. And the second graph shows the densities of prior total score and posterior total score. Comparing to the prior distirbution, the posterior one is more spread-out, but they have the same center value, which is also the mean value, 200. So linear model does will for predicting total score.

Y is math score:

Table 2: Table 2: Math Score

	posterior mean	2.5%	97.5%
test preparation course	9.570539	7.007616	12.05063
parental level of education	7.116272	6.397133	7.83178
race ethnicity	11.875014	11.189641	12.62411

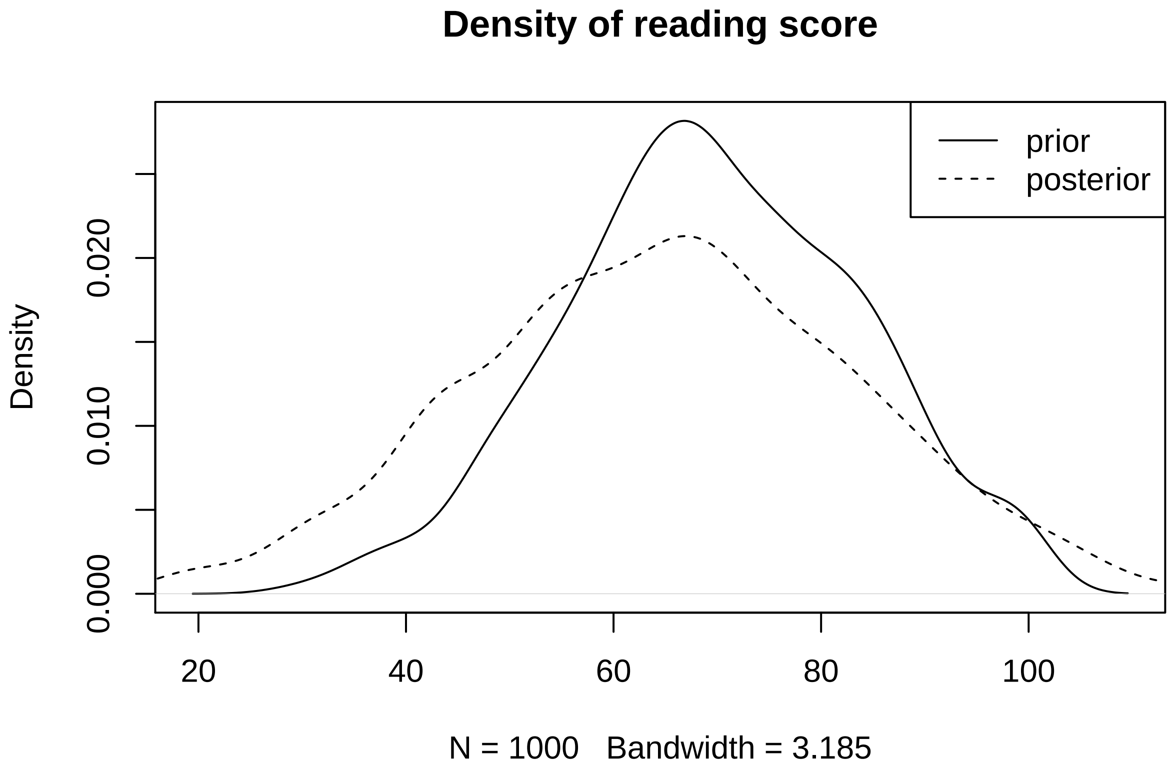


According to the Table 2, all three variables have positive relationship with math score, the increasing of any of these variables will increase the value of math score. And the `parental level of education` still has the smallest slope within three variables, and `race ethnicity` has the largest value of slope, meaning that the `race ethnicity` affects the math score more than other two variables. And in the second graph, the posterior density looks very similar to the prior density, but the posterior one seems to have smaller mean value, as it center at around 65, while the prior one center at around 70. So the linear model does well in math test.

Y is reading score:

Table 3: Table 3: Reading Score

	posterior mean	2.5%	97.5%
test preparation course	11.718924	9.058915	14.292966
parental level of education	7.667204	6.920823	8.409816
race ethnicity	11.905867	11.194532	12.683340

