Final Project

Group members: Aiying Huang (ah4167), Mia Yu (my2838), Eunice Wang(cw3555)

Abstract

In this study, we analyzed test scores from a public school dataset, focusing on math, reading, and writing, alongside 11 socio-economic predictors. After reducing 948 observations to 846 due to missing values, we used AIC-guided backward elimination and LASSO penalization for model refinement. Diagnostic evaluations and 10-fold cross-validation were employed, suggesting a potential overfitting in the Math model. Key findings revealed significant gender impacts on scores and the influence of socio-economic factors such as test preparation, ethnicity, and parental status, with notable variations in math scores among students with different parental marital statuses.

Introduction

This project is based on a dataset which includes three test scores (math, reading and writing) of students at a public school and a variety of personal and socio-economic factors that may have interaction effects upon them. We want to use these factors as the covariates to predict Math, Reading and Writing scores.

Methods (data description and statistical methods)

Data Description and Visualization

There are a total of 14 variables(including 3 response variables[12-14] and 11 predictor variables[1-11]) with 948 observations in this dataset. The 11 predictor variables[1-11] are categorical variables and the 3 response variables[12-14] are continuous variables. The 3 response variables all have a relatively symmetric distributions with several outliers.

Pairwise Relationship and Correlation between varibales

Marginal distributions of each variable is displayed through pariplot with no obvious nonlinearities. We use the following methods to assess the strength of correlation: For categorical vs categorical variables, we use Cramer's V correlation. For continuous vs continuous variables, we use Pearson correlation. For categorical vs continuous variables, we use ANOVA(mcor). The correlation coefficients between variables all vary from 0 (corresponding to no association between the variables) to 1 (complete association) and can reach 1 only when each variable is completely determined by the other.

Missing Value Treatment:

Minimal missing values were observed primarily in qualitative variables such as EthnicGroup, ParentEduc, TestPrep, and others. Mode imputation was used for all except TransportMeans, with samples still showing missing values after imputation excluded.

Model Selection:

Interaction terms were included in the full models, despite limited variable correlations, for theoretical and practical reasons, covering all 11 predictors and their interactions to reduce overfitting. Model refinement began with AIC-guided backward elimination, but to address the retention of excessive variables, we applied LASSO penalization with cross-validation for optimal lambda selection. This process eliminated interaction terms with shrinkage coefficients below 0.5, resulting in three more streamlined, nested models (detailed in the appendix).

Model Assessment:

We utilized various diagnostic plots, including residual vs. fitted, Q-Q plots, scale-location, and residuals vs. leverage, to check whether our model meets the assumptions. We also computed Cook's Distance to identify any influential observations. Furthermore, to assess the presence of multicollinearity issues, we calculated the model's adjusted Generalized Variance Inflation Factor (GVIF).

Model Validation:

We employed a 10-fold cross-validation approach, which involved systematically partitioning the data into 10 subsets. In each cycle of this method, 9 subsets were used for training the model while the remaining one served as the validation set. This rotation continued until each subset had been utilized for validation, ensuring a comprehensive assessment of the model's performance. Additionally, to evaluate the predictive accuracy of our models, we calculated the Mean Squared Prediction Error (MSPE) using separate test data.

Results

There are a total of 14 variables(including 3 response, continuouse variables[12-14] and 11 predictor, categorical variables[1-11]) with 948 observations in this dataset. After missing value treatment(see figure 5), we left 846 obserbvations for further analysis. The 3 response variables all have a relatively symmetric distributions with several outliers.(see figure 1)

From the pairplot, we can see that there is no obvious nonlinearities regarding marginal distributions. (See figure 3) There is no colinearities exist between response variables and explanatory variables, and within 11 explanatory variables(See figure 4). However, there is a high correlation between the 3 response variables.Each explanatory are correlated with the other two with a correlation coefficient higher than 0.8.

Data was split into 80% training and 20% test sets for validity assessment. Diagnostic plots (figures 6-8) show no major issues. Residual vs. leverage plots (figure 9) identified outliers, especially in samples 181 and 268, but with leverage under 0.5 and Cook's distances below 0.1, leading to no data adjustments. Multicollinearity checks revealed no significant concerns(table 2-4).

Cross-validation results indicated RMSEs around 12 for the Math model and 12.5 for Reading and Writing (figure 10). MSPEs were 198.3466, 152.9267, and 142.8281(table 1), respectively, suggesting potential overfitting in the Math model due to its numerous predictors.

Finally, we dive deep into our model and related the statistical results to the initial problem in real world (table 5-7). Here are some important and interesting findings: In all three models, the impact of gender on scores is very significant, specifically reflected in the fact that males perform better in math compared to females, while their reading and writing scores are lower than females. Students

who did not participate in test preparation significantly scored lower in math, while students who enjoy standard lunch types and sometimes practice in sports significantly at the same time scored higher. Futhermore, ethnical Group D performing notably better. Whether taking the school bus and the duration of study also have certain impacts on some scores. Students who are firstborn significantly outperform those who are not the firstborn in reading and writing. The presence and number of siblings also have some impact on the scores, but the changes are not substantial. The lower the parents' education level, the significantly lower the children's reading and writing scores are. Surprisingly, parents' marital status has a significant change in scores. In the math score model, students with widowed parents can actually increase their scores by 32 points! (p-value = 0.019).

Conclusions

Our study on student test scores highlights key influences from socio-economic and personal factors, particularly gender, with robust modeling revealing significant impacts on academic outcomes. These insights are crucial for understanding and enhancing educational performance.

A brief summary on each group member's contribution

Eunice Wang (cw3555) focused on initial data analysis and exploring relationships. Aiying Huang (ah4167) contributed to model construction, validation, and final report structuring. Mia Yu (my2838) was key in model selection and diagnostics, while all team members jointly wrote the report's introduction and conclusion.

