P8130 Final Report (Project 1)

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

Abstract

This study used regression modeling techniques to predict academic performance in math, reading, and writing based on multiple variables, including personal characteristics and socio-environmental factors. Using stepwise regression, criterion-based procedures, and LASSO analyses, we found significant relationships between test scores and several covariants, including gender, race, parent education, lunch type, test preparation, and weekly study time. We further found that including another test score as a predictor (e.g., consider reading scores when predicting math scores) significantly enhanced model performance. This comprehensive analysis reveals the complex relationships between predictors and student performance, providing suggestions to educators and thus achieving overall student progress.

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, test preparation, and weekly study hours, as well as socio-environmental factors like lunch type, transportation and parental education. Furthermore, the study aims to identify potential correlations and regression models 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 to students' overall academic progress.

Methods

This data set 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.

Before pre-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. Then, we have excluded the missing cells because predictor variables we have are factorial data types. 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 plots for each model: Residuals vs Fitted, Q-Q Residuals, Scale-Location, and Residuals vs Leverage (**Figure 3,4,5**). Next, we use BIC-based procedures to select the appropriate subsets of predictors for three subjects (**Figure 6,7,8,9**).

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

First, we conducted the boxcox method (**Figure 10**) to determine if there was any transformation needed. Second, we calculated Cook's distance (**Figure 11**) 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 3,4,5**).

After all the steps above, we conducted model selection using stepwise selection method, criterion-based procedures and LASSO method. For stepwise method, the remaining predictors, coefficients, and p-values are reported in **Table 6,7,8**.

In the selection procedure using LASSO method, for each subject we used cross-validation to decide the optimal value of method parameter λ, and then fitted LASSO model with this optimal value (**Figure 12,13,14**).

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 15**). 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 (**Table 9,10,11,12,13,14**). The VIFs are calculated for all six models generated in this step to reveal the potential multicollinearity (**Table 15,16,17,18,19,20**).

Results

Table 1 gives the summary of the factorial data. **Table 2** provides the mean, deviation, and quantile information about the continuous data (score variables of three subjects). A total of 786 missing values were identified. After removing the missing values from the original dataset, 354 observations remained out of the initial 948. Since categorical variables hold specific meanings on their values, we opted not to impute missing values using the mean.

The distributions of three response score variables are demonstrated in **Figure 1** (histogram) and **Figure 2** (boxplot). Both the histogram and boxplot distributions showed a normal distribution.

Overall, the score distributions for the three subjects are quite similar, with an average around 70 points. As indicated by the boxplot, outliers are predominantly situated within the lower score ranges.

Figure 3, **Figure 4**, and **Figure 5** display the diagnostic plots generated by the model. The diagnosis of full models indicated that the model adheres to the basic assumptions of linear regression, including homoscedasticity, normality, independent residuals with constant variance, and that there was no need for transformations.

Table 3, **Table 4**, and **Table 5** display the regression models for math, reading, and writing scores using "both" strategy stepwise regression. To be specific, `gender`, `ethnic group`, `lunch type`, `test preparation`, `parent marital status`, and `weekly study hours` were found to be significant variables for math scores, with an adjusted R-squared of 0.2742. Similarly, the stepwise regression models for reading and writing scores also revealed statistically significant relationships between test scores and predictors, including `gender`, `ethnicity`, `parental education`, `lunch type`, `test preparation`, `parental marital status`, and `weekly study hours`. These predictors explain the model with an adjusted R-squared of 0.2513 for reading and 0.3298 for writing.

Figure 6 displays the relationship between BIC values and the number of selected model variables. Criterion-based procedures suggest that employing regression models for three test scores with 4–5 predictors will yield optimal results with a minimum Bayesian information criterion. These approaches guide the construction of effective models for the three test scores.

Figure 10 demonstrates the results of boxcox method: the log-likelihood over boxcox method parameter λ, and **Figure 11** demonstrates the Cook's distance as the result of testing of outliers.

Table 3,4,5 display the result of multicollinearity test: VIFs for three full models.

Figure 12,13,14 are plots of mean cross-validation error over LASSO parameter λ, providing information about optimal values of λ for LASSO models. These optimal values are used for the following fitting of LASSO models.

Figure 15 shows the correlation existing among score variables, suggesting the feasibility of using one of the score variables as the auxiliary information to learn the model for another score. **Table 9,10,11,12,13,14** are the results of regression using one subject score as an additional predictor for the prediction of another subject score. **Table 15,16,17,18,19,20** shows the VIFs of these six models, indicating the multicollinearity among predictors.

Conclusions

Overall, we tried several regression models that worked successfully in pointing out the key factors affecting math, reading, and writing scores and in quantifying the specific effects of different classifications in these factors. While the categorical data provided viable results, we found that some points were still not explained by the regression given the relatively low coeficient of determination. This limitation suggests the need for broader and more detailed data collection. With continuous refinement of the data, we may be able to effectively adjust educational strategies in the future.

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

Term	Estimate	P Value
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

```
lambda_seq = 10 ^ seq(-2, 2, by = .1)


#math

cv_object_math =

  cv.glmnet(as.matrix(data[1:11]), data$math_score,

                       lambda = lambda_seq,

                       nfolds = 5)


opt_lambda_math = cv_object_math$lambda.min


#math result

model_math_lasso = glmnet(as.matrix(data[1:11]), data$math_score, lambda = opt_lambda_math, al

res_math_lasso =
```

```r
  coef(model_math_lasso) |>
  as.matrix() |>
  as_tibble() |>
  mutate(term = rownames(coef(model_math_lasso))) |>
  dplyr::select(term, everything()) |>
  filter(term != "(Intercept)") |>
  filter(s0 != 0)
colnames(res_math_lasso) = c("Term", "Estimate")
knitr::kable(x = res_math_lasso, caption = "Math Scores Models by Lasso Model", digits = 1)
```

Table 6: Math Scores Models by Lasso Model

Term	Estimate
gender	-3.4
ethnic_group	2.1
parent_educ	1.0
lunch_type	-11.8
test_prep	5.0
parent_marital_status	-1.0
practice_sport	0.4
nr_siblings	0.7
wkly_study_hours	2.4

```
#reading
cv_object_reading =
  cv.glmnet(as.matrix(data[1:11]), data$reading_score,
                       lambda = lambda_seq,
                       nfolds = 5)

opt_lambda_reading = cv_object_reading$lambda.min

#reading result
model_reading_lasso = glmnet(as.matrix(data[1:11]), data$reading_score, lambda = opt_lambda_rea
res_reading_lasso =
  coef(model_reading_lasso) |>
  as.matrix() |>
  as_tibble() |>
  mutate(term = rownames(coef(model_reading_lasso))) |>
  dplyr::select(term, everything()) |>
  filter(term != "(Intercept)") |>
  filter(s0 != 0)
colnames(res_reading_lasso) = c("Term", "Estimate")
knitr::kable(x = res_reading_lasso, caption = "Reading Scores Models by Lasso Model", digits =
```

Table 7: Reading Scores Models by Lasso Model

Term	Estimate
gender	6.9
ethnic_group	1.0
parent_educ	1.7
lunch_type	-7.2
test_prep	6.3
parent_marital_status	-0.8
wkly_study_hours	0.5

```
#writing
cv_object_writing =
  cv.glmnet(as.matrix(data[1:11]), data$writing_score,
                       lambda = lambda_seq,
                       nfolds = 5)

opt_lambda_writing = cv_object_writing$lambda.min

#writing result
model_writing_lasso = glmnet(as.matrix(data[1:11]), data$writing_score, lambda = opt_lambda_wri
res_writing_lasso =
  coef(model_writing_lasso) |>
  as.matrix() |>
```

```
  as_tibble() |>
  mutate(term = rownames(coef(model_writing_lasso))) |>
  dplyr::select(term, everything()) |>
  filter(term != "(Intercept)") |>
  filter(s0 != 0)
colnames(res_writing_lasso) = c("Term", "Estimate")
knitr::kable(x = res_writing_lasso, caption = "Writing Scores Models by Lasso Model", digits =
```

