Student Performance Analysis

Jacob Thielemier

6 May 2024

1 Introduction

Educational achievement is crucial for personal development and economic advancement. Recent research has highlighted various factors influencing student performance, including socio-economic background, study habits, and school environment. This study will explore the relationships between these factors and academic outcomes among high school senior students. Utilizing a dataset comprising 10,000 observations with 6 variables, we employ a linear regression analysis to identify significant predictors of student performance. By quantifying these relationships, we seek to provide new insights that educators and policymakers can use to enhance education experiences and improve student outcomes. Yang et al. (2018)

Previous works in studying student performance were built around predicting future performance for a student. In education the values normally predicted are performance, knowledge, score or mark. O. & P. (2017) We are going to instead analyze existing data to show what has influenced student performance. We are specifically looking at the relationship between the amount of hours studied by a student and the students overall acadmeic performance. We will accomplish this by reviewing the data we have on hand, developing questions to answer and the models we will use, detailing and describing the dataset, plotting our data, and evaluating the models to best answer our questions of interest.

2 Materials and Methods

Students

Students performance is very crucial in solving issues of the learning process and one of the important matters to measure learning outcomes. Alhazmi & Sheneamer (2023) The Student Performance dataset we selected is designed to examine the factors influencing academic student performance. The dataset consists of 10,000 student records with the outcome measure being a Performance Index rating of 0-100. The covariates measured are: Hours.Studied, Previous.Scores, Extracurricular.Activities, Sleep.Hours, and Sample.Question.Papers.Practiced. The two tables below define the mean and standard deviation of our continuous and categorical variables.

Table 1: Summary Statistics for Continuous Variables

Variable	Mean	SD
Performance.Index	55.2248	19.212558
Hours.Studied	4.9929	2.589309
Previous.Scores	69.4457	17.343152
Sleep.Hours	6.5306	1.695863
Sample.Question.Papers.Practiced	4.5833	2.867348

Table 2: Frequency Distribution for Categorical Variables

Variable	Count	Percentage
Extracurricular.Activities No	5052	50.52
Extracurricular.Activities Yes	4948	49.48

Statistical Analysis

There are two primary goals of the analysis:

1. How does the number of hours studied influence the students performance?
2. What is the interaction of hours studied with hours slept and how does this influence the students performance?

We are going to use simple linear regression to analyze how the covariate of interest, Hours.Studied, impacts the Performance.Index. Our confouders are Sleep.Hours and Extracurricular.Activities. Our precision variables are Previous.Scores and Sample.Question.Papers.Practiced. We expect The proposed model we are using is:

- Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activities + Previous.Scores + Sample.Question.Papers.Practiced

We are using a interaction effect model to answer the second question to determine how the number of Sleep.Hours interacts with Hours.Studied to impact the Performance.Index. The interaction affect model:

- Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activities + Previous.Scores + Sample.Question.Papers.Practiced + Hours.Studied:Sleep.Hours

The Appendix contains plots for both models that show the dataset meets the liner regression assumptions of linearity, independence, normality, and equal variance. This means our dataset is ready for analyze and should provide accurate feedback without issue. We can also look at the historgrams in the Appendix to see that Performance.Index is bell shaped. We see that Hours.Studied and Sleep.Hours tend to be right skewed meaning most students have lower results on average.

3 Results

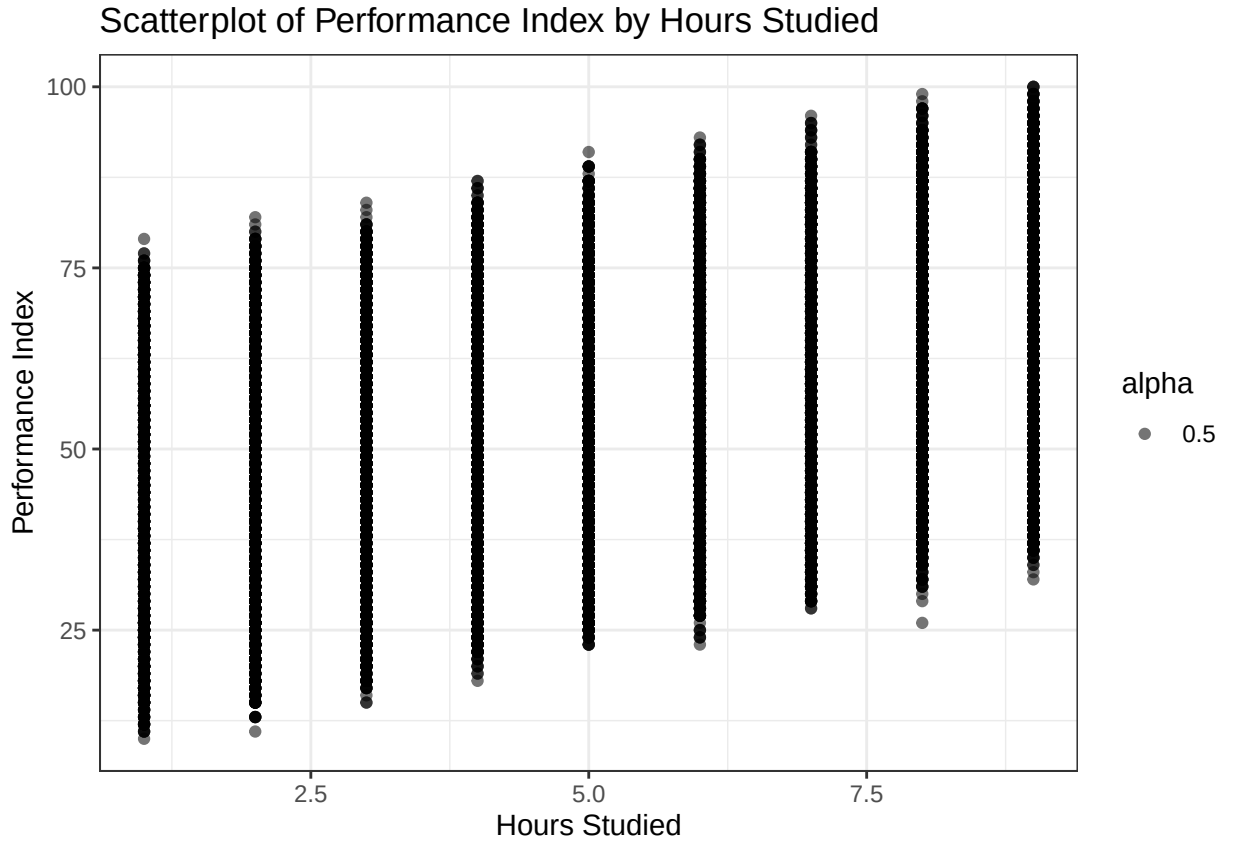
3.1 Question 1

We can see from the Table 3 below that each of our covariates are significant and have a P-value less than 0.01. This answers our Question 1 by defining that Hours.Studied is significant on the Performance.Index. We can see that for each hour studied the student performance increased by about 2.853 points. The scatterplot shows a visual change in the Performance.Index as students Hours.Studied increases.

