Exam Score Prediction For Public School Students

Chhiring Lama, Flora Zang, Jeong Yun Choi, Kate Colvin

Abstract

This research aims to analyse academic performance data from a public school and verify the effectiveness of regression analysis methods in predicting student outcomes. We focus on three critical test scores: Math, Reading, and Writing. After tidying the data, various linear regression models were developed to assess the impact of multiple student characteristics and contextual factors on the test scores. The individual models selected for reading, writing and math score share common main effect variables, with some difference in the interaction modifiers. Our result suggests that failure in test preparation course completion have negative effect on test outcome, while the weekly study hours are beneficial. Similarly, social and family aspects like parents' education, marital status, ethnic group and gender were significantly associated with test scores. The resulting linear regression equations provide insights into future trends and identify the most influential variables affecting academic performance. This study contributes valuable information for designing targeted educational interventions and policies to enhance student success.

Introduction, Background, and Context

Education is the cornerstone of both individual and societal success. Governments and societies have a vested interest in ensuring that learners receive quality education as it opens doors of opportunity for improved quality of life and a progessive society. However, achieving this goal involves addressing the complex challenge of identifying the key components of quality education. Beyond cognitive ability, factors such as socio-economic status, parental involvement, and school environment have recently been appreciated to significantly shape student performance. Despite current literature, there still lacks an absolute model to predict educational outcomes given certain determinants. Understanding those determinants and their specific impact is crucial to educators, policymakers, and researchers aiming to design interventions and policies to enhance student performance.

This analysis aims to: 1. **Construct Predictive Models**: develop robust models for Math, Reading, and Writing scores using an available predictors. 2. **Identify Key Predictors**: Determine predictors that significantly influence each test score and explore interaction. 3. **Model Comparisons**: evaluate the optimal prediction models for each test score and assess the potential for leveraging one test score to improve the prediction of another.

Methods

Data Cleaning and Exploration

The full dataset includes data from 948 students. We created summary tables to describe the distributions of all variables, stratified by gender (Table 1-3). To further investigate the distribution of test scores by subject, we created visualizations including stratified density and boxplots, and a correlation matrix of the test scores (Fig. 1-3). Most variables were missing between 15 to 100 values, spread evenly across the students. Students with missing data were included in all exploratory analysis and visualizations, and were excluded from regression models when necessary. We found a handful of score outliers for each subject, but chose to leave these points in, as they seemed like valid data points (and not errors in data entry). Other minor data cleaning was performed.

Model Development, Diagnostics and Selection

We analyzed covariates with high correlation, we performed a series of chi-square tests and one-way ANOVA to determine whether the covariates were dependent (Table 4-5). These tests helped us identify pairs of variables with the weakest associations, allowing us to select five variables that were most independent of each other. These variables formed the basis of our initial linear model for stepwise-regression model selection. The missing data were dropped because imputed values for categorical variables may complicate the interpretability of the result and it adds uncertainty that needs to be accounted for.

Math Score Prediction

With forward selection, we obtained seven significant predictors. Subsequently,we visualized math score between pairs of categorical variables. Among the pairs where the mean values score are associated within sub-categories, we tested for interaction, and found that weekly study hours and parental marital status has significant combined effect (Fig. 4). Upon examining diagnostic plots for the selected model, the residual plot revealed greater variance for lower values as well as deviation from normality in Q-Q plot (Fig. 5, Table 6). To address this, we applied a transformation of $(Math\ Score + 1)^{1.3}$, as determined by box-cox, to the math scores variable, improving its compatibility with the model (Fig. 6-7, Table 7-9). Step-wise regression also resulted in the same model. Based on the Mallow's Cp and adjusted R^2, an optimal model for math score is suggested to have 15-17 main effect parameters. This aligns with our model selected from forward and backward

model selection. LASSO suggests taking number of siblings into account (Fig. 8, Table 10). So, we decided to cross-validate and compare performances between model with and without number of sibling as a modifier.

Writing Score Prediction

With forward selection, we identified eight significant covariates to predict writing score, among which seven were the same as identified for model predicting math score. We applied similar steps to check interactions and uncovered significant interaction coefficient for `gender:wkly_study_hours` and `lunch_type:wkly_study_hours`(Fig. 9, Table 11-12). The same end model was selected upon step-wise regression selection. The residuals followed homoscedascticity, mean zero assumptions, but deviated from normality assumptions. Therefore, box-cox transformation was used to identify the optimal transformation, and model was re-fitted using $\sqrt{(Writing\ Score)^3}$(Fig. 10-12, Table 13). We also checked for outliers and influential points as a part of the diagnostics. The models were further investigated using test based criteria. Based on the Mallow's Cp and adjusted R^2, an optimal model for writing score will should 13-18 main effect parameters (Fig. 13). This aligns with our model selected from forward and step-wise model selection. Like in math score model, LASSO suggests taking number of siblings into account (Table 14), so, we decided to cross-validate and compare performances between model with and without number of sibling as a modifier and the step-wise model to predict writing score as well.

Reading Score Prediction

We implemented forward selection regression to deduce significant covariates for modeling `reading score`. We applied similar steps to check interactions and uncovered significant interaction coefficient for `gender:wkly_study_hours`. (Fig. 14, Table 15-16). Similar result was selected upon using step-wise regression selection. We observed that the residuals followed the assumptions, except for normality. Therefore, box-cox transformation was used to identify the optimal transformation and the model was refitted. (Fig. 15-17, Table 17). We again checked for outliers and influential points. The model as further deduced using Mallow's Cp and adjusted R^2 for further diagnostics.(Fig. 18). Finally, LASSO demonstrated taking number of siblings into account as did in `writing score`. (Table 18). We again will conduct cross-validation to compare performances between model with and without number of siblings as a modifier to predict reading score.

Model Validation

To evaluate its performance, we conducted 10-fold cross-validation using training datasets, and selected the model based on adjusted R^2 and RMSE values as well as the correlation between predicted and observed outcome.

Results

During the exploratory analysis, we found that most students were the oldest children in their families, had between one to three siblings, rode the bus to school, and had standard (not free/reduced) lunch. The distributions for Math, Reading, and Writing test scores all had similar distributions (Fig. 1), were slightly left-skewed. These distributions were also relatively consistent after grouping by the amount of hours students spent studying weekly (<5, 5-10, >10 hours) (Fig. 2). Within each subject, the median score increased as the amount of time spent studying increased, with the lowest study group having the lowest scores in all subjects. Between the three study groups, there was a statistically significant difference in the distributions (Table 3). Finally, Math, Reading, and Writing scores were all heavily correlated (Fig. 3).

All of the models without sibling counts have lower root mean square error and higher R^2 (excluding reading score) has similar predictive ability as its corresponding more complex model (Table 19-21, Fig. 19-21).

In general, students who spend more time studying and complete test preparation courses do better across math, reading and writing sections. Female students have better scores in reading and writing whereas male students do better in math. Parental information like education and marital status are significant predictors whereby children of educated parents and married parents perform better in the test than children of parent with some high school education and of single parents respectively. Students belonging to ethnic group D and E are performing better across the subjects compared to students from ethnic group A. Getting reduced/free lunch, a proxy for student's socio-economic situation, is associated with poorer test outcome. These observations are made after accounting for other variables in the model. The strong association of the scores with financial, parental socio-economic status and demographics alludes that the support level and time afforded by the student in academics reflects on their test performance.