Figures and Tables

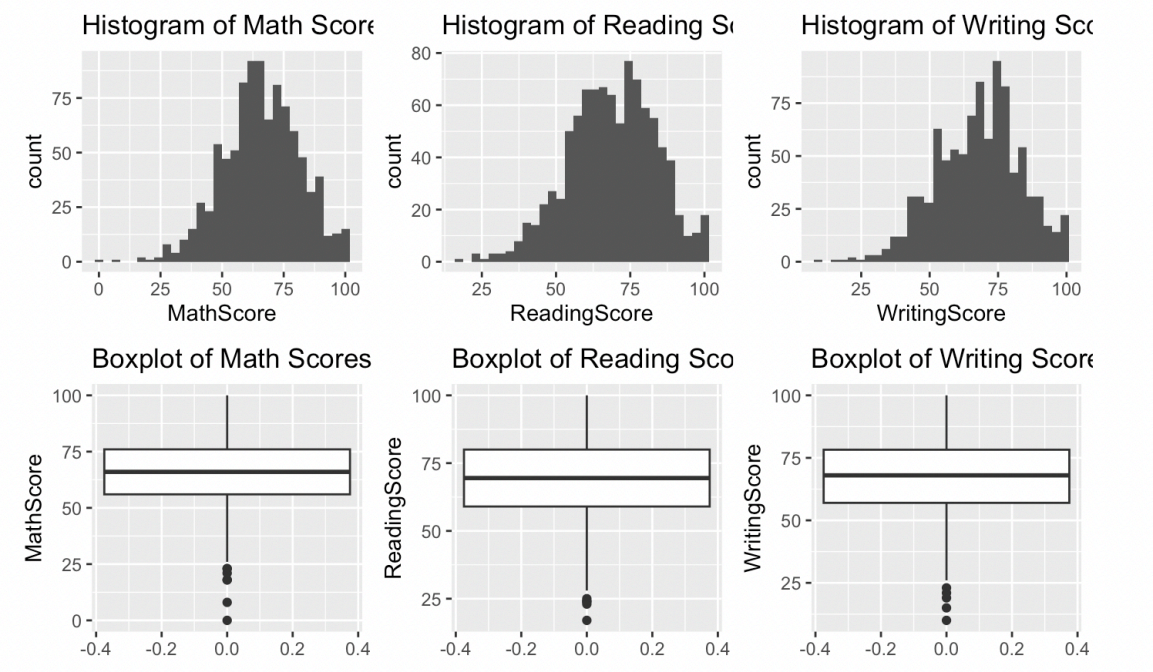

Figure 1: Ys' distribution

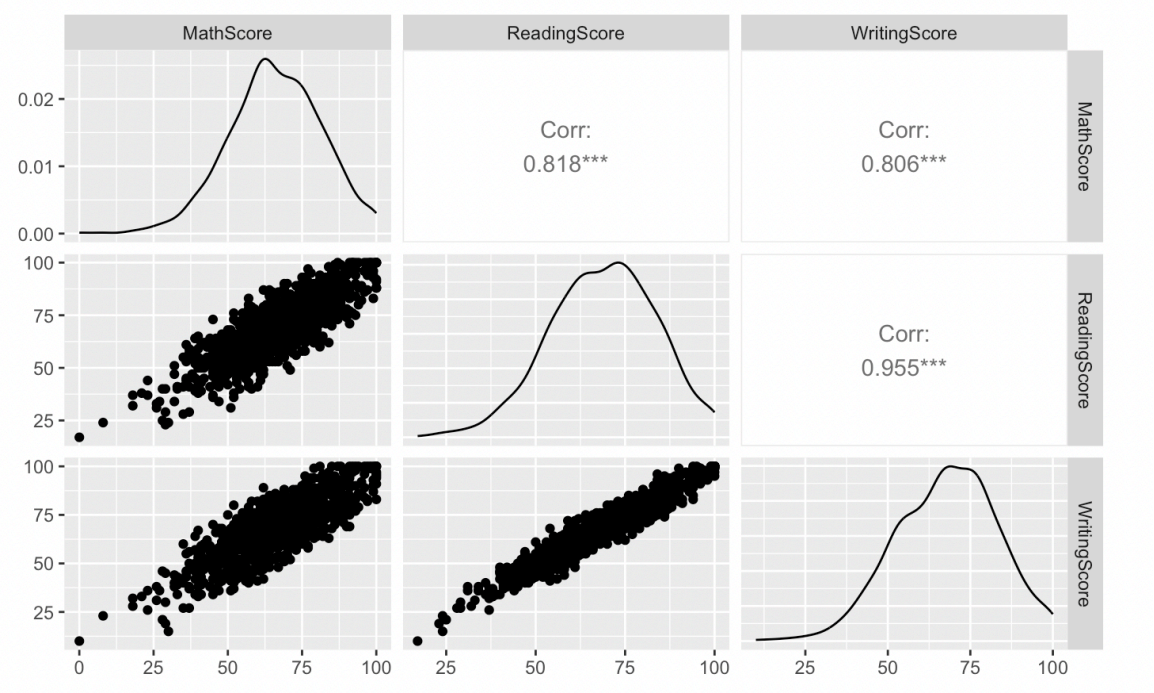

Figure 2: Ys' correlation

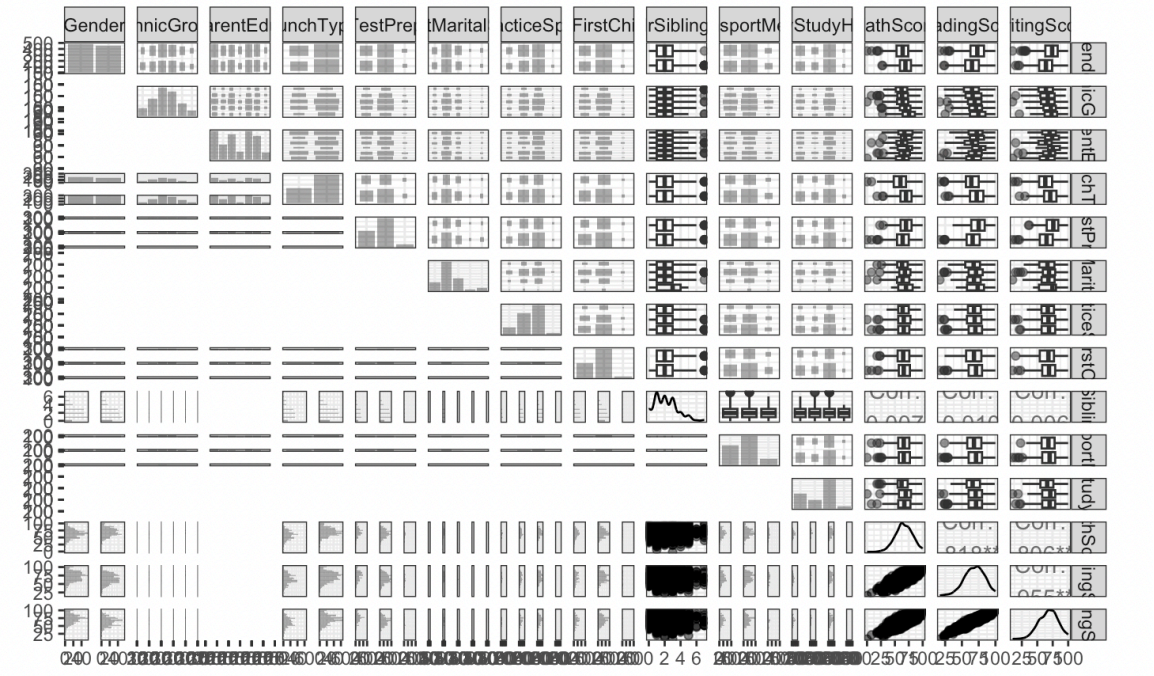

Figure 3: pairwise

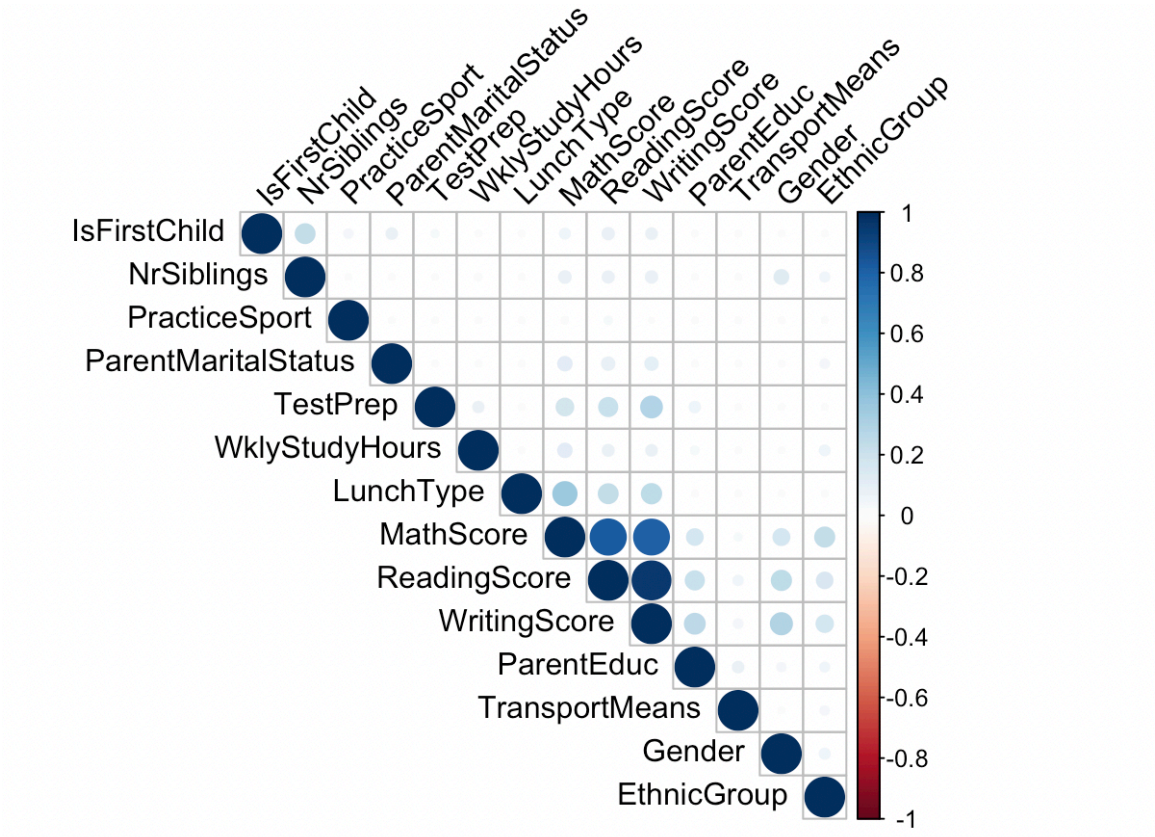

Figure 4: heatmap

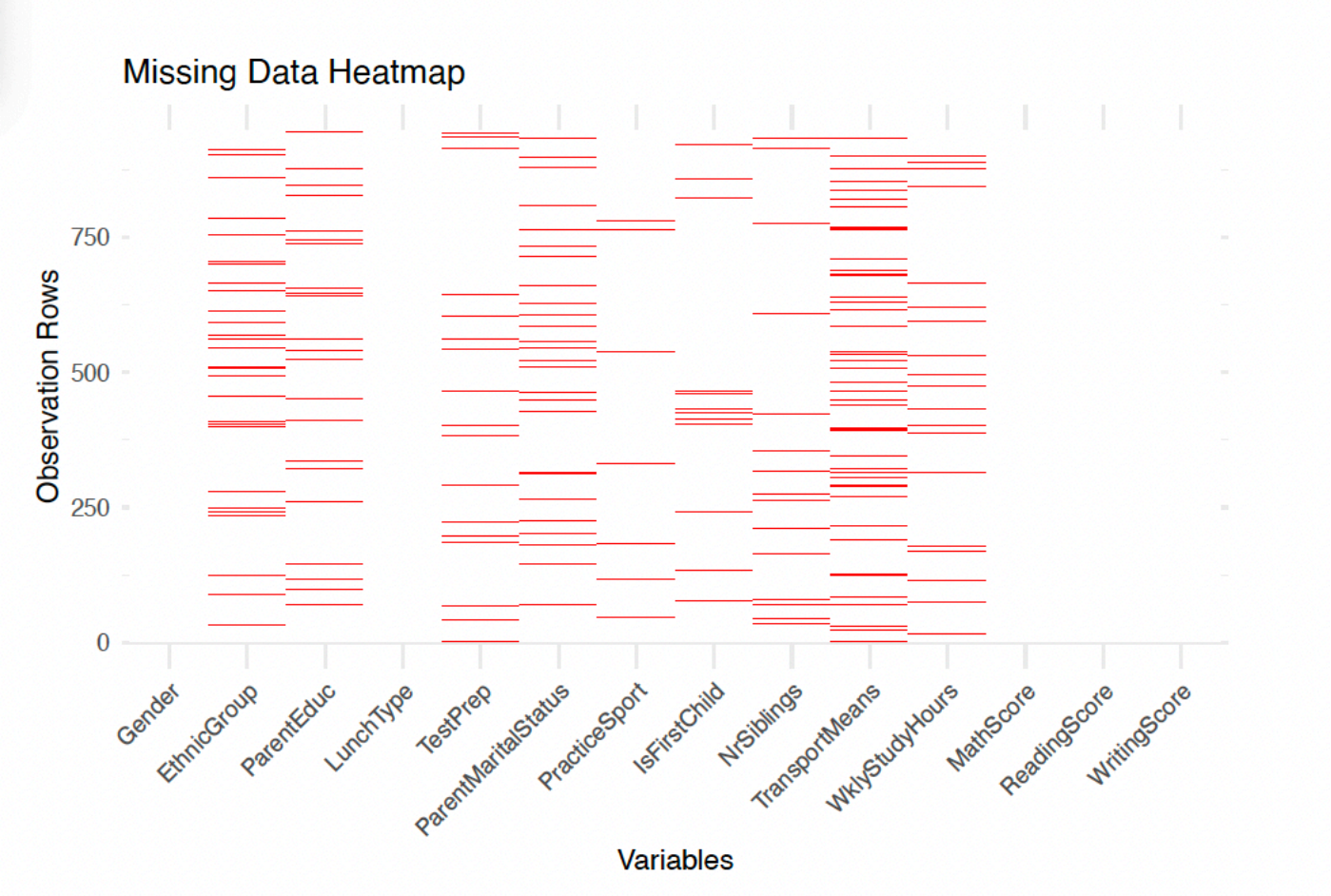

Figure 5: missingdata

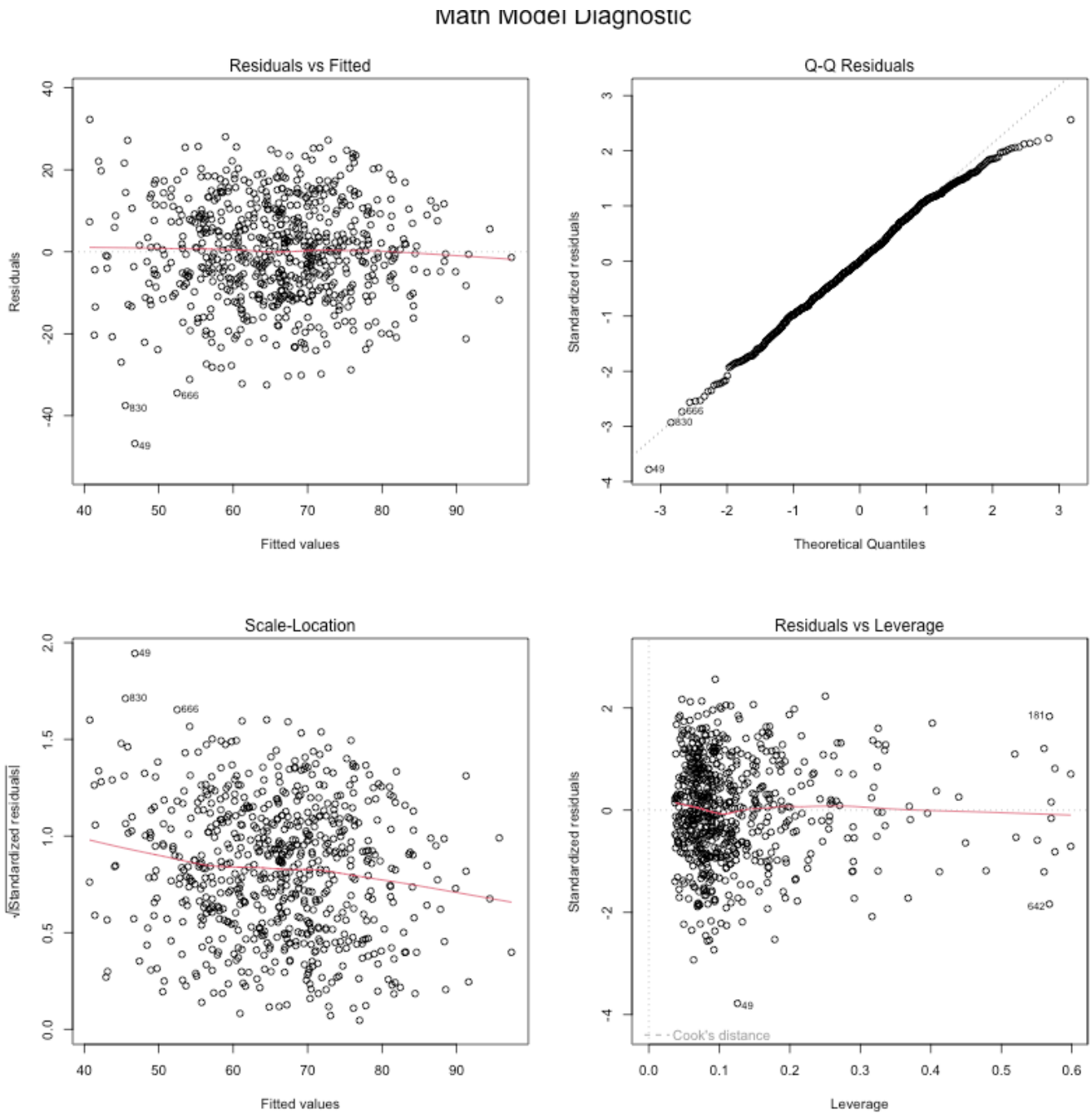

Figure 6: math

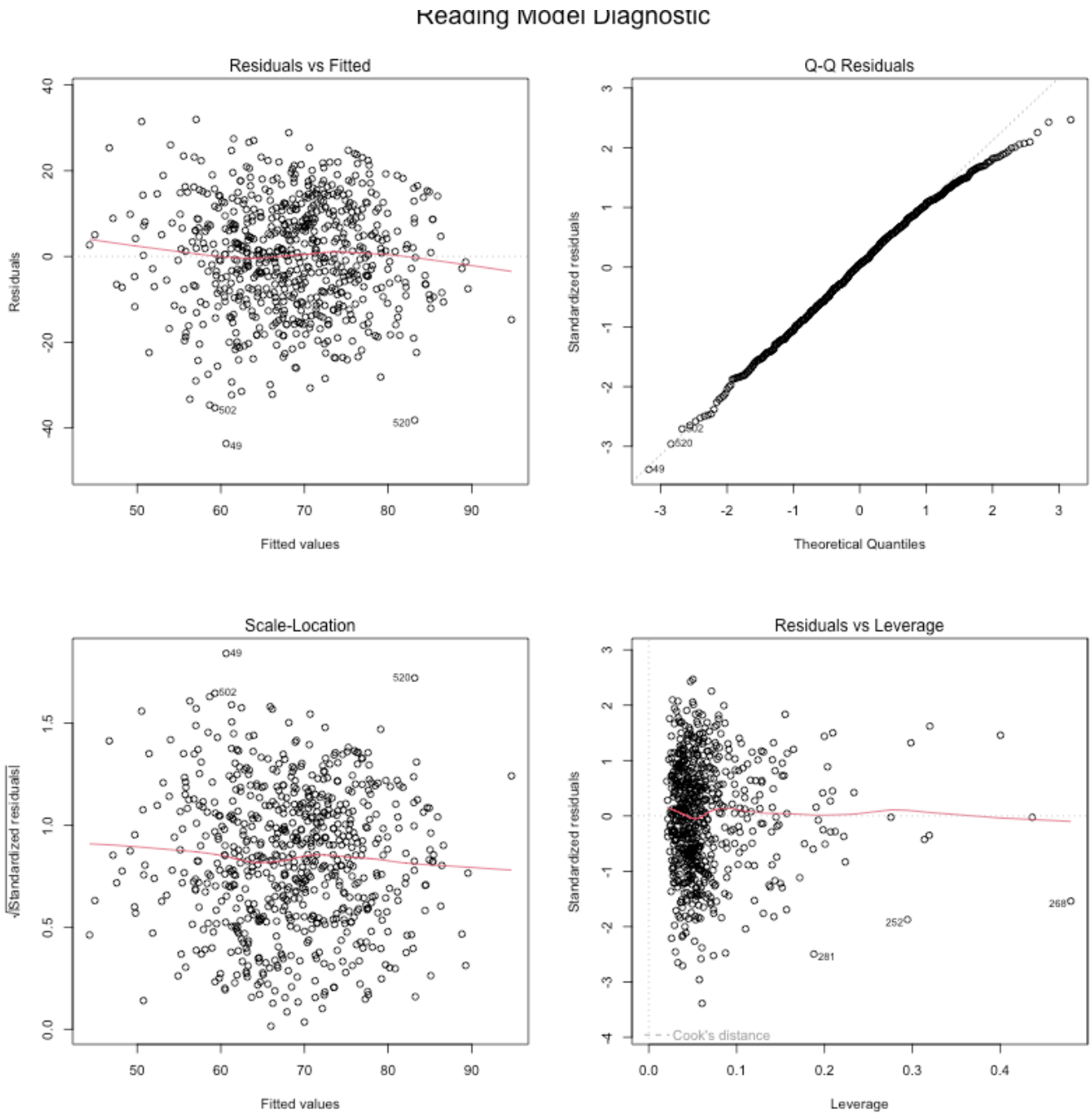

Figure 7: reading

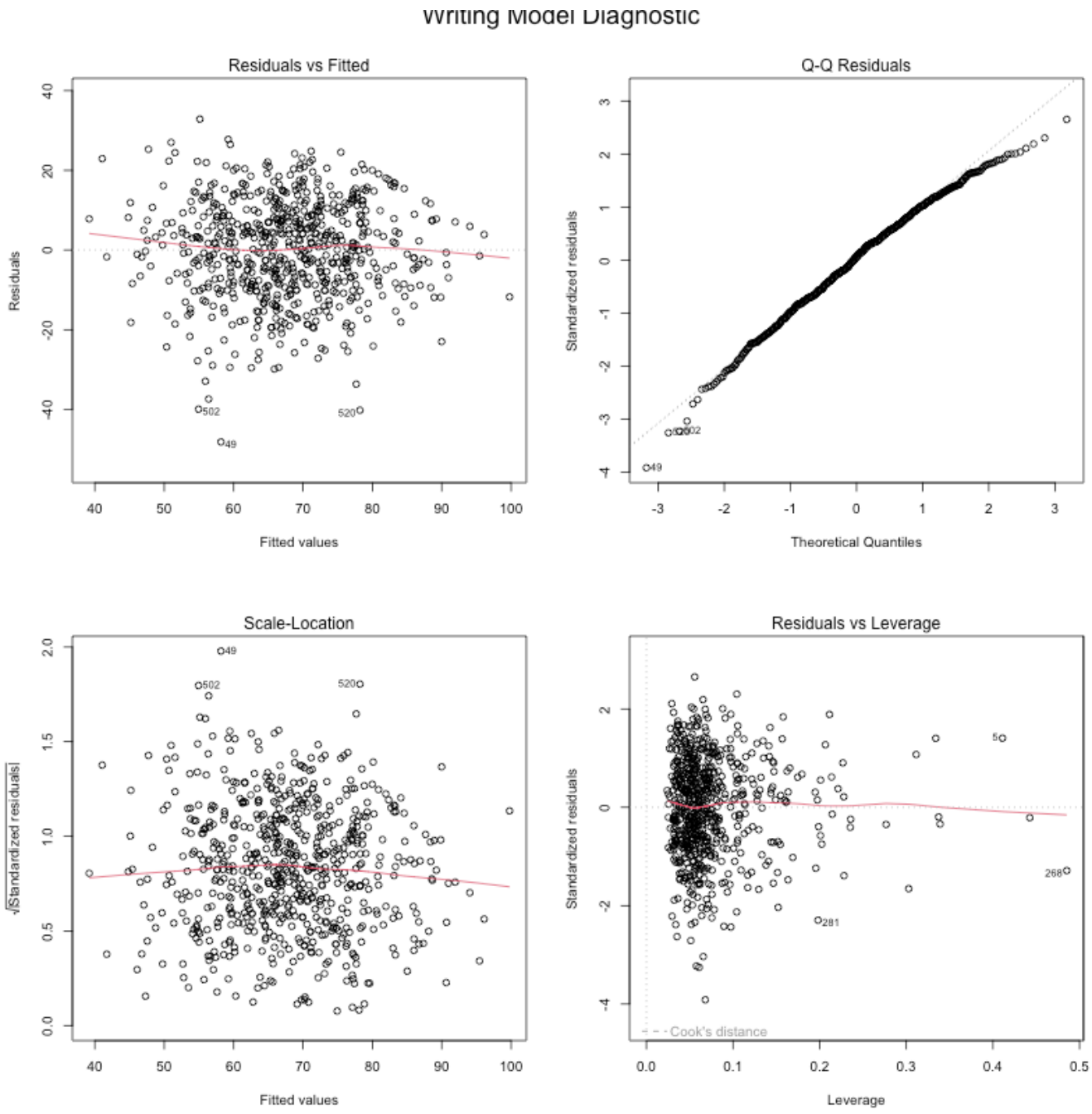

Figure 8: writing

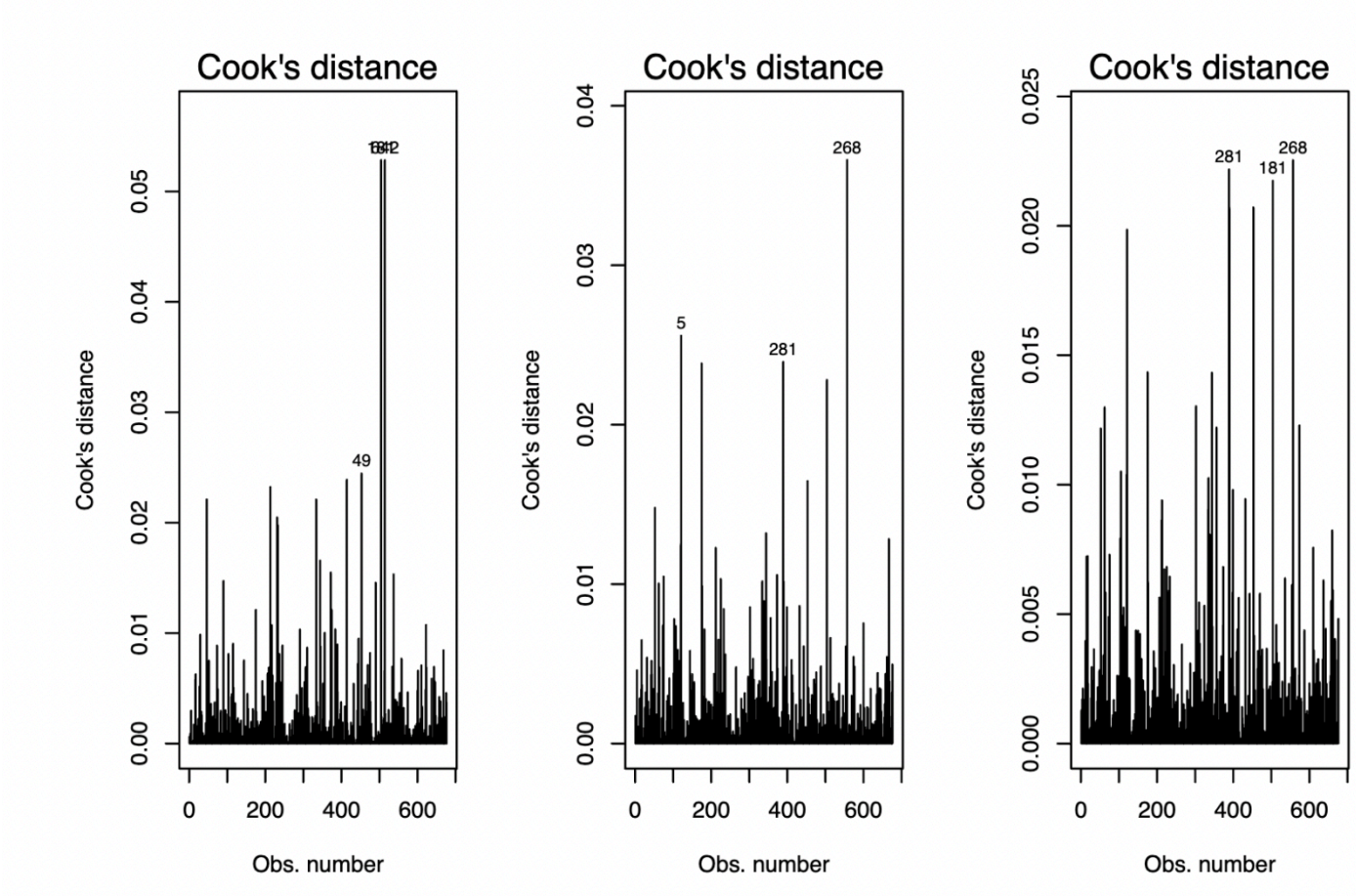

Figure 9: Cook's Distance

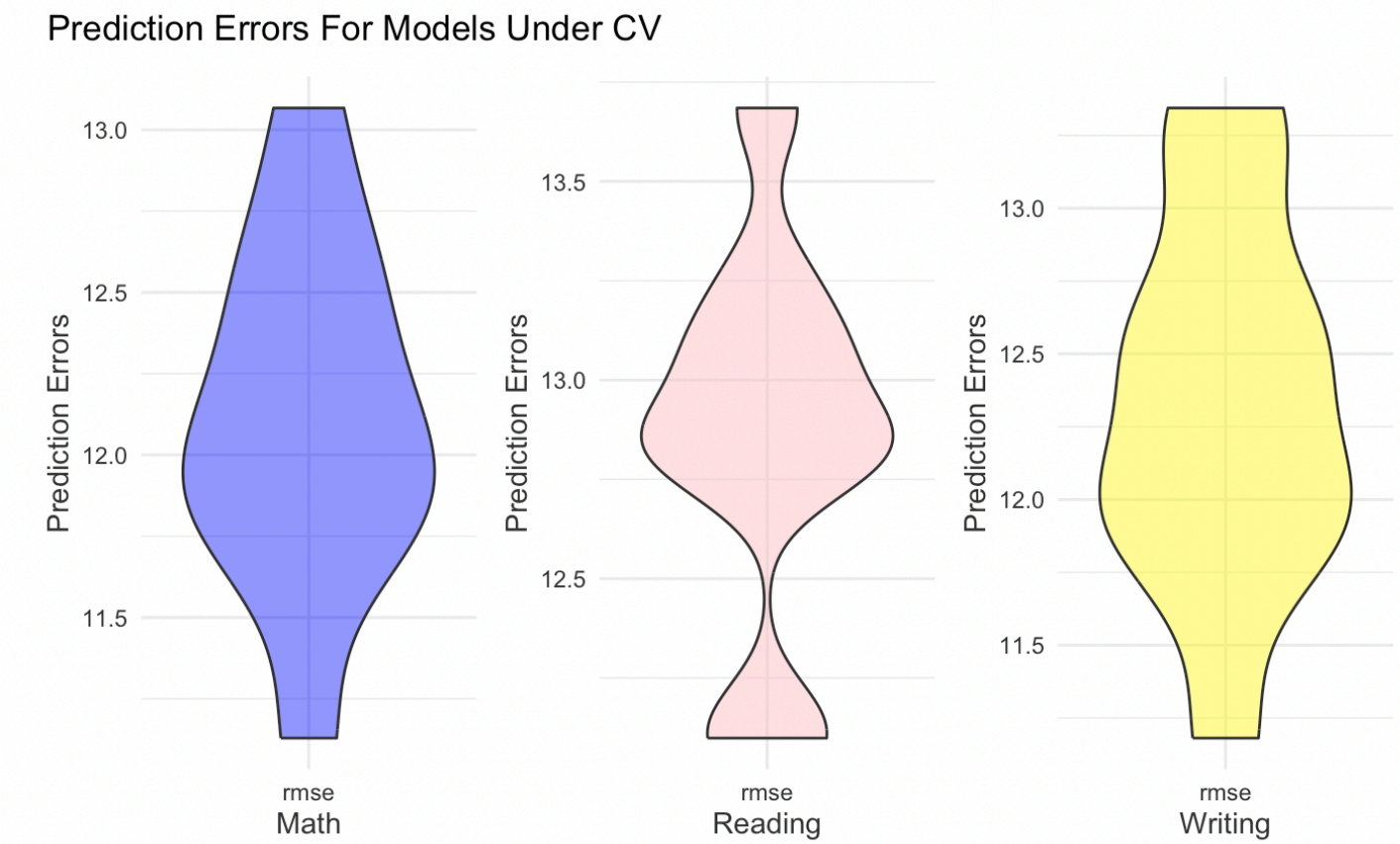

Figure 10: CV outcome

Table 1: MSPE Values for Different Subjects

Subject	MSPE
Math	198.3466
Reading	152.9267
Writing	142.8280

Table 2: Variance Inflation Factors for Math Model

Predictor	GVIF
Gender	1.655040
EthnicGroup	1.353349
ParentEduc	1.081339
LunchType	1.643025
TestPrep	1.074470