Table 3: Proposed Model Linear Regression Summary

Term	Estimate	Std..Error	P.value	CI.2.5.	CI.97.5.
Intercept	-34.076	0.127	< 0.001	-34.325	-33.826
Hours Studied	2.853	0.008	< 0.001	2.838	2.868

Term	Estimate	Std..Error	P.value	CI.2.5.	CI.97.5.
Sleep Hours	0.481	0.012	< 0.001	0.457	0.504
Extracurricular Activities	0.613	0.041	< 0.001	0.533	0.693
Previous Scores	1.018	0.001	< 0.001	1.016	1.021
Sample Question Papers Practiced	0.194	0.007	< 0.001	0.180	0.208



3.2 Question 2

We can see from Table 4 below that once we add the interaction effect of Hours.Studied:Sleep.Hours the P-value is 0.0813. This indicates that the interaction is not significant. We can also determine this based on the small estimate value of 0.008 which indicates a small change and the small t value of 1.743.

This answers our Question 2 about if the interaction between Hours.Studied and Sleep.Hours influences that Performance.Index for the students. There is no significance in the interaction effect. This can seem as a surprise to convential thinking, but the purpose of statistcial analysis.

Table 4: Interaction Effect Model Linear Regression Summary

Term	Estimate	Std..Error	P.value	CI.2.5.	CI.97.5.
Intercept	-33.814	0.197	< 0.001	-34.199	-33.428
Hours Studied	2.800	0.031	< 0.001	2.738	2.862
Sleep Hours	0.440	0.026	< 0.001	0.389	0.491
Extracurricular Activities	0.613	0.041	< 0.001	0.533	0.693

Term	Estimate	Std..Error	P.value	CI.2.5.	CI.97.5.
Previous Scores	1.018	0.001	< 0.001	1.016	1.021
Sample Question Papers Practiced	0.194	0.007	< 0.001	0.180	0.208
Hours Studied:Sleep Hours	0.008	0.005	0.0813	-0.001	0.017

4 Discussion

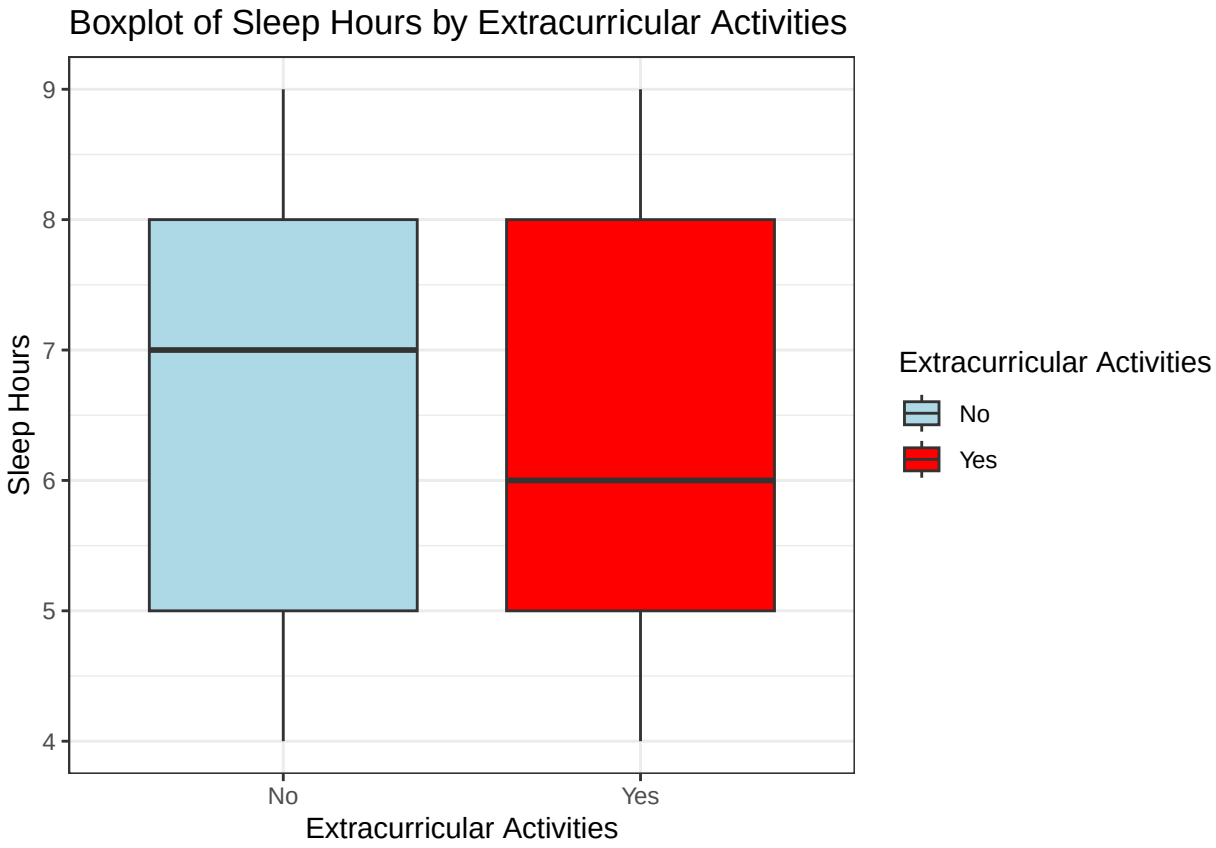
In the Appendix there is a bloxplot of Sleep.Hours by Extracurricular.Activities(#boxplot). This is the only boxplot that shows a direct correlation between Extracurricular.Activities and any other covariate. The plot shows that if they are in Extracurricular.Activities then the mean Sleep.Hours is 6, but if they do not participate in Extracurricular.Activities then the Sleep.Hours is 7. I suggest in future research that this be studied as a interaction effect.

