What Factors are Associated with Student Math Scores?

Team F: Regan Kelly, Rohit Thakur, & Elizabeth Trauger

December 2020

1. Introduction:

Math is a core subject of study in school that children around the world encounter at all levels of education. In most math courses, a student's score is based on their performance on homework, projects, and tests. Naturally, the scores of students vary. Our group seeks to determine the factors associated with the differences in student performance in math.

2. Objective:

The objective of this analysis is to compare mean math scores of different populations and to investigate the relationship between math score and explanatory variables. Our group wishes to identify the key variables that are associated with math scores. This analysis will examine the following factors: sex, desire for higher education, family size, weekly study time, romantic relationship status, presence of family support, presence of extra additional school support, and home address type.

3. Data Source:

To carry out the analysis, we will use a dataset from the UCI datasets repository. The data was collected during the 2005 to 2006 school year from a sample of 395 students from two public secondary schools in the Alentejo region of Portugal. The data is based on a combination of grade reports and questionnaires that collected demographic, social, and school related attributes about the students (Table 1). Students are evaluated in three periods during the school year in Portugal. Thus, each student in the dataset has three scores for each period that range from 0 to 20. (Portugal uses a 20-point grading scale, where 0 is the lowest score and 20 is a perfect score.) In our analysis, we will use the average of the three scores to represent each student's math score. This dataset does not represent a random sample of all public secondary students in the world, and thus all conclusions resulting from the analysis will be applied to the target population of public secondary students in Portugal.

Table 1: Variables and their descriptions from the dataset that we will use in our analysis

Attribute	**Description**
sex	Student's sex (binary: F or M)
famsize	family size (binary: ≤ 3 or > 3)
studytime	weekly study time (numeric: 1 (< 2 hours), 2 (2 -5 hours), 3 (5 to 10 hours) or 4 (> 10 hours))
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
schoolsup	extra educational school support (binary: yes or no)
address	student's home address type (binary: urban or rural)
score	average grade from 3 periods (numeric: from 0 to 20)

famsup	family educational support (binary: yes or no)
Medu	mother's education (numeric: from 0 to 4[a])
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)

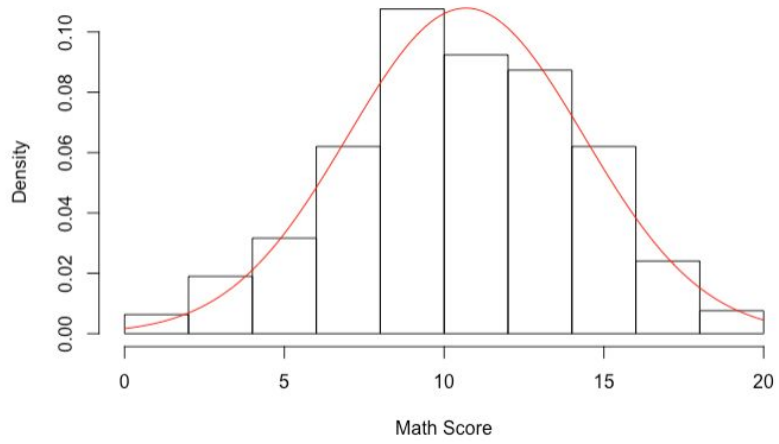
**a* : 0- none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.

4. Data Analysis:

I. Data Assumptions:

From Figure 1, we see that the distribution of the sample population's math scores is approximately normal. Thus, we will assume normality in our analysis.

Figure 1: Histogram of Students' Average Math Score



In addition to normality, we assume the following:

- Data within one group represents a random sample from a population
- The populations are independent
- Within the i-th group, observations are normally distributed with mean μ_i
- Variance is the same for all groups

As the data satisfies these assumptions, we will use analysis of variance (ANOVA) to assess the effect of considered factors on the students' math scores. For each variable, we will test for an overall difference in population means. If there is a difference among the populations, we will pinpoint the difference.