The comparison of prediction models for each test score revealed that while most predictors are

universally significant (weekly study hours, test preparation, status, lunch type, parents' education and marital status, ethnic group and gender), others are subject-specific. Writing score's prediction is also informed by an additional covariate - regular sport practice which is associated with improved score. The individual models for the three scores differ in their interaction terms. The combined effect of parents' marital status and weekly study hours is significant in predicting both math score, while weekly study hours' combined effect with male gender and with lunch type is significantly associated with writing score. Lastly, the combined effect of gender and weekly study hours is significantly associated with reading score. Additionally, levering one test score to predict another showed potential, particularly because all significant predictors, except gender, have common direction of influence on multiple outcomes.

Conclusion

The predictive models developed in this study offer a framework for educators and policymakers to quantify and target key factors that can improve student performance. By understanding the relative importance of various predictors and their interaction, educational interventions can be more effectively designed and implemented. Future research should consider expanding the dataset to include a more diverse student population and exploring additional predictors that may contribute to academic success. Moreover, the application of other advanced statistical and machine learning techniques could further enhance the predictive accuracy and robustness of the models.

In conclusion, this report underscores the importance of a comprehensive approach to analysing academic performance, taking into account a wide range of student characteristics and contextual factors. By holistically addressing the given predictors, we hope to offer meaningful recommendations to the ongoing efforts to enhance the quality of education and support student success through informed, data-driven decision-making.

Contributions

Chhiring:- Math and writing score predictive model, results and methods sections, and compilation of the report.

Flora: Writing up report structure for the group: abstract, introduction and conclusion section.

Jeong Yun: Reading score predictive model, methods and model development section in the report.

Kate: Data cleaning; Exploratory analysis - tables and figures; methods and results sections.

Table 1: Summary of Student Demographic Variables (N=948)

Characteristic	**Overall** N = 948[1]	**female** N = 488[1]	**male** N = 460[1]
Ethnic Group			
group A	80 (9.0%)	32 (7.0%)	48 (11%)
group B	171 (19%)	91 (20%)	80 (19%)
group C	277 (31%)	156 (34%)	121 (28%)
group D	237 (27%)	117 (25%)	120 (28%)
group E	124 (14%)	63 (14%)	61 (14%)
Unknown	59	29	30
Parents' Education			
some high school	163 (18%)	86 (18%)	77 (18%)
associate's degree	198 (22%)	101 (22%)	97 (23%)
bachelor's degree	104 (12%)	57 (12%)	47 (11%)
high school	176 (20%)	83 (18%)	93 (22%)
master's degree	55 (6.1%)	33 (7.1%)	22 (5.1%)
some college	199 (22%)	107 (23%)	92 (21%)
Unknown	53	21	32
Lunch Type			
standard	617 (65%)	309 (63%)	308 (67%)
free/reduced	331 (35%)	179 (37%)	152 (33%)
Parents' Marital Status			
single	213 (24%)	120 (26%)	93 (21%)
divorced	146 (16%)	75 (16%)	71 (16%)
married	516 (57%)	255 (55%)	261 (60%)
widowed	24 (2.7%)	12 (2.6%)	12 (2.7%)
Unknown	49	26	23
Practice Sport			
never	112 (12%)	57 (12%)	55 (12%)
regularly	343 (37%)	177 (37%)	166 (37%)
sometimes	477 (51%)	249 (52%)	228 (51%)
Unknown	16	5	11
Oldest Child (Yes/No)	604 (66%)	321 (68%)	283 (64%)
Unknown	30	13	17
# Siblings	2 (1, 3)	2 (1, 3)	2 (1, 3)
Unknown	46	28	18
Transport Means			
school_bus	509 (60%)	268 (61%)	241 (60%)
private	337 (40%)	174 (39%)	163 (40%)
Unknown	102	46	56

[1] n (%); Median (Q1, Q3)

Table 2: Summary of Student Academic Variables (N=948)

Characteristic	**Overall** N = 948[1]	**female** N = 488[1]	**male** N = 460[1]
Test Prep			
completed	322 (36%)	162 (35%)	160 (37%)
none	571 (64%)	297 (65%)	274 (63%)
Unknown	55	29	26
Weekly Study Hours			
< 5	253 (28%)	131 (28%)	122 (28%)
5-10	508 (56%)	263 (56%)	245 (56%)
> 10	150 (16%)	76 (16%)	74 (17%)
Unknown	37	18	19
Math Score	66 (56, 76)	64 (53, 74)	69 (59, 79)
Reading Score	70 (59, 80)	73 (63, 83)	65 (55, 76)
Writing Score	68 (57, 79)	74 (63, 83)	64 (53, 74)

[1] n (%); Median (Q1, Q3)

Table 3: Academic Variables by Time Spent Studying (N=911)

Characteristic	**Overall** N = 911[1]	**< 5** N = 253[1]	**5-10** N = 508[1]	**> 10** N = 150[1]	**p-value**[2]
Test Prep					0.12
completed	314 (37%)	76 (32%)	175 (37%)	63 (43%)	
none	545 (63%)	160 (68%)	300 (63%)	85 (57%)	
Unknown	52	17	33	2	
Math Score	66 (56, 76)	62 (55, 73)	67 (56, 78)	70 (57, 79)	<0.001
Reading Score	70 (59, 80)	66 (57, 77)	71 (60, 80)	71 (60, 80)	0.025
Writing Score	68 (57, 79)	66 (55, 76)	70 (58, 79)	71 (58, 79)	0.020

[1] n (%); Median (Q1, Q3)
[2] Pearson's Chi-squared test; Kruskal-Wallis rank sum test

