Beyond the Books: Exploring the Impact of Parental Education and Study Time on Students' Academic Performance

Suryam Gupta

Table of contents

1 Abstract **2**

2 Research Question **2**

3 Hypothesis **2**

4 Descriptive Statistics **3**

4.1 About the dataset 3
4.2 Data Collection 3
4.3 Description of Attributes 4
4.4 Importing data 6
4.5 Data Pre-processing 6
4.6 Data Summarization 8

5 Graphs and Charts **10**

5.1 Univariate plot of studytime 10
5.2 Univariate plot of Medu 11
5.3 Univariate plot of Fedu 12
5.4 Univariate plot of final_grade 13
5.5 Relation between studytime and Parental Education 13
5.6 Relation between studytime and final_grade 14
5.7 Correlation Matrix 15

6 Directed Acyclic Graph (DAG) **17**

6.1 DAG Overview 17
6.2 Constructing the DAG 17

7 Hypothesis Testing by Individual Causal Inference Study 19
7.1 Does studytime have a significant effect on final_grade 19
7.1.1 Model Fitting 20
7.1.2 Interpretation of the Model Results 21
7.1.3 Confidence Interval Interpretation 22
7.1.4 Conclusion 23
7.1.5 Diagnostics and Model Evaluation 23
7.2 Does parental education have a significant effect on final_grade 27
7.2.1 Interpretation of Coefficients - A Quick Analysis 28
8 Conclusion 29
9 References 30

1 Abstract

Education is a key factor for achieving a long-term economic progress. This is an important research question because understanding what influences students' grades can help schools, families, and students themselves make better decisions to improve learning outcomes. By looking at how different aspects of a student's life, such as their home address type (rural or urban), their father's/mother's education level, study time, etc., we can gain valuable insights. This topic is important in social sciences because improving student performance has long-term benefits for both individuals and society, such as better career opportunities and greater social equality.

2 Research Question

How parental education and study time affects the academic performance of a student.

3 Hypothesis

"Students with higher parental education levels and more number of hours per week dedicated to studying have better academic performance, as measured by their final grade."

This hypothesis stems from findings in educational research that parental involvement and support, as well as a student's study habits, are key determinants of academic success.

i. Parental Education: Studies have consistently shown that parental education level is positively associated with students' academic performance (and mental health as well). Parents with higher education levels are more likely to provide academic support, create an environment conducive to studying, and have higher expectations for their child's academic success.

ii. Study Time: Academic performance often improves with increased study time, as students who dedicate more hours to studying are better prepared for exams and assignments. This aligns with theories of self-discipline and time investment in education.

This hypothesis answers the research question by specifically considering demographic (parental education) and school-related (study time) factors, both of which are prominent in the dataset and are likely to have a strong influence on overall academic performance.

4 Descriptive Statistics

4.1 About the dataset

This data approaches student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features, and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks.

For this project, we will combine the Mathematics (mat) and Portugal (por) dataset, and conduct the further analysis.

Important note: The target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

Dataset link - https://archive.ics.uci.edu/dataset/320/student+performance

4.2 Data Collection

Research Paper Link - https://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf

Citation - Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

In Portugal, the secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. A 20-point grading scale is used, where 0 is the lowest grade and 20 is the perfect score. During the school year, students are evaluated in three periods and the last evaluation (G3 of Table 1) corresponds to the final grade. This study will consider data collected during the 2005- 2006 school year from two public schools, from the Alentejo region of Portugal. The database was built from two sources: school reports, based on paper sheets and including few attributes (i.e. the three period grades and number of school absences); and questionnaires, used to complement the previous information. The authors designed the latter with closed questions (i.e. with predefined options) related to several demographic (e.g. mother's education, family income), social/emotional (e.g. alcohol consumption) (Pritchard and Wilson 2003) and school related (e.g. number of past class failures) variables that were expected to affect student performance. The questionnaire was reviewed by school professionals and tested on a small set of 15 students in order to get a feedback. The final version contained 37 questions in a single A4 sheet and it was answered in class by 788 students. Later, 111 answers were discarded due to lack of identification details (necessary for merging with the school reports). Finally, the data was integrated into two datasets related to Mathematics (with 395 examples) and the Portuguese language (649 records) classes. During the preprocessing stage, some features were discarded due to the lack of discriminative value. For instance, few respondents answered about their family income (probably due to privacy issues), while almost 100% of the students live with their parents and have a personal computer at home. The remaining attributes are shown in Table 1, where the last four rows denote the variables taken from the school reports.

4.3 Description of Attributes

Attribute	Description	Domain
sex	student's sex	binary: female or male
age	student's age	numeric: from 15 to 22
school	student's school	binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira
address	student's home address type	binary: urban or rural
Pstatus	parent's cohabitation status	binary: living together or apart
Medu	mother's education	numeric: from 0 to 4 (a)
Mjob	mother's job	nominal (b)
Fedu	father's education	numeric: from 0 to 4 (a)
Fjob	father's job	nominal (b)
guardian	student's guardian	nominal: mother, father or other
famsize	family size	binary: 3 or > 3
famrel	quality of family relationships	numeric: from 1 (very bad) to 5 (excellent)

Attribute	Description	Domain
reason	reason to choose this school	nominal: close to home, school reputation, course preference or other
traveltime	home to school travel time	numeric: 1 (< 15 min.), 2 (15 to 30 min.), 3 (30 min. to 1 hour), 4 (> 1 hour)
studytime	weekly study time	numeric: 1 (< 2 hours), 2 (2 to 5 hours), 3 (5 to 10 hours), 4 (> 10 hours)
failures	number of past class failures	numeric: n if 1 n < 3, else 4
schoolsup	extra educational school support	binary: yes or no
famsup	family educational support	binary: yes or no
activities	extra-curricular activities	binary: yes or no
paidclass	extra paid classes	binary: yes or no
internet	Internet access at home	binary: yes or no
nursery	attended nursery school	binary: yes or no
higher	wants to take higher education	binary: yes or no
romantic	with a romantic relationship	binary: yes or no
freetime	free time after school	numeric: from 1 (very low) to 5 (very high)
goout	going out with friends	numeric: from 1 (very low) to 5 (very high)
Walc	weekend alcohol consumption	numeric: from 1 (very low) to 5 (very high)
Dalc	workday alcohol consumption	numeric: from 1 (very low) to 5 (very high)
health	current health status	numeric: from 1 (very bad) to 5 (very good)
absences	number of school absences	numeric: from 0 to 93
G1	first period grade	numeric: from 0 to 20
G2	second period grade	numeric: from 0 to 20
G3	final grade	numeric: from 0 to 20 (target)

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

b: teacher, health care related, civil services (e.g. administrative or police), at home or other.

