Multiple Linear Regression Modeling of Test Scores using Socioeconomical Factors

Miao Fu(mf3593)

Introduction

Education is an essential component of contemporary society, and the assessment of student performance through test scores is commonly seen. This research project focuses on studying the complex relationship between various socioeconomic factors and the test scores of public school students. Key variables, including ethnic group, parents' marital status, and weekly study hours, are considered as potential predictive variables of test score outcomes. Through statistical modeling, we aim to identify significant contributing factors and construct predictive models, providing a thorough understanding of the dynamics between socioeconomic factors and test performance.

The chosen 11 covariates represent a spectrum of influences within students' socioenvironmental contexts and are presumed to be potentially relevant to one's academic performance. By combining all these factors into predictive models, our research aims to select out the variables that significantly shape test performance and recommend a final optimal model for each of three test scores(maths, writing, reading). Additionally, model comparisons will be conducted to evaluate the associations between covariates and test scores in a detailed and comprehensive way. This study contributes to the broader understanding of educational equity and may inform potential interventions to create a better education environment at

public schools.

Methods

The dataset used contained 587 observations and 14 variables.

Data exploration was first done by obtaining descriptive summary statistics for all variables. Percentage and count of each category was obtained for categorical variables(Figure 1). Mean, median, standard deviation, minimum, maximum, q1 and q3 were obtained for numeric variables(Figure 2). The normality of distribution of outcome variables were investigated by plotting the data on histogram, boxplot, and QQ plots. Normality of distribution of all variables were studied by histogram(Figure 3). From the distribution bar graphs, it was decided that no transformation is needed for any predictor variables. Admittedly, the distributions might not be good references for transformation because the variables are all categorical. Since the distribution of the three scores showed slightly left-tailed, three types of transformations were tested: 1) Natural logarithm 2) Square Root 3) Inverse. The resulting plots are plotted in histograms(Figure 4). There is no apparent improvement on the distribution of the outcome through the three transformations. Thus, the original outcome data were chosen to be used in following statistical modeling steps.

By plotting our the pairwise correlation between variables, there is apparent linearity among the three scores(Figure 5). Other correlation coefficients are relatively small, indicating weak linear relationship between the variables.

variable	category	count	percent
gender	female	315	53.662692
gender	male	272	46.337308
ethnic_group	group A	50	8.517888
ethnic_group	group B	123	20.954003
ethnic_group	group C	174	29.642249
ethnic_group	group D	155	26.405451
ethnic_group	group E	85	14.480409
parent_educ	associate's degree	128	21.805792
parent_educ	bachelor's degree	71	12.095400
parent_educ	high school	122	20.783646
parent_educ	master's degree	39	6.643952
parent_educ	some college	116	19.761499
parent_educ	some high school	111	18.909710
lunch_type	free/reduced	206	35.093697
lunch_type	standard	381	64.906303
test_prep	completed	208	35.434412
test_prep	none	379	64.565588
parent_marital_status	divorced	92	15.672913
parent_marital_status	married	343	58.432709
parent_marital_status	single	137	23.339012
parent_marital_status	widowed	15	2.555366
practice_sport	never	68	11.584327
practice_sport	regularly	218	37.137990
practice_sport	sometimes	301	51.277683
is_first_child	no	192	32.708688
is_first_child	yes	395	67.291312
transport_means	private	229	39.011925
transport_means	school_bus	358	60.988075
wkly_study_hours	< 5	154	26.235094
wkly_study_hours	> 10	104	17.717206
wkly_study_hours	5-10	329	56.047700

Figure 1: Summary Statistics of Categorical Variables

variable	mean	median	sd	minimum	maximum	q1	q3
nr_siblings	2.139693	2	1.481712	0	7	1	3
math_score	66.676320	67	16.113744	0	100	56	78
reading_score	69.846678	70	15.166662	17	100	60	81
writing_score	68.901192	69	15.550000	10	100	58	79

