Quantitative Analysis of Early Childhood Scores by Income Group

Abstract

This report investigates the association between income groups and academic performance in reading and mathematics among Kindergarten students. The analysis employs one-way ANCOVAs, controlling for general knowledge to elucidate how income categories may affect educational outcomes.

Introduction

The role of socioeconomic status in educational attainment is a well-documented phenomenon. This study aims to dissect the influence of income on the academic progression of Kindergarten students through a detailed statistical examination of reading and math scores at the end of a school year. Using data from an early childhood longitudinal study, this report seeks to answer the **research question**:

1. How does income impact the reading scores of Kindergarten students when controlling general knowledge?
2. 1.How does income impact the math scores of Kindergarten students when controlling general knowledge?

Data Examination and preparation

The project began by loading and examining the dataset, ensuring there were no missing values and gaining a preliminary understanding of its structure. I excluded “total household income” and “income in thousands” as they are only used to generate “income group”. From table-1 and table-2, it is clear that there is no missing value, no abnormal value in all variables. The dataset is ready for quantitative analysis.

The dataset was carefully prepared for analysis, with income categories derived from continuous income variables to facilitate our ANCOVA.

Table-1: Dataset Structure

Column	Total count	Null value count	Data type
Fall reading score	11933	0	float64
Fall math score	11933	0	float64
Fall general knowledge score	11933	0	float64
spring reading score	11933	0	float64
spring math score	11933	0	float64
spring general knowledge score	11933	0	float64
Total household income	11933	0	float64
Income in thousands	11933	0	float64
Income group	11933	0	int64

Table-2: Dataset summary

	Fall Reading	Fall Math	Fall General	Spring reading	Spring math	Spring general	Income group
count	11933	11933	11933	11933	11933	11933	11933
mean	35.954215	27.128244	23.074	47.511	37.79	28.235	1.895
std	10.473	9.121	7.396	14.327	12.028	7.577	0.822
min	21.01	10.51	6.985	22.35	11.9	7.858	1

25%	29.34	20.68	17.385	38.95	29.27	22.802	1
50%	34.06	25.68	22.954	45.32	36.41	28.583	2
75%	39.89	31.59	28.305	51.77	44.22	33.782	3
max	138.51	115.65	47.691	156.85	113.8	48.345	3.

Data Exploration(EDA)

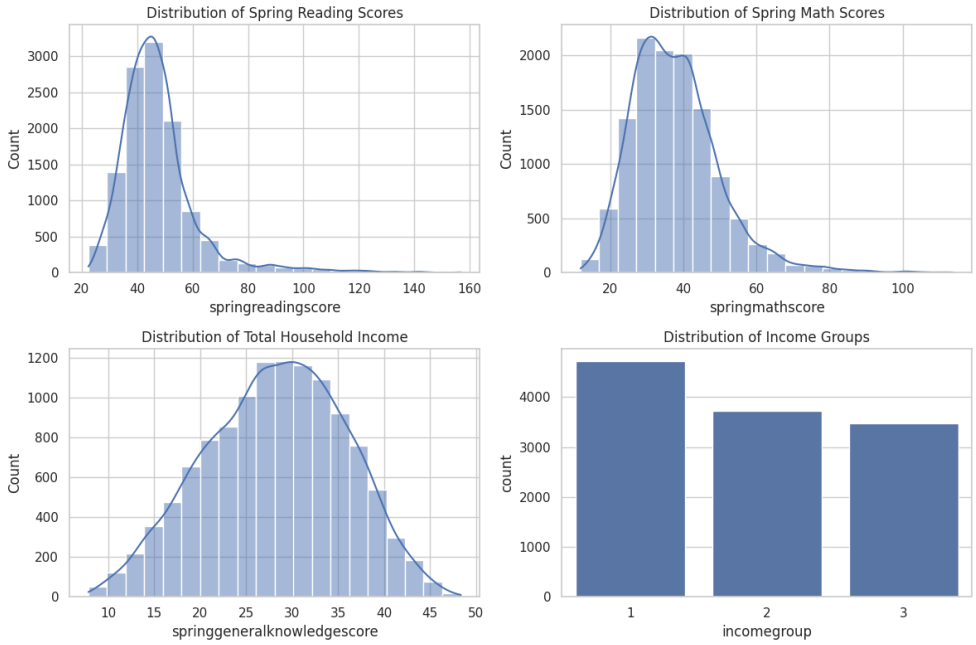

Figure-1: Histograms of Reading and Math Scores

EDA was conducted to understand the distribution and relationship between the academic scores and income. Histograms of the reading and math scores were generated, providing insights into the distribution patterns across different income groups(in figure-1).

