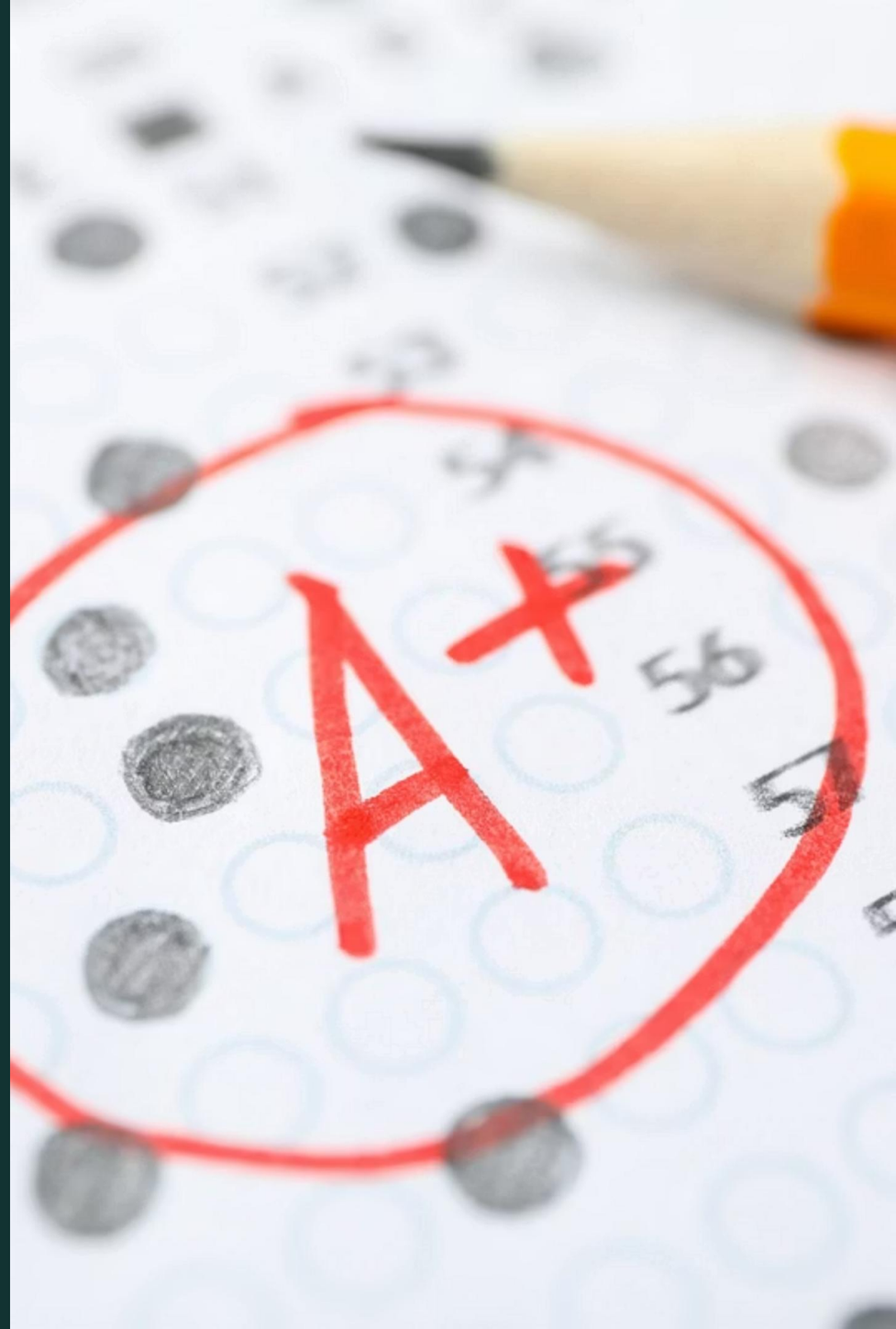
Modeling Exam Scores of Students

Predicting student performance using behavioral, environmental, and parental factors



Research Questions

Strongest Predictors

Which factors have the strongest relationship with exam score?

Combined Effects

Do behavioral, environmental, and parental predictors remain significant when evaluated together?

Model Complexity

Do interaction or nonlinear effects improve prediction accuracy?

Outlier Influence

Are there outliers influencing the results?

