BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

ECONOMETRICS PROJECT

TOPIC: Impact of factors on mathematics test results

Member:	Nguyễn Minh Hiếu	11222333
	Nguyễn Thanh Tùng	11226754
	Phan Đại Tùng	11226767
	Lâm Minh Quang	11225430
	Nguyễn Khánh Toàn	11226276
	Thân Quang Vinh	11226938

Class: DSEB 64B

Hanoi, May 1st, 2024

Table of Contents

Abstract

In today's world, a strong foundation in mathematics is critical for success in numerous fields. However, achieving good educational outcomes remains a significant challenge. Disparities persist in student performance on standardized math tests, highlighting the need to identify factors influencing achievement. This econometric analysis tackles this issue head-on, highlighting the complex relationship between individual characteristics and math scores.

Our study in finding about the factors that influence students' math scores, to improve educational outcomes through targeted interventions. We used multivariate regression analysis to uncover significant relationships between student characteristics and math performance.

We first explored factors like gender, race/ethnicity, parental education level, lunch program participation, and test prep course enrollment. Then, regression analysis was conducted to find the key influences: lunch type (standard vs. subsidized), participation in test prep, race/ethnicity, gender, and parental education all significantly impacted math scores.

While acknowledging limitations in data quality and assumptions, we ensured the model's reliability through careful checks. Our recommendations include ongoing monitoring, data improvement efforts, external validation of the findings, and clear communication of the results.

This study offers a valuable tool for understanding and tackling educational disparities. It allows us to predict math scores and target support for students. Furthermore, it contributes to achieving educational equity by informing policymakers and educators on effective strategies to boost student achievement in mathematics.

I. Introduction:

1. Rationale of the study

We conducted this study to understand the factors influencing math scores among students, aiming to address potential factors that are creating a large gap in education level, specifically math scores. By analyzing data from a survey of 1000 individuals, including variables such as gender, race/ethnicity, parental education, etc, we aim to identify key determinants of math performance. So with the use of Python language and Stata - a statistical software for analysis, the study seeks to provide insights for educators to design goals and allocate resources effectively, ultimately striving for a future where all students have equal opportunities to excel in mathematics.

2. Goals and Purpose

This study will contribute to the ongoing pursuit of educational equity by identifying key factors influencing math scores. The results can inform the development of targeted interventions and resource allocation strategies for improving student achievement in mathematics. Ultimately, this analysis aims to move us closer to a future where all students can excel in this critical subject.

Our motivation for this research is to understand how the factors influencing math scores can inform targeted interventions and resource allocation to close achievement gaps between student groups. This study can provide valuable insights for policymakers and educators when designing educational initiatives and allocating resources to support student learning in mathematics.

3. Research methodology

Python and Stata will be used for this study with some extension libraries. These are our additional libraries for Python: pandas, seaborn, matplotlib.pyplot

- Pandas: allows us to clean, transform, and analyze data efficiently
- Seaborn: provides a high-level interface for drawing attractive and informative statistical graphics, making it easy to visualize datasets directly from DataFrames.
- Matplotlib.pyplot: allowing us to fine-tune the appearance of our plots to meet specific requirements.

For Stata, we have used these functions for our study:

- Ordinary Least Squares (OLS) Regression
- Ramsey RESET Test
- Variance Inflation Factor (VIF)
- Breusch-Pagan Test
- Shapiro-Wilk W Test

The dataset is called ‘Student Study Performance’ and is taken from the website kaggle.com

4. Research subject and scope

We have conducted a survey and gathered each factor and their results in math scores in 1000 individuals. The data set gathered includes the following attributes for each person: gender, race/ethnicity, parental level of education, lunch price, test preparation course, and math score. Here's a breakdown of the key variables:

- Gender: A categorical variable capturing the student's gender identity (e.g., male, female). Previous research suggests potential gender disparities in math performance, warranting closer examination.
- Race_Ethnicity: This categorical variable represents the student's racial or ethnic background. Understanding how race and ethnicity intersect with math scores can help identify potential levels of educational opportunities.
- Parental_Level_of_Education: Categorized by the highest level of education the student's parents attained, this variable is a proxy for socioeconomic status (SES). Research suggests a link between SES and academic achievement, and this analysis aims to quantify that association.
- Lunch: Categorical data indicating eligibility for the free/reduced lunch program. Analyzing this variable alongside parental education can provide a more nuanced understanding of how socioeconomic background influences math scores.
- Test_Preparation_Course: This binary variable indicates whether a student finished a test preparation course. Investigating the impact of such interventions can inform decisions regarding resource allocation and support programs.
- Math Score: Our primary outcome variable is represented by a continuous score on a standardized math test.

II. Result

1. Process and analysis steps

The data used for the study on factors influencing students' math scores has been collected and processed to suit the research objectives. Statistical comparative and regression methods have been applied to assess the relationship between independent and dependent variables. In this case, a multivariate regression analysis using the Ordinary Least Squares (OLS) method has been employed to examine the factors and their impacts on students' math scores.

The structure of the study is organized into sections as follows:

- **Descriptive analysis:**

In this section, our group identifies and explain the significance of the 5 main parameters, including mean, median, variance, standard deviation, and correlation coefficient. These parameters help us understand the distribution and relationship between variables. Identifying outliers and considering how to handle them if necessary to ensure the accuracy of the model.

- **Choosing functional form for variables:**

We determine the appropriate functional form for each variable based on their characteristics and the model's objectives.

- **Variable selection:**

Identifying other influencing factors to include in the model to ensure the study's completeness.

Removing variables that do not significantly affect students' math scores and explaining the relationships between variables to ensure the model's validity.

- Checking the model/residual diagnostics:

At this stage, we evaluate the model's suitability and identify and attempt to address any issues that may arise.

- Forecast application, Limitations, Recommendations:

Forecast application: Once the model has been built and tested, it can be used to predict students' math scores based on the selected independent variables.

Limitations: In this section, we identify and suggest limitations of the research method, as well as any limitations of the data used.

Recommendations: Based on the study's results, we propose recommendations that could help improve academic performance and educational policies. This may include enhancing test preparation, improving students' dietary conditions, and providing educational support for parents with lower levels of education.

