The Impact of Parental Education on Pupil Testing Scores

Team 24

Walker Mellon, Raja Muhammad Shayan, Marie Rudasics, Stavan Shah

DSE 501

Professor Rong Pan

Abstract

This study aims to explore the impact of parental education levels on student performance, with a focus on math, reading, and writing scores. While considering additional factors such as gender, race, lunch program participation, and test preparation course completion, the analysis centers on how parental education influences academic outcomes. Statistical methods are used to identify trends and determine how parental education shapes performance in each subject. The findings can provide valuable insights for policymakers, educators, and parents, enabling the development of targeted strategies to enhance student achievement and address disparities linked to parental education levels. Additionally, this study examines how these secondary factors interact with parental education, further uncovering the multifaceted influences on student success.

Introduction

Education plays a pivotal role in shaping individuals' socioeconomic and cognitive development, and the factors influencing academic success are a focal point for policymakers and educators worldwide. Among these factors, parental education levels have been identified as a key determinant of student performance. Previous studies have highlighted the profound impact of parental education on academic outcomes, as it often correlates with enhanced learning environments, access to resources, and parental engagement in children's education. Understanding the extent to which parental education influences students' achievements in core subjects such as math, reading, and writing can provide actionable insights for reducing disparities and promoting equitable educational opportunities.

This study seeks to analyze the relationship between parental education levels and student performance while accounting for additional factors such as gender, race, socioeconomic status (indicated by participation in lunch programs), and engagement in test preparation courses. By focusing on standardized test scores in math, reading, and writing, this research aims to uncover patterns that reveal the extent and nuances of these influences.

Problem Statement

Despite significant advancements in educational research, disparities in student performance persist, often linked to familial and socio-economic factors. One critical yet underexplored area is how parental education levels impact academic performance across specific subjects. For instance, students whose parents hold higher educational qualifications may benefit from enriched learning environments, better academic support at home, and higher expectations for academic success.

The study's specific problem is to quantify and analyze the relationship between parental education levels and student performance in math, reading, and writing. The analysis also seeks to identify whether additional factors, such as gender, race, lunch program participation, and test preparation course completion, modify or mediate this relationship. By focusing on this specific problem, the study aims to provide a comprehensive understanding of the role of parental education in shaping student outcomes.

Context of the Problem

The context of this research lies in addressing educational disparities through an evidence-based approach. Standardized test scores serve as key performance indicators for students, schools, and policymakers, often shaping educational interventions and resource allocation. However, disparities in these scores often mirror underlying inequities related to parental education levels and socio-demographic characteristics.

Parental education levels often dictate the quality of support, resources, and motivation available to students. Parents with higher educational attainment may have more time and knowledge to assist with homework, provide supplemental learning materials, or seek additional academic support for their children. This study examines these dynamics using data on student performance in standardized tests, focusing on both subject-specific (math, reading, writing) and broader socio-demographic factors.

Data Collection and Characteristics

The dataset used for this study comprises anonymized student records detailing performance in standardized math, reading, and writing tests. Additional attributes include parental education levels, student demographics (gender, race), socioeconomic indicators (participation in free/reduced lunch programs), and engagement in test preparation courses. The data is representative of diverse socio-economic backgrounds and aims to capture a holistic view of the factors influencing academic performance.

Key characteristics of the data include:

1. **Quantitative Nature:** The dataset primarily contains numerical and categorical variables, such as test scores (continuous) and demographic information (categorical).
2. **Multivariate Structure:** The data allows for analysis across multiple dimensions, such as the interaction between parental education and test preparation.
3. **Diversity:** The dataset includes students from various socio-economic and demographic backgrounds, ensuring the analysis is robust and generalizable.

Hypothesis of Interest

The central hypothesis of this study is that parental education levels significantly influence student performance across math, reading, and writing. Specifically:

- **Primary Hypothesis:** Higher parental education levels are positively associated with better test scores in all three subjects.
- **Secondary Hypotheses:** The relationship between parental education and student performance is moderated by factors such as gender, race, lunch program participation, and test preparation course completion.

Importance of the Solution

Understanding the role of parental education in shaping academic outcomes is crucial for designing effective educational policies and interventions. If parental education is shown to be a strong predictor of performance, schools and policymakers can develop targeted programs, such as parental engagement initiatives, tutoring services, and resource allocation to bridge gaps in academic achievement. Moreover, identifying other moderating factors can help in tailoring interventions to the specific needs of different demographic groups, ensuring equitable educational opportunities.

Literature Review

Parental education has been widely recognized as a critical factor influencing student academic performance across diverse educational contexts. Several studies have explored its impact, alongside related demographic and socioeconomic attributes, shedding light on the mechanisms through which parental education shapes student success.

In a study conducted in Kuala Terengganu, Malaysia, researchers identified a strong correlation between parental education levels and students' academic performance. The study found that students with parents holding higher levels of education tended to achieve better academic outcomes in subjects like mathematics and language proficiency. This influence was mediated by improved access to educational resources and parental involvement in academic activities, which fostered better learning environments for students. Furthermore, the study highlighted gender differences, noting that female students demonstrated better academic performance compared to male students when associated with parents possessing higher education levels. This outcome was attributed to the nurturing environments created by such parents, which seemed to particularly benefit female students' learning processes. [1]

Similarly, a study focusing on English proficiency exam scores demonstrated the statistical significance of parental education levels on students' language achievements. Using a chi-square test, the authors revealed that higher levels of parental education were associated with improved English proficiency, indicating that educated parents may provide enhanced language exposure and support for academic activities at home. The study underscores the role of parental education in bridging gaps in student performance, particularly in linguistically demanding subjects. [2]

In addition to exploring direct relationships, researchers have also employed machine learning techniques to predict student performance based on parental and other factors. One study utilized decision tree algorithms to analyze a comprehensive dataset, identifying parental education as one of the top predictors of academic success. The findings highlighted that parents with higher education levels often create environments that emphasize discipline and academic rigor, which contribute positively to student outcomes. Interestingly, the study found that the benefits of parental education disproportionately favored female students, who outperformed their male counterparts under similar conditions. This was attributed to higher parental expectations and the structured support systems that appeared to be more aligned with the learning preferences of female students. [3]

Furthermore, a broader investigation into the factors influencing student performance and career choices provided additional insights into the systemic impact of parental education. This study revealed that parents' educational attainment significantly affects not only academic success but also the career aspirations and decisions of students. High parental education levels were linked to better GPA outcomes and a higher likelihood of pursuing challenging academic and professional paths. The authors also emphasized the intersection of parental education with socioeconomic factors, such as financial stability, which further compounds its influence on student achievement. [4]

Collectively, these studies provide a robust foundation for understanding the pivotal role of parental education in shaping academic outcomes. They also highlight the need for targeted interventions and policy measures aimed at reducing educational disparities caused by variations in parental education levels. By leveraging insights from these findings, stakeholders can better design strategies to foster equitable educational opportunities and enhance student success.

