Predicting Academic Success: A Regression Analysis of Student Performance Factors

Prepared by: Derrick Nyagesuka and Baysa Wakjira

Submitted to: Dr. Iresha Premarathna, STAT 311, Regression Analysis, Fall 2025

1. Introduction

The level of one's education and success is often seen as a determining factor as to whether that person is going to be a success in society or not. However, what many fail to realize is that every student's path is different and after leaving the classroom, other factors come to play which might affect the said student's academic success and excellence. These factors include the number of hours they study outside of class, parental involvement, access to resources, motivation level and sleep habits to name just a few. The main goal of this project is to see what factors are critical for a student's success. To do this, we obtained a dataset from Kaggle called **Student Performance Factors** and analyzed it and answered the question, **"What are the most important predictors of a student's final exam score, and what are the relationships between these predictors?"** Based on our predictors, we had both which show individual effort such as study hours and attendance, and environmental factors such as parental involvement. We conducted regression analysis using the dataset mentioned earlier which contained 6529 student records. At the end of this project, we wanted to build and validate a statistical model and see what exact variables lead to success. This could help students by telling them areas that needed improvement for better results and help schools allocate resources effectively to optimize performance.

2. Data Description

The data for this analysis was obtained from the **Student Performance Factors** dataset which was on Kaggle. The dataset had a comprehensive outlook of the life of a student, tallying up many variables which possibly have a hand in or influence academic success. The original dataset had 6607 observations with 20 predictors(variables). The data satisfied the standard assumptions required for ordinary multiple linear regression, and there were no issues with time-series or spatial autocorrelation because the observations don't have any time or geographic coordinates.

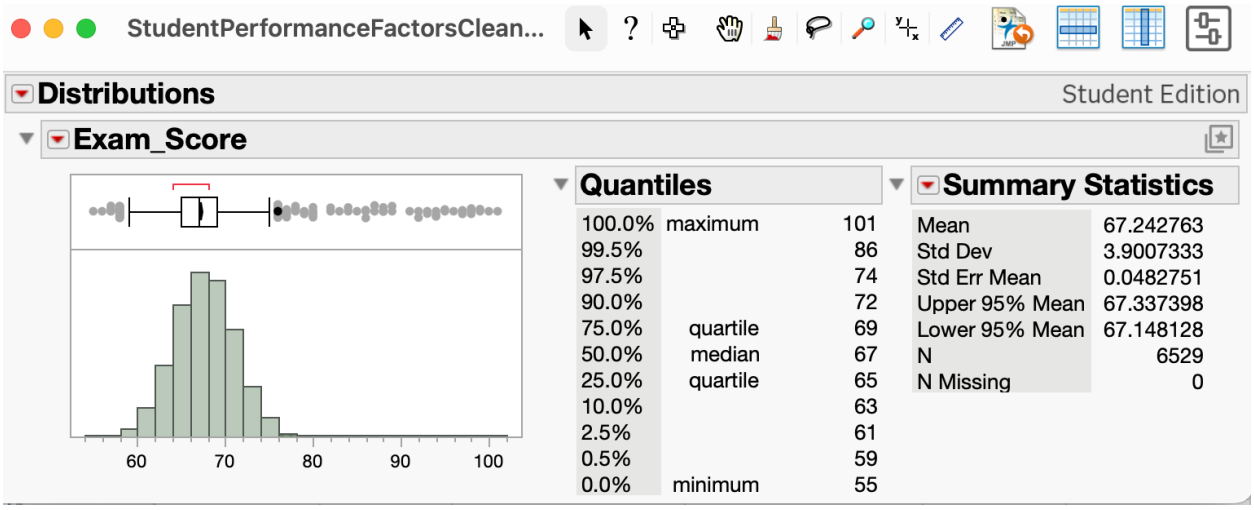
Our dependent variable is Exam_Score which represents the final score for each student, and it is a continuous numerical value. To predict this score, we had predictors grouped into 3 categories which sort of groups a student's life in 3 areas, namely:

- **Academic:** Predictors are Hours_Studied, Attendance, and Previous_Scores
- **Environmental:** Predictors are Parental_Involvement, Teacher_Quality, and Access_to_Resources.
- **Student mood:** Sleep_Hours and Motivation_Level

Before the start of the analysis, we checked to see if the data had any missing values and noticed that the Teacher_Quality predictor had 78 null values. Since 78 is a very small value in 6607 observations, we just did listwise deletion and our final dataset contained 6529 observations, which would still help us do regression analysis.

