MODELING EXAM SCORES OF STUDENTS

STAT 311 Regression Analysis, Fall 2025

11/27/2025

Kate Ellestad, Byron Johnson-Blanchard, Yangpao Vang

Contents

Introduction

The purpose of this project is to look at which factors have the strongest relationship with student exam performance. The dataset includes a mix of behavioral, environmental, and parental variables, and our goal is to see which ones are most influential in predicting exam score once they are all considered together. We focused on predictors that students or schools could realistically influence, like study habits, attendance, sleep, access to resources, and parental involvement. Using this data, educators, parents, and policy makers can focus resources on the most impactful areas to ensure student success.

To answer the research question, we used multiple linear regression and compared different model structures. Since the dataset contains both continuous and categorical variables, we applied stepwise selection to narrow the predictors down to the ones that contribute most meaningfully to the model. After selecting a final model, we checked assumptions, looked at residuals, and evaluated whether the model fit was reasonable.

Data

The dataset is a synthetic dataset generated for analytical and educational purposes and was downloaded from Kaggle and includes 6,607 students and provides a broad view of the factors that can influence academic performance. Each observation represents an individual student and contains information about study behaviors, personal characteristics, family background, and school environment. The goal of the dataset is to capture a realistic set of variables that might contribute to differences in exam outcomes.

Variables

- **Hours_Studied:** Weekly study time.
- **Attendance:** Percentage of classes attended.
- **Parental_Involvement:** Low, Medium, or High involvement in the student's education.
- **Access_to_Resources:** Availability of educational materials/support (Low, Medium, High).

- **Extracurricular_Activities:** Participation in any extracurricular programs (Yes, No).
- **Sleep_Hours:** Average nightly sleep.
- **Previous_Scores:** Scores from earlier exams (percentage).
- **Motivation_Level:** Self-reported motivation (Low, Medium, High).
- **Internet_Access:** Whether reliable internet is available at home (Yes, No)
- **Tutoring_Sessions:** Number of monthly tutoring sessions attended.
- **Family_Income:** Household income (Low, Medium, High).
- **Teacher_Quality:** Perceived quality of teachers (Low, Medium, High).
- **School_Type:** Public or private school enrollment.
- **Peer_Influence:** Whether peers are considered Positive, Neutral, or Negative influences.
- **Physical_Activity:** Weekly physical activity duration.
- **Learning_Disabilities:** Indicator for diagnosed learning disabilities (Yes, No).
- **Parental_Education_Level:** Highest parental education level (High School, College, Postgraduate).
- **Distance_from_Home:** Near, Moderate, or Far distance to school.
- **Gender:** Male or Female.
- **Exam_Score:** Final exam score for the course (Percentage) (our response variable)

Data Structure

	N	Mean	Std Dev	Min	Max	N Missing	Std Err	Median
Exam_Score	6607	67.235659149	3.8904557813	55	101	0	0.0478628252	67
Attendance	6607	79.977448161	11.547474961	60	100	0	0.1420642739	80
Sleep_Hours	6607	7.0290600878	1.4681202267	4	10	0	0.0180617351	7
Previous_Scores	6607	75.070531255	14.399784351	50	100	0	0.1771551715	75
Tutoring_Sessions	6607	1.4937187831	1.2305704213	0	8	0	0.0151392485	1

Table 1 Numerical Data Summary

	N	N Missing	Levels
Access To Resources	6607	0	3
Extracurricular Activities	6607	0	2
Motivation Level	6607	0	3
Internet Access	6607	0	2
Family Income	6607	0	3
Teacher Quality	6607	78	3
School Type	6607	0	2
Peer Influence	6607	0	3
Learning Disabilities	6607	0	2
Parental Education Level	6607	90	3
Distance from home	6607	67	3
Gender	6607	0	2

Table 2 Categorical Data Summary

EDA

Before selecting any models, we checked distribution and frequencies. The goal was to understand the overall patterns in the predictors and identify any unusual or extreme values.

Exam Score Distribution

First, we checked the distribution of the exam score. The scores appeared bell-shaped with no obvious skew, suggesting that a linear model would be appropriate. Some scores on the lower end could act as influential points, but none appeared extreme enough to warrant immediate removal.

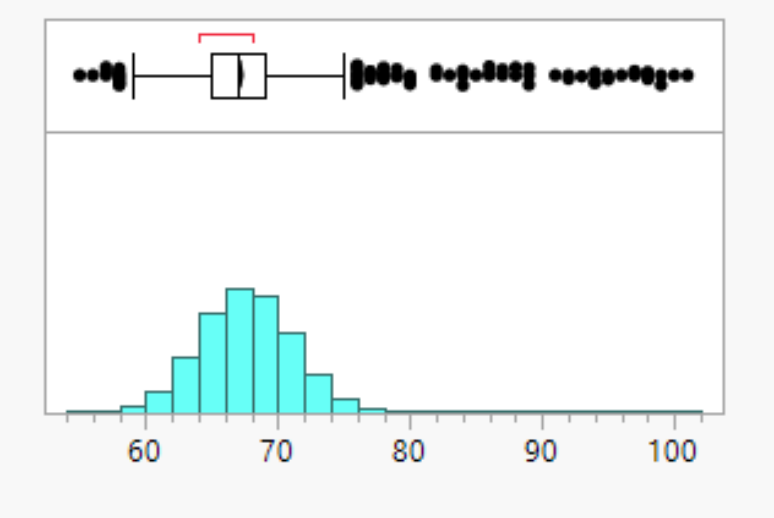


Figure 1 Exam Score Distribution

Next, we looked at each numerical predictor. Attendance was uniform from about 60 to 100. Previous scores had a wider spread between 50 and 101 with a slight right tail. Tutoring showed a large right skew, with most students attending few or no tutoring sessions and a smaller number attending many. Physical Activity appeared to have a fairly normal distribution centered around 3.5 hours on average. Hours studied has a normally distributed shape centered around 20 hours per week, with a small number of higher values that appeared as possible mild outliers. Sleep hours had a very symmetrical normal distribution centered around seven hours per night on average.

