INF 2178

Technical Assignment 3

Yidan Chen

Introduction

A child's educational journey is shaped by various factors, with family income potentially playing a pivotal role. Our project seeks to explore the influence of socioeconomic status on the academic growth of young learners, particularly in reading and math, across a school year. We examine how income levels correlate with academic progress when general knowledge is taken into account. Using the "INF2178_A3_data.csv" dataset, which captures reading and math scores from the fall of 1998 to the spring of 1999 alongside family income data, we apply one-way ANCOVA analysis to explore our **Research Question: Does income group affect changes in reading and math scores for kindergarten students, after controlling for their initial general knowledge?** This analysis is designed to reveal not only the direct influence of income on academic outcomes but also to consider the foundational knowledge that children bring into the classroom. Through this lens, we aim to provide a clearer picture of the educational landscape in early childhood.

Data Preparation

In the data preparation phase, we first confirmed the dataset comprised 11,933 entries across 9 columns. Upon reviewing data types, we converted the 'incomegroup' from an integer to a categorical variable, suitable for ANCOVA. With no missing values detected, we proceeded to simplify the dataset by removing the 'totalhouseholdincome' column to avoid redundancy. Lastly, we added two new variables: 'change_in_reading' and 'change_in_math'. These variables represent the difference in reading and math scores from fall to spring, respectively, capturing the academic progress of students over the academic year.

Exploratory Data Analysis

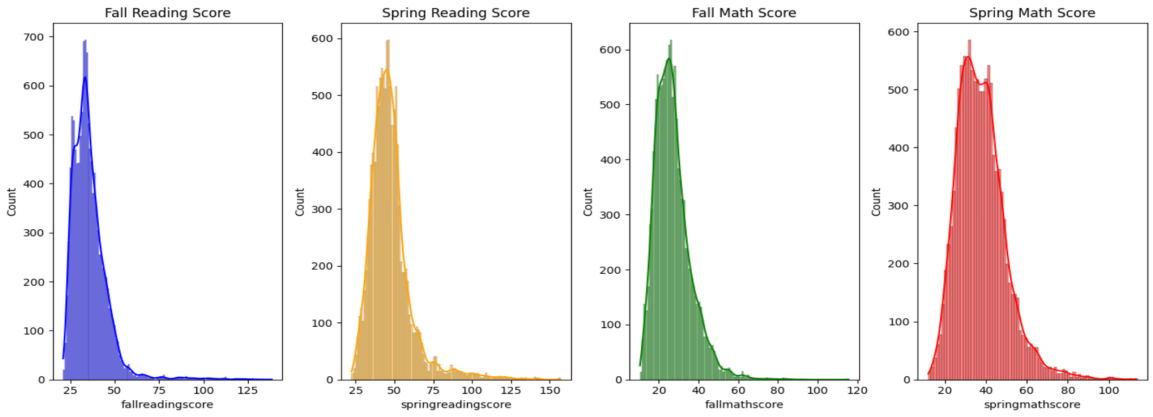


Figure 1 Distribution of Fall and Spring Reading and Math Scores

As shown in Figure 1, the fall and spring reading score histograms show a rightward skew, indicating that a majority of students start with lower scores in the fall. By spring, this skewness diminishes somewhat, suggesting an overall improvement in reading abilities with a shift towards higher scores. The distribution in the spring is wider, reflecting greater variability in students' reading performances. Moreover, the math score histograms for both fall and spring exhibit a pronounced right skew, though there is a notable shift towards higher scores from fall to spring. The concentration of students with lower scores is significant in the fall, which disperses by spring, indicating that students generally improve in math. However, the persistence of skewness in spring suggests that while there is growth, it may not be uniform across the student population.

