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 employed a multi-faceted approach to model building. Initially, we fit simple linear regression (SLR) models for each of the three scores, utilizing all available covariates in our dataset. Subsequently, we conducted a comprehensive exploration by implementing backward, forward, and stepwise regression techniques using p-values, AIC, and BIC as the criterion of interest. We also fit models using the Adjusted R^2 criterion-based approach. This allowed us to assess how variables varied when adjusting for all covariates through multiple linear regression modeling. The combination of these methods provided a nuanced understanding of the relationships between predictors and scores, capturing both individual and collective effects. The optimal models were determined by maximizing the Adjusted R^2, which penalizes the measure as the number of predictors increase. This helps to ensure we choose a better tradeoff between bias and variance.

The chosen models were then tested through model diagnostics and influential observation diagnostics to make sure all linear assumptions are met and no significant outliers were influencing the model.

## Results

From the SLR model, we found that the linear association between MathScore and the predictors LunchType, Gender, TestPrep and WklyStudyHours were significant. The average increase in MathScore was 8.7216 (p=3e-09) for students of ethnic group B and 3.4986 (p=0.00759) for students of ethnic group C compared to students of ethnic group A without adjusting for all other covariates. The average increase in MathScore between students with parents with some high school education and students with parents who completed their high school education was 6.0952 (p=0.000139) without adjusting for all other covariates. The average decrease in MathScore between students with parents with some high school education and students with parents who have their Master’s degree was 2.8793 (p=0.009353) without adjusting for all other covariates. The average increase in MathScore between students with divorced parents and students with married parents was 4.142 (p=0.040) without adjusting for all other covariates. The average decrease in MathScore was 3.6763 (p=0.041) for students with 1 sibling and 3.8757 (p=0.0390) for students with 2 siblings compared to only children without adjusting for all other covariates. Note that the linear association between MathScore and PracticeSport, IsFirstChild, and TransportMeans were not significant.

We fit several Multivariate Linear Regressions (MLR) using a variety of predictor selection techniques. First, we fit the MLR models using backwards, forwards, and stepwise regression with p-values, AIC, and BIC as the criterion of interest. Next, we fit the MLR models using the Adjusted R^2 criterion method, which penalizes the R^2 as the number of predictors increases. In order to determine the best models of best fit, we chose the models with the highest Adjusted R^2 value. For Math Score as the outcome, the optimal model contained Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus, PracticeSport, IsFirstChild, and WklyStudyHours as predictors. For Reading Score as the outcome, the optimal model contained Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus, PracticeSport, IsFirstChild, NrSiblings, TransportMeans, and WklyStudyHours as predictors. For Writing score as the outcome, the optimal model contained Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus, PracticeSport, IsFirstChild, and WklyStudyHours as predictors.

Model diagnosis is an essential step in the modeling process. The models need to be checked through visual and statistical tests to make sure all assumptions are met. Diagnostics of the final models looked good, and we determined that the residual plots looked normal enough to safely consider our results valid. In addition to model diagnostics, we looked for the presence of influential observations using Cook’s distance. We determined that there was no presence of worrisome influential observations in the three final models.

Our estimates for coefficients for each of our significant covariates associated with Math Score, Reading Score, and Writing Score are located in our Appendix in Table 2, 3 and 4.

