Statistical Analysis of
High School Student Academic Performance

By

Jung-a Kim
Mengqi Yin

November, 2018

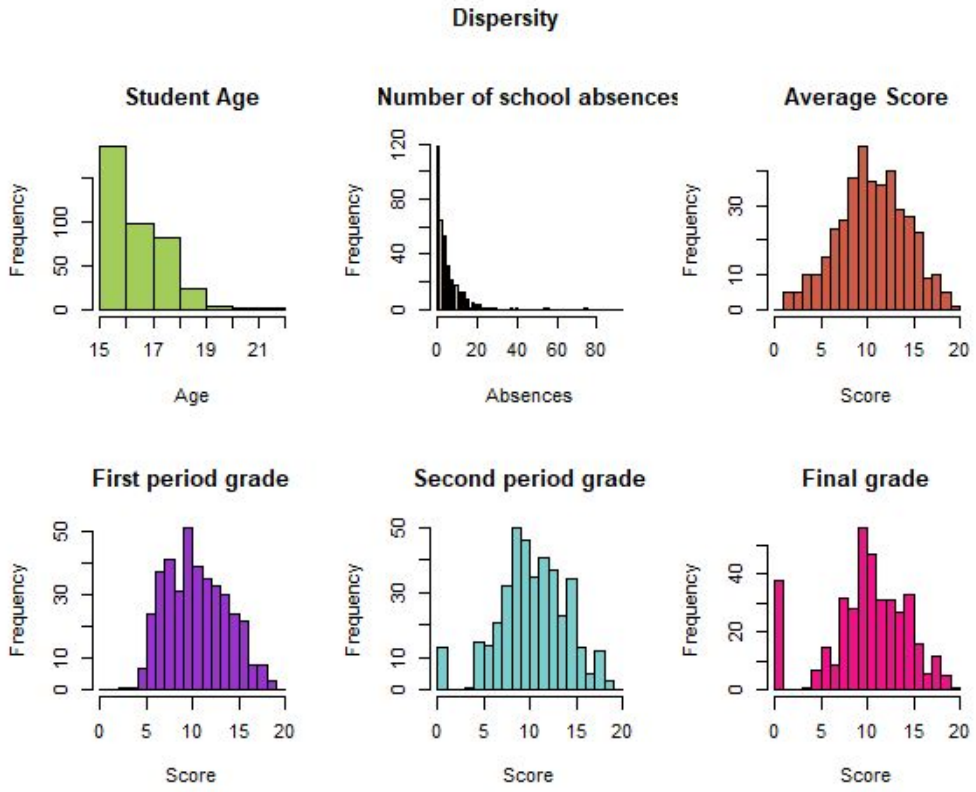
Description of the dataset

This dataset was used for data mining project at University of GuimarÃ£es, Portugal to predict student achievement in secondary education of two Portuguese schools, Gabriel Pereira and Mousinho da Silveira. The dataset has 395 students with the attributes including students' first, second, and final Math exam scores, demographic, social, and school-related features collected by using school reports and questionnaires.The perfect score for each exam is 20.The third exam score(G3) has a strong correlation with the first exam(G1) and the second exam(G2) with r = 0.80 and r = 0.90 respectively.