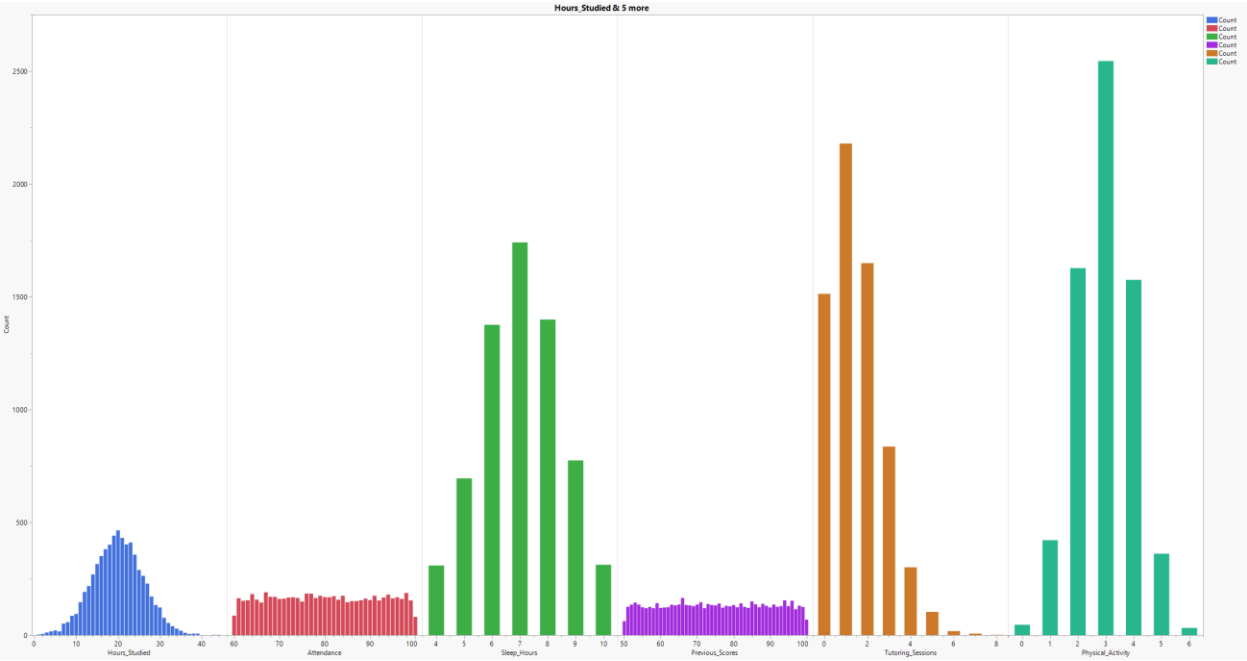


Figure 2 Distribution of Numeric Variables

MODELING EXAM SCORES OF STUDENTS

We constructed a panel of scatterplots comparing exam scores to each numerical predictor. Attendance and Hours Studied had the clearest positive relationship. Sleep Hours and Physical Activity don't have a very strong linear pattern, and Tutoring Sessions showed a wide vertical spread at all values, indicating little to no direct linear relationship.

Figure 3 Scatterplots of Continuous Variables vs Exam Score

To explore how the predictors behave across different exam scores, we created a grid of histograms that show the distribution of each numerical variable within bands of exam score. This highlights several clear patterns. Students with lower exam scores tended to have lower attendance percentages, while the distribution shifted upward for higher scoring groups. A similar pattern appears for hours studied and previous scores, where higher performing students generally showed higher values. In contrast, sleep hours, tutoring sessions, and physical activity showed fairly similar distributions across exam score bands, suggesting weaker relationships with exam performance.

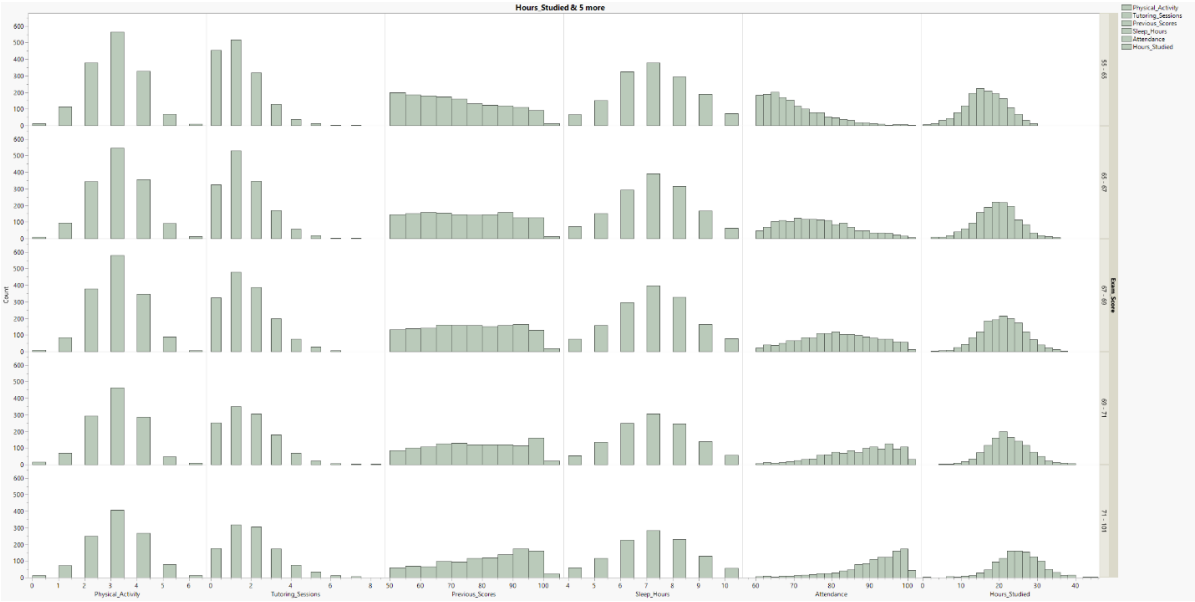


Figure 4 Histogram of Numerical Variables Across Exam Scores

Methods

Variable & Baseline Model Selection

To identify an appropriate starting model, we applied stepwise regression using all 19 available variables. We used a 5% significance level for both entry and removal. All variables appeared to be selected, with the exception of school type, gender, and sleep hours. To determine the number of variables for us to select, we used the all possible models process in JMP. We saved this to a data table and created a plot that shows a clear point of diminishing returns after about six variables.

