Factors Impacting Student Performance: An Applied Regression Analysis

Group 15: Anish Tiwari, Isabel Turner, Eshita Vani, and Shreya Vasant

Abstract

This project applies multiple linear regression to explore factors involved in a student's academic performance. Each team member created a unique research question and tested the significance of predictors on full and reduced models. Regression assumptions were validated with various diagnostic checks: including q-q plots, added variable plots, VIF tests, and Bruesch-Pagan tests. Additionally, all models were validated using K-fold cross validation. This analysis highlights how different variables contribute to academic outcomes and demonstrate the value of statistical modeling in educational research. Ultimately, it was found that a model containing all 5 parameters examined (hours studied, previous scores, extracurriculars, hours slept, and number of papers practiced) was best for predicting student performance.

Background

There are many different factors that impact student's performance at school. Performance at school is typically defined as the test scores or the grade accumulated. Some domains that play a role impacting performance include sleep habits, study habits, extracurricular activities, and previous scores. Students typically get 9 hours of sleep and there have been positive correlations with more sleep having a positive effect on performance, and lower or 'suboptimal' sleep leads to lower performance (Taras, 2005). Study habits also influence performance. Studying more does result in better results; however, this is due largely because more studying increases confidence of the material and reduces test anxiety (Yusefzadeh, 2019). The history of performance can also help to predict future performance results. Previous scores are indicative of habits and if there is consistency then there would be no change but if there is a change in habit then we can see the change from previous score to current scores (Ashenafi, 2015). Finally, extracurricular activities and their participation open up opportunities. These opportunities could lead to gaining skills that result in better performance (Wilson, 2009). All of these domains and factors have individual influences on performance. Our project aims to see if any have a more significant influence on performance. Additionally, we sought to determine an accurate model to predict student performance using the most predictive parameters.

Experiment Design

This study was an exploratory observational study. The selected dataset, entitled "Student Performance Dataset" is a set of data including 10,000 student records meant to examine the impact of several different parameters on student test performance extracted from Kaggle. Each student record contains information on hours studied, previous scores, extracurricular activities, sleep hours, sample test questions practiced, and performance index (Table 1).

Table 1. Variables utilized for analysis of student performance

Variable Identifier	Category	Name	Description
Y	Response Variable	Performance Index	Student academic performance expressed as a range between 10 and 100 with higher numbers indicating better performance
X1	Continuous Explanatory Variable	Hours Studied	Total hours studied by a student in preparation for test
X2	Continuous Explanatory Variable	Previous Scores	Average score by student on previous tests
X3	Categorical Explanatory Variable	Extracurricular Activities	Whether or not a student participates in extracurricular activities (Yes or No)
X4	Continuous Explanatory Variable	Sleep Hours	Average number of hours of sleep per day for a student
X5	Continuous Explanatory Variable	Sample Question Papers Practiced	Number of sample question papers practiced in preparation for test

Prior to any analysis, continuous explanatory variables (X1, X2, X4, and X5) and the response variable (Y) were normalized by utilizing the scale() function in R. This function automatically centers the data using the mean and standard deviation. The formula used to calculate normalized values for each variable was: $X' = \frac{(X-\overline{X})}{s}$ where $\overline{X}$=mean response for a particular explanatory variable and s = standard deviation for a particular explanatory variable. The categorical variable, X3, was not normalized, and was instead converted into a dummy variable where 1=Yes and 0=No.

The full model of the data where Y~X1+X2+X3+X4+X5 was assessed for normality, influential points, heteroscedasticity, and multicollinearity. To assess for influential points, influencePlot(full model) was used in R to create a graph of Y deviations based on studentized residuals and X deviations based on hat values with area of the circle of each point representing Cook's distance. Due to the low maximum Cook's distance in this dataset (0.002), it was decided that the model's influential points are not influential enough to warrant correction. The cutoff utilized was percentile greater than 50% based on F (6, 9994) distribution, which translates to a value of 0.8914. To assess normality, a Q-Q plot of residuals was created (Figure 1b). Because the residuals graphed follow a straight line with no deviations, the model follows a normal assumption, so no corrections are necessary.