Dataset Overview

6,607

Student Observations

Large synthetic sample for robust analysis

19

Predictor Variables

Mix of numeric and categorical data

Response Variable

Exam Score (%) — continuous outcome ranging from 0 to 101

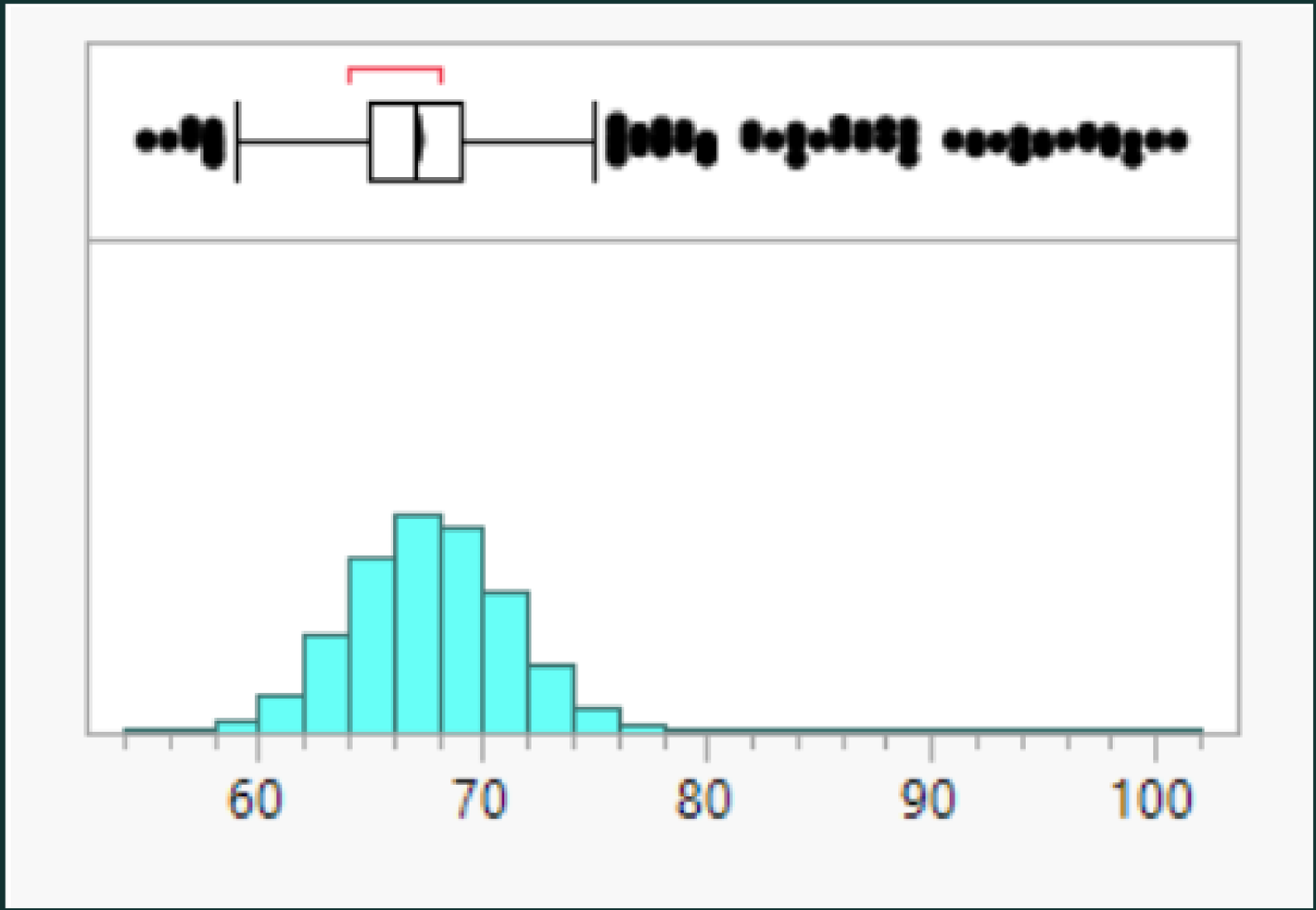
Predictor Categories

- Behavioral: hours studied, attendance, sleep hours, previous exam scores, physical activity
- Support: tutoring sessions, parental involvement, access to resources, parental education levels
- Contextual: income level, disabilities, school type