2. Descriptive analysis

The data provided includes information about 1000 students, with variables such as gender, race ethnicity, parental level of education, lunch type, participation in test preparation courses, and math scores. Analysis of the given data reveals diversity and uneven distribution across variables. Students in this dataset come from various racial/ethnic groups, with wide-ranging diversity in parental education levels. While there are different types of lunches available, it appears that some students receive either free or reduced lunches. Furthermore,

most students do not participate in test preparation courses. Thus, this data provides an overview of factors that may influence students' academic performance, serving as a basis for further research and appropriate intervention strategies to enhance the quality of education.

2.1. Null variables dectection:

First, checking for the number of missing values in the data, and no values were found to be missing. (see Appendix EA for Python source code)

```
gender                       0
race_ethnicity               0
parental_level_of_education  0
lunch                        0
test_preparation_course      0
math_score                   0
```

Table 2.1. Checking null variables

2.2. Outlier:

A box plot was drawn to visualize the distribution of the 'math_score'. This plot helps us identify any potential outliers that may exist in the data.

Subsequently, a statistical method known as the Interquartile Range (IQR) was applied to identify outliers in the 'math_score'. These outlier values were then removed from the data to prepare for further analysis. (see Appendix EB for Python source code)

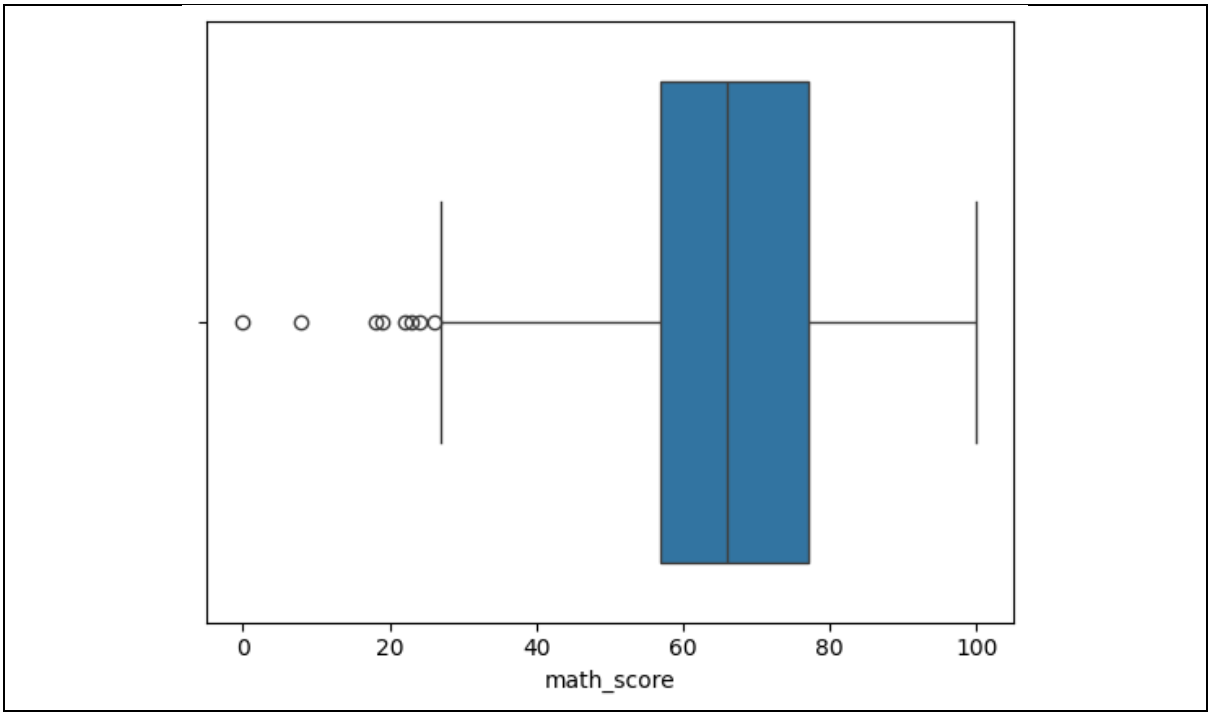


Table 2.2. Checking outliers code

From the initial 1000 data points, after applying the IQR method, we removed 8 outliers.

2.3. Variables

2.3.1. math_score:

The math scores dataset contains 992 data points, with scores ranging from 27 to 100. The average math score is approximately 66.48, and the median is 66, suggesting a roughly symmetric distribution around the mean. The standard deviation is 14.56, indicating a moderate spread around the mean. Additionally, 25% of the scores are 57 or below, and 75% are 77 or below. The distribution appears to be fairly symmetric, given the close alignment of the mean and median. (see Appendix EC for Python source code)

count	992.000000
mean	66.480847
std	14.559999
min	27.000000
25%	57.000000
50%	66.000000
75%	77.000000
max	100.000000

Table 2.3.1 Statistical data of 'math_score'

2.3.2. Gender:

female	51.41129
male	48.58871

Table 2.3.2. Statistical data of 'gender'

The result above shows a slight difference in the percentage (%) of female and male students. (see Appendix ED for Python source code)

2.3.3. Race ethnicity:

group C	31.955645
group D	26.310484
group B	18.649194
group E	14.112903
group A	8.971774

Table 2.3.3. Statistical data of 'race_ethnictity'

Group C has the highest proportion of students, approximately 31.96%, and group A has the lowest student proportion of 8.97%. (see Appendix EE for Python source code)

2.3.4. Parental level of education:

some college	22.580645
associate's degree	22.278226
high school	19.556452
some high school	17.741935
bachelor's degree	11.895161
master's degree	5.947581

Table 2.3.4. Statistical data of 'parental_level_of_education'

- "some colleges" account for the highest proportion among parental education groups, approximately 22.58% of the total. These are typically parents who have completed some college education but have not yet obtained a bachelor's degree.
- "high school" represents parents who have completed high school education, while "some high school" represents parents who have not completed high school education or have only partially completed it.
- "master's degree" represents the smallest proportion, accounting for approximately 5.95% of the total parents, the lowest rate. Representing parents with the highest level of educational attainment.
  (see Appendix EF for Python source code)

2.3.5. Lunch standard:

standard	64.919355
free/reduced	35.080645

Table 2.3.5. Statistical data of 'lunch_standard'

The proportion of students having lunch under the "standard" and "free/reduced" schemes are 64.92% and 35.08% respectively. (see Appendix EG for Python source code)

2.3.6. Test preparation course:

none	64.012097
completed	35.987903

Table 2.3.6. Statistical data of 'test_preparation_course'

The proportion of students not participating and participating in test preparation courses is 64.01% and 35.99%, respectively. (see Appendix EH for Python source code)

2.4. Transforming categorical data:

The dataset consists of various categorical variables such as gender, ethnicity, parental education level, lunch type, and participation in test preparation courses.