Figure 1a. Influence plot of the full model for all research questions

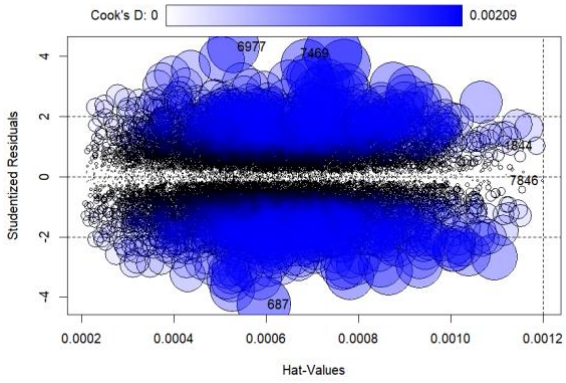
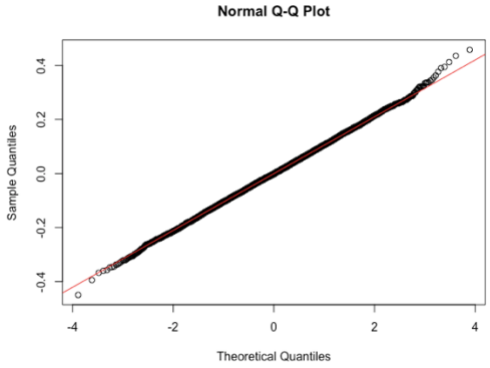
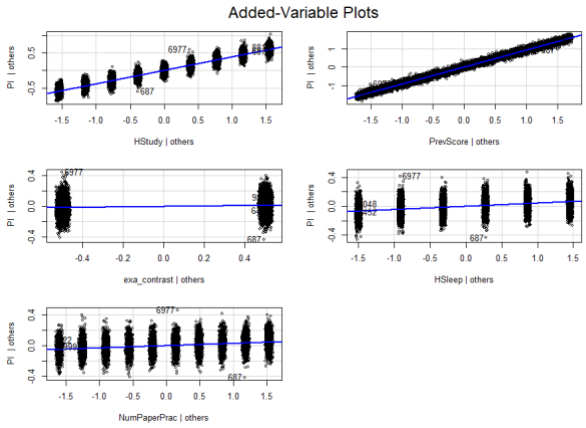


Figure 1b. Q-Q plot of residuals of full model



To assess multicollinearity, variance influence factor (VIF) was calculated for each variable in the model using vif(full model). All vif values were close to 1, so no multicollinearity was detected in this model. To assess heteroscedasticity, the Breusch-Pagan test was used. The BP test statistic was 2.1594 and the p-value was 0.8267. The p-value was greater than 0.05 and therefore the null hypothesis was not rejected, and assumption of heteroscedasticity was not violated and there is constant variance. Added-variable plots were used to test the validity of linear regression (Figure 1c). All variables showed linear or minimal effect given the other predictors in the full model, so linear regression was deemed to be a valid method of analysis. Finally, K-fold cross validation was used to assess model performance. RMSE was calculated to be 0.106 for the full model.

Figure 1c. Added-Variable plots for all variables in the full model



Research Questions

The overarching research question of this project was: what factors impact student test performance the most, and the goal was to find a model that best predicted student performance based on the factors provided. Four research questions were initially tested to see if any parameters could be combined or removed from the model.

1. Research question 1: how does time spent outside of studying impact a student's exam performance?
2. Research question 2: does prior knowledge impact future exam performance the same amount as current effort?
3. Research question 3: does sleep or study hours have the same effect on performance, or not (i.e which healthy habit is valuable)
4. Research question 4: do extra-curriculars affect how beneficial extra studying is on student performance?

Research Question 1

The first research question analyzed was: how does time spent outside of studying impact a student's exam performance? As many students manage sleep and participate in extracurriculars, it would be interesting to see how much of a factor both extracurricular activities and sleep together play a role in exam performance, if at all. To test this question, the impact of extracurricular participation and sleep hours on exam performance was tested in a model already containing hours studied, previous scores, and sample question papers practiced. The null hypothesis tested was $\beta3=\beta4=0$ to test if both extracurricular participation and sleep hours significantly improve the model's ability to explain exam performance. The alternatve hypothesis is that either or both β do not equal zero. The reduced model used is Performance Index = $\beta0 + \beta1$(hours studied) + $\beta2$(previous scores) + $\beta5$(sample question papers practiced) + ϵ. The variables of interest, extracurriculars and sleep hours, are excluded in the reduced model so that this model represents the scenario described by the null hypothesis $\beta3=\beta4=0$ while the

full model is the original full model described above. This allows the F-test to determine if the data provides significant evidence against the null hypothesis.

A Q-Q plot was used to check normality (Figure 2a), with residuals demonstrating a linear pattern. This indicates that the reduced model follows a normal distribution. An assessment of influence using Cook's distance for the reduced model found no observations exceeding the f-test specific cutoff of qf(0.5, 4, 9996) ≈ 0.961. An influence plot was used to visualize this assessment (Figure 2b). The results suggest a lack of exceptionally influential points according to this specific metric. Homoscedasticity for the reduced model was assessed using the Breusch-Pagan test. The test yielded a non-significant result as the p-value (0.942) is greater than the alpha (0.05). This suggests that the assumption of homoscedasticity is satisfied. To assess potential multicollinearity among the predictors in the reduced model, VIF were calculated, with all values being close to 1 (Table 2a).