Figure 1: Distributions of Test Score by Subject

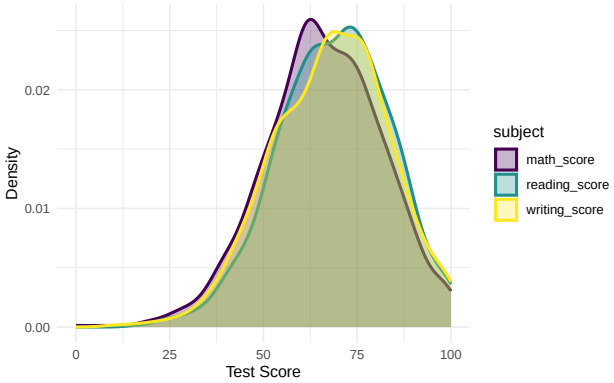

Figure 2: Distributions of Test Scores by Subject and Weekly Study H

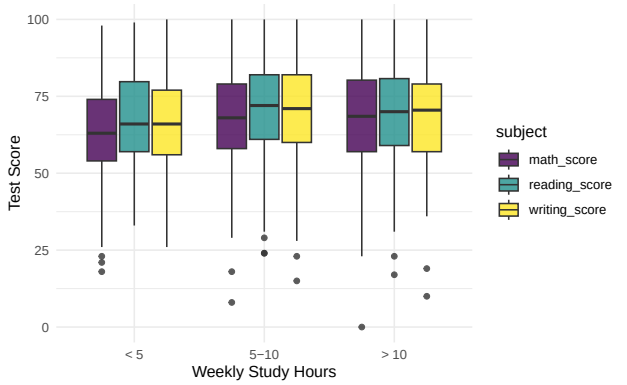

Figure 3: Correlation Between Math, Reading and Writing Score

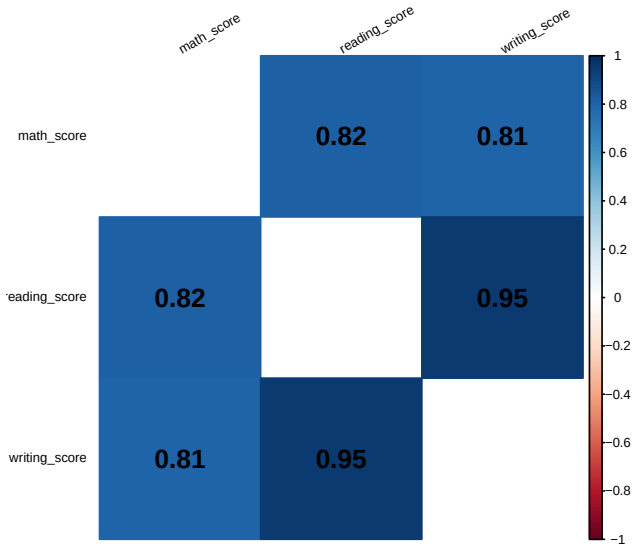

Table 4: Chi-Squared Test: Top 2 results (NS)

statistic	p.value	group
15.206	0.0553	ethnic_group:wkly_study_hours
9.385	0.0947	parent_educ:transport_means
9.282	0.0983	parent_educ:test_prep
7.462	0.1130	gender:ethnic_group
4.251	0.1190	test_prep:wkly_study_hours

Table 5: ANOVA: Number of Siblings v/s Other Covariates (<0.05)

term	df	statistic	p.value
is_first_child	1	16.366	5.68e-05
wkly_study_hours	2	3.024	4.91e-02

Figure 4: Interaction between Covariates (Significant in Prediction)

