Exploring the Impact of Income Levels on Kindergarten Academic Performance: An ANCOVA Approach

Jiachen Liu-INF2178-A3

1.Introduction

To better understand the dynamics between income levels and educational outcomes, this study focuses on an analysis of kindergarten students' academic performances across various disciplines. Utilizing data from an early childhood longitudinal study conducted in 1998-99, we examine the relationships between income groups and changes in reading, math, and general knowledge scores over an academic year.

The essence of this research lies in its quantitative approach, leveraging Analysis of Covariance (ANCOVA) to discern the extent to how income groups influence educational achievements. By controlling for initial knowledge levels and examining score differences from fall to spring, the study aims to provide a nuanced understanding of how economic backgrounds shape educational trajectories from an early age.

Through this analysis, we seek to answer this key research question:

- How does income level affect the changes in kindergarten students' reading and math scores over an academic year, after controlling for their general knowledge scores?

2.Data Cleaning

The initial phase of the analysis involved careful data cleaning to ensure the reliability of the results. The dataset, titled "INF2178_A3_data.csv," required several cleaning procedures:

a. Transforming datatype of "incomegroup" column from integer to category. This transformation helped reflect this column's nature more appropriately.
b. Introducing two new calculated columns to the dataset – the differences in math and reading scores between the fall and spring measurements. This step was important in not only enhancing the dataset's richness for our analyses but also in ensuring models could accurately account for the variances in academic achievement over time.

Through these preparatory measures, we ensured that this dataset was primed for a robust and insightful exploration into the impact of income levels on early educational outcomes, setting a solid foundation for the analytical journey ahead.

3.EDA

The EDA phase of this report serves as the foundation for understanding the underlying patterns and distributions within this dataset, setting the stage for more complex statistical analyses. This crucial phase involved four key tasks:

a. Examined summary statistics for each column: This provided a comprehensive overview of the central tendencies, dispersion, and shape of the dataset's distributions, offering insights into the overall characteristics of the academic scores and income groups.
b. Analyzed the counts for each income group: By visualizing the distribution of students across different income groups, this task helped in understanding the income level landscape of the sample population.
c. Investigated how each type of score is spread out across the three income groups: Through this analysis, I explored the distribution of reading, math, and general knowledge scores within each income category, aiming to identify patterns or disparities in academic performance related to different income groups.
d. Explored changes in reading and math scores by income group using boxplots: This task focused on visualizing the academic progress of students, from fall to spring, across different income groups, providing a visual representation of income groups variations might influence educational advancement over the academic year.

3.1.Analysis of Figure 3.1: Counts for each income group

The bar chart depicting the counts for each income group reveals that the distribution of students across income groups is not uniform. Income group 1, which represents the lowest income category, has the highest count with 4729 students. This is followed by income group 3, with 3478 students, and income group 2 has the lowest count, with 3726 students. The differences in group sizes could have implications for the statistical analyses that follow, especially if the income groups are used as independent variables.