Figure 2a. Q-Q plot of residuals for reduced model of research question 1

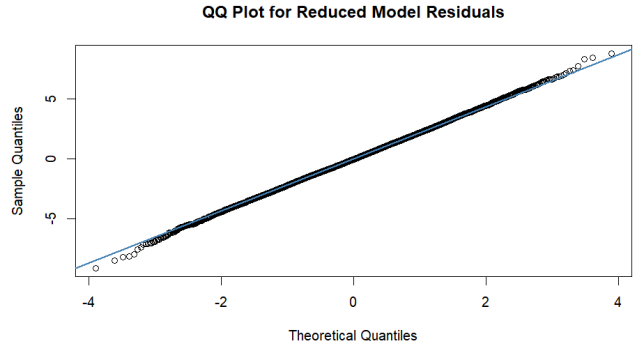


Figure 2b. Influence plot for reduced model of research question 1

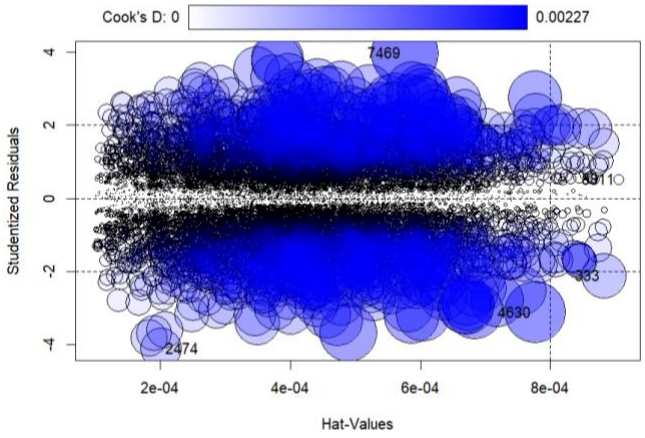


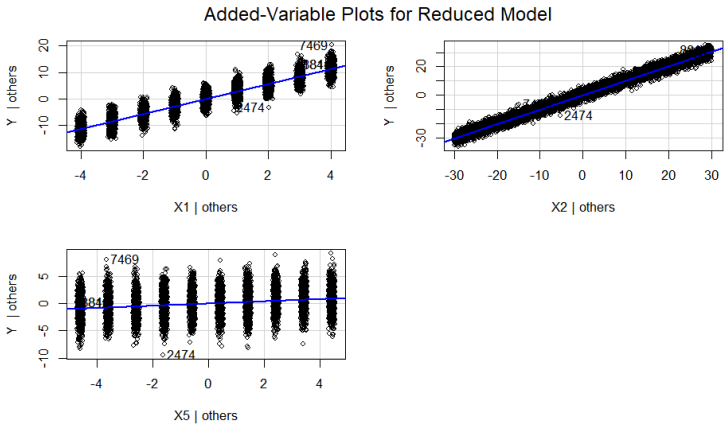
Table 2a. Variance Influence Factor of Variables in Full and Reduced Models for Research Question 1

Variable Name	Variable Symbol	VIF (Full)	VIF (Reduced)
Hours Studied	X1	1.00	1.00

Previous Scores	X2	1.00	1.00
Extracurricular Activities	X3	1.00	N/A
Hours Slept	X4	1.00	N/A
Practice Papers	X5	1.00	1.00

The resulting average RMSE after K-fold cross validation was 0.115. This value suggests an expected average prediction error of approximately 0.115 units of the Performance Index for this model. This RMSE indicates that the reduced model possesses strong predictive accuracy relative to the overall variability in the outcome. Finally, the added-variable plots were generated for each predictor in the reduced model. Visually figure 2c below shows clear linear trends passing through the center of each predictor. This supports the assumption that the partial relationship between exam performance and each of the predictors is modeled as linear.

Figure 2c. Added variable plots for all predictors in the reduced model of research question 2



A partial F-test was done using anova on the reduced model and full model. The test resulted in a highly significant F-statistic ($p < 2.2e\text{-}16$) much less than the threshold of $a=0.05$. Therefore, we reject the null hypothesis that both $\beta 3$ (Extracurricular participation) and $\beta 4$ (Hours slept) are 0. In conclusion, adding information about extracurricular participation and hours slept significantly improves the model's ability to explain exam performance, taking the other variables as a control.

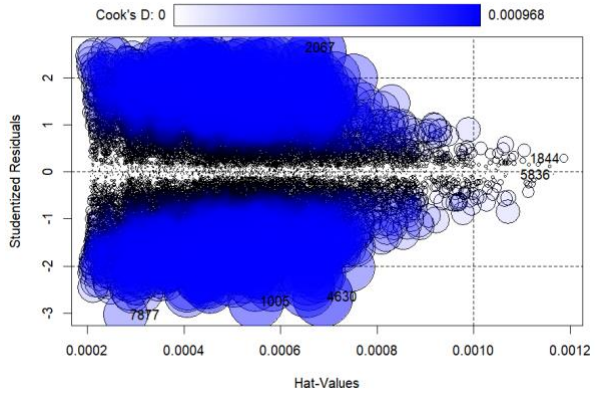
Research question 2

The second research question examined was: does prior knowledge impact future exam performance the same amount as current effort? This question was selected because many people believe that prior knowledge (tested here by previous scores on exam) plays a greater or equivalent role in future performance compared to current effort. Because of this, people who performed poorly in the past will often stop trying or give up instead of putting in more effort. If hours studied and previous scores are equivalent in their impact on future performance, then that gives this theory some truth. However, if these parameters have different impacts on student performance, then the theory is likely incorrect.

To test this question, the impact of hours studied and previous scores on exam performance was compared in a model already containing extracurricular activities, sleep hours, and sample question papers practiced. The null hypothesis tested was that hours studied, and previous scores have an equivalent impact on student performance ($\beta_1 = \beta_2 = \beta_{new}$). The alternative hypothesis tested was that hours studied, and previous scores do not have an equivalent impact on student performance ($\beta_1 \neq \beta_2$). The full model used to represent the alternative hypothesis was Performance Index = β0+ β1(Hours Studied) + β2(Previous Scores) + β3(Extracurricular Activities) +β4(Hours Slept) + β5(Practice Papers) + ε . The reduced model used to represent the null hypothesis was Performance Index = β0 + βcombined(Hours Studied + Previous Scores) + β3(Extracurricular Activities) + β4(Hours Slept) + β5(Practice Papers) + ε .