II. Analysis of Variance (ANOVA) Analysis:

Table 2: Correlation between math score and factors under study

Source of Variable	Null Hypothesis	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Squares (MS)	F-value	P-value
sex	H_0: $\mu_f = \mu_m$	1	55	55.06	4.06	0.0446
Residuals		393	5329	13.56		
higher	H_0: $\mu_{yes} = \mu_{no}$	1	193	193.32	14.64	0.000152
Residuals		393	5191	13.21		
famsize	H_0: $\mu_{\leq 3} = \mu_{>3}$	1	37	36.7	2.697	0.101
Residuals		393	5348	13.61		
address	H_0: $\mu_{rural} = \mu_{urban}$	1	62.0	62.0	4.58	0.033
Residuals		393	5322.5	13.5		
romantic	H_0: $\mu_{yes} = \mu_{no}$	1	56.8	56.8	4.19	0.041
Residuals		393	5327.7	13.6		
study time	H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$	3	121.2	40.4	3.00	0.0305
Residuals		391	5263.3	13.46		
famsup	H_0: $\mu_{yes} = \mu_{no}$	1	20	20.4	1.495	0.222
Residuals		393	5364	13.65		
schoolsup	H_0: $\mu_{yes} = \mu_{no}$	1	102	102.01	7.59	0.00614
Residuals		393	5282	13.44		

As seen in the table above, the null hypothesis is rejected at the 0.05 level of significance for the following sources of variable: sex, higher, address, romantic status, weekly study time, and schoolsup. We fail to reject the null hypothesis at the 0.05 significance level for the following sources of variable: famsize and famsup.

Therefore, we can make the following conclusions:

- There is evidence that there is a difference in the mean math scores for the female and male student populations. Based on summary statistics, males have a greater sample mean math score (11.07) compared to female students (10.33).

- There is evidence that there is a difference in mean math scores for students who want a higher education compared to students who do not want higher education. Students who want a higher education have a larger sample mean math score (10.84) compared to those who do not want higher education (7.65).

- There is evidence that there is a difference in mean math scores for the rural and urban student populations. The urban students have a greater sample mean math score (10.89) compared to the rural students (9.94).

- There is evidence that there is a difference in mean math scores for students who are in a romantic relationship compared to students who are not. The students in a relationship have a lower mean math score (10.14) compared to the other students (10.95).

- There is evidence that there is a difference in mean math scores for students in some of the different study time groups.

- There is evidence that there is a difference in mean math scores for students who have extra additional school support compared to students who do not. The students with extra additional school support have a lower mean math score (9.36) compared to the students without support (10.88).

- There is not sufficient evidence to say the mean score of students who have families size less than 3 people is different from the mean scores of students with family size greater than 3 people.

- There is not sufficient evidence to say the mean score of students with family educational support is different from the mean scores of students without family educational support.

III. Pairwise Comparisons for Study Time Variable:

Since the null hypothesis is rejected at the 0.05 significance level for the variable studytime, we will use a pairwise comparison between the different groups to determine which groups had different math scores. For this analysis, Tukey's Honestly Significant Difference (HSD) multiple testing adjustment was used to conduct the post-hoc pairwise comparisons to control the family-wise error rate.

Pairwise t-tests test the null hypothesis H_0: $\mu_1 = \mu_2$ for all pairs of study time groups. The results of the pairwise comparison with multiple testing adjustment indicate that there is not sufficient evidence that the math scores are different between any of the pairs of study time groups at the 0.05 level of significance. At the 0.1 level of significance, there is evidence that the mean math score of students that study < 2 hours per week is statistically different from the mean math score of students that study 5-10 hours per week. There is also evidence at the 0.1 significance level that the math score of students that study 2-5 hours per week is statistically different from the mean math score of students that study 5-10 hours per week.

Table 3: Pairwise Comparison with Tukey HSD Adjustment for Study Time Variable

Pairwise Comparison Groups	Estimated Difference	Confidence Interval for Difference in Means		Adjusted P-value
		Lower	Upper	
2-1	0.189	-0.954	1.332	0.974
3-1	1.397	-0.097	2.891	0.076
4-1	1.474	-0.568	3.517	0.246
3-2	1.209	-0.144	2.562	0.099
4-2	1.286	-0.656	3.228	0.321
4-3	0.077	-2.09	2.244	0.9997

IV. Regression Analysis:

In order to investigate the relationship between math score and explanatory variables, we used a multiple linear regression model. A multiple linear regression model helps us identify which variables are useful in predicting the math score (response variable) of a student. The model also helps us understand how the math score of a student changes when one of the explanatory variables changes, while the other explanatory variables are held constant.