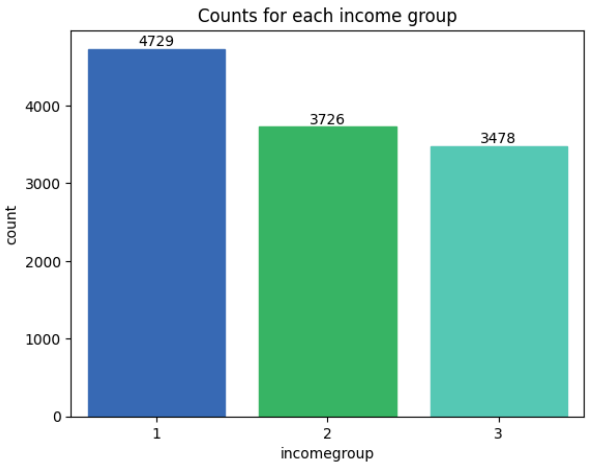

Figure 3.1 Counts for each income group

3.2.Analysis of Figure 3.2-3.7: Boxplots for Each Type of Scores

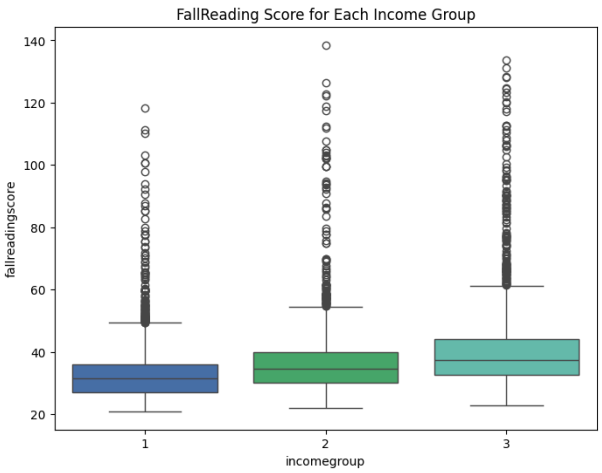

Figure 3.2. Fall Reading Score for Each Income Group in Box Plots

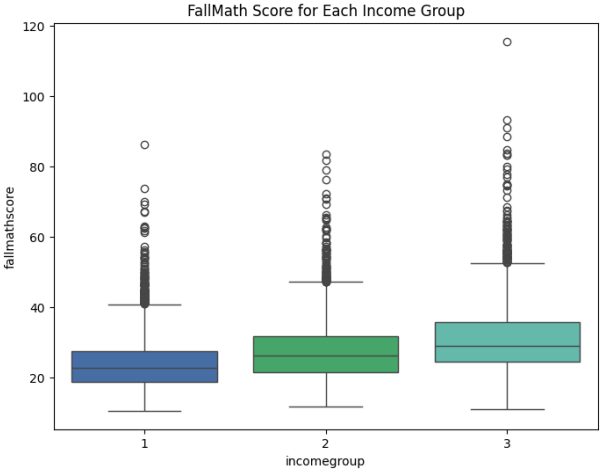

Figure 3.3. Fall Math Score for Each Income Group in Box Plots

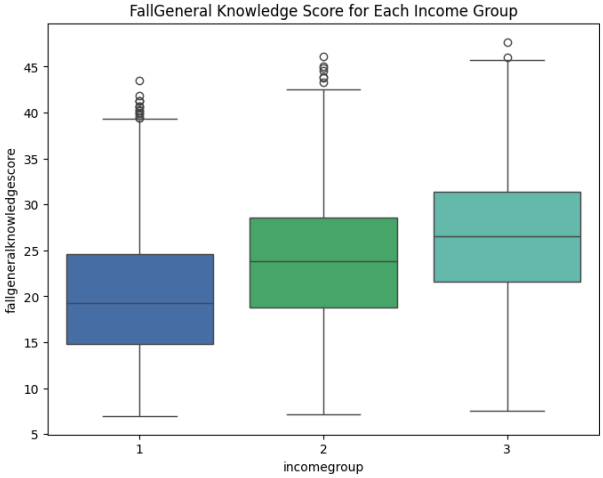

Figure 3.4. Fall General Knowledge Score for Each Income Group

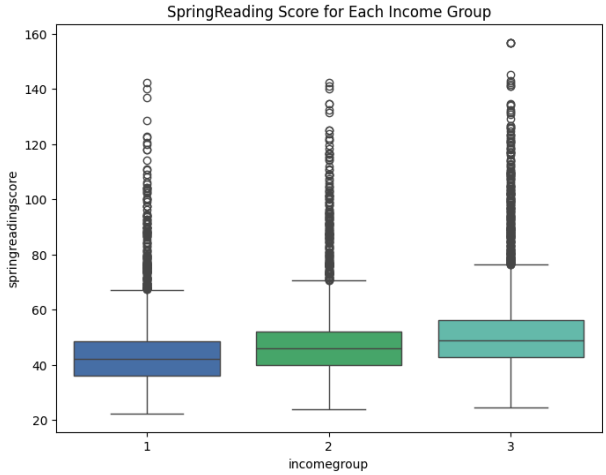

Figure 3.5. Spring Reading Score for Each Income Group in Box Plots

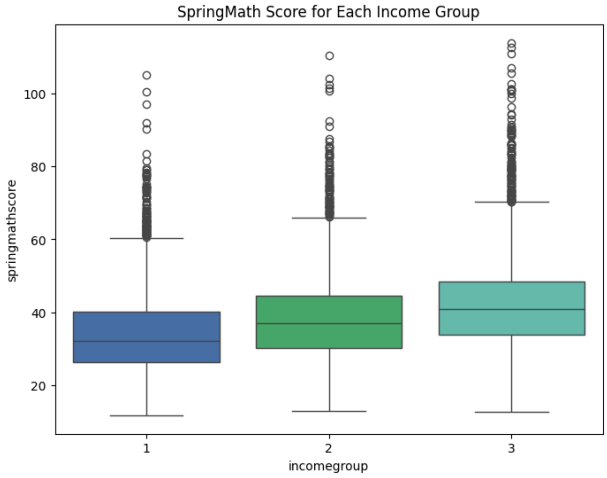

Figure 3.6. Spring Math Score for Each Income Group in Box Plots

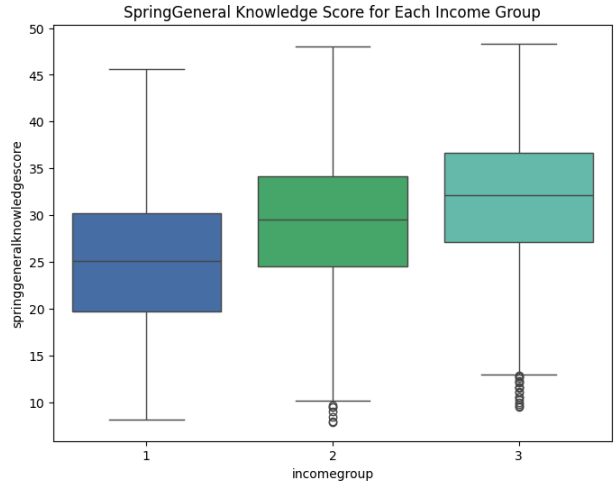

Figure 3.7. Spring General Knowledge Score for Each Income Group

The boxplots for the fall and spring reading, math, and general knowledge scores across the three income groups (as presented in Figure 3.2-3.7 respectively) provide a visual representation of the distribution of scores and offer some key insights.

All the scores in both Fall and Spring indicate a positive association with income groups. The median scores appear to increase from income group 1 to 3, suggesting that students from higher-income families may start and end the school year with better scores in reading, math, and general knowledge. Notably, there are a significant number of outliers, especially in the higher income groups, which could represent exceptional students or data anomalies.