Data Structure - Categorical

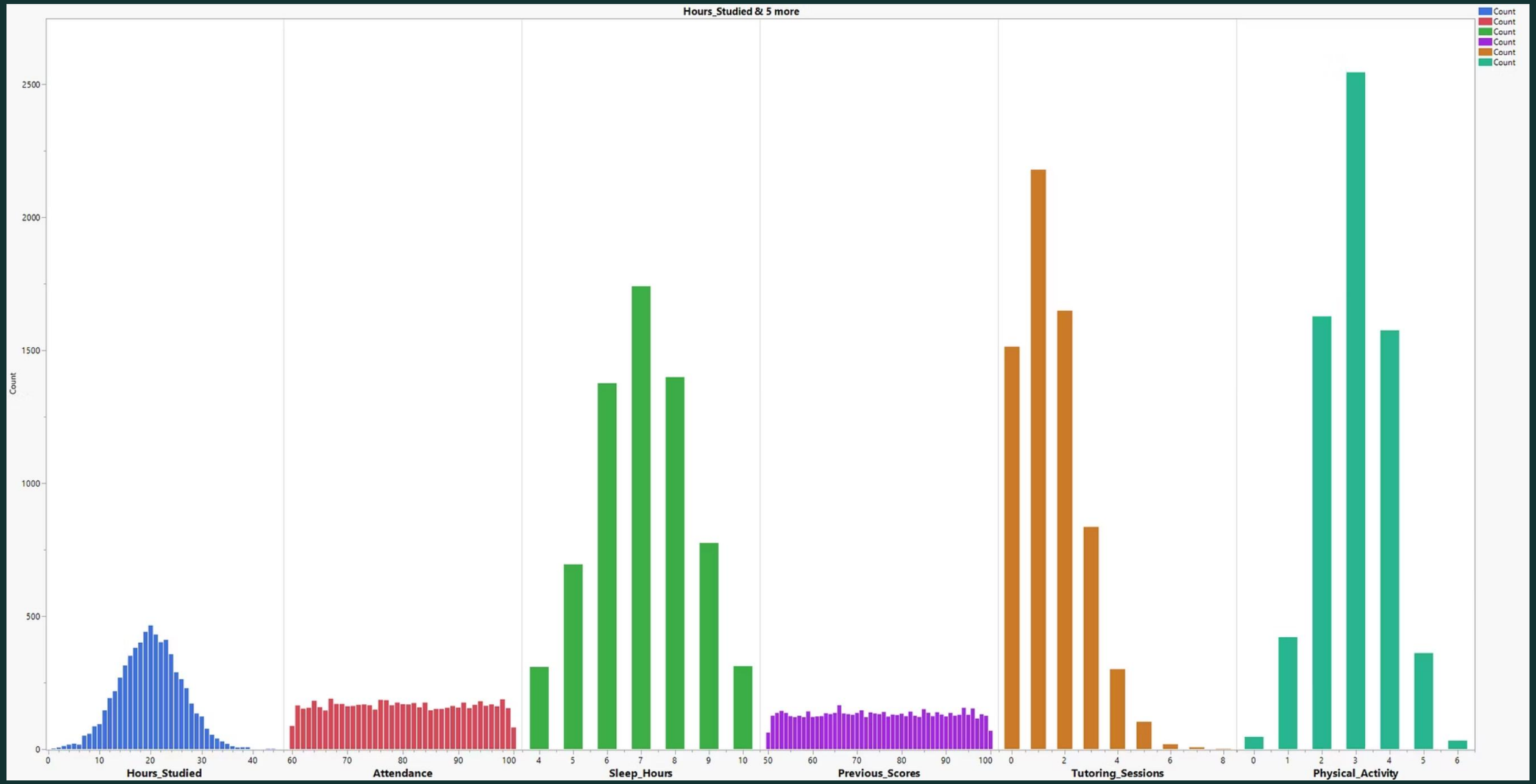
	N	N Missing	Levels
Access To Resources	6607	0	3
Extracurricular Activities	6607	0	2
Motivation Level	6607	0	3
Internet Access	6607	0	2
Family Income	6607	0	3
Teacher Quality	6607	78	3
School Type	6607	0	2
Peer Influence	6607	0	3
Learning Disabilities	6607	0	2
Parental Education Level	6607	90	3
Distance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

Data Structure - Categorical



Response Summary Statistics

- Mean: 67.24
- Std Dev: 3.89
- Std Err Mean: .05
- N: 6607
- N Missing: 0


Hours_Studied & 5 more
Count
Count
Count
Count
Count
Count
Count
2500
2000
1500
1000
500
0
Hours_Studied
Attendance
Sleep_Hours
Previous_Scores
Tutoring_Sessions
Physical_Activity


Exam_Score vs. Hours_Studied & 5 more
Exam_Score
Exam_Score
Exam_Score
Exam_Score
Exam_Score
Exam_Score
Exam_Score
100
90
80
70
60
50
Physical_Activity
Tutoring_Sessions
Previous_Scores
Sleep_Hours
Attendance
Hours_Studied

EDA Takeaways

Strong Positive Relationships

Hours studied, attendance, and previous scores show clear positive correlations with exam performance.

Weaker Patterns

Sleep and physical activity weaker or noisier relationships requiring careful modeling.

High-Achieving Outliers

Identified a distinct group of exceptional performers worth investigating separately.

Modeling Approach

Variable Selection

Stepwise selection method applied to identify most important predictors while avoiding overfitting.

Three Regression Models

Main-effects linear model, interaction model, and quadratic model to test different complexity levels.

Assumption Validation

Checked normality, homoscedasticity, leverage, and outlier influence to ensure model validity.