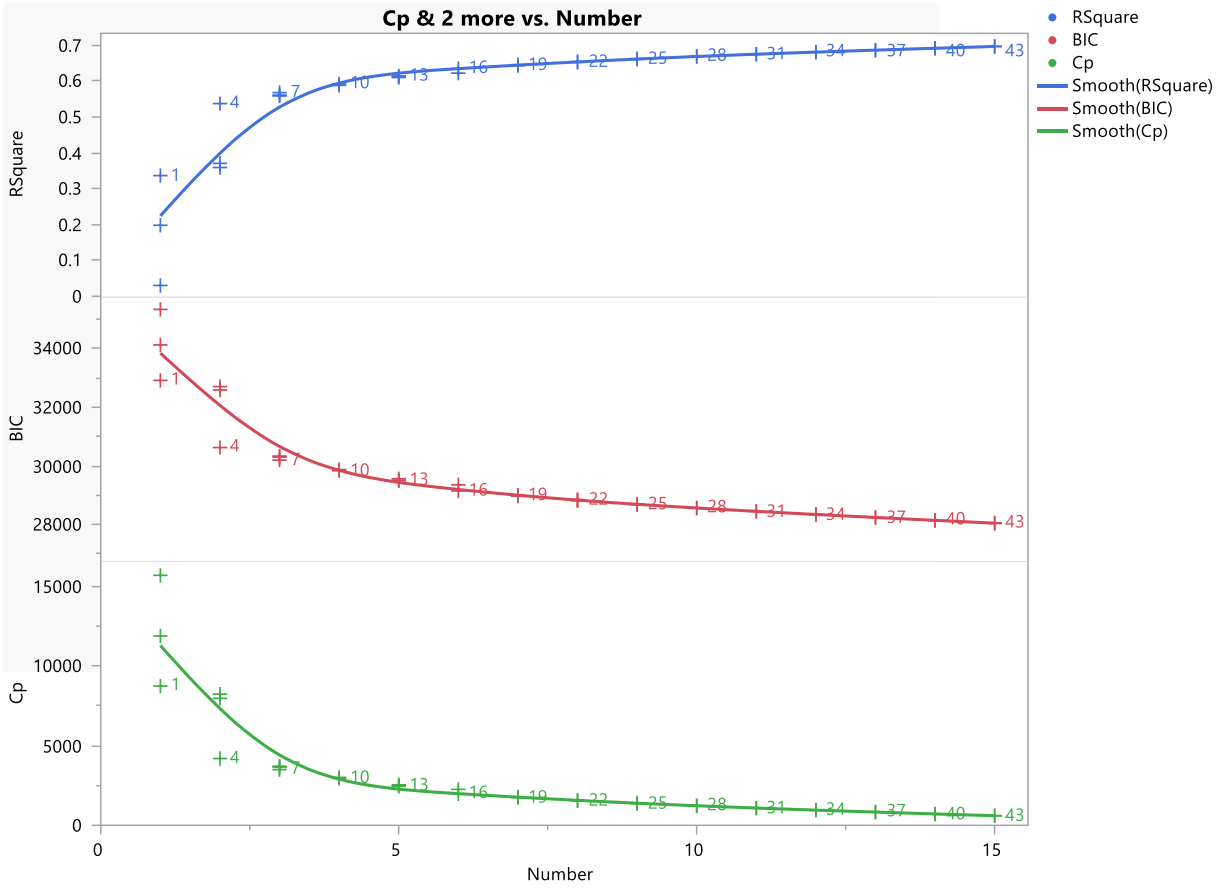


Figure 5 RSquare, BIC, and CP by Number of Predictors in Model

Based on the results of the all possible models process, we looked at the best 3 models that had six predictors and determined that the variables of Attendance, Hours Studied, Previous Scores, Tutoring Sessions, Parental Involvement, and Access to Resources would be most appropriate.

To make sure that the deletion of rows with missing data did not affect the stepwise variable selection and all models process, the above steps were also completed with all of the rows with missing data removed. It was found that the removal of these rows had no influence on our chosen variables and base model.

Modeling

Model 1: Six Predictor Main Effects Model

Our first model was the six predictor model identified by the all possible models output. This model includes the predictors hours studied, attendance, previous scores, tutoring sessions, parental involvement, and access to resources. Dummy variables were created for parental involvement and access to resources. These variables represented the point where the model

achieved a balance between explanatory strength and parsimony. Beyond six predictors, increases in R^2, Cp, and BIC were small.

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

Where:

$$y = Exam\ Score$$

$$x_1 = Hours\ Studied$$

$$x_2 = Attendance$$

$$x_3 = Previous\ Scores$$

$$x_4 = Tutoring\ Sessions$$

$$x_5 = \begin{cases} 1\ if\ parental\ involvement = low \\ else\ 0 \end{cases}$$

$$x_6 = \begin{cases} 1\ if\ parental\ involvement = medium \\ else\ 0 \end{cases}$$

$$x_7 = \begin{cases} 1\ if\ access\ to\ resources = low \\ else\ 0 \end{cases}$$

$$x_8 = \begin{cases} 1\ if\ access\ to\ resources = medium \\ else\ 0 \end{cases}$$

Hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0\ \ vs\ H_1: At\ least\ one\ of\ \beta_i \neq 0$$

Output

Summary of Fit

RSquare	0.659986
RSquare Adj	0.659573
Root Mean Square Error	2.269929
Mean of Response	67.23566
Observations (or Sum Wgts)	6607

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	65989.367	8248.67	1600.882
Error	6598	33996.711	5.15	Prob > F
C. Total	6606	99986.079		<.0001*

Model 1 explains about 66% of variability in exam scores and it's p-value indicates that there is evidence to support that this model is useful in predicting exam scores.