Data Exploration

In this study, we utilize a dataset that captures student performance across three subjects: math, reading, and writing. The dataset also provides details on socioeconomic and demographic factors, including parental education levels, lunch program participation, gender, race/ethnicity, and test preparation course completion. Before conducting statistical analysis, it is essential to explore the data to identify trends, understand variability, and recognize potential outliers. This initial exploration provides critical context for interpreting the relationships between these factors and student performance.

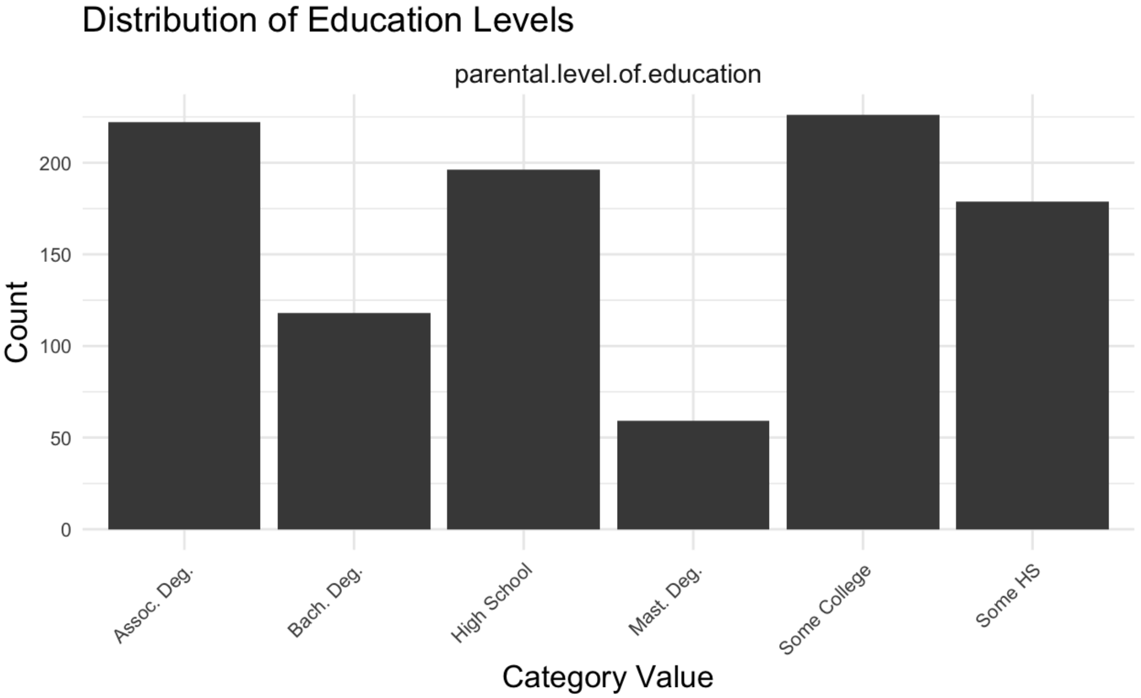


Fig 1. Distribution of Parental Education Levels

In Figure 1, we can see the distribution of parental education levels within our student dataset. This can give us insight into the variation within this variable. The most represented group within this dataset is the "Some College" group with 226 individuals, with the Associate's Degree group following with 222 individuals. The smallest subset of parental education level is the Master's Degree group, with 59 individuals. Parents with bachelor degrees make up the second least out of the group with 118 individuals. The representation of this data within the group could potentially explain overall trends in education levels and how they can reflect upon a child's testing score.

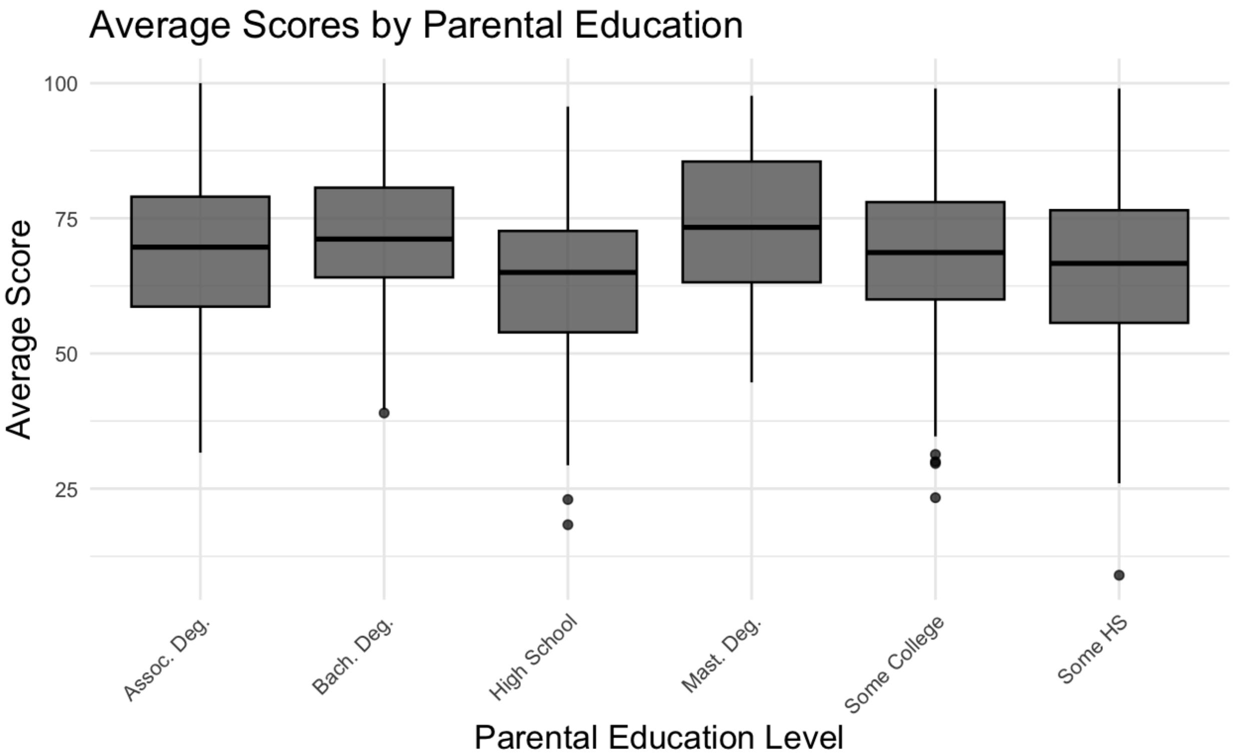


Fig 2. Box plots of Average Scores by Parental Levels of education

Figure 2 highlights the mean and range of average student scores across various parental education levels. The Master's Degree group had the highest mean average score at 73.6%, with a range of 44.67% to 97.67%. Following this, the Bachelor's Degree group had a mean score of 71.92%, with a range of 39% to 100%. The Associate's Degree group reported a mean score of 69.57%, ranging from 31.67% to 100%. Students in the "Some College" group had a mean score of 68.48%, with a range of 23.33% to 99%. The High School group had a lower mean of 63.01%, with a range spanning 18.33% to 95.67%. Finally, the "Some High School" group had the lowest mean average score at 65.11%, but exhibited the widest range, spanning 9% to 99%. This data underscores a general trend of higher parental education correlating with higher student performance and narrower score ranges. This will provide the basis for our initial data testing to understand the statistical difference between these groups. Box plots for all scores are in the provided R markdown.