In figure-1, the distribution of spring scores for reading and math displays a right-skewed trend, implying that a majority of students have scores toward the lower end. In contrast, general knowledge scores are more symmetrically distributed, hinting at a more uniform spread of general knowledge across students. The income groups histogram indicates that most students fall into the lowest income category. The less skewed distribution of general knowledge suggests it may be less influenced by income compared to reading and math scores, but further analysis would be necessary to confirm any such patterns or correlations.

In figure-2, the boxplots, median scores for reading and math rise from the lowest to the highest income groups, hinting at better academic performance among students from wealthier families. The broadest range of reading scores is seen in the lowest income group, denoting diverse abilities. Outliers in all groups, especially in the lowest, reflect varied achievement levels. Meanwhile, higher income groups show more uniform scores, suggesting more consistency in their academic performance. Math scores present a similar trend with slightly less variability than reading. Overall, the boxplots suggest that socioeconomic status may correlate with academic success, particularly in higher income groups which show higher medians and narrower score ranges in the spring assessment.

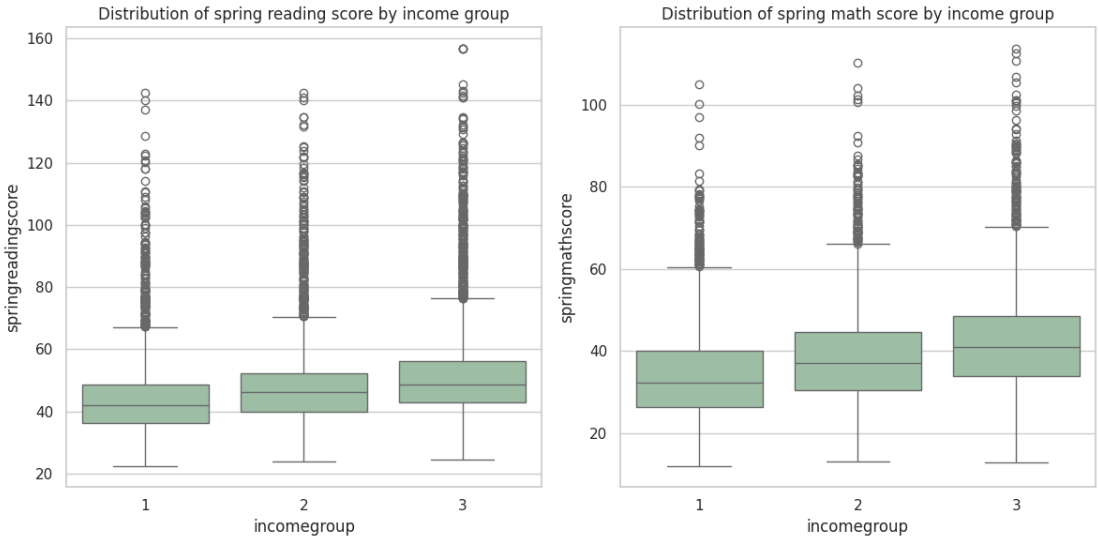

Figure-2: Boxplots of Reading and Math Scores across income group

Methodology

The quantitative analysis centered around multiple one-way ANCOVAs, each designed to measure the impact of income group on reading and math scores, with general knowledge scores serving as a covariate. This approach aimed to isolate the effect of income while accounting for baseline knowledge. I created several functions to conduct ANOVAs, test linearity, test normality, test homogeneity.