Variable Description

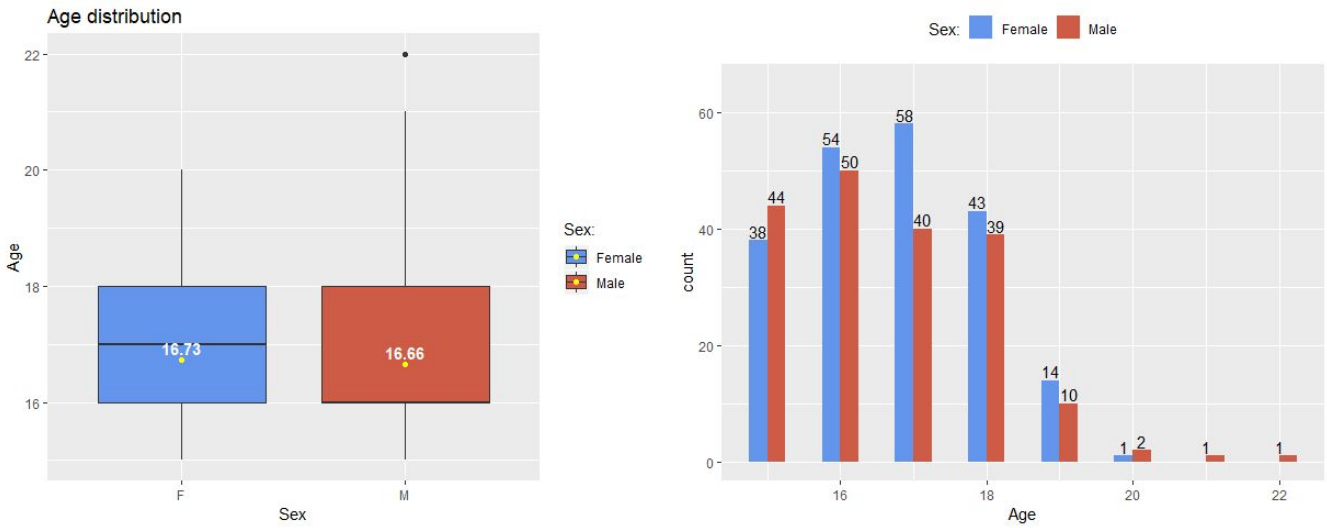
age	student's age (numeric: from 15 to 22)
sex	student's sex (binary: "F" = female, "M" = male)
Pstatus	parent's cohabitation status (binary: "T" = living together, "A" = apart)
studytime	weekly study time: (1) less than 2 hours, (2) 2 to 5 hours, (3) 5 to 10 hours, (4) greater than 10 hours
failures	number of past class failures (numeric: from 1 to 3)
paid	extra paid classes within the course subject (binary: yes or no)
higher	wants to take higher education (binary: 1 = ‘yes’, 0 = ‘no’)
romantic	in a relationship (binary: yes or no)
famrel	quality of family relationships (from 1 = very bad to 5 = excellent)
goout	going out with friends (from 1 = very low to 5 = very high)
absences	number of school absences (numeric: from 0 to 93)
Medu	mother's education: (0) none, (1) primary education (4th grade), (2) 5th to 9th grade, (3) secondary education, (4) higher education
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade
totalG	total grade of the three exams
avgG	average grade of the three exams
avgL	average letter grade of the three exams

Descriptive statistics with graphs



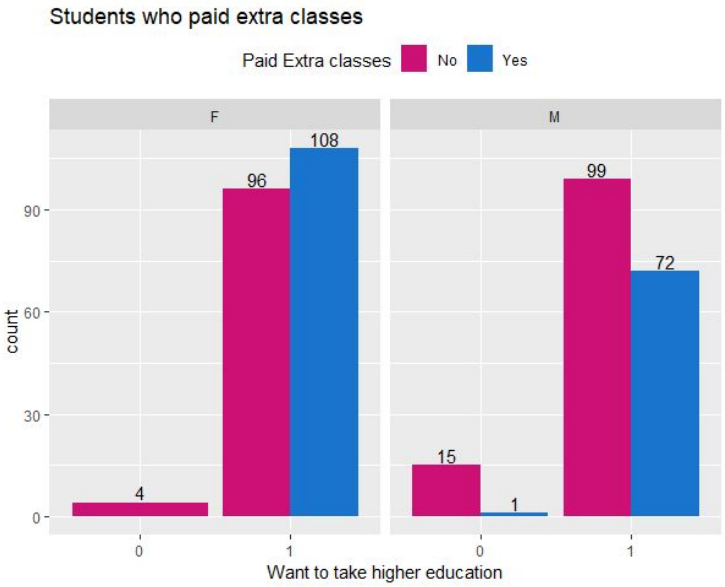
Graph 1. Dispersity in Ages, Absences, and Exam scores

The histograms in Graph 1 show the dispersity of continuous variables age, the number of absences, and grade in the dataset. In "Dispersity in Student Age", it is shown that the majority of students who come to school is about 15-18 years old. For absences, most of the students miss 0-5 classes. The first exam score, second exam score, and final exam score all have approximate symmetric bell-shaped distribution with the same median equal to 11. Those whose second grade is zero have also zero score for their final exams. There are 38 students who have zero score in their final grades.

