Final Project Paper

Ekaterina Hofrenning, Lauren Lazaro, Arthur Starodynov, Yuki Low

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

## Introduction (brief context and background of the problem)

It is well-documented that educational success is influenced by a multitude of factors that extend beyond one’s academic preparation. Our study aims to predict math, reading and writing scores based on personal and socio-economic variables. The dataset used for our study provides us with a diverse range of variables and gives us the opportunity to uncover patterns that influence a student’s test outcome.

## Methods (data description and statistical methods)

The dataset is a sample of 948 individuals from a public school with math, reading, and writing test scores (ranging from 0-100) and the following variables of interest: gender, ethnic group, parental education, lunch type, test prep, parent’s marital status, sports participation, first child, number of siblings, means of transportation to school, number of weekly study hours. Due to significant missing data, we imputed the missing values with the average or most common value per variable.

Math, reading and writing tests scores were the response variables and were scored from 0 to 100. Math test scores ranged from 0 to 100 with a mean score of 65.9821 and a median score of 66. Reading test scores ranged from 17 to 100 with a mean score of 68.8418 and a median score of 69.5. Writing test scores ranged from 10 to 100 with a mean score of 67.9293 and a median score of 68. The histograms of the distributions of math, reading and writing tests scores are slightly skewed to the left. In order to make the distributions more normal, we attempted both logarithmic transformations and square root transformations in which both of these types of transformations resulted in more severely skewed distributions. In the end, we decided to proceed with using no transformation on the test score variables.

We conducted a data description (Table 1), computing the averages and standard deviations for continuous variables and counts and percentages for categorical variables. Our sample is well-balanced across sex. The sample is composed of majority first born children, children on standard fee lunch, children with married parents, school bus riders, and children who study 5-10 hours. The average number of siblings is 2.

| **Characteristic** | **N = 948**1 |
| --- | --- |
| Gender |  |
| female | 488 / 948 (51%) |
| male | 460 / 948 (49%) |
| EthnicGroup |  |
| group A | 80 / 948 (8.4%) |
| group B | 171 / 948 (18%) |
| group C | 336 / 948 (35%) |
| group D | 237 / 948 (25%) |
| group E | 124 / 948 (13%) |
| ParentEduc |  |
| some high school | 163 / 948 (17%) |
| high school | 176 / 948 (19%) |
| associate's degree | 198 / 948 (21%) |
| some college | 252 / 948 (27%) |
| bachelor's degree | 104 / 948 (11%) |
| master's degree | 55 / 948 (5.8%) |
| LunchType |  |
| free/reduced | 331 / 948 (35%) |
| standard | 617 / 948 (65%) |
| TestPrep |  |
| completed | 322 / 948 (34%) |
| none | 626 / 948 (66%) |
| ParentMaritalStatus |  |
| divorced | 146 / 948 (15%) |
| married | 565 / 948 (60%) |
| single | 213 / 948 (22%) |
| widowed | 24 / 948 (2.5%) |
| PracticeSport |  |
| never | 112 / 948 (12%) |
| sometimes | 493 / 948 (52%) |
| regularly | 343 / 948 (36%) |
| IsFirstChild | 634 / 948 (67%) |
| NrSiblings | 2 (1) |
| TransportMeans |  |
| private | 337 / 948 (36%) |
| school\_bus | 611 / 948 (64%) |
| WklyStudyHours |  |
| < 5 | 253 / 948 (27%) |
| 5-10 | 545 / 948 (57%) |
| > 10 | 150 / 948 (16%) |
| MathScore | 66 (16) |
| ReadingScore | 69 (15) |
| WritingScore | 68 (15) |
| 1n / N (%); Mean (SD) | |

After looking at the histograms of the continuous outcomes, we decided that the variables looked fairly normal so we did not use log transformations.

We used a few different methods for model building. First, we decided to use backwards, forwards and stepwise regression. Then, we split our data into training and validation sets for each student (math, reading, writing). We used Lasso regression to select important variables and their coefficients. We also proceeded with building Random Forest models to predict the scores for each student. Our model was then validated using the validation sets.

The chosen model was then tested through model diagnosis and influence diagnosis to make sure all assumptions are met and no significant outliers were influencing the model.