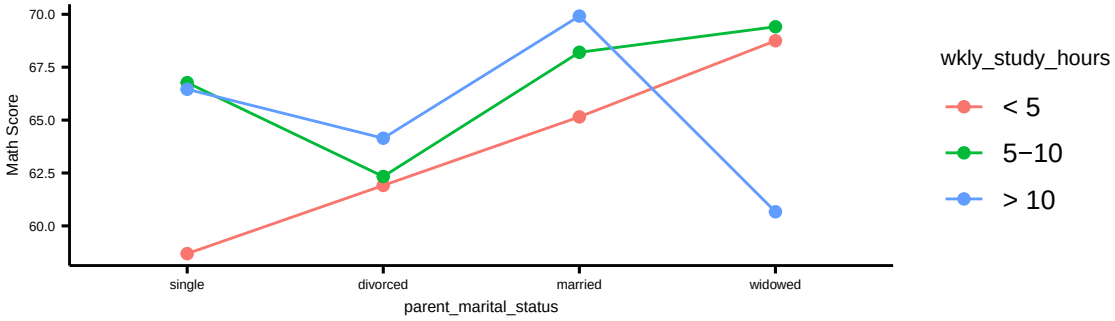

Figure 5: Model from Forward Selection

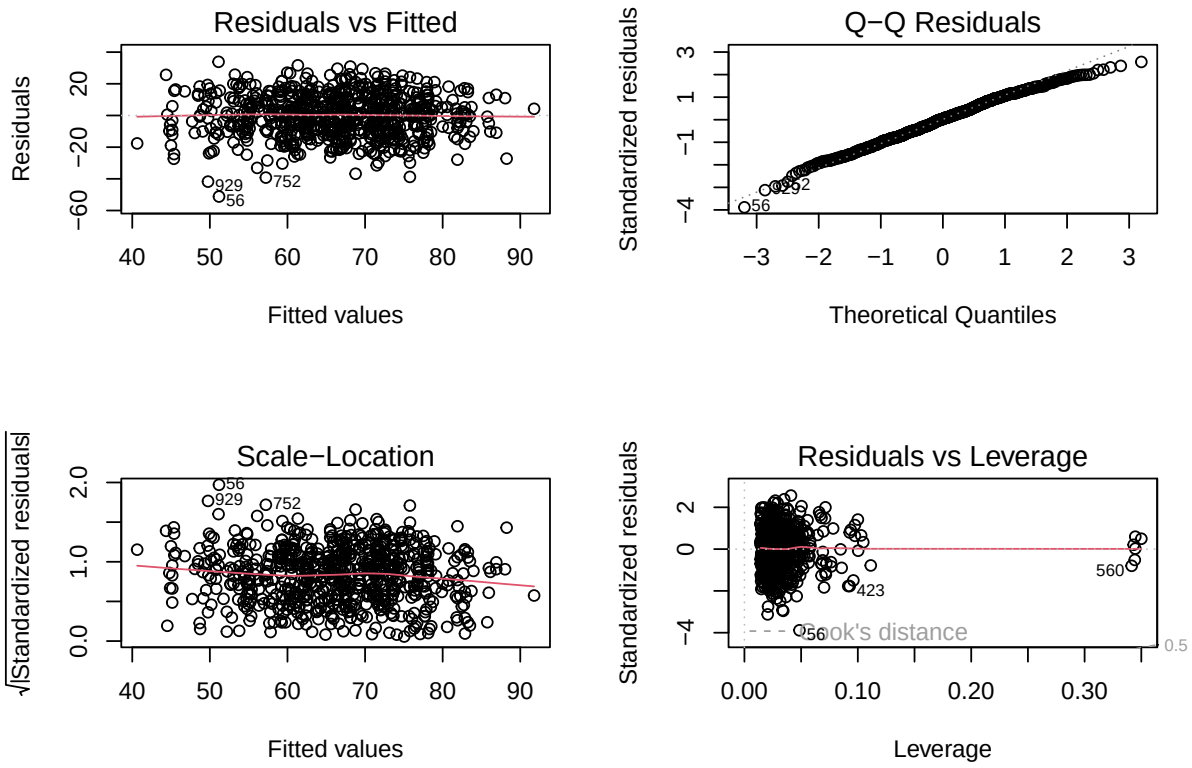

Table 6: Normality Test

statistic	p.value	method
0.9944663	0.010165	Shapiro-Wilk normality test

Figure 6: Box−cox Likelihood

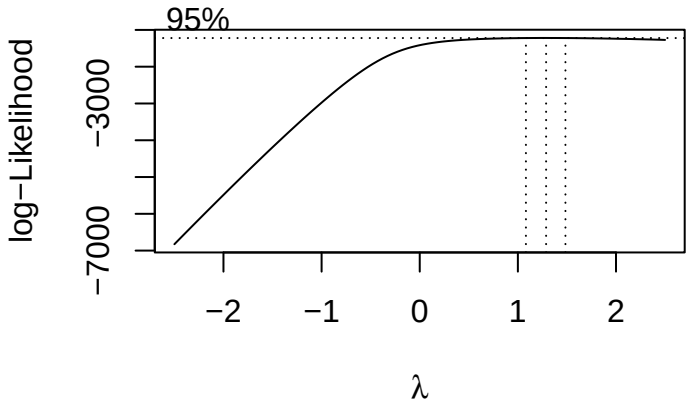

Figure 7: Math Model Assumptions After Transformation

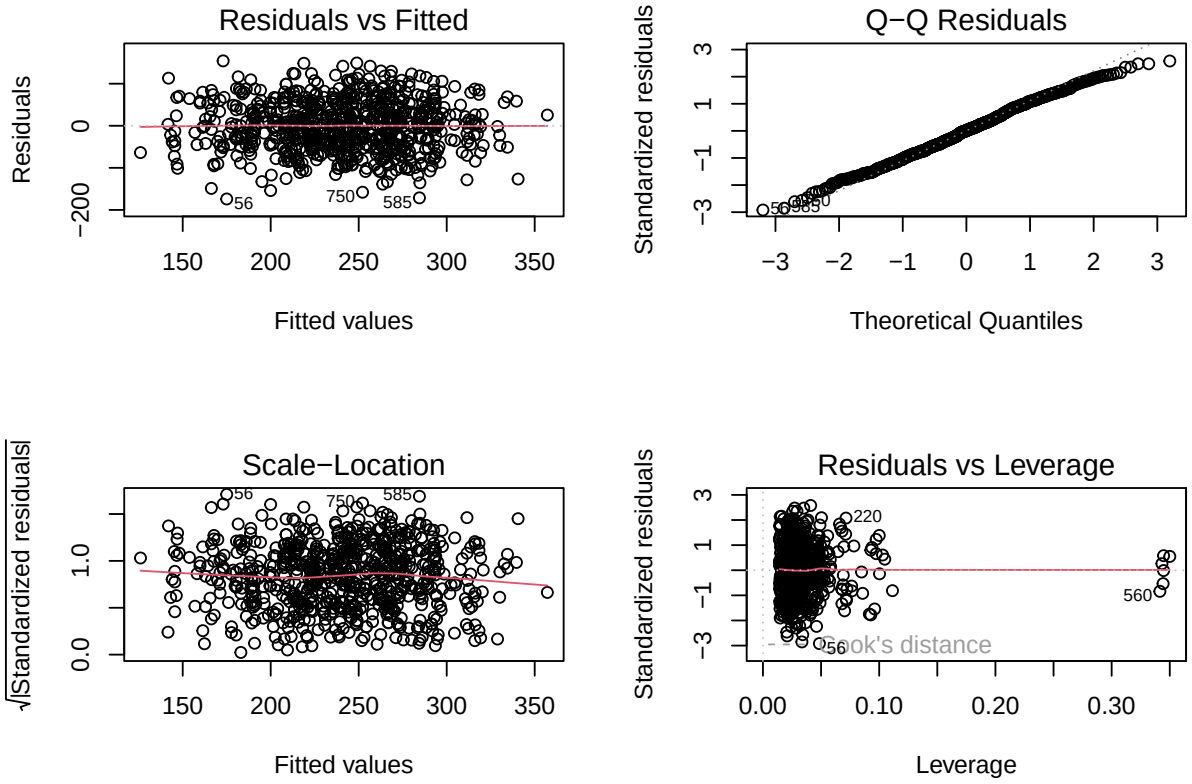

Table 7: Normality Test After Transformation

statistic	p.value	method
0.9958945	0.0558016	Shapiro-Wilk normality test

Table 8: Forward Selection: Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	59.45	2.95	20.17	0.00e+00
lunch_typefree/reduced	-12.25	1.07	-11.47	5.15e-28
test_prepnone	-6.14	1.07	-5.75	1.35e-08
gendermale	5.00	1.03	4.86	1.47e-06
ethnic_groupgroup B	0.57	2.07	0.28	7.82e-01
ethnic_groupgroup C	1.00	1.95	0.51	6.09e-01
ethnic_groupgroup D	4.93	1.97	2.50	1.27e-02
ethnic_groupgroup E	10.31	2.18	4.74	2.62e-06
parent_educassociate's degree	4.53	1.62	2.80	5.29e-03
parent_educbachelor's degree	5.49	1.89	2.91	3.78e-03
parent_educhigh school	-0.63	1.64	-0.39	7.00e-01
parent_educmaster's degree	7.08	2.35	3.01	2.69e-03
parent_educsome college	4.33	1.62	2.67	7.84e-03
parent_marital_statusdivorced	4.46	3.16	1.41	1.58e-01
parent_marital_statusmarried	8.06	2.31	3.48	5.26e-04
parent_marital_statuswidowed	14.38	8.10	1.78	7.62e-02
wkly_study_hours5-10	7.67	2.42	3.17	1.61e-03
wkly_study_hours> 10	10.12	3.19	3.17	1.56e-03
parent_marital_statusdivorced:wkly_study_hours5-10	-7.40	3.83	-1.93	5.37e-02
parent_marital_statusmarried:wkly_study_hours5-10	-5.97	2.87	-2.08	3.80e-02
parent_marital_statuswidowed:wkly_study_hours5-10	-8.55	9.14	-0.94	3.50e-01
parent_marital_statusdivorced:wkly_study_hours>10	-14.20	5.15	-2.76	5.96e-03

term	estimate	std.error	statistic	p.value
parent_marital_statusmarried:wkly_study_hours> 10	-8.22	3.78	-2.18	2.98e-02
parent_marital_statuswidowed:wkly_study_hours> 10	-16.81	11.52	-1.46	1.45e-01

Table 9: Forward Selection: Model Summary

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.32	0.29	14	14	1.8e-43	23	-2879.706	5809	5924	126790.8	695	719

Figure 8: Test−based Procedures For Math Score

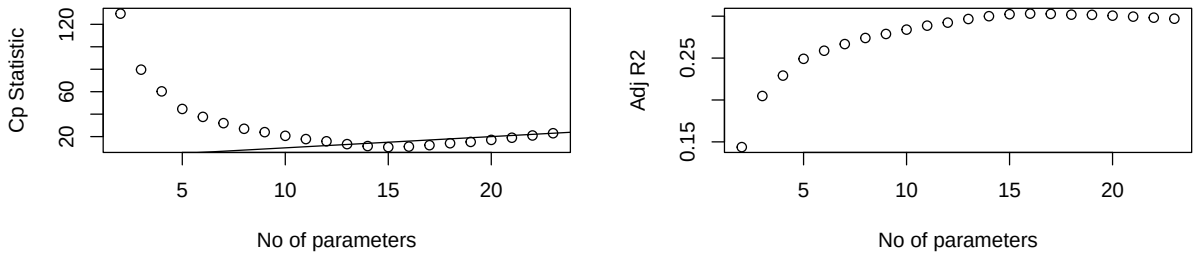

Table 10: LASSO Model For Math Score

term	step	estimate	lambda	dev.ratio
(Intercept)	1	65.269	0.05	0.004
nr_siblings	1	0.658	0.05	0.004

Figure 9: Interaction Between Covariates (Significant in Prediction)