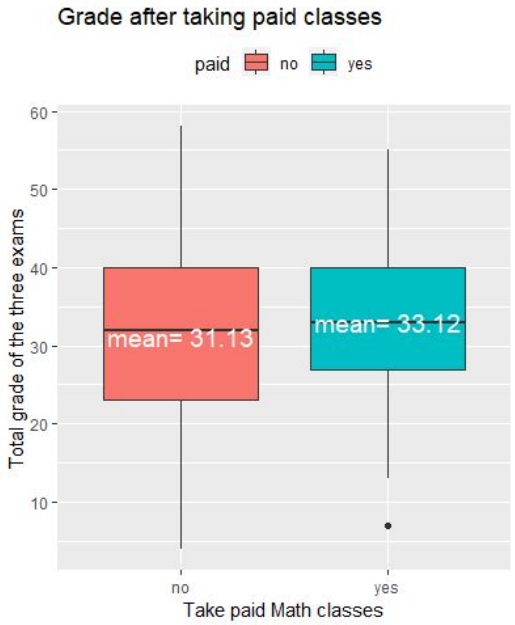


Graph 2. Age distribution in a box-plot and a barplot

The box-plots in Graph 2 shows the difference in mean age between sex. The barplot shows the skewness of the age distribution for both sex. The minimum age is 15, maximum student age is 22, median age is 17 and mean age is 16.7. The mean age is similar between female and male students as the mean age of female students is 16.73 and the mean age of male students is 16.66. An outlier is a 22-year-old male student.



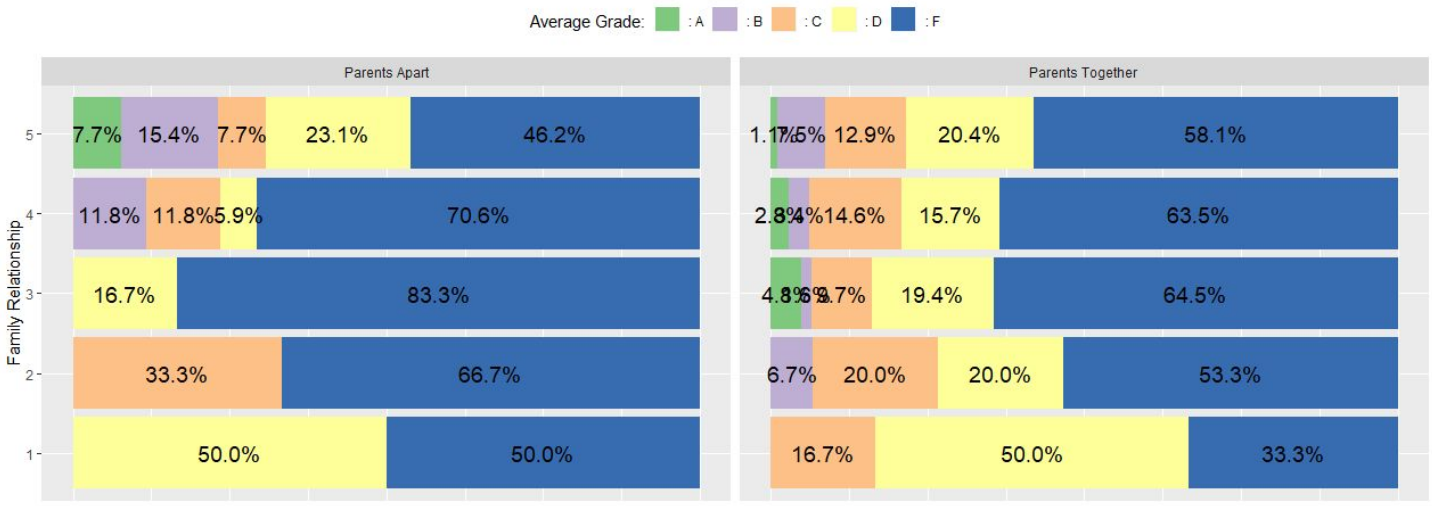
Graph 3. Barplot of students who paid extra classes wrapped by inclination for college