Table 8: Writing Scores Models by Lasso Model

Term	Estimate
gender	8.7
ethnic_group	1.5
parent_educ	1.9
lunch_type	-8.1
test_prep	8.1
parent_marital_status	-0.8
nr_siblings	0.1
wkly_study_hours	0.8

Table 9: VIF for Math Score

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.4]	0.9

Term	VIF	VIF_CI	Tolerance
ethnic_group	1.2	[1.1, 1.4]	0.8
parent_educ	1.2	[1.1, 1.4]	0.8
lunch_type	1.1	[1, 1.4]	1.0
test_prep	1.1	[1, 1.3]	0.9
parent_marital_status	1.2	[1.1, 1.4]	0.9
practice_sport	1.2	[1.1, 1.4]	0.9
is_first_child	1.2	[1.1, 1.3]	0.9
nr_siblings	1.5	[1.4, 1.8]	0.6
transport_means	1.1	[1, 1.3]	0.9
wkly_study_hours	1.1	[1.1, 1.3]	0.9

Table 10: VIF for Reading Score

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.4]	0.9
ethnic_group	1.2	[1.1, 1.4]	0.8
parent_educ	1.2	[1.1, 1.4]	0.8
lunch_type	1.1	[1, 1.4]	1.0
test_prep	1.1	[1, 1.3]	0.9
parent_marital_status	1.2	[1.1, 1.4]	0.9
practice_sport	1.2	[1.1, 1.4]	0.9
is_first_child	1.2	[1.1, 1.3]	0.9

Term	VIF	VIF_CI	Tolerance
nr_siblings	1.5	[1.4, 1.8]	0.6
transport_means	1.1	[1, 1.3]	0.9
wkly_study_hours	1.1	[1.1, 1.3]	0.9

Table 11: VIF for Reading Score

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.4]	0.9
ethnic_group	1.2	[1.1, 1.4]	0.8
parent_educ	1.2	[1.1, 1.4]	0.8
lunch_type	1.1	[1, 1.4]	1.0
test_prep	1.1	[1, 1.3]	0.9
parent_marital_status	1.2	[1.1, 1.4]	0.9
practice_sport	1.2	[1.1, 1.4]	0.9
is_first_child	1.2	[1.1, 1.3]	0.9
nr_siblings	1.5	[1.4, 1.8]	0.6
transport_means	1.1	[1, 1.3]	0.9
wkly_study_hours	1.1	[1.1, 1.3]	0.9

Table 12: Math Scores Models Using Reading Score as Additional Predictor

Term	Estimate	P Value
gender1	-11.5	0.0
ethnic_group1	0.6	0.7
ethnic_group2	-0.2	0.9
ethnic_group3	0.7	0.6
ethnic_group4	4.6	0.0
lunch_type1	-4.7	0.0
test_prep1	-1.2	0.1
practice_sport1	1.4	0.2
practice_sport2	2.4	0.0
wkly_study_hours1	1.4	0.1
wkly_study_hours2	3.4	0.0
reading_score	0.9	0.0

Table 13: Math Scores Models Using Writing Score as Additional Predictor

Term	Estimate	P Value
gender1	-13.7	0.0
ethnic_group1	0.1	0.9
ethnic_group2	-1.6	0.2

Term	Estimate	P Value
ethnic_group3	-2.2	0.1
ethnic_group4	3.5	0.0
parent_educ2	-0.1	0.9
parent_educ3	-1.4	0.1
parent_educ4	-3.8	0.0
lunch_type1	-3.1	0.0
test_prep1	-3.7	0.0
wkly_study_hours1	1.0	0.2
wkly_study_hours2	2.3	0.0
writing_score	1.0	0.0

Table 14: Reading Scores Models Using Math Score as Additional Predictor

Term	Estimate	P Value
gender1	11.3	0.0
ethnic_group1	-0.2	0.9
ethnic_group2	0.2	0.9
ethnic_group3	-0.1	1.0
ethnic_group4	-2.9	0.0
parent_educ2	1.0	0.2
parent_educ3	2.1	0.0

Term	Estimate	P Value
parent_educ4	3.5	0.0
lunch_type1	2.2	0.0
test_prep1	2.5	0.0
is_first_child1	1.1	0.1
wkly_study_hours1	0.0	1.0
wkly_study_hours2	-2.4	0.0
math_score	0.8	0.0

Table 15: Reading Scores Models Using Writing Score as Additional Predictor

Term	Estimate	P Value
gender1	-1.6	0.0
ethnic_group1	-0.2	0.8
ethnic_group2	-1.3	0.2
ethnic_group3	-2.7	0.0
ethnic_group4	-0.7	0.5
lunch_type1	0.8	0.1
test_prep1	-1.9	0.0
practice_sport1	-1.2	0.1
practice_sport2	-1.5	0.0
is_first_child1	0.8	0.1

Term	Estimate	P Value
writing_score	1.0	0.0

Table 16: Writing Scores Models Using Math Score as Additional Predictor

Term	Estimate	P Value
gender1	13.1	0.0
ethnic_group1	0.3	0.8
ethnic_group2	1.8	0.1
ethnic_group3	2.9	0.0
ethnic_group4	-1.9	0.1
parent_educ2	0.3	0.6
parent_educ3	2.0	0.0
parent_educ4	4.3	0.0
lunch_type1	1.2	0.1
test_prep1	4.6	0.0
wkly_study_hours1	0.0	0.9
wkly_study_hours2	-1.5	0.1
math_score	0.8	0.0

Table 17: Writing Scores Models Using Reading Score as Additional Predictor

Term	Estimate	P Value
gender1	2.3	0.0
ethnic_group1	0.4	0.6
ethnic_group2	1.4	0.1
ethnic_group3	3.0	0.0
ethnic_group4	1.3	0.2
parent_educ2	-0.5	0.4
parent_educ3	0.3	0.7
parent_educ4	1.3	0.1
lunch_type1	-1.6	0.0
test_prep1	2.7	0.0
practice_sport1	1.5	0.0
practice_sport2	1.8	0.0
is_first_child1	-0.8	0.1
reading_score	0.9	0.0

Table 18: VIF for Math Score (include reading score)

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.3]	0.9

Term	VIF	VIF_CI	Tolerance
ethnic_group	1.1	[1, 1.3]	0.9
lunch_type	1.1	[1, 1.3]	0.9
test_prep	1.1	[1, 1.3]	0.9
practice_sport	1.0	[1, 1.5]	1.0
wkly_study_hours	1.1	[1, 1.3]	0.9
reading_score	1.3	[1.2, 1.5]	0.8

Table 19: VIF for Math Score (include writing score)

Term	VIF	VIF_CI	Tolerance
gender	1.2	[1.1, 1.4]	0.8
ethnic_group	1.1	[1, 1.3]	0.9
parent_educ	1.1	[1, 1.3]	0.9
lunch_type	1.1	[1.1, 1.4]	0.9
test_prep	1.2	[1.1, 1.4]	0.8
wkly_study_hours	1.1	[1, 1.3]	0.9
writing_score	1.5	[1.3, 1.7]	0.7

Table 20: VIF for Reading Score (include math score)

