Middle East Technical University

Department of Statistics

STAT 250
TERM PROJECT
Academic and Socioeconomic Factors Influencing Student Success

June 2025

Submitted to Prof. Dr. Burçak Başbuğ Erkan

Yusuf Özcan 2614782

Dilara Yıldırım 2614907

Kemal Can Yoloğlu 2614915

Table of Contents

Aim & Objectives

This research includes academic and personal information for 382 high school students with two different lessons: Mathematics and Portuguese language courses. Also this dataset has many variables such as gender, parental educational levels, study time, age, alcohol consumption, relationship status and so on. Also it includes three main grades: G1, G2, G3 (First period grade, second period grade. final grade) for both lessons.

For this research we are aiming to find academic and socioeconomic factors that influence student success. R-studio was employed to perform all the analysis and generate graphs and figures in this report. Firstly, in order to learn more about the overall structure of the data, exploratory data analysis was done with basic graphs and figures to visualize assumptions. To further analyze and form a conclusion about these relationships, some analysis techniques were used.

For our aim for that report we created new variables called FinalGrade_Math and FinalGrade_Lang. These variables are created to represent 3 grades to weighted one grade:

$$\text{FinalGrade\_Math} = 0.3xG1 + 0.3xG2 + 0.4xG3$$

$$\text{FinalGrade\_Lang} = 0.3xG1 + 0.3xG2 + 0.4xG3$$

Introduction

The dataset used in this study provides an inclusive overview of students' academic performance as well as family and social backgrounds. It allows to do comparison analysis of the factors affecting student success across two different lessons with their grade data. Moreover, the presence of various variables such as parental education, personal habits, and socioeconomic status provides significant advantages in terms of analysis flexibility.

In this context, the priority is to identify the key factors influencing student performance and to statistically determine the scope to which these factors are significant. This type of study is expected to offer worthy insights for teachers, education policymakers, and guidance professionals in making more informed decisions. In addition, this research can also provide help to raising individual awareness among students and of course their families.

Although our findings are based on this particular dataset, they have the potential to extend the results more generally for other student groups with similar age ranges and similar educational systems. Hence, it is expected that the outcomes of this study may contribute to a useful foundation for comprehensive research in the future.

Before starting our research findings, it is crucial to understand the dataset clearly. Below table has variables in the dataset which are also used in research questions.

variable name	description	scale
Romantic	In a romantic relationship	binary
Studytime	Weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)	numeric
Walc	Weekend alcohol consumption (1-5)	numeric
Higher	Intention to pursue higher education	binary
FinalGrade_Lang	Weighted final language grade	numeric
FinalGrade_Math	Weighted final math grade	numeric
Schoolsup	Extra educational support	binary
Famsup	Family educational support	binary
Paid	Paid subject-related classes	binary
Famrel	Quality of family relationships (1-5)	numeric
Pstatus	Parental cohabitation status	binary
Famsize	Family size (≤3 or >3)	binary
Guardian	Legal guardian of the student	nominal
Medu, Fedu	Mother's and father's education level	numeric
Mjob, Fjob	Parent occupation	nominal

For these variables, first of all we have looked at the dataset whether it is clean or not. Finally, we have done some tests to check that issue and we reached that the dataset is clean. So let's move on to the Exploratory Data Analysis section.

Exploratory Data Analysis

After deciding the main goal of that research and being sure the dataset is clean, we wanted to look at several research questions by achieving our goal. Questions will be given in every individual step but before passing through to questions we wanted to explain descriptive statistics of numerical variables that we form the dataset. Notice that the variables written with italic font state that these variables are from that lessons' variables and data belong to that lesson as well.

Variable	**Mean**	**StdDev**	**Min**	**Q1**	**Median**	**Q3**	**Max**
age	16.59	1.17	15	16	17	17	22
Medu	2.81	1.09	0	2	3	4	4
Fedu	2.57	1.10	0	2	3	4	4
traveltime	1.44	0.70	1	1	1	2	4
studytime	2.03	0.85	1	1	2	2	4
failures	0.29	0.73	0	0	0	0	3
famrel	3.94	0.92	1	4	4	5	5
freetime	3.22	0.99	1	3	3	4	5
goout	3.11	1.13	1	2	3	4	5
Dalc	1.47	0.89	1	1	1	2	5
Walc	2.28	1.28	1	1	2	3	5
health	3.58	1.40	1	3	4	5	5
absences	5.32	7.63	0	0	3	8	75
Final Grade for Math Lesson	10.63	3.84	1.2	8.3	10.6	13.36	19.4

Figure 1: Descriptive Statistics for Mathematics Lesson

By looking at Figure 1 above; for mathematics lessons, the average age variable is 16.6 years, with a maximum of 22 and minimum of 15. The parental education levels are moderately high, with average values of 2.57 for fathers and 2.80 for mothers on a 0-4 scale, where 4 represents higher education.