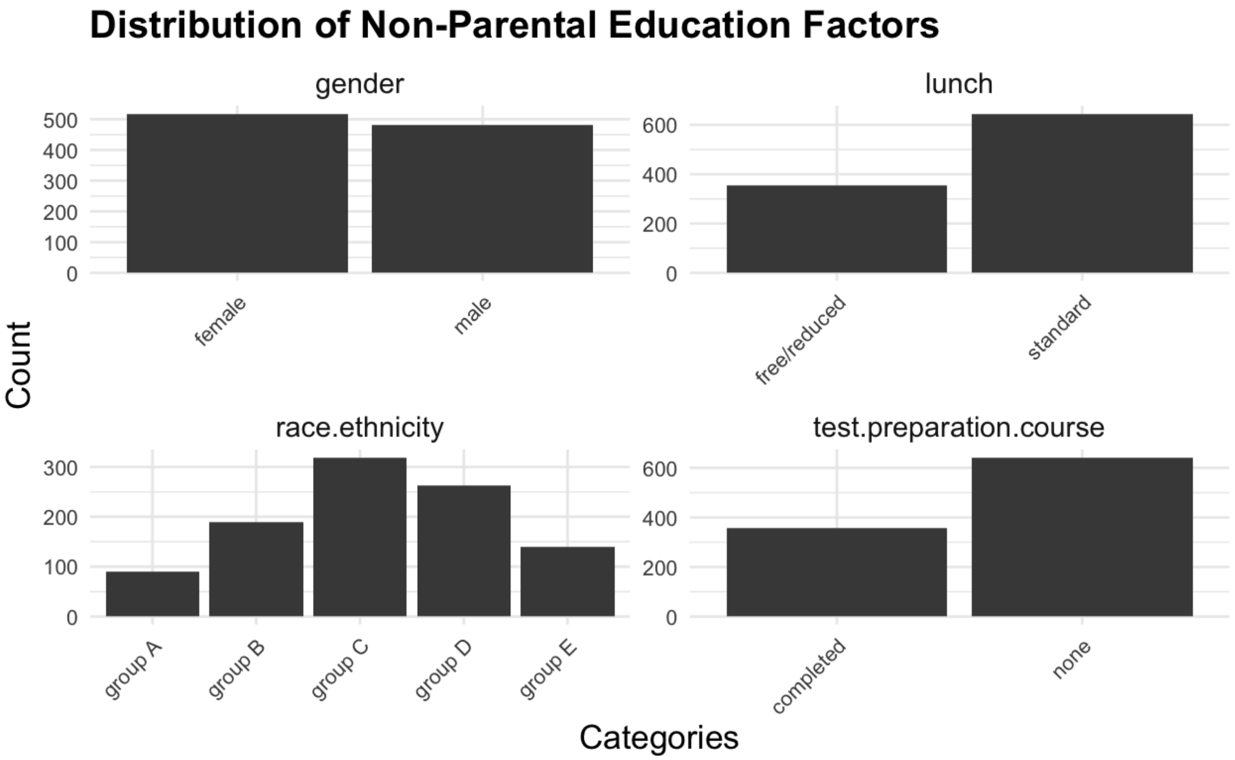
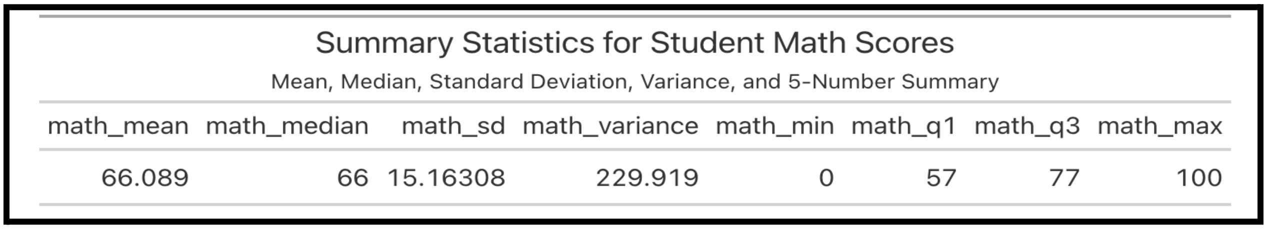


Fig 3. Distribution of demographic variables

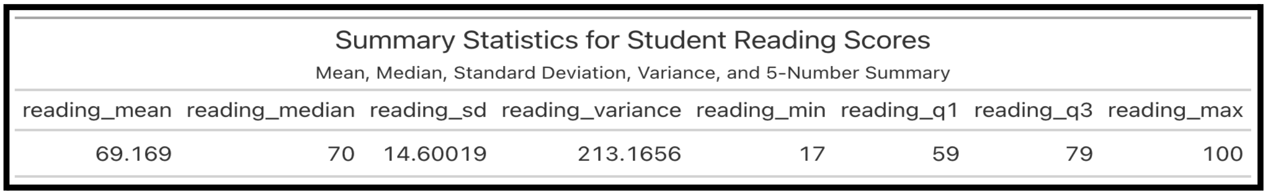
Figure 3 displays the distribution of the variables within our data that are not parental level of education. We can see that the gender distribution is nearly balanced, with 518 females and 482 males. For lunch program participation, a significant majority (645) of students are on the standard lunch program, while 355 participate in free or reduced lunch programs, reflecting differences in socioeconomic status. Regarding race/ethnicity, group C is the largest group with 319 students, followed by group D (262), group B (190), group E (140), and group A (89), indicating varied representation across racial/ethnic groups. Finally, in the test preparation course category, 358 students completed the course, while the remaining majority did not.