Indicator Function Parameterization				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	42.906806	0.271365	158.11	<.0001*
Hours_Studied	0.2942981	0.004665	63.08	<.0001*
Attendance	0.1989845	0.00242	82.23	<.0001*
Previous_Scores	0.0478584	0.001942	24.65	<.0001*
Tutoring_Sessions	0.5035489	0.022705	22.18	<.0001*
Parental_Involvement[Low]	-1.996425	0.081051	-24.63	<.0001*
Parental_Involvement[Medium]	-1.034538	0.065082	-15.90	<.0001*
Access_to_Resources[Low]	-2.036606	0.080906	-25.17	<.0001*
Access_to_Resources[Medium]	-0.94334	0.064518	-14.62	<.0001*

Every parameter in model 1 is statistically significant at $\alpha = .05$

Conclusion

Model 1 is a statistically significant model. All predictors included have meaningful and significant relationships with exam scores. The adjusted R^2 of 0.6596 indicates that this model explains a substantial portion of exam score variation.

Because this model performed well, we next explored whether adding interaction terms (Model 2) or quadratic components (Model 3) could meaningfully improve performance.

Model 2 – Two-Way First Order Interaction Model

After establishing Model 1 as a strong baseline, we evaluated whether adding all first-order interaction terms among the predictors would meaningfully improve predictive performance. This step was motivated by the idea that the effect of behavioral variables (such as hours studied or tutoring) might depend on socio-environmental context (such as access to resources or parental involvement). Model 2 includes all main effects from Model 1 and all two-way first order interaction terms among those predictors.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \beta_{15} x_1 x_5 + \beta_{16} x_1 x_6 + \beta_{17} x_1 x_7 + \beta_{18} x_1 x_8 + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{25} x_2 x_5 + \beta_{26} x_2 x_6 + \beta_{27} x_2 x_7 + \beta_{28} x_2 x_8 + \beta_{34} x_3 x_4 + \beta_{35} x_3 x_5 + \beta_{36} x_3 x_6 + \beta_{37} x_3 x_7 + \beta_{38} x_3 x_8 + \beta_{45} x_4 x_5 + \beta_{46} x_4 x_6 + \beta_{47} x_4 x_7 + \beta_{48} x_4 x_8 + \beta_{57} x_5 x_7 + \beta_{58} x_5 x_8 + \beta_{67} x_6 x_7 + \beta_{68} x_6 x_8$$

Where:

$$y = Exam\ Score$$

$$x_1 = Hours\ Studied$$

$$x_2 = Attendance$$

$$x_3 = Previous\ Scores$$

$$x_4 = Tutoring\ Sessions$$

$$x_5 = \begin{cases} 1\ if\ parental\ involvement = low \\ else;\ 0 \end{cases}$$

$$x_6 = \begin{cases} 1 \; if \; parental \; involvement = medium \\ else; \; 0 \end{cases}$$

$$x_7 = \begin{cases} 1 \; if \; access \; to \; resources = low \\ else; \; 0 \end{cases}$$

$$x_8 = \begin{cases} 1 \; if \; access \; to \; resources = medium \\ else; \; 0 \end{cases}$$

Hypothesis

$$H_0: \beta_1 = \beta_2 = \cdots . \beta_{68} = 0 \;\; vs \; H_1: At \; least \; one \; of \; \beta_j \neq 0$$

Output

Summary of Fit

RSquare	0.660894
RSquare Adj	0.65914
Root Mean Square Error	2.271375
Mean of Response	67.23566
Observations (or Sum Wgts)	6607

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	34	66080.195	1943.54	376.7167
Error	6572	33905.884	5.16	Prob > F
C. Total	6606	99986.079		<.0001*

Conclusion

To evaluate whether including all two-way interaction terms improved model performance, we compared Model 2 (interaction model) to the simpler main-effects model (Model 1). Although Model 2 adds a large number of additional predictors, the overall fit did not improve. Model 1 had an MSE of 5.15 and an R_a^2 of 0.6596, while Model 2 had a slightly higher MSE of 5.16 and a slightly lower R_a^2 of 0.6591. The RMSE values were essentially identical (2.27), and the vast majority of interaction terms in Model 2 were not statistically significant. While Model 2's overall F-test remained significant (as expected with a large sample size), the decrease in R_a^2 and increase in MSE indicate that the interaction terms did not provide meaningful explanatory power. Therefore, the added model complexity is not justified, and model 1 remains the best model.

Model 3 – Quadratic Model

After evaluating the full first-order interaction model, we next tested whether adding nonlinear curvature to the predictors would improve model performance. The idea was that variables such as hours studied, attendance, or previous scores might have diminishing returns or nonlinear effects on exam performance. To check for this, we added all quadratic terms for the predictors from Model 1.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_1^2 + \beta_{10} x_2^2 + \beta_{11} x_3^2 + \beta_{12} x_4^2$$

Where:

MODELING EXAM SCORES OF STUDENTS

$$y = Exam\ Score$$

$$x_1 = Hours\ Studied$$

$$x_2 = Attendance$$

$$x_3 = Previous\ Scores$$

$$x_4 = Tutoring\ Sessions$$

$$x_5 = \begin{cases} 1\ if\ parental\ involvement = low \\ else;\ 0 \end{cases}$$

$$x_6 = \begin{cases} 1\ if\ parental\ involvement = medium \\ else;\ 0 \end{cases}$$

$$x_7 = \begin{cases} 1\ if\ access\ to\ resources = low \\ else;\ 0 \end{cases}$$

$$x_8 = \begin{cases} 1\ if\ access\ to\ resources = medium \\ else;\ 0 \end{cases}$$

Hypothesis

$$H_0: \beta_9 = \beta_{10} = \cdots = \beta_{12} = 0\ \ vs\ H_1: At\ least\ one\ of\ the\ quadratic\ coefficients \neq 0$$

Output

Summary of Fit

RSquare	0.660209
RSquare Adj	0.659591
Root Mean Square Error	2.269871
Mean of Response	67.23566
Observations (or Sum Wgts)	6607

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	66011.720	5500.98	1067.671
Error	6594	33974.359	5.15	Prob > F
C. Total	6606	99986.079		<.0001*