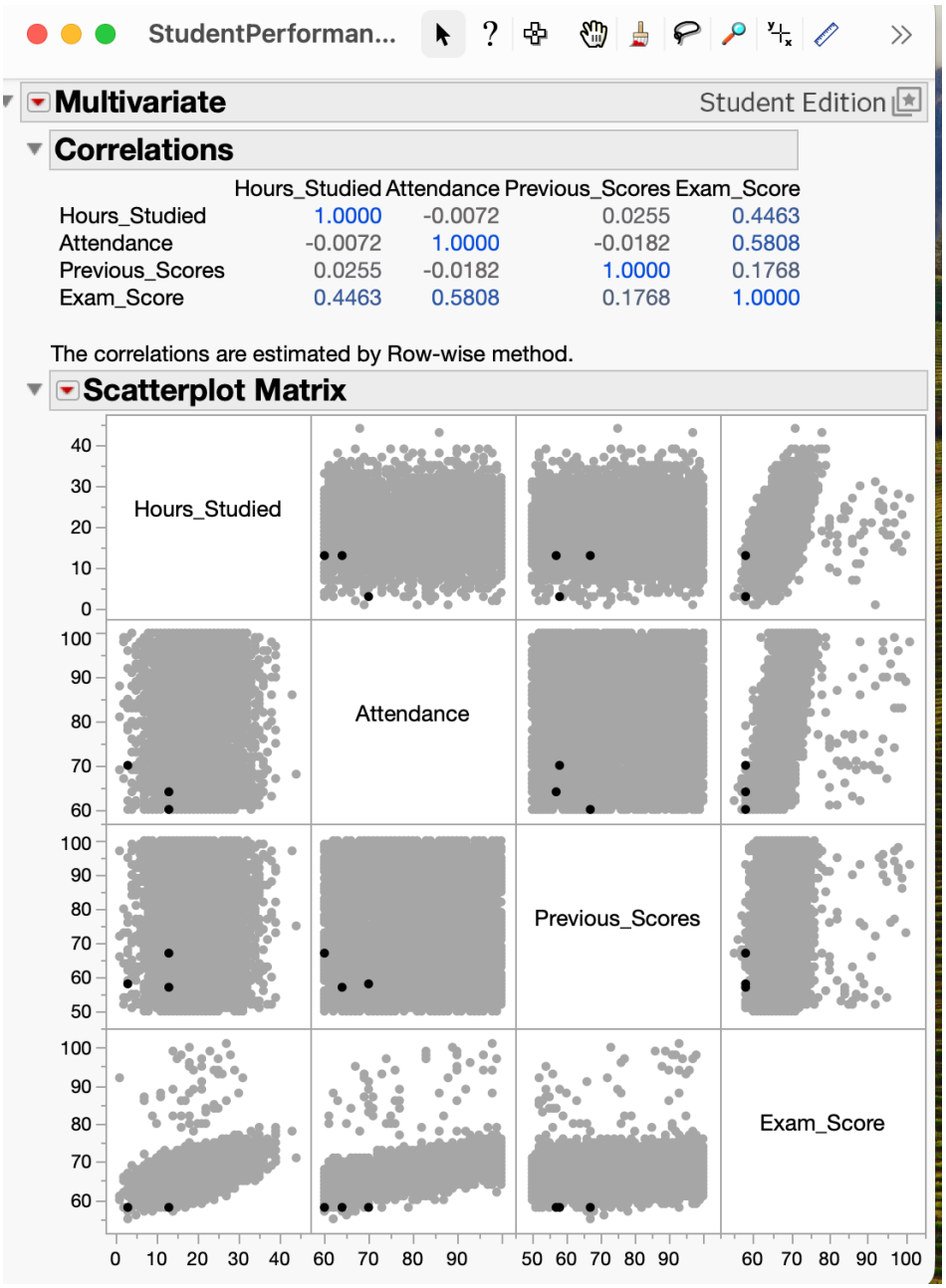
3. Exploratory Data Analysis

To fully understand the data before we began to fit the model, we did exploratory data analysis which mainly focused on our dependent variable Exam_Score and its relationship with other key predictors. We started with checking the distribution of Exam_Scores to check for normality and outliers as shown below:



The distribution of Exam_Scores appears to be bell-shaped with a mean of around 67.2428. Also, it was important to note that there are outliers for scores below 59 and those above 75. We decided to keep them as they are likely to be real scores for low performing and high performing students respectively.

Next, we checked the linear relationships between numerical predictors and Exam_Score. These numerical predictors are: Hours_Studied, Previous_Scores and Attendance. The following scatterplot matrix shows these correlations:



The scatterplot matrix above shows a positive linear relationship between Exam_Score and a few predictors especially Attendance which has a positive correlation of 0.5808 followed by Hours_Studied. Already, this tells us students who show up for class and put in more study hours perform better, which is the norm for many students in many schools. This aligns with our expectations.

4. Methods

This analysis used existing data to build a predictive model for Exam_Score using a set of 8 predictors. The analysis was done using JMP software. Before the analysis, the dataset was screened to check for any missing values, and 78 null values were noted in the Teacher_Quality predictor, and we decided to apply listwise deletion. Categorical predictors such as Parental_Involvement and Motivation_Level were treated as nominal which allowed JMP to create the required dummy variables for regression.

In model building and selection, we used multiple linear regression with standard least squares method. The initial approach was fitting a full model with all 8 potential predictors to establish a baseline for performance. In refining the model, we used a backward selection strategy. We evaluated predictors based on significance of their t-tests and monitored changes in Adjusted R^2.

To test how good our model was able to predict without risk of overfitting, we used Train/Test validation. The now cleaned dataset was randomly split into 2 sets in 80/20 split, training set and validation set respectively. The training set was used to fit the regression coefficients and validation set was used to evaluate how well the model predicts scores for new data. Similar values in the Root Mean Square Error between training and validation sets of data would indicate a model that integrates well to new data.

The first regression model, which is a full model is as follows:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$

Where:

x_1 is Hours_Studied

x_2 is Attendance

x_3 is Parental_Involvement

x_4 is Access_to_Resources

x_5 is Sleep_Hours

x_6 is Previous_Scores

x_7 is Motivation_Level

x_8 is Teacher_Quality

To determine the significance of the overall model and individual predictors, we tested the following hypotheses:

H_0: $\beta_0 = \beta_1 = \beta_2 = \cdots \beta_8 = 0$ vs H_a: $at\ least\ one\ \ \beta_i \neq 0$

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.6660	2.2356	5172
Validation Set	0.5881	2.5799	1357

Summary of Fit

RSquare	0.666048
RSquare Adj	0.665271
Root Mean Square Error	2.238398
Mean of Response	67.23898
Observations (or Sum Wgts)	5172

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	51553.846	4296.15	857.4432
Error	5159	25848.775	5.01	**Prob > F**
C. Total	5171	77402.622		<.0001*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	5158	25848.275	5.01130	10.0226
Pure Error	1	0.500	0.50000	**Prob > F**
Total Error	5159	25848.775		0.2479
				Max RSq
				1.0000