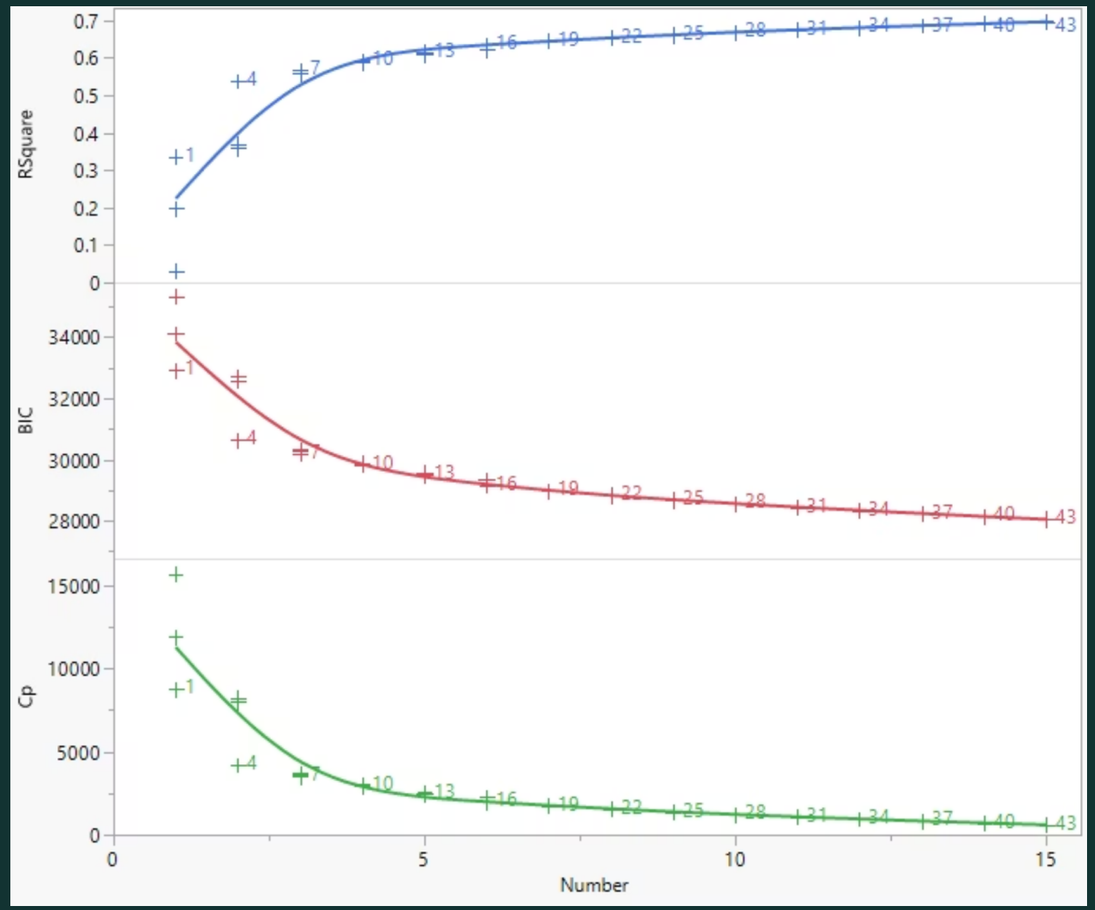
Cross-Validation

80/20 train-test split to assess model generalization and prevent overfitting to sample data.

Variable Selection

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	Attendance	Entered	0.0000	3289.646	0.3367	8758.2	2	32893.7	32913.9
2	Hours_Studied	Entered	0.0000	1961.0.1	0.5374	4181.3	3	30597.1	30624.1
3	Previous_Scores	Entered	0.0000	2985.719	0.5680	3486.1	4	30163.2	30197
4	Tutoring_Sessions	Entered	0.0000	2389.283	0.5924	2930.2	5	29793.6	29834.1
5	Parental_Involvement{Low&Medium-High}	Entered	0.0000	2122.947	0.6142	2436.5	6	29446.1	29493.4
6	Access_to_Resources{Low&Medium-High}	Entered	0.0000	2057.435	0.6352	1958.1	7	29090.2	29144.3
7	Access_to_Resources{Low-Medium}	Entered	0.0000	1037.595	0.6458	1717.8	8	28903.8	28964.6
8	Parental_Involvement{Low-Medium}	Entered	0.0000	838.7662	0.6544	1524	9	28749.3	28816.8
9	Family_Income{Low-Medium&High}	Entered	0.0000	793.3444	0.6625	1340.7	10	28599.6	28673.9
10	Parental_Education_Level{High School&College-Postgraduate}	Entered	0.0000	645.6547	0.6691	1192	11	28475.5	28556.6
11	Motivation_Level{Low-Medium&High}	Entered	0.0000	608.4679	0.6754	1051.9	12	28356.3	28444.1
12	Peer_Influence{Negative-Neutral&Positive}	Entered	0.0000	570.3046	0.6812	920.71	13	28242.6	28337.2
13	Extracurricular_Activities{No-Yes}	Entered	0.0000	522.8195	0.6866	800.63	14	28136.6	28238
14	Distance_from_Home{Far&Moderate-Near}	Entered	0.0000	529.7326	0.6920	678.94	15	28027.3	28135.4
15	Teacher_Quality{Low&Medium-High}	Entered	0.0000	513.2339	0.6972	561.1	16	27919.6	28034.5
16	Learning_Disabilities{Yes-No}	Entered	0.0000	428.4011	0.7016	463.06	17	27828.6	27950.2
17	Peer_Influence{Neutral-Positive}	Entered	0.0000	352.7917	0.7052	382.69	18	27753	27881.3
18	Internet_Access{No-Yes}	Entered	0.0000	352.7888	0.7088	302.31	19	27676.4	27811.4
19	Parental_Education_Level{High School-College}	Entered	0.0000	276.9121	0.7117	239.65	20	27616	27757.8
20	Motivation_Level{Medium-High}	Entered	0.0000	264.1599	0.7144	179.97	21	27557.9	27706.5
21	Family_Income{Medium-High}	Entered	0.0000	266.375	0.7171	119.77	22	27498.8	27654.1
22	Physical_Activity	Entered	0.0000	229.2186	0.7195	68.25	23	27447.7	27609.7
23	Teacher_Quality{Low-Medium}	Entered	0.0000	139.161	0.7209	37.756	24	27417.2	27586
24	Distance_from_Home{Far-Moderate}	Entered	0.0000	71.77172	0.7216	22.998	25	27402.4	27578



Model 1: Main-Effects Linear Model

Selected Predictors

- Hours studied
- Attendance percentage
- Previous exam scores
- Tutoring sessions
- Parental involvement
- Access to resources