4.4 Importing data

```
library(readr)
student_mat <- read_csv("student_performance\\student\\student_mat.csv",
                        show_col_types = FALSE)
student_por <- read_csv("student_performance\\student\\student_por.csv",
                        show_col_types = FALSE)
```

4.5 Data Pre-processing

Making a new subject column, as we will combine the datasets and would like to have a distinction:

```
student_mat$subj <- "M"
student_por$subj <- "P"
dim(student_mat)
```

```
[1] 395  34
```

```
dim(student_por)
```

```
[1] 649  34
```

Combining Math and Portugal language datasets:

```
student <- rbind(student_mat, student_por)
```

According to the hypothesis, the variables of interest are Medu (mother's education), Fedu (father's education), studytime (weekly study time (in hours)), subject (subj), and G3 (final grade). Hence, to have no discrepancies or ambiguity in the analysis later in the project and keep things straightforward, we are only keeping these columns in the dataset.

```
student <- student[,c('Medu', 'Fedu', 'studytime', 'subj', 'G3')]
```

Renaming "G3" column to "final_grade" for clarity.

```r
names(student)[names(student) == "G3"] <- "final_grade"
```

According to the hypothesis, we would need to combine Mother's education and Father's education into a single unit called "Parental Education." The two ways of doing that are:

i. average - We consider that the education level of both parent have equal contribution to the child's academic performance and give equal weights to both. Here, we do run into the problem of getting values such as 3.5 (average of 3 and 4), which is inconsistent with the Medu and Fedu data. Hence, since one of the parent indeed has a higher education level value and does contribute to/affect their child's education accordingly, we can round off the average to the next higher integral value.
ii. max - We consider that the highest education of either of the parent is enough to independently determine its affect on the child's academic performance.

For this study, we will explore only the latter method. The maximum education level of either parent often represents the highest educational attainment in the household. Research suggests that a parent with higher education can act as the strongest role model or source of academic support, setting higher aspirations and expectations for the child. The more highly educated parent may contribute disproportionately to the child's academic environment, offering targeted guidance, better resources, and strategic support for learning. This method is more interpretable because it focuses on the most influential educational level in the household, avoiding dilution of the effect that might happen when averaging. Studies have shown that the educational attainment of the more educated parent (especially the mother, in many cases) tends to have a significant impact on children's academic success, particularly in lower-income or resource-constrained households.

Note: We can definitely make Pedu_max and Pedu_avg and conduct 2 separate analysis However, considering the hypothesis which already has "studytime" factor, and later we consider "subject" in our DAG as well, this might over complicate things and go out of the scope of this project.

```r
# student$Pedu_avg <- ceiling(rowMeans(cbind(student$Medu, student$Fedu)))
# student$Pedu_avg <- ceiling((student$Medu + student$Fedu) / 2)
student$Pedu <- pmax(student$Medu, student$Fedu)
```

```r
table(student$Pedu)
```

```
  0   1   2   3   4 
  1 138 288 252 365 
```

```
any(is.na(student))
```

```
[1] FALSE
```

There are no NA values in the dataset.

```
head(student)
```

```
# A tibble: 6 x 6
   Medu  Fedu studytime subj  final_grade  Pedu
  <dbl> <dbl>     <dbl> <chr>       <dbl> <dbl>
1     4     4         2 M               6     4
2     1     1         2 M               6     1
3     1     1         2 M              10     1
4     4     2         3 M              15     4
5     3     3         2 M              10     3
6     4     3         2 M              15     4
```

4.6 Data Summarization

```
dim(student)
```

```
[1] 1044    6
```

The student dataset now has 1044 rows and 6 columns.

We now use 2 summarization function:

1. summary() - Provides a statistical summary of each variable in a dataset. For numeric variables, it returns statistics like Min, 1st Qu., Median, Mean, 3rd Qu., and Max.

```
summary(student)
```

```
      Medu            Fedu          studytime         subj          
 Min.   :0.000   Min.   :0.000   Min.   :1.00   Length:1044       
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   Class :character  
 Median :3.000   Median :2.000   Median :2.00   Mode  :character  
 Mean   :2.603   Mean   :2.388   Mean   :1.97                     
```

	Mean	SD	Min	Max	Median	N
Medu	2.60	1.12	0.00	4.00	3.00	1044
Fedu	2.39	1.10	0.00	4.00	2.00	1044
studytime	1.97	0.83	1.00	4.00	2.00	1044
final_grade	11.34	3.86	0.00	20.00	11.00	1044
Pedu	2.81	1.06	0.00	4.00	3.00	1044

```
 3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:2.00  
 Max.   :4.000   Max.   :4.000   Max.   :4.00  
  final_grade         Pedu      
 Min.   : 0.00   Min.   :0.000  
 1st Qu.:10.00   1st Qu.:2.000  
 Median :11.00   Median :3.000  
 Mean   :11.34   Mean   :2.807  
 3rd Qu.:14.00   3rd Qu.:4.000  
 Max.   :20.00   Max.   :4.000  
```

2. kableExtra - Designed to enhance the visual appearance of tables created with knitr::kable(). It adds advanced styling options to make tables more visually appealing for reports, presentations, or dashboards.

3. glimpse(): Provides a quick overview of a dataset in a more compact, horizontal format. It displays the structure of the dataset, i.e. the number of rows and columns, followed by each column's name, data type, and a preview of its values.

```
suppressPackageStartupMessages(library(dplyr))
library(dplyr)
glimpse(student)
```

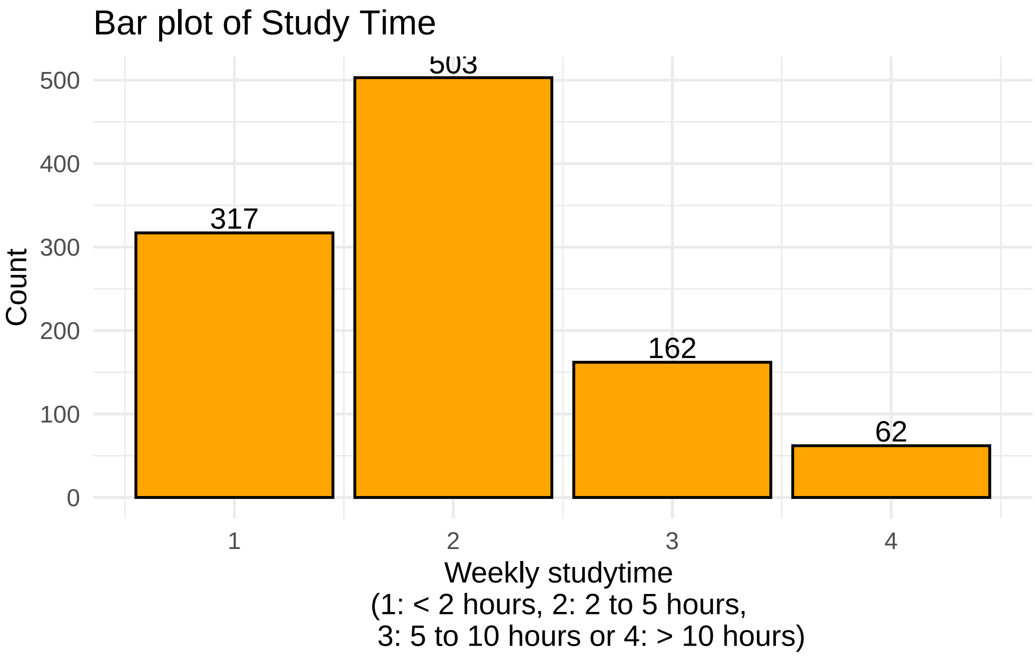
```
Rows: 1,044
Columns: 6
$ Medu        <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4~
$ Fedu        <dbl> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2, 3~
$ studytime   <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, 1~
$ subj        <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M"~
$ final_grade <dbl> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16, 14~
$ Pedu        <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 4, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4~
```

