Final Project

Abstract

a brief introduction, brief description of methods, and main results into a one-paragraph summary

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 variables and description is listed below:

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)

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.(See figure 1)

Marginal Distributions and Pairwise Relationship

From the pariplot of marginal distributions, we can see that there is no obvious nonlinearities.(See figure 3)

Correlation between varibales

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).

This correlation coefficient all varies 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.

From the table and the heatmap we generated, we can see that there is no colinearities exist between response variables and explanatory variables. There are also colinearities exist within 11 explanatory variables. 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.

Missing Value Treatment:

Minimal missing values were observed as figure 1, 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:

LASSO(Least Absolute Shrinkage and Selection Operator)

LASSO ($\min_\beta \left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right)$), introduced by Robert Tibshirani in 1996, is a regression method that enhances model accuracy and interpretability by performing variable selection and regularization. Ideal for models with numerous predictors or high collinearity, LASSO shrinks some coefficients to zero, thus simplifying the model by penalizing the sum of absolute parameter values.

Model Assessment:

Cook's Distance (Residual vs Leverage Plot)

Cook's Distance($D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE}$.), combined with the Residual vs. Leverage Plot, identifies influential observations in regression models. It measures each observation's impact on the fitted values to highlight potential outliers.

VIF (Variance Inflation Factor)

Adjusted Generalized Variance Inflation Factor ($GVIF^{(1/(2\times Df))} = \left(\frac{1}{1-R_j^2}\right)^{(1/(2\times Df_j))}$.) was used instead of traditional VIF to assess multicollinearity, especially suitable for categorical variables and models with interaction terms.

An adjusted GVIF value ranging from 1 to 3 indicates no significant issues, while values exceeding 5 or 10 suggest notable multicollinearity.

Model Validation:

10-fold Cross Validation

A 10-fold cross-validation approach was employed. This method involves partitioning the data into 10 subsets, using 9 for training and 1 for validation in a rotating fashion. This process is repeated until each subset has been used for validation.

The Mean Squared Prediction Errors (MSPE)

The Mean Squared Prediction Error ($MSPE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$) is a statistical measure used to evaluate the accuracy of a predictive model. It calculates the average of the squares of the prediction errors, where a prediction error is the difference between the observed values and the predicted values made by the model.

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, we left 846 obserbvations for further analysis. The 3 response variables all have a relatively symmetric distributions with several outliers.

There is no colinearities exist between response variables and explanatory variables, and within 11 explanatory variables. 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.

To facilitate internal validity assessment in subsequent modeling, the data was split into a training set (80%) and a test set (20%).

Despite the lack of evident correlations among variables, interaction terms were included in the full models based on theoretical and practical considerations. These models encompassed all 11 predictors with their pairwise interactions, aiming to further simplify the variable subset to mitigate overfitting risks.

Utilizing a combination of automated methods and criteria-based selection, the model refinement process began with backward elimination guided by the AIC criterion, which favored fewer variables. However, this approach initially retained too many variables. Consequently, we employed LASSO penalization, complemented by cross-validation to determine the optimal lambda. This led to the removal of interaction terms with shrinkage coefficients under 0.5, yielding three more efficient, nested models. (see in appendix)

Diagnostic plots for these models indicate no issues.

Examination of the residual vs leverage plots for the three models revealed a few outlier observations, notably in samples 181 and 268 according to figure 2. However, closer inspection showed that their leverage did not exceed 0.5, and Cook's distances were below 0.1. Additionally, no data entry anomalies were identified in these samples. Removing these samples and reconstructing the models showed negligible differences from the original models. Therefore, no adjustments were made, and the original data was used for modeling.

As for the multicollinearity check, we found no substantial multicollinearity concerns in any of the model variables.

The cross-validation results indicated that the RMSE for the Math model was concentrated around

12, while for the Reading and Writing models, RMSE values were centered around 12.5, according to figure 3. To test for potential overfitting, the initially separated test data was employed to evaluate the predictive performance of the models. The Mean Squared Prediction Errors (MSPE) for the Math, Reading, and Writing models were 198.3466, 152.9267, and 142.8281, respectively. Contrary to the RMSE results, the Math model exhibited a lower performance compared to the Reading and Writing models. Given that the Math model incorporated the most predictors, it suggests a potential issue with overfitting.

Conclusions/Discussion

Your content here.

A brief summary on each group member's contribution

Aiying Huang (ah4167) notably contributed to the project by handling literature research for methods, model construction and validation, and structuring the final report. Mia Yu (my2838) played a key role in model selection, additional methods research, and enhancing model diagnostics. Eunice Wang (cw3555) was crucial in initial data analysis, variable selection, relationship exploration, and writing the report's introduction and conclusion.