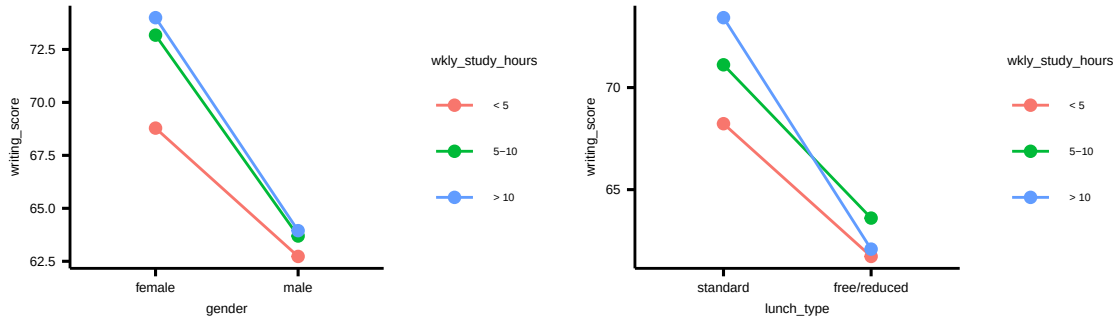

Table 11: Forward Selection: Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	568.09	35.27	16.11	1.00e-49
test_prepnone	-117.48	12.25	-9.59	1.61e-20
gendermale	-68.47	22.35	-3.06	2.27e-03
lunch_typefree/reduced	-97.08	22.94	-4.23	2.64e-05
parent_educassociate's degree	68.07	18.45	3.69	2.43e-04
parent_educbachelor's degree	99.07	21.40	4.63	4.39e-06
parent_educhigh school	-7.95	18.67	-0.43	6.71e-01
parent_educmaster's degree	151.74	27.12	5.60	3.19e-08
parent_educsome college	64.77	18.67	3.47	5.56e-04
ethnic_groupgroup B	-11.32	23.63	-0.48	6.32e-01
ethnic_groupgroup C	7.60	22.30	0.34	7.33e-01
ethnic_groupgroup D	65.57	22.66	2.89	3.93e-03
ethnic_groupgroup E	64.00	24.99	2.56	1.07e-02
parent_marital_statusdivorced	-25.76	18.76	-1.37	1.70e-01
parent_marital_statusmarried	46.44	14.25	3.26	1.18e-03
parent_marital_statuswidowed	63.62	39.53	1.61	1.08e-01
practice_sportregularly	44.60	19.25	2.32	2.08e-02
practice_sportsometimes	38.46	18.61	2.07	3.92e-02

term	estimate	std.error	statistic	p.value
wkly_study_hours5-10	58.65	21.93	2.68	7.65e-03
wkly_study_hours> 10	80.58	29.73	2.71	6.88e-03
gendermale:wkly_study_hours5-10	-63.27	27.40	-2.31	2.13e-02
gendermale:wkly_study_hours> 10	-67.54	36.09	-1.87	6.17e-02
lunch_typefree/reduced:wkly_study_hours5-10	-6.36	28.40	-0.22	8.23e-01
lunch_typefree/reduced:wkly_study_hours> 10	-77.34	37.21	-2.08	3.80e-02

Table 12: Forward Selection: Model Summary

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.38	0.36	150	18	0	23	-4542.314	9135	9249	16022165	682	706