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.4]	1.0

Term	VIF	VIF_CI	Tolerance
ethnic_group	1.1	[1.1, 1.4]	0.9
parent_educ	1.1	[1, 1.3]	0.9
lunch_type	1.2	[1.1, 1.4]	0.8
test_prep	1.1	[1, 1.3]	0.9
is_first_child	1.0	[1, 3.7]	1.0
wkly_study_hours	1.1	[1, 1.3]	0.9
math_score	1.4	[1.2, 1.6]	0.7

Table 21: VIF for Reading Score (include writing score)

Term	VIF	VIF_CI	Tolerance
gender	1.2	[1.1, 1.4]	0.9
ethnic_group	1.1	[1, 1.3]	0.9
lunch_type	1.1	[1.1, 1.3]	0.9
test_prep	1.2	[1.1, 1.4]	0.9
practice_sport	1.1	[1, 1.4]	0.9
is_first_child	1.0	[1, 1.6]	1.0
writing_score	1.4	[1.3, 1.6]	0.7

Table 22: VIF for Writing Score (include math score)

Term	VIF	VIF_CI	Tolerance
gender	1.0	[1, 1.5]	1.0
ethnic_group	1.1	[1.1, 1.3]	0.9
parent_educ	1.1	[1, 1.4]	0.9
lunch_type	1.2	[1.1, 1.4]	0.8
test_prep	1.1	[1, 1.3]	0.9
wkly_study_hours	1.1	[1, 1.3]	0.9
math_score	1.4	[1.2, 1.6]	0.7

Table 23: VIF for Writing Score (include reading score)

Term	VIF	VIF_CI	Tolerance
gender	1.1	[1, 1.3]	0.9
ethnic_group	1.1	[1, 1.3]	0.9
parent_educ	1.1	[1, 1.3]	0.9
lunch_type	1.1	[1, 1.3]	0.9
test_prep	1.1	[1, 1.3]	0.9
practice_sport	1.1	[1, 1.3]	0.9
is_first_child	1.0	[1, 1.5]	1.0
reading_score	1.3	[1.2, 1.5]	0.8

Figure

Figure 1: Scores Histograms by Subjects

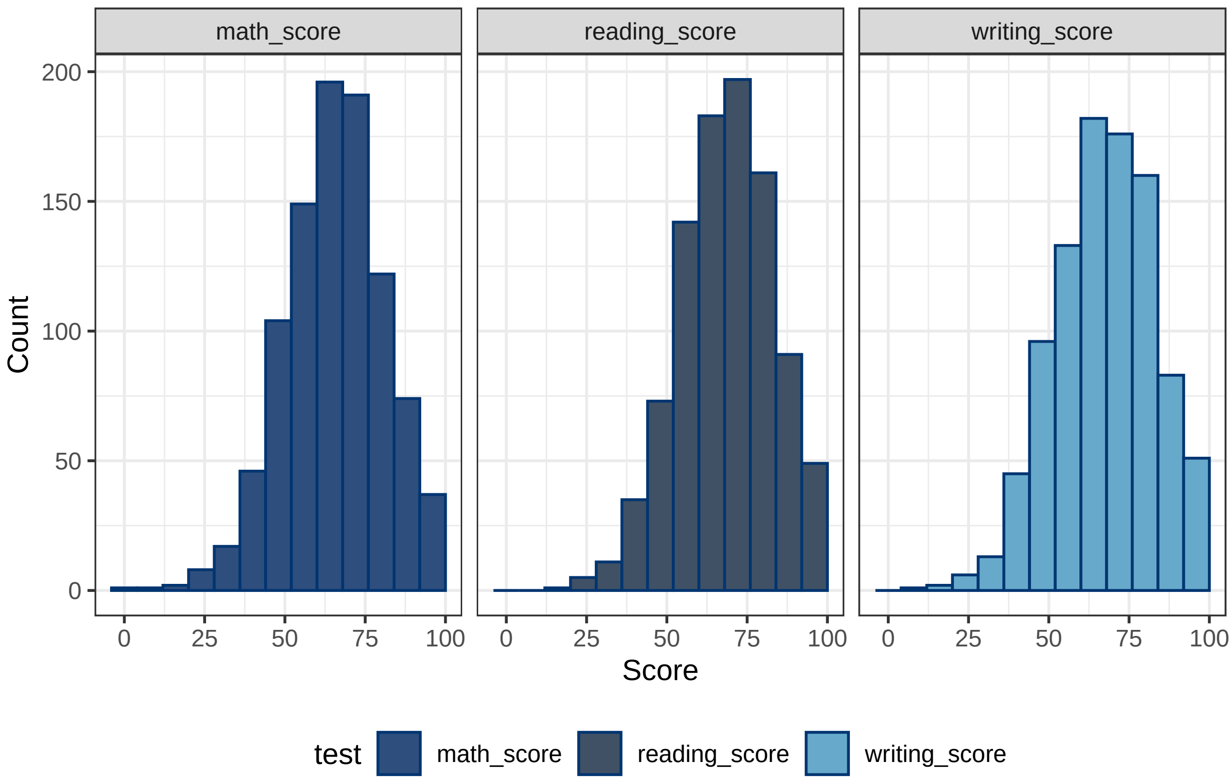

Figure 2: Scores Boxplot by Subjects

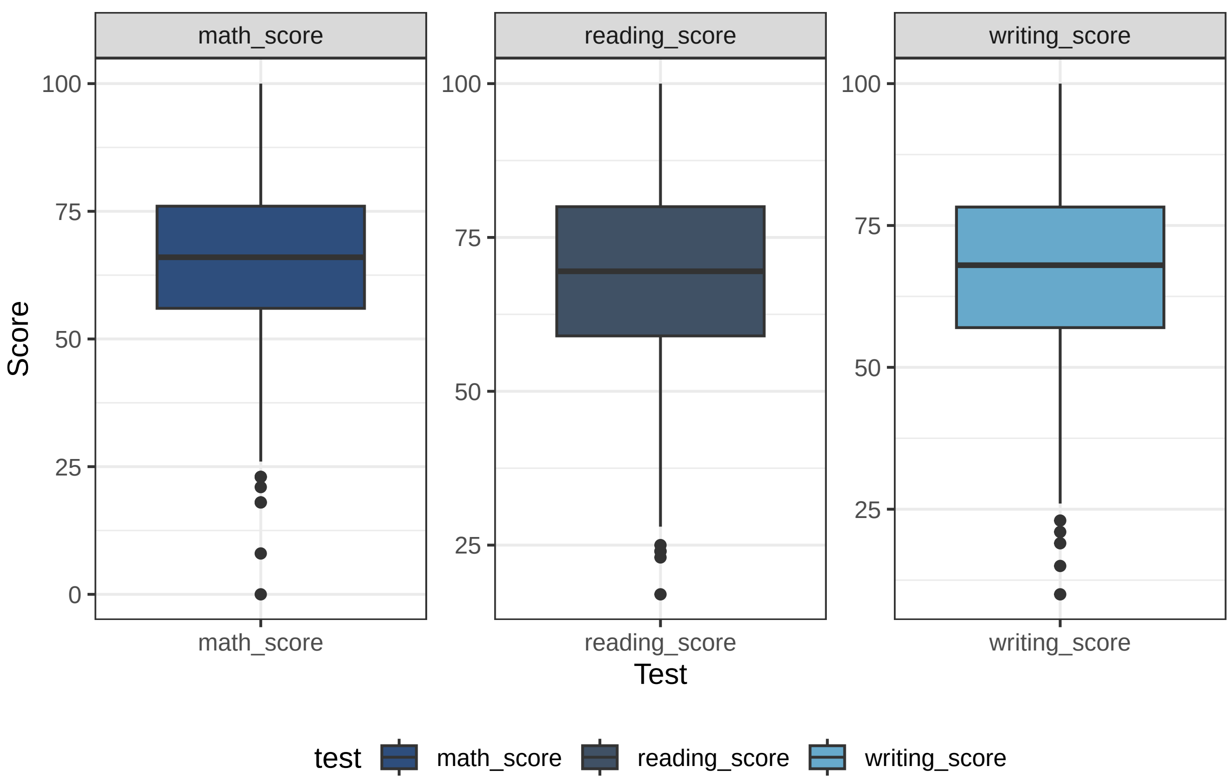

Figure 3: Diagnostic Plots for Math Test Score (Stepwise)