Summary Statistics for Student Math Scores

Mean, Median, Standard Deviation, Variance, and 5-Number Summary

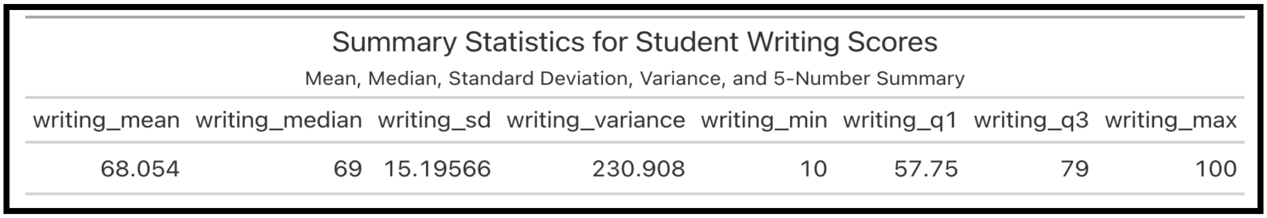
math_mean	math_median	math_sd	math_variance	math_min	math_q1	math_q3	math_max
66.089	66	15.16308	229.919	0	57	77	100



Summary Statistics for Student Reading Scores

Mean, Median, Standard Deviation, Variance, and 5-Number Summary

reading_mean	reading_median	reading_sd	reading_variance	reading_min	reading_q1	reading_q3	reading_max
69.169	70	14.60019	213.1656	17	59	79	100



Summary Statistics for Student Writing Scores

Mean, Median, Standard Deviation, Variance, and 5-Number Summary

writing_mean	writing_median	writing_sd	writing_variance	writing_min	writing_q1	writing_q3	writing_max
68.054	69	15.19566	230.908	10	57.75	79	100

Table 1. Summary Statistics of Student Dataset

Table 1 presents the summary statistics for each of the scores from the dataset, offering insight into performance trends and variability among students. The ranges for math, reading, and writing scores highlight significant diversity in student outcomes, with some students achieving perfect scores while others score considerably lower. The summaries show that while median scores are fairly consistent across subjects, there is notable variability in the lower and upper quartiles, reflecting a wide distribution of scores.

Problems Found From Initial Data Analysis

The initial data analysis of the provided dataset revealed several challenges and issues that must be addressed to ensure accurate and meaningful results. One key issue is the imbalance in the spread of parental education levels represented in the data. Whereas "some college" has a very high number, "master's degree" has a very low frequency. This can lead to biased analysis and reduce the generalizability of the findings. Such imbalances could result in overemphasizing the more frequent categories and underrepresenting the less frequent categories, thus affecting the reliability of any conclusions drawn. This balance is also expected since higher levels of education are harder to complete. This distribution could display realistic educational attainment levels within a larger population.

There is an overlap in the performance of students across different parental education levels. Students whose parents have "some high school" as their highest education level sometimes perform like students whose parents have a "bachelor's degree." For instance, they sometimes post the same scores in math, reading, and writing. Such an overlap confuses the identification of a clear and consistent trend between parental education and student performance, meaning other factors may also be crucial.

Other observations include variables unrelated to parental education, such as lunch type and participation in test preparation courses, which also significantly influence academic performance. Economically disadvantaged students receiving free or reduced-price lunches tend to score lower across all subjects, reflecting the broader impact of socioeconomic status on educational outcomes. Conversely, students who completed a test preparation course consistently perform better in all subjects, regardless of their parent's education levels, suggesting that access to resources and preparation opportunities can play a critical role in academic success. These findings highlight the multifaceted nature of student performance, demonstrating that it is shaped by a combination of social, economic, and resource-based factors beyond parental education.

Summary statistics indicate the presence of outliers within the dataset, which is common and expected in educational data. Students facing challenging socioeconomic conditions may outperform peers from more privileged backgrounds, while some students from privileged circumstances may score unexpectedly poorly. The reasons behind these variations are often influenced by complex social and personal factors, which are beyond the scope of this dataset and analysis. This highlights the importance of interpreting educational outcomes with an understanding that individual performance cannot always be predicted solely by demographic or socioeconomic factors.

Potential multicollinearity between variables also needs attention. The correlation between "race/ethnicity" and "parental level of education" can be used to suggest that some racial groups are overrepresented in certain categories of education. This interaction may well confound the analysis to result in effects being attributed to parental education when, in reality, racial or cultural variables are responsible.

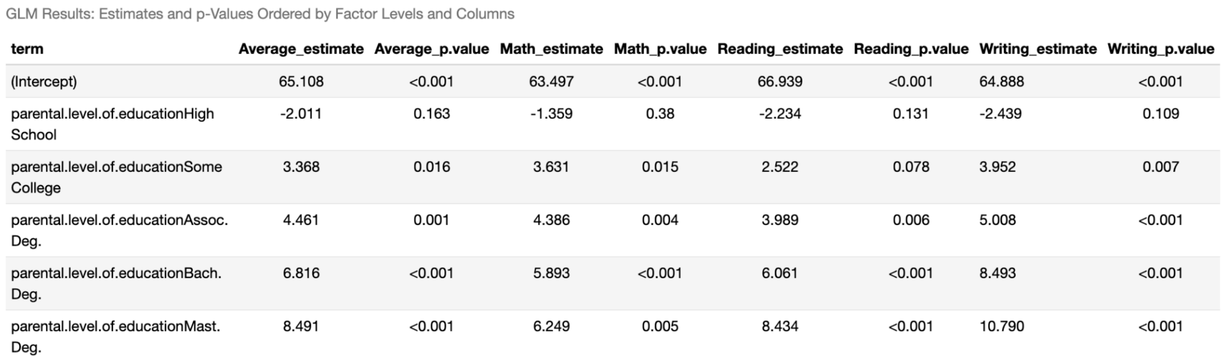
Moreover, the dataset also does not provide context on the performance metrics. It is not known if the scores are for standardized tests, in-class assessments, or some form of evaluation. This confusion can lead to misinterpretations since it is possible that the factors affecting standardized tests, such as test anxiety, may be different compared to classroom assessments.

Among all, the dataset comprises variables like "parental level of education" and "test preparation course," which require categorical encoding into appropriate numerical formats for both statistical and machine learning analyses. Improper encoding could introduce biases or inaccuracies, undermining the interpretability of results. For this study, we chose “Some High School” as the baseline factor for the parental level of education. This is because this is the lowest level of education compared to other present levels.

These are very important aspects to be considered for a comprehensive and unbiased investigation of how parental education level affects the performance of students. Steps to balance the dataset, handle outliers, explore additional influencing variables, and ensure proper encoding of data will serve to enhance the validity and reliability of this analysis.

Statistical Analysis

In this study, we will analyze the relationship between students’ academic performance across math, reading, and writing. We also investigate how other factors such as lunch program performance participation, race/ethnicity, test preparation course completion, and gender can affect test scores alongside parental education level. We implemented univariate and multivariate regression models to measure the impact of educational attainment from parents on student scores. This allows us to see how each level contributes to student performance while controlling for other factors.