Assumption Testing for ANCOVA

Before interpreting the results of the ANCOVA analysis, it is crucial to ensure that several assumptions underlying the statistical test are met. These assumptions include:

1. **Linearity**: The relationship between the covariates and the dependent variable should be linear. This can be checked by plotting the residuals against predicted values or covariates and examining for any discernible patterns in the plot.
2. **Homogeneity of Variances**: The variances of the dependent variable should be equal across different groups. This assumption can be tested using Levene's test, which assesses whether the variability in scores is similar across groups.
3. **Normality of Residuals**: The residuals (errors) from the model should be normally distributed. This assumption can be checked using statistical tests like the Shapiro-Wilk test for normality, as well as visual methods such as Q-Q plots or histograms of the residuals.

By conducting these assumption checks, we can ensure the validity and reliability of the ANCOVA analysis results. Violations of these assumptions may require further data transformation or alternative statistical methods for accurate interpretation.

One-way ANCOVA(research question 1)

Assumption Testing

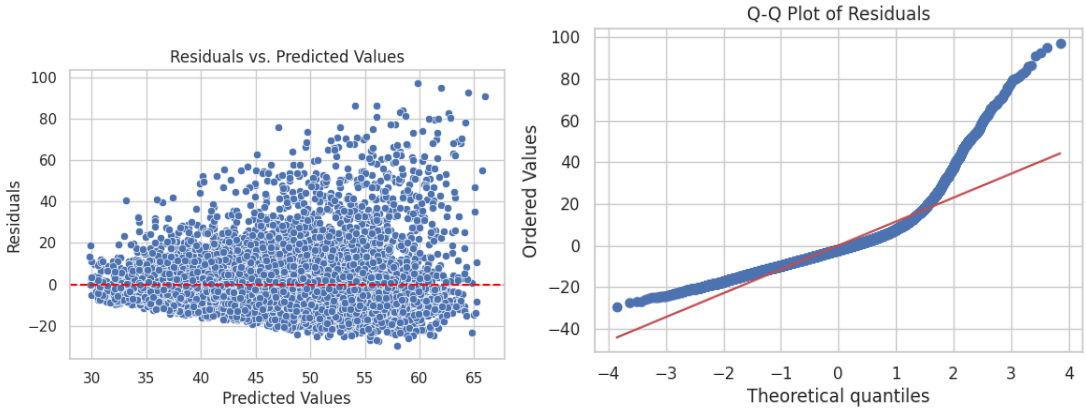

Figure-3:residual plot and QQ plot(reading)

Table-3 Assumption testing for one-way ANCOVA (reading)

Test	Statistics	P-value	Interpretation
Shapiro-Wilk	0.827	**< 0.001***	Violation of normality
Levene's Test	39.553	**< 0.001***	Violation of homogeneity of variances

Interpretation:

- The **Shapiro-Wilk test** resulted in a statistic of approximately 0.827 with a p-value less than 0.001, indicating a rejection of the null hypothesis of normality. This means the residuals from our ANCOVA model do not follow a normal distribution, which is a violation of one of the assumptions of ANCOVA.
- **Levene's test** produced a statistic of approximately 39.553 with a p-value less than 0.001, leading to a rejection of the null hypothesis of equal variances. This suggests that the assumption of homogeneity of variances is not met across the different income groups.
- **Residuals vs. Predicted Values Plot**: The plot demonstrates a pattern where the residuals randomly fall into both side of the predicted value, indicating potential linearity, which meets one of ANCOVA assumptions: linearity.
- **Q-Q Plot of Residuals**: The plot shows a clear deviation from the line, especially at the tails, confirming the results from the Shapiro-Wilk test that the residuals do not follow a normal distribution. This deviation suggests the presence of outliers or that the residual distribution has heavier tails than a normal distribution.

In light of these results, the assumptions necessary for the validity of ANCOVA are not fully met. This suggests **the need for caution in interpreting the ANCOVA results** and may warrant the use of alternative statistical methods that are not dependent on these assumptions.

ANCOVA Results and Interpretation

Table-4 one-way ANCOVA Results(reading)

Source	**Sum of Squares**	**df**	**F**	**p-value**
C(income group)	18,661.443	2	58.533	<0.001
Spring General Knowledge Score	400,061.996	1	2509.637	<0.001

Residual	1,901,605.882	11,929		

The ANCOVA results demonstrate a statistically significant effect of the income group on spring reading scores, controlling for general knowledge ($F(2) = 58.533$, $p < 0.001$). This suggests notable differences in reading proficiency across income groups. Moreover, general knowledge scores as a covariate have an exceedingly significant effect on spring reading scores ($F(1) = 2509.637$, $p < 0.001$), implying a strong association between students' general knowledge and their reading abilities.