Study-related variables: For studytime variable, it has 2.03 mean on 1-4 scale means that most of students are studying weekly 2-5 hours, for failures variable on a 0-3 scale most of them have not

failed any courses before, then for travel time variable on a 1-4 scale, mean is 1.44 which can be considered as much of students have 15-30 minutes travel time.

Social behavior indicators such as goout variable and alcohol consumption indicates moderate levels of socialization.

Health variable seems to be good such that 3.58 means on a 1-5 scale.

Absences are very low on the 0-75 scale which means that students are mostly attending classes.

Finally, final grades seem to average 10.63 on 0-20 scale, which means that there are many kinds of students for success.

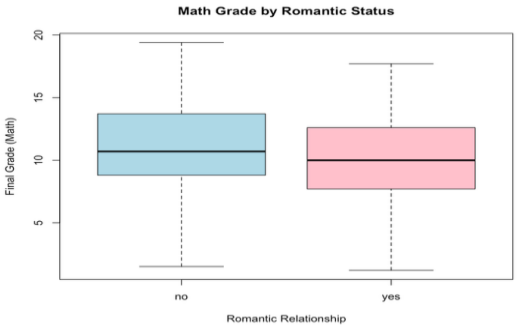
Variable	**Mean**	**StdDev**	**Min**	**Q1**	**Median**	**Q3**	**Max**
age	16.59	1.17	15	16	17	17	22
Medu	2.81	1.09	0	2	3	4	4
Fedu	2.57	1.10	0	2	3	4	4
traveltime	1.44	0.70	1	1	1	2	4
studytime	2.04	0.85	1	1	2	2	4
failures	0.14	0.51	0	0	0	0	3
famrel	3.94	0.91	1	4	4	5	5
freetime	3.23	0.99	1	3	3	4	5
goout	3.12	1.13	1	2	3	4	5
Dalc	1.48	0.89	1	1	1	2	5
Walc	2.29	1.28	1	1	2	3	5
health	3.58	1.40	1	3	4	5	5
absences	3.67	4.91	0	0	2	6	32
Final Grade for Lang Lesson	12.31	2.56	3.6	10.7	12.3	14	18.7

Figure 2: Descriptive Statistics for Language Lesson

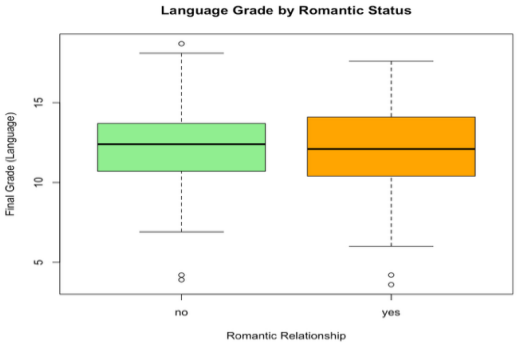
For language stats we can say similar things as we said in the Math grades but differently final grades for language lessons seem to be slightly above average (12.31) on 0-20 scale, which means students are more successful in language lessons than math lessons which is reasonable.

1) Does being in a romantic relationship affect academic performance?

Before going deeply to analysis we wanted to look at the boxplot graphs of each lessons.



Math Grades: Median value of being in a romantic relationship seems to be lower than students who are not being. Moreover, the width of boxes seems to be equal which means that variances seem to be equal but to be sure we need to test it. Additionally, there is no outlier.



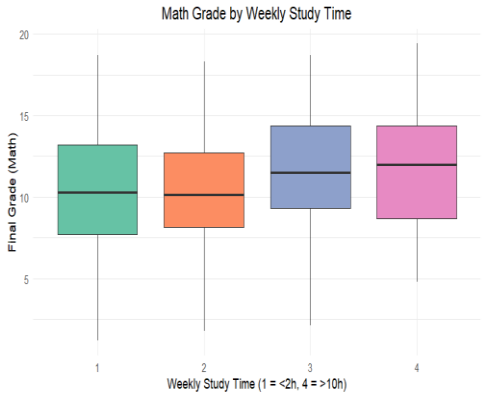
Language Grades: Median values seem to be equal but width of boxes are not equal. Which means that students who are in a relationship have more variance than who are not. Moreover, there are 3 outliers for no answers and 2 outliers for yes answers.

Before applying any kind of hypothesis tests, we need to check the normality of given data. After conducting normality by using the Shapiro-Wilk test, we see that the final grade distributions for both subjects were not normally distributed in at least one group. Afterwards, we applied a non-parametric test which is Mann-Whitney U test, and we see that romantic relationship status did not significantly affect language performance and had a marginally non-significant effect on Math performance.