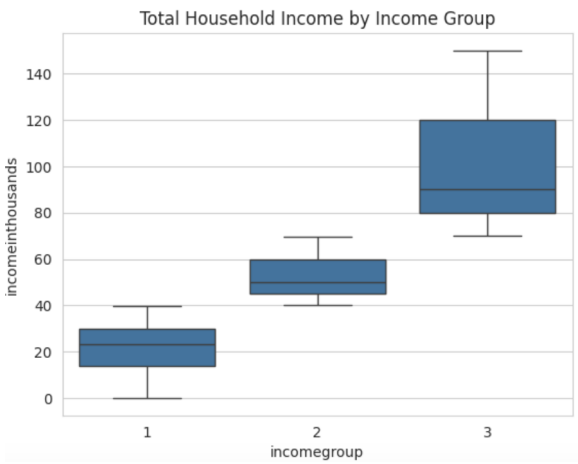


Figure 2 displays a boxplot for each income group, delineating the distribution of household incomes within thousands. Income Group 1 shows the lowest median household income with a relatively tight interquartile range, suggesting less variation within this group. Income Group 2 exhibits a higher median income and a slightly narrower IQR. Income Group 3 has the highest median income with a considerably wider range, which points to a greater diversity in income levels within this category. Additionally, Group 3's boxplot reveals the presence of higher-income outliers, indicating that some households earn significantly more than their counterparts within the same group.

Figure 2 Total Household Income by Income group

Then, we explore the relationship between independent variables (income group), dependent variables (math/reading scores) and Covariate(general knowledge score). The scatter plot illustrates the relationship between students' general knowledge scores in the fall and their math scores. There is a visible positive trend, indicating that higher general knowledge scores tend to correspond to higher math scores. However, the data points are quite dispersed, suggesting that while general knowledge is related to math performance, other factors may also play a significant role.

The boxplots Figure 4 on the right compare fall math scores across the three income groups. The median math score appears to increase with higher income groups.Notably, the higher income group (Group 3) also exhibits greater variance in math scores than the lower groups, as seen by the longer box and wider spread of scores. Additionally, there are several outliers in all income groups, indicating some students' scores deviate substantially from the median of their respective group.

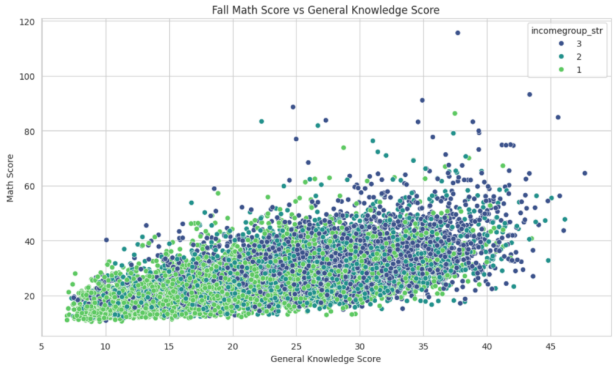


Figure 3 Fall Math Score vs General knowledge score

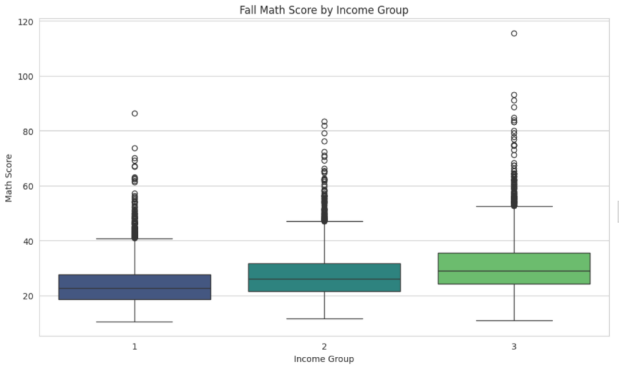


Figure 4 Fall Math Score by Income Group

As for reading score, figure 5 also suggests a positive correlation between fall general knowledge scores and fall reading scores—students with higher general knowledge tend to have higher reading scores. This trend appears consistent across all income groups, which suggests that general knowledge might be predictor of reading achievement. Figure 6 highlights the distribution of fall reading scores across different income groups. Similar to the findings in math scores, the median reading score tends to increase with higher income groups. This trend suggests that socioeconomic status, as approximated by income group, may have a relationship with reading proficiency. There is a notable number of outliers in the higher income groups, indicating that there are students with

reading scores significantly above the group's median. The spread of scores, particularly in the higher income group, is also wider, implying more variability in reading achievement within this group compared to the lower income groups.