The residual sum of squares represents the variation in reading scores that is unexplained by the model, with the degrees of freedom corresponding to the number of observations minus the number of estimated parameters.

These findings underline the significance of socioeconomic factors, as indicated by income groups, as well as the role of general knowledge in educational outcomes related to reading. The extremely low p-values, reported as "<0.001," emphasize the robustness of these findings, albeit with a caution against inferring an exact zero probability of such results occurring by random chance.

One-way ANCOVA(research question 2)

Assumption Testing

Table-5 Assumption testing for one-way ANCOVA (math)

Test	Statistics	P-value	Interpretation
Shapiro-Wilk	0.948	**< 0.001***	Violation of normality
Levene's Test	18.899	**< 0.001***	Violation of homogeneity of variances

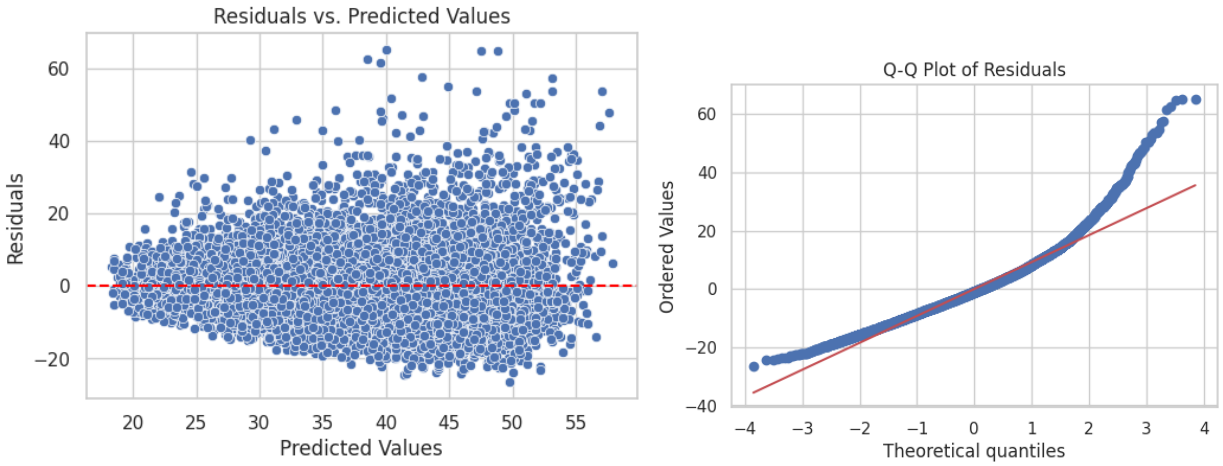

Figure-4:residual plot and QQ plot(math)

Interpretation:

- Shapiro-Wilk Test: A statistic of 0.948 with a p-value less than 0.001 suggests that the residuals do not follow a normal distribution. This indicates a potential violation of normality, which is a fundamental assumption of ANCOVA, and may impact the reliability of the test's results.
- Levene's Test: A statistic of 18.899, also with a p-value less than 0.001, implies that the variances of the groups are not equal, signifying a breach of the homogeneity of variances assumption. This could affect the ANCOVA's robustness and might lead to increased Type I or Type II errors.

- Residuals vs. Predicted Values Plot for Math Scores: I observe a pattern where residuals are not evenly distributed, especially in the middle range of predicted values. This pattern supports the results from Levene's test, suggesting heteroscedasticity, or unequal variances, across the range of predicted values.
- The plot shows a clear deviation from the line, especially at the tails, confirming the results from the Shapiro-Wilk test that the residuals do not follow a normal distribution. This deviation suggests the presence of outliers or that the residual distribution has heavier tails than a normal distribution.

Given these findings, caution is advised when interpreting the ANCOVA results for math scores. Alternative approaches, such as data transformation or the use of non-parametric methods, may be warranted to accommodate the violations of the ANCOVA assumptions.