Figures and Tables

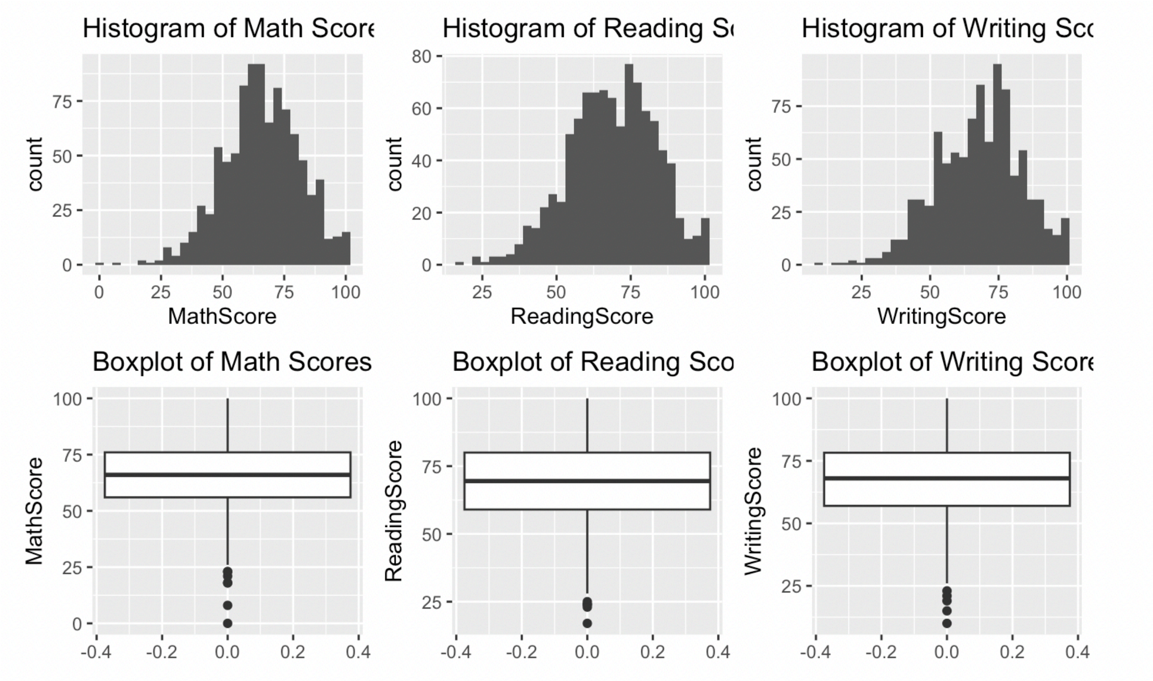

Figure 1: Ys' distribution

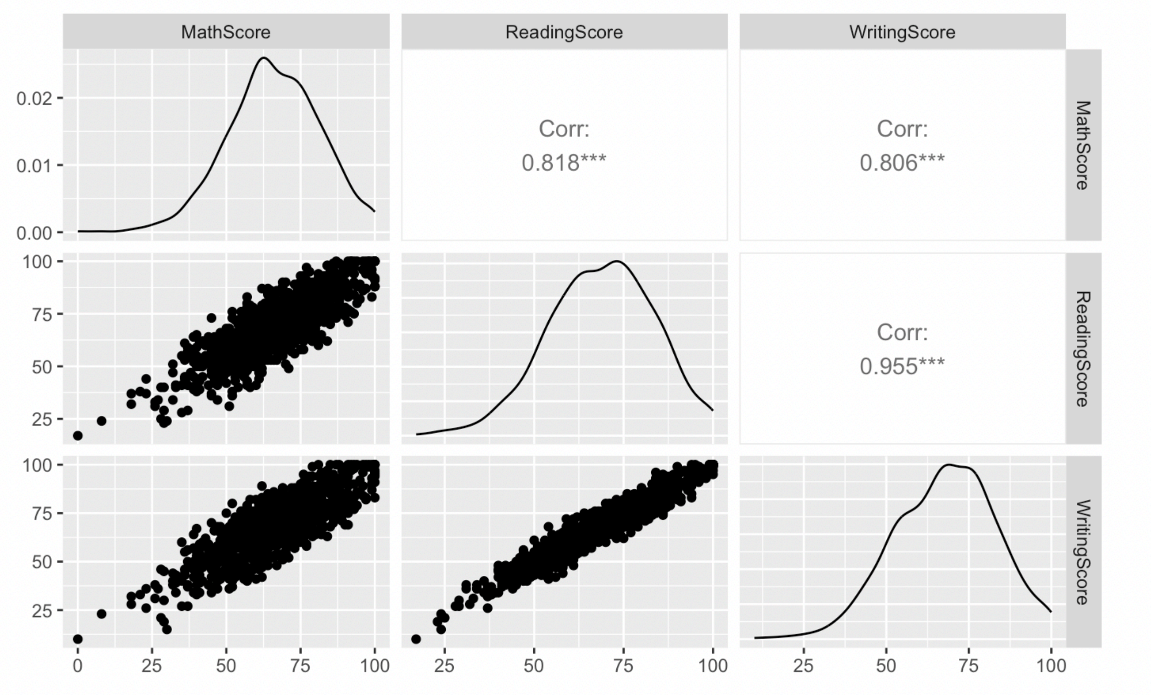