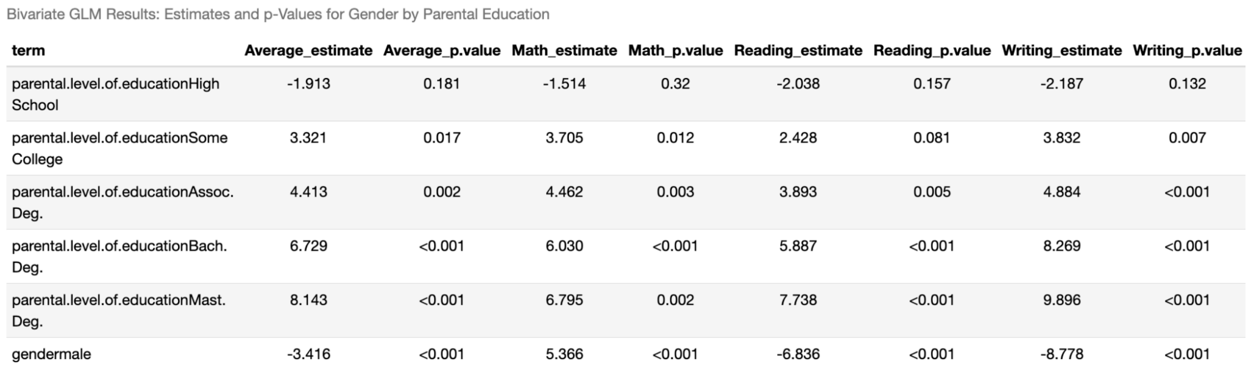
GLM Results: Estimates and p-Values Ordered by Factor Levels and Columns

term	Average_estimate	Average_p.value	Math_estimate	Math_p.value	Reading_estimate	Reading_p.value	Writing_estimate	Writing_p.value
(Intercept)	65.108	<0.001	63.497	<0.001	66.939	<0.001	64.888	<0.001
parental.level.of.educationHigh School	-2.011	0.163	-1.359	0.38	-2.234	0.131	-2.439	0.109
parental.level.of.educationSome College	3.368	0.016	3.631	0.015	2.522	0.078	3.952	0.007
parental.level.of.educationAssoc. Deg.	4.461	0.001	4.386	0.004	3.989	0.006	5.008	<0.001
parental.level.of.educationBach. Deg.	6.816	<0.001	5.893	<0.001	6.061	<0.001	8.493	<0.001
parental.level.of.educationMast. Deg.	8.491	<0.001	6.249	0.005	8.434	<0.001	10.790	<0.001

Table 2. Univariate Generalized Linear Model Results for Writing, Reading, Math, and Average Student scores and Parental Education Level

In the first set of regressions in Table 2, we evaluate the relationship between parental education levels and student performance across reading, math, and writing. We included an average score measure to track overall performance. There was a negative, non-significant trend across all 4 scores between the parental level of education of "High School" and test scores. Students with parents who went to "Some College" had a positive relationship with test scores. Test scores for the "Some College" group were all significant except for the reading scores with a p-value of 0.078. The higher groups all displayed significantly positive trends in all scores.

Within our literature review, we found that there is evidence of demographic differences in test scores across students. With significant trends found within the parental education level, we can investigate how different variables interact with parental education. First, we investigated how gender coincides with

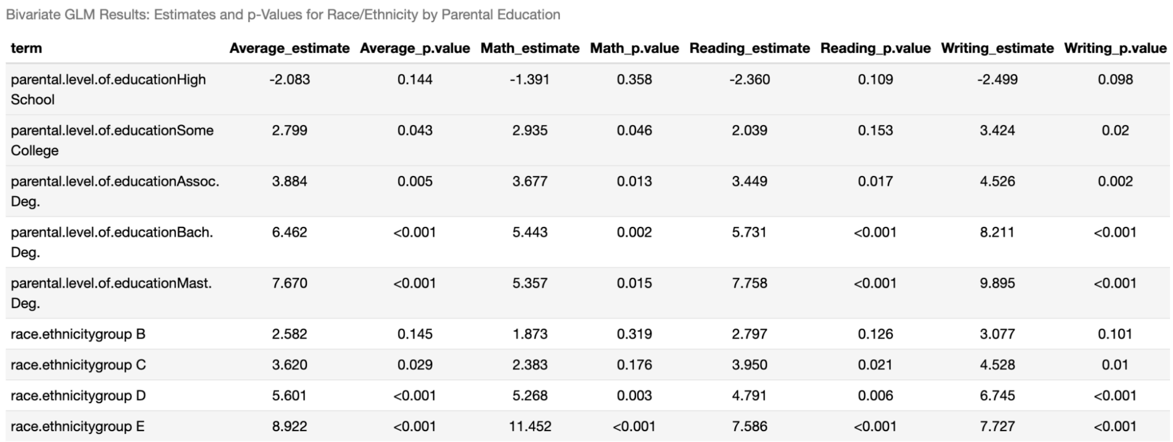


Bivariate GLM Results: Estimates and p-Values for Gender by Parental Education

term	Average_estimate	Average_p.value	Math_estimate	Math_p.value	Reading_estimate	Reading_p.value	Writing_estimate	Writing_p.value
parental.level.of.educationHigh School	-1.913	0.181	-1.514	0.32	-2.038	0.157	-2.187	0.132
parental.level.of.educationSome College	3.321	0.017	3.705	0.012	2.428	0.081	3.832	0.007
parental.level.of.educationAssoc. Deg.	4.413	0.002	4.462	0.003	3.893	0.005	4.884	<0.001
parental.level.of.educationBach. Deg.	6.729	<0.001	6.030	<0.001	5.887	<0.001	8.269	<0.001
parental.level.of.educationMast. Deg.	8.143	<0.001	6.795	0.002	7.738	<0.001	9.896	<0.001
gendermale	-3.416	<0.001	5.366	<0.001	-6.836	<0.001	-8.778	<0.001

Table 3. Bivariate Generalized Linear Model Regression Results for Average Score vs Parental Education and Gender

When introducing gender to the regression we can see that trends and significance of parental education levels remain similar to the univariate model. In some cases, p-values even drop, but not enough for non-significant levels to become significant. In conjunction with parental education, male students had significantly worse average, reading, and writing scores. In this model, male math scores were positively and significantly trending.

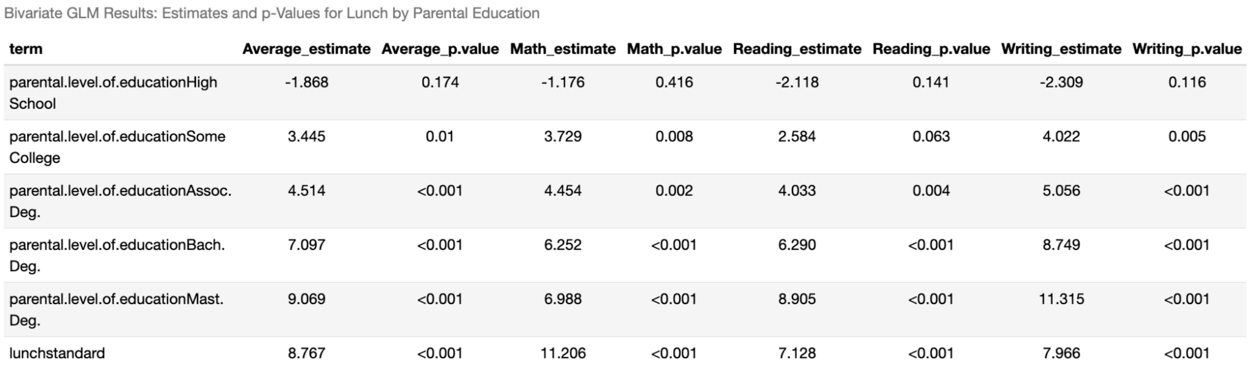


Bivariate GLM Results: Estimates and p-Values for Race/Ethnicity by Parental Education