Figure 10: Writing Score Model from Forward Selection

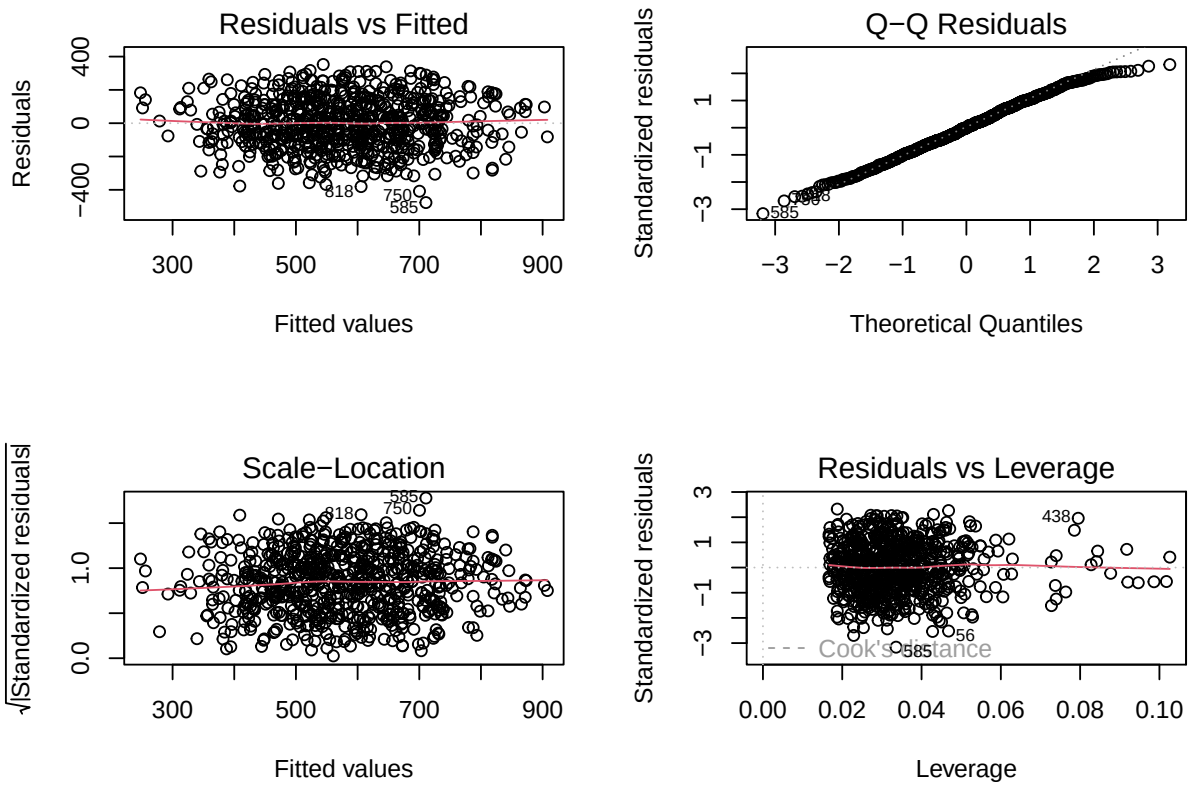

Table 13: Writing Score Model Normality Test

statistic	p.value	method
0.9897292	7.61e-05	Shapiro-Wilk normality test

Figure 11: Box−cox Likelihood

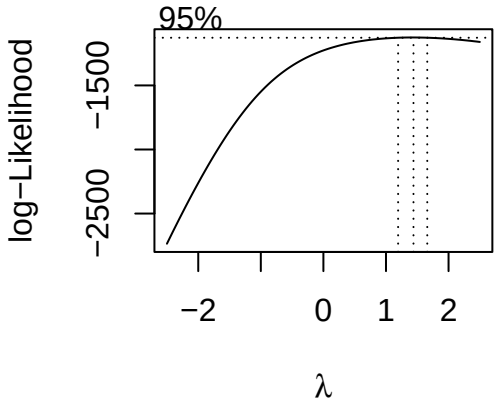

Figure 12: Assumptions for Writing Score Model After Transformation

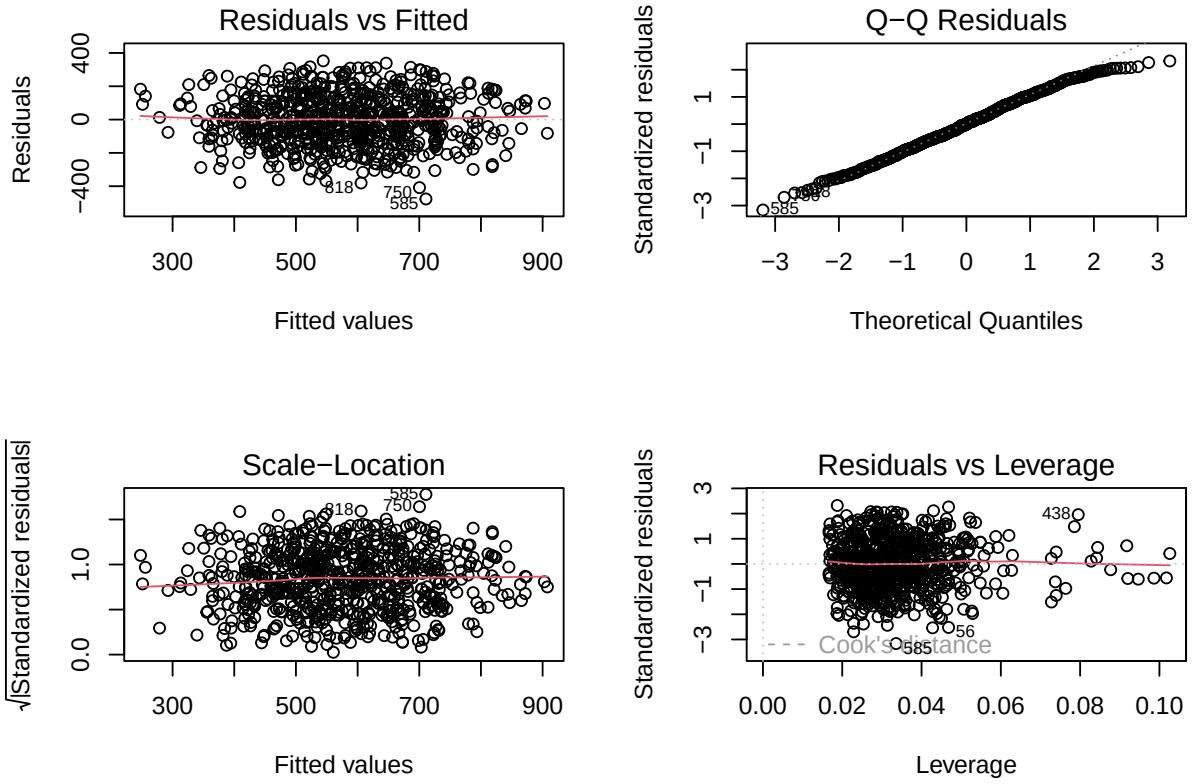

Figure 13: Test−based Procedures For Writing Score

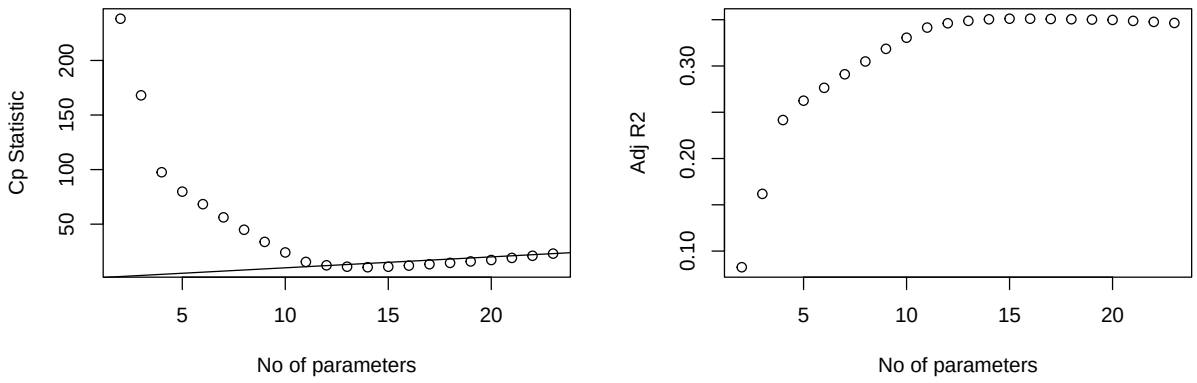

Table 14: LASSO Model For Writing Score

term	step	estimate	lambda	dev.ratio
(Intercept)	1	565.413	0.501	0.005
nr_siblings	1	8.268	0.501	0.005

Figure 14: Interaction Between Covariates for Reading Score (Significant in Prediction)

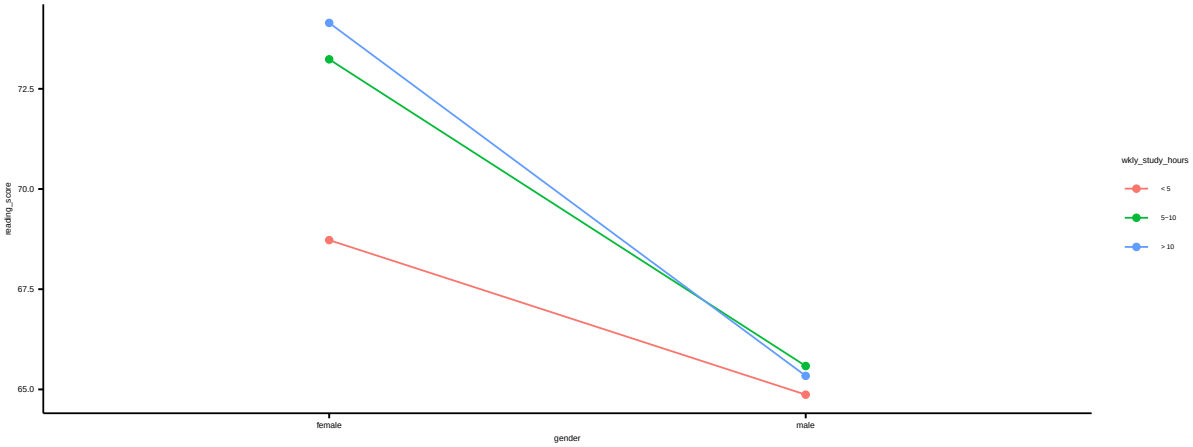

Table 15: Forward Selection: Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	379.44	19.09	19.87	0.00e+00
gendermale	-27.51	13.67	-2.01	4.46e-02
test_prepnone	-49.90	7.52	-6.64	6.62e-11
lunch_typefree/reduced	-59.38	7.52	-7.90	1.14e-14
parent_educassociate's degree	31.20	11.26	2.77	5.76e-03
parent_educbachelor's degree	47.41	13.21	3.59	3.54e-04
parent_educhigh school	-6.49	11.43	-0.57	5.70e-01
parent_educmaster's degree	62.13	16.39	3.79	1.63e-04
parent_educsome college	22.73	11.36	2.00	4.58e-02
ethnic_groupgroup B	-8.57	14.40	-0.60	5.52e-01
ethnic_groupgroup C	-5.89	13.62	-0.43	6.65e-01
ethnic_groupgroup D	18.40	13.78	1.34	1.82e-01
ethnic_groupgroup E	33.01	15.17	2.18	2.99e-02
is_first_childyes	10.66	7.58	1.41	1.60e-01
parent_marital_statusdivorced	-12.99	11.54	-1.13	2.61e-01
parent_marital_statusmarried	23.68	8.76	2.70	7.04e-03
parent_marital_statuswidowed	26.15	23.42	1.12	2.65e-01

term	estimate	std.error	statistic	p.value
wkly_study_hours5-10	31.55	11.52	2.74	6.33e-03
wkly_study_hours> 10	31.02	15.42	2.01	4.46e-02
gendermale:wkly_study_hours5-10	-35.28	16.77	-2.10	3.57e-02
gendermale:wkly_study_hours> 10	-46.50	21.91	-2.12	3.41e-02