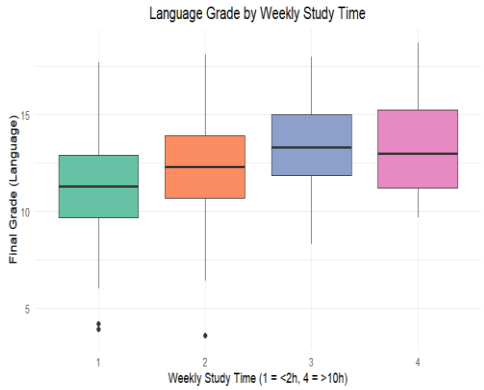
Although our analysis did not show any significance of being in a romantic relationship or not for grades, remember that this data is taken from high school students and there might be a non-significance effect for these age groups but on the other hand, there are many studies related with university students and in that type of studies, being romantic relationship effects grades both positively and negatively according to type of these relationships and of course psychology. If you go forward you can search deeply but our study needs to pass through to 2nd question.

2) Does weekly study time significantly affect academic performance?

Before diving into statistical analysis, we inspected the boxplot for grades by weekly study time.



Math Grades: Students who studied more (levels 3 and 4) had visibly higher median grades. The box widths appear similar, indicating similar variances, and there were no significant outliers.



Language Grades: Median values also increased slightly with more study time. However, variance seemed less consistent across categories. Some groups had wider boxes, but again, there were no significant outliers.

These boxplots suggest that study time may influence grades, but statistical tests should be done to confirm this.

We conducted a One-Way ANOVA test since the assumptions which are given below are satisfied.

1. **Independence**: Each row in the dataset represents a unique student, and no students are measured repeatedly. So, the independence assumption is met.
2. **Normality**: The final grade distributions within each *studytime* group for both Math and Language did not strongly violate normality (since p-values of all groups > 0.05).
3. **Homogeneity of Variances**: By conducting Levene's test it can be seen that there is no significant difference in variances across groups for both Math and Language (Math p = 0.21, Language p = 0.32).

	df	sumsq	meansq	F value	p.value
studytime	3	112,9143	37,63809	2,581526	0,053195
Residuals	378	5511,158	14,57978		

For the Math grades, the ANOVA gave a p value of 0.053. Since this is slightly above 0.05, it suggests that weekly study time might have some impact on math performance, but the evidence is not strong enough to confidently state so.

	df	sumsq	meansq	F value	p.value
studytime	3	248,8437	82,9479	13,90703	1,28E-08
Residuals	378	2254,566	5,964459		

For language grades, the ANOVA gave a very low p value which is close to zero. This result suggests that there is clear and strong evidence to state that how much a student studies weekly significantly affects their performance in language in a positive way.

The differing effects of study time on Math and Language may be due to:

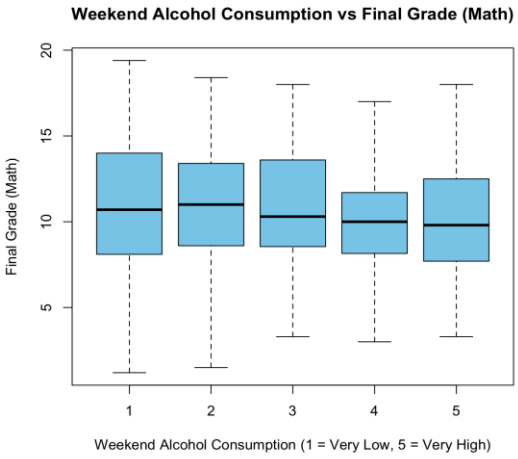
1- Language skills generally improve more directly with consistent study, which makes study time a convenient predictor of success in language.

2- Success in math can be influenced by other factors such as problem-solving skills or test difficulty, which cannot be captured by study time alone.

3- Students might use different strategies for studying math, such as focusing on problem solving or taking private lessons, instead of simply spending more hours studying.

3) Does weekend alcohol consumption have a significant impact on academic performance?

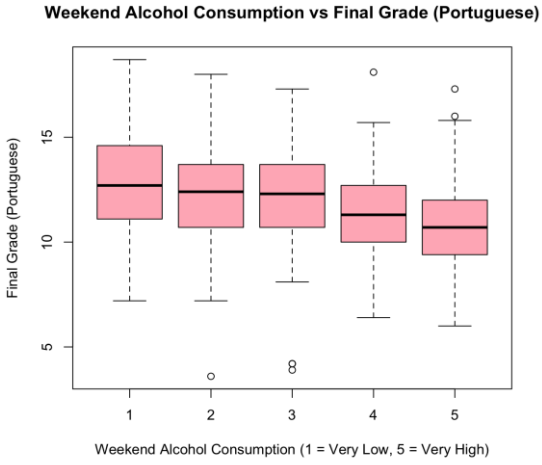
Before proceeding to the in-depth statistical analysis, we examined the box plots showing the relationship between weekend alcohol consumption and academic performance.



Math Grades: The median values are very close across all groups and there is no significant change in academic performance as weekend alcohol consumption increases. The spread of the data is also similar across groups.

There are no statistical outliers in the graph, indicating that the grade distribution is balanced and free of outliers.

Overall, this visualization does not show a significant relationship between alcohol consumption and academic achievement in mathematics.