Hypothesis Testing

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$

$$H_a : \beta_i = 0; i = 1, 2, 3, 4, 5, 6, 7, 8$$

Model

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

y = Exam Score

x_1 = Hours Studied

x_2 = Attendance

x_3 = Previous Scores

x_4 = Tutoring Sessions

x_5 = {1 if parental involvement = low, else 0}

x_6 = {1 if parental involvement = medium, else 0}

x_7 = {1 if access to resources = low, else 0}

x_8 = {1 if access to resources = medium, else 0}

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*

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*

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

Model 2: Two Way Interaction Model

Model

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

Hypothesis Testing

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \ldots \beta_{68} = 0$$

$$H_a : \beta_i = 0; i = 9, 10, 11, 12$$

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*

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

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	45.018747	1.443429	31.19	<.0001*
Hours_Studied	0.2149257	0.042437	5.06	<.0001*
Attendance	0.1787465	0.015951	11.21	<.0001*
Parental_Involvement[Low]	-2.344612	0.784926	-2.99	0.0028*
Parental_Involvement[Medium]	-0.831753	0.627208	-1.33	0.1848
Access_to_Resources[Low]	-1.603792	0.782749	-2.05	0.0405*
Access_to_Resources[Medium]	-1.256349	0.616722	-2.04	0.0417*
Previous_Scores	0.0349124	0.015768	2.21	0.0269*
Tutoring_Sessions	0.5023968	0.217377	2.31	0.0209*
Hours_Studied*Attendance	0.000526	0.000401	1.31	0.1894
Hours_Studied*Parental_Involvement[Low]	0.0035064	0.01369	0.26	0.7979
Hours_Studied*Parental_Involvement[Medium]	0.0040677	0.010749	0.38	0.7051
Hours_Studied*Access_to_Resources[Low]	0.0181074	0.013667	1.32	0.1852
Hours_Studied*Access_to_Resources[Medium]	0.015304	0.010656	1.44	0.1510
Hours_Studied*Previous_Scores	0.0003122	0.000324	0.96	0.3358
Hours_Studied*Tutoring_Sessions	-4.525e-5	0.003799	-0.01	0.9905
Attendance*Parental_Involvement[Low]	0.0049726	0.007061	0.70	0.4813
Attendance*Parental_Involvement[Medium]	-0.002206	0.005672	-0.39	0.6974
Attendance*Access_to_Resources[Low]	-0.007534	0.007008	-1.07	0.2824
Attendance*Access_to_Resources[Medium]	-0.002241	0.00561	-0.40	0.6896
Attendance*Previous_Scores	0.0001504	0.000169	0.89	0.3747
Attendance*Tutoring_Sessions	0.000755	0.001958	0.39	0.6998
Parental_Involvement[Low]*Access_to_Resources[Low]	0.2419609	0.237725	1.02	0.3088
Parental_Involvement[Low]*Access_to_Resources[Medium]	0.1898932	0.186201	1.02	0.3078
Parental_Involvement[Medium]*Access_to_Resources[Low]	0.2599236	0.185912	1.40	0.1621
Parental_Involvement[Medium]*Access_to_Resources[Medium]	0.2051995	0.151405	1.36	0.1754
Parental_Involvement[Low]*Previous_Scores	-0.004111	0.005654	-0.73	0.4672
Parental_Involvement[Medium]*Previous_Scores	-0.00329	0.004557	-0.72	0.4704
Parental_Involvement[Low]*Tutoring_Sessions	0.0293223	0.064527	0.45	0.6495
Parental_Involvement[Medium]*Tutoring_Sessions	-0.011101	0.053212	-0.21	0.8348
Access_to_Resources[Low]*Previous_Scores	-0.004113	0.005706	-0.72	0.4711
Access_to_Resources[Medium]*Previous_Scores	-0.000712	0.004485	-0.16	0.8738
Access_to_Resources[Low]*Tutoring_Sessions	-0.040457	0.066897	-0.60	0.5454
Access_to_Resources[Medium]*Tutoring_Sessions	0.0644477	0.052675	1.22	0.2212
Previous_Scores*Tutoring_Sessions	-0.001133	0.001574	-0.72	0.4716

Model 3: Second Order Model

Model

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

Hypothesis Testing

$$H_0 : \beta_9 = \beta_10 = \beta_12 = 0$$

$$H_a : \beta_i = 0; i = 9, 10, 11, 12$$

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 Comparison Summary

Model	*MSE*	Adj. RSquare	*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

Best Balance

Model 1 achieves optimal balance of fit and interpretability without unnecessary complexity.