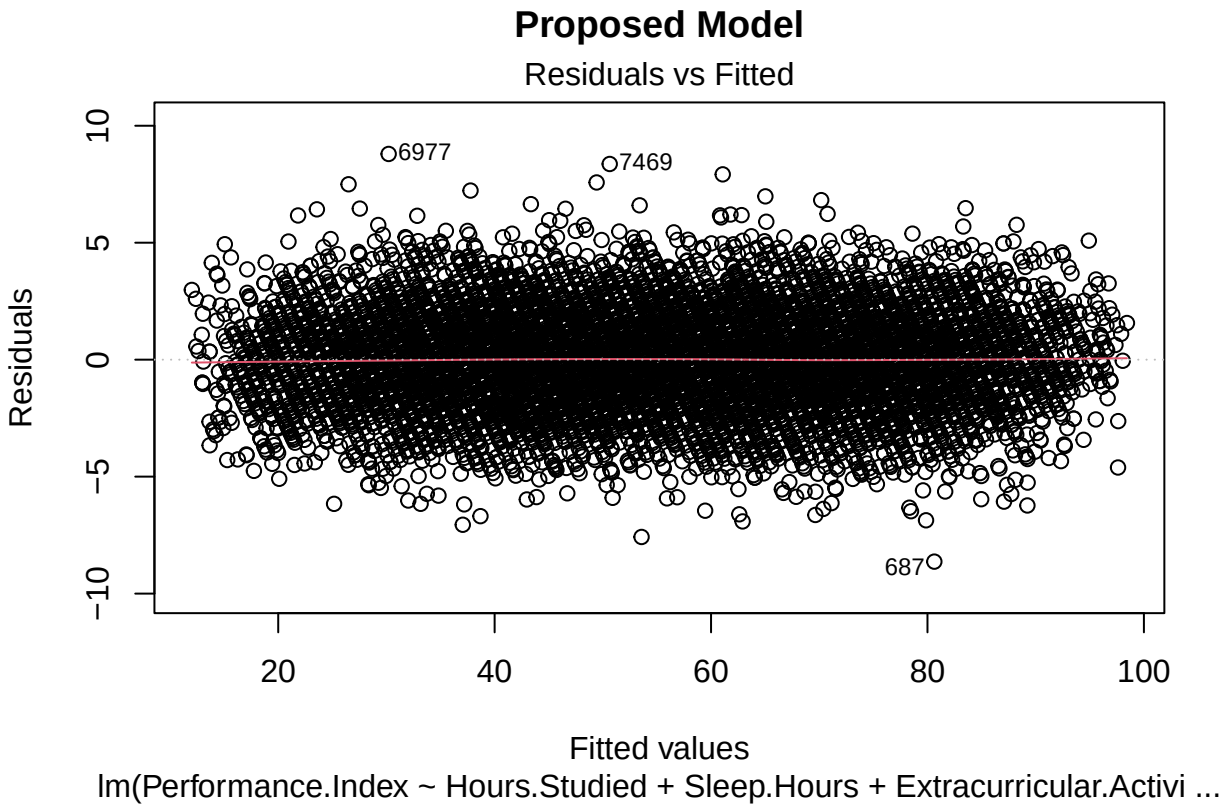
When comparing the two models we used ANOVA with the results in Table 5 below. The Sum of Squares due to the interaction term is 12.62 which means that adding the interaction term explains a small additional amount of variance in Performance.Index. The P-value associated with the F-statistic is 0.08133, which is greater than the normal significance level of 0.05. This means that the addition of the interaction term between Hours.Studied and Sleep.Hours does not improve the model at the 5% significance level.

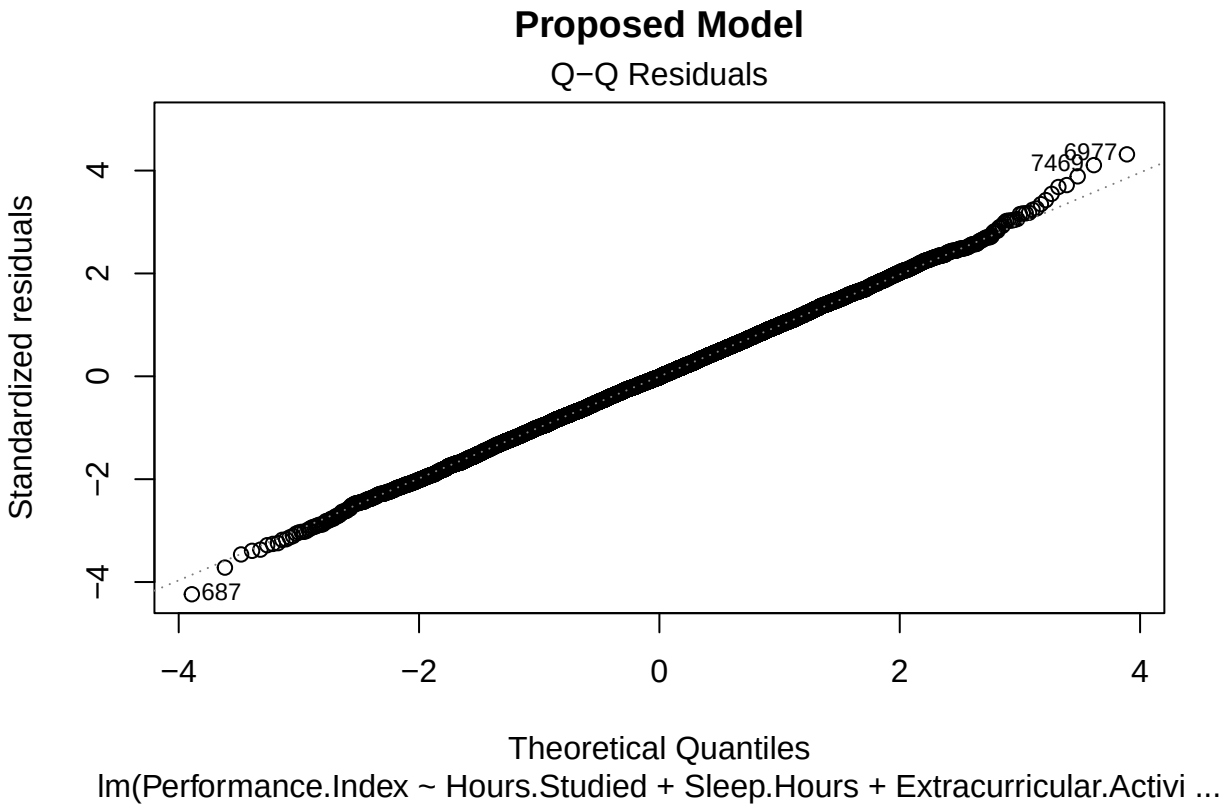
Table 5: ANOVA Results for Comparing the Two Models

