Quantitative Analysis of the Effect of Household Income on Students' Reading and Math Scores with the Baseline of General Knowledge Scores

1. Introduction

In recent years, the educational community has shown increasing interest in the relationship between family socioeconomic status and the academic performance of early childhood. This interest is fueled by growing evidence suggesting that children's early academic experiences can significantly influence their long-term educational trajectories (Kern, 2008). Against this backdrop, this report aims to dissect the nuances of how socioeconomic factors influence the academic performance of kindergarten students over the course of the school year.

This report conducts an exploratory data analysis using data from the Early Childhood Longitudinal Study (INF2178_A3_data.csv) conducted during the 1998-1999 school year. This dataset records kindergarten students' reading, math, and general knowledge scores at two key moments (fall 1998 and spring 1999), as well as information about their family income. The income data allow us to classify students into different socioeconomic groups, thereby providing a lens through which to examine the impact of socioeconomic status on educational performance.

This analysis will discuss one research question, serving as a guide to finding out patterns of students' scores in relation to household income:

1. **Research Question:** How does household income affect students' reading, math, and general knowledge scores?

2. Data Cleaning and Data Wrangling

The raw dataset contains a total of 9 columns and 11,933 entities(rows). Through an initial examination of the dataset, the number of nulls in the data is 0, but "total household income" and "income in thousands" have the same effect on the content of this study. Therefore, we will discard the redundant variable, "income in thousands," and use only "total household income" for the next step of analysis.

From the examination, we find that "incomegroup" is a variable indicating the income level of the respondent, which includes categories such as "level 1", "level 2", and "level 3". Therefore, "incomegroup" should be treated as a categorical variable; otherwise, the model

will incorrectly assume it is a continuous variable. This assumption would mean that the differences between income groups are treated as linear, which is not accurate.