Graph 4. Boxplot of Grade after taking the paid class

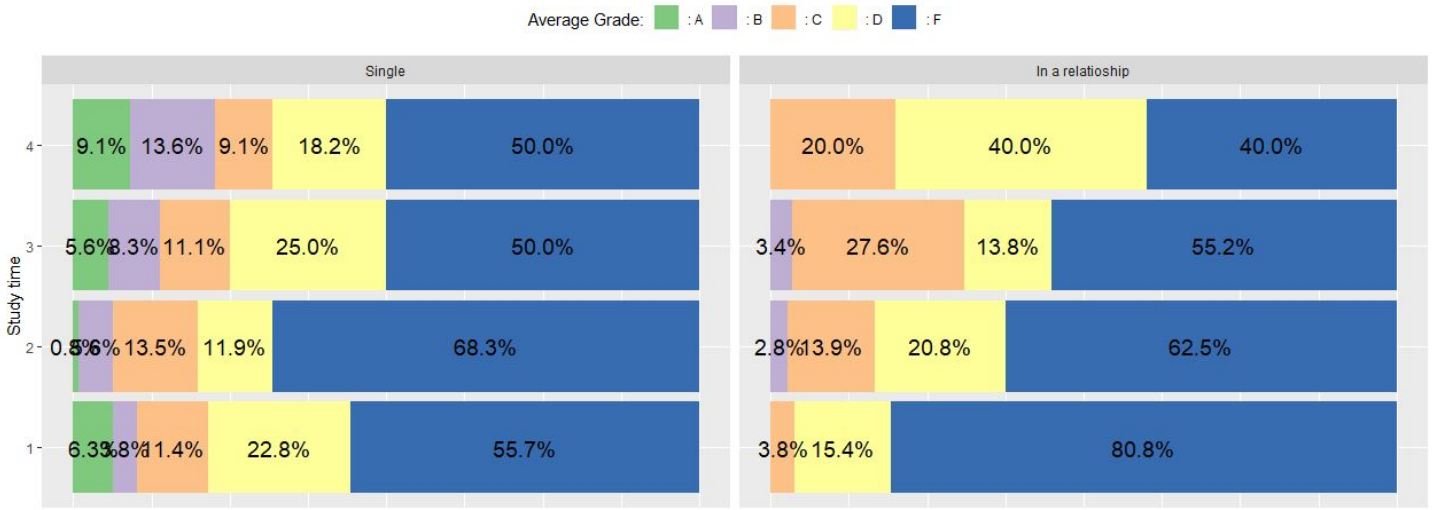
Graph 3 suggests that for female students who want to take higher education, about 53%(108 out of 204 female students) take extra paid math classes. Male students who want to take higher education about 42.1%(73 out of 171 male students) take extra paid math classes. Thus we used two independent sample proportion test to check whether the proportion of taking the paid classes are different between the male and female students who incline to go to college. The null hypothesis that the proportion of taking paid classes between male and female students who want to take higher education is the same. With p-value = 0.03645 < 0.05, we reject H0 and conclude that female students who want to take higher education is more likely to take paid classes than male students.

From Graph 4, it is noticeable that the mean total score for those who took paid math classes and those who did not are very close given that the total score is 60. The median of the score of students who take paid courses is 33 and the median of the score of students who do not take paid courses is 32. Thus, we conducted two-independent-sample t-test to reassure our assumption. With p-value = 0.06922 and significance level = 0.05, we failed to reject that the mean total score of students who take paid math classes is different from the mean total score of those who did not take paid math classes and concluded that the mean total scores are the same.