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	44.303622	1.709596	25.91	<.0001*
Hours_Studied	0.2807889	0.022456	12.50	<.0001*
Hours_Studied*Hours_Studied	0.0003359	0.000549	0.61	0.5405
Attendance	0.2071811	0.037476	5.53	<.0001*
Attendance*Attendance	-5.12e-5	0.000234	-0.22	0.8265
Parental_Involvement[Low]	-1.996646	0.081091	-24.62	<.0001*
Parental_Involvement[Medium]	-1.035646	0.065105	-15.91	<.0001*
Access_to_Resources[Low]	-2.03395	0.080937	-25.13	<.0001*
Access_to_Resources[Medium]	-0.943582	0.064531	-14.62	<.0001*
Previous_Scores	0.003619	0.022567	0.16	0.8726
Previous_Scores*Previous_Scores	0.0002946	0.00015	1.97	0.0493*
Tutoring_Sessions	0.5139161	0.057108	9.00	<.0001*
Tutoring_Sessions*Tutoring_Sessions	-0.002647	0.013129	-0.20	0.8402

Model 3 explains about 66% of variability in exam scores and its p-value indicates that there is evidence to support that this model is useful in predicting exam scores, however, it did not improve the model enough.

Conclusion

The quadratic model did not improve predictive performance from Model 1. The MSE increased slightly, and the adjusted R^2 decreased, indicating worse overall fit when curvature was added. Most of the quadratic terms were not statistically significant, and the few that showed borderline significance did not meaningfully improve prediction accuracy.

Because the quadratic model adds considerable complexity without providing a meaningful reduction in error or increase in explained variance, we concluded that curvature is **not** an important feature in this dataset. The relationship between the predictors and exam score appears to be adequately captured by the linear effects in Model 1.

Partial F-Test

To add evidence to our claim that model 1 is the best model, partial F-tests were run on both of the nested models in JMP. Both of the tests resulted in p-values that are greater than alpha=0.05, showing that both the interaction terms (p-value=0.889) and quadradic terms (p-value=0.354) are not significant.

Global F-Test

All three models were statistically significant overall, but neither model 2 nor model 3 improved fit enough to justify the additional terms. For the remainder of the analysis, we therefore focused on model 1.

Model	*MSE*	R_a^2	*s*	F Ratio	p-value
1	5.15	.66	2.27	1600.882	<.0001
2	5.16	.66	2.27	376.7167	<.0001
3	5.15	.66	2.27	1067.671	<.0001

Model Diagnostics

After determining that model one was the best balance of fit and parsimony, we moved on to validating the model. This included checking for outliers and influential observations, checking for multicollinearity, confirming that regression assumptions were met, and performing cross validation to confirm predictive performance.

Identifying Outliers

We used several diagnostics to evaluate whether any observations strongly influenced the fitted model.

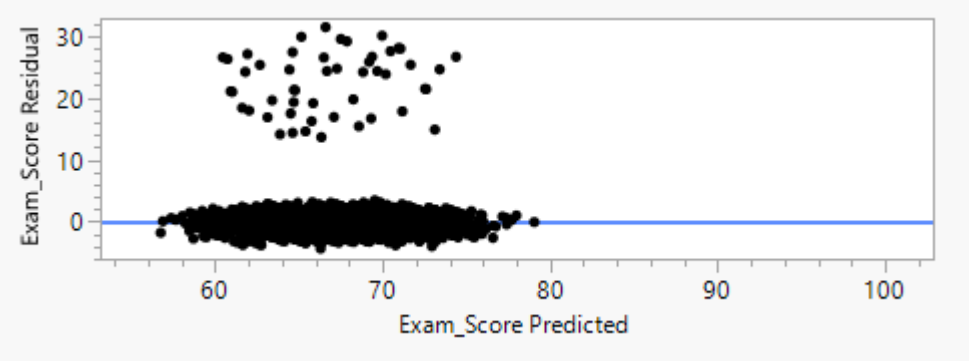


Figure 6 Residuals by Predicted

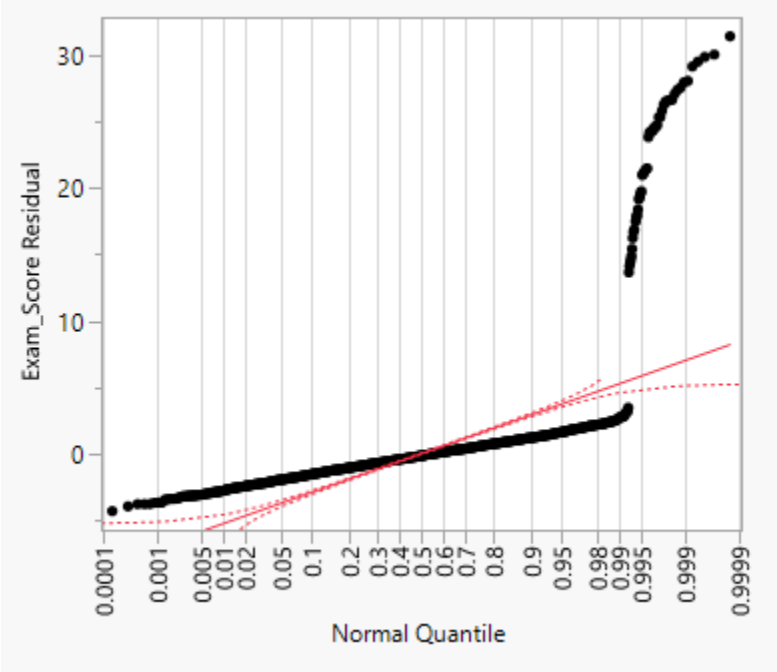


Figure 7 Normal Quantile by Residual Plot

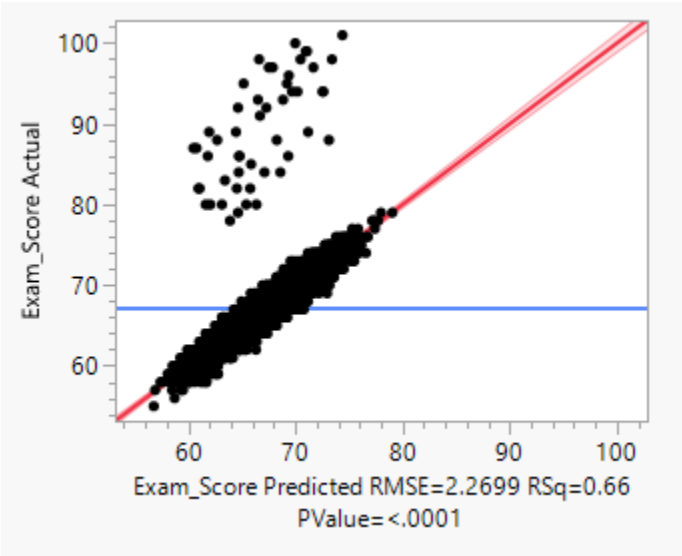


Figure 8 Actual vs. Predicted

Figure 9 Distribution of Residuals

The results show that the model is very close to satisfying the assumption of error distribution. There doesn't appear to be the need for us to transform the dependent variable values, but the residual plots reveal a cluster of outliers that warranted further investigation.

