INF 2178 Technical Assignment 3

Jinhang Luo (1005124777)

Faculty of Information
University of Toronto
INF2178: Experimental Design for Data Science
Prof.Shion Guha

March 23rd, 2024

Introduction

1.1 Background

This dataset, sourced from an early child longitudinal study (1998-99), recorded kindergarten students' reading, math, and general knowledge scores as measured during six months from the Fall 1998 to Spring 1999. Particularly noteworthy is that the dataset categorizes total household income by income group. These comprehensive data are invaluable to policymakers, educators, and practitioners in developing targeted strategies to create more inclusive and supportive educational environments that close gaps in accessibility and affordability, ensure that children's developmental needs are effectively met, and support the economic needs of families.

1.2 Dataset description

The dataset contains 11,933 records and numerous attributes. However, our main focus is exploring the interaction between variables such as reading score (1998-99), math score (1998-99), general knowledge score (1998-99), total household income, and income group.

1.3 Research question

In analyzing the dataset, we aimed to explore the effect of income groups on changes in academic scores after controlling for baseline general knowledge scores. Consequently, two research questions have been formed.

1. How do income groups affect changes in reading scores from fall 1998 to spring 1999?
2. How do income groups affect changes in math scores from fall 1998 to spring 1999?

Data Cleaning

We tidied up the dataset by removing unnecessary columns and renaming the remaining ones. Then, we calculated the reading score change and math score change for our main analysis.

Exploratory data analysis

3.1 Descriptive Statistics

Firstly, we performed a descriptive analysis that included the dataset's sample size, mean, and standard deviation for certain attributes.

As shown in **Table 1**, the sample size of the data is 11,933. Moreover, the variability in means and standard deviations among various variables suggests that the dataset is robust enough to provide meaningful statistical conclusions.

Table 1. Descriptive Statistics

Variables	N	Mean	SD
Fall reading score	11,933	35.954	10.473
Fall math score	11,933	27.128	9.121
Fall general knowledge score	11,933	23.074	7.397
Spring reading score	11,933	47.511	14.327
Spring math score	11,933	37.799	12.028
Spring general knowledge score	11,933	28.236	7.577
Total household income	11,933	54317.200	36639.061
Reading score change	11,933	11.557	8.085
Math score change	11,933	10.671	6.855

Source: Early Child Longitudinal Study (1998-99).

3.2 Distribution Visualizations

As shown in **Figure 1**, these histograms visually illustrate the distribution of the data across the key variables, as well as the change in academic scores from 1998-99 by income group. It is worth noting that from fall 1998 to spring 1999, the score changes in reading and math were mostly positive, indicating a mostly improving trend in children's scores. Additionally, the distribution of total household income is right-skewed, indicating that more households have lower incomes and fewer have higher incomes.

Furthermore, as shown in **Figure 2**, the box plots indicate that in reading and math scores, the median change across all income groups showed positive scores, suggesting an improvement in overall scores. However, spreads and outliers indicate differences in score change.