Graph 5. Barplot of Average Grade according to Family relationship and Parent cohabitation status

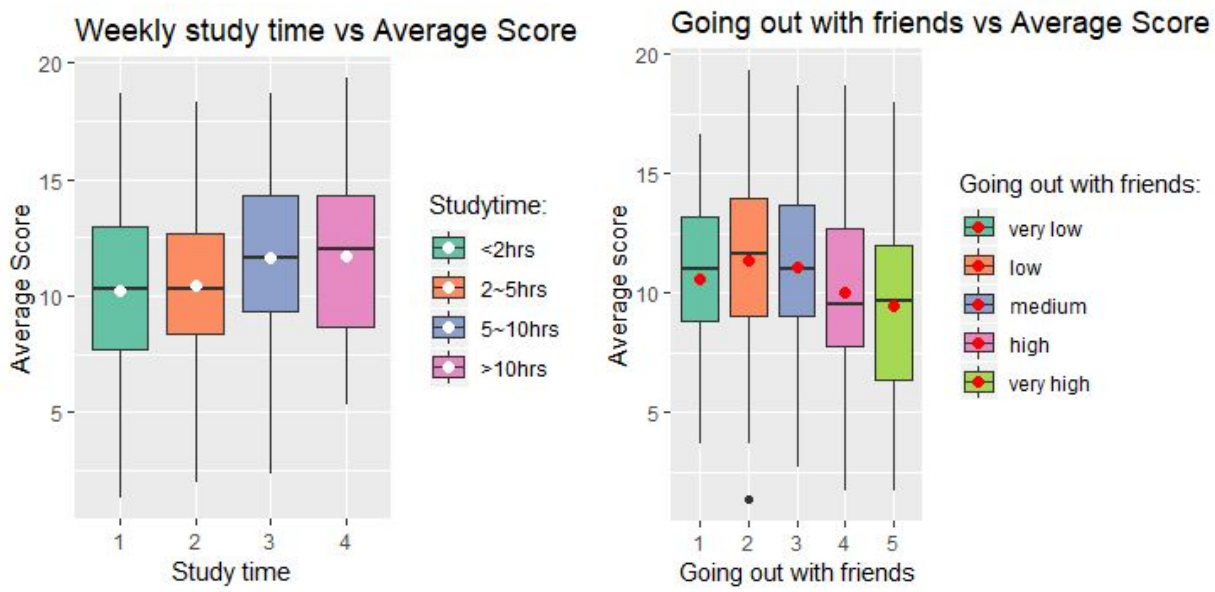
The barplot in Graph 5 implies if the parents are together, the failure ratio is moderate across all the family relationship levels. Overall, failure ratio is greater when the parents are apart than when the parents are together. If the family relationship is excellent, the ratio of getting “A” grade is higher when parents are apart. But since there's only 41 out of 395 students whose parents are apart, so there is some bias in comparing the ratio between the two groups.



Graph 6. Barplot of Average Grade according to weekly Studying hours and Relationship status

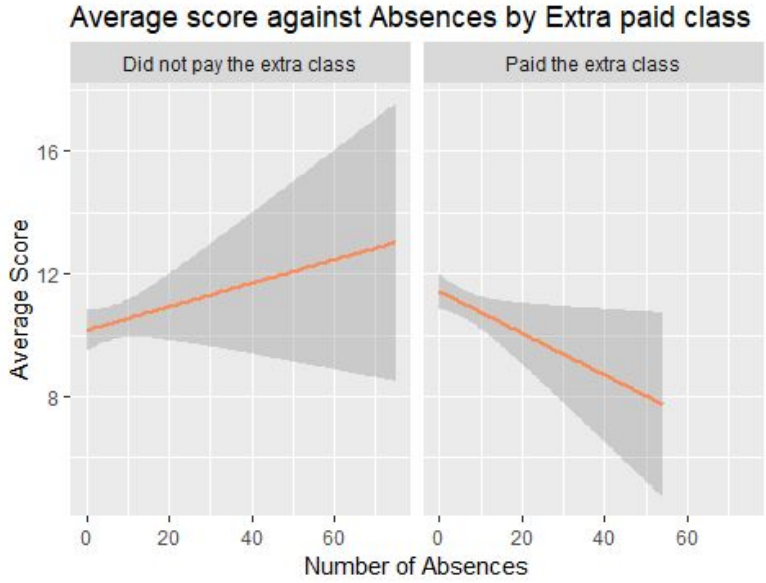
In Graph 6, for the students in a relationship, those who spend more than 10 hours on studying per week don't have 'A' or 'B' as their average grades. Furthermore, there is no ‘A’ grade across all the studying levels. 'B' is also rare for those who study more than 2 hours per week. Those who study less than 2 hours per week have grade C or lower. Also, they are more likely to fail than those who are single, if they spend less than 2 hours on studying per week.