ParentMaritalStatus	1.331176
PracticeSport	1.357566
TransportMeans	1.516250
WklyStudyHours	1.366449

Table 3: Variance Inflation Factors for Writing Model

Predictor	GVIF
Gender	1.042331
EthnicGroup	1.041528
ParentEduc	1.157013
LunchType	1.659319
TestPrep	1.040662
ParentMaritalStatus	1.183376
PracticeSport	1.200553
IsFirstChild	1.339793
NrSiblings	1.040662
TransportMeans	1.034014
WklyStudyHours	1.413038

Table 4: Variance Inflation Factors for Reading Model

Predictor	GVIF
Gender	1.036102
EthnicGroup	1.039638
ParentEduc	1.032325
LunchType	1.648075
TestPrep	1.044683
ParentMaritalStatus	1.117825
PracticeSport	1.199588
IsFirstChild	1.619041
NrSiblings	1.608364
TransportMeans	1.034064
WklyStudyHours	1.255155

Table 5: Math Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	59.824	7.825	7.646	0.000
Gendermale	4.999	3.390	1.475	0.141
EthnicGroupgroup B	2.116	5.391	0.393	0.695
EthnicGroupgroup C	-0.392	5.104	-0.077	0.939
EthnicGroupgroup D	-0.270	5.456	-0.049	0.961
EthnicGroupgroup E	4.856	5.556	0.874	0.382
ParentEducbachelor’s degree	13.216	10.457	1.264	0.207
ParentEduchigh school	-1.379	8.703	-0.158	0.874
ParentEducmaster’s degree	-11.182	13.313	-0.840	0.401
ParentEducsome college	1.269	7.904	0.161	0.872

term	estimate	std.error	statistic	p.value
ParentEducsome high school	2.884	8.905	0.324	0.746
LunchTypestandard	2.582	3.430	0.753	0.452
TestPrepnone	-5.467	1.165	-4.693	0.000
ParentMaritalStatusmarried	4.477	3.718	1.204	0.229
ParentMaritalStatusnone	5.149	6.330	0.813	0.416
ParentMaritalStatussingle	7.105	4.344	1.636	0.102
ParentMaritalStatuswidowed	32.183	13.768	2.338	0.020
PracticeSportregularly	-7.032	6.093	-1.154	0.249
PracticeSportsometimes	-6.203	5.900	-1.051	0.294
TransportMeansschool_bus	-2.996	2.957	-1.013	0.312
WklyStudyHours> 10	2.174	5.306	0.410	0.682
WklyStudyHours10-May	-3.016	3.856	-0.782	0.435