5 Graphs and Charts

```
library(ggplot2)
```

5.1 Univariate plot of studytime

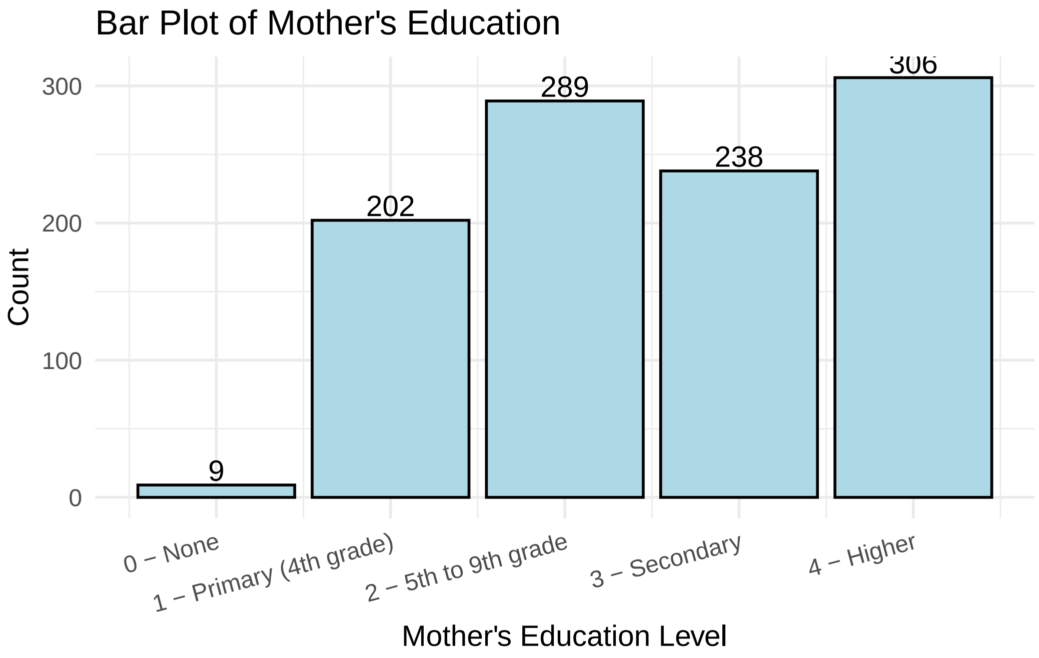
```
ggplot(student, aes(x=studytime)) +
  geom_bar(fill = "orange", color = "black") +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.2) +
  labs(title = "Bar plot of Study Time",
       x = "Weekly studytime \n(1: < 2 hours, 2: 2 to 5 hours,
       3: 5 to 10 hours or 4: > 10 hours)", y = "Count") +
  theme_minimal()
```



Around half of the total students (503) have a studytime level of 2, i.e., they study between 2 to 5 hours in a week. This is followed by 317 students having a studytime of 1 (less than 2 hours per week), 162 students with studytime 3 (5 to 10 hours per week), and lastly, 62 students with studytime 4 (more than 10 hours per week).

5.2 Univariate plot of Medu

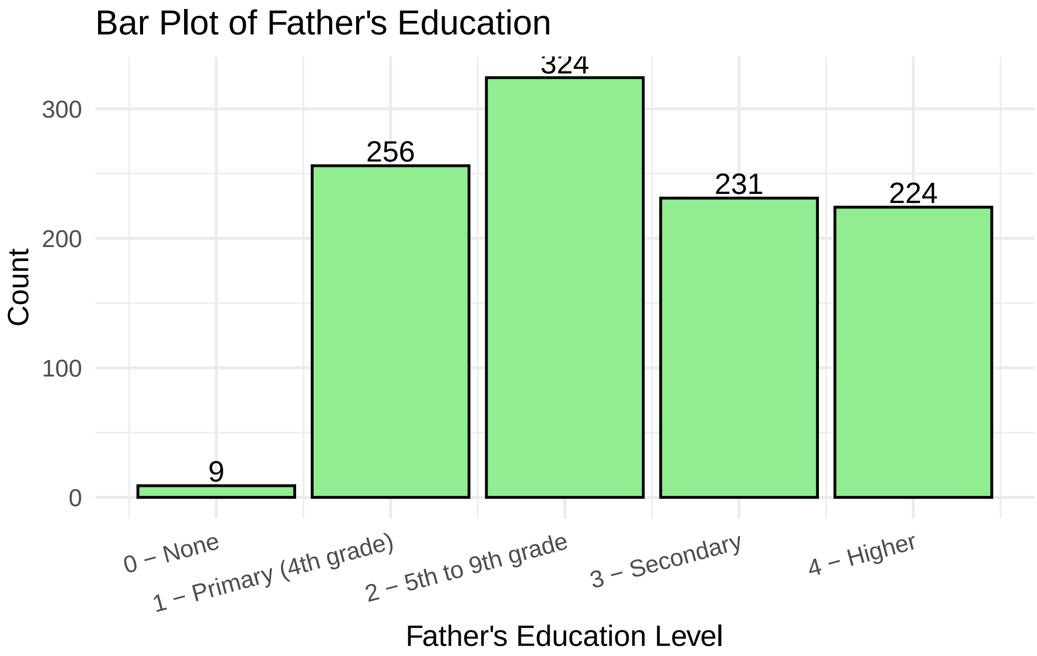
```
ggplot(student, aes(x = Medu)) +
  geom_bar(fill = "lightblue", color = "black") +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.2) +
  labs(title = "Bar Plot of Mother's Education",
       x = "Mother's Education Level", y = "Count") +
  scale_x_continuous(
    breaks = c(0, 1, 2, 3, 4),
    labels = c("0 - None", "1 - Primary (4th grade)",
      "2 - 5th to 9th grade", "3 - Secondary", "4 - Higher")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 15, hjust = 1))
```



Out of 1044, 306 students have mothers who have education level of 4, followed by 289 and 238 students with mother's education of 2 and 3 respectively. Comparatively few of 202 have Medu level 1, and only a negligible amount of 9 have Medu as 0.

5.3 Univariate plot of Fedu

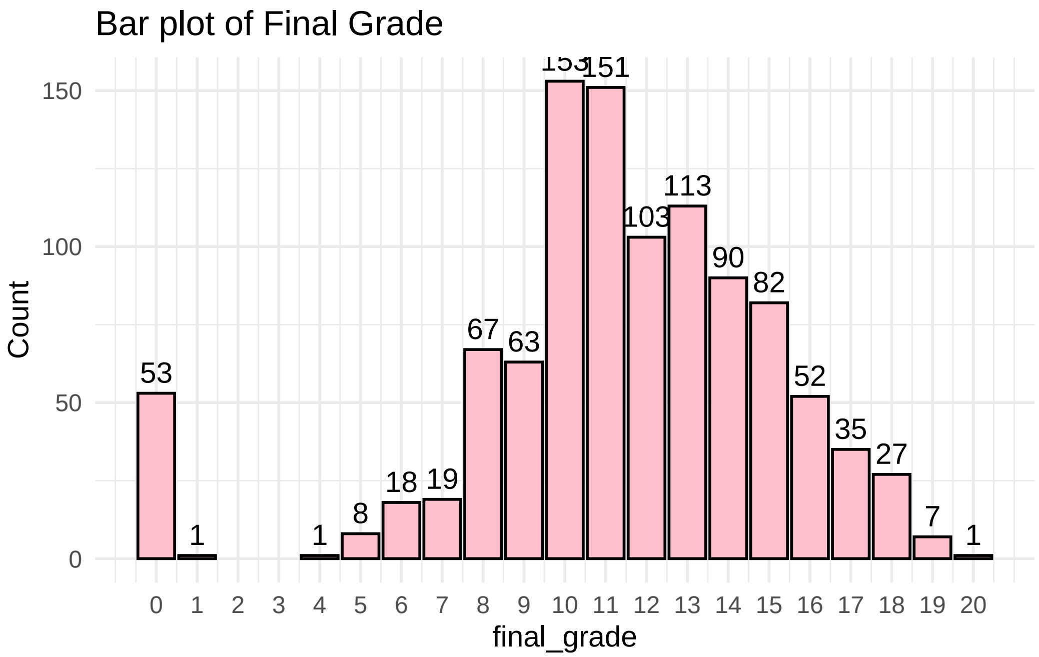
```r
ggplot(student, aes(x = Fedu)) +
  geom_bar(fill = "lightgreen", color = "black") +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.2) +
  labs(title = "Bar Plot of Father's Education",
       x = "Father's Education Level", y = "Count") +
  scale_x_continuous(
    breaks = c(0, 1, 2, 3, 4),
    labels = c("0 - None", "1 - Primary (4th grade)",
      "2 - 5th to 9th grade", "3 - Secondary", "4 - Higher")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 15, hjust = 1))
```



