

# P8130 Final Report (Project 1)

Huanyu Chen (hc3451) Xiaoting Tang (xt2288)  
Yifei Liu (yl5508) Longyu Zhang (lz2951)

## Abstract

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

## Introduction

The objective of this study is to use regression models to predict academic performance in math, reading, and writing based on various variables, including personal characteristics such as gender, ethnicity, and parental education, as well as environmental factors like lunch type, test preparation, and weekly study hours. Furthermore, the study aims to identify potential correlations and regression model between scores in different subjects. The combination of these analyses is intended to provide educators and policymakers with practical insights for tailoring interventions, improving educational programs, and building strong support structures that promote students' overall academic progress.

## Methods

This dataset provides information on public school students, including three test scores and various personal and socioeconomic factors. To facilitate analysis, categorical data have been converted to numerical representations based on their ordinal order or type. We have excluded the missing cells because they are factorial data types.

After processing the data, we created **Table 1**, which presents a summary of the factorial data, including the number of missing data, the number of categories under each variable, and the top

counts. For the numeric data (three test scores), we constructed a comprehensive descriptive table (**Table 2**) to provide a snapshot of central tendencies and variability. The distribution of the three response variables (test scores) is presented in **Figure 1** (histogram) and **Figure 2** (boxplot), indicating a normal distribution.

Then we fitted the “full model” using the score of three subjects respectively as the response variables, which consists of all 11 categorical variables as predictors. The model diagnostics are conducted by generating four plot for each model: Residuals vs Fitted, Q-Q Residuals, Scale-Location and Residuals vs Leverage (**Figure 3-5**). Next, we use BIC-based procedures to select the appropriate subsets of predictors for three subjects (**Figure 6-9**).

Based on the full models, we did some tests and calculations:

First, we conducted boxcox method (**Figure 6**) to determine if there’s any transformation needed. Second, calculated Cook’s distance (**Figure 7**) to check the existence of outliers and influence points.

Finally, in order to test the multicollinearity among predictors, we calculated VIF as the criterion of multicollinearity (**Table 3-5**).

After all the steps above, we conducted model selection using both stepwise selection method and LASSO method. For stepwise method, the remaining predictors, coefficients and p-values are reported in **Table 7-9**.

In the selection procedure using LASSO method, for each subject we used cross-validation to decide the optimal value of method parameter  $\lambda$ , and then fitted LASSO model with this optimal value (**Figure 8-10**).

Finally, we tried to figure out if it is possible to leverage one score as the auxiliary information to learn the model for another score (still its model against variables 1-11) better. we plotted the correlation among three score variables (**Figure 11**). Then we refitted the linear models for the scores of three subjects using eleven categorical variables and one other score variable of a different subject as predictors (**Figure 12-17**). The VIFs are calculated for all six models generated in this step to reveal the potential multicollinearity (**Table 17-22**).

## Results

**Table 3**, **Table 4**, and **Table 5** display the regression models for math, reading, and writing scores using both forward and backward stepwise regression. Moreover, **Figures 3**, **Figures 4**, and **Figures 5** display the diagnostic plots generated by the model.

## Conclusions/Discussion

## Contribution

**Xiaoting Tang**: Method, **Yifei Liu**: Result Display

**Longyu Zhang**: Interpretation, **Huanyu Chen**: Writing

## Appendix

### Table

Table 1: Categorical Variables pre-analysis

Variable	Missing	Unique	Top Counts
gender	0	2	1: 488, 0: 460
ethnic_group	59	5	2: 277, 3: 237, 1: 171, 4: 124
parent_educ	392	4	1: 199, 2: 198, 3: 104, 4: 55
lunch_type	0	2	0: 617, 1: 331
test_prep	55	2	0: 571, 1: 322
parent_marital_status	49	4	0: 516, 1: 213, 3: 146, 2: 24
practice_sport	16	3	1: 477, 2: 343, 0: 112
is_first_child	30	2	1: 604, 0: 314
nr_siblings	46	8	1: 245, 2: 213, 3: 198, 0: 101
transport_means	102	2	0: 509, 1: 337
wkly_study_hours	37	3	1: 508, 0: 253, 2: 150

Table 2: Continuous Variables pre-analysis

Variable	Mean	SD	Min	Q1	Median	Q3	Max
math_score	68.7	15.9	18	57	69.0	81	100
reading_score	72.3	14.8	23	61	73.0	84	100
writing_score	72.0	15.2	19	62	72.5	84	100