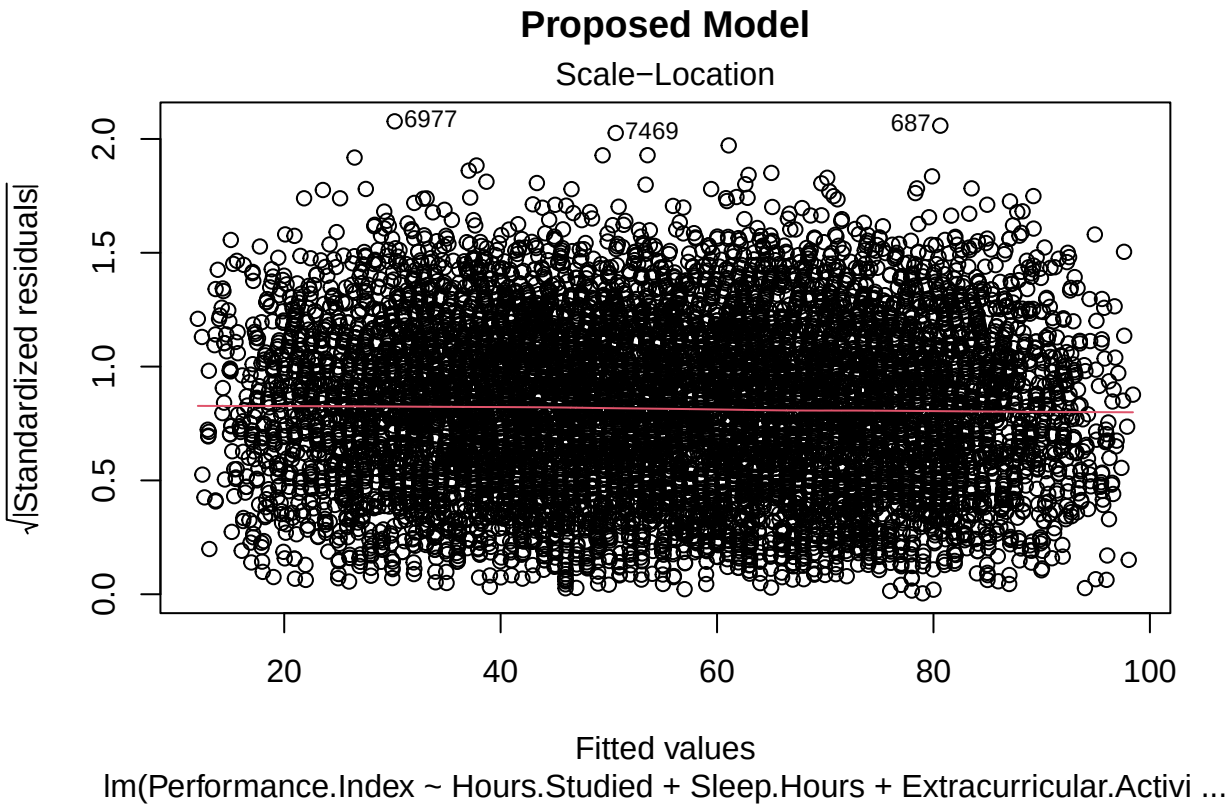
Model	RSS	Df	Sum of Sq	F	P-value
Model 1: Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activities + Previous.Scores + Sample.Question.Papers.Practiced	41514	-	-	-	-
Model 2: Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activities + Previous.Scores + Sample.Question.Papers.Practiced + Hours.Studied:Sleep.Hours	41501	1	12.62	3.039	0.081