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	41.555851	0.332098	125.13	<.0001*
Hours_Studied	0.2965202	0.005247	56.51	<.0001*
Attendance	0.1996928	0.002685	74.38	<.0001*
Parental_Involvement[Low]	-1.005362	0.052144	-19.28	<.0001*
Parental_Involvement[Medium]	-0.009794	0.041873	-0.23	0.8151
Access_to_Resources[Low]	-1.029709	0.052408	-19.65	<.0001*
Access_to_Resources[Medium]	0.0374599	0.042052	0.89	0.3731
Sleep_Hours	0.0022601	0.021337	0.11	0.9156
Previous_Scores	0.0472414	0.002176	21.71	<.0001*
Motivation_Level[Low]	-0.548298	0.047263	-11.60	<.0001*
Motivation_Level[Medium]	0.0113584	0.041909	0.27	0.7864
Teacher_Quality[Low]	-0.481919	0.068914	-6.99	<.0001*
Teacher_Quality[Medium]	-0.03986	0.046172	-0.86	0.3880

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Hours_Studied	1	1	15998.507	3193.045	<.0001*
Attendance	1	1	27721.139	5532.694	<.0001*
Parental_Involvement	2	2	2557.763	255.2442	<.0001*
Access_to_Resources	2	2	2577.791	257.2428	<.0001*
Sleep_Hours	1	1	0.056	0.0112	0.9156
Previous_Scores	1	1	2362.267	471.4706	<.0001*
Motivation_Level	2	2	741.679	74.0136	<.0001*
Teacher_Quality	2	2	509.039	50.7980	<.0001*

The second regression model, which is reduced by removing Sleep_Hours, is as follows:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$

Where:

x_1 is Hours_Studied
x_2 is Attendance
x_3 is Parental_Involvement
x_4 is Access_to_Resources
x_6 is Previous_Scores
x_7 is Motivation_Level
x_8 is Teacher_Quality

To determine the significance of the overall model and individual predictors, we tested the following hypotheses:

H_0: $\beta_0 = \beta_1 = \beta_2 = \cdots \beta_8 = 0$ vs H_a: $at\ least\ one\ \ \beta_i \neq 0$

Summary of Fit

RSquare	0.648914
RSquare Adj	0.648321
Root Mean Square Error	2.313232
Mean of Response	67.24276
Observations (or Sum Wgts)	6529

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	11	64455.464	5859.59	1095.036
Error	6517	34872.756	5.35	**Prob > F**
C. Total	6528	99328.221		<.0001*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	6503	34842.256	5.35787	2.4594
Pure Error	14	30.500	2.17857	**Prob > F**
Total Error	6517	34872.756		0.0264*
				Max RSq
				0.9997

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	41.605645	0.271271	153.37	<.0001*
Hours_Studied	0.2936872	0.004787	61.35	<.0001*
Attendance	0.1999374	0.002481	80.60	<.0001*
Parental_Involvement[Low]	-0.981939	0.047972	-20.47	<.0001*
Parental_Involvement[Medium]	-0.033394	0.038496	-0.87	0.3857
Access_to_Resources[Low]	-1.035845	0.048204	-21.49	<.0001*
Access_to_Resources[Medium]	0.0520001	0.0386	1.35	0.1780
Previous_Scores	0.0472077	0.001991	23.71	<.0001*
Motivation_Level[Low]	-0.54198	0.043396	-12.49	<.0001*
Motivation_Level[Medium]	-0.005421	0.038579	-0.14	0.8883
Teacher_Quality[Low]	-0.501605	0.063942	-7.84	<.0001*
Teacher_Quality[Medium]	-0.022482	0.042676	-0.53	0.5983

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Hours_Studied	1	1	20142.392	3764.198	<.0001*
Attendance	1	1	34765.235	6496.906	<.0001*
Parental_Involvement	2	2	3171.934	296.3846	<.0001*
Access_to_Resources	2	2	3197.787	298.8002	<.0001*
Previous_Scores	1	1	3007.959	562.1256	<.0001*
Motivation_Level	2	2	933.695	87.2442	<.0001*
Teacher_Quality	2	2	645.628	60.3273	<.0001*

The third regression model, which tests for non-linearity by adding a squared term for Hours_Studied, is as follows:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9$$

Where:

x_1 is Hours_Studied

x_2 is Attendance

x_3 is Parental_Involvement

x_4 is Access_to_Resources

x_6 is Previous_Scores

x_7 is Motivation_Level

x_8 is Teacher_Quality

x_9 is Hours_Studied squared (x_1^2)