In the case of father's education, most students (324) have Fedu level of 2. Fedu level of 1, 3, and 4 are quite similar at 256, 231, and 224 respectively. Only 9 students have Fedu level of 0.

5.4 Univariate plot of final_grade

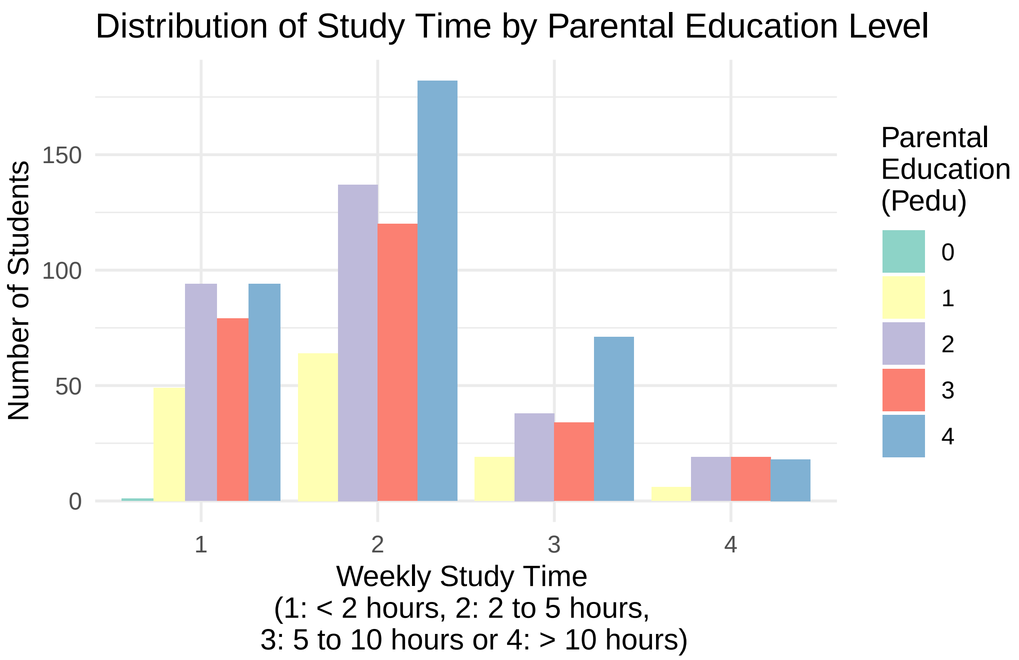
```r
ggplot(student, aes(x=final_grade)) +
  geom_bar(fill = "pink", color = "black") +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.5) +
  labs(title = "Bar plot of Final Grade",
       x = "final_grade", y = "Count") +
  scale_x_continuous(breaks = seq(0, 20, by = 1)) +
  theme_minimal()
```



Except a few students receiving 0, the final_grade variables has quite a normal distribution.

5.5 Relation between studytime and Parental Education

```r
ggplot(student, aes(x = factor(studytime), fill = factor(Pedu))) +
  geom_bar(position = "dodge") +
  labs(x = "Weekly Study Time \n(1: < 2 hours, 2: 2 to 5 hours,
  3: 5 to 10 hours or 4: > 10 hours)", y = "Number of Students",
       fill = "Parental \nEducation \n(Pedu)",
       title = "Distribution of Study Time by Parental Education Level") +
  scale_fill_brewer(palette = "Set3") +
  theme_minimal()
```



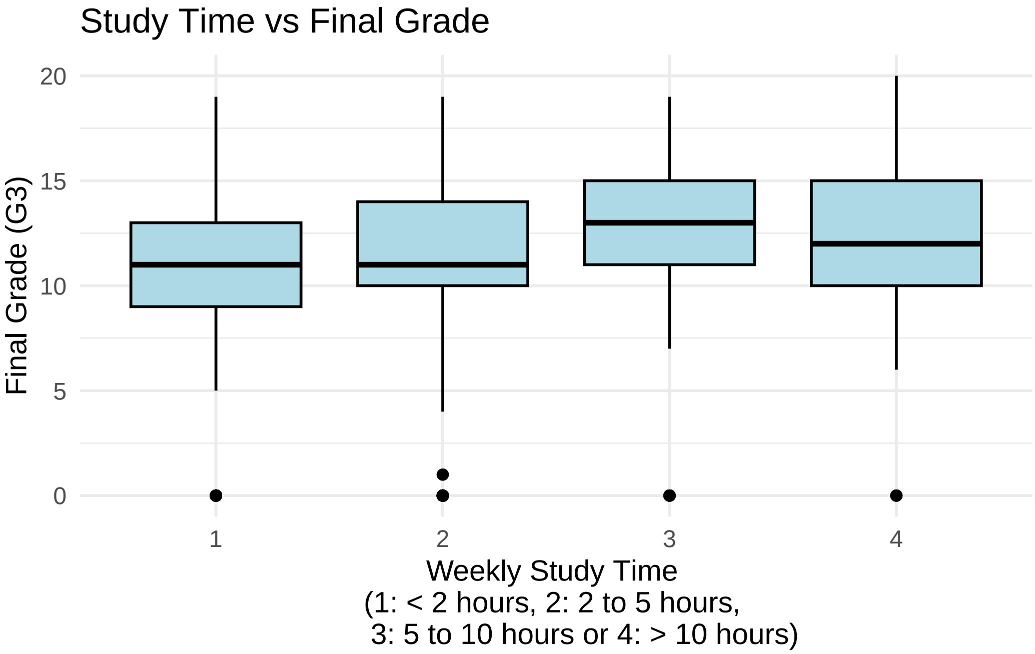
Here, in almost all cases, irrespective of the number of hours a student is dedicating to this studies in a week, the most number of students always seem to be a Pedu value of 4, suggesting that even if one of the parent is highly educated, they are able to have a positive effect on their child's studytime. It is also interesting to note that the second most prominent Pedu value is 2, not 3.