Gendermale:PracticeSportregularly	2.409	3.826	0.630	0.529
Gendermale:PracticeSportsometimes	-3.109	3.682	-0.844	0.399
EthnicGroupgroup B:ParentEducbachelor’s degree	12.330	10.104	1.220	0.223
EthnicGroupgroup C:ParentEducbachelor’s degree	15.648	9.442	1.657	0.098
EthnicGroupgroup D:ParentEducbachelor’s degree	11.103	9.858	1.126	0.260
EthnicGroupgroup E:ParentEducbachelor’s degree	17.593	10.628	1.655	0.098
EthnicGroupgroup B:ParentEduchigh school	-6.634	7.086	-0.936	0.350
EthnicGroupgroup C:ParentEduchigh school	1.002	6.695	0.150	0.881
EthnicGroupgroup D:ParentEduchigh school	1.024	7.063	0.145	0.885

term	estimate	std.error	statistic	p.value
EthnicGroupgroup E:ParentEduchigh school	4.168	7.576	0.550	0.582
EthnicGroupgroup B:ParentEducmaster’s degree	23.280	13.356	1.743	0.082
EthnicGroupgroup C:ParentEducmaster’s degree	15.317	12.186	1.257	0.209
EthnicGroupgroup D:ParentEducmaster’s degree	26.857	11.969	2.244	0.025