To determine if there is a curvilinear relationship between studying and scores, we tested the following hypothesis:

H_0: $\beta_9 = 0$ vs H_a: $\beta_9 \neq 0$

Summary of Fit

RSquare	0.648928
RSquare Adj	0.648281
Root Mean Square Error	2.313365
Mean of Response	67.24276
Observations (or Sum Wgts)	6529

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	64456.821	5371.40	1003.689
Error	6516	34871.400	5.35	Prob > F
C. Total	6528	99328.221		<.0001*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	6502	34840.900	5.35849	2.4596
Pure Error	14	30.500	2.17857	Prob > F
Total Error	6516	34871.400		0.0264*
				Max RSq
				0.9997

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	41.711674	0.343435	121.45	<.0001*
Hours_Studied*Hours_Studied	0.0002839	0.000564	0.50	0.6147
Attendance	0.1999344	0.002481	80.60	<.0001*
Parental_Involvement[Low]	-0.981263	0.047993	-20.45	<.0001*
Parental_Involvement[Medium]	-0.033446	0.038498	-0.87	0.3850
Access_to_Resources[Low]	-1.035359	0.048217	-21.47	<.0001*
Access_to_Resources[Medium]	0.0520605	0.038603	1.35	0.1775
Previous_Scores	0.0471838	0.001992	23.69	<.0001*
Motivation_Level[Low]	-0.541709	0.043402	-12.48	<.0001*
Motivation_Level[Medium]	-0.005639	0.038584	-0.15	0.8838
Teacher_Quality[Low]	-0.501331	0.063948	-7.84	<.0001*
Teacher_Quality[Medium]	-0.022678	0.042681	-0.53	0.5952
Hours_Studied	0.2823147	0.02309	12.23	<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Hours_Studied*Hours_Studied	1	1	1.357	0.2535	0.6147
Attendance	1	1	34764.001	6495.932	<.0001*
Parental_Involvement	2	2	3164.876	295.6912	<.0001*
Access_to_Resources	2	2	3192.439	298.2663	<.0001*
Previous_Scores	1	1	3003.211	561.1740	<.0001*
Motivation_Level	2	2	932.961	87.1656	<.0001*
Teacher_Quality	2	2	645.452	60.3040	<.0001*
Hours_Studied	1	1	800.036	149.4932	<.0001*

The fourth regression model, which tests for an interaction between Hours_Studied and Motivation_Level, is as follows:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_{10} x_{10}$$

Where:

x_1 is Hours_Studied

x_2 is Attendance

x_3 is Parental_Involvement

x_4 is Access_to_Resources

x_6 is Previous_Scores

x_7 is Motivation_Level

x_8 is Teacher_Quality

x_9 is Hours_Studied * Motivation_Level

To determine if the effect of studying depends on a student's Motivation_Level, we tested the following hypothesis:

H_0: $\beta_{10} = 0$ vs H_a: $\beta_{10} \neq 0$

Summary of Fit

RSquare	0.649181
RSquare Adj	0.648481
Root Mean Square Error	2.312708
Mean of Response	67.24276
Observations (or Sum Wgts)	6529

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	13	64481.962	4960.15	927.3703
Error	6515	34846.258	5.35	Prob > F
C. Total	6528	99328.221		<.0001*

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	6501	34815.758	5.35545	2.4582
Pure Error	14	30.500	2.17857	Prob > F
Total Error	6515	34846.258		0.0265*
				Max RSq
				0.9997