Figure 2: Ys’ correlation

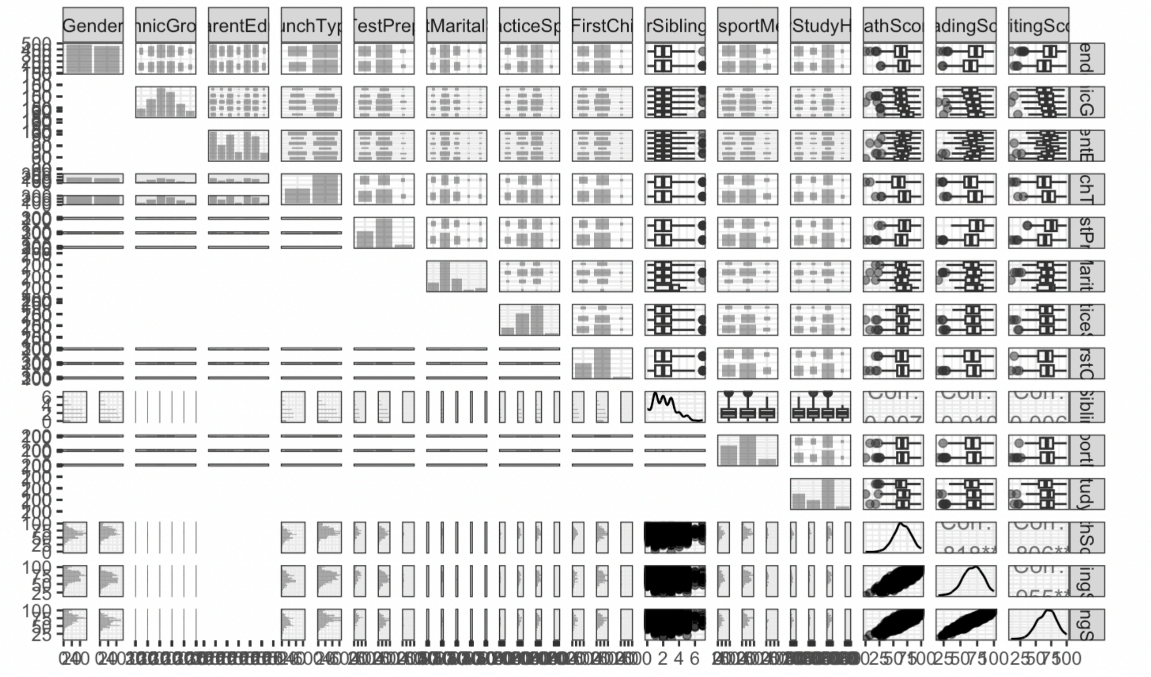

Figure 3: pairwise

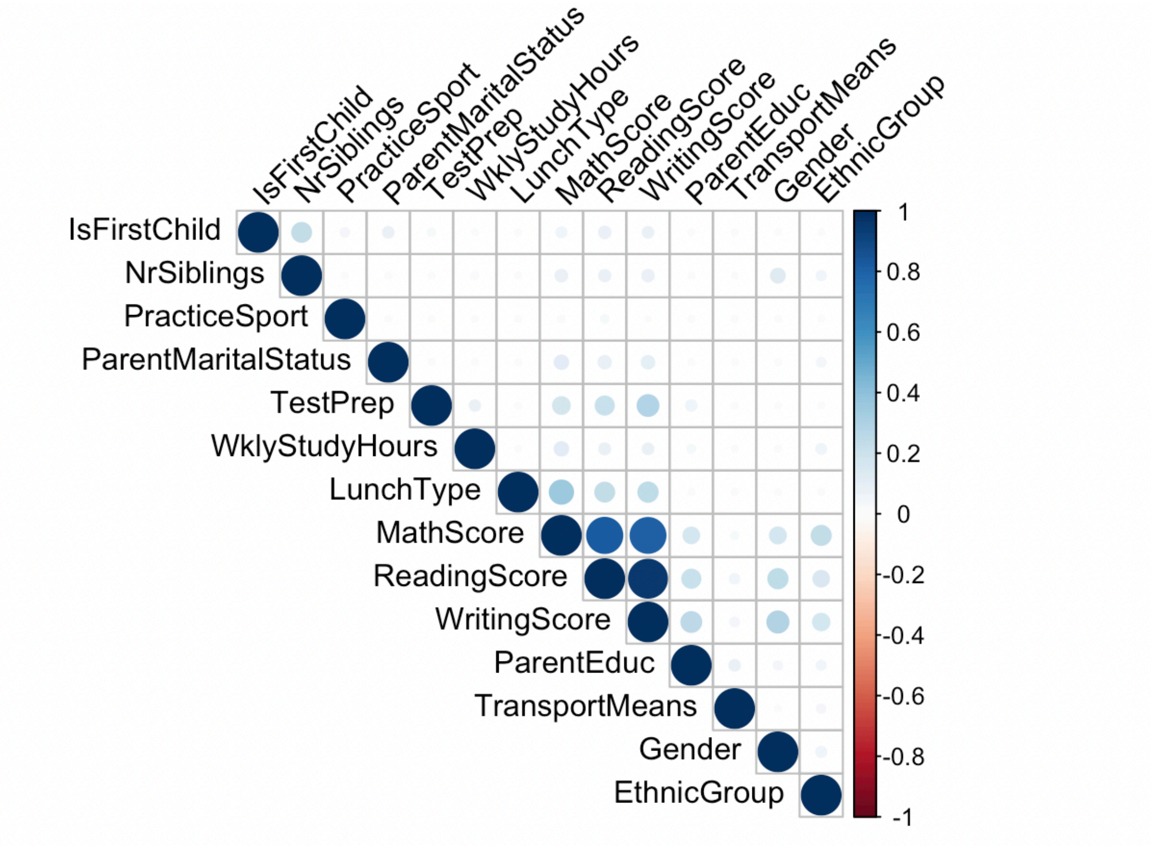