Prior to comparing ANOVA for the full and reduced models listed above, both models were checked for influential points visually using influence plots that graph Y deviations based on studentized residuals and X deviations based on hat values with area of the circle of each point representing Cook's distance. The code used in R was influencePlot(full model) (Figure 1a) or influencePlot(reduced model) (Figure 3a). Because of the low Cook's distance values previously determined for the full model and the low Cook's distance values determined here for the reduced model for this research question, it was decided that any outliers in the models aren't influential enough to require correction. The cutoff utilized was percentile greater than 50% based on F (5, 9995) distribution, which translates to a value of 0.8703.

Figure 3a. Influence plot of reduced model for research question 2



The presence of multicollinearity was evaluated using variance influence factor (VIF) calculated for all the variables in each model using vif(full model) or vif(reduced model) in R. The values are presented in Table 3a and because all VIF values are close to 1, it was concluded that no multicollinearity issue exists in either model. Both models were checked for heteroscedasticity using a Breusch-Pagan test. The results for the full model were discussed above, and the reduced model had a BP test statistic of 2.8672, corresponding to a p-value of 0.5803. Because the p-value was greater than 0.05, the null hypothesis was not rejected and the assumption of heteroscedasticity was not violated. Additionally, added variable plots were used to assess the

validity of linear regression assumptions. The full model was checked above, and the reduced model for this research question was checked as well (Figure 3b). All variables in the full and reduced models showed a linear or no relationship between the marginal effect of X and Y. This demonstrates that a linear model is appropriate. Models were checked for normality using Q-Q plots, the full model was assessed above and found to be normal, but the reduced model showed some variation from normality (Figure 3c), however, with the large sample size of 10,000, the model is close enough to normality to be used.

Table 3a. Variance Influence Factor of Variables in Full and Reduced Models for Research Question 2

Variable Name	Variable Symbol	VIF (Full)	VIF (Reduced)
Hours Studied	X1	1.00	N/A
Previous Scores	X2	1.00	N/A
(Hours Studied + Previous Scores)	(X1+X2)	N/A	1.00
Extracurricular Activities	X3	1.00	1.00
Hours Slept	X4	1.00	1.00
Practice Papers	X5	1.00	1.00

Figure 3b. Reduced model added-variable plots

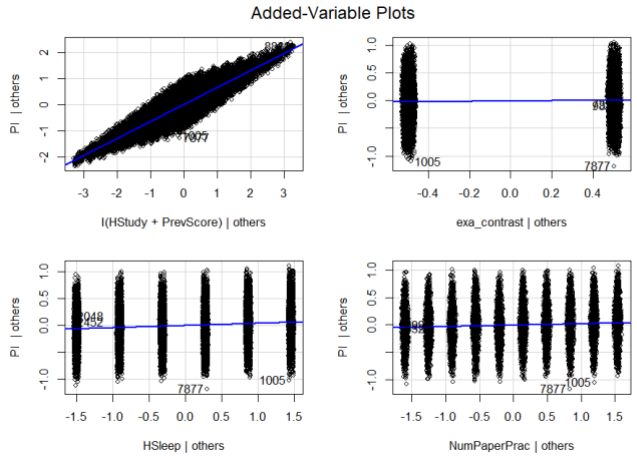
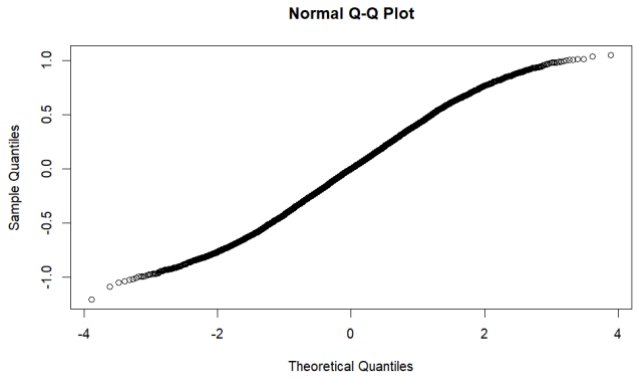


Figure 3c. Q-Q plot for reduced model



Upon finding no issues with model fit or assumption violations, ANOVA was used to compare the full and reduced models with the code anova(reduced, full). The f-value for the comparison between the reduced and full models was 128,650 and the p-value was <0.001. Because of this, the null hypothesis that hours studied and previous scores have an equivalent impact on performance index was rejected. Hours studied and previous scores have different impacts on performance index, so the model cannot be simplified to include one β for both variables. Finally, K-fold cross validation was utilized to assess model performance. RMSE was found to be 0.395, which is higher than the RMSE for the full model, so the reduced model performance worse than the full model.