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	41.627273	0.273621	152.13	<.0001*
Hours_Studied	0.2928024	0.005161	56.74	<.0001*
Attendance	0.1999085	0.00248	80.61	<.0001*
Parental_Involvement[Low]	-0.982243	0.047966	-20.48	<.0001*
Parental_Involvement[Medium]	-0.030903	0.038504	-0.80	0.4222
Access_to_Resources[Low]	-1.033768	0.048205	-21.45	<.0001*
Access_to_Resources[Medium]	0.0514313	0.038592	1.33	0.1827
Previous_Scores	0.0471589	0.001991	23.69	<.0001*
Motivation_Level[Low]	-0.858106	0.152114	-5.64	<.0001*
Motivation_Level[Medium]	0.0249383	0.134055	0.19	0.8524
Teacher_Quality[Low]	-0.503521	0.063951	-7.87	<.0001*
Teacher_Quality[Medium]	-0.021231	0.042671	-0.50	0.6188
Hours_Studied*Motivation_Level[Low]	0.0158802	0.007316	2.17	0.0300*
Hours_Studied*Motivation_Level[Medium]	-0.001454	0.006436	-0.23	0.8213

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Hours_Studied	1	1	17216.733	3218.911	<.0001*
Attendance	1	1	34753.932	6497.738	<.0001*
Parental_Involvement	2	2	3161.993	295.5896	<.0001*
Access_to_Resources	2	2	3185.165	297.7557	<.0001*
Previous_Scores	1	1	3000.697	561.0226	<.0001*
Motivation_Level	2	2	186.692	17.4523	<.0001*
Teacher_Quality	2	2	646.652	60.4503	<.0001*
Hours_Studied*Motivation_Level	2	2	26.498	2.4771	0.0841