Figure 5 Fall Reading Score vs General knowledge score

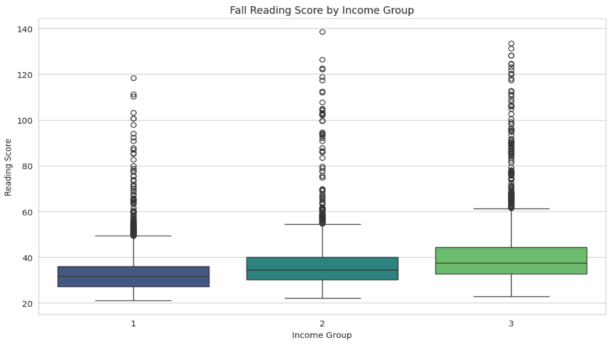


Figure 6 Fall Reading Score by Income Group

Next, we made two boxplots to present the changes in reading and math scores by income group. For reading scores, all income groups show positive median changes, indicating improvement overall. The IQR for all income group suggest similar variability in reading progress with similar median changes. Interestingly, Groups 2 and 3 have outliers indicating both significant improvements and decreases, with Group 3 showing particularly high gains for some students.

Math score changes also exhibit an overall positive median for all groups. The spread of changes is relatively consistent across three groups. Outliers in Groups 2 and 3 suggest individual variations, with some students experiencing declines. Group 2's boxplot is notable for its upper outliers, pointing to exceptional improvements in math scores. The two figures suggest that while students across all income groups generally improved in both subjects, the degree of change varies, with higher income groups demonstrating a tendency for more significant improvements, especially in math.

Figure 7 Change in Reading Score by Income Group

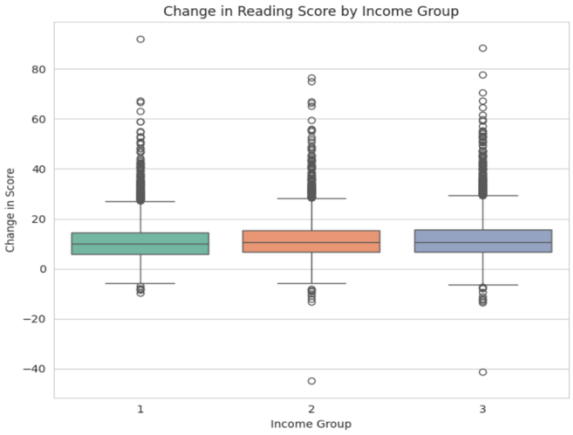
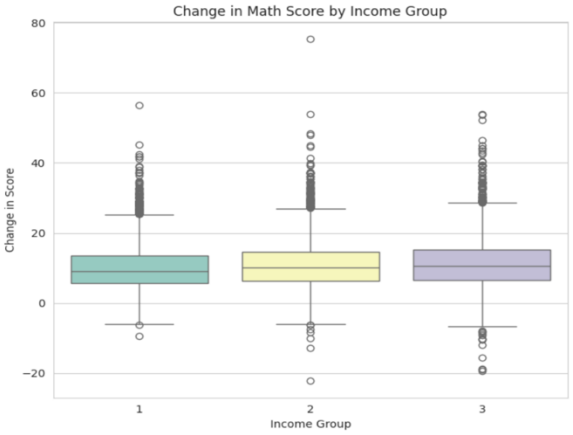


Figure 8 Change in Math Score by Income Group



To explore relationships between variables, the correlation matrix was made in **Figure 9(Appendix)**, we can see that fall reading scores have a strong positive correlation with spring reading scores, indicating that students who score well in the fall are likely to also score well in the spring. Similarly, fall math scores strongly correlate with spring math scores. General knowledge correlates with both subjects, suggesting it may be a foundational skill influencing overall academic

performance. Notably, the correlation between income and changes in academic scores is relatively weak. This suggests that while income may play a role, it is not the sole predictor of academic progress, reinforcing the need for ANCOVA to control for confounding variables like general knowledge.