Language Grades: There is a clear decrease in academic achievement as weekend alcohol consumption increases. Median values decrease gradually with increasing weekend alcohol consumption and the grade distribution become more compressed in the higher alcohol consumption groups.

Moreover, in some groups, statistical outliers are observed.

This suggests that weekend alcohol consumption has a negative impact on academic achievement in Portuguese and some students are more affected.

One-Way ANOVA test assumptions:

1. **Independence** is satisfied for both groups. Because each observation corresponds to a different student.
2. **Normality**: For the Math course, the normality assumption was not satisfied for the group with Walc = 2 ($p = 0.01082$).For the Portuguese course, the normality assumption was not satisfied for the group with Walc = 3 ($p = 0.002678$).
3. **Homogeneity of Variances**: For the Portuguese course, Levene's Test showed that the assumption of equal variances was satisfied ($p = 0.7743$).But, in the Math course Levene's Test showed that the assumption of equal variances was not satisfied ($p = 0.009831$).

Because the assumptions were not fully satisfied, Kruskal-Wallis was a better choice than ANOVA for this question.

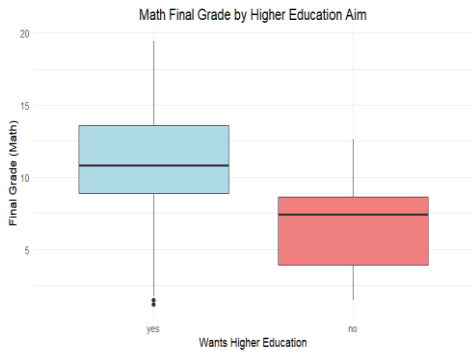
	df	chi-squared	p.value
Final Grade Math	4	3.585	0.4651

	df	chi-squared	p.value
Final Grade Lang	4	21.824	0.0002172

When we examined the results of Kruskal-Wallis test, for the Portuguese course, the result was statistically significant ($\chi^2(4) = 21.824$, $p < 0.0002172$), indicating that students' final grades differ across alcohol consumption levels. This suggests that higher weekend alcohol consumption is associated with lower academic performance in Portuguese.

For the Math course, no significant difference was found between the groups ($\chi^2(4) = 3.585$, $p = 0.4651$), suggesting that alcohol consumption does not have a notable effect on students' performance in Mathematics.

4) Do students who want to pursue higher education perform better academically than those who do not?



Math Grades: The "yes" group has a wider spread of grades, which shows there is more variability in performance. However, the "no" group has a smaller range. Students who want to pursue higher education have a higher median for final Math grades compared to those who do not want to. So, the boxplot shows that students who aim to pursue higher education tend to perform better in Math compared to students who do not.

Then we checked the assumptions for independent samples t-test

1. Normality: The Shapiro-Wilk test was conducted for both groups. The p- value of the "yes" group was 0.02, showing that this group deviates from normality. The "no" group had a p-value of 0.41, which indicates normality. Despite the issue of normality in the "yes" group

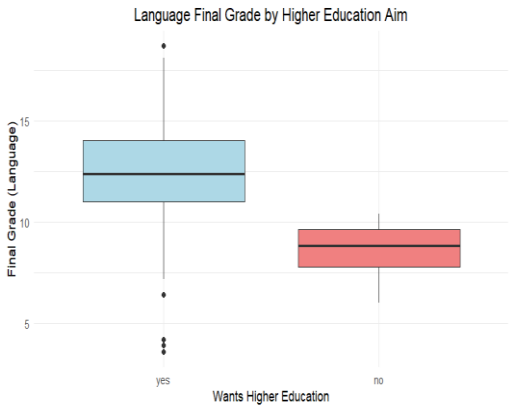
we can still use the t-test because of the central limit theorem (since the sample size of the "yes" group is large.

2. Equal Variances: Levene's test showed no significant difference in variances (p-value = 0.28). So, the assumption of equal variances is valid.

Since these assumptions are met, we proceed with a standard independent t-test to compare the Math grades of students who want to pursue higher education and those who do not.

Statistic	t-statistic	df	p-value
Value	4,80182	380	2,27E-06

The t-statistic is 4.80 and the p-value is close to 0, which is highly significant since $p < 0.05$. This result shows that there is a statistically significant difference in the Math grades between the groups. So students who aim for higher education perform better compared to other students that do not based on their grades.



Language Grades: In the boxplot it can be seen that students who want a higher education to have a higher median Language grades compared to those who do not want. The "yes" group has a greater variability in grades, with a wider interquartile range and longer whiskers. Moreover, there are some outliers in the "yes" group, which shows that there are some exceptional grades. On the other hand, students who do not want higher education to have lower grades with less variability. This graph supports that students aiming for higher education tend to perform better in Language.

Then we checked the assumptions for independent samples t-test

1. **Normality**: The Shapiro-Wilk test was conducted for both groups. For the students who aim for higher education the p-value was 0.0007, which shows that it deviates from normality. For the students who do not aim for higher education the p-value was 0.31, which shows that it is approximately normal.
2. **Equal Variances**: Levene's test revealed that there is a significant difference in variances between the groups since the p-value is 0.03.