Figure 4: heatmap

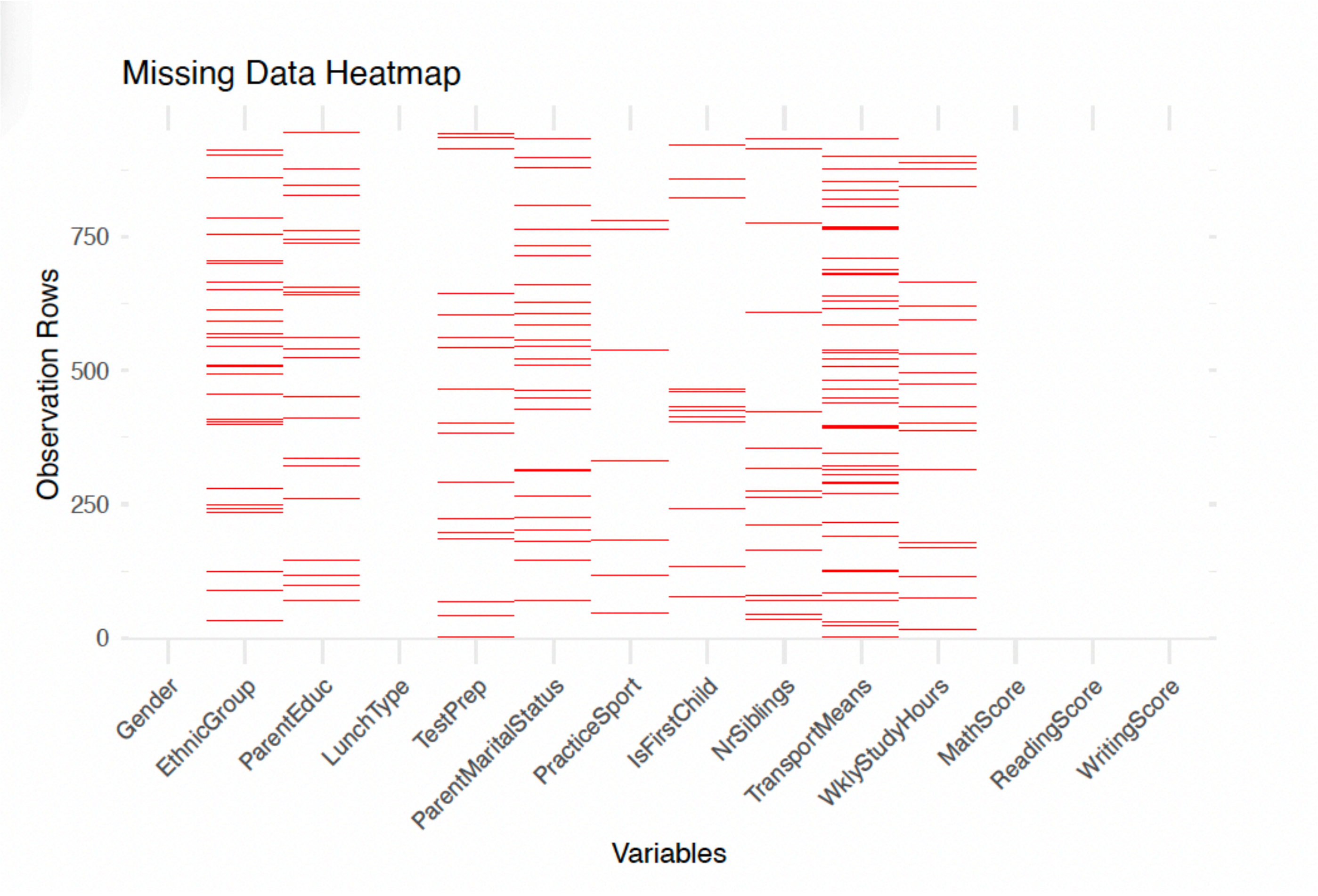

Figure 5: missingdata