One-Way ANCOVAs Analysis

Changes in Reading Scores

We conducted a one-way ANCOVA analysis with the dependent variable as the change in reading score, the independent variable as income group, and the covariate as the fall general knowledge score. We found that the impact of income group on the change in reading scores was not statistically significant, as indicated by a p-value of 0.105, which exceeds the common alpha of 0.05. The effect size for income group was also negligible, with a partial eta squared value of 0.000377. Conversely, the fall general knowledge score was a significant predictor of the change in reading scores, with an F statistic of approximately 220.11 and a highly significant p-value (2.35e-49), suggesting a strong relationship. The effect size of fall general knowledge score on reading score change was small to moderate, with a partial eta squared value of 0.018117. This analysis indicates that while income group does not significantly affect reading progress, the general knowledge that students have at the start of the school year plays a crucial role in their reading development.

Source	SS	DF	F	p-unc	np2
incomegroup	287.485906	2	2.251247	1.05E-01	0.000377
fallgeneralknowledgescore	14054.12468	1	220.11032	2.35E-49	0.018117
Residual	761671.0364	11929	NaN	NaN	NaN

Figure 10 One-Way ANCOVA results - Reading Score

Model Summary: After fitting ANCOVA to OLS model, we found that the R-squared value is 0.023, indicating that approximately 2.3% of the variance in the change in reading scores is explained by the model. This suggests that while there is some association, a large portion of the variance in reading score changes are not explained by income group and fall general knowledge scores alone. The coefficients for income groups 2 and 3 are 0.2169 and 0.4038, respectively. However, only the coefficient for income group 3 is statistically significant at the 0.05 level (p=0.035), suggesting that the average change in reading scores for students in income group 3 is significantly different from that of students in income group 1, after controlling for general knowledge scores. Income group 2's effect is not statistically significant (p=0.228).

Assumption check:

Statistics	
w	0.899631738
p-value	<0.001

Assumption 1: residuals are normally distributed

Then, we performed **Shapiro-Wilk test**, While the w value being below 1 suggests some deviation from normal distribution, it's the extremely small p-value suggests that this deviation is statistically significant. Consequently, this violates the normality assumption of our ANCOVA model, implying that the changes in reading scores may not be normally distributed across the sample. This finding highlights the need for caution when interpreting the ANCOVA results.

Statistics	
Levene's test statistic	19.72801037
p-value	2.79493E-09

#Assumption 2: variances are homogenous:

The Levene's test was conducted to verify if the assumption of homogeneity of variances was met for the change in reading scores across the three income

groups. The results showed that the variances are not homogeneous, as indicated by a significant Levene's test statistic and a p-value close to zero. This suggests that the variability of reading score changes differs between the income groups, which could potentially affect the results of the ANCOVA.

Changes in Math Scores

Then, we conducted a one-way ANCOVA analysis with the dependent variable as the change in math score, the independent variable as income group, and the covariate as fall general knowledge score. The analysis revealed that the income group had a nonsignificant effect on the change in math scores, as indicated by an F-statistic of 0.624286 and a p-value of 0.536614, which is greater than the typical alpha level of 0.05. Additionally, the partial eta squared (np2) for income group was very small (0.000105), suggesting a negligible effect size.

In contrast, the fall general knowledge score was a significant predictor of change in math scores, with an F-statistic of 501.083959 and a highly significant p-value (9.425259e-109). The effect size for the fall general knowledge score was moderate (0.040312), indicating that general knowledge in the fall is a substantial factor in math score improvement over the school year.