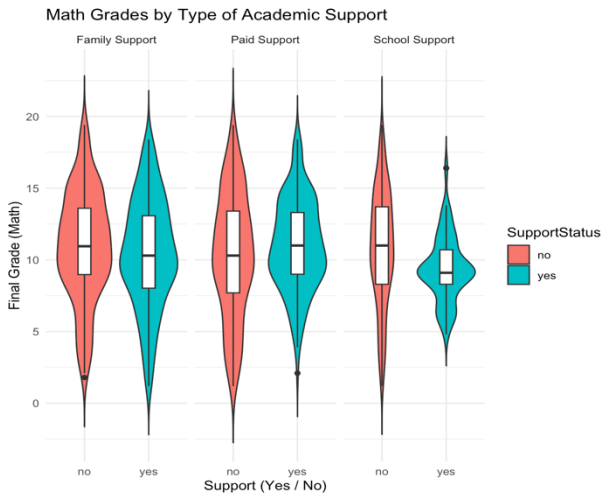
These results do not satisfy the assumptions required for a standard t-test so due to the violated assumption of equal variances we have conducted Welch's t-test to compare Language performance between groups.

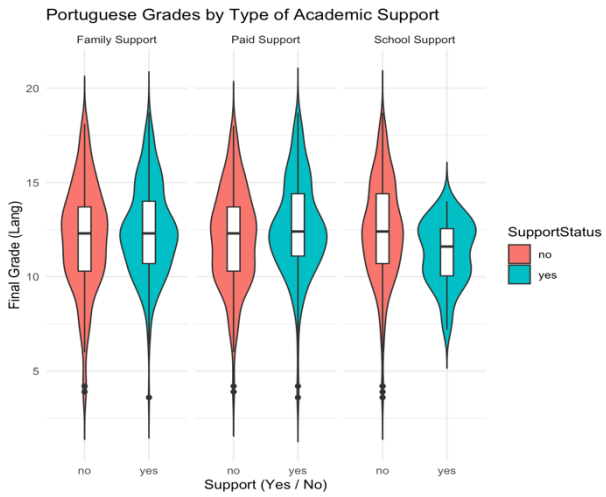
Statistic	t-statistic	df	p-value
Value	11,20059	23,12693	8,08E-11

Welch's t-test gave a very small p-value which shows that there is a statistically significant difference between the two groups. Therefore, we can conclude that students who aim for higher education have significantly higher language grades than students who do not aim for higher education.

As a conclusion, students who plan to take higher education perform significantly better in both Math and Language compared to those who do not. This difference may be because students with aims for higher education are generally more motivated and focused on their studies. These students can also give more time and effort to learn. These factors can lead to better grades and academic performance.

5) Does receiving any kind of academic support (school, family, or paid) affect students' academic performance?





The violin plots show that the Math and Language grades of students receiving school, family or paid academic support have similar distributions, with minor differences in the means. The mean grade of students who receive school support is less than others in both courses. Moreover, there are a few outliers in some groups. The results of detailed statistical analysis will help confirm these observations. Since the distributions of the Math group are more different in the violin plots, further analysis was conducted specifically for this group.

We examined the impact of school, family, and paid academic support on students' Math performance using a multiple linear regression model.
Firstly, a multiple linear regression model was built including all three types of academic support as predictors.

Variable	Estimate	Std. Error	p-values
(Intercept)	10.8396	0.3433	<2e-16
***schoolsup* (yes)**	-1.3366	574	0.0204
***famsup* (yes)**	-0.6783	418	0.1055
***paid* (yes)**	0.8382	0.4049	0.0391

Then, as seen in the table, a new model was built using only the significant variables; school support and paid classes as seen in below:

$$Y_{\{\{Math\}\}} = \beta_0 + \beta_1 \cdot \{schoolsup(yes)\} + \beta_2 \cdot \{paid(yes)\}$$

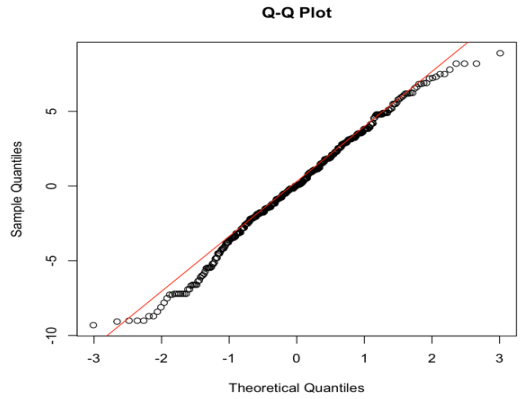
Variable	Estimate	Std. Error	p-values
Intercept	10.5111	0.2779	<2e-16
***schoolsup* (yes)**	-1.4233	0.5727	0.0134
***paid* (yes)**	0.6602	0.3906	0.0919