Table 3: Math Scores Models by Stepwise Regression

Term	Estimate	P Value
gender1	-3.70	0.01
ethnic_group1	2.45	0.45
ethnic_group2	0.30	0.92
ethnic_group3	4.17	0.18
ethnic_group4	10.18	0.00
lunch_type1	-12.38	0.00
test_prep1	6.08	0.00
parent_marital_status1	-4.08	0.02
parent_marital_status2	6.80	0.14
parent_marital_status3	-5.25	0.01
wkly_study_hours1	5.92	0.00
wkly_study_hours2	3.83	0.08

Table 4: Reading Scores Models by Stepwise Regression

Term	Estimate	P Value
gender1	8.18	0.00
ethnic_group1	1.89	0.54
ethnic_group2	0.38	0.90
ethnic_group3	3.38	0.26

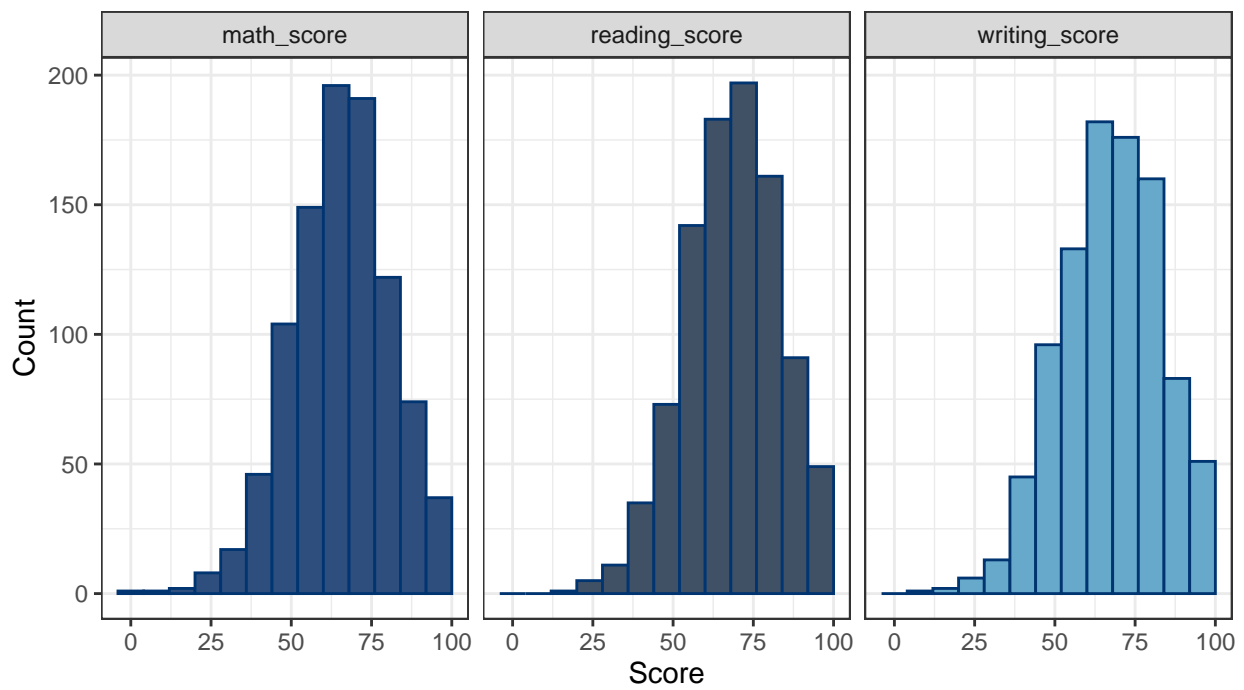
Term	Estimate	P Value
ethnic_group4	5.69	0.07
parent_educ2	2.40	0.15
parent_educ3	4.67	0.02
parent_educ4	6.49	0.01
lunch_type1	-8.26	0.00
test_prep1	7.62	0.00
parent_marital_status1	-4.60	0.01
parent_marital_status2	4.18	0.34
parent_marital_status3	-4.30	0.03
wkly_study_hours1	5.16	0.00
wkly_study_hours2	1.05	0.62

Table 5: Writing Scores Models by Stepwise Regression

Term	Estimate	P Value
gender1	10.03	0.00
ethnic_group1	2.21	0.46
ethnic_group2	1.85	0.52
ethnic_group3	6.34	0.03
ethnic_group4	6.62	0.03
parent_educ2	1.79	0.27
parent_educ3	4.60	0.02
parent_educ4	7.21	0.00
lunch_type1	-9.26	0.00
test_prep1	9.61	0.00
parent_marital_status1	-4.42	0.01
parent_marital_status2	4.67	0.28
parent_marital_status3	-4.64	0.02
wkly_study_hours1	5.17	0.00
wkly_study_hours2	1.89	0.36