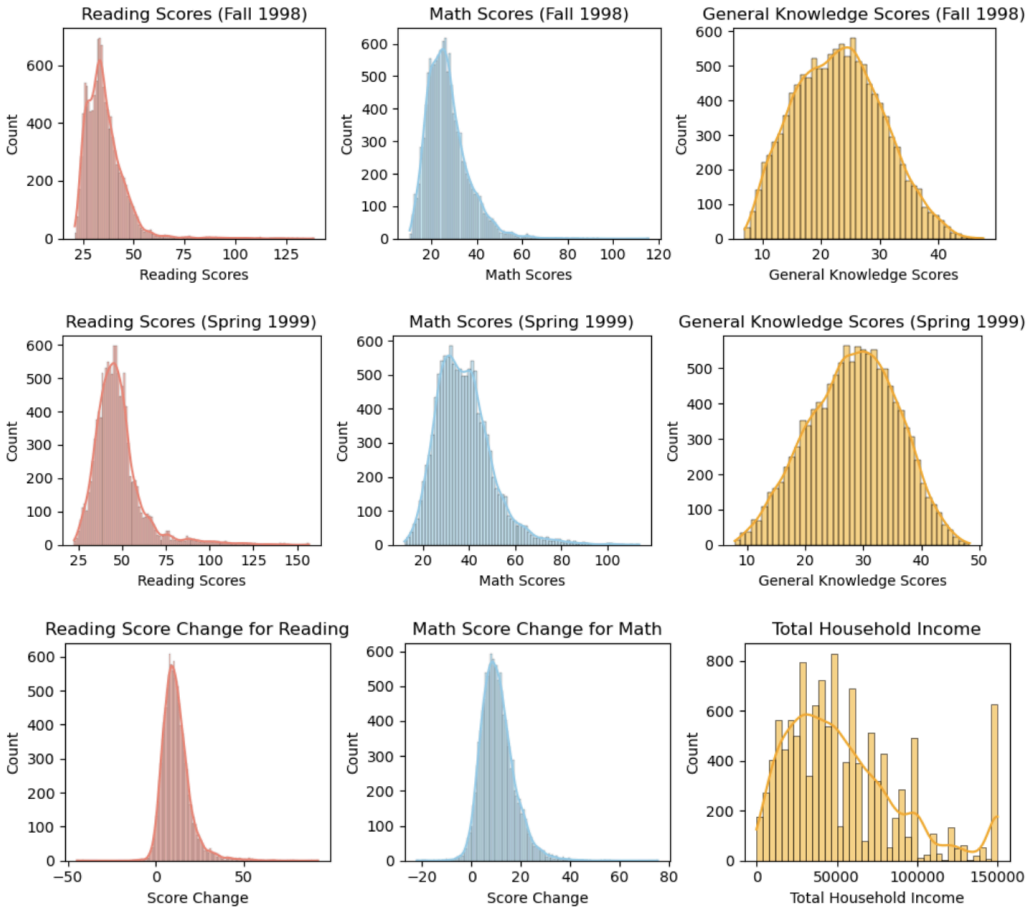


Figure 1: Histogram Plots of all Variables

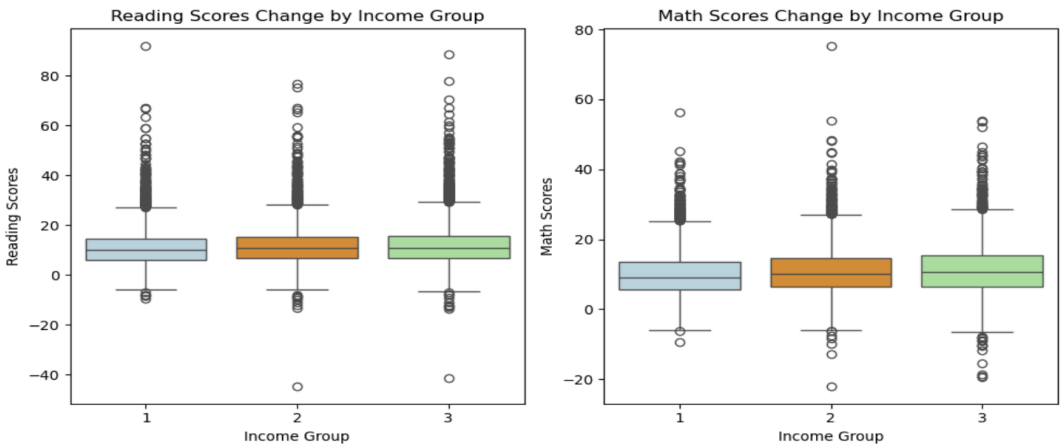


Figure 2: Boxplot Plots of Reading/Math Score Change by Income Group

Quantitative Analysis (One-way ANCOVA)

We will use the one-way ANCOVA to determine whether the income group affects the changes in reading and math scores from fall 1998 to spring 1999 while controlling for the baseline general knowledge scores.

As seen in **Table 2**, the p-value for the moderate income group is less than $\alpha = 0.05$. Also, since 0 is not in the confidence interval, this indicates that there is strong evidence to reject the null hypothesis of the ANCOVA. This also suggests a statistically significant difference between the income groups and the change in reading scores, controlling for baseline general knowledge scores.

Table 2. The One-way ANCOVA Table (Reading Score Change as Outcome Variable)

	Std err	T	P>|t|	[0.025 0.975]
Income [T.2]	0.180	1.205	0.228	-0.136 0.570
Income [T.3]	0.191	2.110	0.035	0.029 0.779

Source: Early Child Longitudinal Study (1998-99).

As seen in **Table 3**, the p-values for all income groups are higher than $\alpha = 0.05$. Since 0 is within the confidence interval, this indicates that there is no strong evidence to reject the null hypothesis of the ANCOVA. This indicates that there might be no statistically significant difference between the income groups and the change in math scores, controlling for baseline general knowledge scores.

Table 3. The One-way ANCOVA Table (Math Score Change as Outcome Variable)

	Std err	T	P>|t|	[0.025 0.975]
Income [T.2]	0.151	1.011	0.312	-0.143 0.448
Income [T.3]	0.160	0.900	0.368	-0.170 0.458

Source: Early Child Longitudinal Study (1998-99).