EthnicGroupgroup E:ParentEducmaster’s degree	21.688	13.395	1.619	0.106
EthnicGroupgroup B:ParentEducsome college	0.901	6.924	0.130	0.896
EthnicGroupgroup C:ParentEducsome college	3.771	6.460	0.584	0.560
EthnicGroupgroup D:ParentEducsome college	11.184	6.774	1.651	0.099
EthnicGroupgroup E:ParentEducsome college	2.569	7.050	0.364	0.716
EthnicGroupgroup B:ParentEducsome high school	-5.232	7.387	-0.708	0.479
EthnicGroupgroup C:ParentEducsome high school	-0.626	7.268	-0.086	0.931
EthnicGroupgroup D:ParentEducsome high school	1.506	7.256	0.208	0.836
EthnicGroupgroup E:ParentEducsome high school	9.711	8.279	1.173	0.241
ParentEducbachelor’s degree:ParentMaritalStatusmarried	-21.643	7.246	-2.987	0.003
ParentEduchigh school:ParentMaritalStatusmarried	0.258	4.638	0.056	0.956
ParentEducmaster’s degree:ParentMaritalStatusmarried	-5.689	8.430	-0.675	0.500
ParentEducsome college:ParentMaritalStatusmarried	-7.459	4.428	-1.685	0.093