In preparation for regression analysis, it is essential to transform these categorical variables into dummy variables. Dummy variables are binary variables representing categories in a categorical variable. They take the value 0 or 1, indicating the absence or presence of a category, respectively.

In our dataset, the following categorical variables will be converted into dummy variables:

- Gender (male, female)
- Ethnicity (A, B, C, D, E)
- Parental education level (associate degree, bachelor's degree, high school, master's degree, some college, some high school)
- Lunch type (standard, free/reduced)
- Test preparation course (none, completed)

Once these categorical variables are transformed into dummy variables, they will replace the original categorical variables in the dataset.

	math_score	gender	Race_ethnicity		lunch	Test_..._ course
0	72	female	group B		standard	none
1	69	female	group C	…	standard	completed
2	90	female	group B		standard	none
…						
989	59	male	group C		free	completed
990	68	male	group D	…	standard	completed
991	77	male	group D		free	none

Table 2.4.1. Original dataset

	math_score	Male	Female	Race…A	Race…B		Test..one	Test..ted
0	72	0	1	0	1		1	0
1	69	0	1	0	0	…	0	1
2	90	0	1	0	1		1	0
…								
989	59	1	0	0	0		0	1
990	68	1	0	0	0	…	0	1
991	77	1	0	0	0		1	0

Table 2.4.2. Categorical variables transformed dataset

2.5. Correlation between variables:

By using Pearson correlation, we can observe the linear correlation between variables. Linear correlation between variables is not too strong, except for 3 pairs:

- male and female
- lunch_standard and free/reduced_lunch
- test_preparation_course_none and test_preparation_course_completed

with a correlation coefficient of -1, meaning that if one variable occurs, the other cannot occur. (see Appendix EI for result table)

3. Choosing the Function:

For the multiple regression analysis, choosing a linear function of the independent variables to model the relationship between the math scores and various factors. The function takes the form:

$$\text{math\_score} = \beta_0 + \beta_1 \cdot \text{race\_ethnicity} + \beta_2 \cdot \text{lunch} + \beta_3 \cdot \text{parental\_level\_of\_education} + \beta_4 \cdot \text{gender} + \beta_5 \cdot \text{test\_preparation\_course} + \epsilon$$

Where:

- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the coefficients corresponding to each variable.
- ϵ represents the random error.

This choice is based on the assumption that the relationship between math scores and the independent variables is linear.

Each variable here will be replaced by the corresponding set of dummies, with the base category removed to avoid dummy variable trap.

4. Stepwise regression

In summary, bidirectional elimination approach are chosen. At first, math_score is simply regressed on each variable sets gender (male, female) race_ethnicity (Group A, B, C, D, E), lunch (standard, free/reduced), test_preparation_course (none, completed), parental_level_education (some college, associate's degree, high school, some high school, bachelor's degree, master's degree). If t–test P-value of variables in variable groups smaller than alpha-to-enter (0.15), they will be added to candidates group. In that candidate group, if a variable's t-test P-value is min, absolute t-value is max, all variables at the same variable sets will be added to the function. Afterward, the affect of added variables will be tested. If p-values of available variables are lower than alpha-to-out (0.15) and the adjusted r-squared, r-squared are increase, they will be kept. Otherwise, they will be eliminated from the function. The procedure will be continued by regress math_score, added variables with othe other variables untill no valid variables left.

4.1. First variable

While no independent variables exist, performing a regression of math scores on gender (male, female) race_ethnicity (Group A, B, C, D, E), lunch (standard, free/reduced), test_preparation_course (none, completed), parental_level_education (some college, associate's degree, high school, some high school, bachelor's degree, master's degree).Variables like female, race_ethnicity_A, bachelor_degree, free/reduced, and test_preparation_completed as base groups.

The analysis reveals that the P-value of lunch_standard is extremely low (equal to 0), indicating a significant correlation. Though other variables meet this condition, lunch_standard has the highest absolute t-test value (see Appendix FA for all result tables). Therefore, lunch_standard becomes the first variable added to the model:

$$\text{math\_score} = \beta_0 + \beta_1 \cdot \text{lunch\_standard} + u$$

4.2. Second variable

While there's independent variables exist, regress math_score, lunch_standard, sequentially with the remaining variables.

Similarly, test_preparation_none not only has a p-value smaller than Alpha to Enter but also has the highest absolute t-test value (see Appendix FB for all result tables). Therefore, test_prepartion_none will be the next variable added. The model now takes the form:

$$\text{math\_score} = \beta_0 + \beta_1 \cdot \text{lunch\_standard} + \beta 2 \cdot \text{test\_preparation\_none} + u$$

Looking back, after adding the test_preparation_none variable, we need to check if the variables are significant.

We see that the p-value = 0 and is smaller than Alpha to Remove. Not only that, the Adjusted R-squared also increases, so no variables are removed.

4.3. Third variable

Running regression on the three variables math_score, lunch_standard, test_preparation with the remaining 3 variables. Race_ethnicity_D and E have p-values smaller than Alpha to Enter, and have the highest absolute t-test value. So the race_ethnicity variable is added to the model. Although race_ethnicity_B and C are in the same group as D and E, their p-values are higher than Alpha to Enter, so these 2 variables do not have a significant impact on the model and will be added to the cons (see Appendix FC for all result tables). The model after being added is:

$$\text{math\_score} = \beta_0 + \beta_1 \cdot \text{lunch\_standard} + \beta 2 \cdot \text{test\_preparation\_none} + \beta 3 \cdot \text{race\_ethnicity\_D} + \beta 4 \cdot \text{race\_ethnicity\_E} + u$$

The p-value = 0 for the t-test and is smaller than Alpha to Remove for both pairs. Not only that, the Adjusted R-squared also increases, so no variables are removed.