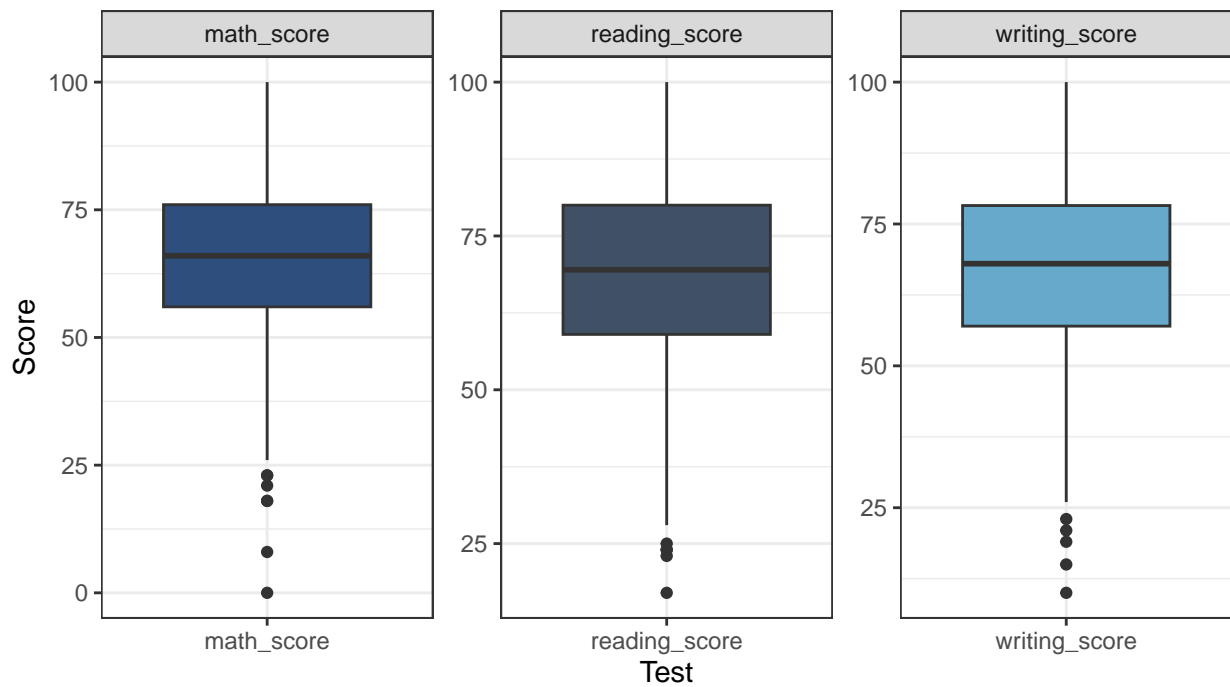
Figure

**Figure 1: Scores Histograms by Subjects**



test math\_score reading\_score writing\_score

**Figure 2: Scores Boxplot by Subjects**



test math\_score reading\_score writing\_score