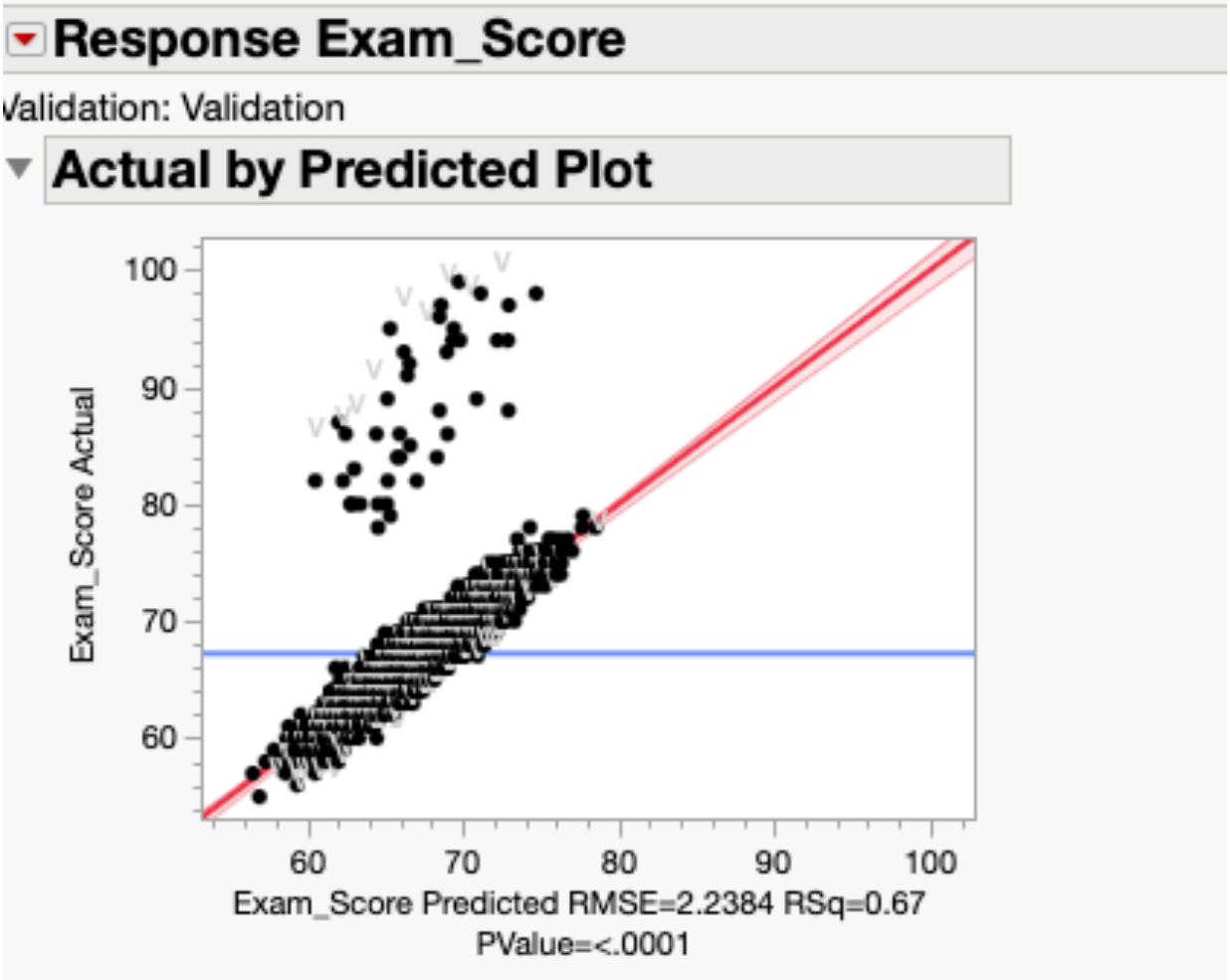
5. Results

We fit a multiple linear regression model using Exam_Score as the dependent variable and 8 predictors, namely: Hours_Studied, Attendance, Previous_Scores, Sleep_Hours, Parental_Involvement, Motivation_Level, Teacher_Quality and Access_to_Resources.

As seen below, the best overall model was statistically significant as $p < 0.0001$ and explained approximately 66.53% of the variance in Exam_Score as Adjusted $R^2 = 0.6653$. The Root Mean Square Error for the training set was roughly 2.2383, suggesting our model's predictions fell within 2 points of the actual exam score.