4.4. Fourth variable

Continuing as above with the remaining 2 variables (see Appendix FD for all result tables). Thus, male will be the next variable to meet the conditions and will be added to the model. The model after adding male is:

math_score = $\beta_0 + \beta_1 \cdot$ lunch_standard + $\beta 2 \cdot$ test_preparation_none + $\beta 3 \cdot$ race_ethnicity_D + $\beta 4 \cdot$ race_ethnicity_E + $\beta 5 \cdot$ male + u

Adjusted R-squared increases, and the p-values of all t-tests are greater than Alpha to Remove, so no variables are removed.

4.5. Fifth variable

Similarly with the last variable, parental_education_level. From the table, the variables in parental_education_level, except associate_degree, master_degree, some_college, do not have a significant impact and are added to the cons (see Appendix FE for all result tables). The 2 variables high_school and some_high_school, meeting the conditions to be added:

math_score = $\beta_0 + \beta_1 \cdot$ lunch_standard + $\beta 2 \cdot$ test_preparation_none + $\beta 3 \cdot$ race_ethnicity_D + $\beta 4 \cdot$ race_ethnicity_E + $\beta 5 \cdot$ male + $\beta 6 \cdot$ high_school + $\beta 7 \cdot$ some_high_school

Adjusted R-squared increases, and the p-values of all t-tests are greater than Alpha to Remove, so all variables are retained.

4.6. Final function form:

The variable high_school and some_high_school causing difficulty in distinguishing, will be combined into one variable, called high_school_merged (see Appendix FE for all result tables). At this point, the model takes the form:

math_score = $\beta_0 + \beta_1 \cdot$ lunch_standard + $\beta 2 \cdot$ test_preparation_none + $\beta 3 \cdot$ race_ethnicity_D + $\beta 4 \cdot$ race_ethnicity_E + $\beta 5 \cdot$ male + $\beta 6 \cdot$ high_school_merged

We will get a complete model:

math_score = 60.95309 + 4.254953 * male + 3.144104 * race_ethnicity_D + 7.814428 * race_ethnicity_E - 4.725267 * high_school_merged + 10.23981 * lunch_standard - 5.241064 * test_preparation_course_none

5. Estimated results and statical inferences

5.1. Running OLS

5.1.1. Model Performance:

From the below table, our group saw that P-value (P > |t|) < 0,05. Therefore, all independent variables affect the dependent variable and model is statistically significance level at 5%.

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.24	.852	12.02	0	8.567	11.912	***
test_preparation_c~e	-5.241	.848	-6.18	0	-6.906	-3.576	***
race_ethnicity_D	3.144	.952	3.30	.001	1.275	5.013	***
race_ethnicity_E	7.814	1.209	6.46	0	5.442	10.187	***
male	4.255	.814	5.23	0	2.658	5.852	***
high_school_merged	-4.725	.843	-5.60	0	-6.38	-3.07	***
Constant	60.953	1.043	58.44	0	58.906	63	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.233	Number of obs	992
F-test	49.974	Prob > F	0.000
Akaike crit. (AIC)	7878.252	Bayesian crit. (BIC)	7912.550

*** p<.01, ** p<.05, * p<.1

Table 5.1.1. Regression of 'math_score' on all other variables

5.1.2. Lunch Standard:

On average, students with standard lunches achieve 10.24 points more in math score than those with free/reduced lunches, after controlling for all other variables. This implies that providing standard lunches may be beneficial for improving students' math scores.

5.1.3. Test Preparation Course None:

On average, not participating in test preparation courses is associated with a decrease in math scores by approximately 5.24 points compared to students who do participate, keeping all other variables constant.

5.1.4. Race Ethnicity D and Ethnicity E:

On average, the estimated coefficient for ethnicity D is 3.144104, indicating that students of ethnicity D have math scores a 3.14 points higher than ethnicities A, B, C. For ethnicity E, the estimated coefficient is 7.814428, indicating that students of ethnicity E have math scores approximately 7.81 points higher than students of other ethnicities

5.1.5. Male:

The coefficient for the "Male" variable is positive (4.254953), indicating that male students, on average, have math scores 4.25 points higher than female students.

5.1.6. High School Merged:

The coefficient for the "High School Merged" variable is negative (-4.725267), indicating that students whose parents have a high school level of education, on average, have math scores 4.73 points lower than students whose parents have a higher level of education (such as an associate degree, master's degree, or bachelor's degree), after controlling for all other variables.

5.1.7. Constant:

The average score of students who are female, etc. (i.e. all other variables are 0) is 60.95.

5.2. Check omitted variables

We conduct Ramsey RESET Test for omitted independent variables, which result is noted in the below table:

Ramsey RESET test for omitted variables Omitted: Powers of fitted values of math_score H0: Model has no omitted variables F(3, 982) = 0.48

Table 5.2.1. Ramsey RESET test result

Prob > F = 0.6970

H: Model has no omitted variables.

H1: Model has omitted variables.

f-stat = 0.6970 > 0.05 => not reject null hypothesis => model has no omitted variables

If the calculated F-statistic is less than the critical value corresponding to the chosen significance level (usually 0.05), we fail to reject the null hypothesis. However, we can only implying that the model may have no omitted variables, since Ramsey RESET test is used (further interpretion in ‘Limitation and prediction’).

5.3. Check Multicollinearity

Regarding detecting multicollinearity, we employed the technique of Variance Inflation Factor (VIF) to diagnose the issue

Variable	VIF	1/VIF
race_ethnicity_E	1.070	0.930
race_ethnicity_D	1.070	0.937
high_school_merged	1.010	0.991
test_preparation_course_none	1.010	0.994
male	1.000	0.996
lunch_standard	1.000	0.996
Mean VIF	1.030	

Table 5.3.1. VIF test result

In conclusion, variables with VIF indicators lower than 10 show no signs of multicollinearity. Therefore, our model does not exhibit errors related to the mentioned issue.

5.4. Check Heteroskedasticity

We have conducted the Breusch-Pagan test, of which result is shown in this figure:

```
Breusch-Pagan / Cook–Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of math_score
H0: Constant variance
    chi2(1) =   0.24
Prob > chi2 = 0.6241
```

Table 5.4.1. Breusch-Pagan / Cook–Weisberg test result

with the null hypothesis of:

H0: Homoscedasticity

H1: Heteroscedasticity

We have: p-value = 0.6241 > 0.05

→there is no heteroskedasticity.