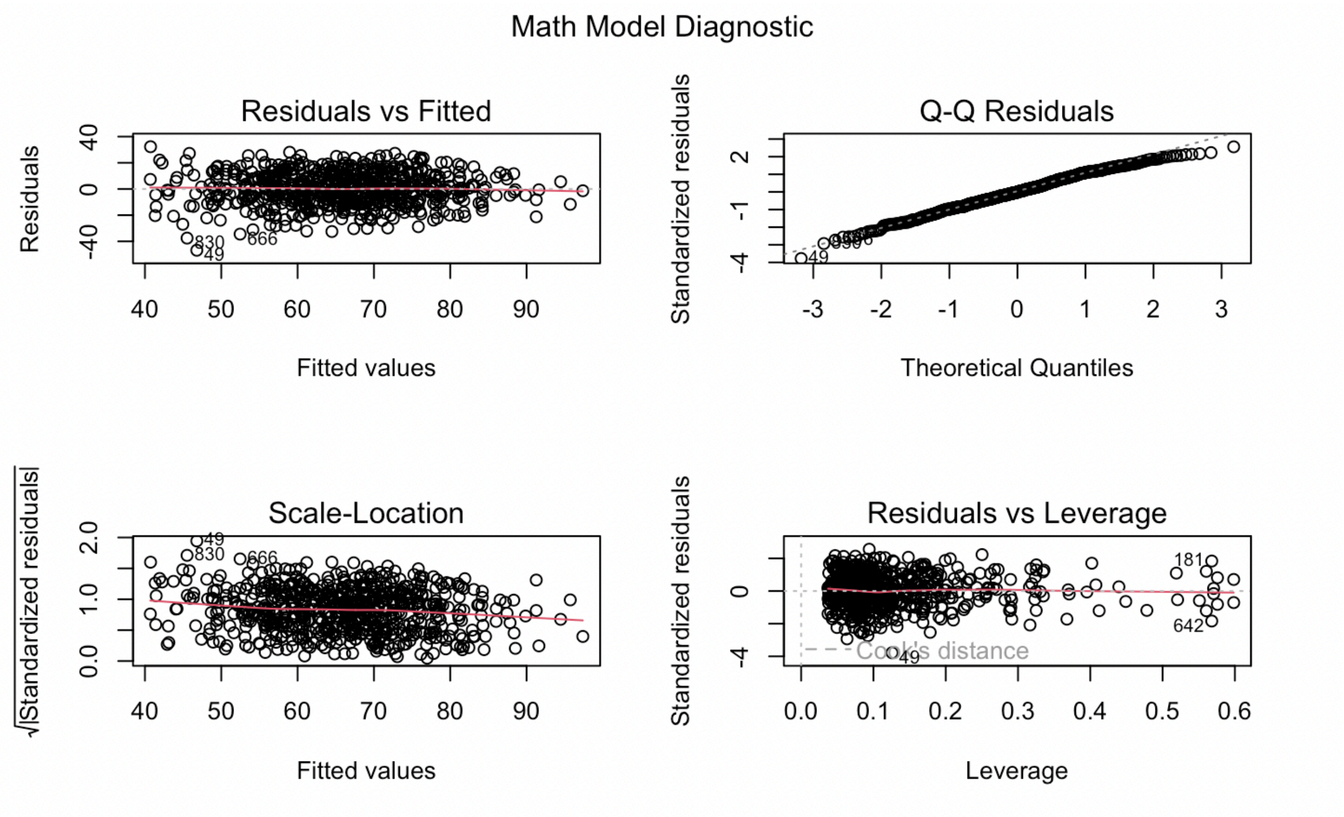

Figure 6: math

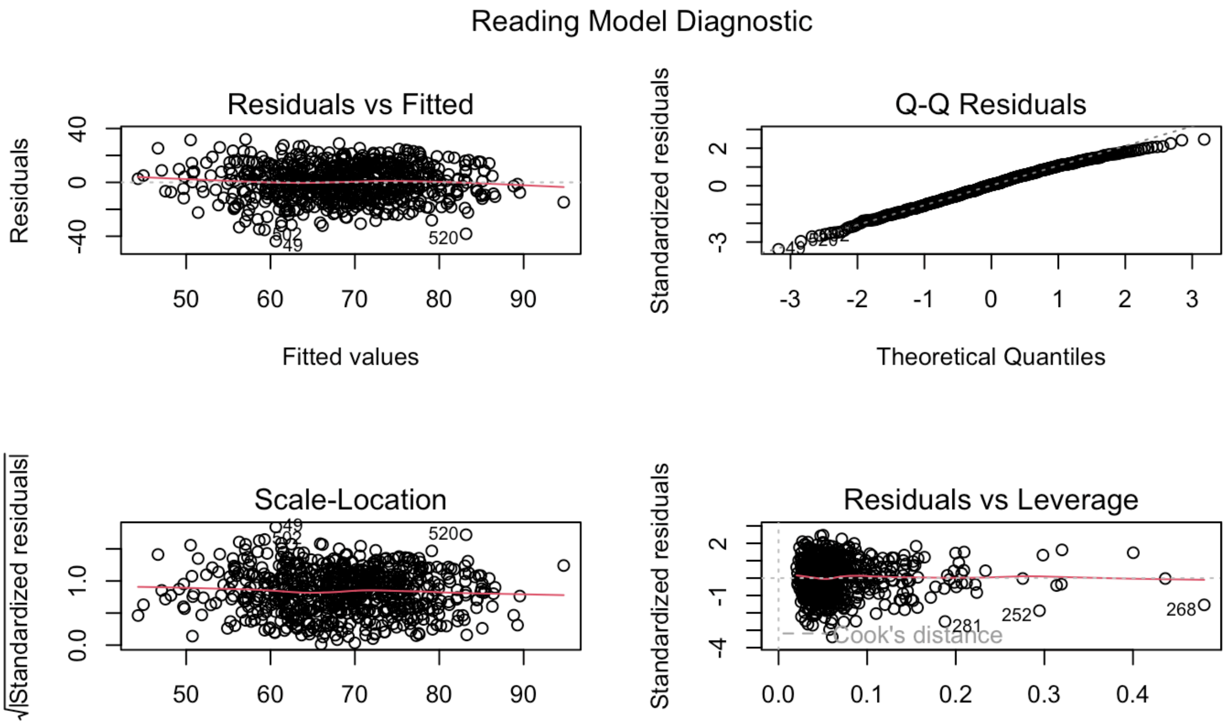

Figure 7: reading

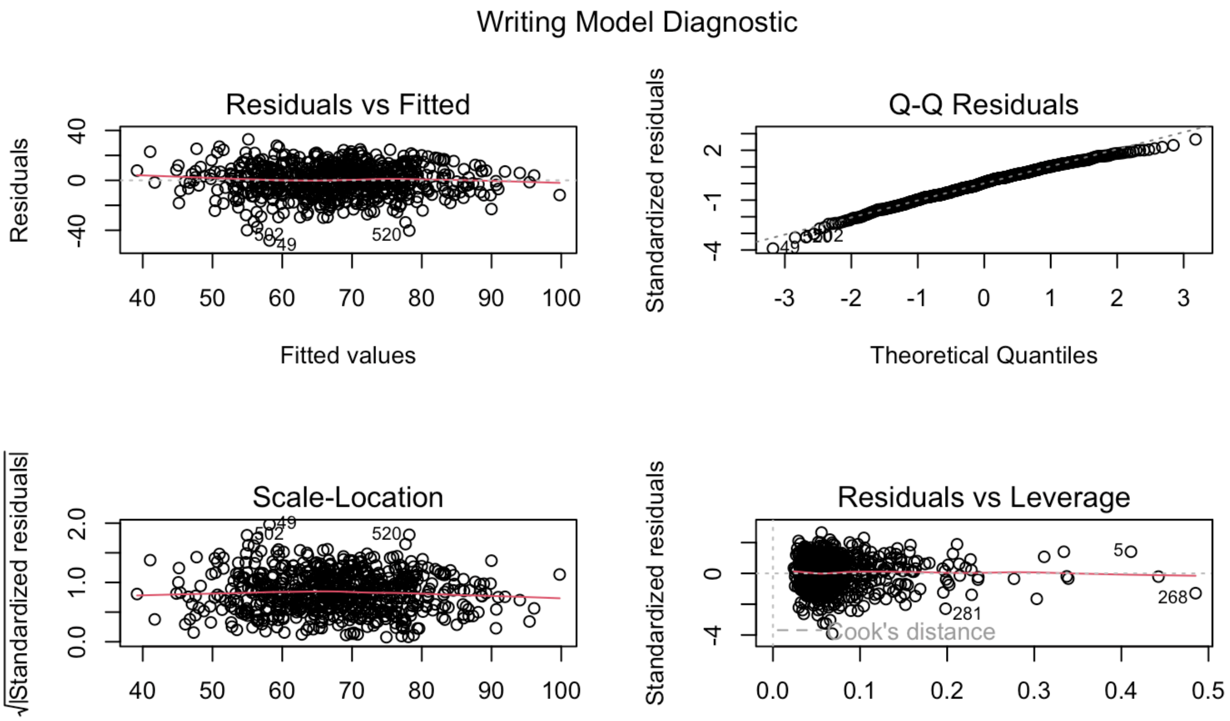