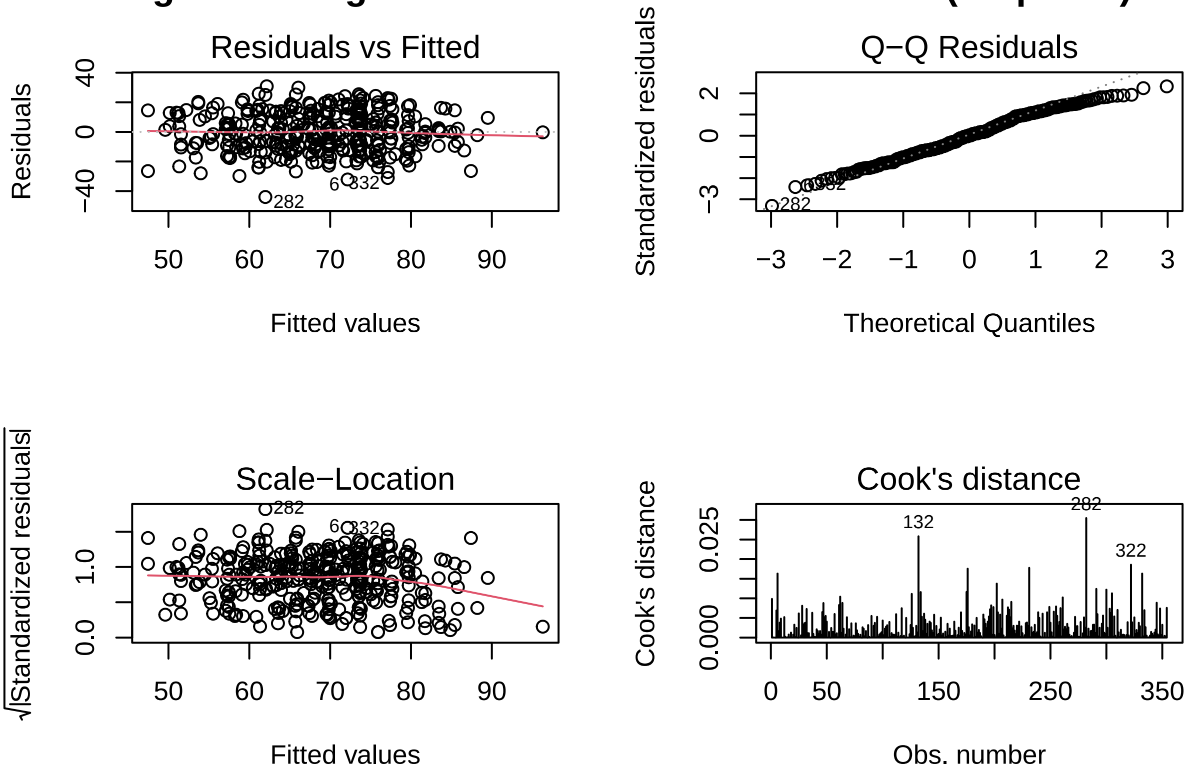

Figure 4: Diagnostic Plots for Reading Test Score (Stepwise)

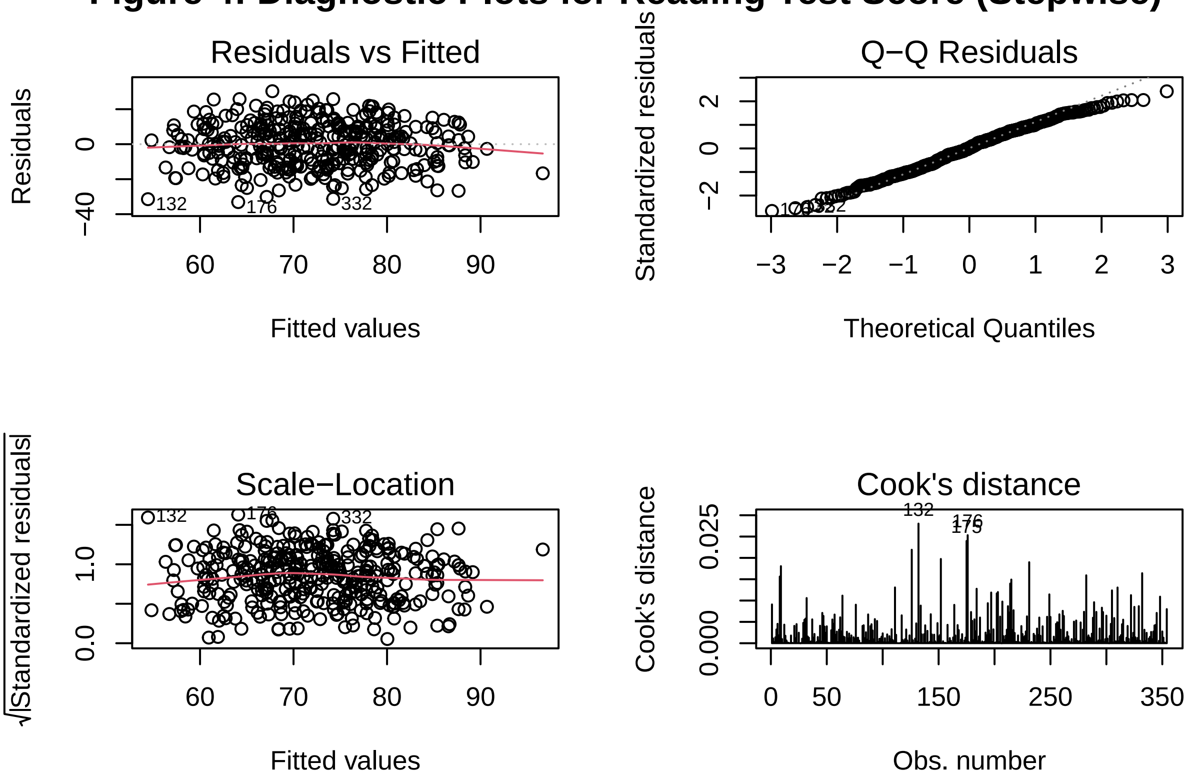

Figure 5: Diagnostic Plots for Writing Test Score (Stepwise)