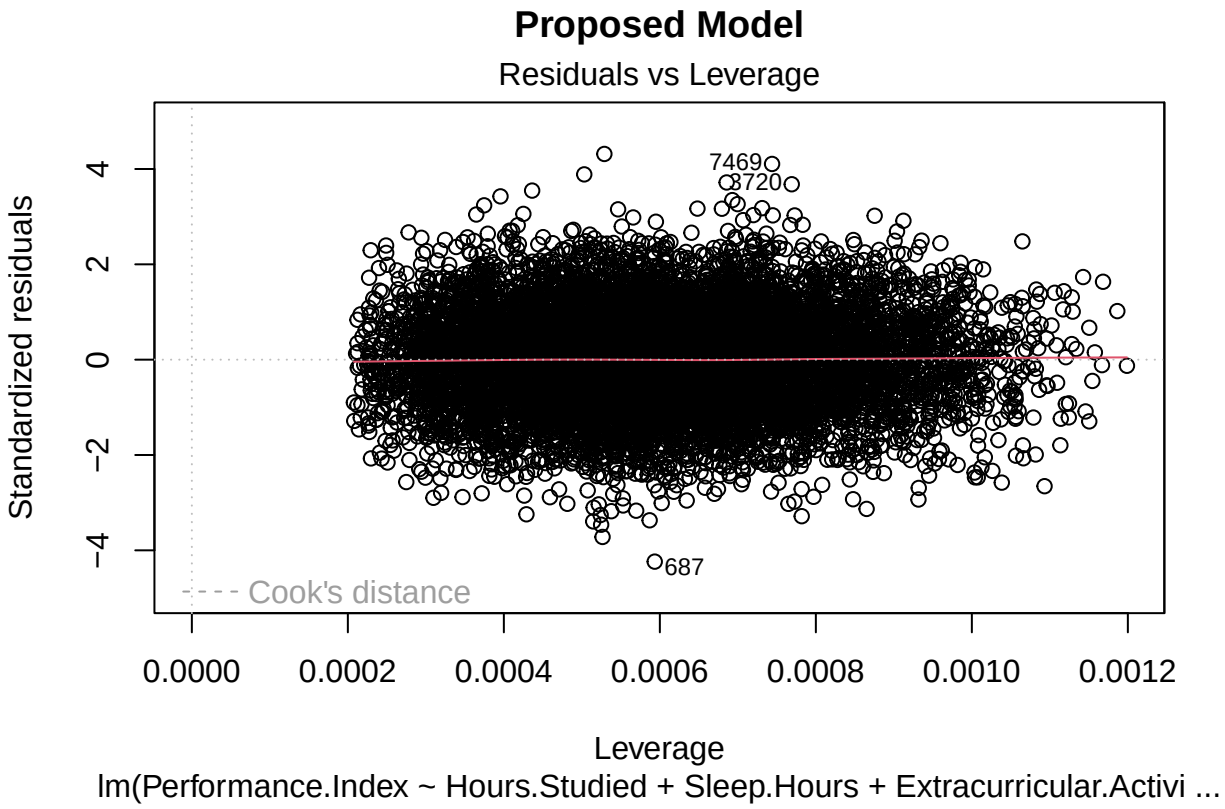
5 Appendix

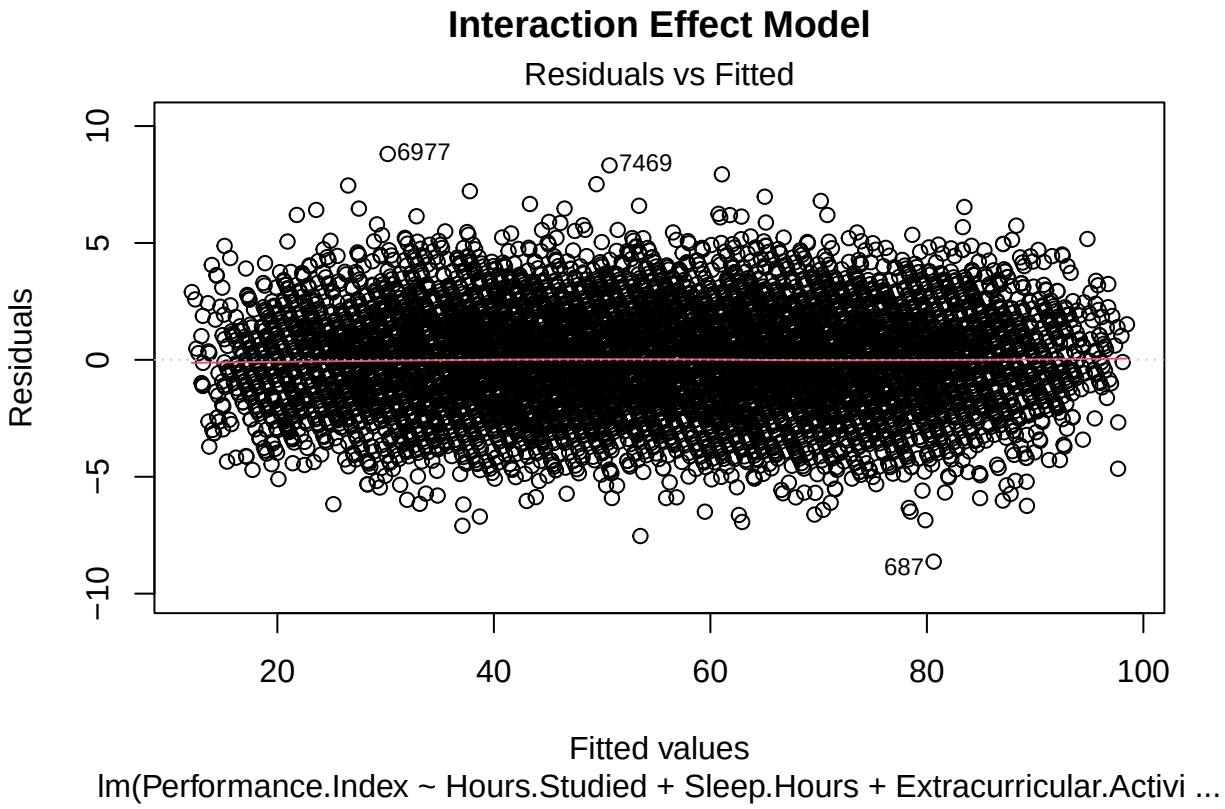

Boxplot of Sleep Hours by Extracurricular Activities
Sleep Hours
9
8
7
6
5
4
No
Yes
Extracurricular Activities
Extracurricular Activities
No
Yes

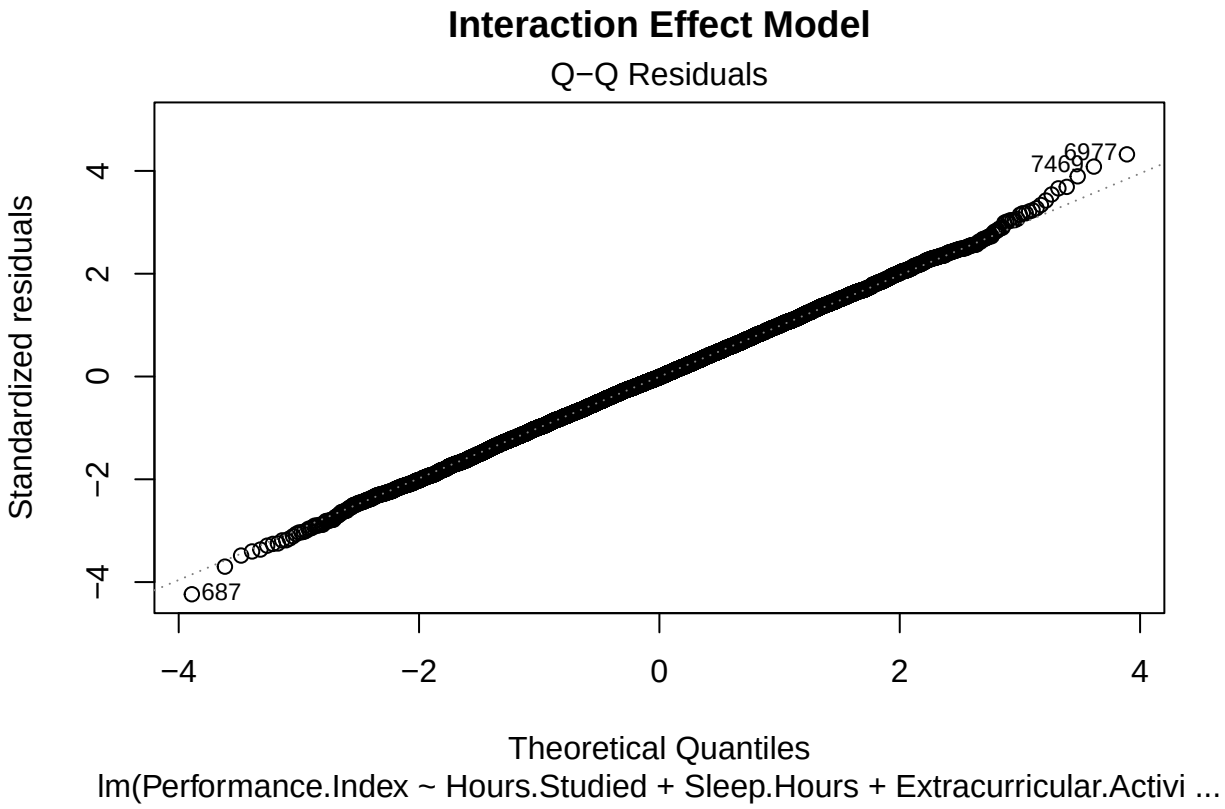

Proposed Model
Residuals vs Fitted
Residuals
Fitted values
lm(Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activi ...
6977
7469
687
10
5
0
−5
−10
20
40
60
80
100