3. EDA

Data Visualization

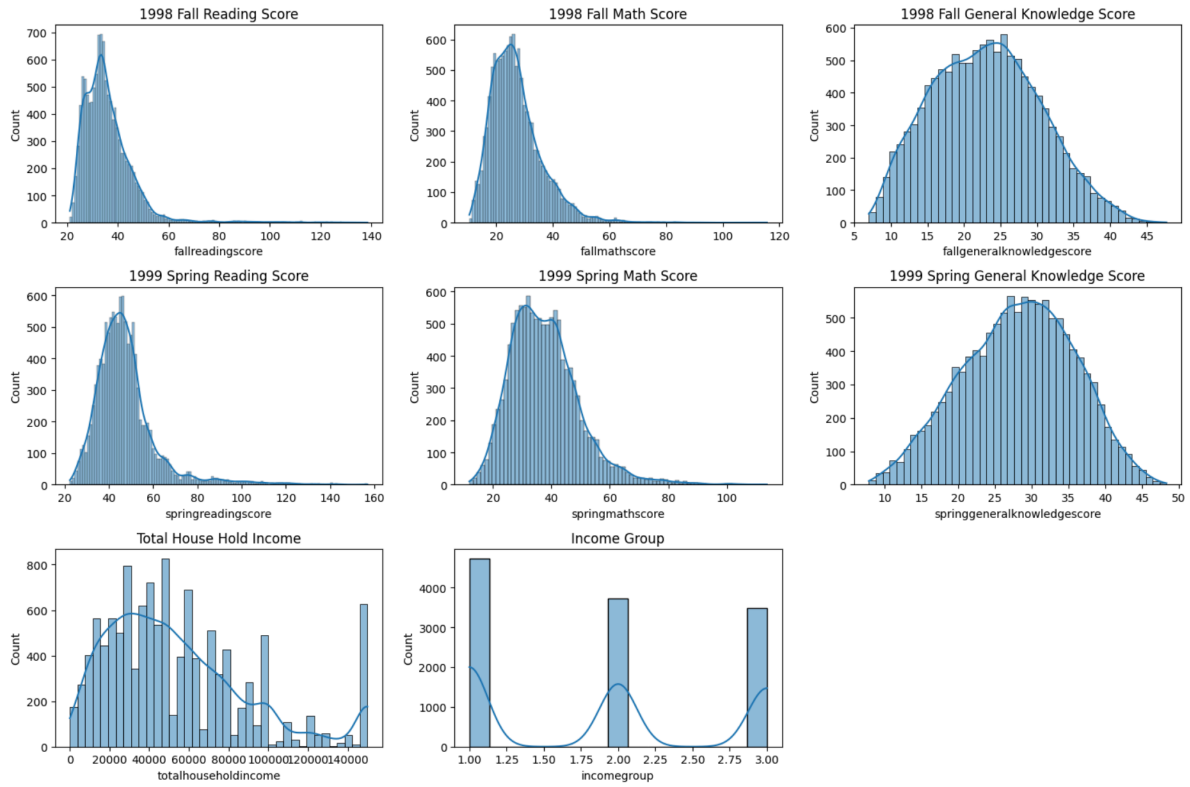

Figure 1: Exploratory Data Analysis

Based on Figure 1, it is observed that the scores for reading, mathematics, and general knowledge in fall 1998 and spring 1999 were generally normally distributed, albeit with potential outliers as evidenced by the long tails in their distributions. The overall scores for these early childhood tests have seen a significant increase, with the distribution graphs moving to the right when compared to those from 1998 and 1999.

The distribution of household income is right-skewed, which suggests that a majority of households fall into the lower-middle income bracket, with fewer households situated in the higher-income brackets.

Furthermore, income groups are distinctly divided into three levels, indicating variation in the sample across different income groups. The majority of the data is concentrated in the low- and middle-income categories, highlighting a clear disparity in income distribution within the sample.

4. ANCOVA

We will explore the impact of different income groups (independent variables) on students' reading and math test scores (dependent variables) over time. To analyze the data, we plan to use two separate one-way ANCOVAs, examining the changes in reading and math scores individually to determine the effect of income groups on student performance over time. We will control for the fall general knowledge score as a covariate to account for the initial state. This approach allows us to measure changes in student performance in reading and math from fall to spring more accurately, as it helps reduce error variance and prevents the influence of varying general knowledge scores on students' outcomes.

Model Fitness Testing

Before conducting ANCOVA analysis, we will first test the necessary assumptions, including the normality of residuals and the homogeneity of variances.

Normality of residuals

w	0.8996317386627197
p-value	<0.001

Table 1: Shapiro-Wilk Test for Reading Score

w	0.9664044380187988
p-value	<0.001

Table 2: Shapiro-Wilk Test for Math Score

For the changes in reading and math scores, as shown in Table 1 and Table 2, the Shapiro-Wilk test indicated non-normality in the residuals (both p-values < 0.000001). However, a warning message stated, "p-value may not be accurate for N > 5000." This implies that for data points exceeding 5000, the p-value of the Shapiro-Wilk test might not be accurate. In large samples, minor deviations from normality can be deemed acceptable.

Homogeneity of variances

statistic	39.552829596478
p-value	7.570499656193536e-18

Table 3: Levene's Test for Reading Score

statistic	18.899850238111785
p-value	6.380804527269148e-09

Table 4: Levene's Test for Math Score

The results of Levene's test as shown in Table 3 and Table 4 indicate that the variances in both reading scores ($p < 0.001$) and math scores ($p < 0.001$) across different income groups violate the assumption of homogeneity of variances. This suggests that the variances of scores are not equal between groups, which may impact the reliability of ANCOVA results.

OLS Regression Model

In order to explore the impact of different income groups on students' reading and math test scores over time, controlling for fall general knowledge scores as a covariate, an OLS Regression Model is fitted, and here are the key findings:

Reading Score

	Coefficient	Standard Error	t statistic	p-value	[0.025	0.975]
Income Group 1 (Reference)	7.7308	0.242	31.960	<0.001	7.257	8.205
Income Group 2	0.2169	0.180	1.205	0.228	-0.136	0.570
Income Group 3	0.4038	0.191	2.110	0.035	0.029	0.779
General Knowledge	0.1578	0.011	14.836	<0.001	0.137	0.179

Table 5: OLS Regression Result for Reading Score

As shown in Table 5, the coefficient for income group 2 (0.2169) is not statistically significant (p-value = 0.228), indicating that income group 2 does not have a significant effect on improving reading scores compared to income group 1 (the reference group). In contrast, the coefficient for income group 3 (0.4038) is statistically significant (p-value = 0.035), suggesting that income group 3 is associated with improved reading scores compared to income group 1, after controlling for fall general knowledge scores.