We began by identifying potential outliers, using JMP's explore outliers function, which flagged 50 observations using the Normal Quantile method. These matched the observations identified in our diagnostic plots. We isolated these 50 values to explore if we should be excluding them from the model.

Leverage

To determine the leverage cut off we calculate the average hat value

$$\bar{h} = \frac{9}{6607} \approx .00136$$

If $h_i > 2\bar{h} \approx .0027$ then it is considered a high leverage point. All of the 50 observations flagged as outliers were below this threshold with the exception of one, indicating that leverage was not the main concern for these observations.

Cook's Distance

To look at if any of the flagged observations greatly influenced the model, we examined Cook's Distance for each of the 50 observations. A value greater than 1 would indicate an influential observation. All of the values for the flagged observations were far below 1 indicating that while they had large residuals, none had large influence on the estimated coefficients. Removing them would not distort the model.

Externally Studentized Residuals

Next, we tried to identify observations whose response value is unusual based on model predictions. We found that the externally studentized residuals for the flagged responses were between 6 and 14, well above the standard cutoff of 3. This helps to explain the large deviations between the residual vs predicted plot and the heavy right tail in the normal quantile plot.

Outlier Patterns

The 50 outliers appeared to be clustered and high performing students. The range of the flagged observations was 78-101 with a mean of 89, whereas the full dataset clusters scores between 60-75 with a mean of 67. This indicates that these are data points that our model is unable to explain and could indicate a missing variable.

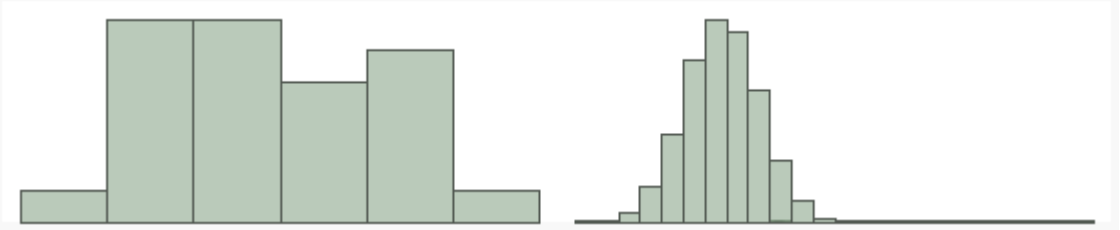


Figure 10 Distribution of Exam Scores: Outliers vs Full Model

To confirm that this pattern was not due to the removal of variables for the sake of model simplicity, we reran the model with all 19 variables included. The same cluster of high performing students continued to exhibit large residuals. This confirms that the clustering is not caused by the exclusion of any measured predictors but indicates the presence of an important unobserved factor.

Interpretation

An examination of the distributions revealed that the outliers flagged were not random anomalies, but a distinct cluster of high-performing students with exam scores much higher than their peers. This indicates that the model is underpredicting exam scores for a subset of high-achieving students.

Because the predictors in the model do not fully account for the performance levels of these students, it is likely that an explanatory variable is missing. As a result, the residuals for these students are very large, which leads to them being flagged as outliers. While their residuals did violate the assumption of normality, their Cook's Distance values were all far below 1, indicating that they were not influential in determining the model coefficients. Their removal improves the model's adherence to regression assumptions without distorting the underlying relationships among the included predictors.

Refitting The Model

After removing outliers, we did another stepwise regression to ensure that the chosen model was still the best starting point, and the output still indicated the model we chose was the best option. We then refit model 1 and saw a marked improvement in model fit. The R_a^2 improved to .898945, meaning the refit model explains nearly 90% of the variation in exam scores for the average student. The mean square was reduced to 1.13 and the F ratio increased to 7290.912. We found the Normal Quantile Plot, Residual by Predicted Plot and Studentized Residuals Plot all much more matching the assumptions for regression.

Summary of Fit

RSquare	0.899068
RSquare Adj	0.898945
Root Mean Square Error	1.064164
Mean of Response	67.06802
Observations (or Sum Wgts)	6557

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	66052.418	8256.55	7290.912
Error	6548	7415.246	1.13	Prob > F
C. Total	6556	73467.664		<.0001*

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	42.502821	0.128013	332.02	<.0001*
Hours_Studied	0.297996	0.002197	135.66	<.0001*
Attendance	0.2000058	0.001139	175.66	<.0001*
Parental_Involvement[Low]	-2.007478	0.038166	-52.60	<.0001*
Parental_Involvement[Medium]	-0.984173	0.03063	-32.13	<.0001*
Access_to_Resources[Low]	-1.987256	0.038079	-52.19	<.0001*
Access_to_Resources[Medium]	-0.920676	0.030365	-30.32	<.0001*
Previous_Scores	0.0482489	0.000915	52.75	<.0001*
Tutoring_Sessions	0.5058041	0.010682	47.35	<.0001*