term	Average_estimate	Average_p.value	Math_estimate	Math_p.value	Reading_estimate	Reading_p.value	Writing_estimate	Writing_p.value
parental.level.of.educationHigh School	-2.083	0.144	-1.391	0.358	-2.360	0.109	-2.499	0.098
parental.level.of.educationSome College	2.799	0.043	2.935	0.046	2.039	0.153	3.424	0.02
parental.level.of.educationAssoc. Deg.	3.884	0.005	3.677	0.013	3.449	0.017	4.526	0.002
parental.level.of.educationBach. Deg.	6.462	<0.001	5.443	0.002	5.731	<0.001	8.211	<0.001
parental.level.of.educationMast. Deg.	7.670	<0.001	5.357	0.015	7.758	<0.001	9.895	<0.001
race.ethnicitygroup B	2.582	0.145	1.873	0.319	2.797	0.126	3.077	0.101
race.ethnicitygroup C	3.620	0.029	2.383	0.176	3.950	0.021	4.528	0.01
race.ethnicitygroup D	5.601	<0.001	5.268	0.003	4.791	0.006	6.745	<0.001
race.ethnicitygroup E	8.922	<0.001	11.452	<0.001	7.586	<0.001	7.727	<0.001

Table 4. Bivariate Generalized Linear Model Regression Results for Average Score vs Parental Education and Race/Ethnicity

When incorporating race/ethnicity into the base parental education model, we observed similar trends, although p-values for the levels of education increased. Importantly, these changes did not alter the significance of the parental education levels relative to the 0.05 threshold. Within the racial groups, we did not identify significant trends for group B across any of the scores when combined with parental education. However, in the model, race/ethnicity group C showed a significant positive association with average, reading, and writing scores. Additionally, race/ethnicity groups D and E exhibited significantly positive trends across all scores.

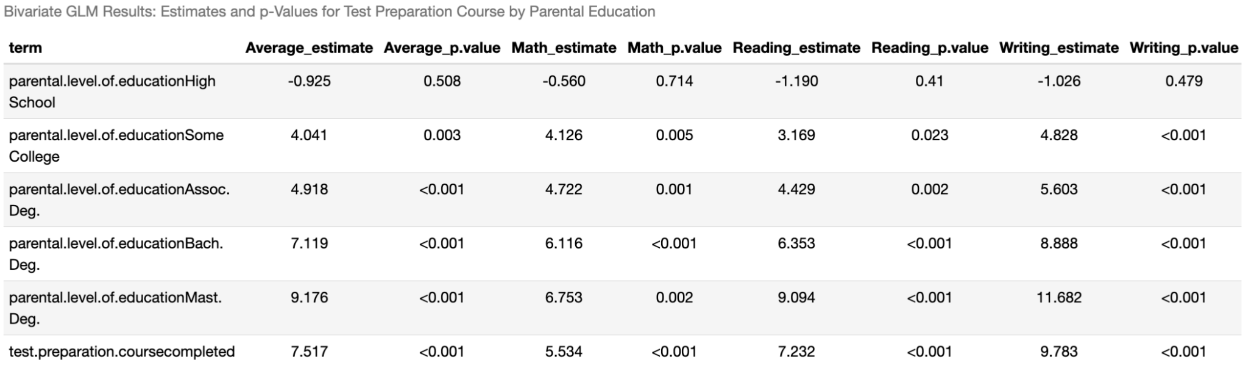


Bivariate GLM Results: Estimates and p-Values for Lunch by Parental Education

term	Average_estimate	Average_p.value	Math_estimate	Math_p.value	Reading_estimate	Reading_p.value	Writing_estimate	Writing_p.value
parental.level.of.educationHigh School	-1.868	0.174	-1.176	0.416	-2.118	0.141	-2.309	0.116
parental.level.of.educationSome College	3.445	0.01	3.729	0.008	2.584	0.063	4.022	0.005
parental.level.of.educationAssoc. Deg.	4.514	<0.001	4.454	0.002	4.033	0.004	5.056	<0.001
parental.level.of.educationBach. Deg.	7.097	<0.001	6.252	<0.001	6.290	<0.001	8.749	<0.001
parental.level.of.educationMast. Deg.	9.069	<0.001	6.988	<0.001	8.905	<0.001	11.315	<0.001
lunchstandard	8.767	<0.001	11.206	<0.001	7.128	<0.001	7.966	<0.001

Table 5. Bivariate Generalized Linear Model Regression Results for Average Score vs Parental Education and Lunch Program Participation

The bivariate model of parental education levels and lunch participation programs reveals similar trends and significance for levels of education and scores. We find that the standard, non-reduced, lunch participants had significantly better scores than their counterparts. Since lunch program eligibility is typically based upon family income, this could create a link to show how poverty and socioeconomic status can contribute to student success, alongside parental education level.



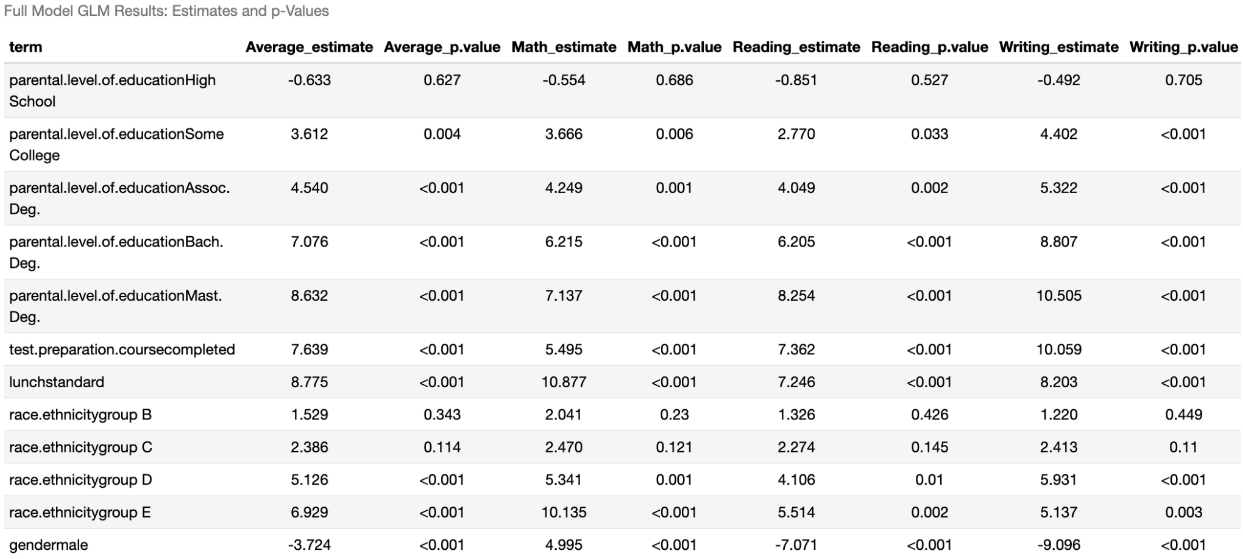
Bivariate GLM Results: Estimates and p-Values for Test Preparation Course by Parental Education