term	estimate	std.error	statistic	p.value
ParentEducsome high school:ParentMaritalStatusmarried	-6.891	5.205	-1.324	0.186
ParentEducbachelor’s degree:ParentMaritalStatusnone	-23.763	13.279	-1.790	0.074
ParentEduchigh school:ParentMaritalStatusnone	-8.760	8.413	-1.041	0.298
ParentEducmaster’s degree:ParentMaritalStatusnone	0.260	14.410	0.018	0.986
ParentEducsome college:ParentMaritalStatusnone	-5.382	7.549	-0.713	0.476
ParentEducsome high school:ParentMaritalStatusnone	-8.130	11.424	-0.712	0.477
ParentEducbachelor’s degree:ParentMaritalStatussingle	-27.998	7.996	-3.502	0.000
ParentEduchigh school:ParentMaritalStatussingle	1.151	5.420	0.212	0.832
ParentEducmaster’s degree:ParentMaritalStatussingle	-10.498	9.553	-1.099	0.272
ParentEducsome college:ParentMaritalStatussingle	-13.882	5.171	-2.685	0.007
ParentEducsome high school:ParentMaritalStatussingle	-8.442	5.906	-1.429	0.153
ParentEducbachelor’s degree:ParentMaritalStatuswidowed	-14.566	14.065	-1.036	0.301

term	estimate	std.error	statistic	p.value
ParentEduchigh school:ParentMaritalStatuswidowed	-22.354	12.997	-1.720	0.086
ParentEducmaster’s degree:ParentMaritalStatuswidowed	-32.978	21.234	-1.553	0.121
ParentEducsome college:ParentMaritalStatuswidowed	-5.989	12.354	-0.485	0.628
ParentEducsome high school:ParentMaritalStatuswidowed	-31.184	13.997	-2.228	0.026
ParentEducbachelor’s degree:PracticeSportregularly	2.492	7.601	0.328	0.743
ParentEduchigh school:PracticeSportregularly	-4.748	5.953	-0.798	0.425
ParentEducmaster’s degree:PracticeSportregularly	-11.583	9.179	-1.262	0.207
ParentEducsome college:PracticeSportregularly	-2.810	5.126	-0.548	0.584
ParentEducsome high school:PracticeSportregularly	-3.974	5.912	-0.672	0.502
ParentEducbachelor’s degree:PracticeSportsometimes	-9.838	7.379	-1.333	0.183
ParentEduchigh school:PracticeSportsometimes	-3.297	5.803	-0.568	0.570
ParentEducmaster’s degree:PracticeSportsometimes	3.476	7.539	0.461	0.645
ParentEducsome college:PracticeSportsometimes	0.573	4.983	0.115	0.909
ParentEducsome high school:PracticeSportsometimes	-1.793	5.768	-0.311	0.756
LunchTypestandard:PracticeSportregularly	7.862	3.902	2.015	0.044

term	estimate	std.error	statistic	p.value
LunchTypestandard:PracticeSportsometimes	10.908	3.775	2.890	0.004
ParentMaritalStatusmarried:TransportMeansschool_bus	5.998	3.308	1.813	0.070
ParentMaritalStatusnone:TransportMeansschool_bus	4.167	6.035	0.691	0.490
ParentMaritalStatussingle:TransportMeansschool_bus	1.949	3.747	0.520	0.603
ParentMaritalStatuswidowed:TransportMeansschool_bus	-13.871	9.598	-1.445	0.149
PracticeSportregularly:WklyStudyHours> 10	1.787	6.017	0.297	0.767
PracticeSportsometimes:WklyStudyHours> 10	2.813	5.808	0.484	0.628
PracticeSportregularly:WklyStudyHours10-May	9.574	4.392	2.180	0.030
PracticeSportsometimes:WklyStudyHours10-May	4.841	4.247	1.140	0.255

Table 6: Reading Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	66.729	6.513	10.246	0.000
Gendermale	-8.345	1.059	-7.878	0.000
EthnicGroupgroup B	0.926	2.207	0.419	0.675
EthnicGroupgroup C	1.768	2.067	0.855	0.393
EthnicGroupgroup D	4.369	2.123	2.058	0.040
EthnicGroupgroup E	5.569	2.345	2.375	0.018
ParentEducbachelor’s degree	0.959	1.958	0.490	0.624
ParentEduchigh school	-5.970	1.644	-3.630	0.000
ParentEducmaster’s degree	3.074	2.423	1.268	0.205
ParentEducsome college	-3.423	1.529	-2.238	0.026
ParentEducsome high school	-6.227	1.753	-3.551	0.000
LunchTypestandard	1.547	3.254	0.475	0.635
TestPrepnone	-6.711	1.137	-5.903	0.000
ParentMaritalStatusmarried	12.998	5.371	2.420	0.016
ParentMaritalStatusnone	25.053	9.693	2.585	0.010
ParentMaritalStatussingle	17.190	6.298	2.729	0.007
ParentMaritalStatuswidowed	20.365	9.681	2.104	0.036
PracticeSportregularly	-8.473	6.238	-1.358	0.175
PracticeSportsometimes	-2.865	5.911	-0.485	0.628

term	estimate	std.error	statistic	p.value
IsFirstChildyes	10.083	3.409	2.958	0.003
NrSiblings	-1.433	0.727	-1.971	0.049
TransportMeansschool_bus	2.028	1.078	1.881	0.060
WklyStudyHours> 10	-4.077	5.529	-0.737	0.461
WklyStudyHours10-May	-8.780	4.118	-2.132	0.033
LunchTypestandard:PracticeSportregularly	4.275	3.713	1.151	0.250
LunchTypestandard:PracticeSportsometimes	7.422	3.592	2.067	0.039
ParentMaritalStatusmarried:PracticeSportregularly	2.117	5.152	0.411	0.681
ParentMaritalStatusnone:PracticeSportregularly	-13.255	8.424	-1.573	0.116
ParentMaritalStatussingle:PracticeSportregularly	-10.545	6.229	-1.693	0.091
ParentMaritalStatuswidowed:PracticeSportregularly	-15.747	10.733	-1.467	0.143
ParentMaritalStatusmarried:PracticeSportsometimes	-3.407	4.862	-0.701	0.484
ParentMaritalStatusnone:PracticeSportsometimes	-14.406	7.960	-1.810	0.071
ParentMaritalStatussingle:PracticeSportsometimes	-12.944	5.980	-2.165	0.031
ParentMaritalStatuswidowed:PracticeSportsometimes	-10.387	10.919	-0.951	0.342
ParentMaritalStatusmarried:IsFirstChildyes	-10.457	3.716	-2.814	0.005
ParentMaritalStatusnone:IsFirstChildyes	-15.278	7.288	-2.096	0.036
ParentMaritalStatussingle:IsFirstChildyes	-5.087	4.135	-1.230	0.219
ParentMaritalStatuswidowed:IsFirstChildyes	-4.553	7.980	-0.571	0.568
PracticeSportregularly:WklyStudyHours> 10	3.568	5.748	0.621	0.535
PracticeSportsometimes:WklyStudyHours> 10	2.870	5.566	0.516	0.606

term	estimate	std.error	statistic	p.value
PracticeSportregularly:WklyStudyHours10-May	11.501	4.309	2.669	0.008
PracticeSportsometimes:WklyStudyHours10-May	5.483	4.140	1.324	0.186
NrSiblings:WklyStudyHours> 10	1.560	1.118	1.396	0.163
NrSiblings:WklyStudyHours10-May	2.070	0.858	2.411	0.016