\

\

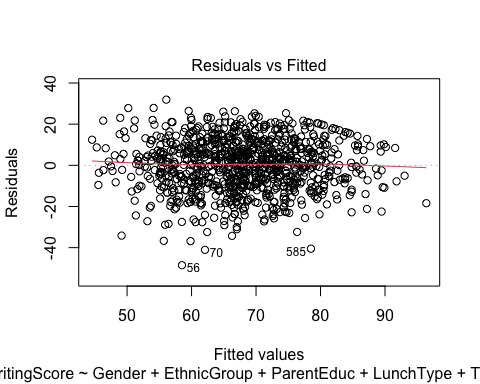
\

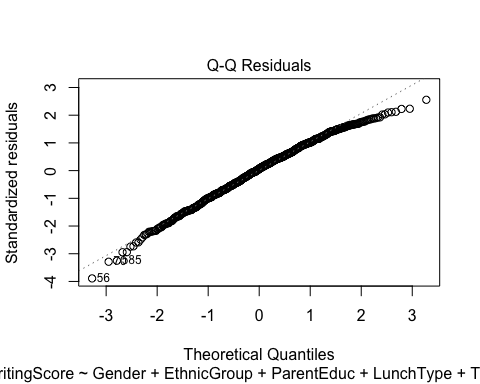
\

\

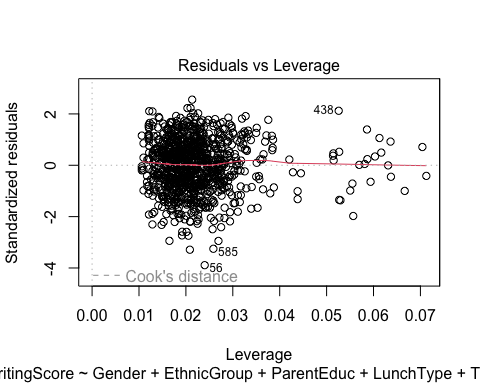
\

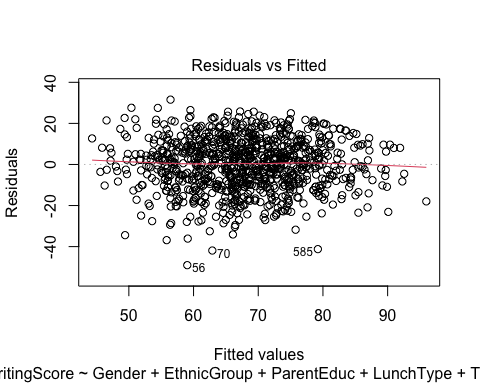
\

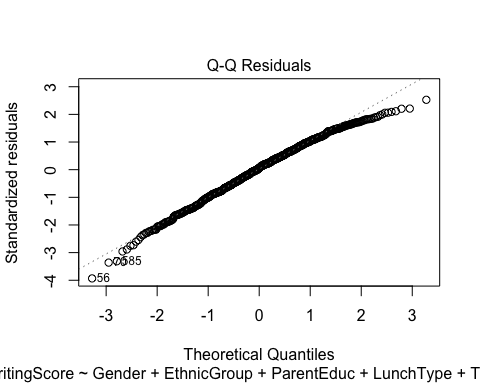


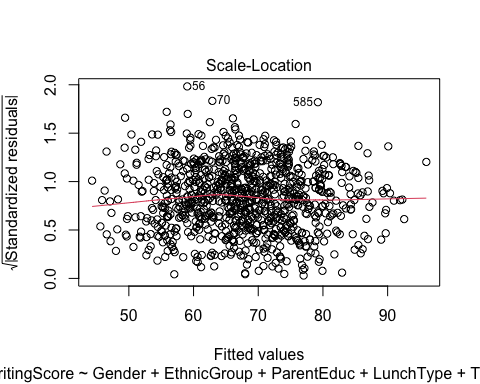


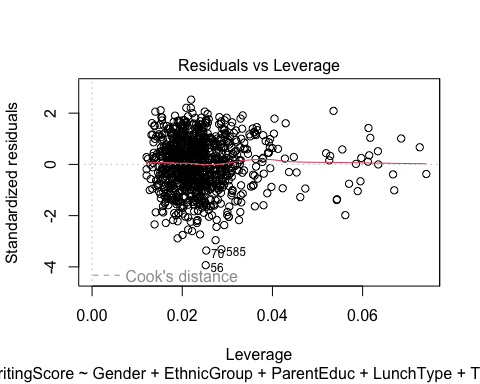


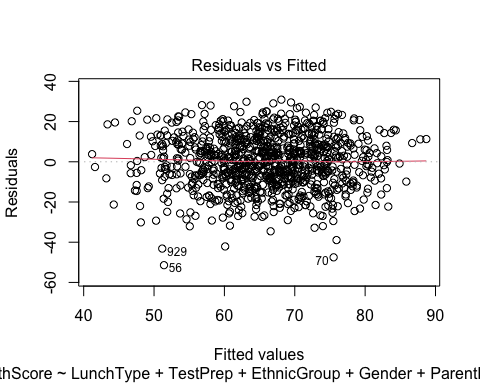


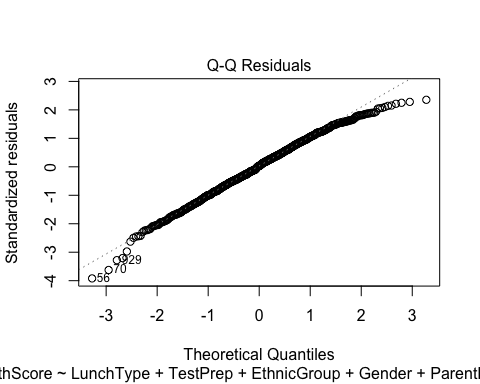


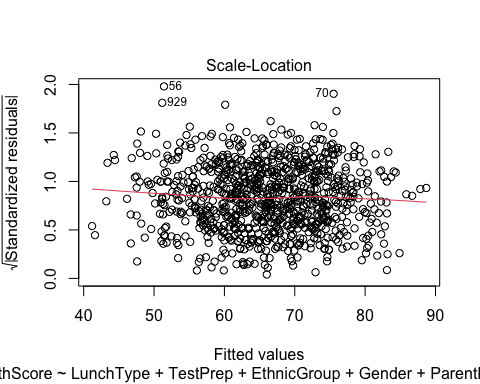


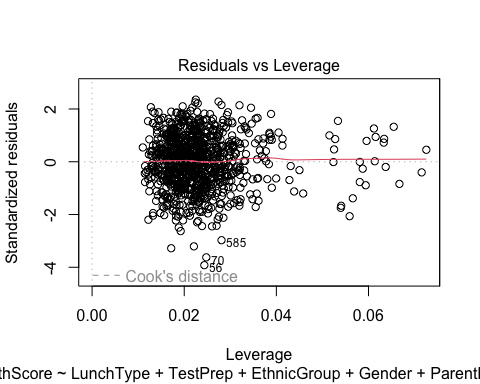












## Conclusions/Discussion

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

## Appendix

writing\_both %>%  
 broom::tidy() %>%  
 dplyr::select(term, estimate, p.value) %>%  
 mutate(estimate = round(estimate, 3),  
 p.value = round(p.value, 3),  
 p.value = case\_when(p.value < .001 ~ "<.001", p.value >= .001 ~ as.character(p.value))) %>%  
 kableExtra::kable(caption = "Table 2. Writing Score Final Model")

Table 2. Writing Score Final 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 |

reading\_adjR2 %>%  
 broom::tidy() %>%  
 dplyr::select(term, estimate, p.value) %>%  
 mutate(estimate = round(estimate, 3),  
 p.value = round(p.value, 3),  
 p.value = case\_when(p.value < .001 ~ "<.001", p.value >= .001 ~ as.character(p.value))) %>%  
 kableExtra::kable(caption = "Table 3. Reading Score Final Model")

Table 3. Reading Score Final Model

| term | estimate | p.value |
| --- | --- | --- |
| (Intercept) | 68.356 | <.001 |
| Gendermale | -9.191 | <.001 |
| EthnicGroup.L | 4.729 | <.001 |
| EthnicGroup.Q | 0.560 | 0.612 |
| EthnicGroup.C | -1.873 | 0.061 |
| EthnicGroup^4 | -1.654 | 0.043 |
| ParentEduc.L | 10.007 | <.001 |
| ParentEduc.Q | 1.442 | 0.24 |
| ParentEduc.C | 0.309 | 0.79 |