5.6 Relation between studytime and final_grade

```
for (i in c(1:4)) {
  print(mean( student$final_grade[student$studytime==i] ))
}
```

```
[1] 10.58044
[1] 11.33598
[1] 12.49383
[1] 12.27419
```

```
ggplot(student, aes(x = factor(studytime), y = final_grade)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Study Time vs Final Grade",
       x = "Weekly Study Time \n(1: < 2 hours, 2: 2 to 5 hours,
       3: 5 to 10 hours or 4: > 10 hours)", y = "Final Grade (G3)") +
  theme_minimal()
```



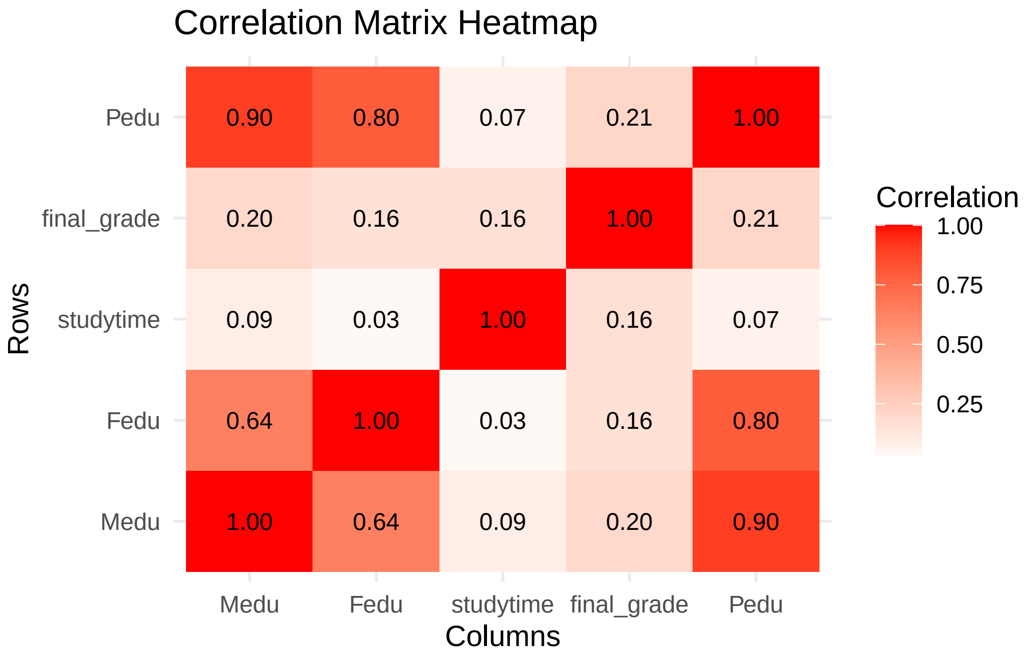
We notice that for weekly study time level of 3, we have the highest median of final grade at 12.49, which makes sense as the more time one dedicates to studying in a week, the better they perform. However, for that of level 4, the median is slightly less at 12.27. While still pretty high, a slight decline from the previous result could be because studying for more than 10 hours in a week, along with doing other activities, could lead to exhaustion and worse mental and physical health, and consequently a bad academic performance. Studytime 1 and 2 have median final_grade median values of 10.58 and 11.33 respectively, again aligning with the hypothesis.

5.7 Correlation Matrix

```
correlation_matrix <- cor(student[sapply(student, is.numeric)],
                          use = "complete.obs")
```

```r
# Converting the correlation matrix to a long format
library(reshape2)
data_long <- melt(correlation_matrix)

ggplot(data_long, aes(Var2, Var1, fill = value)) +
  geom_tile() +
  geom_text(aes(label = sprintf("%.2f", value)), color = "black", size = 3) +
  scale_fill_gradient2(low ="blue", mid ="white", high ="red", midpoint = 0) +
  labs(title = "Correlation Matrix Heatmap",
       x = "Columns", y = "Rows", fill = "Correlation") +
  theme_minimal()
```



There is an expected high correlation among Medu/Fedu & Pedu variables, but since we are never going to use any combination of these variables together (in fact, from hereon, only Pedu will be analyzed), this will not pose an issue later in the study. Other than that, there is notably no correlation among any combination of variables except for Medu and Fedu themselves. This can be explained by the sociology concept of **Homophily**, which describes the tendency of people to seek out or be attracted to those who are similar to themselves.

6 Directed Acyclic Graph (DAG)

6.1 DAG Overview

A Directed Acyclic Graph (DAG) is a graphical representation used to illustrate assumptions about causal relationships between variables in a system. It consists of nodes (variables) and directed edges (arrows), where an arrow from one node to another indicates a hypothesized causal influence.

Key Characteristics of a DAG:

1. Directed: The edges (arrows) have a specific direction, showing the causal flow from one variable to another.
2. Acyclic: The graph contains no cycles; you cannot start at a variable and return to it by following the arrows.
3. Nodes: Represent variables of interest, including independent, dependent, and control variables.

Elements in a DAG:

1. Exposure: The main independent variable whose effect on the outcome is being studied.
2. Outcome: The dependent variable or response of interest.
3. Confounders: Variables that affect both the exposure and the outcome, potentially biasing the causal estimate.
4. Mediators: Intermediate variables through which the exposure affects the outcome.
5. Colliders: Variables caused by two or more other variables; conditioning on them can introduce bias.

6.2 Constructing the DAG

To hypothesize a DAG, we make the following logical reasoning based on common educational insights:

1. Pedu -> studytime: Higher parental education levels might influence students to prioritize their studies, leading to more study time. Parents with higher education often emphasize the importance of education, encourage disciplined study routines, and might provide better academic guidance.

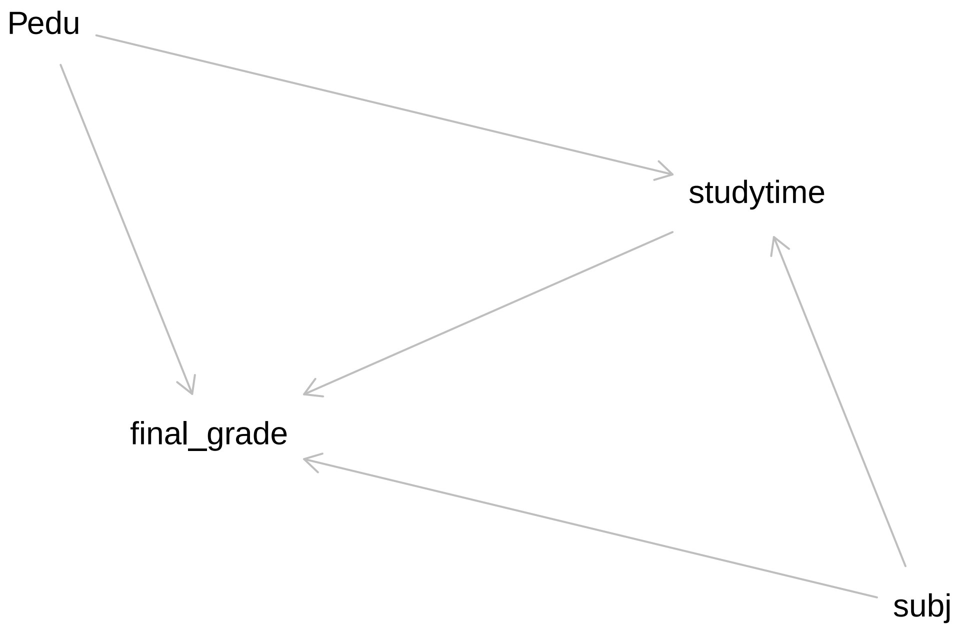
2. Pedu -> final_grade: Parental education has been shown to positively influence a student's academic success, as parents with higher education are often more involved in their child's education and can provide resources or a more supportive learning environment. Such parents may help the student with learning strategies, academic support, and a greater focus on educational achievement.
3. studytime -> final_grade: More time spent studying generally leads to better academic performance. This is a common assumption in educational research, as increased effort and preparation are typically associated with higher grades.
4. subj -> final_grade: Different subjects may have varying levels of difficulty, grading standards, and teacher expectations. Thus, the subject a student is studying could influence their final grade due to these differences in assessment.
5. subj -> studytime: The subject being studied could influence the amount of time spent on it. More challenging or time-intensive subjects (e.g., mathematics) might require more study time compared to subjects perceived as easier, like arts or humanities.

```r
library(dagitty)

dag <- dagitty('
  dag {
    Pedu -> studytime
    Pedu -> final_grade
    studytime -> final_grade
    subj -> final_grade
    subj -> studytime
  }
')

plot(dag)
```

```
Plot coordinates for graph not supplied! Generating coordinates, see ?coordinates for how to
```



7 Hypothesis Testing by Individual Causal Inference Study

7.1 Does studytime have a significant effect on final_grade

Null Hypothesis: Study time has no effect on final_grade.