3.1 Assumption Checks

A. Normality of Residuals

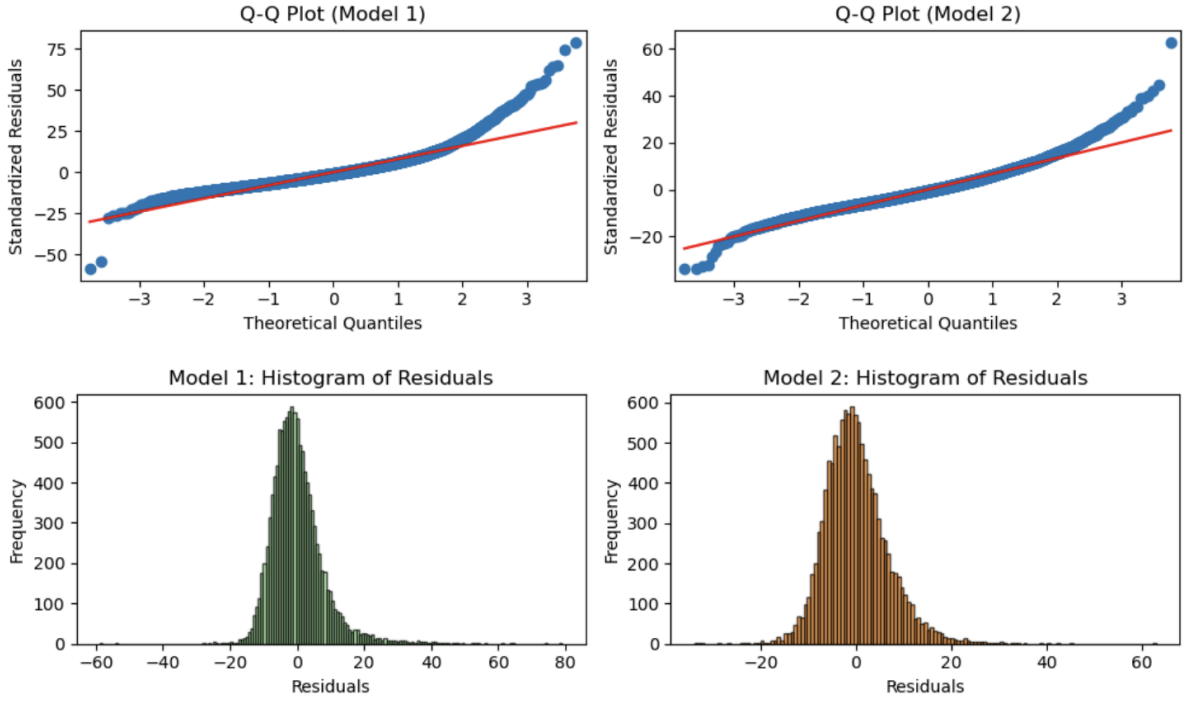


Figure 3: The Q-Q Plot and Histogram Plot from ANCOVA Tables (Reading/ Math Score Change)

As shown in **Figure 3**, both model's diagnostic plots show that the residuals are very close to the normality assumption. The Q-Q plot shows that most points are on the normal line. The shape of both model's histograms also generally supports that residuals follow a normal distribution.

Table 4. The Shapiro-Wilk Test. (Model1:Reading Score Change, Model2: Math Score Change)

	Test statistics (W)	P-Value
Value(Model 1)	0.900	<0.001
Value (Model 2)	0.966	<0.001

Source: Early Child Longitudinal Study (1998-99).

As shown in **Table 4**, the Shapiro-Wilk test for Model 1 has a test statistic (W) of 0.900 and a p-value of less than 0.001, indicating that the residuals deviate from normality. For Model 2, the test statistic (W) is 0.966 with a p-value of less than 0.001, suggesting a closer to normality and indicating a statistical deviation. However, considering the robustness of ANCOVA to certain violations of its assumptions, we decided to continue with the analysis and will consider alternative approaches for specific explanations in the limitation section at the end.

B. Homogeneity of Variances

As shown in **Table 5**, Levene's test statistic (W) for both Models has a p-value of less than 0.05, indicating a homogeneity assumption violation. However, considering the robustness of ANCOVA to certain violations of its assumptions, we decided to continue with the analysis and will consider alternative approaches for specific explanations in the limitation section at the end.

Table 5. The Levene's Test. (Model1:Reading Score Change, Model2: Math Score Change)

	Test statistics (W)	P-Value
Value(Model 1)	19.728	2.795e-09
Value (Model 2)	22.215	2.344e-10

Source: Early Child Longitudinal Study (1998-99).