Figure 3: Diagnostic Plots for Math Test Score Regression Model

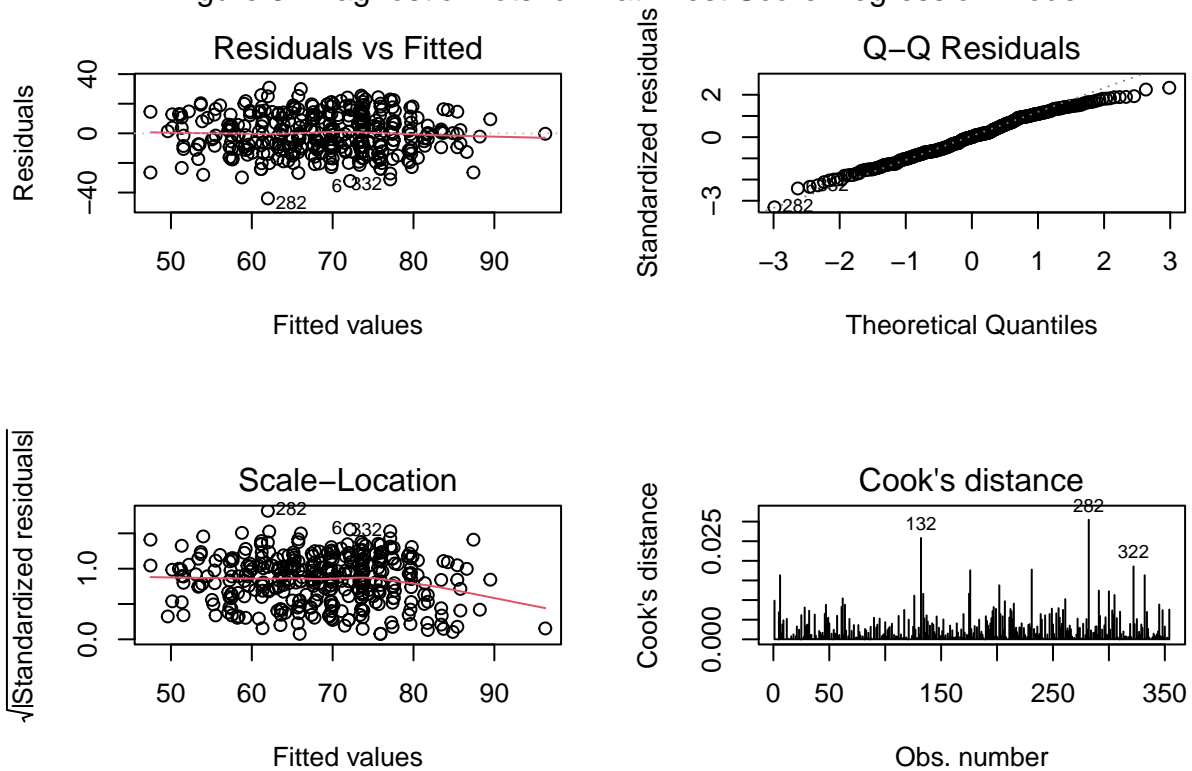


Figure 4: Diagnostic Plots for Reading Test Score Regression Model

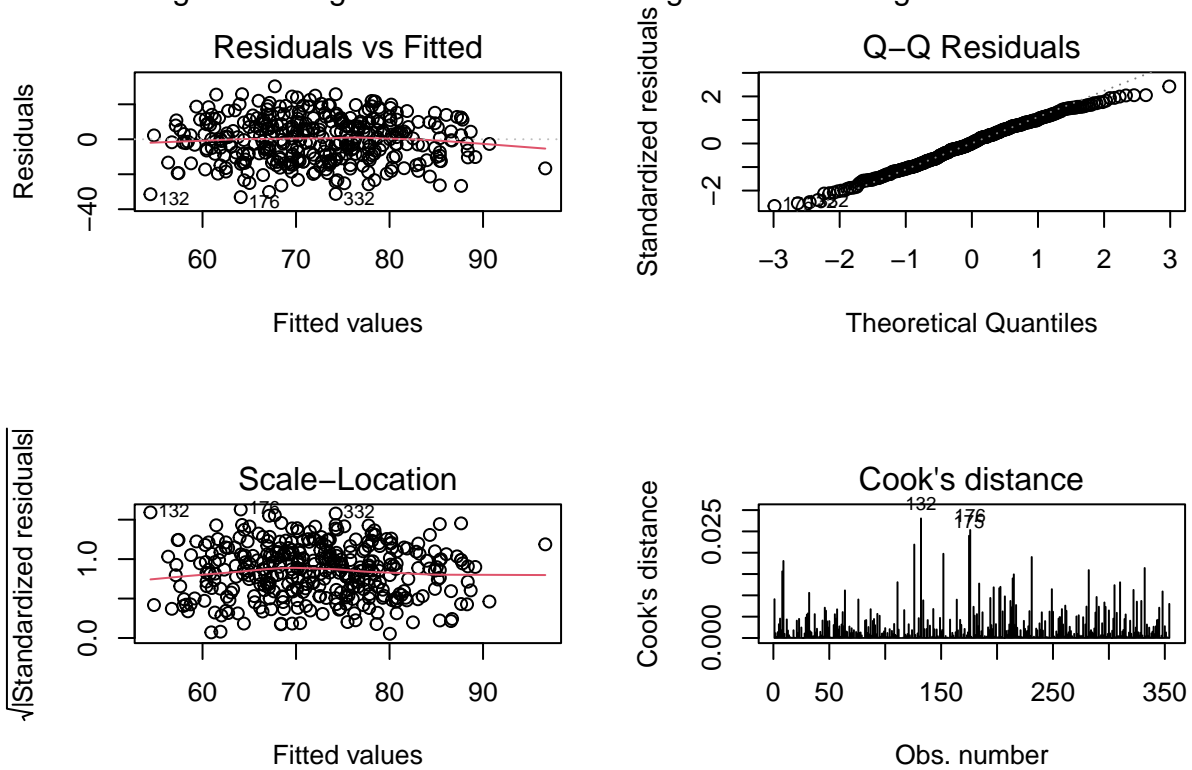


Figure 5: Diagnostic Plots for Writing Test Score Regression Model