Research Question 3

The third research question focused on hours of sleep and hours of study and their effects on performance index. We decided to research this because many people in college have a tendency to prioritize hours of study over hours of sleep the night before a test. By researching this question, we would then be able to see if one does truly have a benefit over the other or if there is equal effect of both sleep and study on performance index. The answer to this question could impact the decisions students make in how they prepare and review for a test and their personal habits prior to testing. We compared the impact of hours of study and hours of sleep on exam performance with a model that had all other predictors in our data set (sample practice papers completed, extracurricular activities participation, and previous scores). The null hypothesis tested if hours of study and sleep have an equivalent effect on performance index ($\beta_1=\beta_4=\beta_{combined}$), while the alternative hypothesis tested whether hours of study and sleep do not have an equivalent effect on performance index ($\beta_1\neq\beta_4$). The full model was used to represent the alternative hypothesis: Performance Index = $\beta 0 + \beta 1$(Hours Studied) + $\beta 2$(Previous Scores) + $\beta 3$(Extracurricular Activities) + $\beta 4$(Hours Slept) + $\beta 5$(Sample Papers Practiced) + ε .
The reduced was model used to represent the null hypothesis: Performance Index = $\beta 0$ + βcombined(Hours Study + Hours Slept) + $\beta 2$(Previous Scores) + $\beta 3$(Extracurricular Activities) +$\beta 5$(Sample Papers Practiced) + ε .

Prior to comparing ANOVA for the full and reduced models listed above, both models were checked for influential points visually using influence plots that graph Y deviations based on studentized residuals and X deviations based on hat values with area of the circle of each point representing Cook's distance. The code used in R was influencePlot(full model) (Figure 1a) or influencePlot(reduced model) (Figure 4a). Because of the low Cook's distance values previously determined for the full model and the low Cook's distance values determined here for the reduced model for this research question, it was decided that any outliers in the models aren't influential enough to require correction. The cutoff utilized was percentile greater than 50% based on F (6, 9994) distribution, which translates to a value of 0.8914. The presence of multicollinearity was evaluated using variance influence factor (VIF) calculated for all the variables in each model using vif(full model) or vif(reduced model) in R. The values are presented in Table 4a and because all VIF values are close to 1 (rounded to two decimal points), it was concluded that no multicollinearity issue exists in either model. Both models were checked for heteroscedasticity using a Breusch-Pagan test. The results for the full model were discussed above, and the reduced model had a BP test statistic of 1.743, corresponding to a p-value of 0.783. Because the p-value was greater than 0.05, the null hypothesis was not rejected, and the assumption of heteroscedasticity was not violated and there is constant variance. Additionally, added variable plots were used to assess the validity of linear regression assumptions. The full model was checked above, and the reduced model for this research question was checked as well (Figure 4b). All variables in the full and reduced models showed a linear or no relationship between the marginal effect of X and Y. This demonstrates that a linear model is appropriate. Models were checked for normality using Q-Q plots, the full model was assessed above and found to be normal, but the reduced model showed some variation from normality (Figure 4c), however, with the large sample size of 10,000, the model is close enough to normality to be used.

Figure 4a. Influence plot of reduced model for research question 3

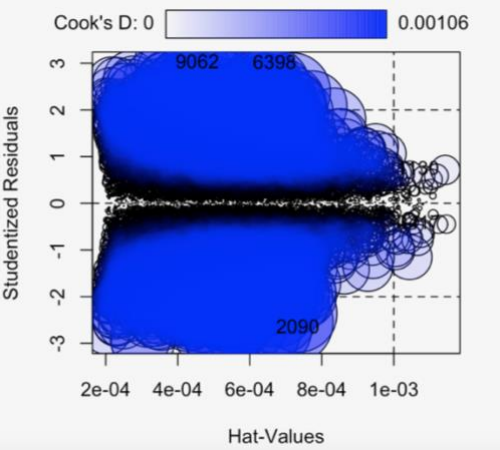


Table 4a. Variance Influence Factor of Variables in Full and Reduced Models for Research Question 3

Variable Name	Variable Symbol	VIF (Full)	VIF (Reduced)
Hours Studied	X1	1.00	N/A
Previous Scores	X2	1.00	1.00
Extracurricular Activities	X3	1.00	1.00
Hours Slept	X4	1.00	N/A
Hours Studied + Hours Slept	X1+X4	N/A	1.00
Practice Papers	X5	1.00	1.00

Figure 4b. Reduced model added-variable plots

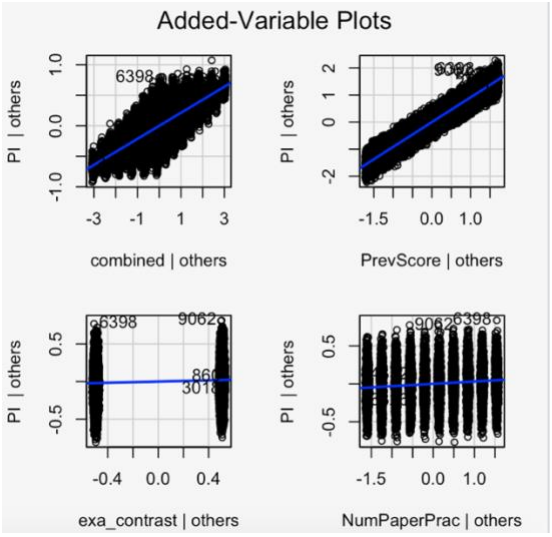
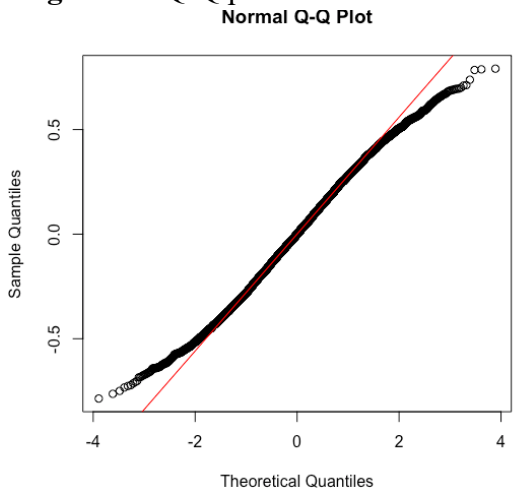