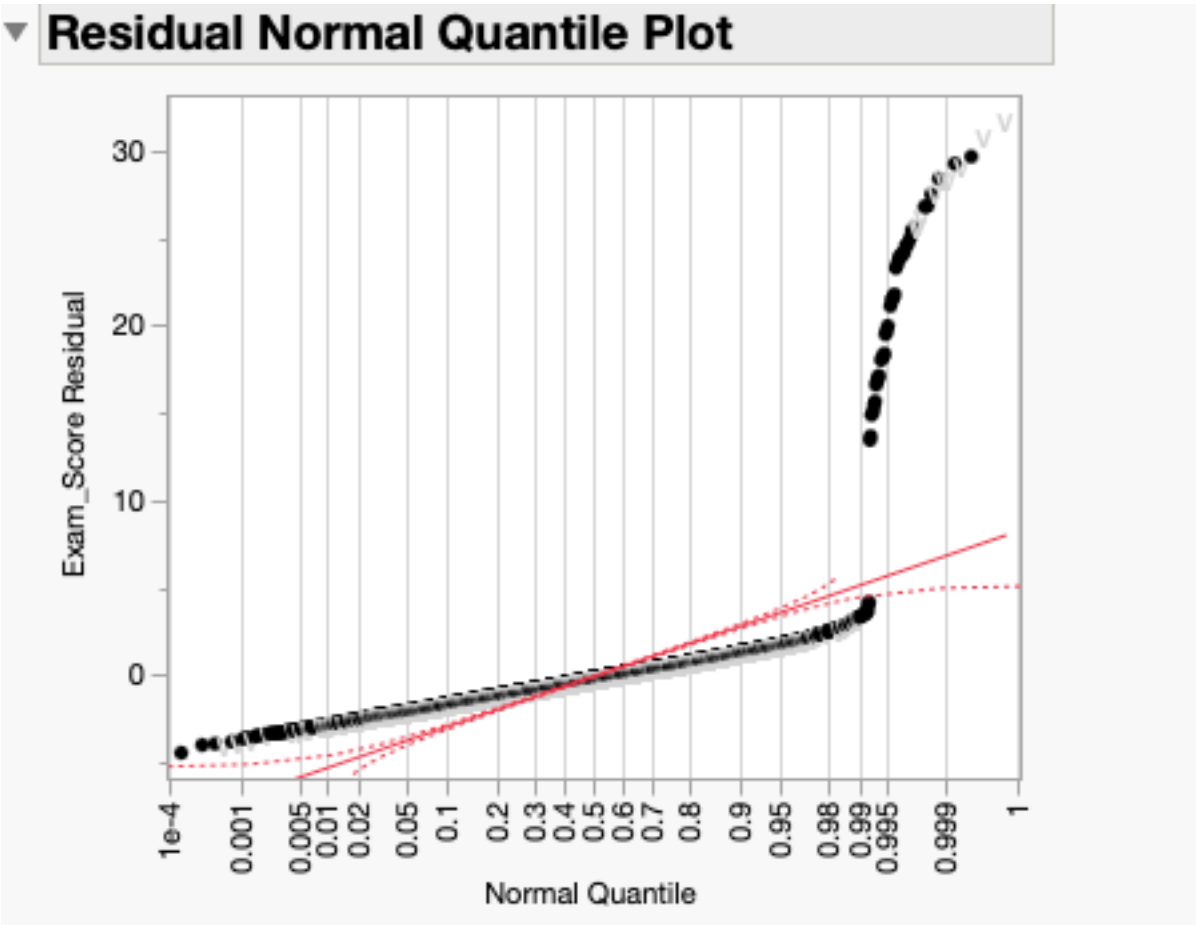
The predictor Attendance had a positive coefficient indicating that for every 1% increase in attendance, the Exam_Score increased by roughly 0.1997 points, holding all other variables constant. The predictor Hours_Studied also showed a strong positive effect as each additional hour of study suggested an increase of around 0.2965 points. A very interesting observation was noted in the Sleep_Hours predictor as it had a $p > 0.05$ around 0.9156 indicating to us that it is not statistically significant to the model.