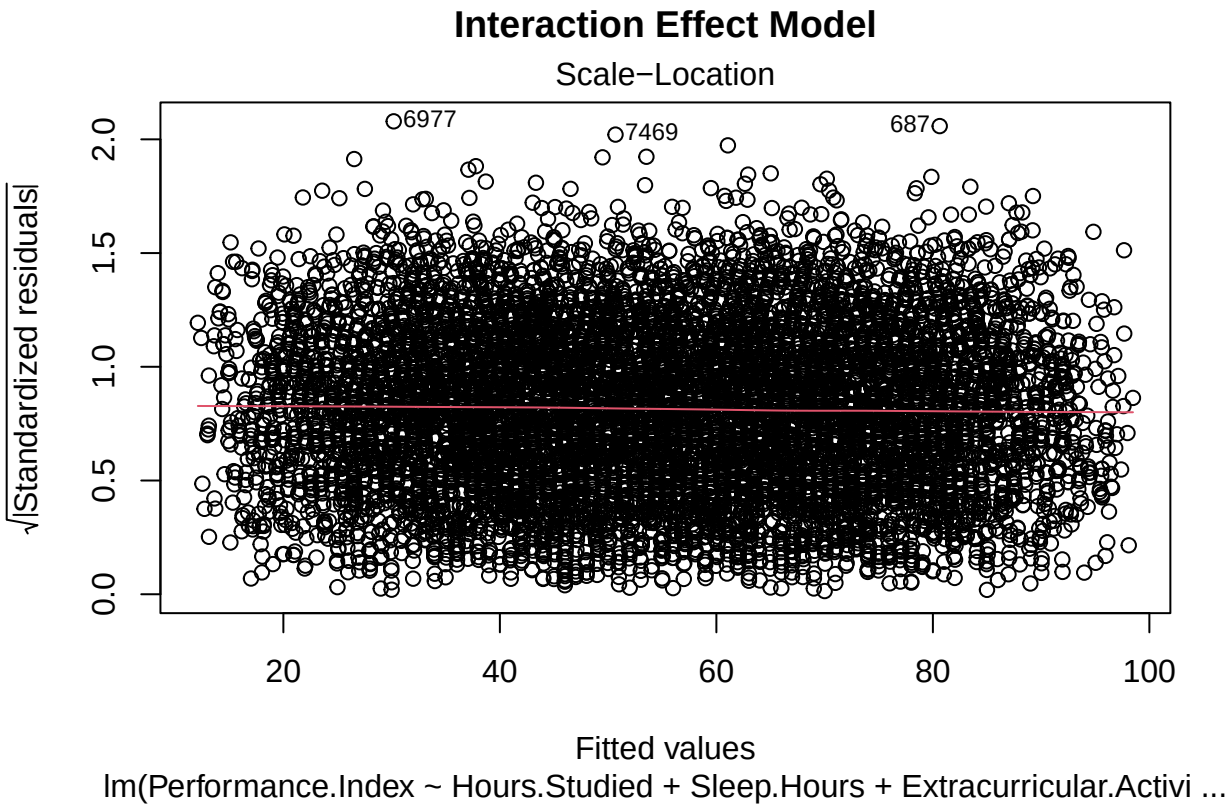

Proposed Model
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activi ...
6977
7469
687
−4
−2
0
2
4

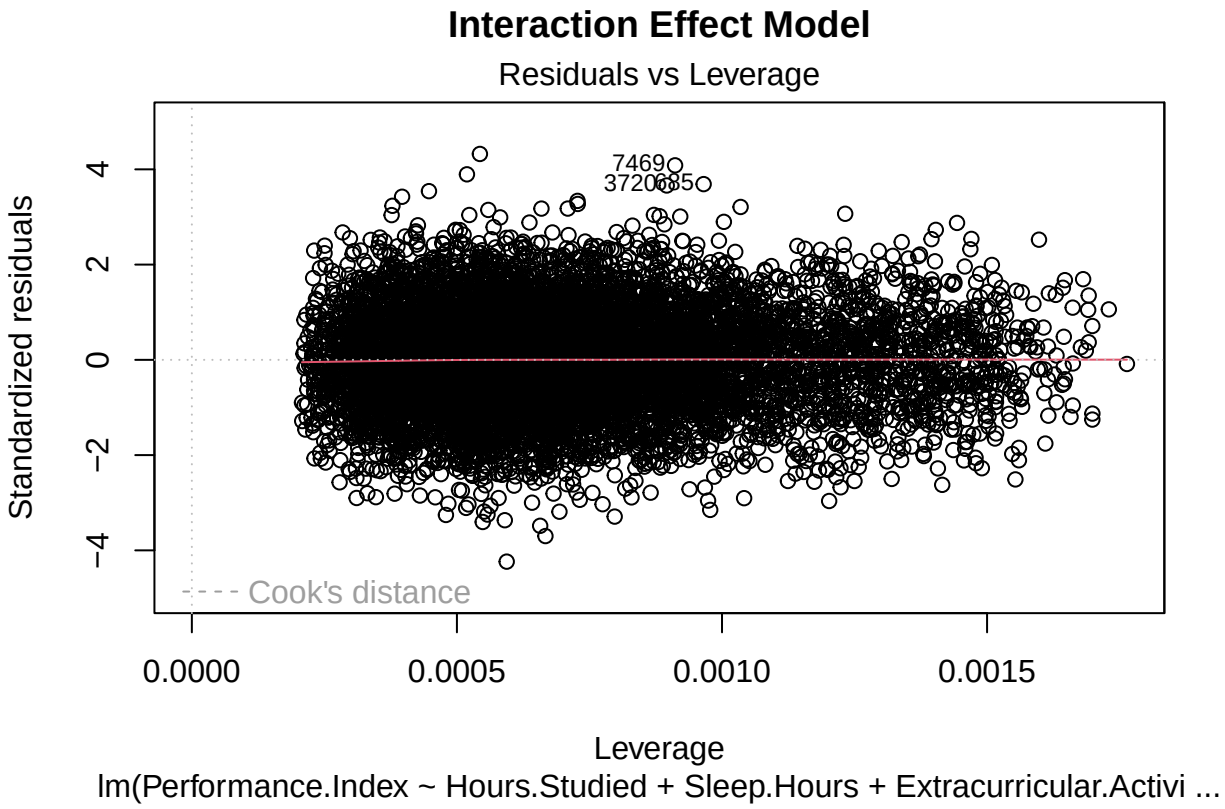

Proposed Model
Scale−Location
6977
7469
687
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
20
40
60
80
100
Fitted values
lm(Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activi ...