Figure 4c. Q-Q plot for reduced model



Upon finding no issues with model fit or assumption violations, ANOVA was used to compare the full and reduced models with the code anova(reduced, full). The f-value for the comparison between the reduced and full models was 51,891 and the p-value was <2.2e-16. Because of this, the null hypothesis that hours studied, and hours slept have an equivalent impact on performance index was rejected. Hours studied and hours slept have different impacts on performance index, so the model cannot be simplified to include one β for both variables. Finally, K-fold cross

validation was utilized to assess model performance. RMSE was found to be 0.264, which is higher than the RMSE for the full model, so the reduced model performed worse than the full model.

Research Question 4

The fourth research question: focuses on whether participation in extra-curricular activities improves or changes the linear impact of study hours on student's academic performance. Prime motivation behind this question lies in the fact that extra-curricular activities are often encouraged throughout school and college as they offer cognitive, emotional, and social development that can complement academic efforts and boost the performance. We hypothesize that these activities do not only improve performance on their own but also determine how effective study hours are. Students could study more depending on whether they are involved in such activities. To test this, we introduced an interaction term between study hours and extracurricular participation in a model which also controls for prior academic performance, sleep hours and number of sample papers practiced. The null hypothesis (reduced model) was there is no significant interaction between study hours and extracurricular activities on performance: Performance Index = $\beta 0 + \beta 1$(Hours Studied) + $\beta 2$(Previous Scores) + $\beta 3$(Extracurricular Activities) + $\beta 4$(Hours Slept) + $\beta 5$(Question Papers Practiced) + $\beta 6$(Hours Studied * Extracurricular Activities) + ε . The alternative hypothesis (full model) was that a significant interaction exists: Performance Index = $\beta 0 + \beta 1$(Hours Studied) + $\beta 2$(Previous Scores) + $\beta 3$(Extracurricular Activities) + $\beta 4$(Hours Slept) + $\beta 5$(Question Papers Practiced) + ε .

Prior to comparing ANOVA for the full and reduced models listed above, both models were checked for influential points visually using influence plots that graph Y deviations based on studentized residuals and X deviations based on hat values with area of the circle of each point representing Cook's distance. The code used in R was influencePlot(full model) (Figure 1a) or influencePlot(reduced_model) (Figure 5a). The cutoff utilized was percentile greater than 50% based on F (7, 9993) distribution, which translates to a value of 0.9014. Since the maximum Cook's Distance observed in the reduced model was only 0.00209, no data points were considered influential enough to remove. The presence of multicollinearity was evaluated using VIF calculated for all the variables in each model using vif(full model) or vif(reduced model) in R. The values are presented in Table 5a and because all VIF values are close to 1 (rounded to two decimal points), it was concluded that no multicollinearity issue exists in either model. Both models were checked for heteroscedasticity using a Breusch-Pagan test. The results for the reduced model were discussed above, and the full model for this research question had a BP test statistic of 2.375, corresponding to a p-value of 0.8821. Because the p-value was greater than 0.05, the null hypothesis was not rejected, and the assumption of heteroscedasticity was not violated and there is constant variance. Additionally, added variable plots were used to assess the validity of linear regression assumptions. The full model was checked above, and the reduced model for this research question was checked as well (Figure 5b). All variables in the full and reduced models showed a linear or no relationship between the marginal effect of X and Y. This

demonstrates that a linear model is appropriate. Models were checked for normality using Q-Q plots, the full model was assessed above and found to be normal, but the reduced model showed some variation from normality (Figure 5c), however, with the large sample size of 10,000, the model is close enough to normality to be used.

Figure 5a. Influence plot of full model for research question 4

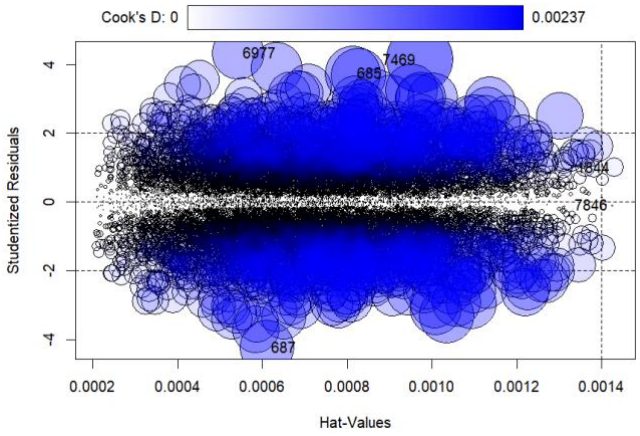


Table 5a. Variance Influence Factor of Variables in Full and Reduced Models for Research Question 3