Alternate Hypothesis: Study time has an effect on final_grade.

Level of Significance: 0.05

To study the individual/independent effect of studytime on final_grade, we consider all the paths between studytime and final_grade from our DAG and find all the open backdoor paths. To close these paths, we simply control for a particular variable between the path.

We can directly find the variables we need to control for by using the adjustmentSets() function.

```r
adjustmentSets(dag, exposure = "studytime", outcome = "final_grade")
```

```
{ Pedu, subj }
```

Hence, we need to control for Pedu and subj in this scenario.

7.1.1 Model Fitting

```
lm_studytime <- lm(final_grade ~ studytime + Pedu + subj, data = student)
summary(lm_studytime)
```

```
Call:
lm(formula = final_grade ~ studytime + Pedu + subj, data = student)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9903  -1.6197   0.1842   2.3167   9.1206

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5314     0.4466  14.623  < 2e-16 ***
studytime     0.7403     0.1358   5.451 6.24e-08 ***
Pedu          0.8039     0.1072   7.498 1.39e-13 ***
subjP         1.7629     0.2344   7.521 1.17e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.645 on 1040 degrees of freedom
Multiple R-squared:  0.113, Adjusted R-squared:  0.1105
F-statistic: 44.18 on 3 and 1040 DF,  p-value: < 2.2e-16
```

Regression Summary of lm_studytime

Variable	Estimate	Std. Error	t-Value	P-Value
(Intercept)	6.53e+00	4.47e-01	1.46e+01	3.52e-44
studytime	7.40e-01	1.36e-01	5.45e+00	6.24e-08
Pedu	8.04e-01	1.07e-01	7.50e+00	1.39e-13
subjP	1.76e+00	2.34e-01	7.52e+00	1.17e-13

```r
#install.packages("broom")
library(broom)
```

```
Warning: package 'broom' was built under R version 4.4.2
```

```r
#install.packages("gt")
library(gt)
```

```
Warning: package 'gt' was built under R version 4.4.2
```

```r
tidy_model <- tidy(lm_studytime)

tidy_model |>
  gt() |>
  tab_header(title = "Regression Summary of lm_studytime") |>
  fmt(columns = c(estimate, std.error, statistic, p.value),
      fns = scales::scientific) |>
  cols_label(term = "Variable", estimate = "Estimate", std.error =
             "Std. Error", statistic = "t-Value", p.value = "P-Value")
```

7.1.2 Interpretation of the Model Results

1. **Model Overview**:

    - The linear model predicts `final_grade` using `studytime`, `Pedu`, and `subj`.
    - The adjusted (R^2 = 0.1105) indicates that approximately 11.05% of the variance in `final_grade` is explained by the predictors in the model. While this value is relatively low, it is common in social science data and survey data where many unobserved factors influence the outcome. Also, since studytime and Pedu are categorical variables with values ranging from 0 to 4 and subj is also a categorical

variable with 2 unique values (M and P), it makes it more complex and difficult for the lm model to make predictions for an outcome variable which has discrete count values ranging from 0 to 20, resulting in a low R2 score.

2. **Coefficients**:
   - **Intercept (Estimate = 6.5314, p < 0.001)**:
     - The average `final_grade` for students with `studytime = 0`, `Pedu = 0`, and the reference level of `subj` ("Math") is 6.53.
   - **studytime (Estimate = 0.7403, p < 0.001)**:
     - For every one-unit increase in `studytime`, the `final_grade` increases by 0.74 points, holding `Pedu` and `subj` constant.
     - This positive and significant relationship (as p is less than 0.05, we reject the null hypothesis at 0.05 significance level) supports the hypothesis that more time dedicated to studying improves academic performance.
   - **Pedu (Estimate = 0.8039, p < 0.001)**:
     - For every one-unit increase in `Pedu`, the `final_grade` increases by 0.80 points, holding other variables constant.
     - This highlights the importance of higher parental education levels in predicting better academic performance.
   - **subjP (Estimate = 1.7629, p < 0.001)**:
     - Students studying Portuguese (`subjP`) have grades that are, on average, 1.76 points higher than those studying the reference subject (Math), holding other variables constant.
3. **Standard Errors**:
   - The standard errors for `studytime` (0.1358), `Pedu` (0.1072), and `subjP` (0.2344) are relatively small compared to their coefficients, indicating precise estimates of these effects.
4. **Statistical Significance (p-values)**:
   - All three predictors (`studytime`, `Pedu`, and `subjP`) have p-values much smaller than 0.05, confirming that their relationships with `final_grade` are statistically significant.
5. **Residual Standard Error**:
   - The residual standard error is 3.645, which reflects the average deviation of observed grades from the model's predictions.

7.1.3 Confidence Interval Interpretation

```
confint(lm_studytime, level = 0.95)
```

```
                2.5 %   97.5 %
(Intercept) 5.6549802 7.407824
studytime   0.4738261 1.006758
Pedu        0.5934777 1.014230
subjP       1.3030025 2.222864
```

For each one-unit increase in studytime, the final_grade is estimated to increase by 0.74 (point estimate), with a plausible range of 0.47 to 1.01. Since the entire interval is positive and does not include 0, this reinforces the conclusion that studytime has a statistically significant and positive effect on final_grade at the 95% confidence level.

Note: I am not reporting AIC and BIC as we are not comparing models. Additionally, not reporting RMSE as well because the main aim of this study is to do causal inferencing and not accurate predictions.

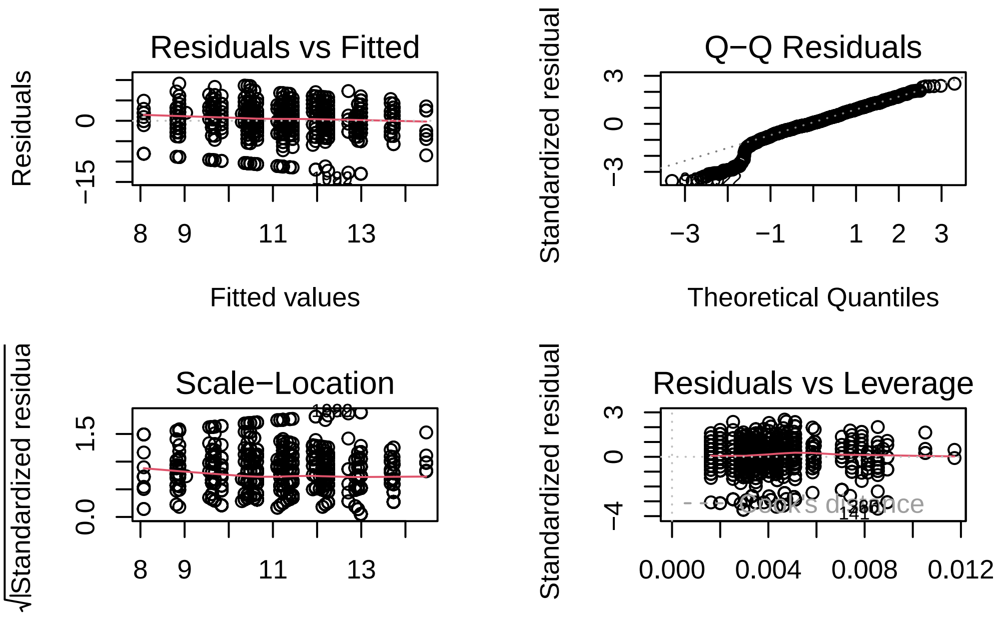
```
# AIC(lm_studytime)
# BIC(lm_studytime)
#
# predictions <- predict(lm_studytime, newdata = student)
# residuals <- student$final_grade - predictions
# RMSE <- sqrt(mean(residuals^2))
# RMSE
```