3.3.Analysis of Figure 3.8&3.9: Changes in Academic Scores by Income Group

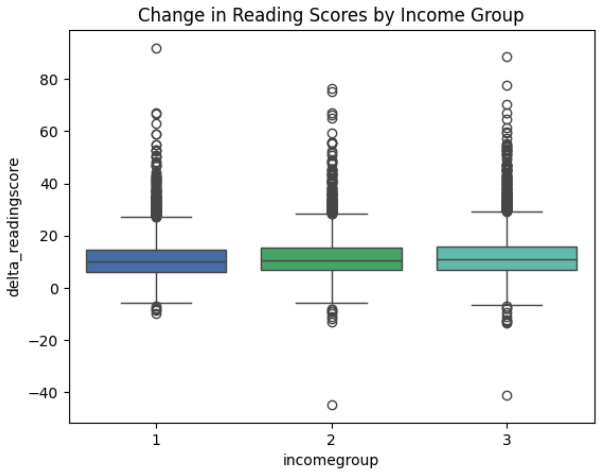

Figure 3.8. Changes in reading scores by Income Group in Box Plots

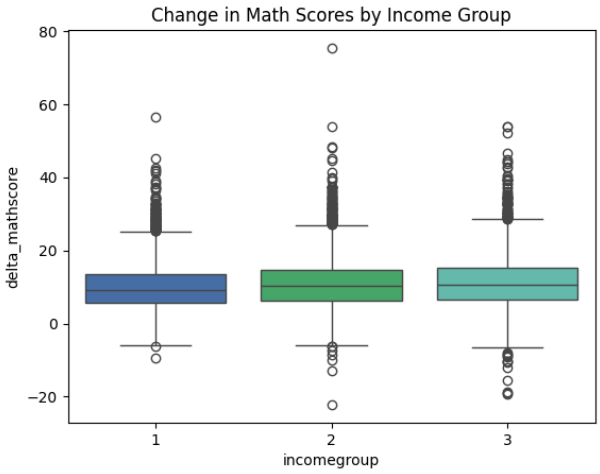

Figure 3.9. Changes in math scores by Income Group in Box Plots

From figure 3.8 and 3.9, several key insights could be generated:

a. The median increase in both reading and math scores from fall to spring appears to be consistent across all income groups, as indicated by the median line within each boxplot being around the same level.
b. There are outliers in each group, with some students showing exceptionally high increases or decreases in scores, indicating individual variations in both reading and math score improvements that are not explained solely by income group.

It's important to note that while the boxplots provide a visual assessment of the data, formal statistical tests would be necessary to determine if any observed differences are statistically significant. The next steps of the analysis will involve using one-way ANCOVA to control for other variables and to more precisely evaluate the impact of income on the changes in academic scores.

4.One-Way ANCOVA for Reading Scores Improvement

This part of the ANCOVA analysis sought to determine how the income group affects the changes in kindergarten students' reading scores over an academic year, after controlling for their general knowledge scores.

- Null Hypothesis (H0): There is no difference in the change in reading scores from fall to spring among the different income groups after controlling for the general knowledge score.
- Alternative Hypothesis (H1): There is a difference in the change in reading scores from fall to spring among the different income groups after controlling for the general knowledge score.

An OLS regression model was fitted to predict the change in reading scores based on income groups and spring general knowledge scores. The model, as shown by Table 4.1, indicated an adjusted R-squared value of 0.045, suggesting that the independent variables explain approximately 4.5% of the variance in the change in reading scores. The F-statistic and its associated p-value indicate that the model is statistically significant.

	Coefficient	**Std. Error**	**t-value**	**P>|t|**	**95% Confidence Interval**
Intercept	5.172	0.281	18.394	<0.001	[4.620, 5.723]
C(incomegroup)[T.2]	-0.089	0.178	-0.499	0.618	[-0.438, 0.260]
C(incomegroup)[T.3]	-0.049	0.189	-0.260	0.795	[-0.419, 0.321]
Spring general knowledge score	0.228	0.010	22.240	<0.001	[0.208, 0.248]

Table 4.1 OLS Regression Results for Reading Score Improvements

	Sum of Squares	**df**	**F-value**	**P-value**
C(incomegroup)	15.564	2.0	0.125	8.828e-01
Spring general knowledge score	30884.839	1.0	494.637	2.094e-107
Residual	744840.322	11929.0		

Table 4.2 ANCOVA result table

The coefficients for income groups 2 and 3 were not significantly different from the reference group (income group 1), with p-values well above the typical alpha level of 0.05, indicating that

changes in reading scores are not significantly different among the income groups after controlling for general knowledge scores.The coefficient for spring general knowledge scores was significant ($p < 0.0001$), suggesting that general knowledge is a significant predictor of the change in reading scores.