## Results

In order to predict test score from our variables of interest, we decided to use backwards, forwards, and stepwise regression. Each of the three methods selected the same results for each outcome. For math score as the outcome, the optimal predictors of Gender + EthnicGroup + ParentEduc + LunchType + TestPrep + ParentMaritalStatus + PracticeSport + IsFirstChild + WklyStudyHours. For Reading score as the outcome, the optimal predictors are Gender + EthnicGroup + ParentEduc + LunchType + TestPrep + ParentMaritalStatus + IsFirstChild + WklyStudyHours. For writing score as the outcome, the optimal predictors are Gender + EthnicGroup + ParentEduc + LunchType + TestPrep + ParentMaritalStatus + PracticeSport + IsFirstChild + WklyStudyHours.

Table 2. Math score backwards stepwise model

| term | estimate | p.value |
| --- | --- | --- |
| (Intercept) | 55.976 | <.001 |
| Gendermale | 4.981 | <.001 |
| EthnicGroup.L | 7.401 | <.001 |
| EthnicGroup.Q | 2.940 | 0.011 |
| EthnicGroup.C | 0.737 | 0.482 |
| EthnicGroup^4 | -0.985 | 0.251 |
| ParentEduc.L | 6.642 | <.001 |
| ParentEduc.Q | -0.045 | 0.972 |
| ParentEduc.C | -0.300 | 0.806 |
| ParentEduc^4 | 1.402 | 0.213 |
| ParentEduc^5 | -2.644 | 0.006 |
| LunchTypestandard | 11.155 | <.001 |
| TestPrepnone | -5.582 | <.001 |
| ParentMaritalStatusmarried | 3.876 | 0.002 |
| ParentMaritalStatussingle | 1.102 | 0.444 |
| ParentMaritalStatuswidowed | 5.066 | 0.087 |
| PracticeSport.L | 2.514 | 0.015 |
| PracticeSport.Q | -0.518 | 0.503 |
| IsFirstChildyes | 2.368 | 0.011 |
| WklyStudyHours.L | 2.621 | 0.008 |
| WklyStudyHours.Q | -1.032 | 0.158 |

Table 3. Reading score backwards stepwise model

| term | estimate | p.value |
| --- | --- | --- |
| (Intercept) | 68.495 | <.001 |
| Gendermale | -7.282 | <.001 |
| EthnicGroup.L | 4.149 | 0.001 |
| EthnicGroup.Q | 1.438 | 0.203 |
| EthnicGroup.C | -0.471 | 0.645 |
| EthnicGroup^4 | -0.942 | 0.261 |
| ParentEduc.L | 7.638 | <.001 |
| ParentEduc.Q | 1.535 | 0.22 |
| ParentEduc.C | 0.566 | 0.635 |
| ParentEduc^4 | 1.547 | 0.158 |
| ParentEduc^5 | -3.033 | 0.001 |
| LunchTypestandard | 7.494 | <.001 |
| TestPrepnone | -6.972 | <.001 |
| ParentMaritalStatusmarried | 4.113 | 0.001 |
| ParentMaritalStatussingle | 1.275 | 0.363 |
| ParentMaritalStatuswidowed | 4.645 | 0.106 |
| IsFirstChildyes | 2.446 | 0.007 |
| WklyStudyHours.L | 1.431 | 0.135 |
| WklyStudyHours.Q | -0.933 | 0.191 |

Table 4. Writing score backwards stepwise model

| term | estimate | p.value |
| --- | --- | --- |
| (Intercept) | 69.347 | <.001 |
| Gendermale | -9.209 | <.001 |
| EthnicGroup.L | 4.673 | <.001 |
| EthnicGroup.Q | 0.627 | 0.569 |
| EthnicGroup.C | -1.891 | 0.058 |
| EthnicGroup^4 | -1.649 | 0.043 |
| ParentEduc.L | 9.983 | <.001 |
| ParentEduc.Q | 1.365 | 0.265 |
| ParentEduc.C | 0.289 | 0.803 |
| ParentEduc^4 | 1.715 | 0.109 |
| ParentEduc^5 | -3.106 | 0.001 |
| LunchTypestandard | 8.388 | <.001 |
| TestPrepnone | -9.629 | <.001 |
| ParentMaritalStatusmarried | 4.135 | 0.001 |
| ParentMaritalStatussingle | 1.056 | 0.44 |
| ParentMaritalStatuswidowed | 3.950 | 0.16 |
| PracticeSport.L | 2.251 | 0.022 |
| PracticeSport.Q | -0.708 | 0.335 |
| IsFirstChildyes | 2.208 | 0.012 |
| WklyStudyHours.L | 1.338 | 0.152 |
| WklyStudyHours.Q | -0.960 | 0.167 |

## Conclusions/Discussion

## A brief summary on each group member’s contribution (method, data analysis, writing, etc).