7.1.4 Conclusion

This model supports the hypothesis regarding the positive impact of both `studytime` and `Pedu` on `final_grade`. The inclusion of `subj` as a control also reveals that subject choice is an important factor. However, the low (R^2) suggests that additional variables or alternative modeling approaches might be needed to better capture the variance in academic performance.

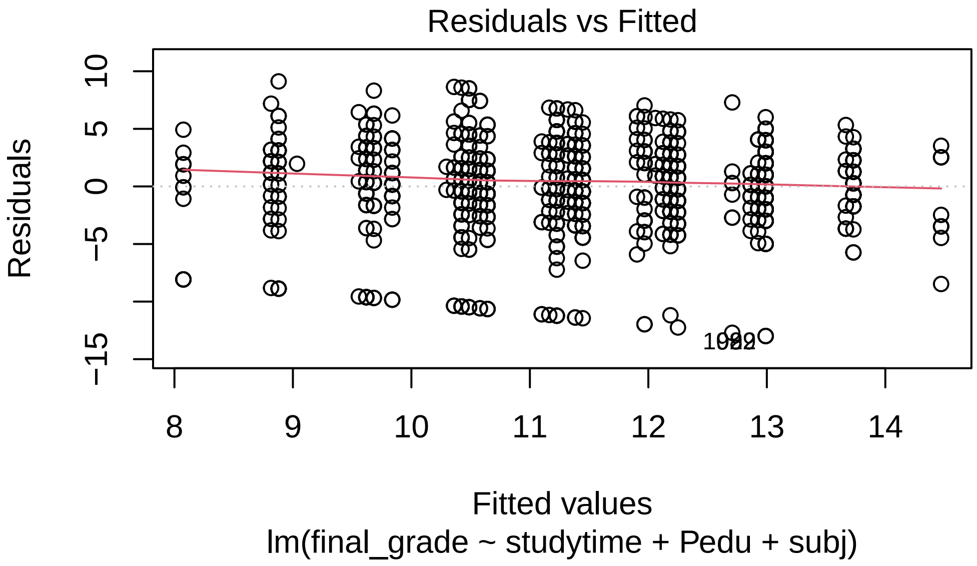
7.1.5 Diagnostics and Model Evaluation

```
par(mfrow = c(2, 2), mar = c(5, 5, 2, 2))
plot(lm_studytime)
```



Residuals vs Fitted (Top-Left)

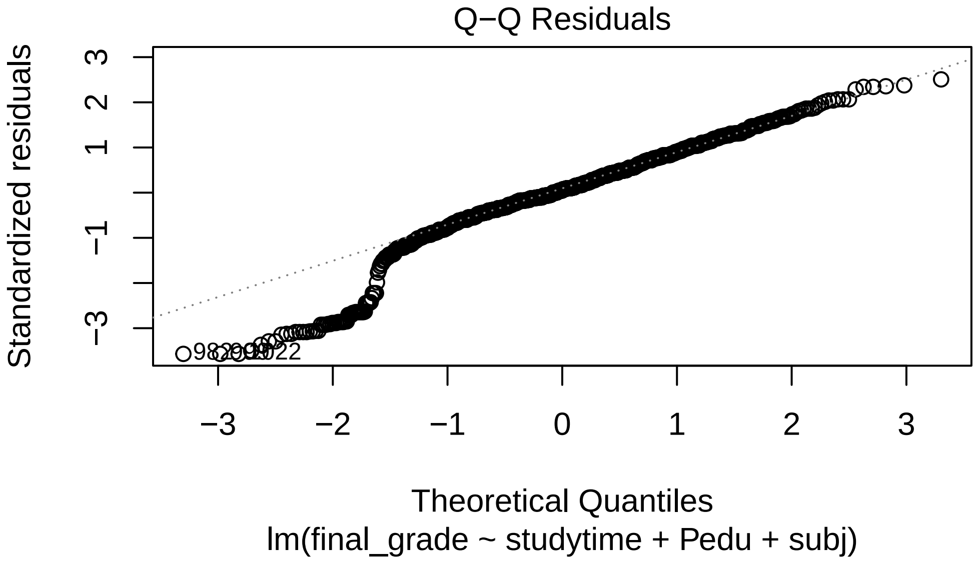
```
par(mfrow = c(1, 1))
plot(lm_studytime, which = 1)
```



- There are no clear patterns or curves, but the spread of residuals seems larger for certain fitted values, indicating some potential **heteroscedasticity**.
- A few extreme residuals (e.g., near -15) may suggest outliers.
- **Conclusion**: While the model seems to fit moderately well, the spread of residuals is uneven, which might hint at a violation of homoscedasticity.

Q-Q Plot of Residuals (Top-Right)

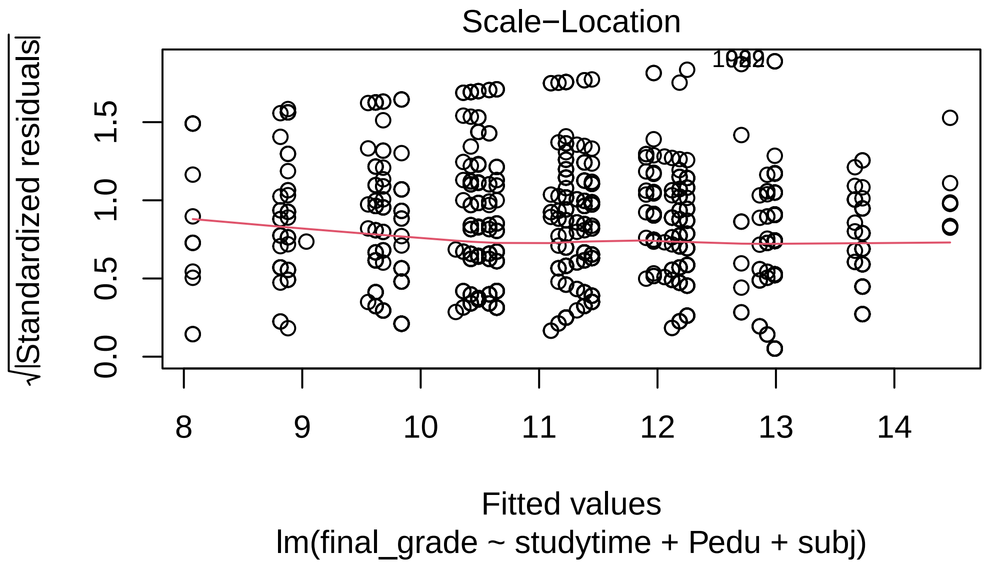
```r
plot(lm_studytime, which = 2)
```



- The residuals deviate from the line, especially at the extremes (tails), which indicates **non-normality**.
- The presence of outliers in both tails suggests that the assumption of normality may be violated.
- **Conclusion**: Residuals are not perfectly normal, but minor deviations are common.