Levene's Test Statistic	19.728
p-value	2.795e-09

Table 4.3 Levene's Test result to check homogeneity of variances assumption

	Sum of Squares	**df**	**F-value**	**P-value**
C(income group)	15.564	2.0	0.125	8.827e-01
spring general knowledge score	30884.839	1.0	495.023	1.739e-107
C(income group):spring general knowledge score	706.993	2.0	5.666	3.471e-03
Residual	744133.329	11927.0		

Table 4.4.ANOVA results for a regression analysis

Shapiro-Wilk Test Statistic	0.902
p-value	<0.001

Table 4.5 Shapiro-Wilk Test result for normality of residuals

It's also important to look at if the four assumptions of ANCOVA analysis were met in this dataset. Levene's test (table 4.3) that's used resulted in a significant p-value ($p < 0.01$), indicating that the assumption of equal variances across groups is violated. As shown by table 4.4, the interaction between income groups and spring general knowledge scores was significant ($p < 0.01$), suggesting that the relationship between general knowledge scores and changes in reading scores may differ across income groups. The linearity assumption was checked visually using a scatter plot (figure 4.1) of predicted change in reading scores versus residuals. Also, the widespread residual around the red reference line is concerning for the linearity assumption. The spread of residuals around the zero line suggests a possible concern for non-linearity. The Shapiro-Wilk test statistic (table 4.5) indicated that the residuals are not normally distributed ($p < 0.01$), violating the normality assumption.