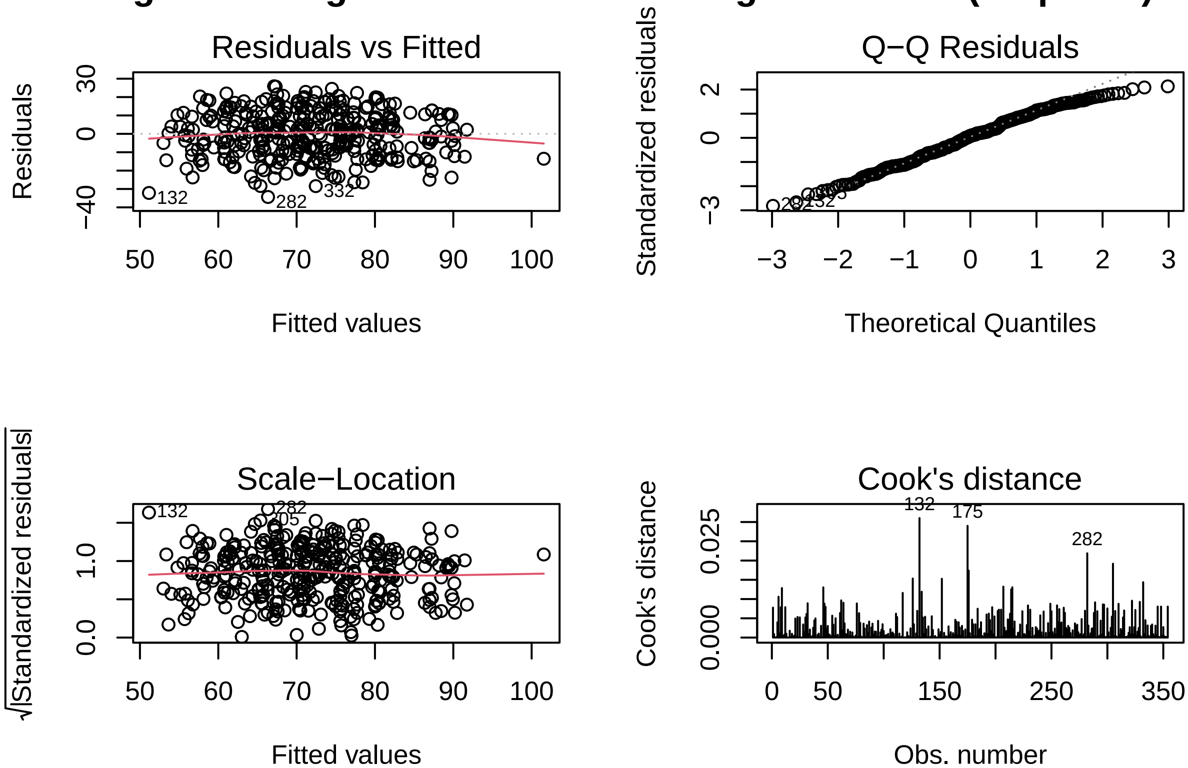

Figure 6: BIC Over Number of Parameters for Models of Three Subjects

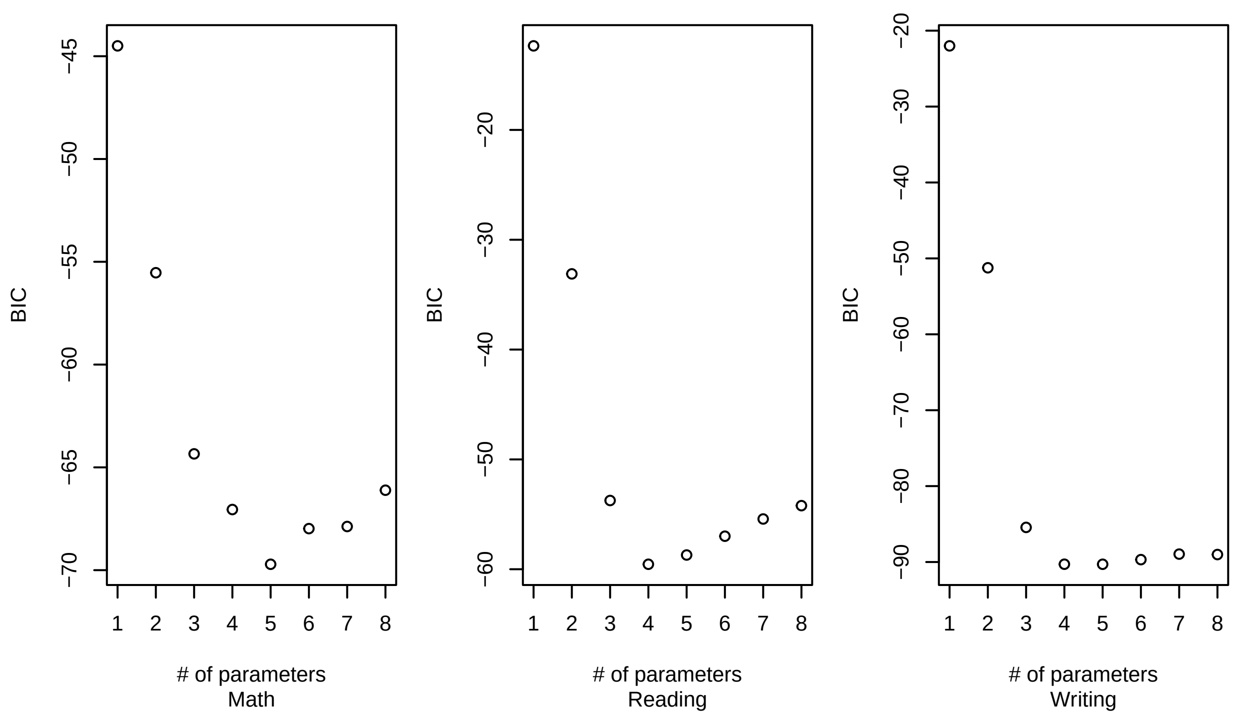

Figure 7: Boxcox Method

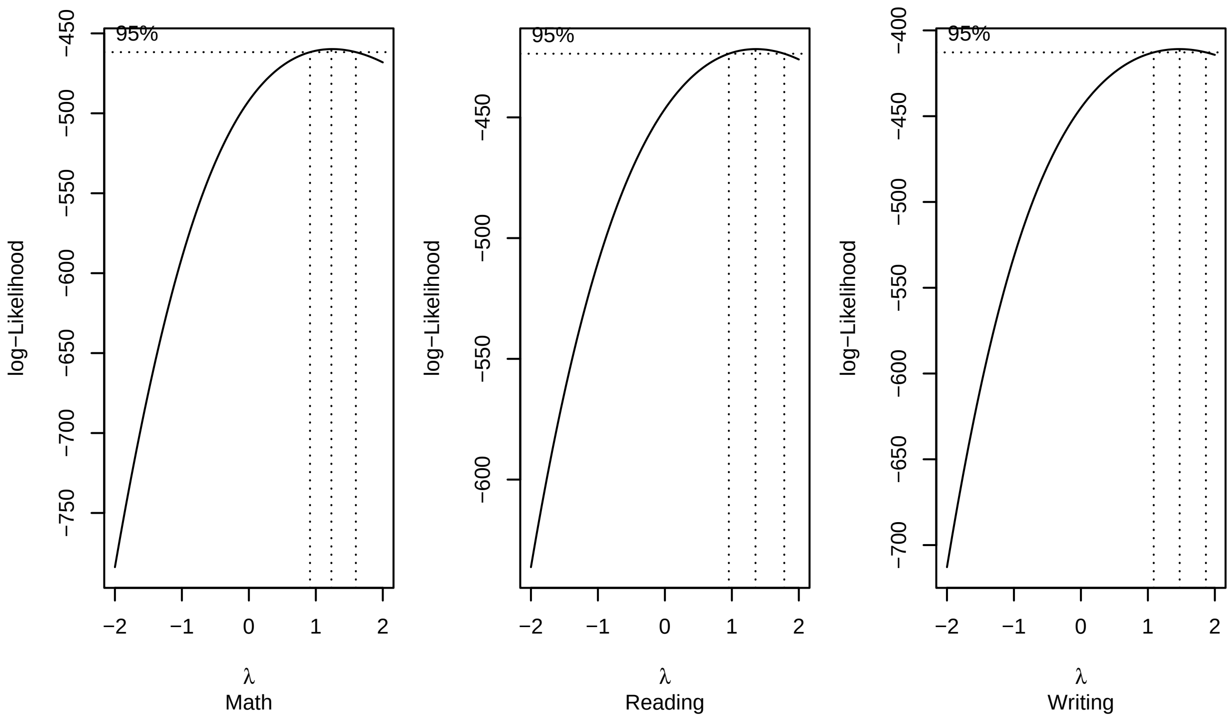