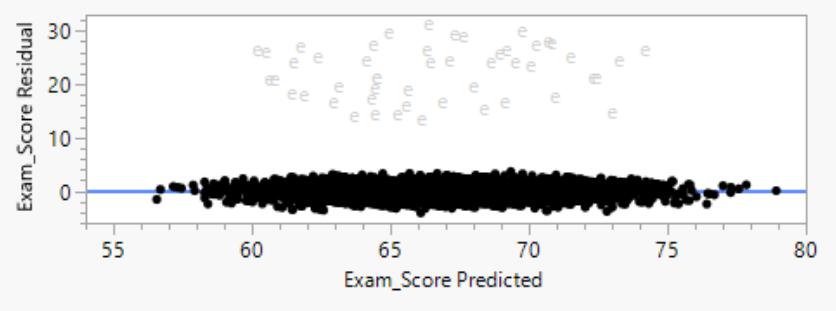


Figure 11 Refit Predicted vs Residual

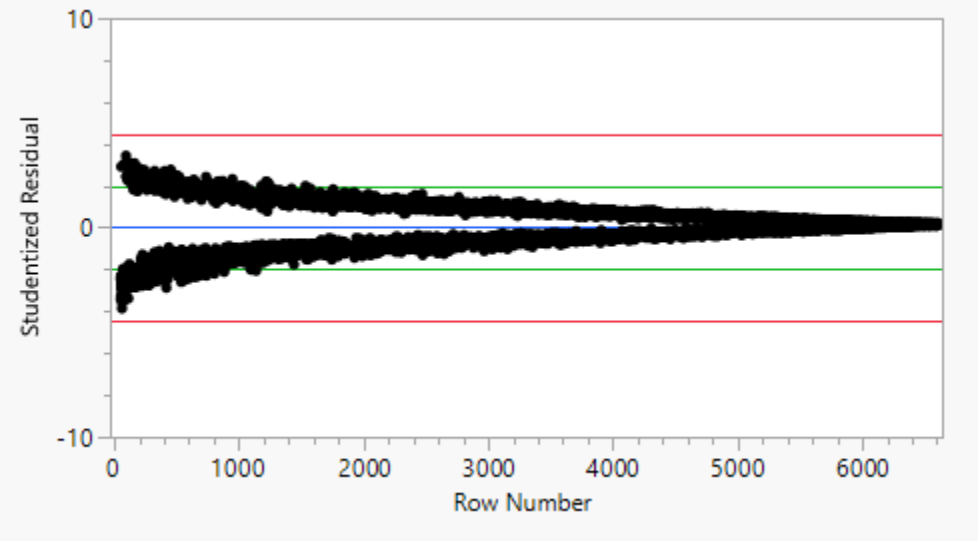


Figure 12 Refit Studentized Residual Plot

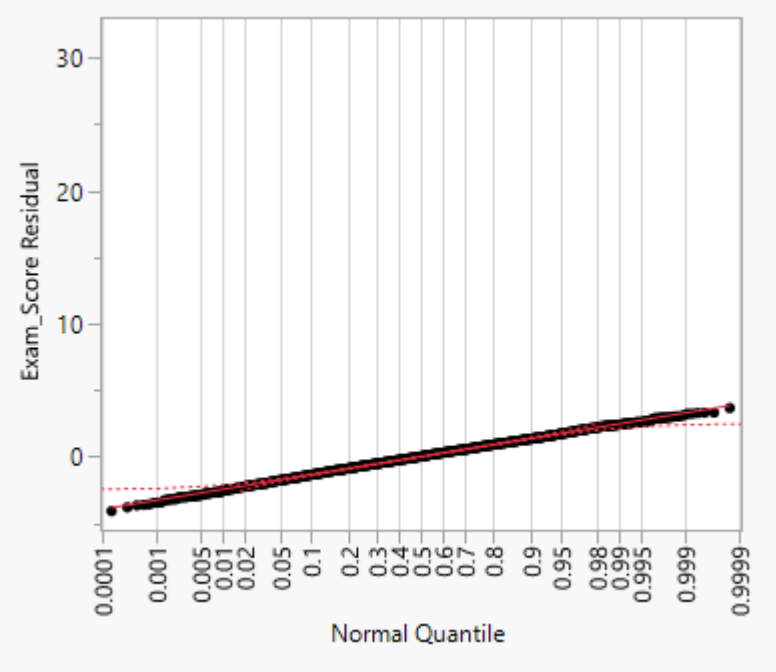


Figure 13 Refit Normal Quantile Plot

Next, we checked for multicollinearity in our numeric variables by creating a color map of correlations. There did not appear to be any strong correlation among predictors, so multicollinearity does not appear to be present.

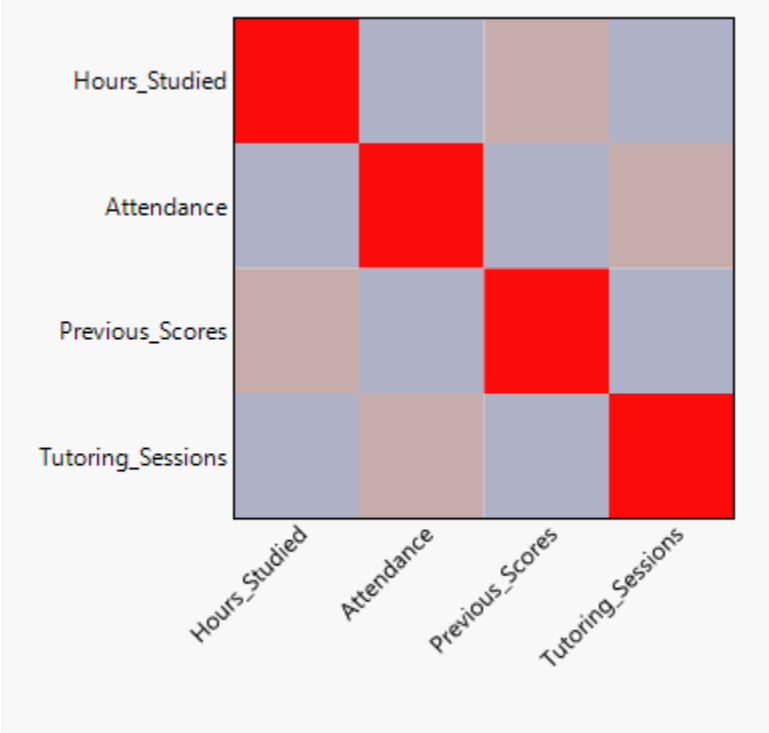


Figure 14 Color Map on Correlations

Finally, we ran a cross validation with an 80/20 split and found that the refit model had very little difference in the two groups, indicating that the model is useful for generalizing new student exam scores for the average student.