Based on the Breusch-Pagan test, with a p-value of 0.6241, which exceeds the significance level of 0.05, we fail to reject the null hypothesis, indicating no evidence of heteroskedascity in the model.

5.5. Check distribution of residual

Hypothesis Pair:

H0: The residuals follow a normal distribution.

H1: The residuals do not follow a normal distribution.

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
residuals	992	0.995	2.895	2.631	0.004

Table 5.5.1. Shapiro-Wilk W test result

p-value = 0.000425 < 0.05 => not normally distributed

The Shapiro-Wilk test results indicate that the p-value (0.00425) is less than the significance level (0.05).

Therefore, we reject the null hypothesis, suggesting that the residuals are not normally distributed.

6. Limitation and prediction:

6.1. Predictional Application:

Once the regression model has been built and validated, it can be utilized to forecast students' math scores based on the selected independent variables. By inputting the relevant values of the independent variables into the model, educators and policymakers can obtain predicted math scores for individual students or groups of students. This forecasting capability enables educational institutions to anticipate students' academic performance and tailor interventions and support accordingly. For example, if a student is identified as being at risk of low math scores based on the model's predictions, targeted interventions such as additional tutoring or personalized learning plans can be implemented to improve their academic outcomes.

6.2. Limitations:

- Omitted variables: Omitted varaibles of a Ramsey RESET are just the interaction variables and higher orders of the variables already included in the analysis. It can’t say if the model is

free from other omitted variables such as family income, number of siblings, etc, which might be important to explain students' math performance. So the model might still be subject to omitted variable issue

- Distribution of residuals: The result from residual distribution test previously shows that residuals are not normally distributed.

- Generalizability: The findings and recommendations derived from the regression analysis may not be generalizable to all student populations. The study's sample size, demographic composition, and contextual factors may limit the applicability of the results to other educational settings.

- Data Quality: The accuracy and reliability of the regression model depend heavily on the quality of the data used for analysis. Issues such as missing data, measurement errors, or sampling biases can introduce inaccuracies and uncertainties into the model's predictions.

- Model Assumptions: The regression model relies on several assumptions, including linearity, independence of errors, and normality of residuals. Violations of these assumptions can undermine the validity and reliability of the model's predictions.

- Extraneous Variables: The regression model may not account for all potential factors influencing students' math scores. Unobserved variables or omitted variables could introduce confounding effects and bias the model's predictions.

- Temporal Dynamics: The regression model assumes that the relationships between independent and dependent variables remain stable over time. Changes in educational policies, teaching methodologies, or socio-economic conditions may affect the predictive accuracy of the model over time.

6.3. Recommendations:

- Continuous Monitoring: Educational institutions should regularly monitor students' academic performance and update the regression model accordingly. By collecting longitudinal data and incorporating new variables or modifying existing ones, the model can adapt to changing dynamics and improve its predictive accuracy.

- Data Enhancement: Efforts should be made to enhance the quality and completeness of the data used for regression analysis. This may involve implementing data collection protocols, addressing missing data issues, and conducting sensitivity analyses to assess the robustness of the model to data variations.

- External Validation: The regression model's predictions should be validated using external datasets or cross-validation techniques. By comparing the model's predictions with observed outcomes in independent samples, educators and researchers can assess its generalizability and reliability.

- Interpretation and Communication: The findings and predictions generated by the regression model should be communicated effectively to relevant stakeholders, including educators, policymakers, and parents. Clear explanations of the model's assumptions, limitations, and uncertainties can facilitate informed decision-making and ensure the appropriate use of the model's predictions in practice.

- Targeted Interventions: Based on the model's predictions, targeted interventions and support programs should be implemented to address the specific needs of students at risk of academic underachievement. These interventions may include personalized tutoring, mentoring, academic enrichment programs, and parental involvement initiatives.

IV. Conclusion and Recommendations:

1. Conclusion

In summary, analyzing the data of 1000 students on various factors influencing students' math scores provides valuable insights into the educational landscape. The statistical examination of math scores reveals an overall distribution centered around the mean, indicating moderate consistency across the dataset.

Gender distribution demonstrates a slight disparity between male and female students, while ethnicity reveals varying proportions across different groups, with Group C being the largest segment and Group A the smallest. Parental level of education exhibits diverse backgrounds, with "some college" being the most common category, potentially impacting the level of support students receive in their math education. The gap in lunch programs underscores the potential influence of socioeconomic factors on academic performance, with students under the "standard" plan likely to benefit from improved nutrition compared to their peers under the "free/reduced" plan. Furthermore, participation in test preparation courses emerges as a significant factor, with completion of such courses correlating with higher math scores. This highlights the importance of academic support structures in enhancing student performance and suggests avenues for improving educational outcomes. Overall, these findings indicate the multifaceted nature of the educational gap and the need for targeted interventions to improve the education level of students.

2. Recommendations

Based on our analysis, we also gave some forecast and recommendations for better math score and inform educational policies:

2.1 Enhanced Test Preparation Programs

Given the significant impact of test preparation courses on math scores, we recommend enhancing access to and participation in these programs. Schools and educational institutions should provide comprehensive and tailored test preparation resources to support students in achieving academic excellence.

2.2 Nutritional Support for Students

The disparity in academic performance between students receiving standard and free/reduced lunches underscores the importance of nutritional support. Efforts should be made to ensure all students have access to healthy and nutritious meals, which can positively influence cognitive function and academic outcomes.

2.3 Parental Engagement and Support

Recognizing the varying levels of parental education and its impact on student performance, initiatives should be implemented to enhance parental engagement in their children's education. Providing resources and guidance to parents, especially those with lower education levels, can empower them to better support their children's learning journey.

In general, by implementing these recommendations and leveraging insights from our analysis, stakeholders can work collaboratively to create a more conducive learning environment and empower students to achieve their full academic potential.