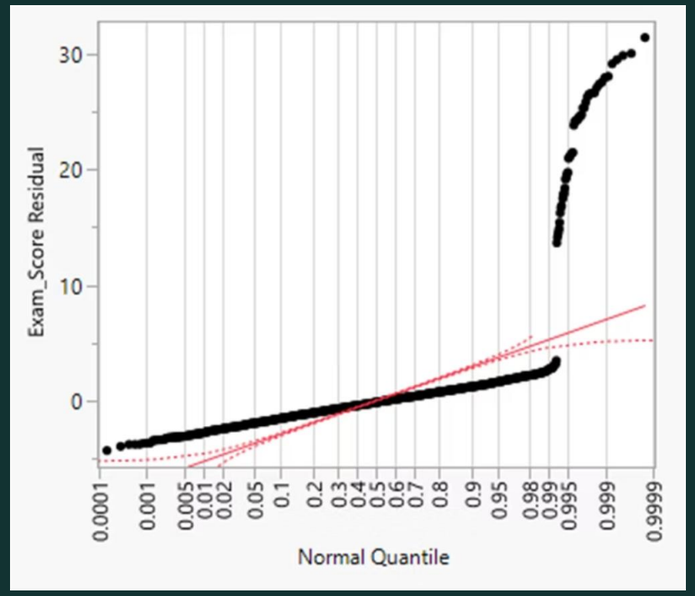
No Added Value

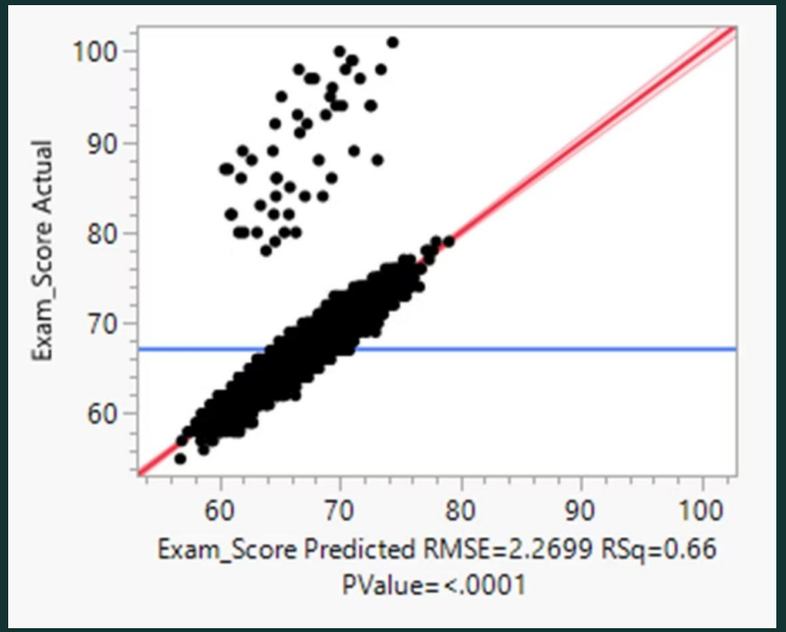
Both interaction and quadratic models fail to improve prediction—partial F-tests confirm this.

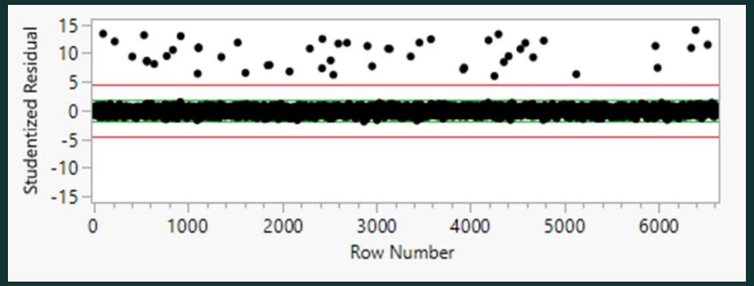
Final Selection

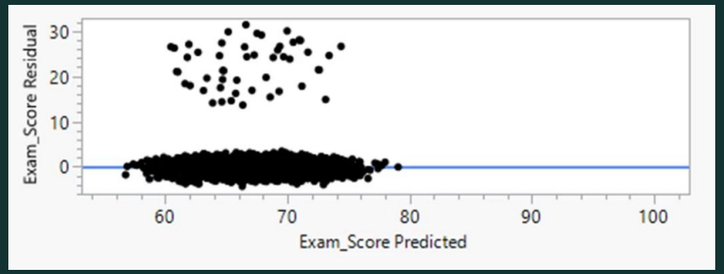
Model 1 selected as final model for interpretation and practical application.

Model Diagnostics









Outlier Detection & Analysis

50

Extreme Outliers

Students identified with unusual residuals

6-14

Studentized Residuals

Standard deviations from predicted

78-101

Score Range

Mostly high-achieving students

Outlier Influence Assessment

Low Leverage

Not unusual in predictor space

Cook's D < 1

Not influential on coefficients

Important Subgroup

Represent meaningful pattern

Conclusion: **T**he outliers do not distort model or invalidate our regression estimates. However, they do represent an important subgroup of high achievers whose success isn't captured by our current variables, warranting further investigation.

Refitting the Model

Model Improvement

0.90

New Adjusted R^2

Dramatic improvement from 0.66

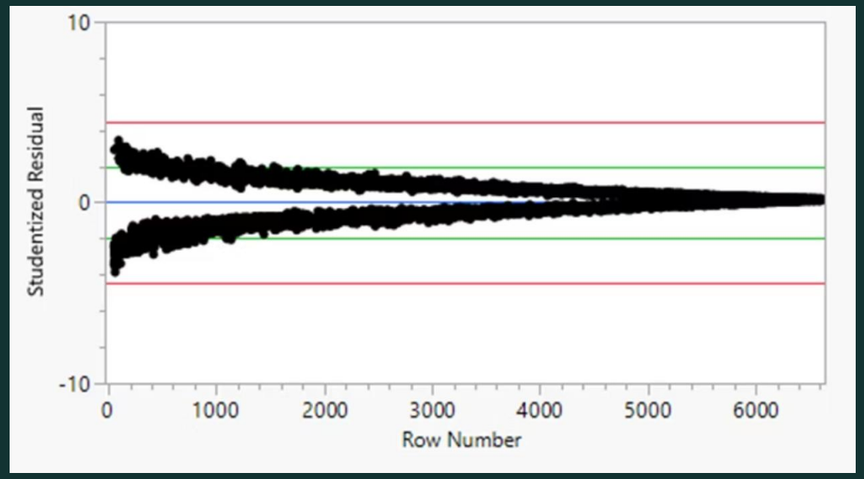
Summary of Fit	
RSquare	0.899068
RSquare Adj	0.898945
Root Mean Square Error	1.064164
Mean of Response	67.06802
Observations (or Sum Wgts)	6557