term	Average_estimate	Average_p.value	Math_estimate	Math_p.value	Reading_estimate	Reading_p.value	Writing_estimate	Writing_p.value
parental.level.of.educationHigh School	-0.925	0.508	-0.560	0.714	-1.190	0.41	-1.026	0.479
parental.level.of.educationSome College	4.041	0.003	4.126	0.005	3.169	0.023	4.828	<0.001
parental.level.of.educationAssoc. Deg.	4.918	<0.001	4.722	0.001	4.429	0.002	5.603	<0.001
parental.level.of.educationBach. Deg.	7.119	<0.001	6.116	<0.001	6.353	<0.001	8.888	<0.001
parental.level.of.educationMast. Deg.	9.176	<0.001	6.753	0.002	9.094	<0.001	11.682	<0.001
test.preparation.coursecompleted	7.517	<0.001	5.534	<0.001	7.232	<0.001	9.783	<0.001

Table 6. Bivariate Generalized Linear Model Regression Results for Average Score vs Parental Education and Test Preparation Course Completion

The last bivariate model of test preparation course completion and education level increased the significance of the "Some College" factor. We also found that not completing the test preparation had a significantly increasing effect on student scores. This shows that test preparation, in conjunction with parental education levels, has a significant relationship to student achievement.

The bivariate analyses demonstrate that parental education levels significantly impact student scores. This relationship is further supported and influenced by introducing demographic factors such as lunch program participation, race/ethnicity, gender, and test course completion. Each variable interacts uniquely with parental education levels, which could reveal insight into how student success is achieved. Since these all interact with each other in different ways, while remaining significant, we created a model with all of these variables to estimate how all of these variables work together to predict student scores.



Full Model GLM Results: Estimates and p-Values

term	Average_estimate	Average_p.value	Math_estimate	Math_p.value	Reading_estimate	Reading_p.value	Writing_estimate	Writing_p.value
parental.level.of.educationHigh School	-0.633	0.627	-0.554	0.686	-0.851	0.527	-0.492	0.705
parental.level.of.educationSome College	3.612	0.004	3.666	0.006	2.770	0.033	4.402	<0.001
parental.level.of.educationAssoc. Deg.	4.540	<0.001	4.249	0.001	4.049	0.002	5.322	<0.001
parental.level.of.educationBach. Deg.	7.076	<0.001	6.215	<0.001	6.205	<0.001	8.807	<0.001
parental.level.of.educationMast. Deg.	8.632	<0.001	7.137	<0.001	8.254	<0.001	10.505	<0.001
test.preparation.coursecompleted	7.639	<0.001	5.495	<0.001	7.362	<0.001	10.059	<0.001
lunchstandard	8.775	<0.001	10.877	<0.001	7.246	<0.001	8.203	<0.001
race.ethnicitygroup B	1.529	0.343	2.041	0.23	1.326	0.426	1.220	0.449
race.ethnicitygroup C	2.386	0.114	2.470	0.121	2.274	0.145	2.413	0.11
race.ethnicitygroup D	5.126	<0.001	5.341	0.001	4.106	0.01	5.931	<0.001
race.ethnicitygroup E	6.929	<0.001	10.135	<0.001	5.514	0.002	5.137	0.003
gendermale	-3.724	<0.001	4.995	<0.001	-7.071	<0.001	-9.096	<0.001

Table 7. Generalized Linear Model Regression Results for Average, Writing, Math, and Reading Scores for Parental Level of Education, Lunch Program Participation, Test Preparation Course Completion, Race/Ethnicity, and Gender.

In this final, full GLM model, we find a significant increase in scores in every level of education except for “High School”, compared to our baseline of “Some High School”. Lunch program participation, test preparation course completion, and racial groups D and E have significantly increasing relationships with student scores in this full model. Males still have significantly decreasing trends in scores compared to females across the board. Racial groups B and C had increasing, non-significant, trends in student scores.

With a full model developed and analyzed, we checked for multicollinearity using VIF. There was no evidence of multicollinearity between any of the variables. With no multicollinearity found, we then used a stepwise approach to create the best-fit model using the Akaike Information Criterion. This process returned the initial full model as the best-fit model of student score trends. With this, we can say that parental education, alongside additional demographic variables, provides significant insight and trend estimations of student scores in reading, writing, and math.

The results of our analysis provide enough evidence to reject both null hypotheses. The primary null hypothesis, which states that parental education levels have no significant effect on student performance across math, reading, and writing scores, is disproven by consistent findings showing that higher parental education levels are associated with significantly better test scores. Similarly, the secondary null hypothesis, which claims that factors such as gender, race/ethnicity, lunch program participation, and test preparation course completion do not influence the relationship between parental education and student performance, is also rejected. Our multivariate analysis shows that these factors not only interact with parental education but also independently affect student performance, emphasizing the complex and multifaceted nature of academic success

Experimentation

With a fully optimized multivariate model, alongside a simpler univariate model, the team was interested in seeing how this information can be used to predict student performance. With these models in place, we designed a bootstrap experiment to create robust estimators that could accurately predict student pass rates. The team utilized a logistic regression, instead of the measure of continuous student scores, to predict if students pass or not. This was accomplished by first creating a binary indicator for each score, determining if a student had gotten a 70% or higher within the class. From here, a bootstrap loop, with 1000 iterations was created. In each iteration, a new dataset of 1000 rows was generated by sampling with replacement from the original data. Logistic regression models were then fit to each of these samples. All of the coefficients of these models over the bootstrap iterations were then averaged to create the bootstrap estimates for each of the model parameters. This was completed for a univariate model between student average, math, writing, and reading scores and parental education levels, as well as the full model. The bootstrap models were then utilized to predict the probability of each student passing or not passing, based on either parental education as a standalone predictor or in combination with all the available variables.