For choosing the predictor variables, stepwise model selection was performed based on AIC score. For this study, the independent variables are student's sex (x_1), mother's education (x_2), weekly study time (x_3), number of past failures (x_4), extra educational school support (x_5) family educational support (x_6), and weekly hours spent with friends (x_7).

The model is specified as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$

Table 4: Multiple Linear Regression Coefficient with Score as Response Variable

Variable	Estimate β_j	Std. Error	t-value	Pr (> |t|)
Intercept	10.284	0.817	12.59	< 2e-16
sex (x_1)	0.873	0.356	2.45	0.0147
Medu (x_2)	0.525	0.160	3.27	0.001
studytime (x_3)	0.512	0.212	2.417	0.016
failures (x_4)	-1.555	0.235	-6.626	1.16e-10
schoolsup (x_5)	-1.268	0.501	-2.532	0.012
famsup (x_6)	-0.730	0.354	-2.065	0.039
goout (x_7)	-0.441	0.151	-2.920	0.0037

For each β_j, we test the null hypothesis H_0: $\beta_j = 0$ against H_A: $\beta_j \neq 0$ at the 0.05 level of significance. As seen in Table 4, all of the p-values are less than 0.05. Thus, we reject the null hypothesis and conclude that all of the variables are statistically significant in the model.

From Table 4, we can see that the regression model is
$y = 10.284 + 0.873x_1 + 0.525x_2 + 0.512x_3 - 1.555x_4 - 1.268x_5 - 0.73x_6 - 0.441x_7$

The constant value (10.284) is the intercept which represents the math score of a student given that all the predictors are zero. The variables with a positive coefficient value (x_1, x_2, x_3) indicate how much the math score of a student would increase per unit change in x_1, x_2 and x_3. Similarly, the variables with a negative coefficient value (x_4, x_5, x_6, x_7) indicate how much the math score of a student would decrease per unit change in x_4, x_5, x_6, and x_7.

Table 5: Multiple Correlation and Coefficient of Determination

Model Summary		
Multiple R-squared	Adjusted R-squared	p-value
0.2223	0.2082	< 2.2e-16

The p-value of 2.2e-16 shows that there exists a strong positive relationship between math score as dependent variable and the independent variables mentioned above. This implies that there is a relationship between behavioural patterns/family background and math score. The adjusted R^2 value is 0.2082 which indicates that 20.82% of the variations in math scores of the students is explained by the predictor variables.

Furthermore, the residuals were plotted against the fitted values in Figure 2 to verify model assumptions. The residual plot showed random scatter which verifies that our linear model is appropriate for the data. The scatterplot shows that there is no evidence that the assumption of homoscedasticity has been violated. Also, the QQ plot in Figure 3 showed normal quantiles thus verifying the normality assumption.