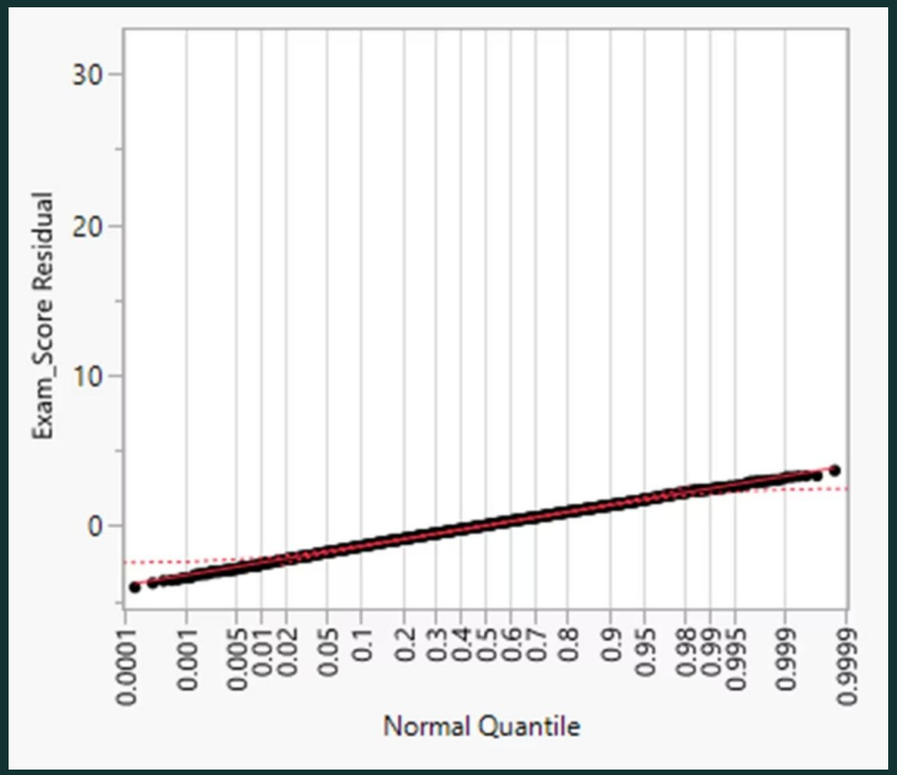
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	66052.418	8256.55	7290.912
Error	6548	7415.246	1.13	Prob > F
C. Total	6556	73467.664		<.0001*

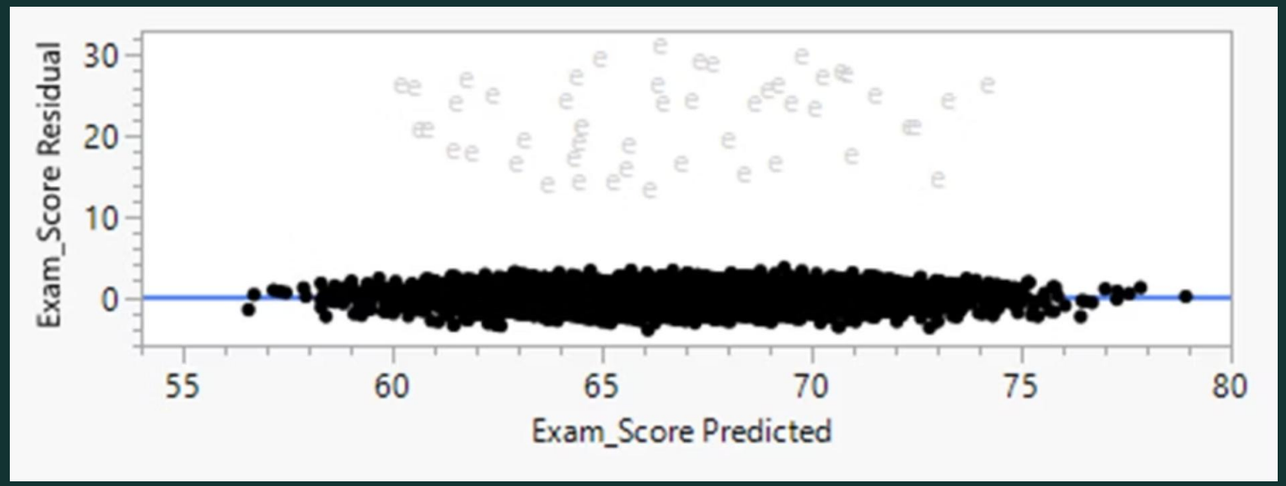
Indicator Function Parameterization				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	42.502821	0.128013	332.02	<.0001*
Hours_Studied	0.297996	0.002197	135.66	<.0001*
Attendance	0.2000058	0.001139	175.66	<.0001*
Parental_Involvement[Low]	-2.007478	0.038166	-52.60	<.0001*
Parental_Involvement[Medium]	-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*