For those who are single and study more than 10 hours per week, 9% of them received 'A'. Failure rate of those who study 2~10 hours per week is similar to that of those in a relationship.



Graph 7. Boxplots of Average Score against weekly studying time and level of going out with friends

From the first box plot in Graph 7, we can verify that the mean scores of those who study 5~10 hours is not much different than those who study more than 10 hours per week. From the second box plot, we can see that those who hang out very much has the lowest average score. But those who hang out with friends at medium level is about similar to those who hang out occasionally.

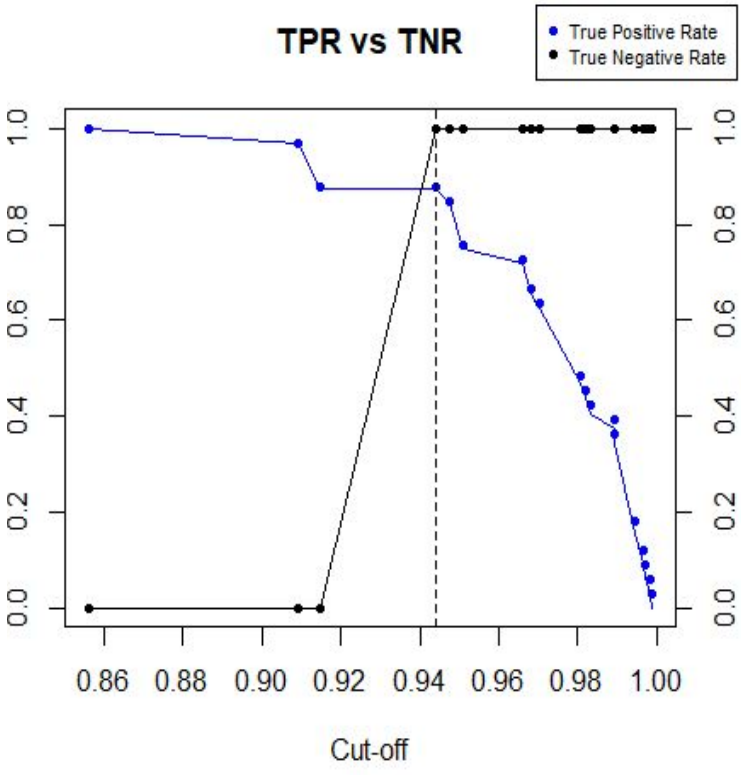


Graph 8. Simple linear regression of Average Score against the number of Absences wrapped by the payment for extra classes