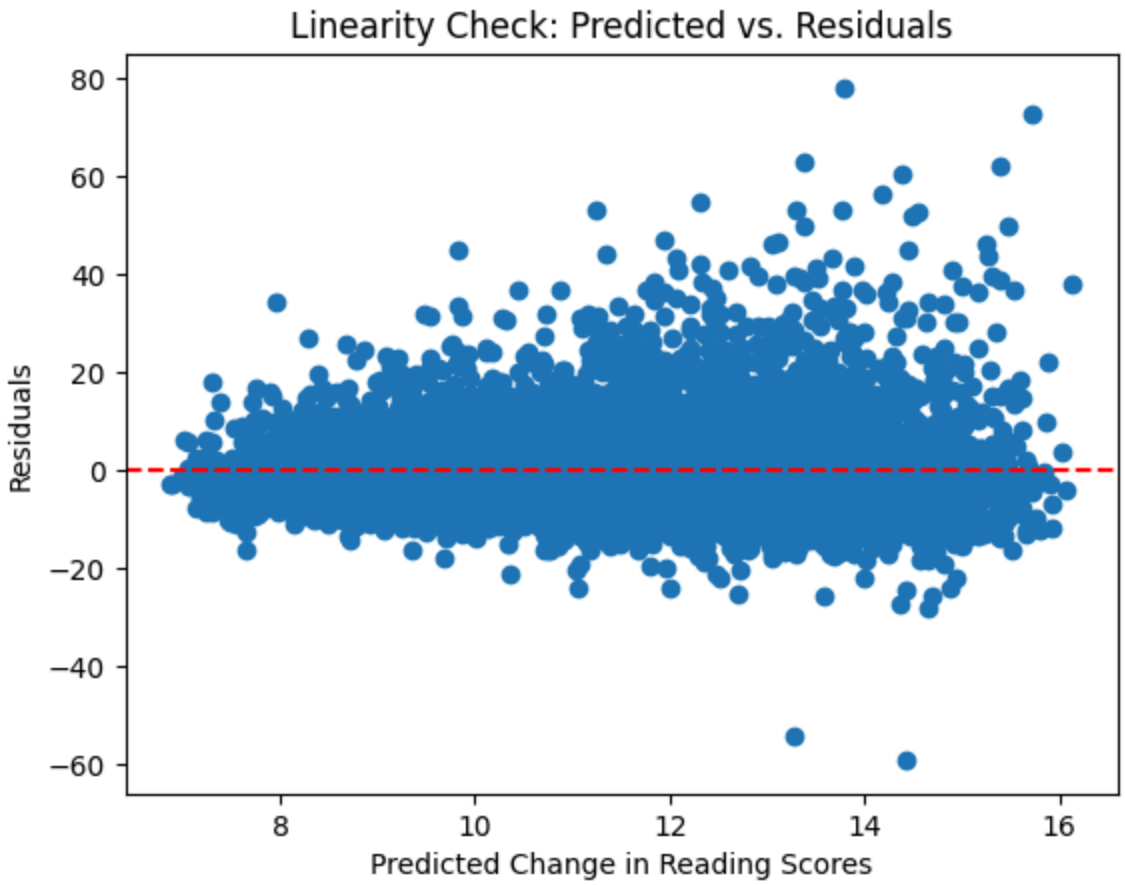

Figure 4.1 Linearity check for the reading score improvement

To sum up, the ANCOVA analysis of reading scores found no evidence that income group significantly affects the change in reading scores after controlling for general knowledge. While the model's overall significance suggests that general knowledge plays an important role in predicting reading score changes, the homogeneity of variance and normality assumptions were violated, which could affect the validity of the ANCOVA results.

Given the violations of assumptions, it may be necessary to consider alternative statistical methods or transformations of the data. Furthermore, the significant interaction term suggests that future analysis could explore the differential impact of general knowledge on reading score changes across income groups. The linearity check plot also suggests a need for further investigation into the potential non-linear relationships within the data.

5.One-Way ANCOVA for Math Scores Improvement

This section of the ANCOVA analysis focuses on understanding the influence of income group on the improvement of kindergarten students' math scores over an academic year, with control for their general knowledge scores.

- Null Hypothesis (H0):There is no difference in the change in math scores from fall to spring among the different income groups after controlling for the general knowledge score.
- Alternative Hypothesis (H1):There is a difference in the change in math scores from fall to spring among the different income groups after controlling for the general knowledge score.

An OLS regression model was fitted to predict the change in math scores based on income groups and spring general knowledge scores. The model indicated an adjusted R-squared value of 0.086, suggesting that the independent variables explain approximately 8.6% of the variance in the change in math scores. The F-statistic and its associated p-value indicate that the model is statistically significant.

	Coefficient	**Std. Error**	**t-value**	**P>|t|**	**95% Confidence Interval**
Intercept	3.183	0.233	13.647	<0.001	[2.726, 3.641]
C(incomegroup)[T.2]	-0.163	0.148	-1.106	0.269	[-0.453, 0.126]
C(incomegroup)[T.3]	-0.316	0.157	-2.016	0.044	[-0.622, -0.009]
Spring general knowledge score	0.270	0.008	31.822	<0.001	[0.254, 0.287]

Table 5.1.OLS Regression Results for Math Score Improvements

	Sum of Squares	**df**	**F-value**	**P-value**
C(income group)	175.900	2.0	2.047	0.129
Spring general knowledge score	43,530.309	1.0	1,013	<0.001
C(income group): Spring general knowledge score	325.899	2.0	3.793	0.023
Residual	512,450.225	11,927		

Table 5.2ANCOVA Result Table for Math Score Improvement

The coefficients for income groups 2 and 3 were not significantly different from the reference group (income group 1) for math scores, with p-values indicating that income group 3 had a significant effect on math score changes. The coefficient for spring general knowledge scores

was significant ($p < 0.0001$), suggesting that general knowledge is a significant predictor of the change in math scores.

Levene's Test Statistic	22.215
p-value	2.344e-10

Table 5.3.Levene's Test Result for Math Score Improvement

	Sum of Squares	**df**	**F-value**	**P-value**
C(income group)	175.900	2.0	2.047	1.292e-01
spring general knowledge score	43530.309	1.0	1013.144	1.825e-213
C(income group):spring general knowledge score	325.899	2.0	3.793	2.257e-02
Residual	512450.225	11927.0		