Table 7: Writing Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	63.371	6.641	9.543	0.000
Gendermale	-9.867	1.022	-9.651	0.000
EthnicGroupgroup B	-0.207	2.115	-0.098	0.922
EthnicGroupgroup C	1.344	1.987	0.676	0.499
EthnicGroupgroup D	5.798	2.037	2.847	0.005
EthnicGroupgroup E	4.438	2.261	1.962	0.050
ParentEducbachelor’s degree	3.801	3.386	1.123	0.262
ParentEduchigh school	-11.410	2.774	-4.113	0.000
ParentEducmaster’s degree	4.501	4.052	1.111	0.267
ParentEducsome college	-8.814	2.458	-3.586	0.000
ParentEducsome high school	-8.939	2.801	-3.191	0.001
LunchTypestandard	1.852	3.148	0.588	0.557
TestPrepnone	-7.346	1.887	-3.893	0.000
ParentMaritalStatusmarried	11.693	5.174	2.260	0.024
ParentMaritalStatusnone	17.080	9.344	1.828	0.068
ParentMaritalStatussingle	13.982	6.062	2.306	0.021
ParentMaritalStatuswidowed	18.427	9.274	1.987	0.047
PracticeSportregularly	-8.365	6.050	-1.383	0.167
PracticeSportsometimes	-3.454	5.724	-0.603	0.546

term	estimate	std.error	statistic	p.value
IsFirstChildyes	8.790	4.242	2.072	0.039
NrSiblings	1.007	0.573	1.758	0.079
TransportMeansschool_bus	2.180	1.034	2.108	0.035
WklyStudyHours> 10	-0.797	5.586	-0.143	0.887
WklyStudyHours10-May	-0.400	4.175	-0.096	0.924
ParentEducbachelor’s degree:IsFirstChildyes	-2.490	4.077	-0.611	0.542
ParentEduchigh school:IsFirstChildyes	6.186	3.368	1.837	0.067
ParentEducmaster’s degree:IsFirstChildyes	1.528	4.887	0.313	0.755
ParentEducsome college:IsFirstChildyes	8.674	3.054	2.840	0.005
ParentEducsome high school:IsFirstChildyes	2.957	3.468	0.853	0.394
LunchTypestandard:PracticeSportregularly	5.442	3.601	1.511	0.131
LunchTypestandard:PracticeSportsometimes	7.838	3.475	2.256	0.024
TestPrepnone:NrSiblings	-1.036	0.710	-1.460	0.145
ParentMaritalStatusmarried:PracticeSportregularly	4.327	4.958	0.873	0.383
ParentMaritalStatusnone:PracticeSportregularly	-11.610	8.131	-1.428	0.154
ParentMaritalStatussingle:PracticeSportregularly	-6.106	6.032	-1.012	0.312
ParentMaritalStatuswidowed:PracticeSportregularly	-16.487	10.332	-1.596	0.111
ParentMaritalStatusmarried:PracticeSportsometimes	-1.608	4.676	-0.344	0.731
ParentMaritalStatusnone:PracticeSportsometimes	-10.538	7.662	-1.375	0.170
ParentMaritalStatussingle:PracticeSportsometimes	-9.158	5.769	-1.587	0.113
ParentMaritalStatuswidowed:PracticeSportsometimes	-11.440	10.545	-1.085	0.278

term	estimate	std.error	statistic	p.value
ParentMaritalStatusmarried:IsFirstChildyes	-9.697	3.596	-2.697	0.007
ParentMaritalStatusnone:IsFirstChildyes	-8.688	7.015	-1.239	0.216
ParentMaritalStatussingle:IsFirstChildyes	-5.121	4.008	-1.278	0.202
ParentMaritalStatuswidowed:IsFirstChildyes	-1.651	7.678	-0.215	0.830
PracticeSportregularly:WklyStudyHours> 10	3.192	5.548	0.575	0.565
PracticeSportsometimes:WklyStudyHours> 10	3.941	5.420	0.727	0.467
PracticeSportregularly:WklyStudyHours10-May	11.030	4.169	2.645	0.008
PracticeSportsometimes:WklyStudyHours10-May	5.510	3.994	1.379	0.168
IsFirstChildyes:WklyStudyHours> 10	-0.880	3.363	-0.262	0.794
IsFirstChildyes:WklyStudyHours10-May	-5.469	2.580	-2.119	0.034

References

Bollinger, G. (1981). Book Review: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Journal of Marketing Research, 18(3), 392-393. https://doi.org/10.1177/002224378101800318

Fox, J., & Monette, G. (1992). Generalized Collinearity Diagnostics. Journal of the American Statistical Association, 87(417), 178-183. https://doi.org/10.2307/2290467

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.).OTexts.

Appendix

Three final regression models

```
MathScore ~  Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
ParentMaritalStatus + PracticeSport + TransportMeans + WklyStudyHours +
Gender:PracticeSport + EthnicGroup:ParentEduc + ParentEduc:ParentMaritalStatus
+ + ParentEduc:PracticeSport + LunchType:PracticeSport + ParentMaritalStatus:TransportMe
+ PracticeSport:WklyStudyHours
```

```
ReadingScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings + TransportMeans
+ WklyStudyHours + LunchType:PracticeSport + ParentMaritalStatus:PracticeSport
+ ParentMaritalStatus:IsFirstChild + PracticeSport:WklyStudyHours + NrSiblings:WklyStudy
```

```
WritingScore ~ Gender + EthnicGroup + ParentEduc + LunchType + TestPrep +
ParentMaritalStatus + PracticeSport + IsFirstChild + NrSiblings + TransportMeans
+ WklyStudyHours + ParentEduc:IsFirstChild + LunchType:PracticeSport +
TestPrep:NrSiblings + ParentMaritalStatus:PracticeSport + ParentMaritalStatus:IsFirstChi
+ PracticeSport:WklyStudyHours + IsFirstChild:WklyStudyHours
```

1. Gender: Gender of the student (male/female)
2. EthnicGroup: Ethnic group of the student (group A to E)
3. ParentEduc: Parent(s) education background (from some_highschool to master's degree)
4. LunchType: School lunch type (standard or free/reduced)

5. TestPrep: Test preparation course followed (completed or none)
6. ParentMaritalStatus: Parent(s) marital status (married/single/widowed/divorced)
7. PracticeSport: How often the student parctice sport (never/sometimes/regularly))
8. IsFirstChild: If the child is first child in the family or not (yes/no)
9. NrSiblings: Number of siblings the student has (0 to 7)
10. TransportMeans: Means of transport to school (schoolbus/private)
11. WklyStudyHours: Weekly self-study hours(less that 5hrs; between 5 and 10hrs; more than 10hrs)
12. MathScore: math test score(0-100)
13. ReadingScore: reading test score(0-100)
14. WritingScore: writing test score(0-100)