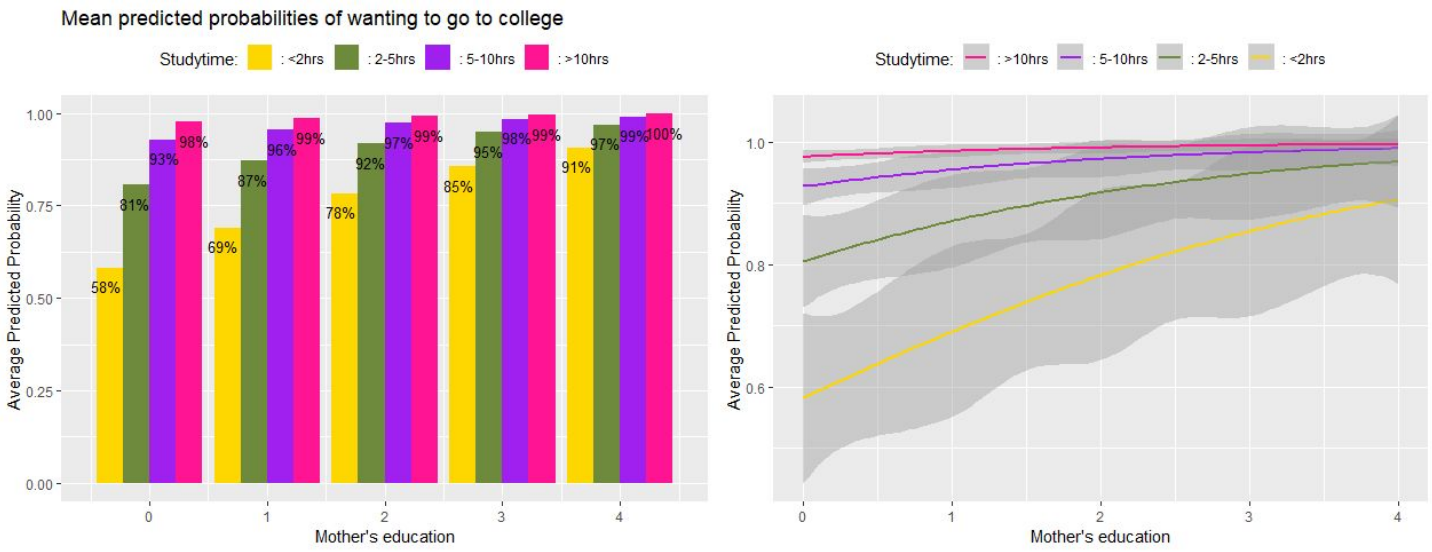
Graph 8 shows the simple linear regression line with the average score as the response regressiong on the number of absences. The right plot indicates that if the students paid the extra class, the number of absences and the average score has a negative linear relationship. From the left plot, if they did not pay the extra class, the absences and the average score have a positive relationship. It is noticeable that those two plots have a large variance for the large number of absences. There are only 26 students out of 395 students who were absent more than 16 times. Thus, the error is large for the number of absences greater than 16.

Logistic Regression

The binary variable 'higher' indicates whether a student wants to proceed to higher education or not. We wanted to predict how likely a student would want to go to college given the mother's education, the weekly studytime, and the number of failures in the past classes. So we splitted the dataset into two parts: one for modeling and the other for validation set. In the student dataset, there were 20 students who did not want to go to college and 375 students who wanted to go to college. Thus to prevent the class bias, we separated the sample by the 'higher' values(either 0 or 1) and randomly selected 90% of the sample with higher = 1, assigning them to the training data. The remaining 10% was reserved as the new dataset for validation use. Then we used 'glm' function in R to create the logistic regression model. We checked the variation inflation factor for each predictor and they were all close to 1, which means they were linearly independent. For the threshold to determine whether the student wants to proceed higher education(success), we randomly selected 10% of training data 100 times and computed average of the 100 optimal cut-offs. Each optimal cut-off was where the product of 'True Positive Rate' and 'True Negative Rate' were maximum. In Graph 9, this point is where the lines of the True Positive Rate and True Negative Rate meet against the unique predicted probabilities of the logit function. The resulting threshold was 0.8880844 and the misclassification error was 2.5% when we tested the model for the new dataset. This indicated the percentage of mismatch between the predicted values and the actual values of the 40 new observations in the test dataset.