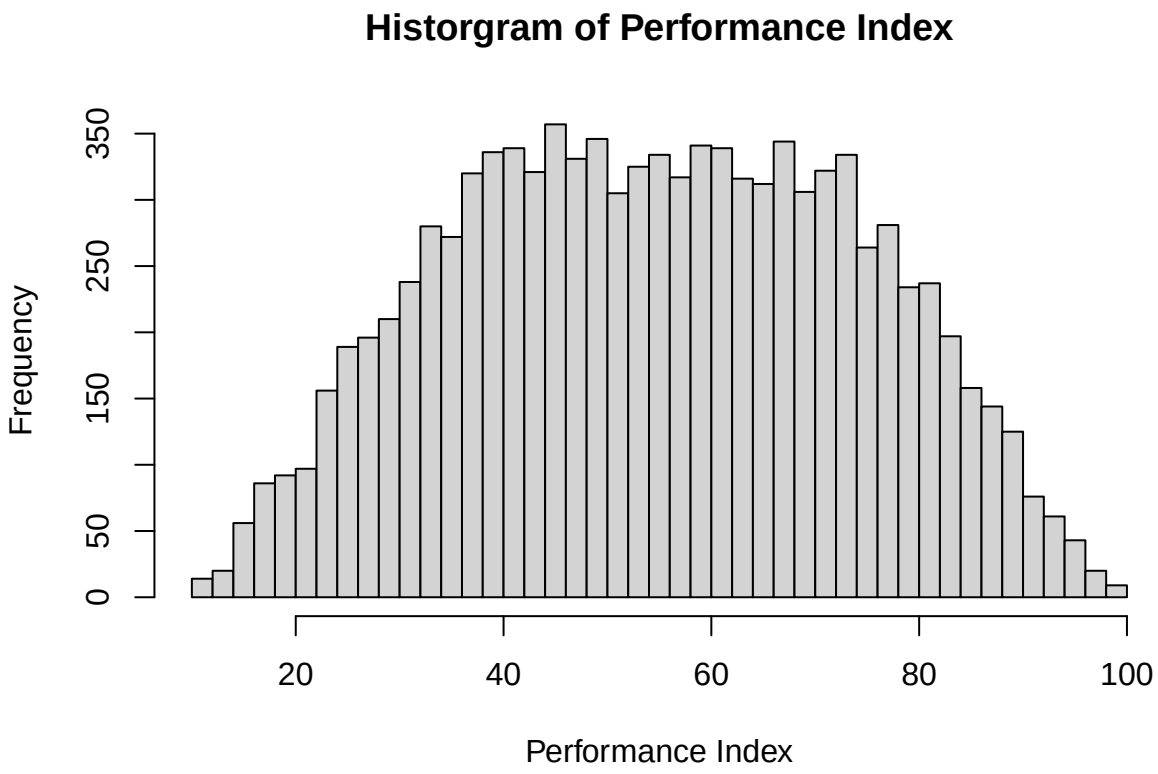

Proposed Model
Residuals vs Leverage
Standardized residuals
4
2
0
−2
−4
7469
8720
687
Cook's distance
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
Leverage
lm(Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activi ...

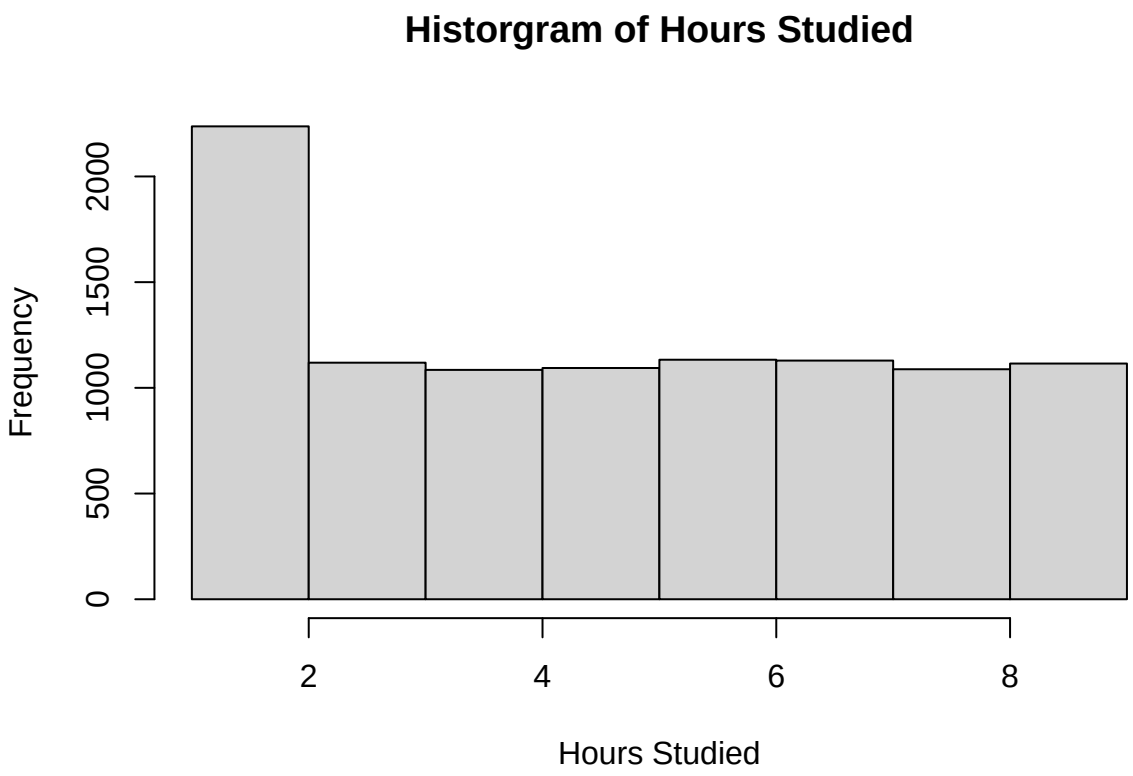

Interaction Effect Model
Residuals vs Fitted
Residuals
Fitted values
lm(Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activi ...
6977
7469
687
10
5
0
−5
−10
20
40
60
80
100


Interaction Effect Model
Q−Q Residuals
Standardized residuals
4
2
0
−2
−4
6977
7469
687
−4
−2
0
2
4
Theoretical Quantiles
lm(Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activi ...