V. Appendices:

Appendix A

Study performance dataset

Appendix B

Categorical variables transformed dataset

Appendix C

Econometrics python code

Appendix D

Econometrics stata code

Appendix EA

Python code:

```
print(data.isnull().sum())
```

Result:

```
gender                       0
race_ethnicity               0
parental_level_of_education  0
lunch                        0
test_preparation_course      0
math_score                   0
```

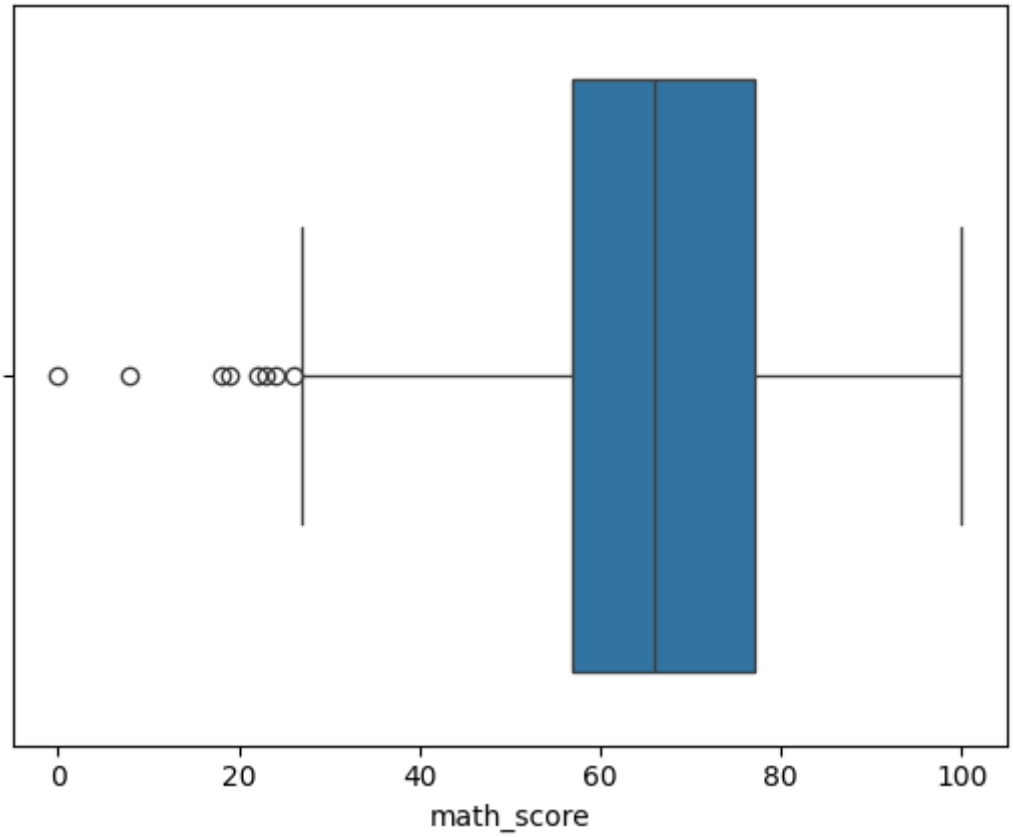
Appendix EB

Python code:

```
sns.boxplot(x=data['math_score'])
plt.show()

Q1 = data['math_score'].quantile(0.25)
Q3 = data['math_score'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
data = data[(data['math_score'] >= lower_bound) & (data['math_score'] <=
upper_bound)]
```

Result:



Appendix EC

Python code:

```
data['math_score'].describe()
```

Result:

count	992.000000
mean	66.480847
std	14.559999
min	27.000000
25%	57.000000
50%	66.000000
75%	77.000000
max	100.000000

Appendix ED

Python code:

```
gender_counts = data['gender'].value_counts()
percentages = (gender_counts / len(data)) * 100

print(percentages)
```

Result:

female	51.41129
male	48.58871

Appendix EE

Python code:

```
race_ethnicity_counts = data['race_ethnicity'].value_counts()
percentages = (race_ethnicity_counts / len(data)) * 100

print(percentages)
```

Result:

group C	31.955645
group D	26.310484
group B	18.649194
group E	14.112903
group A	8.971774

Appendix EF

Python code:

```
parental_level_of_education_counts =
data['parental_level_of_education'].value_counts()
percentages = (parental_level_of_education_counts / len(data)) * 100

print(percentages)
```

Result:

```
some college          22.580645
associate's degree    22.278226
high school           19.556452
some high school      17.741935
bachelor's degree     11.895161
master's degree        5.947581
```

Appendix EG

Python code:

```
lunch_counts = data['lunch'].value_counts()
percentages = (lunch_counts / len(data)) * 100

print(percentages)
```

Result:

```
standard       64.919355
free/reduced   35.080645
```

Appendix EH

Python code:

```
test_preparation_course_counts =
data['test_preparation_course'].value_counts()
percentages = (test_preparation_course_counts / len(data)) * 100

print(percentages)
```

Result:

```
none        64.012097
completed   35.987903
```

Appendix EI

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
(1) math_score	1.000																	
(2) male	0.150	1.000																
(3) female	-0.150	-1.000	1.000															
(4) race_ethnicity_A	-0.105	0.069	-0.069	1.000														
(5) race_ethnicity_B	-0.060	-0.020	0.020	-0.150	1.000													
(6) race_ethnicity_C	-0.079	-0.065	0.065	-0.215	-0.328	1.000												
(7) race_ethnicity_D	0.043	0.028	-0.028	-0.188	-0.286	-0.409	1.000											
(8) race_ethnicity_E	0.204	0.017	-0.017	-0.127	-0.194	-0.278	-0.242	1.000										
(9) associate_degree	0.059	-0.007	0.007	-0.049	-0.001	0.038	-0.050	0.054	1.000									
(10) bachelor_degree	0.073	-0.015	0.015	0.015	-0.016	0.015	-0.022	0.012	-0.197	1.000								
(11) high_school	-0.131	0.039	-0.039	0.005	0.064	0.011	-0.041	-0.039	-0.264	-0.181	1.000							
(12) master_degree	0.056	-0.048	0.048	-0.034	-0.055	0.001	0.072	-0.004	-0.135	-0.092	-0.124	1.000						
(13) some_college	0.039	-0.004	0.004	-0.018	-0.036	-0.019	0.044	0.023	-0.289	-0.198	-0.266	-0.136	1.000					
(14) some_high_sch~l	-0.068	0.013	-0.013	0.076	0.022	-0.047	0.022	-0.052	-0.249	-0.171	-0.229	-0.117	-0.251	1.000				
(15) lunch_standard	0.340	0.013	-0.013	-0.035	-0.001	-0.004	-0.012	0.049	0.008	-0.017	0.000	-0.029	0.003	0.021	1.000			
(16) free_reduced	-0.340	-0.013	0.013	0.035	0.001	0.004	0.012	-0.049	-0.008	0.017	-0.000	0.029	-0.003	-0.021	-1.000	1.000		
(17) test_preparat~n	-0.172	-0.002	0.002	0.008	-0.002	-0.013	0.057	-0.058	-0.012	-0.023	0.078	0.011	0.018	-0.075	0.021	-0.021	1.000	
(18) test_preparat~l	0.172	0.002	-0.002	-0.008	0.002	0.013	-0.057	0.058	0.012	0.023	-0.078	-0.011	-0.018	0.075	-0.021	0.021	-1.000	1.000