Figure 2: Summary Statistics of Numeric Variables

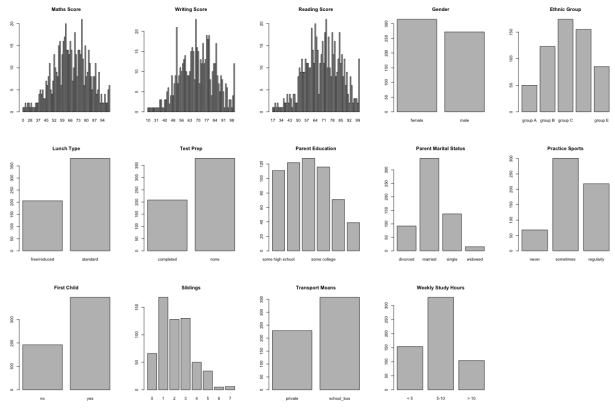

Figure 3: Histogram Distribution of All Variables

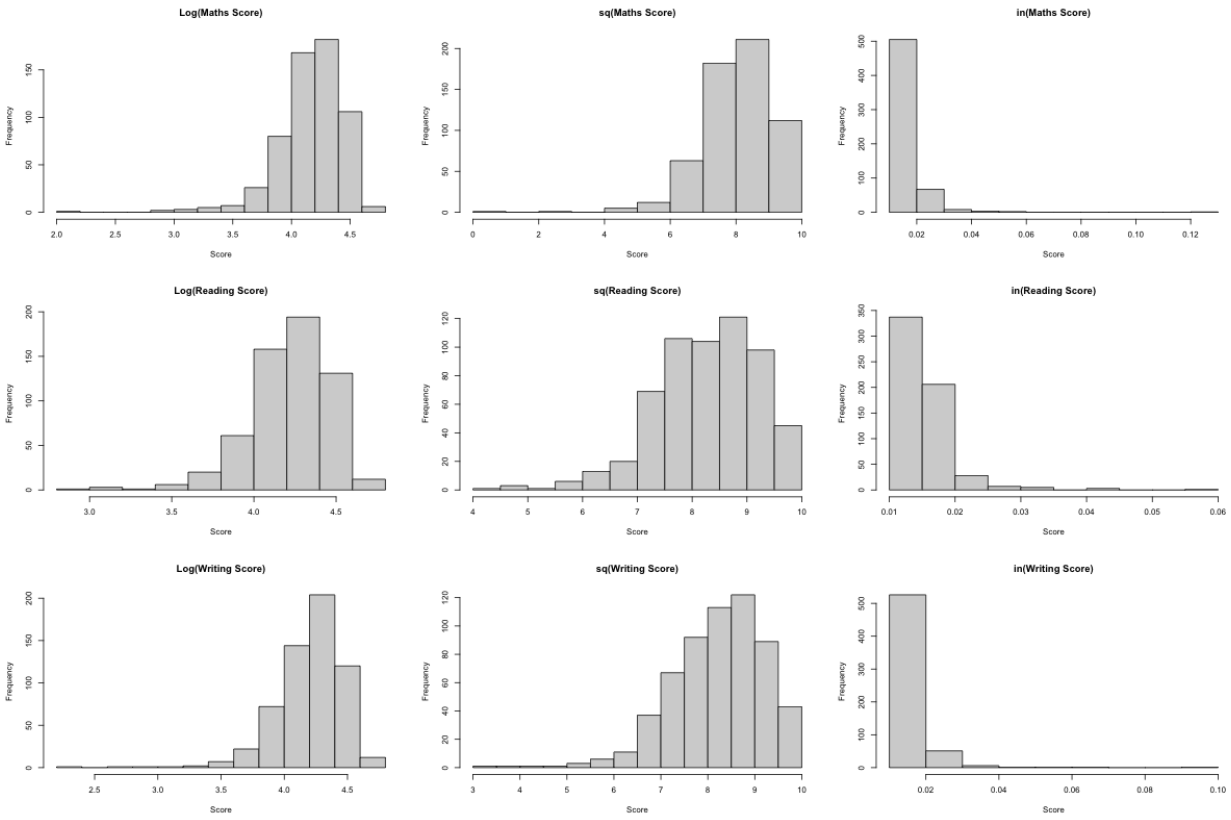

Figure 4: Log, sqrt, and inverse transformation of the outcome variables

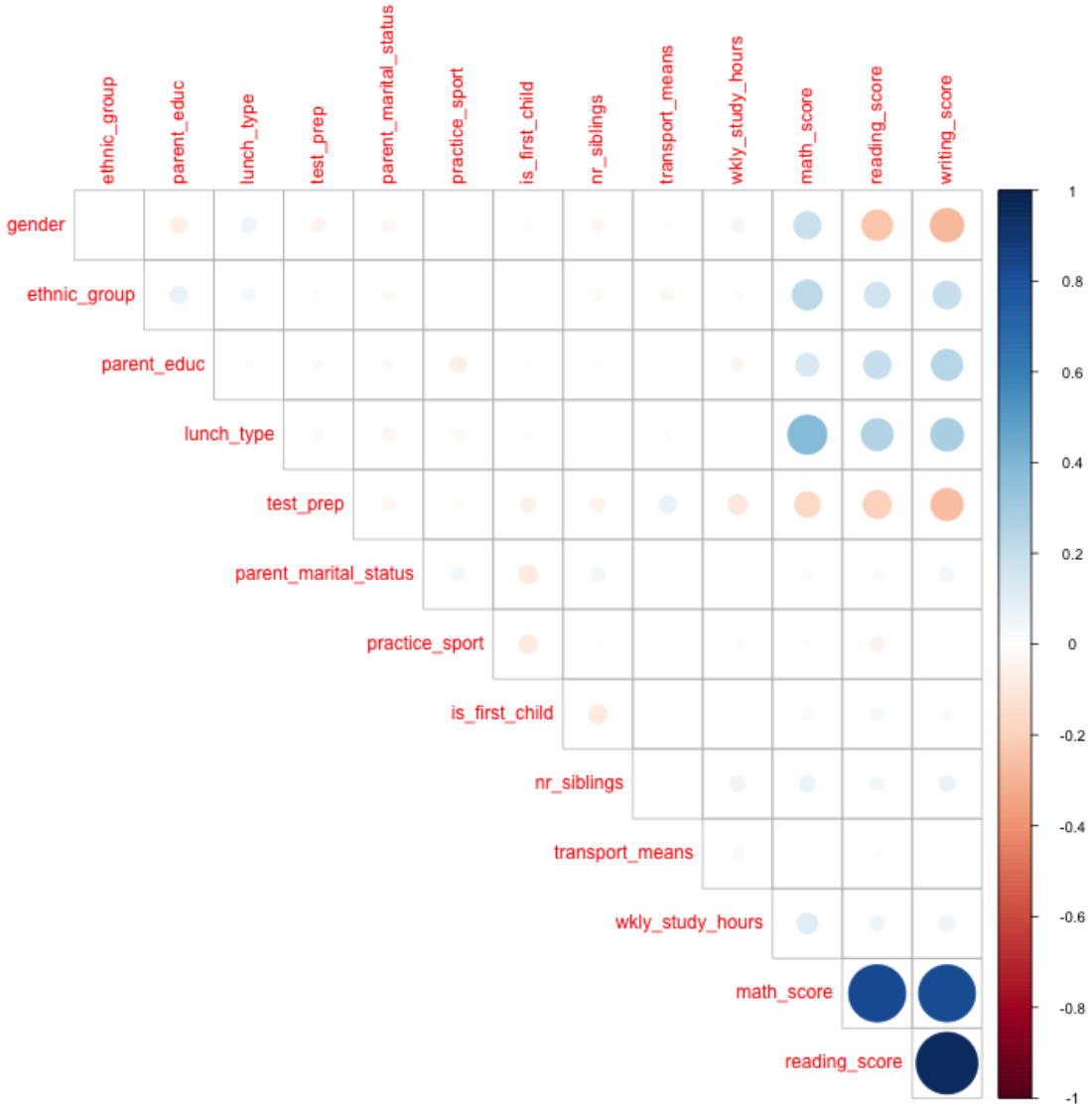

Figure 5: Correlation plot between variables

Discussion

To study whether it is possible to leverage one or two scores to improve the model of scores, a new MLR for each score was created by adding the rest two scores as predictors alongside with the original covariates obtianed through previous model selection. MSE was calculated for each model and compared (Figure 6). We can see that the MSE all significantly decreased after adding other scores to fit one score's model, indicating that leveraging other scores to enhance one score's model is possible and successful .

model_name	MSE
maths_reading+writing	31.67574
maths_original	187.71873
reading_maths+writing	16.36549
reading_original	177.64685
writing_maths+reading	12.79564
writing_original	163.35751

Figure 6: MSE of models after addition of other scores