Residual Std. Error	Adj. R-squared	p-value
3.806	0.01867	0.01037

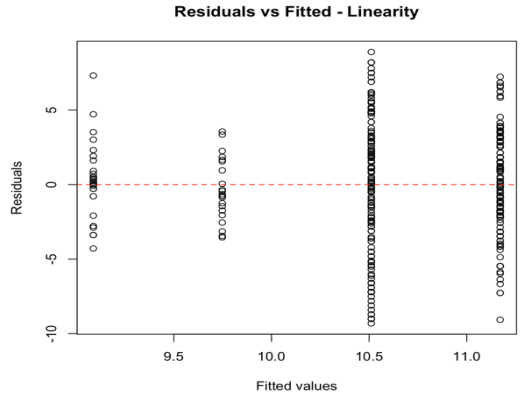
$$\boldsymbol{Y_{\{\{Math\}\}} = 10.5111 - 1.4233 \cdot \{schoolsup(yes)\} + 0.6602 \cdot \{paid(yes)\}}$$

According to this regression model. School support [*schoolsup* (yes)] has a significant negative effect on final math grades (β1 = -1.42, p = 0.013). This suggests that students receiving school support tend to score, on average, 1.42 points lower than those who do not. This may indicate that students who need extra support already struggle academically. Paid classes [*paid* (yes)] have a positive but marginally insignificant effect (β2 = 0.66, p = 0.092). Because it is not statistically significant at the 0.05 level. The low R^2 value (0.0238) indicates that the model explains only a small portion of the variation in math grades. Therefore, other factors may be more influential.

To ensure the validity of the model, assumptions were checked as follows:



Q-Q Plot shows that most of the residuals are close to the red line. This indicates that the residuals are approximately normally distributed, and the normality assumption is reasonably satisfied.



The residuals appear to be randomly scattered around the horizontal axis without any clear pattern. Therefore, the assumption of linearity is reasonably satisfied. There is no strong evidence of heteroscedasticity, meaning the variance of residuals appears to be roughly constant (homoscedasticity is satisfied).

To check for multicollinearity, the VIF values were examined. Both predictors had VIF values of approximately 1.00, which is less than 5. Therefore, there is no multicollinearity issue among the predictors in this model.

The Durbin-Watson test was conducted. The test result was DW = 1.868, p = 0.097, indicating that there is no significant autocorrelation. Thus, the assumption of independence of errors is reasonably satisfied.

Since assumptions of multiple linear regression were reasonably satisfied, the model can be considered valid.

6) Does the student's family background affect academic performance?

For this question we have several family-related variables such as: *famrel*, *pstatus*, *famsize*, *guardian*, *Medu*, *Fedu*, *Mjob* and *Fjob*. In order to understand which of these variables are meaningful predictors of academic performance, we constructed multiple linear regression models using students' final Math grades as the response variable.

Below is a comparison of three regression models for **Math grades**:

Model ID	Fitted Equation	Significant Variable(s)	Adjusted R^2	Model p-value
Model 1	Math – full	*famsize*	0.066	0.0004085
Model 2	Math – simplified	*famsize, Medu*	0.059	2.08e-05
Model 3	Math – log	*famsize, Medu*	0.053	3.27e-05

Model 1:

$$Y_{\{Math\}} = \beta_0 + \beta_1 \cdot famrel + \beta_2 \cdot Pstatus + \beta_3 \cdot famsize + \beta_4 \cdot guardian + \beta_5 \cdot Medu + \beta_6 \cdot Fedu + \beta_7 \cdot Mjob + \beta_8 \cdot Fjob$$

First of all, All p-values are below 0.05, which means all three models are statistically significant. In Model 1, which includes all family-related variables, the only significant variable is *famsize* with an estimated coefficient of β_3 = +0.98. Although this model has the highest adjusted R^2, we decided to conduct a simpler model since most variables were not significant.

Model 2:

$$Y_{\{Math\}} = \beta_0 + \beta_1 \cdot famsize + \beta_2 \cdot Medu + \beta_3 \cdot Fedu + \beta_4 \cdot famrel$$

In Model 2, we removed the non-significant variables and built a simplified model including only *famsize, Medu, Fedu* and *famrel.* This time, both *famsize* and *Medu* appeared as significant predictors, with estimated coefficients of +1.02 and +0.63, respectively. The adjusted R^2 value slightly decreased but remained acceptable. Therefore, this model indicates a good balance between simplicity and explanatory power.

Model 3:

$$log(Y_{\{Math\}} + 1) = \beta_0 + \beta_1 \cdot Medu + \beta_2 \cdot famsize + \beta_3 \cdot Fedu$$

To further improve model assumptions, we applied a logarithmic transformation to the response variable. In Model 3, again *famsize* and *Medu* were significant with estimated coefficients of +0.11 and +0.06, respectively. However, removing *Fedu* caused a drop in adjusted R^2, so we kept it even though it was not significant.

