1. Introduction

Educational attainment is influenced by various factors, including socioeconomic status. The relationship between a family's income and a child's educational performance is complex and multifaceted. In this study, exploratory data analysis and one-way ANCOVA are used to explore how income levels relate to educational outcome progression in reading, math, and general knowledge by comparing performance scores at two timestamps. This investigation seeks to evaluate whether educational achievements in children vary across income groups after accounting for baseline abilities.

2. Exploratory Data Analysis

2.1 Dataset Description

The dataset consists of scores from educational assessments of kindergarten children in three categories, reading, math and general knowledge, conducted in Fall 1988 and Spring 1999, as well as household income and income group categorized into 3 levels, comprising a total of 11,933 instances.

The fall reading scores exhibit a roughly normal distribution with a skew towards the lower end, indicating that a larger number of students scored lower on the reading assessment. The spring reading scores also follow a normal distribution but with a noticeable shift towards higher scores. This shift suggests a general improvement in reading abilities from Fall 1988 to Spring 1999..

Similar to reading, the fall math scores are distributed normally with a skewness towards the lower scores (Figure 2). In spring, the distribution maintains its normal shape, shifting slightly towards higher scores. This indicates an overall increase in math scores from fall to spring, though the shift is not as pronounced as in reading.

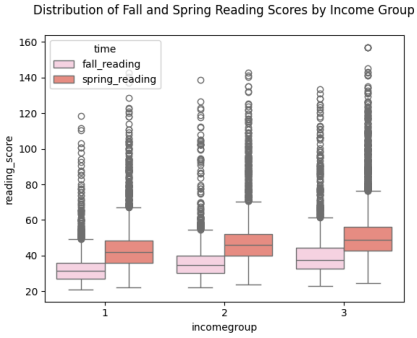


Figure 1. Distribution of Reading Scores

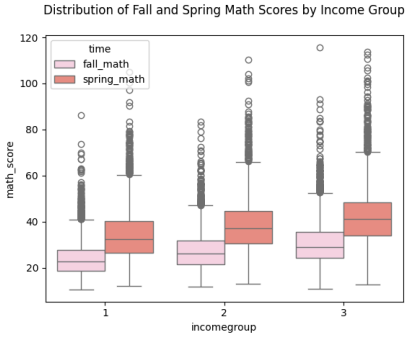


Figure 2. Distribution of General Knowledge

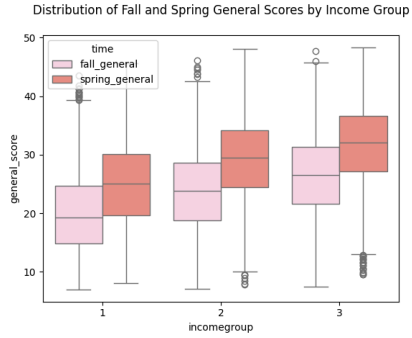


Figure 3. Distribution of General Knowledge

The distribution of fall general knowledge scores appears normal (Figure 3) but with a more pronounced peak than the reading and math scores, suggesting less variance among students in their general knowledge at the beginning of the year. The spring general knowledge scores also show a normal distribution with a slight increase in the mean score, suggesting an improvement in general knowledge over the academic year.

The distributions of all three subjects indicate an improvement from fall to spring. The histograms reveal that while the majority of students improve, the extent of this improvement varies across subjects, with reading scores showing the most noticeable enhancement. These visual findings will be further explored through statistical analysis to understand the factors contributing to these changes.

Household income has an overall skewed distribution (Figure 3), with a large standard deviation, a minimum of no household income and a maximum of $150K, significantly higher from the mean household income of $54.3K.