The coefficient for the baseline fall general knowledge score (0.1578) is statistically significant ($p < 0.00001$), indicating that fall general knowledge scores significantly predict gains in reading achievement regardless of income group

Math Score

	Coefficient	Standard Error	t statistic	p-value	[0.025	0.975]
Income Group 1	5.9826	0.203	29.542	<0.001	5.586	6.380
Income Group 2	0.1523	0.151	1.011	0.312	-0.143	0.448
Income Group 3	0.1442	0.160	0.900	0.368	-0.170	0.458
General Knowledge	0.1993	0.009	22.385	<0.001	0.182	0.217

Table 6: OLS Regression Result for Math Score

The coefficient for income group 2 is 0.1523 with a p-value of 0.312. This indicates that belonging to income group 2 does not significantly affect improvements in mathematics performance compared to income group 1 (the reference group). Similarly, the coefficient for income group 3 is 0.1442 with a p-value of 0.368. Just like income group 2, belonging to income group 3 does not significantly affect gains in math performance compared to income group 1.

The coefficient of 0.1993 for the fall general knowledge score is statistically significant ($p < 0.00001$), suggesting that a 1-unit increase in the fall general knowledge score is associated with a corresponding increase of 0.1993 units in the math score.

Visualization

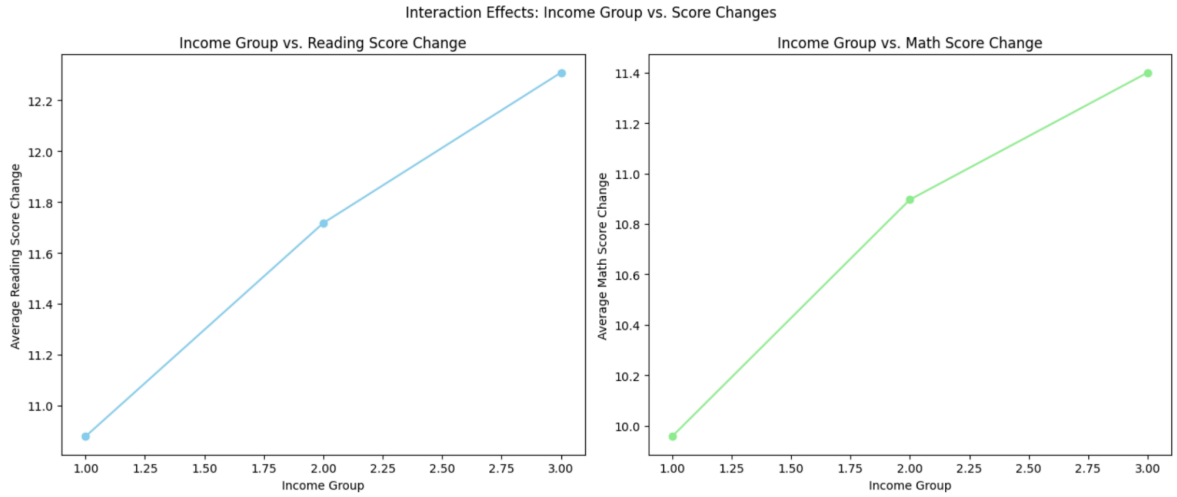

Figure 2: Interaction Plot for Reading Score Change and Math Score Change

As shown in Figure 2, the graphs offer a visual representation of the average changes in reading and math scores associated with students' income groups. Although the ANCOVA analysis indicates that there are no statistically significant differences in score changes by income group, these plots still offer a visual depiction of how score changes are distributed across income levels. However, the statistical analysis did not find these differences to be significant, suggesting that any observed differences in these plots should be interpreted with caution within the context of the broader analysis.

5. Conclusion

In conclusion, the ANCOVA analyses indicate that the income group has little direct effect on gains in reading and math scores, however, baseline general knowledge scores play an important role in academic development in these areas. This finding highlights the importance of early education and building a strong knowledge base for students' continued academic success. The same insignificant effect of the income group on score changes also reflects the complexity of the dynamics of educational achievement, in which factors such as school resources, teaching quality, and family environment may also play a significant role.

6. Reference

Kern, M. L., & Friedman, H. S. (2008). Early educational milestones as predictors of lifelong academic achievement, midlife adjustment, and longevity. Journal of applied developmental psychology, 30(4), 419–430. https://doi.org/10.1016/j.appdev.2008.12.025