Scale-Location Plot (Bottom-Left)

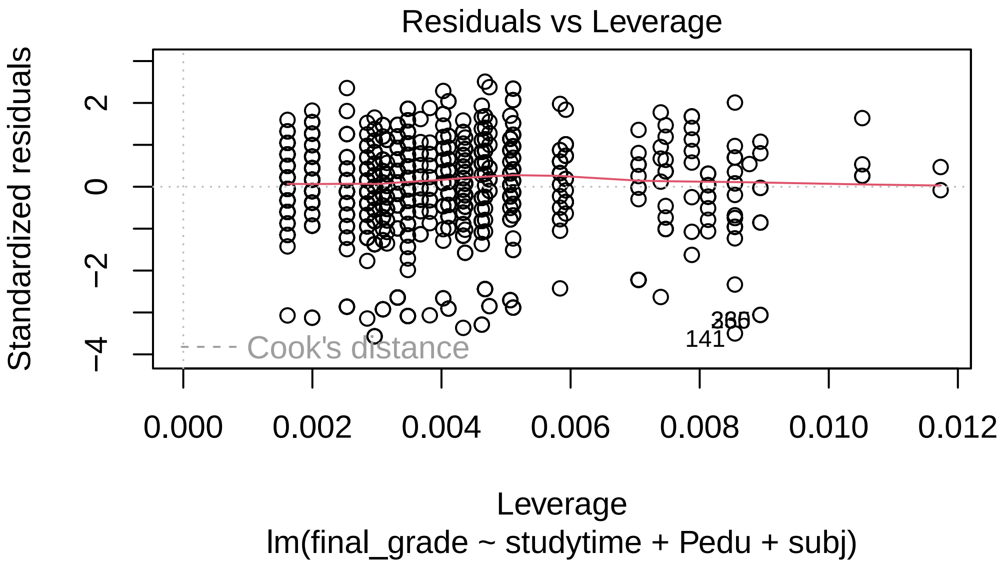
```r
plot(lm_studytime, which = 3)
```



- The residuals appear somewhat spread out unevenly across fitted values, which might suggest **heteroscedasticity**.
- While the red line is approximately horizontal, at higher fitted values, residuals appear to have slightly lower spread.
- **Conclusion**: There is evidence of slight heteroscedasticity.

Residuals vs Leverage (Bottom-Right)

```r
plot(lm_studytime, which = 5)
```



- Most points have low leverage, but there are a few with moderate leverage and larger residuals.
- Points near the Cook's distance threshold might warrant closer inspection for influence.
- **Conclusion**: While most data points are not influential, a few data points might significantly impact the model.

7.2 Does parental education have a significant effect on final_grade

Null Hypothesis: Parental education has no effect on final_grade.

Alternate Hypothesis: Parental education has an effect on final_grade.

Level of Significance: 0.05

```r
adjustmentSets(dag, exposure = "Pedu", outcome = "final_grade")
```

```
 {}
```

Hence, we do not need to control for any variables in this scenario.

```r
lm_Pedu <- lm(final_grade ~ Pedu, data = student)
summary(lm_Pedu)
```

```

Call:
lm(formula = final_grade ~ Pedu, data = student)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.243  -1.488   0.267   2.267   8.267 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.2230     0.3311  27.856  < 2e-16 ***
Pedu          0.7550     0.1104   6.842 1.33e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.783 on 1042 degrees of freedom
Multiple R-squared:  0.04299,	Adjusted R-squared:  0.04207 
F-statistic: 46.81 on 1 and 1042 DF,  p-value: 1.333e-11
```

```r
confint(lm_Pedu, level = 0.95)
```

```
                2.5 %    97.5 %
(Intercept) 8.5733090 9.8727000
Pedu        0.5384693 0.9715535
```

7.2.1 Interpretation of Coefficients - A Quick Analysis

1. **Pedu (Estimate = 0.7550, Std. Error = 0.1104, p < 0.001)**:

   - **Effect**: For every **1-unit increase in parental education**, the **final grade** increases by **0.755 points**.
   - **Statistical Significance**: The p-value (**1.33e-11**) is far below the 0.05 threshold, indicating that the effect of Pedu is **highly significant**, hence we reject the null hypothesis at 0.05 significance level.
   - **Practical Interpretation**: Parental education has a positive and meaningful effect on student performance.

2. **Confidence Interval**

- The 95% confidence interval for the effect of Pedu is: [0.538, 0.972]
- **Interpretation**: We are 95% confident that the true effect of parental education lies between **0.538 and 0.972**.
- Since the entire interval is positive (does not include 0), this reinforces the significance of Pedu.

3. Model Fit (R-squared and Adjusted R-squared)

- **Multiple R-squared = 0.043 (4.3%)**:
  - Approximately **4.3% of the variation in final grades** is explained by parental education alone.
  - While this is a small proportion, it is expected when considering a single predictor, as student performance is influenced by many other factors.
- **Adjusted R-squared = 0.042**:
  - Adjusted R^2 corrects for model complexity and remains almost identical here, confirming that the predictor (Pedu) is contributing to the model meaningfully.

4. Conclusion

- **Parental education (Pedu) has a statistically significant and positive effect** on student final grades.
- While the model explains only a small portion of the variation (4.3%), the effect size (0.755) is meaningful.
- The confidence interval further confirms the robustness of the estimate.

8 Conclusion

This project investigated the factors influencing student performance, measured by final grades, using parental education, study time, and subject as key predictors. The analysis was guided by a well-constructed Directed Acyclic Graph (DAG) to account for hypothesized relationships between variables.

Key findings include:
1. **Parental education** (Pedu) has a significant positive effect on final grades. A 1-unit increase in parental education corresponds to an increase of approximately **0.76 points** in the final grade.
2. **Study time** significantly impacts academic performance, even after controlling for parental education and subject, supporting the hypothesis that more study hours improve outcomes.
3. **Subject** (Math vs. Portuguese) influences final grades, with students in Portuguese scoring higher on average compared to Math.

The final model demonstrated a moderate explanatory power, highlighting that while parental education and study time are influential, other unobserved factors also contribute to student performance. Diagnostics confirmed the assumptions of linear regression, validating the model results.

Overall, the findings emphasize the importance of parental education and dedicated study time in fostering academic success, offering actionable insights for educators and policymakers to support students effectively.

9 References

1. https://archive.ics.uci.edu/dataset/320/student+performance

2. https://repositorium.sdum.uminho.pt/bitstream/1822/8024/1/student.pdf

3. Cortez, P. (2008). Student Performance [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

4. Pinquart, M., & Ebeling, M. (2020). Parental educational expectations and academic achievement in children and adolescents—a meta-analysis. Educational Psychology Review, 32(2), 463-480.

5. Davis-Kean, P. E., Tighe, L. A., & Waters, N. E. (2021). The role of parent educational attainment in parenting and children's development. Current Directions in Psychological Science, 30(2), 186-192.

6. Hammerstein, S., König, C., Dreisörner, T., & Frey, A. (2021). Effects of COVID-19-related school closures on student achievement-a systematic review. Frontiers in psychology, 12, 746289.

7. Wilder, S. (2023). Effects of parental involvement on academic achievement: a meta-synthesis. In Mapping the field (pp. 137-157). Routledge.