Variable Name	Variable Symbol	VIF (Full)	VIF (Reduced)
Hours Studied	X1	1.96	1.00
Previous Scores	X2	1.00	1.00
Extracurricular Activities	X3	1.00	1.00
Hours Slept	X4	1.00	1.00
Practice Papers	X5	1.00	1.00
Interaction (Hstudied x Extracurricular)	X1 x X3	1.96	N/A

Figure 5b. Full model added-variable plots for research question 4

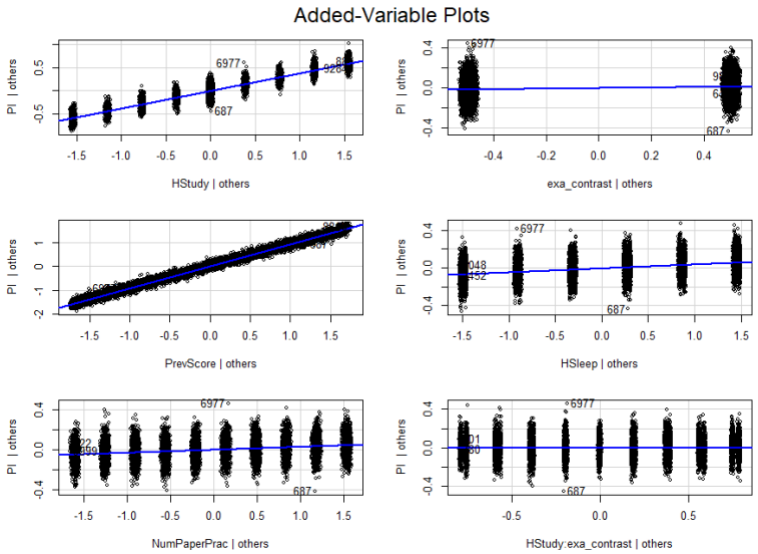
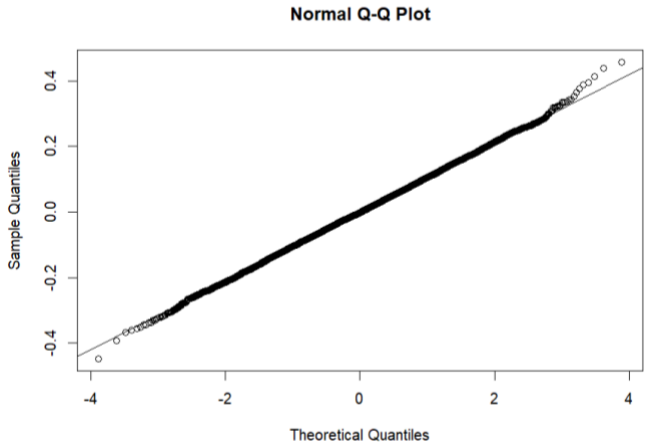


Figure 5c. Q-Q plot for full model for research question 4



Upon finding no issues with model fit or assumption violations, ANOVA was used to compare the full and reduced models with the code anova(reduced, full). The f-statistic for the comparison between the reduced and full models was 2.3164 and the p-value was 0.128. Because of this, the null hypothesis that there is no significant interaction between study hours and extra-curriculars was not rejected. Finally, K-fold cross validation was utilized to assess model performance. RMSE was found to be 0.1061, which is roughly equal to the RMSE for the full model, so the reduced model performed better than the full model.

Discussion and Conclusion

Overall, the results from each of the research questions explored indicate that none of the reduced models are better than the full model selected. As one final step to determine what model best fits the data, and to determine what parameters are necessary to best predict student performance, a best subset algorithm was ran using the BestSub function in R. The results from

this test confirmed that the model with all parameters is the best model for this data (Table 6a). The results from research question 4 indicate that the interaction between extracurriculars and hours studied should not be included in the model.

Table 6a. Best Subset Algorithm

	p	1	2	3	4	5	SSEp	r2	r2 adjusted	Cp	AICp	SBCp	PRESSp
1	2	0	1	0	0	0	1624.13	0.838	0.838	134328.48	-18172.15	-18157.73	1624.78
2	3	1	1	0	0	0	141.26	0.986	0.986	2558.91	-42591.27	-42569.64	141.34
3	4	1	1	1	0	0	123.49	0.988	0.988	981.67	-43933.81	-43904.97	123.58
4	5	1	1	1	1	0	115.01	0.988	0.988	229.88	-44643.45	-44607.4	115.12
5	6	1	1	1	1	1	112.47	0.989	0.989	6.00	-44864.94	-44821.68	112.60