We choose Model 2 as the final model for interpretation since it includes only two significant variables and achieves the highest explanatory power without transformation. The fitted equation is:

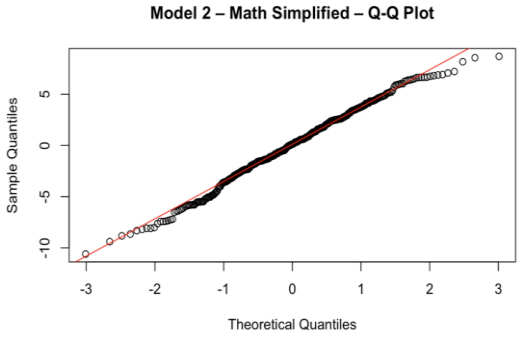
$$\boldsymbol{Y_{\{Math\}} = 7.29 + 1.02 \cdot famsize + 0.63 \cdot Medu}$$

This model explains approximately 5.9% of the variation in Math grades. As *famsize* increases by 1 unit, the grade increases by 1.02 points as well. Similarly, for every 1 unit increase in *Medu*, the grade increases by 0.63 points. The intercept value β_0 = 7.29 represents the expected

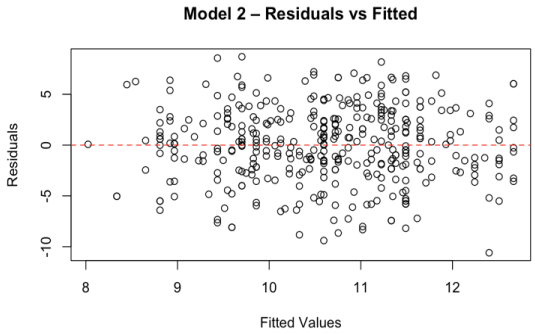
Math grade when all predictors are at their reference levels. Now let's move on to assumption checking for Model 2 but before that let's remember briefly the assumptions of Multiple Linear Regression:

1. **Linearity,** between predictors and the response variable.
2. **Normality,** of residuals
3. **Homoscedasticity,** (i.e. constant variance of residuals)
4. **No multicollinearity,** among predictors
5. **Independence** of errors

We visually checked the normality and homoscedasticity assumptions using QQ plots and Residuals vs Fitted plots and you can see at below:



This QQ plot shows that most of the residual values are positioned to the neighborhood of the red line and only maximum values are deviated from the line. This situation indicates that most of the residuals are distributed as normal and normality assumption is provided at a reasonable level.



This plot shows that the residuals appear to be scattered randomly around the horizontal axis without forming any clear pattern. This indicates that the assumption of constant variance (homoscedasticity) is reasonably satisfied, and there is no evidence of heteroscedasticity in the model.

Variables	**VIF values**
famsize	1.00
Medu	1.73
Fedu	1.73
famrel	1.00

For multicollinearity we used a VIF test and all results are less than 5, we can say that there is no multicollinearity among predictors.

Although we did not directly visualize the **linearity assumption** with individual scatterplots, the absence of curved or systematic patterns in the Residuals vs Fitted plot suggests that the relationship between the predictors and the response variable is reasonably linear. Therefore, the linearity assumption appears to be satisfied. On the other hand, **the independence of errors** is assumed based on the data structure, as the observations represent different students and are not repeated or time-based. Since the dataset does not include time-series or grouped data, we consider the independence assumption to be reasonably satisfied.

Therefore, the Model 2 is a statistically significant model with best estimators for predicting Math grades.

Below is a comparison of three regression models for **Language grades**:

Model ID	Fitted Equation	Significant Variable(s)	Adjusted R^2	Model p-value
Model 1	Lang – full	*none*	0.036	0.0174
Model 2	Lang – simplified	*Medu*	0.045	0.000255
Model 3	Lang – log	*Medu*	0.046	0.0001225

Model 1:

$$Y_{\{Lang\}} = \beta_0 + \beta_1 \cdot famrel + \beta_2 \cdot Pstatus + \beta_3 \cdot famsize + \beta_4 \cdot guardian + \beta_5 \cdot Medu + \beta_6 \cdot Fedu + \beta_7 \cdot Mjob + \beta_8 \cdot Fjob$$

Firstly, again all models are statistically significant models since all p-values are less than 0.05. In Model 1, which includes all family related variables, there is no significant variable and its adjusted R^2 is lowest one. So we decided to conduct a simpler model which is,

Model 2:

$$Y_{\{Lang\}} = \beta_0 + \beta_1 \cdot famsize + \beta_2 \cdot Medu + \beta_3 \cdot Fedu + \beta_4 \cdot famrel$$

In this model, we have decided to take only near values of 0.05 which are *famsize*, *Medu*, *Fedu* and *famrel*. This time we have a new significant variable which is *Medu* and its estimated coefficient β_2 is + 0.41. For adjusted R^2 we have a larger number and we wondered that could we reach a new larger number for this value and decided to conduct a new log transformation equation which is,