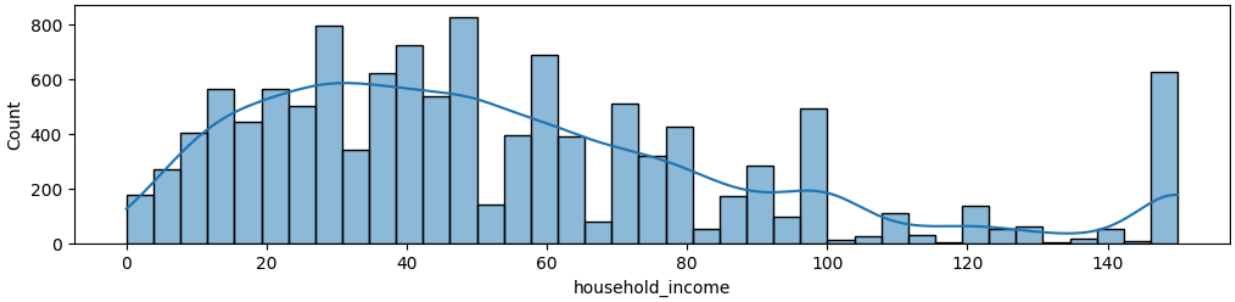


Figure 4. Distribution of Household Income

An overview of income distribution by income group shows that income group 1 includes families with household income from zero to 40k, income group 2 includes families with household income from 40k to 70k, income group 3 includes families with household income from 70k to over 150k, the biggest range among the three income groups.
A closer look at each distribution shows that although income distribution is roughly normal for the low to mid income groups, the highest income group has significant outliers.

Statistics Summary of Income by Income Group

	mean	median	min	max	count
Income group					
1	22.02	23.0	0.0	39.8	4729
2	51.74	50.0	40.0	69.7	3726
3	100.99	90.0	70.0	150.0	3478

3. Research Questions and Hypotheses

Research Question: Does household income level have a significant effect on children's score performance after controlling for their baseline scores?

H0 (Null Hypothesis): Household income level does not have a significant effect on children's score performance after controlling for their baseline scores.

Ha (Alternative Hypothesis): Household income level has a significant effect on children's score performance after controlling for their baseline scores.

4. Analysis

4.1 Overview

One-way Analysis of Covariance (ANCOVA) with interaction was used to determine if income groups have significant effects in children's academic performance across all three categories (reading, math, and general knowledge), after controlling for baseline scores captured in the fall.
Assumptions of linearity, normality, and homogeneity of variances were checked for one-way ANCOVA. The checks involved plotting observed versus predicted values, Q-Q plots for residuals, and Levene's test for equality of variances.

4.2 Reading

The One-way ANCOVA with spring reading score as dependent variable, fall reading score as covariate show the following result:

	sum_sq	df	F	PR(>F)
incomegroup	469.981	1	7.432	0.006
fall_reading	1547012.917	1	24463.318	0
incomegroup:fall_reading	301.795	1	4.772	0.029
Residual	754366.896	11929	NaN	NaN

- The F-statistic for the income group is 7.432 with a p-value of 0.006, indicating that there are statistically significant differences in spring reading scores across income groups when controlling for fall reading scores.
- The fall reading scores (covariate) are significantly associated with the spring reading scores with an F-statistic of 24633.32 and a p-value of < 0.001. This strong association confirms the importance of controlling for baseline ability when examining the impact of income groups.
- The interaction term has an F-value of 4.772 and a p-value of 0.029, which is statistically significant. This indicates that the relationship between baseline reading scores and spring reading scores differs by income group. There is an interaction effect at play, meaning the impact of a child's initial reading level on their subsequent reading score depends on their income group.

The significant effect of income groups after controlling for baseline scores in fall implies that socioeconomic factors which are captured in part by the income group, may influence reading outcomes. These could include access to resources, educational support, and other environmental factors associated with higher income levels. The significant interaction suggests that income-based disparities in reading achievement may be more pronounced for children with certain levels of initial reading ability. For instance, children from higher-income families with already good baseline reading scores may have more opportunities to advance their skills than peers from lower-income groups.