Figure 8: writing

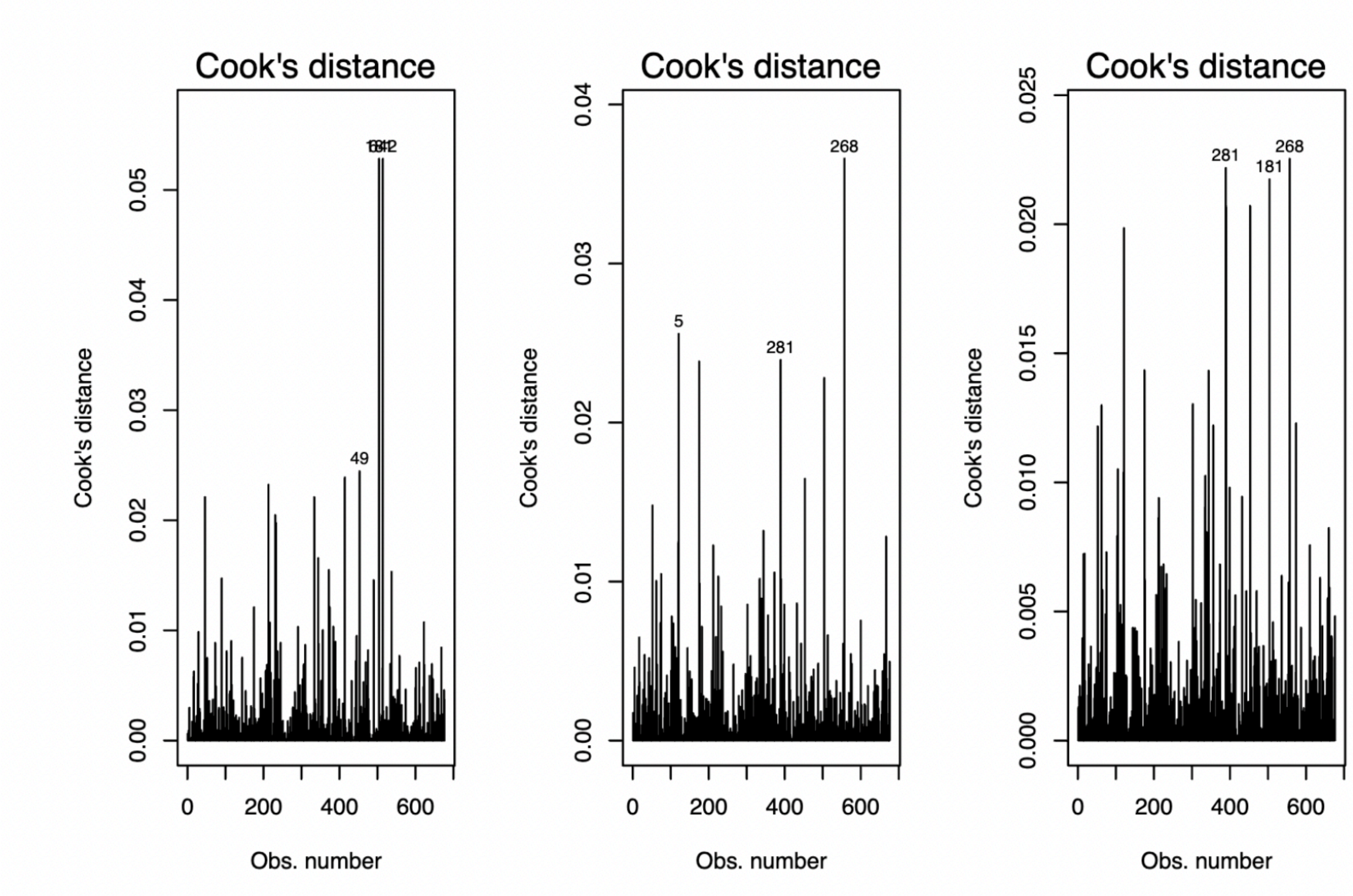

Figure 9: Cook's Distance