As k-fold validation was already run with this model, it was not necessary to examine again. The final step was getting an estimate of intercept and parameters for the model, which was done by running the summary() code in R. The equation for the model utilizing normalized data was PI=(-0.016) + 0.385(HStudy) + 0.919(PrevScore) + 0.032(exa_contrast) + 0.042(HSleep) + 0.029(NumPaperPrac). The coefficients represent the effect of one standard deviation increase in the parameter on performance index, with the coefficient 0.385 indicating that for every one standard deviation increase in hours studied, the performance index increased by 0.385 standard deviations. To better interpret this model, the coefficients and intercept were converted back to original units. The intercept was converted to original performance index by multiplying it by the standard deviation of the original data and adding that to the mean performance index. The coefficients were transformed by multiplying by the standard deviation of the performance index and then dividing by the standard deviation of the respective parameter. The coefficient for extracurricular activities was not transformed because extracurricular activities is a categorical variable, so the coefficient represents the difference between 1(Yes) and 0 (No). The final equation used to represent this data was Performance Index = 54.92 + 2.86(Hours Studied) + 1.02(Previous Scores) + 0.03(Extracurricular Activities) + 0.48(Sleep Hours) + 0.33(Number of practice papers). Overall, we have shown that all parameters in the model are significant and must be included. Hours studied, previous scores on exams, whether or not a student participates in extracurricular activities, hours of sleep, and number of practice papers done all contribute to the student's performance on an exam. Using this model, we can predict a student's performance given these factors.

Appendix

Works Cited

Ashenafi, Michael Mogessie, et al. "Predicting Students' Final Exam Scores from Their Course Activities." 2015 IEEE Frontiers in Education Conference (FIE), Oct. 2015, https://doi.org/10.1109/fie.2015.7344081. Accessed 10 May 2023.

Taras, Howard, and William Potts-Datema. "Sleep and Student Performance at School." Journal of School Health, vol. 75, no. 7, Sept. 2005, pp. 248–254, https://doi.org/10.1111/j.1746-1561.2005.tb06685.x.

Wilson, Nikki. "Impact of Extracurricular Activities on Students." May 2009.

Yusefzadeh, Hasan, et al. "The Effect of Study Preparation on Test Anxiety and Performance: A Quasi-Experimental Study." Advances in Medical Education and Practice, vol. Volume 10, no. 10, May 2019, pp. 245–251, https://doi.org/10.2147/amep.s192053.

Data Source

The dataset used in this project was obtained from Kaggle.
Dataset Title: *Student Performance Prediction Dataset*
Retrieved from: https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression

Relevant R output

Full Model Model Summary

```
Call:
lm(formula = PI ~ HStudy + PrevScore + exa_contrast + HSleep +
    NumPaperPrac, data = Final)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44935 -0.07123 -0.00162  0.07056  0.45768

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.015785   0.001493  -10.57   <2e-16 ***
HStudy        0.384501   0.001061  362.35   <2e-16 ***
PrevScore     0.919339   0.001061  866.45   <2e-16 ***
exa_contrast  0.031901   0.002123   15.03   <2e-16 ***
HSleep        0.042418   0.001061   39.97   <2e-16 ***
NumPaperPrac  0.028924   0.001061   27.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1061 on 9994 degrees of freedom
Multiple R-squared:  0.9888,   Adjusted R-squared:  0.9887
F-statistic: 1.757e+05 on 5 and 9994 DF,  p-value: < 2.2e-16
```

Research question 1 full and reduced model ANOVA

```
Analysis of Variance Table

Model 1: PI ~ HStudy + PrevScore + NumPaperPrac
Model 2: PI ~ HStudy + PrevScore + exa_contrast + HSleep + NumPaperPrac
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   9996 132.68
2   9994 112.47  2    20.217 898.28 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Research question 2 full and reduced model ANOVA

```
Analysis of Variance Table

Model 1: PI ~ I(HStudy + PrevScore) + exa_contrast + HSleep + NumPaperPrac
Model 2: PI ~ HStudy + PrevScore + exa_contrast + HSleep + NumPaperPrac
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   9995 1560.20
2   9994  112.47  1    1447.7 128650 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Research question 3 full and reduced model ANOVA

```
Analysis of Variance Table

Model 1: PI ~ I(HStudy + HSleep) + PrevScore + exa_contrast + NumPaperPrac
Model 2: PI ~ HStudy + PrevScore + exa_contrast + HSleep + NumPaperPrac
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1   9995 696.41
2   9994 112.47  1    583.95 51891 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Research question 4 full and reduced model ANOVA

```
Analysis of Variance Table

Model 1: PI ~ HStudy + PrevScore + exa_contrast + HSleep + NumPaperPrac
Model 2: PI ~ HStudy * exa_contrast + HSleep + PrevScore + NumPaperPrac
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   9994 112.47
2   9993 112.44  1  0.026064 2.3164  0.128
```

Best subset algorithm

```
  p 1 2 3 4 5      SSEp        r2    r2.adj          Cp      AICp      SBCp
PRESSp
1 2 0 1 0 0 0 1624.1259 0.8375712 0.8375549 134328.4804 -18172.15 -18157.73
1624.7817
2 3 1 1 0 0 0  141.2616 0.9858724 0.9858696   2558.9129 -42591.27 -42569.64
141.3474
3 4 1 1 1 0 0  123.4899 0.9876498 0.9876461    981.6685 -43933.81 -43904.97
123.5894
4 5 1 1 1 1 0  115.0073 0.9884981 0.9884935    229.8763 -44643.45 -44607.40
115.1230
5 6 1 1 1 1 1  112.4654 0.9887523 0.9887467      6.0000 -44864.94 -44821.68
112.6010
```