Model 3:

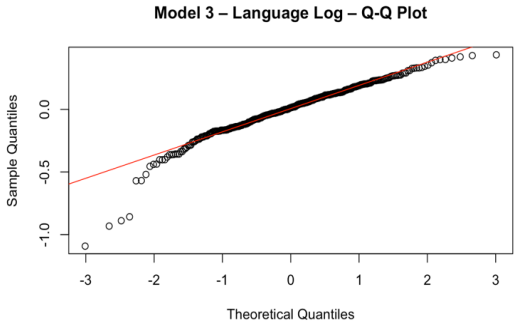
$$\log(Y_{\{Lang\}} + 1) = \beta_0 + \beta_1 \cdot Medu + \beta_2 \cdot famsize + \beta_3 \cdot Fedu$$

In this model firstly we omitted *famrel* since its p-value is larger than the others and we have a larger adjusted R^2 as we wish and it is 0.046. Again, the same significant variable which is *Medu* and its estimated coefficient β_2 is + 0.03.

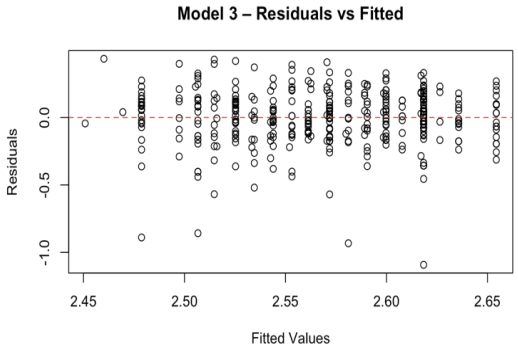
Finally our best estimated equation is:

$$\log(Y_{Lang} + 1) = 2.43 + 0.03 \times Medu$$

This model explains approximately 4.6% of the variation in Language grades. As *Medu* increases by 1 unit, the grade increases by 0.03 points as well. The intercept value β_0 = 2.43 represents the expected Language grade when all predictors are at their reference levels. Now, similar as above, let's move on assumptions for this model:



This QQ plot shows that most of the residual values are positioned to the neighborhood of the red line, except for only some deviations in the tail. Therefore, the residuals are approximately normally distributed and because of that, we can say that normality assumption is reasonably satisfied.



This plot shows that the residuals appear to be scattered randomly around the horizontal axes without composing any figure or pattern. This suggests that the assumptions of constant variance (homoscedasticity) are reasonably satisfied.

For multicollinearity we used again the VIF test and since all results are less than 5 (*Medu = Fedu =1.73* & *Famsize = 1*) we can say that there is no multicollinearity among predictors.

For other assumptions (**Linearity** between predictors and the response variable, and **Independence** of errors), we can say similar things as Math grades.

Therefore, the Model 3 is a statistically significant model with best estimators for predicting Lang grades.

Results & Conclusion

Our analysis provides valuable insights into what influences educational success. Using data from high school students across Math and Language we conducted a statistical analysis. The key findings are as follows:

To begin with, we analyzed **romantic relationships** and found being in a relationship showed no significant impact on grades. In terms of **study time**, study time significantly affects Language grades, with more hours being associated with higher grades. The impact on Math is less significant. Regarding **alcohol consumption**, weekend alcohol consumption negatively affects Language performance while its impact on Math is insignificant. Then **higher education motivation** emerged as an important factor, students who aim to pursue higher education perform better in both Math and Language, highlighting the role of long-term academic goals. Next, we investigated **academic support** and found that students who receive school support had significantly lower Math grades, showing that they may be struggling. Paid support showed a borderline positive effect while family support was insignificant. None of these had a meaningful impact on Language. Lastly, we considered **family background**, where among the family related variables, mother's education level and family size emerged as significant factors influencing Math grades, for Lang we had only *Medu.*

In conclusion, academic success is influenced by personal habits, family background and motivation. While key factors were identified, other elements like personality, teacher quality and friends may also play a role and we have clarified such reasons deeply as below.

Suggestion for Future Works

In this research, we have clarified several important factors affecting students' success, indeed there are several things to improve this research and generalize the results. First of all, expanding the observations leads to improved explanatory and reliability of models. Also, using time series and panel data allows researchers to watch changes in students' success in the long run and hence, causality relationships are constructed clearly.

To better explain individual differences of students', psychological variables such as anxiety levels, motivation, learning strategies and self-confidence to the model and of course the data. Additionally, school factors should also be searched such as classroom environment, teacher quality and instructional methods. In today's world, we have tech everywhere. Therefore, digital learning context variables should also be considered such as access to digital resources and students' online study habits.

Although this research focused on main effects only, interaction effects (e.g. *Medu x Studytime*) also can be included to conduct more complex and deeper models. Finally, beyond the linear models, researchers may consider using machine learning techniques to catch nonlinear relationships more efficiently.