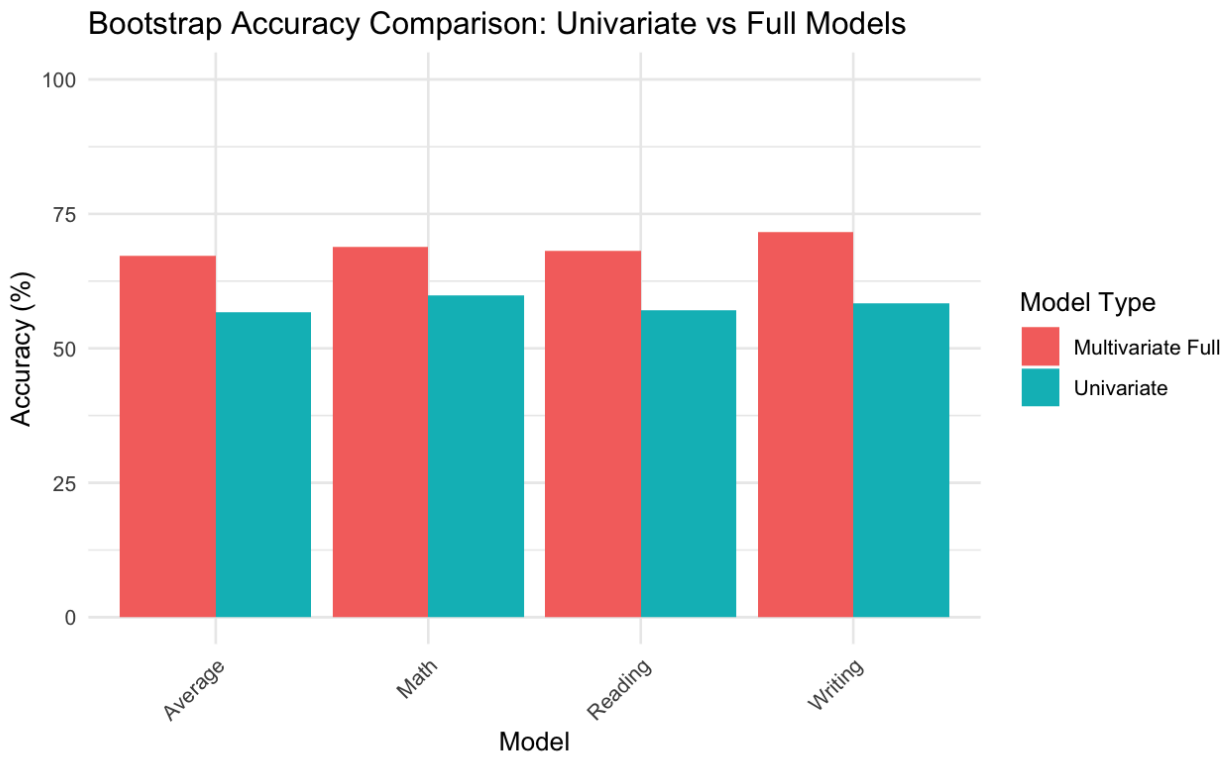


Fig 4. Bootstrap Accuracy

The bootstrap results, as shown in the figure, compare the accuracy of univariate and full multivariate models in predicting student performance across average, math, reading, and writing scores. The full multivariate model consistently outperformed the univariate model across all categories. For average scores, the multivariate model achieved an accuracy of 67.2%, compared to 56.8% for the univariate model. In math, the multivariate model reached 68.8% accuracy, while the univariate model lagged at 59.8%. Similarly, for reading scores, the multivariate model demonstrated an accuracy of 68.1%, significantly higher than the univariate model's 57%. The most pronounced difference was observed in writing, where the multivariate model achieved 71.6% accuracy compared to 57% for the univariate model. These results show how a more complete model can help predict student performance, compared to a univariate model.

Conclusion and Discussion

Our findings demonstrate that parental education levels have a significant and measurable impact on student academic performance across math, reading, and writing. Through comprehensive statistical analysis and bootstrap validation, we found that higher levels of parental education consistently correlate with better student outcomes, even when controlling for other demographic and socioeconomic factors. This relationship proved especially strong for parents with bachelor's and master's degrees, where students showed marked improvements in all three subject areas. With this, we can say that a parent's educational background can strongly influence how students perform in school. This can be helpful for educators, policymakers, and parents in understanding and ensuring student success.

However, the story is more complex than a simple direct relationship. Our secondary findings revealed that other factors, including gender, race/ethnicity, lunch program participation, and test preparation course completion, all play crucial roles in student success. Particularly noteworthy was the finding that test preparation courses showed a substantial positive impact regardless of parental education level, suggesting a potential avenue for intervention to support students from all backgrounds. This was further evidenced by the full model being the best fit for our data and most efficient in predicting student scores when subjected to a bootstrap. While the single variable of parental education can give educators and policymakers significant insight into what drives student success, the combined influence of multiple factors can create a more comprehensive understanding.

Our bootstrap analysis, which included 1,000 iterations predicting pass/fail outcomes based on a 70-point threshold, revealed varying levels of predictive accuracy across different subjects. While writing scores showed the highest predictive accuracy at around 72%, followed closely by math and reading at 70%, the average score model demonstrated lower accuracy at approximately 55%. These results suggest that while our metrics are valuable indicators of student performance, they shouldn't be treated as definitive predictors of student success or failure. The complexity of academic achievement extends beyond what can be captured in numerical scores alone, highlighting the importance of considering multiple factors when evaluating student potential.

Future research should explore several key areas: First, longitudinal studies tracking student performance over time could provide insights into how the influence of parental education evolves throughout a student's academic career. Second, investigating the specific mechanisms through which parental education affects student performance – such as homework help, study habits, or academic expectations – could help develop more targeted interventions. The insights gained from our bootstrap analysis could be particularly valuable in designing these interventions, as they suggest that subject-specific approaches might be more effective than general academic support. Future studies could focus on understanding why writing scores showed higher predictive accuracy and how this knowledge could be applied to improve pass rates across all subjects. Finally, examining the effectiveness of various support programs in mitigating the effects of lower parental education levels could provide valuable guidance for educational policymakers and administrators. The underlying socioeconomic dynamics of variables such as lunch program participation and race/ethnicity could provide further insight into why this specific factor could be driving student achievement. Further investigation could reveal bias or mismatch between students of different groups and help develop more equitable educational strategies that improve pass rates for all demographic groups.

References

[1] Y. Sun, H. Wang and C. Yang, "Exploring the impact of parental involvement on student academic achievement in China," 2023 13th International Conference on Information Technology in Medicine and Education (ITME), Wuyishan, China, 2023, pp. 237-241, doi: 10.1109/ITME60234.2023.00056.

[2] M. Mathakiya and S. Verma, "Association between Students English Proficiency Exam Score and their Parental Education Level Using Chi-Square Test," 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2023, pp. 1538-1542, doi: 10.1109/ICPCSN58827.2023.00256.

[3] M. Karmagatri, D. Kurnianingrum, M. R. Suciana and S. Aulia Utami, "Predicting Factors Related to Student Performance Using Decision Tree Algorithm," 2023 5th International Conference on Cybernetics and Intelligent System (ICORIS), Pangkalpinang, Indonesia, 2023, pp. 1-6, doi: 10.1109/ICORIS60118.2023.10352269.

[4] A. H. Hoti and X. Zenuni, "Factors influencing student academic performance and career choices," 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkiye, 2024, pp. 1-8, doi: 10.1109/IDAP64064.2024.10710702.