The below scatterplot of observed vs. predicted values (Figure 5) demonstrates a strong linear relationship, supporting the linearity assumption in ANCOVA. The normality of the residuals assumption for ANCOVA was confirmed using the Shapiro-Wilk test, with a statistic of 0.912 and a p-value of <0.001. This indicates that the residuals of our model are normally distributed, which was confirmed by the histogram of residuals (Figure 6). Homogeneity of variances was tested using Levene's test, which produced a statistic of 39.553 and p-value of <0.001.

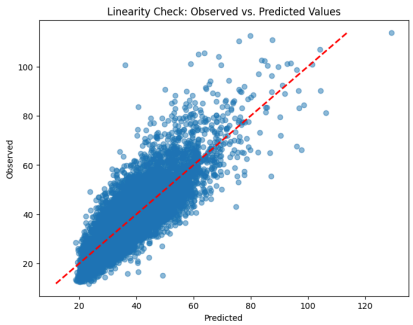


Figure 6. Observed vs. Predicted Value (Reading)

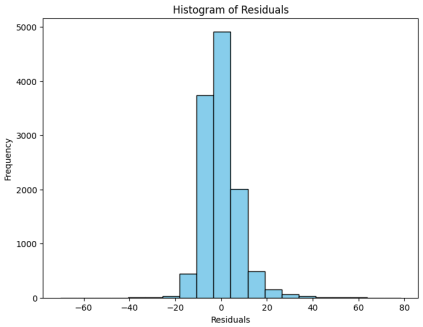


Figure 7. Histogram of Residuals (Reading)

Figure 8. Interaction Plot (Reading)

The interaction plot shows the relationship between fall reading scores and spring reading scores across three distinct income groups. The plotted regression lines for each income group appear parallel, suggesting that the slope of improvement from fall to spring is consistent across income groups, indicating that the effect of the fall reading score on the spring reading score is similar regardless of income group.

4.3 Math

The same process was repeated with spring **math** score as dependent variable, fall math score as covariate:

	sum_sq	df	F	PR(>F)
incomegroup	1600.557	1	34.719	0
fall_math	1026632.731	1	22269.515	0
incomegroup:fall_math	1680.414	1	36.451	0
Residual	549931.229	11929	NaN	NaN

- incomgroup has an F-value of 34.719 and a p-value of <0.001. This suggests that there are statistically significant differences in children's math performance across different income groups, even after accounting for the variance due to baseline math scores..
- For the covariate in the ANCOVA model, children's baseline math scores in fall, F-value is very high 22269.515, with a p-value of <0.001, showing that the baseline scores are a highly significant predictor of the spring math scores. This suggests that initial mathematical ability might be a powerful indicator of subsequent performance.
- The interaction term has an F-value of 4.772 and a p-value of 0.029, which is statistically significant. This indicates that the relationship between baseline math scores and spring math scores differs by income group. There is an interaction effect at play, meaning the impact of a child's initial math level on their subsequent math score depends on their income group.

The assessment of ANCOVA assumptions was checked first with a Quantile-Quantile plot (Appendix 1) supported the normality of residuals, despite slight deviations at the distribution tails hinting at outlier presence. A scatterplot of observed versus predicted math scores (Appendix 2) confirming a linear relationship, as data points clustered tightly along the reference line. The residuals' histogram (Appendix 3), centered around zero with minor skewness, further corroborated the normality assumption necessary for ANCOVA.

Figure 9. Interaction Plot (Math)

An interaction plot (Figure 9) for math scores across income groups presented lines with slightly varying slopes, indicating a small interaction effect on spring scores. This indicates that the relationship between the baseline math scores and the spring math scores is fairly consistent across different income groups, with the slope of the line slightly steeper for higher income groups between baseline (fall) and spring math scores, suggesting that for children with a higher baseline math score, their spring math scores increase more dramatically for students from higher-income groups compared to lower-income groups.