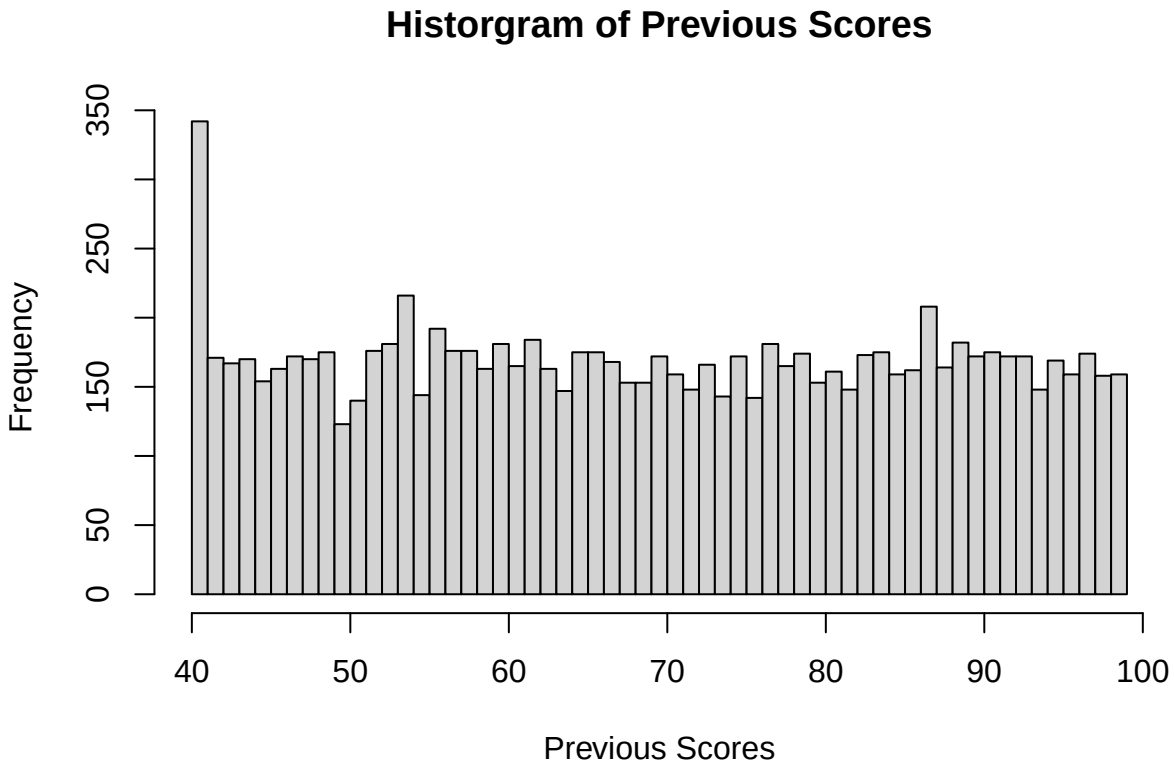

Interaction Effect Model
Scale−Location
6977
7469
687
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
20
40
60
80
100
Fitted values
lm(Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activi ...

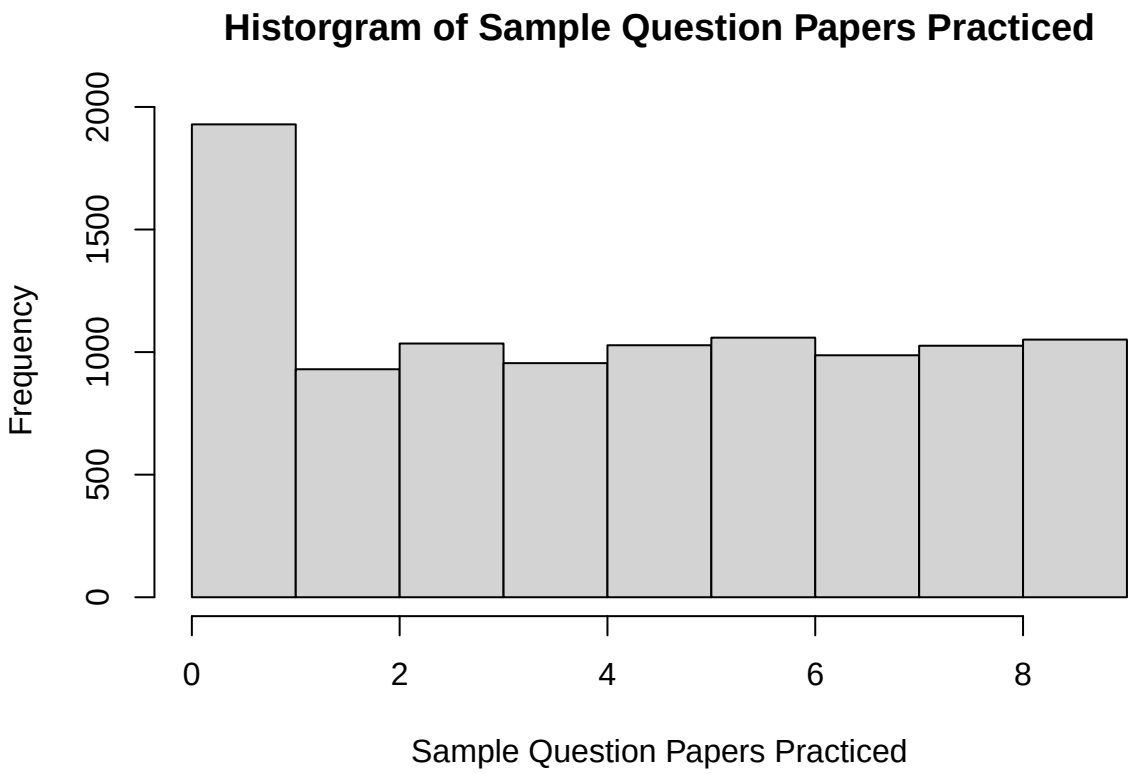

Interaction Effect Model
Residuals vs Leverage
Standardized residuals
4
2
0
−2
−4
7469
3720685
Cook's distance
0.0000
0.0005
0.0010
0.0015
Leverage
lm(Performance.Index ~ Hours.Studied + Sleep.Hours + Extracurricular.Activi ...


Historgram of Performance Index
Frequency
0
50
150
250
350
20
40
60
80
100
Performance Index


Historgram of Hours Studied
Frequency
0
500
1000
1500
2000
2
4
6
8
Hours Studied


Historgram of Sleep Hours
Frequency
0
500
1500
2500
4
5
6
7
8
9
Sleep Hours


Historgram of Previous Scores
Frequency
0
50
150
250
350
40
50
60
70
80
90
100
Previous Scores



6 References

Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. *IEEE Access*, *11*. https://doi.org/10.1109/ACCESS.2023.3250702

O., O., & P., C. (2017). Predicting students' academic performances – a learning analytics approach using multiple linear regression. *International Journal of Computer Applications*, *157*. https://doi.org/10.5120/ijca2017912671

Yang, S. J. H., Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Ogata, H., & Lin, A. J. Q. (2018). Predicting students' academic performance using multiple linear regression and principal component analysis. *Journal of Information Processing*, *26*. https://doi.org/10.2197/ipsjjip.26.170