Table 5.4.ANOVA results for a regression analysis with Interaction Term

Shapiro-Wilk Test Statistic	0.969
p-value	<0.001

Table 5.5 Shapiro-Wilk Test result for normality of residuals

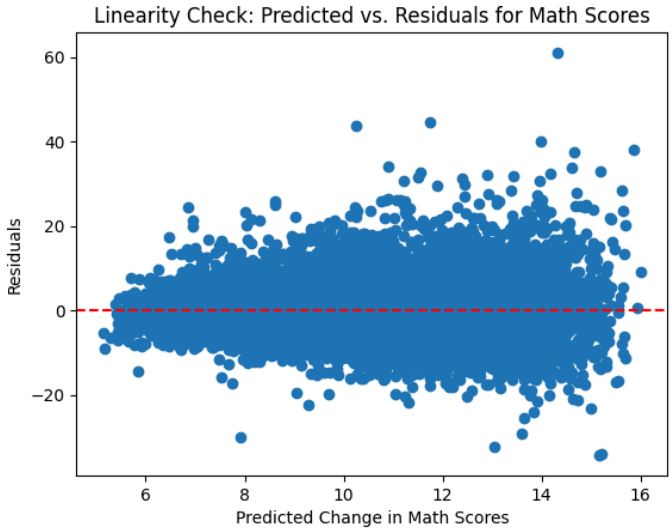

Figure 5.1 Linearity Check for Math Score Improvement

For the four assumptions in this ANCOVA analysis, Levene's Test (table 5.3) for Math Score Improvement showed a significant p-value (2.344e-10), indicating a violation of the homogeneity of variances assumption. The interaction term in the ANCOVA model (table 5.4) was significant ($p = 2.257e-02$), suggesting differences in the relationship between general

knowledge scores and math score improvements across income groups. Visual inspection of a scatter plot comparing predicted changes in math scores against residuals (figure 5.1) raised concerns about the linearity assumption. Moreover, the Shapiro-Wilk Test for Math Score Improvement (table 5.5) showed that the residuals were not normally distributed ($p < 0.001$), violating the normality assumption.

The ANCOVA analysis on math scores showed that while the general knowledge score is a strong predictor of math score improvements, the impact of income group is not uniform across the groups, with group 3 showing a statistically significant effect. The analysis, however, faced challenges with the assumptions of homogeneity of variances and normality of residuals, which can affect the interpretation of the results.

Similar to the reading score's ANCOVA analysis, due to the violations of key assumptions, alternative methods or data transformations should be considered for future analysis. The significant interaction suggests that the effect of general knowledge on math score improvements may vary across different income groups. Further research could investigate these potential differential effects and examine non-linear relationships suggested by the linearity check plot.

6.Conclusion

The ANCOVA analysis for reading scores did not reveal a statistically significant difference in score improvements across the income groups once general knowledge scores were controlled for. This suggests that, in terms of reading, income level may not be a determining factor in academic progression over the course of a kindergarten year. In contrast, the ANCOVA analysis for math scores indicated a significant difference for income group 3, hinting at a potential impact of higher income level on math score improvements. The general knowledge scores were found to be a significant predictor in both reading and math score improvements, emphasizing the role of initial knowledge in academic progression.

Despite these insights, there are several limitations to this study which must be acknowledged:

a. Assumption Violations: Both analyses encountered issues with the homogeneity of variances and normality of residuals. These violations can impact the trustworthiness of ANCOVA results and suggest the need for cautious interpretation.
b. Statistical Power and Variance Explanation: The adjusted R-squared values in both models were relatively low, indicating that only a small percentage of the variance in the dependent variables was accounted for by the models, also implying that the scores of students have variations for different individuals.
c. Potential Confounding Variables: There are likely other variables not included in this study that could influence academic progress, such as parental involvement, quality of instruction, or access to educational resources.

d. Methodological Constraints: The significant interaction terms suggest that the relationship between general knowledge and score improvements may differ across income groups, which the ANCOVA may not fully capture.

Future research should consider these limitations and explore alternative statistical approaches that can handle assumption violations, such as robust regression or transformation of variables. Moreover, further investigation into potential confounding variables and a deeper dive into the interaction effects could yield more nuanced understandings of the factors that influence academic progress.