ANCOVA Results and Interpretation

Table-6 ANCOVA Summary Table for Spring Math Scores

Source	**Sum of Squares**	**df**	**F**	**p-value**
C(income group)	11,669.913	2	64.830	< 0.001
spring general knowledge score	504,336.297	1	5603.517	< 0.001
Residual	1,073,652.115	11,929		

Interpretation:

The ANCOVA analysis for spring math scores, controlling for spring general knowledge scores, demonstrates statistically significant effects for both the income group ($F(2) = 64.830$, $p < 0.001$) and general knowledge scores ($F(1) = 5603.517$, $p < 0.001$). The significance of the income group suggests that there are differences in math scores across different income levels even after controlling for general knowledge. The very high F-statistic for general knowledge indicates a strong association with math scores, which implies that as general knowledge scores increase, math scores also tend to increase, regardless of the income group.

The results suggest that while income group is a significant factor in the variability of math scores among students, general knowledge has a more substantial effect. The residual term represents the unexplained variation after accounting for the model's factors. The high significance of the covariate also underscores the importance of general knowledge in the context of math achievement.

These findings should be interpreted with caution given the earlier noted potential violations of ANCOVA assumptions, such as normality of residuals and homogeneity of variances. It may be beneficial to explore these relationships further using different models or transformation techniques that can handle the assumption violations.

Interaction plot

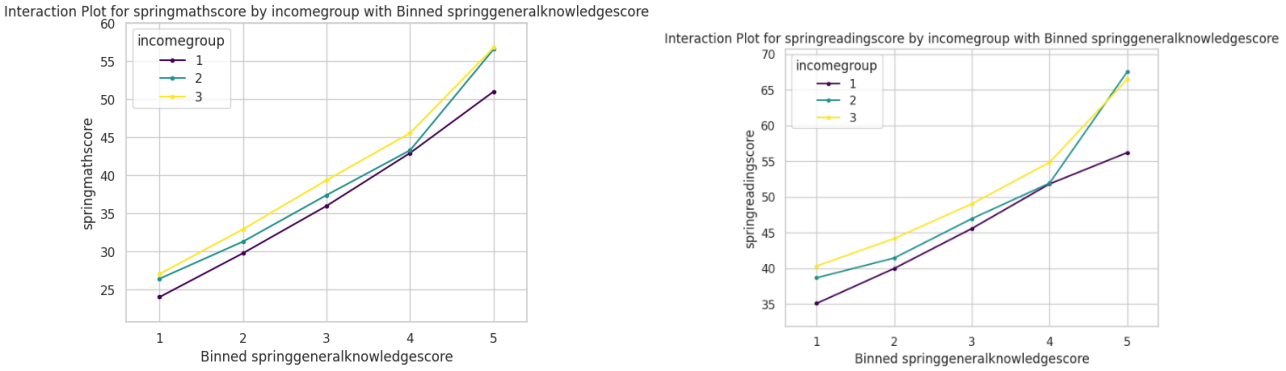

Figure-5:residual plot and QQ plot(reading)

This interaction plot(reading) is pivotal in illustrating the interplay between socioeconomic status and educational outcomes. It demonstrates that while general knowledge is an important factor for reading achievement across all income groups, the advantage of higher general knowledge is more pronounced for students from more affluent backgrounds. These insights could be instrumental in informing educational policy and interventions aimed at closing the achievement gap related to socioeconomic status.

This interaction plot underscores the significant correlation between general knowledge and math achievement while also highlighting the impact of socioeconomic status. It illustrates that students from higher-income families are likely to experience greater benefits in math performance with increased general knowledge compared to their lower-income peers. This data supports the narrative that socioeconomic status can influence the degree to which educational outcomes can be maximized and points towards the necessity for targeted educational strategies to mitigate these disparities.

Discussion

The preliminary findings suggest that income has a discernible effect on the academic outcomes of Kindergarten students, moderated by their general knowledge. The inability to fully satisfy ANCOVA's assumptions calls for a cautious interpretation of the results. It indicates that while the results signal a potential pattern, they may not be as precise as desired.

Conclusion

The ANCOVA analysis indicates a relationship between income and academic performance, contingent upon students' general knowledge. Future studies may consider alternative statistical methods or transformations to fully meet the assumptions of ANCOVA, thereby ensuring the robustness of the findings.