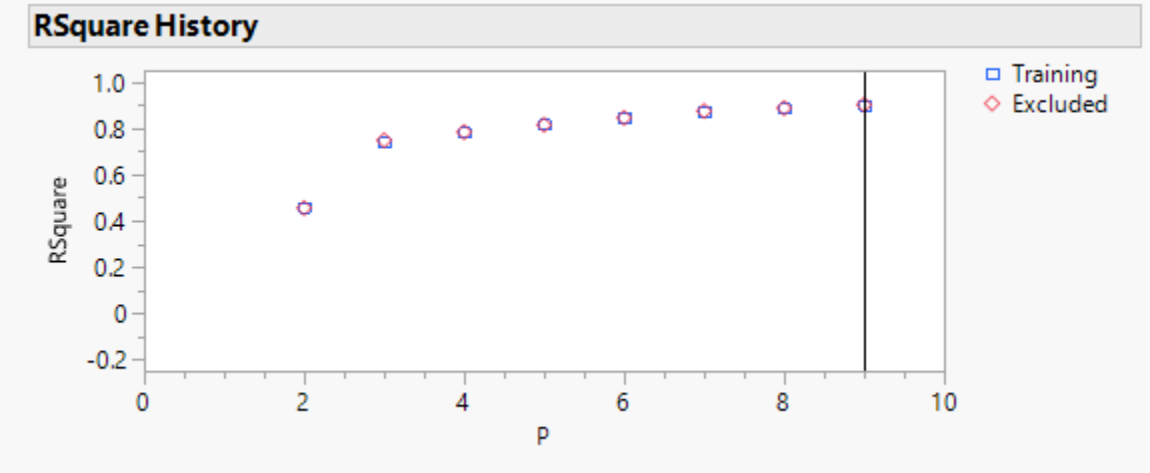


Figure 15 RSquare History Cross Validation Test

Conclusion

In this project, we modeled student exam scores using a combination of behavioral, parental, and environmental factors. After comparing several model structures, our final chosen model included hours studied, attendance, previous exam scores, tutoring sessions, parental involvement, and access to resources. This model explained about 66% of variation in exam scores in the full dataset and nearly 90% of the variation after removing a small cluster (.76% of observations) of high performing outliers.

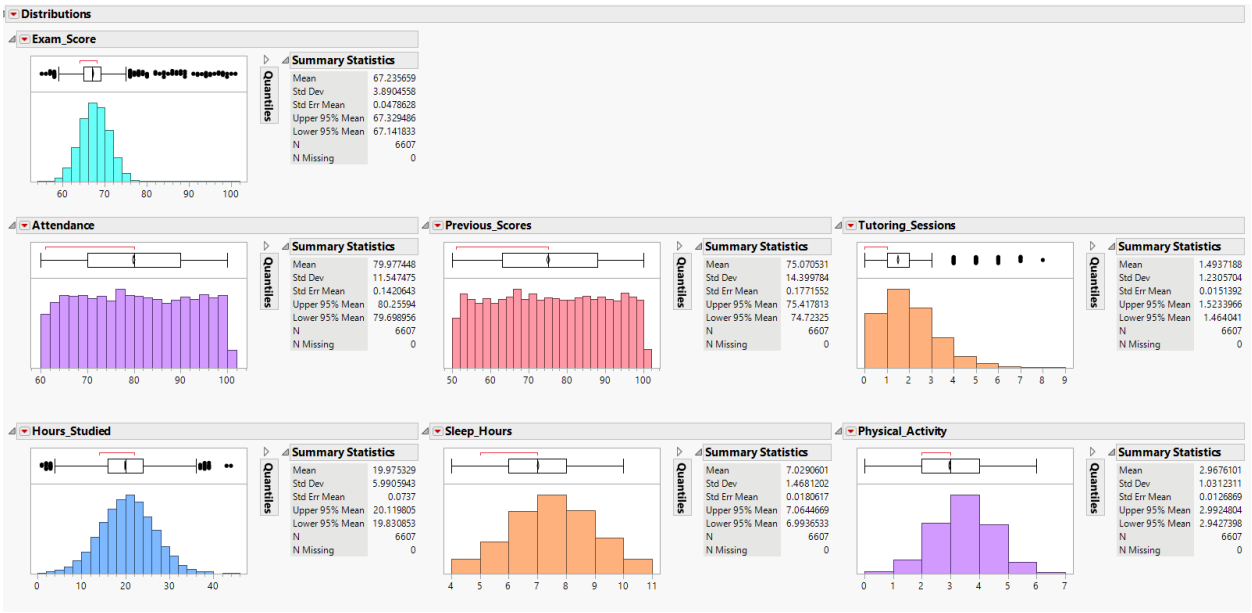
The strongest predictors of exam performance were previous exam scores, hours studied, and attendance. Students who studied more hours per week and attended a higher percentage of classes tended to earn higher exam scores, even after controlling prior performance. Access to resources and parental involvement also mattered. Students with higher parental support and access to resources typically did better than those with fewer resources or lower involvement.

We identified an interesting pattern in the outliers. The observations flagged as outliers did not appear to be random mistakes or data errors, but instead a distinct group of high achieving students whose scores were much higher than the model predicted. Their large positive residuals suggest that there could be additional unobserved factors such as study strategies, course difficulty, or subject matter that help explain why these students perform so well. Removing these observations did make the regression assumptions much better satisfy the assumption of normality without greatly changing the estimated relationship for the large majority of students (over 99%).

Because the data is observational, we cannot claim that these predictors cause higher exam scores. However, the results do highlight several areas where students and schools can potentially focus attention and resources to increase chances of academic success. Encouraging consistent attendance and regular study habits and ensuring students have access to basic learning resources with few barriers, are likely to support better exam scores. Future analysis could enhance the model by adding more factors to the observation to better identify the greatest predictor on students who are high achievers or by fitting different models for different subgroups of students.

Appendix

Individual Distribution Graphs



Individual Frequency Graphs