Figure 8: Cook's Distance

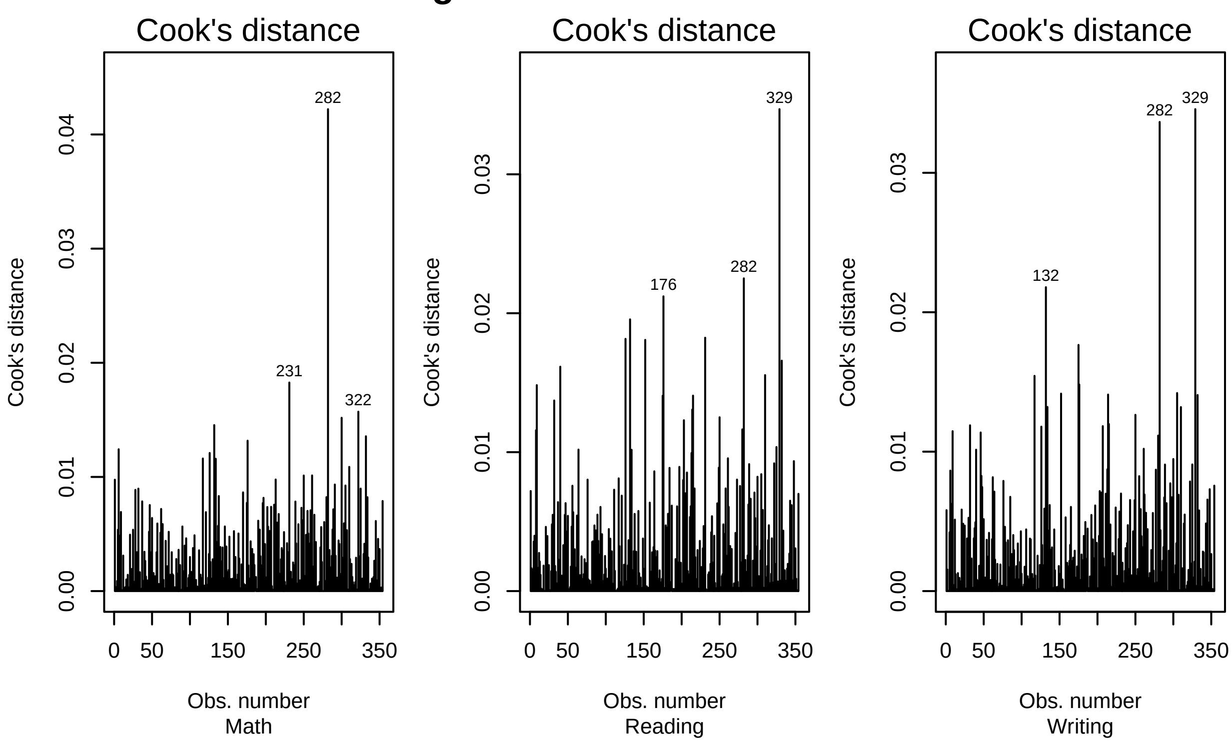

Figure 9: Mean CV Error vs. Lambda for Math

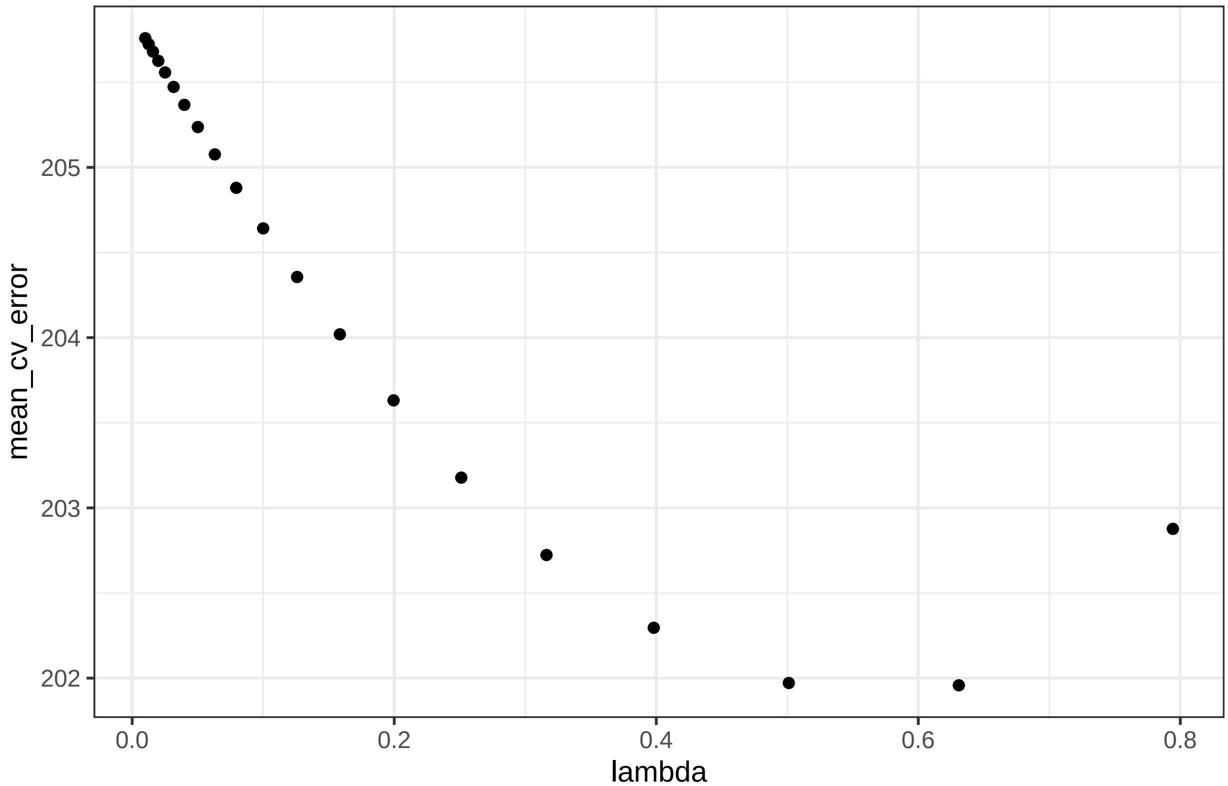

Figure 10: Mean CV Error vs. Lambda for Reading