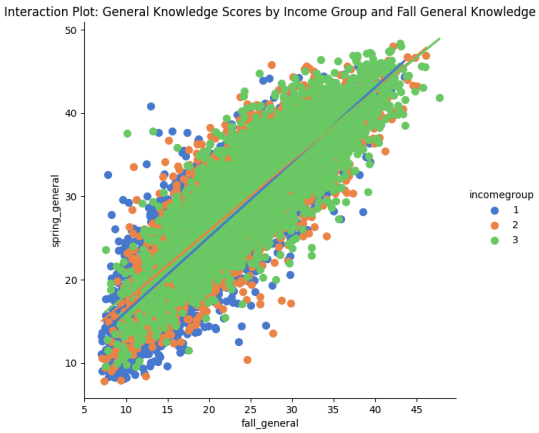
4.4 General Knowledge

One-way ANOVA results with spring general knowledge score as dependent variable, fall general knowledge score as covariate show the following result:

	sum_sq	df	F	PR(>F)
incomegroup	1614.073	1	104.936	0
fall_general	413465.233	1	26880.795	0
incomegroup:fall_general	797.862	1	51.872	0
Residual	183485.15	11929	NaN	NaN

- incomegroup has an F-value of 104.936 and a p-value of < 0.001, suggesting that income level has a distinct impact on children's knowledge acquisition. This affirms the research hypothesis that household income level significantly affects children's performance in general knowledge after controlling for their baseline scores.
- The 'fall_general' term, which represents the baseline general knowledge scores, has an F-value of 26880.795 with a p-value of 0.001, indicating that the baseline general knowledge scores are a very strong predictor of the spring scores and underscores the importance of children's initial knowledge level in their subsequent academic performance.
- The interaction term has an F-value of 51.872 with a p-value of < 0.001, suggesting that the relationship between baseline knowledge and the improvement in scores is not uniform across income levels. This implies that children from different income groups may benefit differently from their initial levels of general knowledge.

To check the validity of ANCOVA assumptions for this model, the Q-Q plot (Appendix) displays the residuals' alignment with the theoretical normal distribution line, deviating only slightly at the tails. a scatter plot (Appendix 4) showed a concentration of data points along the dashed line and confirmed a strong linear association, fulfilling the assumption of linearity. This deviation might suggest the presence of outliers or slight non-normality, but generally supports the assumption that residuals are normally distributed. A scatter plot (Appendix 5) showed a concentration of data points along the dashed line and confirmed a strong linear association, fulfilling the assumption of linearity. The histogram (Appendix 6) shows the residuals distributed around the central value, mostly symmetrical, suggesting normality. There is some evidence of slight skewness, which is common in real-world data and does not necessarily violate the normality assumption if not extreme.



The interaction plot indicates the slopes for the regression lines across income groups are close to parallel, implying minimal interaction between income groups and baseline general knowledge on spring scores. This suggests the effect of baseline knowledge on the spring scores is consistent across different income groups.

In summary, the assumption checks for ANCOVA were satisfactorily met. The ANCOVA model results suggest that household income level has a significant impact on children's general knowledge performance, even after accounting for their baseline knowledge. Further investigation

Figure 10. Interaction Plot (General Knowledge)

reveals that general knowledge consists of science and social studies knowledge. This implied the influence of socioeconomic factors on educational outcomes and the necessity of considering such factors in educational strategies and policy-making.

5. Findings

To investigate the effect of household income on children's performance across educational assessment in reading, math, and general knowledge, the ANCOVA results from controlling for respective baseline scores in all three categories must be evaluated individually and holistically, while taken into consideration of the soundness of assumptions such as linearity, normality, and the homogeneity of variances.

In reading, significant disparities in scores were observed across income groups, corroborated by a statistically significant interaction effect, indicating that household economics status plays a crucial role in reading achievements. The analysis for math reflected similar patterns of income-related differences, with both income group and baseline math scores demonstrating substantial effects on spring scores. A minor interaction effect was noted, suggesting a slight variance in progress attributable to income levels, particularly for those with higher baseline scores. For general knowledge, comprising science and social studies, the effects of income groups were significant, and a notable interaction with baseline scores was identified. The baseline general knowledge scores were highly predictive of spring scores, reinforcing the premise that early foundational knowledge in science and social studies is vital for later academic success.