| ParentEduc^4 | 1.678 | 0.117 |
| ParentEduc^5 | -3.073 | 0.001 |
| LunchTypestandard | 8.409 | <.001 |
| TestPrepnone | -9.649 | <.001 |
| ParentMaritalStatusmarried | 4.102 | 0.001 |
| ParentMaritalStatussingle | 1.055 | 0.441 |
| ParentMaritalStatuswidowed | 3.675 | 0.192 |
| PracticeSport.L | 2.255 | 0.022 |
| PracticeSport.Q | -0.685 | 0.351 |
| IsFirstChildyes | 2.283 | 0.01 |
| NrSiblings | 0.196 | 0.492 |
| TransportMeansschool\_bus | 0.840 | 0.331 |
| WklyStudyHours.L | 1.300 | 0.165 |
| WklyStudyHours.Q | -0.972 | 0.162 |

math\_pval\_for %>%  
 broom::tidy() %>%  
 dplyr::select(term, estimate, p.value) %>%  
 mutate(estimate = round(estimate, 3),  
 p.value = round(p.value, 3),  
 p.value = case\_when(p.value < .001 ~ "<.001", p.value >= .001 ~ as.character(p.value))) %>%  
 kableExtra::kable(caption = "Table 4. Math Score Final Model")

Table 4. Math Score Final Model

| term | estimate | p.value |
| --- | --- | --- |
| (Intercept) | 55.167 | <.001 |
| LunchTypestandard | 11.196 | <.001 |
| TestPrepnone | -5.577 | <.001 |
| EthnicGroup.L | 7.416 | <.001 |
| EthnicGroup.Q | 2.901 | 0.013 |
| EthnicGroup.C | 0.754 | 0.472 |
| EthnicGroup^4 | -1.020 | 0.235 |
| Gendermale | 5.009 | <.001 |
| ParentEduc.L | 6.676 | <.001 |
| ParentEduc.Q | 0.023 | 0.986 |
| ParentEduc.C | -0.333 | 0.785 |
| ParentEduc^4 | 1.413 | 0.209 |
| ParentEduc^5 | -2.700 | 0.005 |
| WklyStudyHours.L | 2.540 | 0.01 |
| WklyStudyHours.Q | -1.048 | 0.152 |
| ParentMaritalStatusmarried | 3.842 | 0.002 |
| ParentMaritalStatussingle | 1.063 | 0.46 |
| ParentMaritalStatuswidowed | 4.858 | 0.101 |
| IsFirstChildyes | 2.509 | 0.007 |
| PracticeSport.L | 2.520 | 0.015 |
| PracticeSport.Q | -0.503 | 0.515 |
| NrSiblings | 0.335 | 0.263 |