Table 2.5. Correlation between variables

Appendix FA

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
male	4.371	.915	4.78	0	2.576	6.167	***
Constant	64.357	.638	100.91	0	63.105	65.608	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.023	Number of obs	992
F-test	22.828	Prob > F	0.000
Akaike crit. (AIC)	8109.262	Bayesian crit. (BIC)	8119.062

*** $p<.01$, ** $p<.05$, * $p<.1$

Table 4.1.1 Regression of 'math_score' on 'male'

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
race_ethnicity_B	3.041	1.83	1.66	.097	-.55	6.633	*
race_ethnicity_C	3.172	1.702	1.86	.063	-.168	6.512	*
race_ethnicity_D	5.892	1.741	3.38	.001	2.474	9.309	***
race_ethnicity_E	12.192	1.923	6.34	0	8.418	15.967	***
Constant	61.629	1.504	40.98	0	58.678	64.58	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.054	Number of obs	992
F-test	14.186	Prob > F	0.000
Akaike crit. (AIC)	8082.423	Bayesian crit. (BIC)	8106.922

*** p<.01, ** p<.05, * p<.1

Table 4.1.2. Regression of 'math_score' on 'race_ethnicity'

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
associate_degree	-1.317	1.64	-0.80	.422	-4.535	1.9	
high_school	-6.771	1.679	-4.03	0	-10.066	-3.476	***
master_degree	.356	2.293	0.16	.877	-4.144	4.856	
some_college	-1.845	1.636	-1.13	.26	-5.056	1.365	
some_high_school	-5.049	1.711	-2.95	.003	-8.407	-1.691	***
Constant	69.39	1.324	52.41	0	66.792	71.988	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.029	Number of obs	992
F-test	5.938	Prob > F	0.000
Akaike crit. (AIC)	8110.449	Bayesian crit. (BIC)	8139.848

*** p<.01, ** p<.05, * p<.1

Table 4.1.3. Regression of 'math_score' on 'parental_level_of_education'

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.355	.912	11.36	0	8.566	12.144	***
Constant	59.759	.734	81.36	0	58.317	61.2	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.115	Number of obs	992
F-test	129.025	Prob > F	0.000
Akaike crit. (AIC)	8010.349	Bayesian crit. (BIC)	8020.148

*** *p<.01, ** p<.05, * p<.1*

Table 4.1.4. Regression of 'math_score' on 'lunch'

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
test_preparation_c~e	-5.226	.949	-5.51	0	-7.089	-3.364	***
Constant	69.826	.759	91.94	0	68.336	71.317	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.030	Number of obs	992
F-test	30.315	Prob > F	0.000
Akaike crit. (AIC)	8101.956	Bayesian crit. (BIC)	8111.756

*** *p<.01, ** p<.05, * p<.1*

Table 4.1.5. Regression of 'math_score' on 'test_preparation_course'

Appendix FB

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.297	.901	11.43	0	8.528	12.065	***
male	4.243	.86	4.93	0	2.555	5.931	***
Constant	57.735	.834	69.23	0	56.098	59.371	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.137	Number of obs	992
F-test	78.191	Prob > F	0.000
Akaike crit. (AIC)	7988.252	Bayesian crit. (BIC)	8002.951

*** *p<.01, ** p<.05, * p<.1*

Table 4.2.1. Regression of 'math_score' on 'lunch' and 'male'

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.026	.89	11.26	0	8.279	11.774	***
race_ethnicity_B	2.508	1.724	1.45	.146	-.875	5.892	
race_ethnicity_C	2.659	1.603	1.66	.098	-.488	5.806	*
race_ethnicity_D	5.447	1.641	3.32	.001	2.228	8.667	***
race_ethnicity_E	11.073	1.814	6.10	0	7.513	14.633	***
Constant	55.658	1.512	36.80	0	52.691	58.626	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.162	Number of obs	992
F-test	38.154	Prob > F	0.000
Akaike crit. (AIC)	7964.420	Bayesian crit. (BIC)	7993.818

*** p<.01, ** p<.05, * p<.1

Table 4.2.2. Regression of 'math_score' on 'lunch' and 'race_ethnicity'

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.474	.898	11.66	0	8.712	12.237	***
associate_degree	-1.621	1.538	-1.05	.292	-4.639	1.397	
high_school	-7.006	1.575	-4.45	0	-10.096	-3.915	***
master_degree	.711	2.151	0.33	.741	-3.51	4.932	
some_college	-2.104	1.534	-1.37	.171	-5.115	.908	
some_high_school	-5.503	1.605	-3.43	.001	-8.653	-2.353	***
Constant	62.821	1.363	46.08	0	60.146	65.497	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.147	Number of obs	992
F-test	28.300	Prob > F	0.000
Akaike crit. (AIC)	7984.111	Bayesian crit. (BIC)	8018.409

*** p<.01, ** p<.05, * p<.1

Table 4.2.3. Regression of 'math_score' on 'lunch' and 'parental_level_of_education'

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.469	.895	11.69	0	8.712	12.227	***
test_preparation_c~e	-5.444	.89	-6.11	0	-7.192	-3.697	***
Constant	63.169	.912	69.28	0	61.38	64.959	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.148	Number of obs	992
F-test	85.580	Prob > F	0.000
Akaike crit. (AIC)	7975.535	Bayesian crit. (BIC)	7990.234

*** *p<.01, ** p<.05, * p<.1*

Table 4.2.4. Regression of 'math_score' on 'lunch' and 'test_preparation_course'

