P8130 Final Report (Project 1)

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

2023-12-16

not exceed 5 pages

Abstract

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

Introduction

This study aims to predict student performance in math, reading, and writing using regression models. The study aims to identify the factors that influence academic success by exploring 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. Additionally, the study aims to identify potential correlations between scores in different subjects. The study aims to provide educators and policymakers with practical insights to customize interventions, improve educational programs, and establish

strong support structures that promote comprehensive academic progress among students by combining these analyses.

Methods

This dataset contains information about students in a public school, including three test scores and various personal and socioeconomic factors. To aid analysis, categorical data has been converted into numerical representations based on their ordial order. To handle missing cells, the mean for each column is computed, and missing cells are replaced with the mean for that column in order to keep other variables still available to use in the regression model. This process ensures a more complete and representative dataset for further analysis.

After processing the data, we constructed a comprehensive descriptive table (**Table 1**) that summarizes key statistical measures of this dataset, providing a snapshot of central tendencies and variability. Then the distribution of the three response variables (test scores) is presented in **Figure 1**, indicating an approximate normal distribution.

- in diagnostics, all linear
- no transformation needed for all three
- no ouliters showed in plot, but for influential points
- multicollinearity checked. no correlated variables, VIF all small
- two models created for each response. one backward (better compared to forward), one lasso

Results

- final model for each test score
- analyze between models

Conclusions/Discussion

Contribution

Xiaoting Tang: Method

Yifei Liu: Result Display

Longyu Zhang: Interpretation

Huanyu Chen: Writing

Appendix

Table

```
## Rows: 948 Columns: 14
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMarita...
## dbl  (4): NrSiblings, MathScore, ReadingScore, WritingScore
```

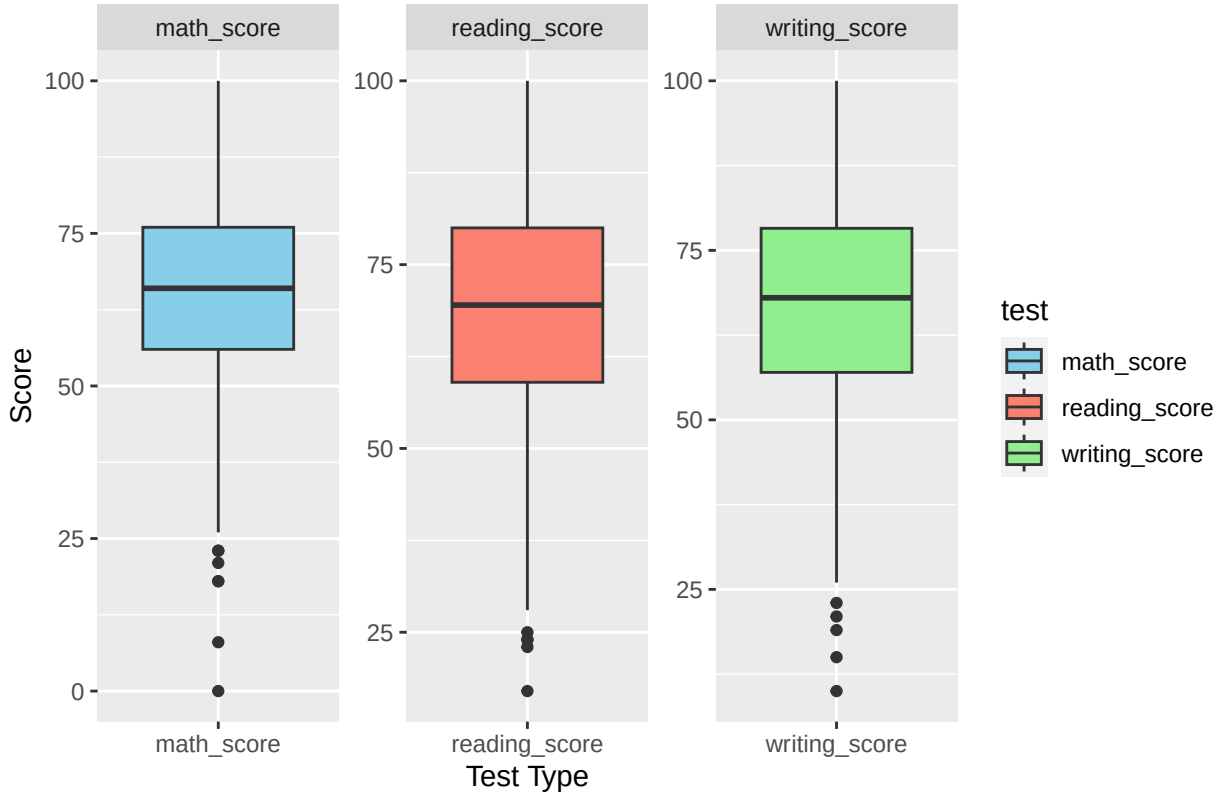
```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

## [1] "\n# Deal with NA -- Calculate the column mean (round to integer) and plug it int
```

Table 1: Categorical Variables pre-analysis

Variable	Missing	Complete Rate	Unique	Top Counts
gender	0	1.0	2	1: 488, 0: 460
ethnic_group	59	0.9	5	2: 277, 3: 237, 1: 171, 4: 124
parent_educ	392	0.6	4	1: 199, 2: 198, 3: 104, 4: 55
lunch_type	0	1.0	2	0: 617, 1: 331
test_prep	55	0.9	2	0: 571, 1: 322
parent_marital_status	49	0.9	4	0: 516, 1: 213, 3: 146, 2: 24
practice_sport	16	1.0	3	1: 477, 2: 343, 0: 112
is_first_child	30	1.0	2	1: 604, 0: 314
nr_siblings	46	1.0	8	1: 245, 2: 213, 3: 198, 0: 101
transport_means	102	0.9	2	0: 509, 1: 337
wkly_study_hours	37	1.0	3	1: 508, 0: 253, 2: 150

Figure

Figure 1: Scores Distribution (Boxplot)



Analytical questions (you may only answer some of them but properly addressing more will improve the rate of your report): 1. Using variables 1-11 as the covariates to predict Math, Reading and Writing scores. 2. Which factors (features) affect each test score signficantly? Are there interacting effects? 3. Are the optimal prediction models similar or different across the three test scores? Is it possible to leverage one score as the auxiliary information to learn the model for another score (still its model against variables 1-11) better?

describe your final model and interpret its parameters. examine the marginal distributions and pairwise relationships between variables (e.g., nonlinearities) explore several candidate models, and explain why you selected your model. be clear about your motivation for carrying out certain analyses. Your report should include a table summarizing parameter

estimates associated with your final fitted model, characterizing predictor variables. • Data exploration: descriptive statistics and visualization. • In your regression model, be watchful for variables that are highly correlated and be selective in the variables you will include in your analysis. selective interactions between variables.

Grading: Method and interpretation: Understand the background and task of the project; Preprocess of the data; Proper choice and decision on models' assumption, specification, diagnostic, selection, comparison, and validation; Strategies and reasons of your designs and decisions should be included. Correct and informative interpretation of the results.

Writing and result display: Well-organized report with professional writing and structuring; Proper introduction of the method with helpful and informative explanations; Display the data and analysis results with tables and plots. Be concise and avoid displaying redundant or useless results; Do not exceed the page limit.