Diagnostic Improvements

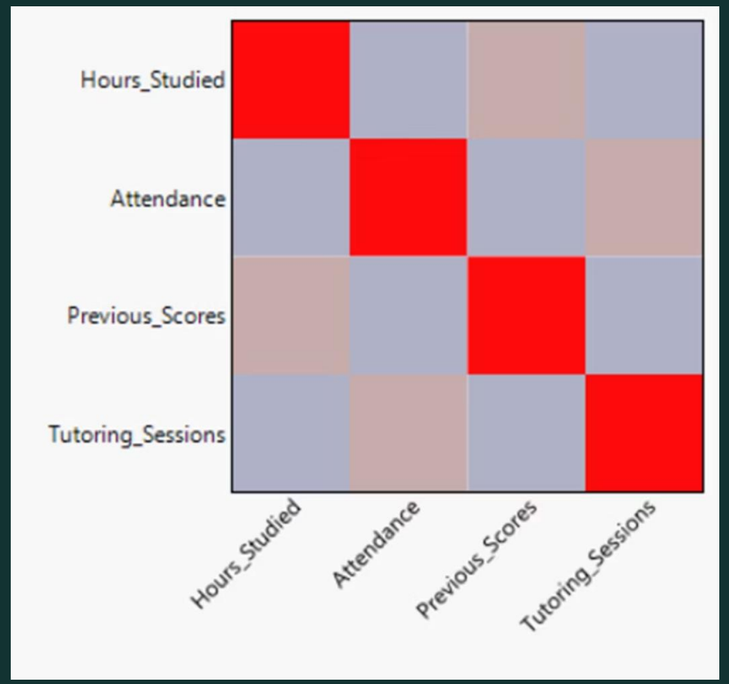
- Residuals more normally distributed
- Greatly reduced skewness
- Homoscedasticity substantially improved
- Clearer linear patterns

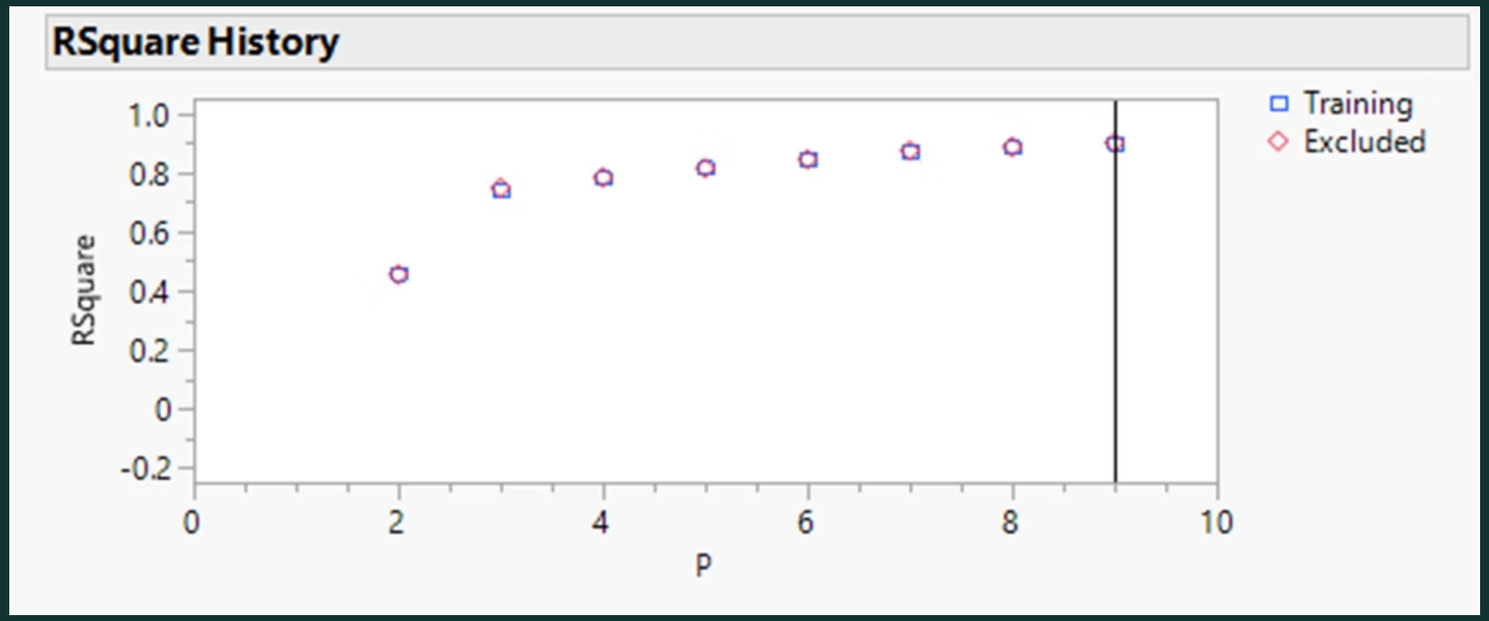






Cross-Validation & Diagnostics





Final Conclusions

Behavioral Factors

- **Previous exam scores**
- **Hours studied**
- **Attendance**

Support Factors

Parental involvement, Tutoring sessions, and access to resources have meaningful but smaller effects on outcomes.

High Achievers

Outliers represent high achieving students not fully explained by measured variables.

Thank You!