Table 16: Forward Selection: Model Summary

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.27	0.25	93	13	2e-35	20	-4133.14	8310	8410	5863823	675	696

Figure 15: Reading Score Model from Forward Selection

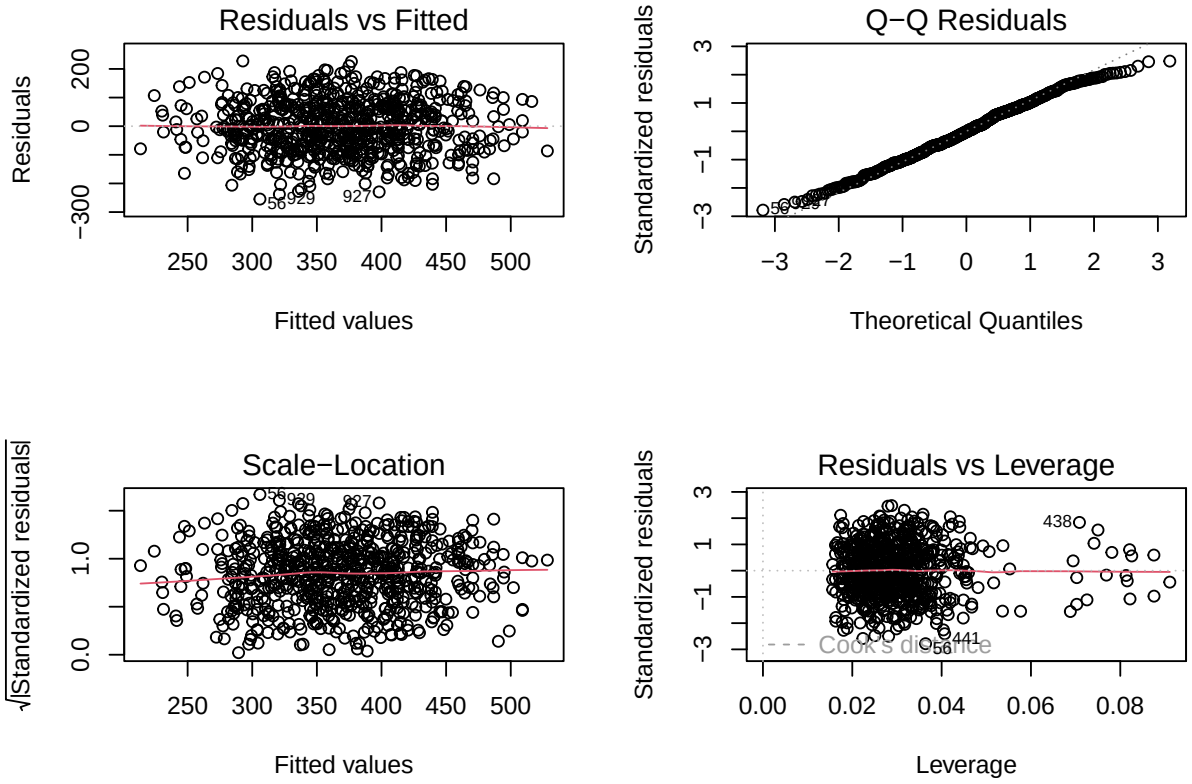

Table 17: Reading Score Model Normality Test

statistic	p.value	method
0.993125	0.0027654	Shapiro-Wilk normality test

Figure 16: Box−cox Likelihood

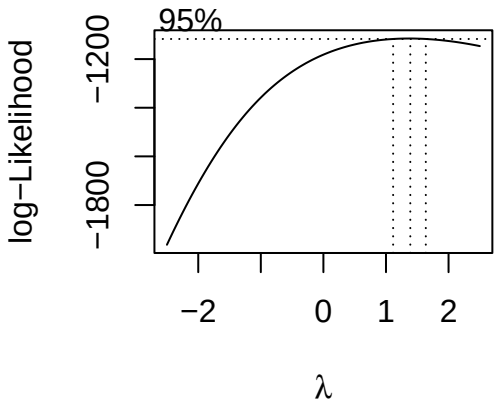

Figure 17: Assumptions for Reading Score Model After Transformation

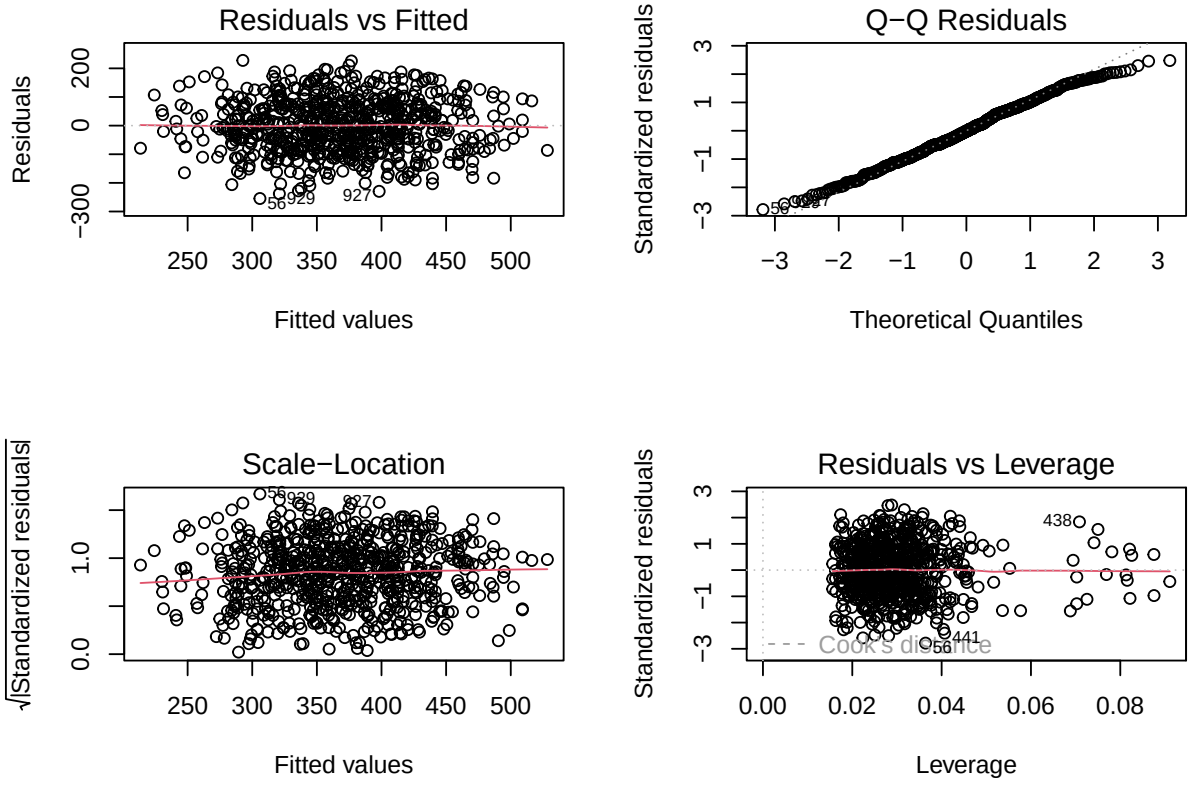

Figure 18: Test−based Procedures For Reading Score

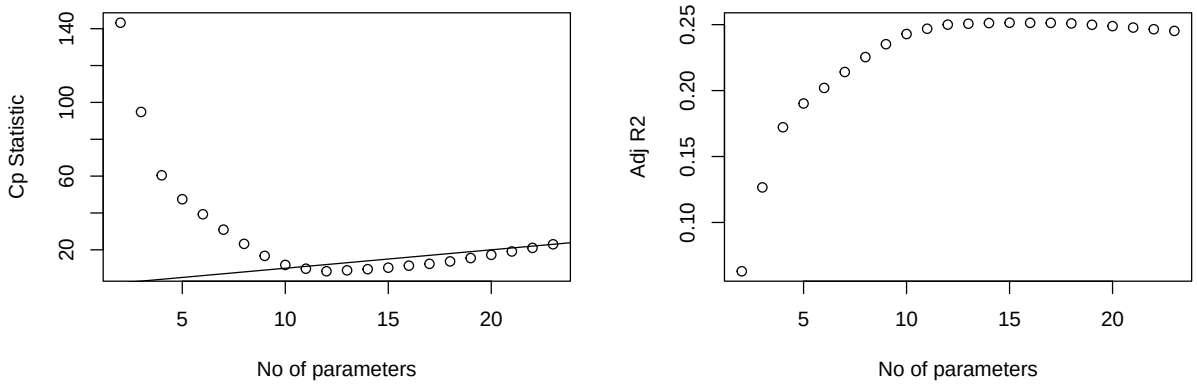

Table 18: LASSO Model For Reading Score

term	step	estimate	lambda	dev.ratio
(Intercept)	1	363.898	0.398	0.002
nr_siblings	1	3.176	0.398	0.002

Figure 19: Correlation of Observed and Predicted Math Score

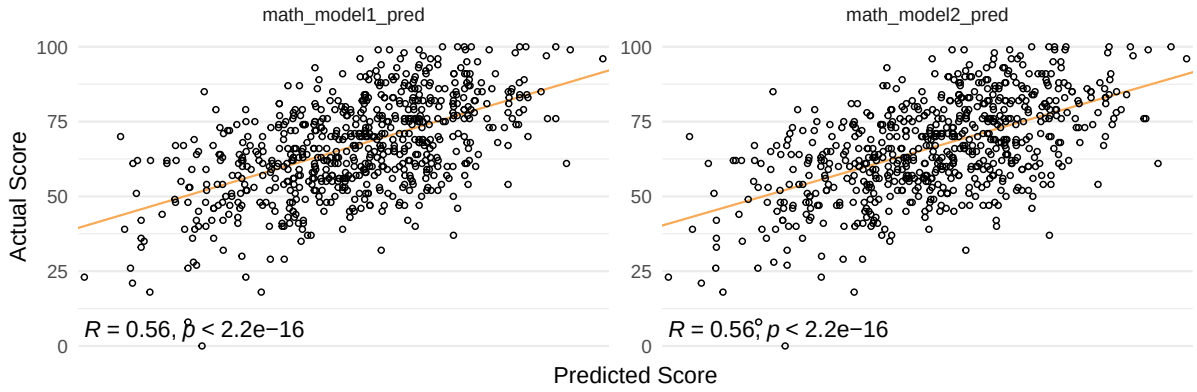

Table 19: Performance matrices of the 2 Models in Predicting (Math Score + 1)ˆ1.3