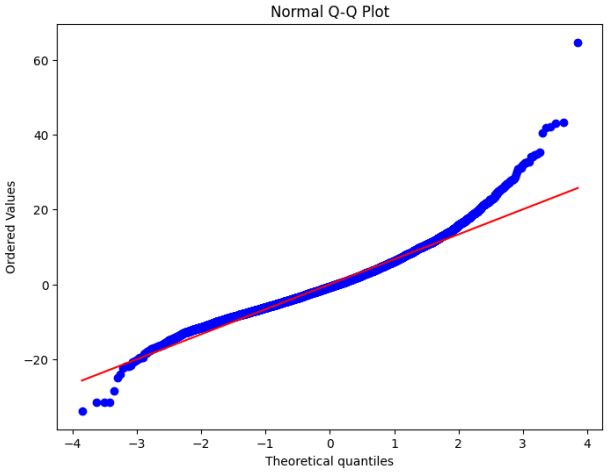
Throughout the analyses, the foundational assumptions of ANCOVA held firm, with data points closely aligning along the reference lines in the scatter plots, residuals distributed normally in histograms, and Q-Q plots showing only minor deviations from theoretical quantiles, thus validating the robustness of the models used.

6. Conclusion

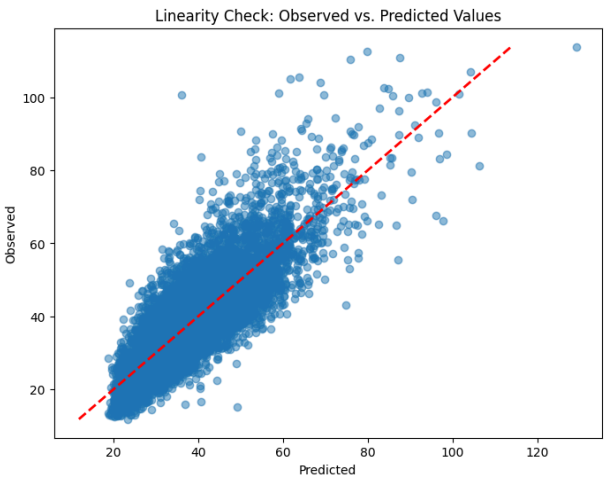
The findings of this study highlight the compelling implication of household income on children's educational outcomes, on top of the influence of their initial academic abilities. From the data, we have learned that household economic status, as measured by income group, significantly affects educational outcomes across children’s reading, math and general knowledge performance progression, even when baseline proficiencies are accounted for. The notable interaction effects further suggest that the relationship between baseline abilities and academic growth is influenced and compounded by socioeconomic status, with potential implications for educational interventions aimed at leveling the playing field. The disparities observed across income groups and the interaction of these groups with baseline knowledge necessitate a multi-faceted approach to educational policy and practice. Tailored interventions that take into account not only economic inequalities but also the initial academic standings of children are essential to mitigate the identified achievement gaps. This study underscores the need for strategic educational planning and resource allocation that foster equitable learning opportunities for all children, irrespective of their household income levels.

However, it is important to note some of the limitations of this analysis. ANCOVA, by design, can control for covariates but cannot determine causal relationships and therefore does not establish causation but rather identifies correlations. Further study should be conducted by 1. taking into consideration the longitudinal study design that tracks long-term progress over-time, and 2. employing methods like propensity score matching to control for confounding variables that might correlate or affect both household socioeconomic status and educational outcomes.