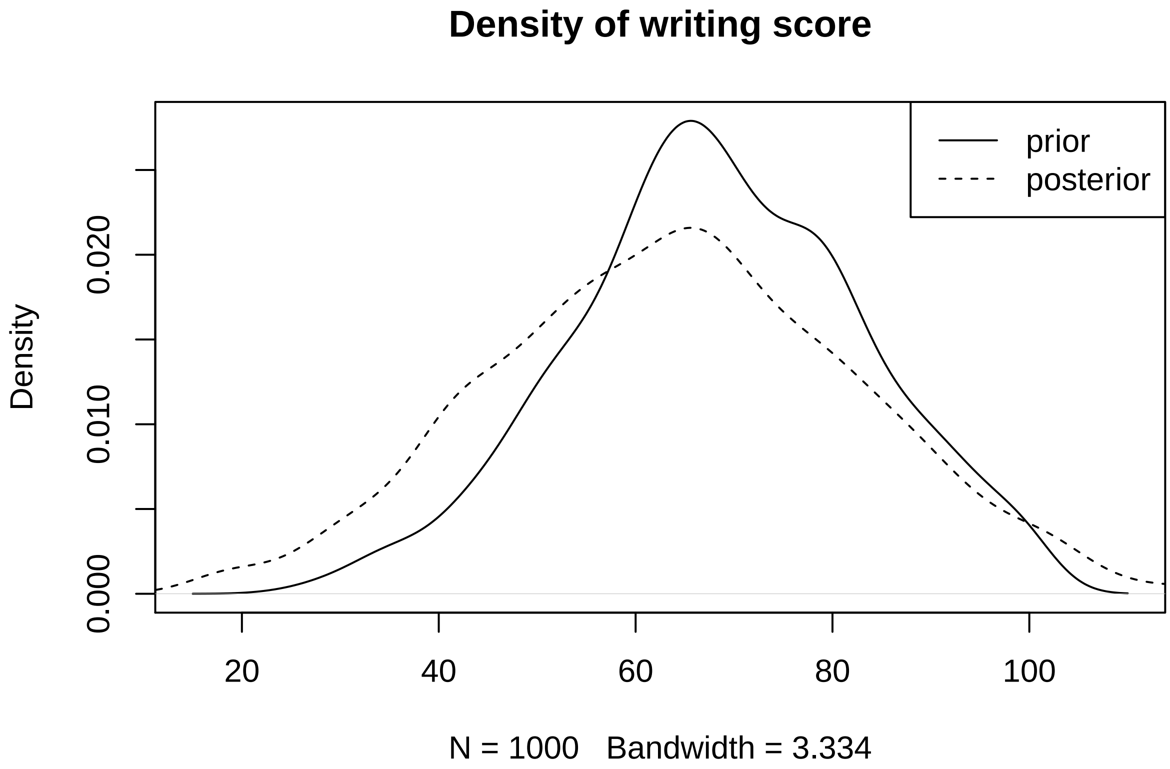


According to the Table 3, it's still the `parental level of education` has the lowest value of slope, and `race ethnicity` still has the largest value of slope, and all of the slopes are postitive. Thus, it could be concluded that for reading test, all these three variables could effect the reading score postitively, and the `race ethnicity` will affect most, and `parental level of education` will affect least. The second graph is the comparison of density between prior and posterior. For the reading test, the density of posterior is more spread-out, but it still has the same center point with the prior one. So the liner model does well in reading test.

Y is writing score:

Table 4: Table 4: Writing Score

	posterior mean	2.5%	97.5%
test preparation course	14.421105	11.863483	16.896071
parental level of education	7.635683	6.918031	8.349712
race ethnicity	11.361258	10.677302	12.108805



According to the Table 4, all these three variables have postive slope, and `parental level of education` still has the lowest value of slope, but for writing test, `test preparation course` has the highest value of slope, showing that all there variables have positive relationship with writing score, and `test preparation course` will affect the writing score most, while `parental level of education` will affect least. In the second graph, for writing tes, the posterior density is more spread-out than the prior density, while they center at the same value, which is around 67. The linear model does well in predicting the writing score.

Conclusion:

Overall, parental level of education and race ethinicity are able to affect the grades or school preformance of students. Their parents have higher level of education could make them have higher grades or better performance in school. And different groups of race ethinicity could also affect studnets' grades and performance, as the collector of this set of data did not annotate which groups in `race ethnicity` represents which race groups, it could not be told that which race would preform better. Although parental level of education and race ethnicity could affect studnet's grades and performance in school positively, these two factors do not granrantee that students could have good grades and performance if they have good background. Comparing to parental level of education and race ethnicity, whether the students have taken the test perparation course matters more. If a studnets could take the test preparation course, he/she would have higher probability to do well in school. Students cannot change the conditions that they born with, like the parental level of education and race, however, they could improve themselves by learning, as learning is more helpful and efficient to imprave their performance than the conditions they are born with.

Furthermore, as this project does not discuss about how the gender of students could affect their grade and performance in school, if there could be a investigation following this project, it will be interesting to find out whehter different gender will have different level of affect on the results in this linear model. Especially there is a stereotype that female does not perform as well as male in math, but do better in reading and writing. It's fun to explore that if the `parental level of education` and `race ethinicity` will have the same effects on score for all genders.