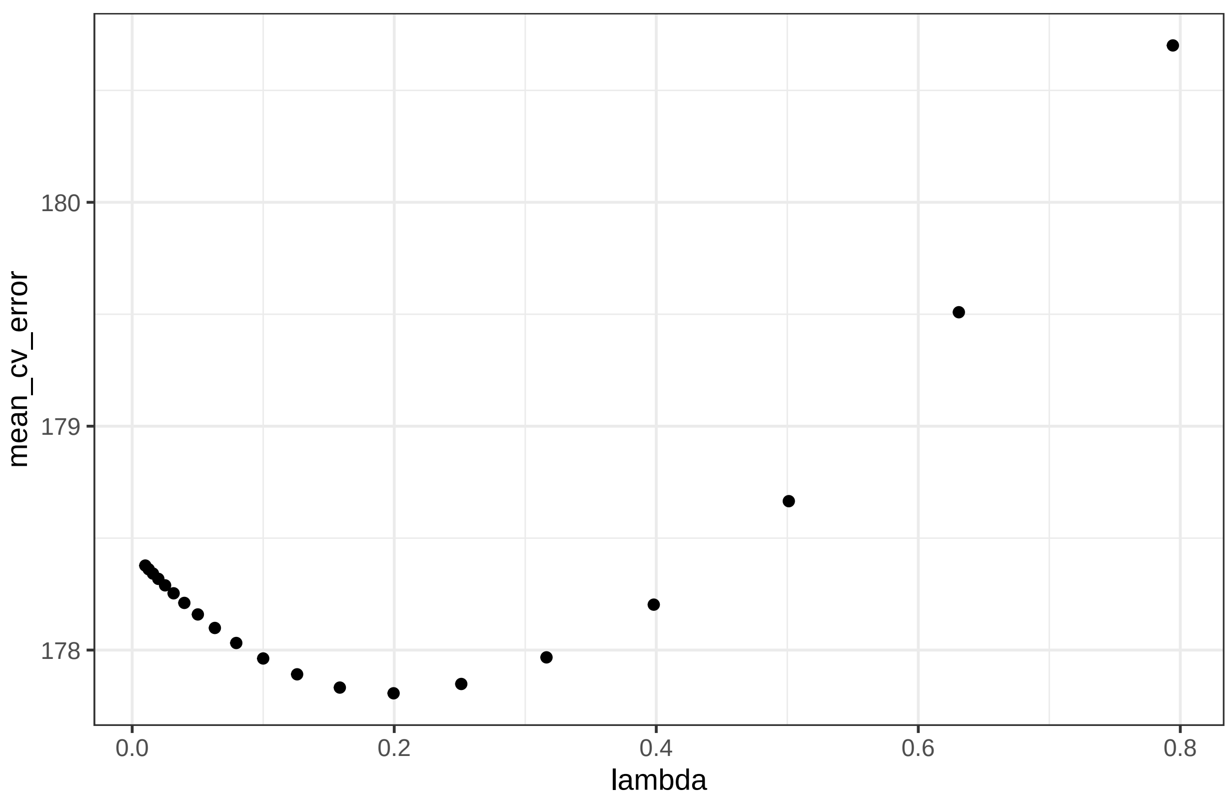

Figure 11: Mean CV Error vs. Lambda for Writing

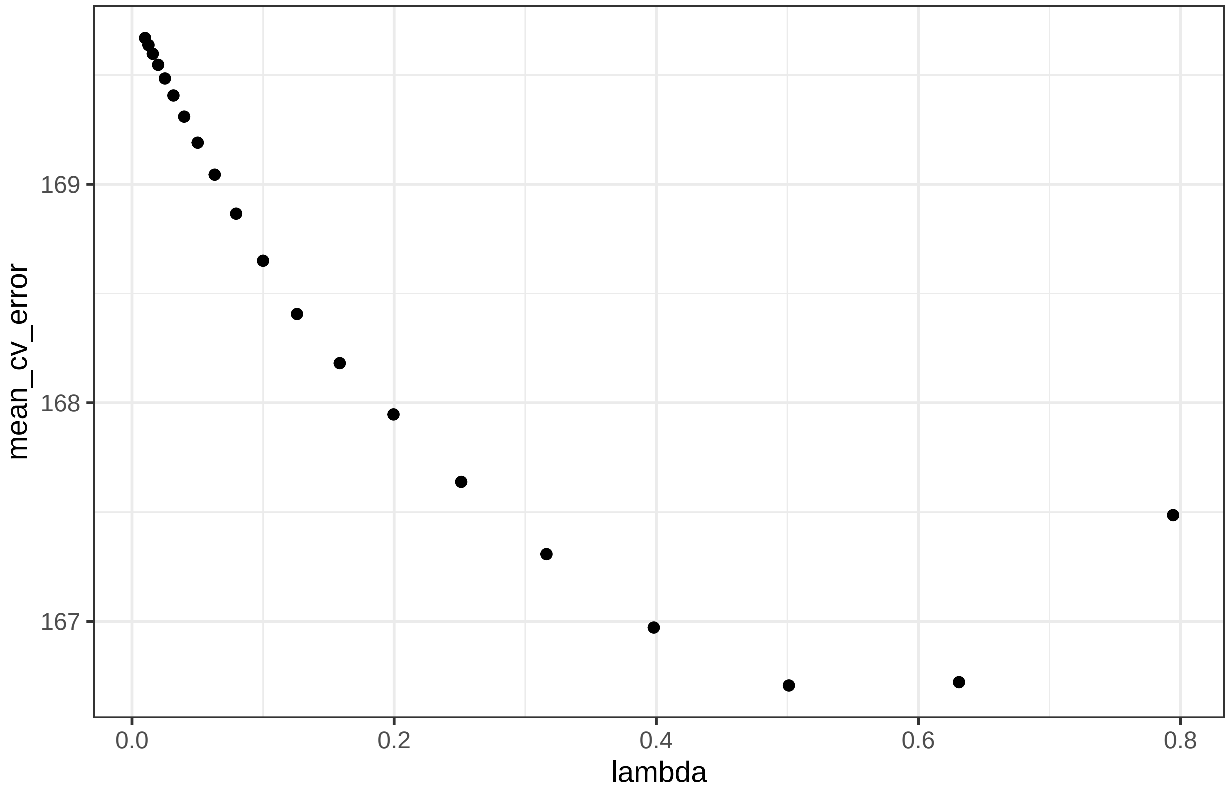

Figure 12: Lasso Variables Contraction for Math

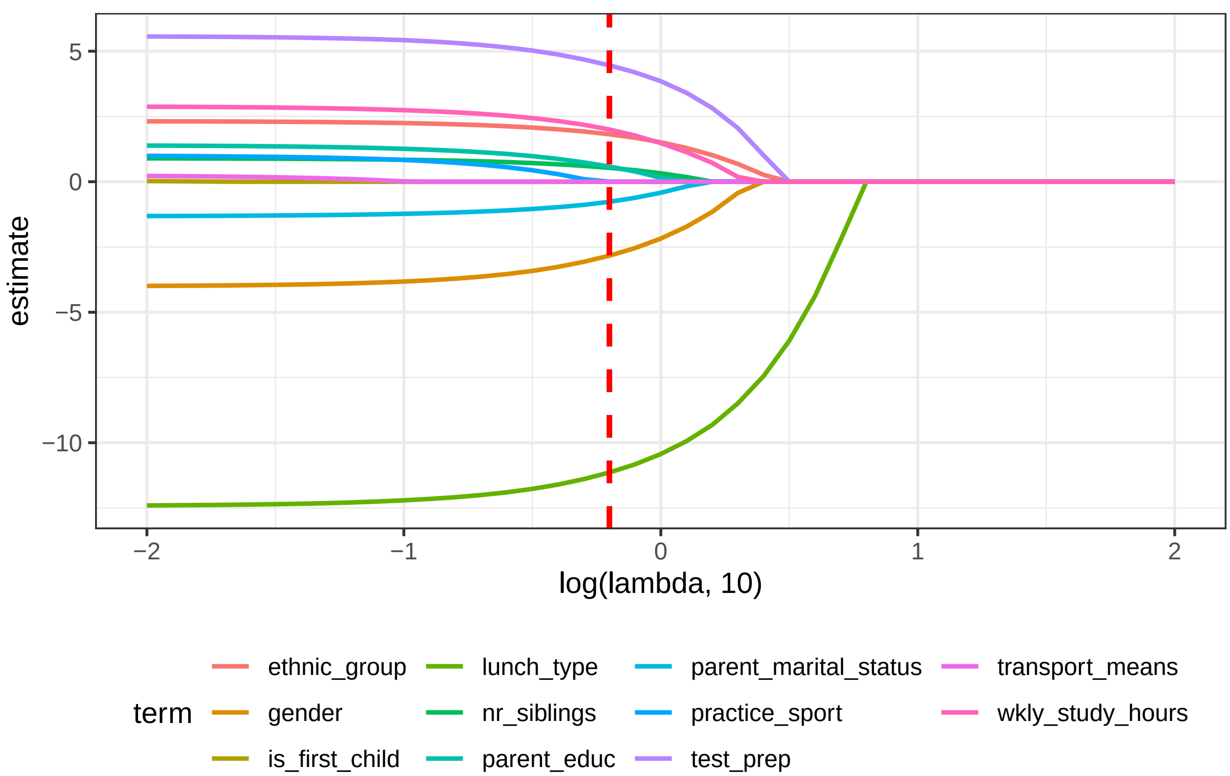