To check for overfitting, we compared the performance of the model on the training set against the validation set. You will notice the difference between RMSE and RASE is very close, within a 0.1 difference, 2.2383 and 2.2356, which confirms our model is not overfitting the data.



The “Actual by Predicted Plot” showed points tightly clustered around the diagonal line indicating a good fit. It however, underpredicts for top performers

The residual distribution was checked using “Residual Normal Quantile Plot” and found to be seriously violated and not normally distributed. This means the Exam_Score did not satisfy normality assumption in ordinary least squares regression.



6. Conclusion and Discussion

The primary objective of this research was to identify the most significant factors of student academic performance. The analysis of 6529 student records revealed that academic performance is not determined by a single factor, but rather a combination of consistent effort and environmental support. The final regression model explained approximately 66.53% of the variation in Exam_Score. Both Attendance and Hours_Studied stood out as strong positive predictors for high exam scores as they outweighed every other variable.

Another important factor to note was how important environmental factors played a crucial role. Students with “high” Parental_Involvement had better performance than students with “low” Parental_Involvement, considering study habits stay constant.

While sleep has often been regarded very important for health, our analysis nullified the idea that number of hours a student sleeps will have an impact on their performance. The variable Sleep_Hours was not statistically significant in our model. This showed that different sleep schedules did not lead to high or low scores in the Exam_Score even when the other factors are accounted for.

While our model is statistically significant, there are limitations to this study. First, the data is observational, meaning we can identify correlations but cannot definitively prove causation. For example, while higher attendance predicts higher scores, we cannot rule out that a third factor (like student discipline) causes both. Second, our model explains around 66% of the variance, leaving 34% unexplained This suggests there are other unmeasured

variables—such as innate aptitude, classroom peer dynamics, or test anxiety—that also play a major role in student performance.

In conclusion, this project demonstrates that while we cannot control every aspect of a student's life, schools and parents can significantly do their part by focusing on the fundamentals. Policies that prioritize attendance and programs that encourage parental engagement are likely to yield the highest return on investment for improving student outcomes.

7. References

i. Kaggle

ii. JMP Software

iii. STAT 311 Fall 2025 Course Materials and Project Guidelines