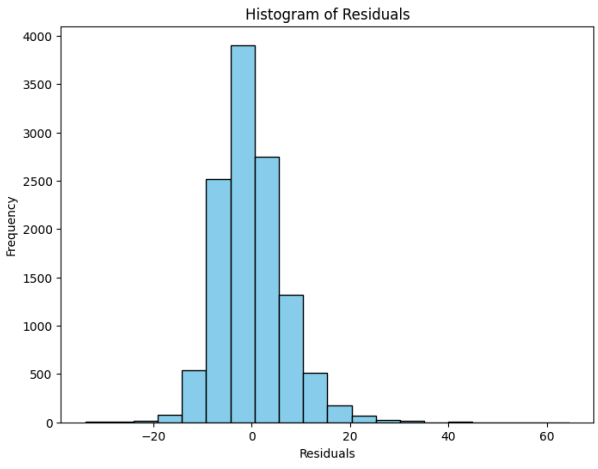
Appendix



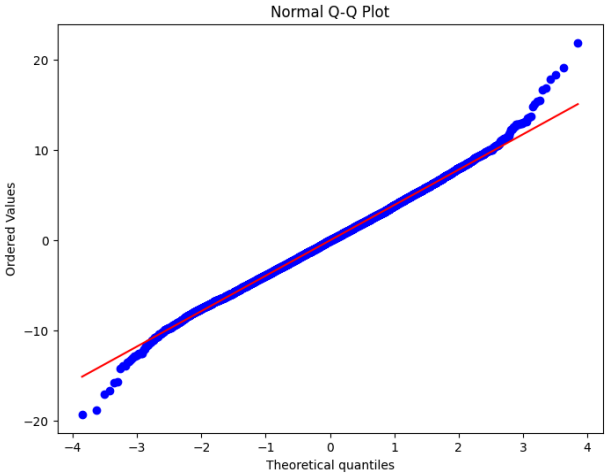
Appendix 1. Q-Q Plot (Math)



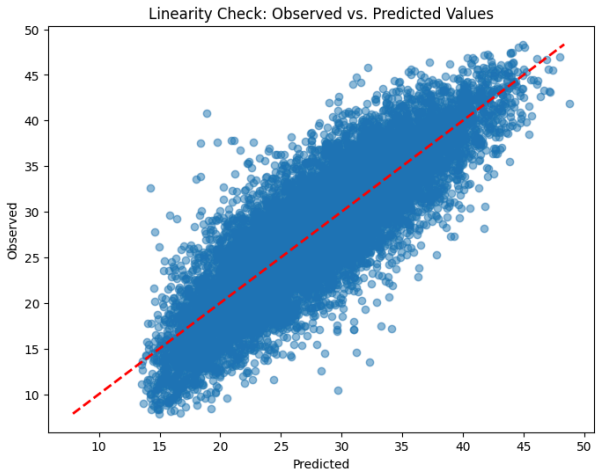
Appendix 2. Scatterplot (Math)



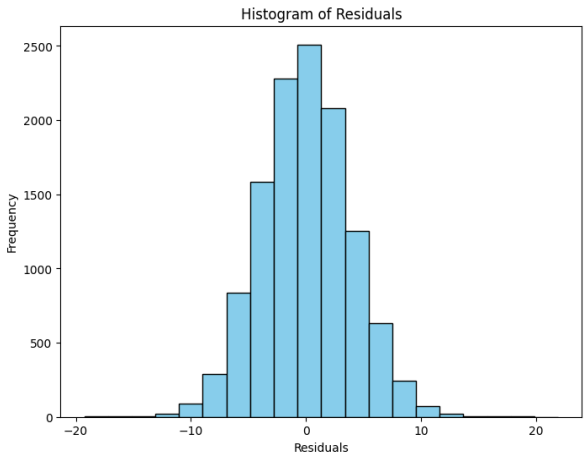
Appendix 3. Histogram of Residuals (Math)



Appendix 4. Q-Q Plot (General Knowledge)



Appendix 5. Scatterplot (General Knowledge)



Appendix 6. Histogram of Residuals (General Knowledge)