Figure 13: Lasso Variables Contraction for Reading

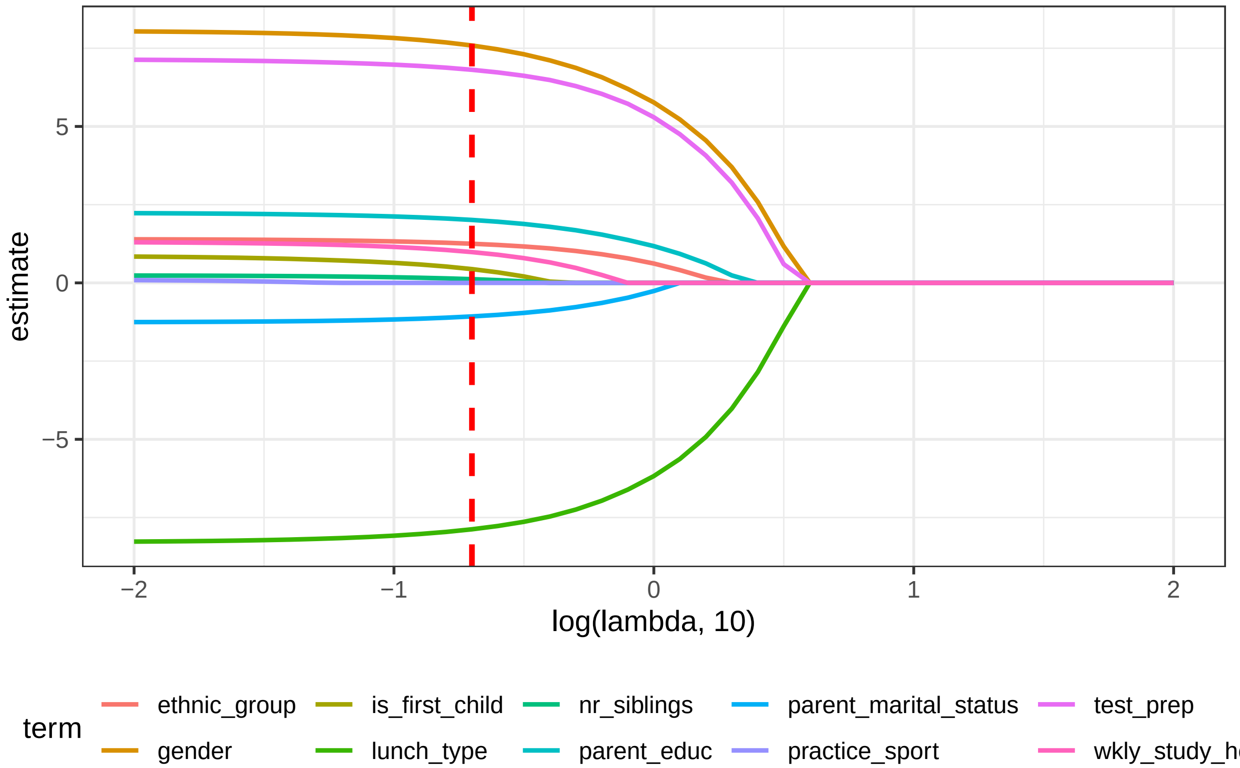

Figure 14: Lasso Variables Contraction for writing

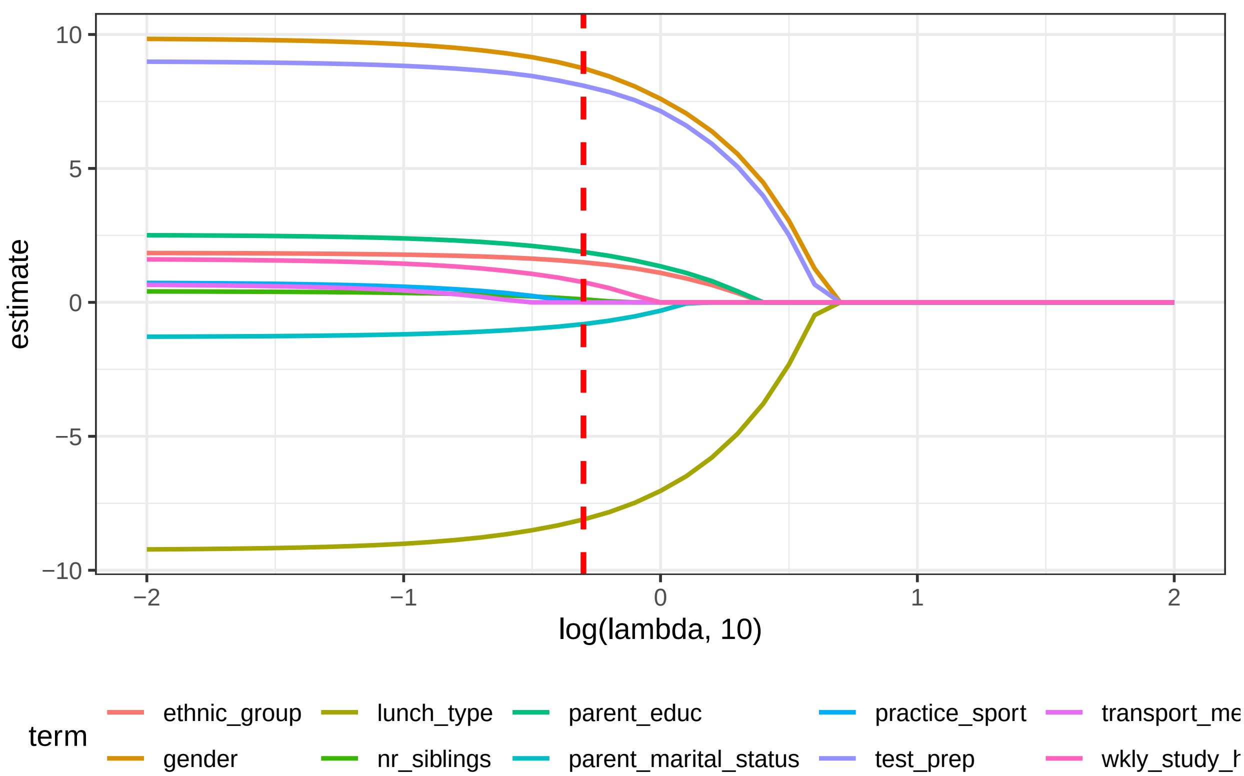

Figure 15: Correlation Plot among Score Variables