Figure 2: Plot of Residuals vs Fitted Values

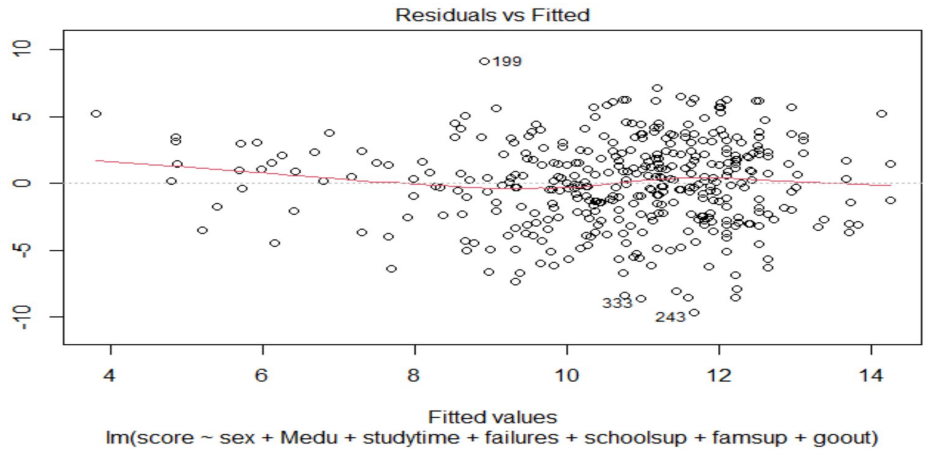
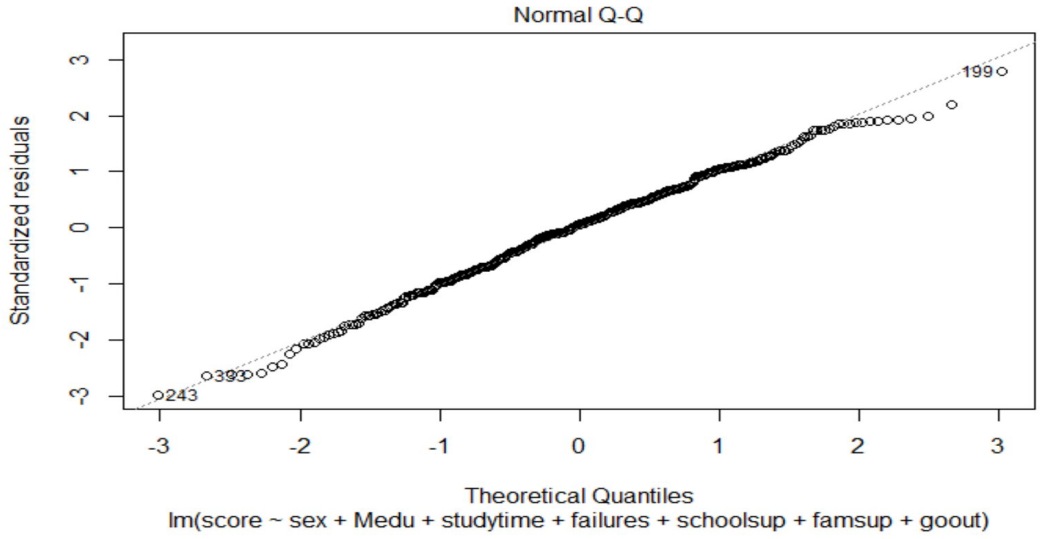


Figure 3: QQ Plot



5. Conclusion:

Based on the ANOVA tests, we can conclude that the mean math scores for public secondary school students in Portugal differ for the subgroups in the following variables: sex, higher, address, romantic status, weekly study time, and schoolsup. Furthermore, the multiple linear regression model shows that behavioural qualities, as well as by family background, including a student's sex, mother's education level, weekly study time, number of past failures, extra educational school support, family educational support, and weekly hours spent with friends all influence a student's math score. Based on the R^2 value (0.2082), 20.82% of variance in the target variable is explained by our model. This means around 79% of changes in average math score are not accounted for by our explanatory variables.

The sample may not be a true random sample, as the specific details of the collected data including student and school selection are not known. For this reason, the validity of the conclusions presented in this paper are dependent on the assumption that the sample represents a random sample from the population.

Statement of Contributions:

The work of this project was evenly distributed among all three team members.
Each member has contributed equally in brainstorming the ideas for the project and analysis.
Following are the specific contributions
Elizabeth Trauger: Tested for normality, performed ANOVA tests on 5 of the variables and interpreted the results of those tests.
Regan Kelly: Performed ANOVA tests on 3 of the variables, did the pairwise comparisons for the study time variable, and interpreted those results.
Rohit Thakur: Performed exploratory data analysis through visualizations, regression analysis in continuation with ANOVA tests and interpretation of results.
All members have contributed equally in documenting the report.

Appendix:

Data: https://www.kaggle.com/janiobachmann/math-students

Citations:
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.
http://www3.dsi.uminho.pt/pcortez/student.pdf

Project repository on GitHub:
https://github.com/thakur-ro/statistical-data-analyis