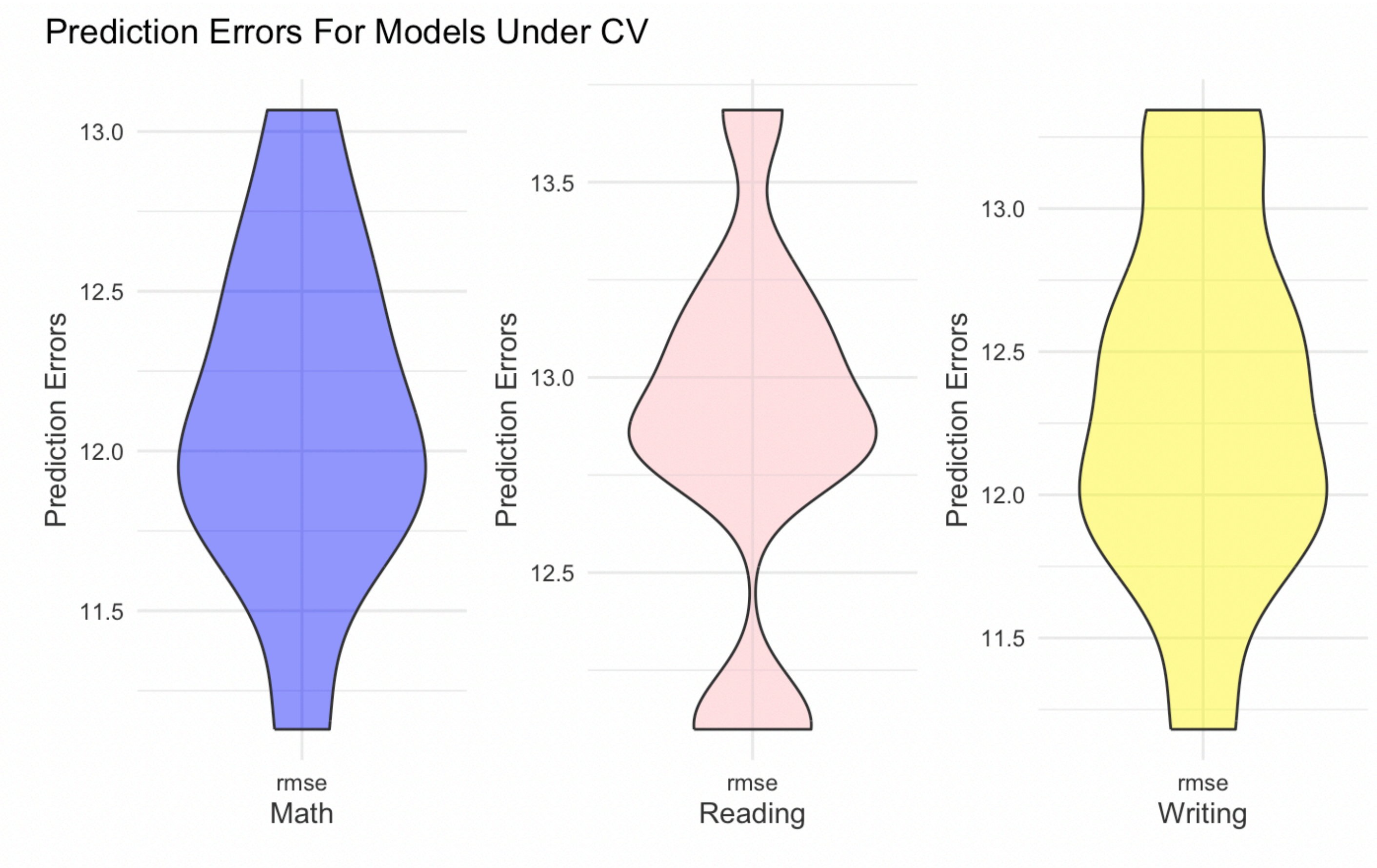

Figure 10: CV outcome

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

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-□□□□□□coefficient□□□

```
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
```