3.2 Post-hoc tests (Tukey's HSD)

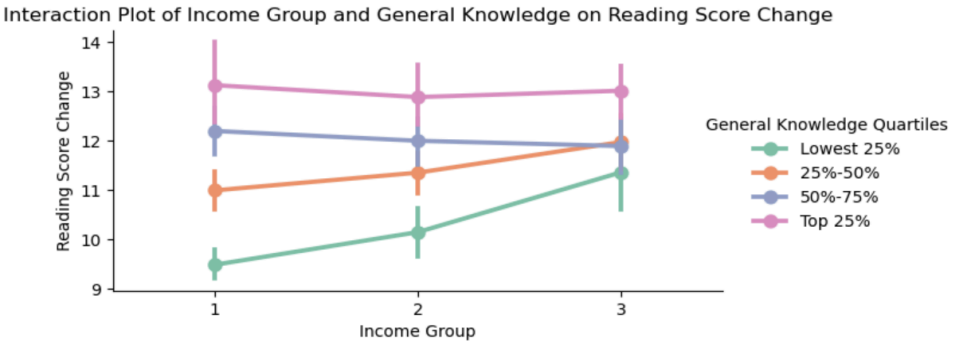
As shown in Table 6, there is a statistically significant difference between the income groups using change in reading and math scores as the dependent variable. For both models, all comparisons indicate a rejection of the null hypothesis.

Table 6. The Tukey HSD Table. (Model1:Reading Score Change, Model2: Math Score Change)

	1 : 2	1 : 3	2: 3
Reject (Model 1)	True	True	True
Reject (Model2)	True	True	True

Source: Early Child Longitudinal Study (1998-99).

3.3 Interaction Plot



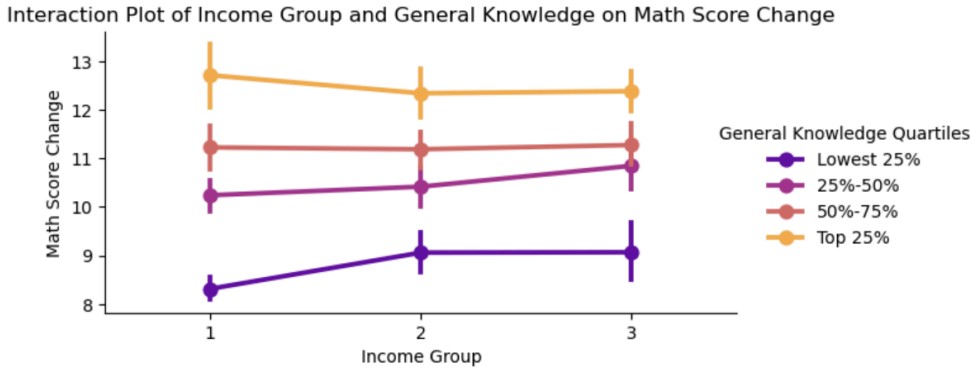


Figure 4: The Interaction Plots of Income Group and General Knowledge Score On Reading/Math Score Change

As shown in **Figure 4**, the interaction plot shows how income groups with different quartiles of general knowledge affect changes in reading and math scores. The interaction plots show that children within the **'Top 25%'** of general knowledge show consistent score changes across income groups, suggesting a stabilizing effect of prior knowledge on educational performance. Conversely, those in the '**Lowest 25%**' show greater variability, with changes in scores being more sensitive to changes in income groups. This implies that lower baseline knowledge may make reading and math score improvements more susceptible to income-related differences, thus suggesting a complex interaction between socioeconomic status and baseline knowledge in educational advancement.

Limitation

1. The findings from this dataset may be limited to other cities or regions.
2. Shapiro-Wilk test and Levene's tests indicate that the assumption of normality and homogeneity of variance are violated. While ANCOVA is robust to some deviations from this assumption, the extent of the deviation here may have influenced the investigation's results. Alternative methods, such as Welch's ANOVA or Generalized Least Squares, may be considered to address this issue in later stages.

Conclusion

In summary, our investigation demonstrates a complex correlation between income groups and baseline general knowledge in affecting changes in children's reading and math scores. The investigation shows that children with lower baseline knowledge significantly improve reading and math scores as income groups rise. This suggests that while higher baseline knowledge provides a certain level of educational adaptability, those children with lower initial knowledge levels benefit more from better economic conditions. These findings highlight the multi-dimensional effects of socioeconomic status and initial general knowledge level on academic improvement and provide valuable insights for educators and policymakers working to promote equitable educational advancement.