model_id	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
w/o no. of sibling	62.188	0.283	51.084	3.817	0.060	3.556
w/ no. of sibling	62.479	0.277	51.456	4.739	0.085	3.737

Figure 20: Correlation of Observed and Predicted Writing Score

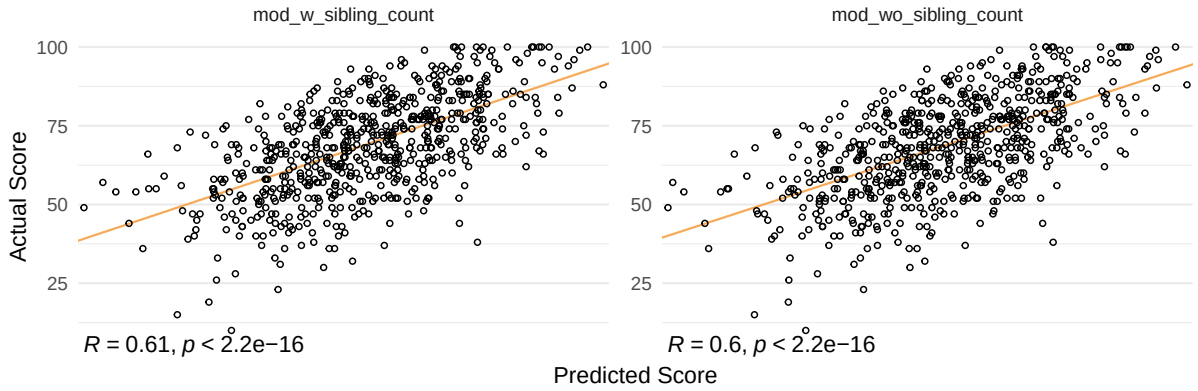

Table 20: Performance matrices of the 2 Models in Predicting (Writing Score)ˆ1.5

model_id	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
w/o no. of sibling	157.769	0.330	128.940	10.013	0.097	8.118
w/ no. of sibling	158.340	0.325	129.863	10.698	0.063	9.883

Figure 21: Correlation of Observed and Predicted Reading Score

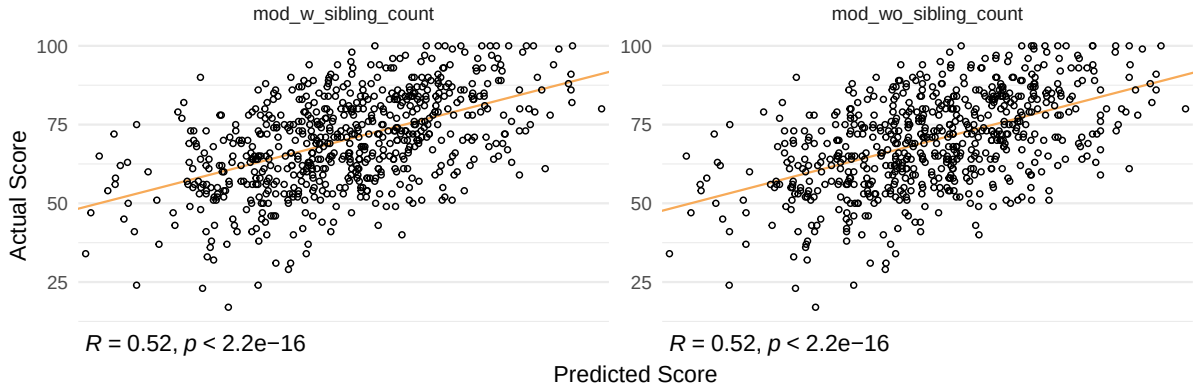

Table 21: Performance matrices of the 2 Models in Predicting (Reading Score) ˆ1.39

model_id	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
w/o no. of sibling	95.564	0.230	78.528	5.218	0.094	3.44
w no. of sibling	95.762	0.236	78.553	5.062	0.095	3.98