Appendix FC

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.411	.885	11.77	0	8.675	12.148	***
test_preparation_c~e	-5.433	.88	-6.18	0	-7.16	-3.707	***
male	4.23	.845	5.01	0	2.572	5.887	***
Constant	61.145	.988	61.92	0	59.207	63.082	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.169	Number of obs	992
F-test	66.800	Prob > F	0.000
Akaike crit. (AIC)	7952.673	Bayesian crit. (BIC)	7972.272

*** *p<.01, ** p<.05, * p<.1*

Table 4.3.1. Regression of 'math_score' on 'lunch', 'test_preparation_course' and 'male'

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.154	.875	11.60	0	8.437	11.871	***
test_preparation_c~e	-5.247	.871	-6.02	0	-6.956	-3.537	***
race_ethnicity_B	2.429	1.694	1.43	.152	-.896	5.754	
race_ethnicity_C	2.543	1.576	1.61	.107	-.549	5.635	
race_ethnicity_D	5.621	1.612	3.49	.001	2.457	8.785	***
race_ethnicity_E	10.638	1.784	5.96	0	7.137	14.138	***
Constant	59.002	1.586	37.20	0	55.889	62.114	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.192	Number of obs	992
F-test	38.979	Prob > F	0.000
Akaike crit. (AIC)	7930.542	Bayesian crit. (BIC)	7964.840

*** p<.01, ** p<.05, * p<.1

Table 4.3.2. Regression of 'math_score' on 'lunch', 'test_preparation_course' and 'race_ethnicity'

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.594	.882	12.01	0	8.863	12.326	***
test_preparation_c~e	-5.367	.881	-6.09	0	-7.096	-3.637	***
associate_degree	-1.524	1.511	-1.01	.313	-4.488	1.441	
high_school	-6.438	1.549	-4.15	0	-9.478	-3.397	***
master_degree	.988	2.113	0.47	.64	-3.159	5.134	
some_college	-1.859	1.508	-1.23	.218	-4.818	1.099	
some_high_school	-5.764	1.577	-3.65	0	-8.859	-2.669	***
Constant	66.02	1.438	45.90	0	63.198	68.843	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.178	Number of obs	992
F-test	30.444	Prob > F	0.000
Akaike crit. (AIC)	7949.407	Bayesian crit. (BIC)	7988.604

*** p<.01, ** p<.05, * p<.1

Table 4.3.3. Regression of 'math_score' on 'lunch', 'test_preparation_course' and 'parental_level_of_education'

Appendix FD

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.146	.865	11.73	0	8.448	11.843	***
test_preparation_c~e	-5.25	.861	-6.10	0	-6.941	-3.56	***
race_ethnicity_D	3.338	.966	3.45	.001	1.441	5.234	***
race_ethnicity_E	8.365	1.223	6.84	0	5.964	10.765	***
male	4.05	.825	4.91	0	2.43	5.669	***
Constant	59.229	1.012	58.53	0	57.243	61.215	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.209	Number of obs	992
F-test	52.084	Prob > F	0.000
Akaike crit. (AIC)	7907.378	Bayesian crit. (BIC)	7936.776

*** p<.01, ** p<.05, * p<.1

Table 4.4.1. Regression of 'math_score' on 'lunch', 'test_preparation_course', 'race_ethnicity' and 'gender'

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.33	.864	11.95	0	8.634	12.025	***
test_preparation_c~e	-5.196	.864	-6.01	0	-6.892	-3.499	***
race_ethnicity_D	3.268	.969	3.37	.001	1.365	5.17	***
race_ethnicity_E	8.018	1.225	6.55	0	5.615	10.422	***
associate_degree	-1.66	1.478	-1.12	.262	-4.561	1.24	
high_school	-6.102	1.517	-4.02	0	-9.078	-3.126	***
master_degree	.608	2.072	0.29	.769	-3.458	4.673	
some_college	-2.092	1.476	-1.42	.157	-4.989	.804	
some_high_school	-5.494	1.545	-3.56	0	-8.525	-2.463	***
Constant	64.084	1.441	44.47	0	61.256	66.912	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.215	Number of obs	992
F-test	29.888	Prob > F	0.000
Akaike crit. (AIC)	7907.714	Bayesian crit. (BIC)	7956.712

*** $p<.01$, ** $p<.05$, * $p<.1$

Table 4.4.2. Regression of 'math_score' on 'lunch', 'test_preparation_course', 'race_ethnicity' and 'parental_level_of_education

Appendix FE

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.282	.853	12.06	0	8.609	11.955	***
test_preparation_c~e	-5.175	.853	-6.07	0	-6.848	-3.501	***
race_ethnicity_D	3.072	.957	3.21	.001	1.194	4.95	***
race_ethnicity_E	7.842	1.209	6.49	0	5.47	10.215	***
male	4.322	.814	5.31	0	2.723	5.92	***
associate_degree	-1.717	1.458	-1.18	.239	-4.578	1.145	
high_school	-6.37	1.497	-4.26	0	-9.307	-3.432	***
master_degree	.962	2.045	0.47	.638	-3.051	4.974	
some_college	-2.149	1.456	-1.48	.14	-5.006	.709	
some_high_school	-5.637	1.524	-3.70	0	-8.628	-2.647	***
Constant	62.159	1.467	42.37	0	59.28	65.038	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.237	Number of obs	992
F-test	30.459	Prob > F	0.000
Akaike crit. (AIC)	7881.642	Bayesian crit. (BIC)	7935.539

*** $p<.01$, ** $p<.05$, * $p<.1$

Table 4.5.1. Regression of 'math_score' on all other variables

Appendix FF

math_score	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
lunch_standard	10.24	.852	12.02	0	8.567	11.912	***
test_preparation_c~e	-5.241	.848	-6.18	0	-6.906	-3.576	***
race_ethnicity_D	3.144	.952	3.30	.001	1.275	5.013	***
race_ethnicity_E	7.814	1.209	6.46	0	5.442	10.187	***
male	4.255	.814	5.23	0	2.658	5.852	***
high_school_merged	-4.725	.843	-5.60	0	-6.38	-3.07	***
Constant	60.953	1.043	58.44	0	58.906	63	***

Mean dependent var	66.481	SD dependent var	14.560
R-squared	0.233	Number of obs	992
F-test	49.974	Prob > F	0.000
Akaike crit. (AIC)	7878.252	Bayesian crit. (BIC)	7912.550

*** p<.01, ** p<.05, * p<.1

Table 4.5.2. Regression of 'math_score' on all other variables