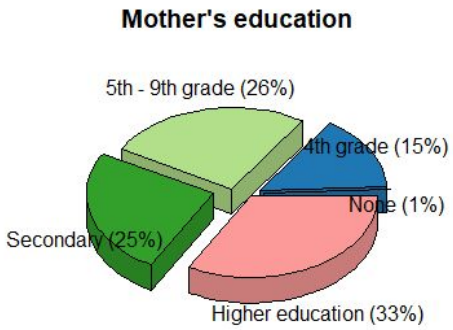


Graph 9. True Positive Rate and True Negative Rate



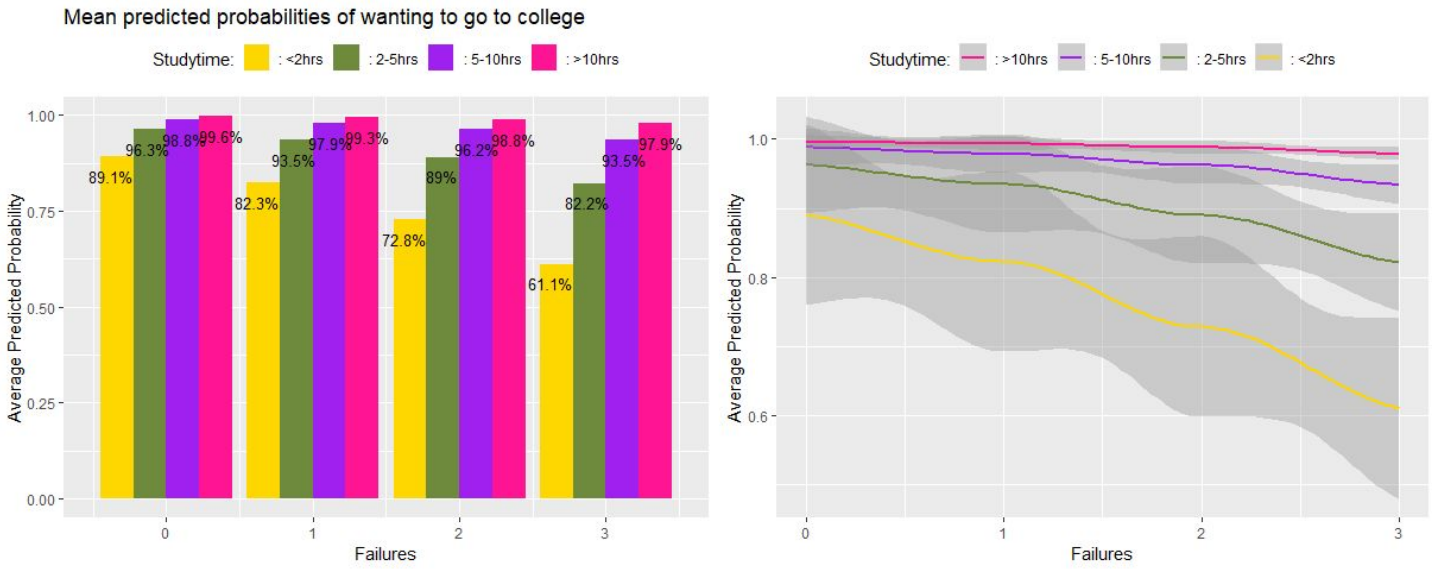
Graph 10. The mean predicted probabilities of inclining to proceed a higher education degree according to the mother's education

In Graph 10, if the Mother's education is college or higher, a student who studies less than 2 hours a week would want to go to college with 90% probability on average, but the variation is very large for those who study less than 2 hours. A student who studies more than 10 hours a week would almost 100% want to go to college, regardless of the mother's education degree.



Graph 11. Piechart of Mother's education

The pie chart in Graph 11 shows that the Mother's education is evenly spread across the four levels from 4th grade to higher education. This data supports that the positive linear relationship between the mother's education and the average probability of a student's willingness to go to college.



Graph 12. The mean predicted probabilities of inclining to proceed a higher education degree according to the number of class failures

In Graph 12, a student who studies less than 2 hours with zero class failure are more likely to want to go to college than those who study 2~5 hours with 2~3 class failures. It is noticeable that if a student studies more than 10 hours a week, the student is likely to want to go to college regardless of the number of failures. If a student studies less than 2 hours a week, the probability of wanting to go to college decreases steeply as the number of failures increases.

References

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
Available at: https://archive.ics.uci.edu/ml/datasets/student+performance

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Prabhakaran, S. (2016), " Logistic Regression", r-statistics.co, Available at http://r-statistics.co/Logistic-Regression-With-R.html