Source	SS	DF	F	p-unc	np2
incomegroup	55.879616	2	0.624286	5.3566E-01	0.000105
fallgeneralknowledgescore	22425.93296	1	501.08396	9.4253E-109	0.040312
Residual	533880.4998	11929	NaN	NaN	NaN

Figure 11 One-Way ANCOVA results - Math Score

Model Summary: After fitting ANCOVA to OLS model for the change in math scores, we observed an R-squared value of 0.048. Meaning that about 4.8% of the variance in the change in math scores is accounted for by the model. This percentage, although slightly higher than that for reading scores, still leaves a majority of the variance unexplained by our model's variables. Looking at the income group effects, the coefficients for income groups 2 and 3 are 0.1523 and 0.1442 respectively, but neither is statistically significant (p=0.312 for group 2 and p=0.368 for group 3). This indicates that after adjusting for general knowledge scores, there is no significant difference in the average change in math scores between students from income groups 2 or 3 compared to group 1. The fall general knowledge score, with a coefficient of 0.1993, is statistically significant (p<0.001), reinforcing its predictive value for changes in math scores similar to reading scores. It indicates that a one-unit increase in the fall general knowledge score is associated with an average increase of about 0.2 points in the change in math scores, controlling for income group.

Assumption check:

Statistics	
w	0.966404438
p-value	<0.001

Assumption 1: residuals are normally distributed

The result for the math scores, with a w statistic of 0.9664 and a p-value of less than 0.001, tells us that the residuals from the ANCOVA model do not follow a normal distribution. Although the w statistic is closer to 1, which would suggest normality, the very low p-value provides strong evidence to reject the null hypothesis that the residuals are normally distributed.

Statistics	
Levene's test statistic	22.21518018
p-value	2.34418E-10

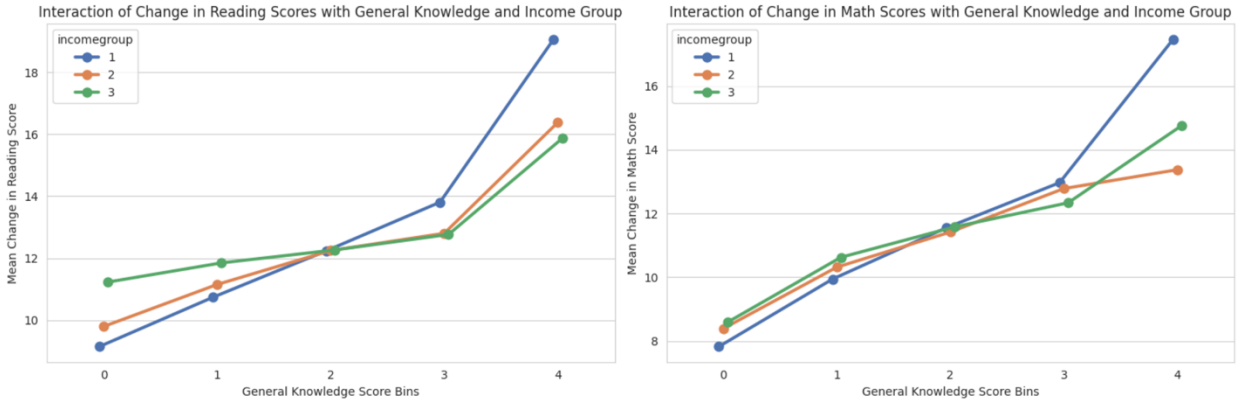
#Assumption 2: variances are homogenous: Levene's test

The results from **Levene's test** suggest that the assumption of equal variances is violated. Specifically, the

very low p-value indicates there is strong statistical evidence to reject the null hypothesis that the variances of change in math scores are the same across the different income groups. This result calls for caution in interpreting the ANCOVA results and may require adjustments to the analysis or the use of alternative statistical methods that do not assume homogeneity of variances.

Lastly, we made two **interaction plots** illustrate how changes in reading and math scores relate to students' general knowledge scores across different income groups. For **reading scores**, as general knowledge increases, all income groups show an upward trend in mean change scores, suggesting that students with higher initial general knowledge tend to have greater improvements in reading regardless. Notably, Group 3 shows a particularly steep increase, suggesting that higher income may amplify the positive impact of general knowledge on reading score improvements.

In **math scores**, we see a similar pattern, with all groups improving as general knowledge scores increase. Again, Group 3 has a steeper slope, especially from the third bin to the fourth, indicating a more pronounced effect of general knowledge on math score improvements for students from higher-income families.



Conclusion

To understand how income group impacts the change in reading and math scores from fall to spring, after controlling for general knowledge scores. The findings indicate that while income group alone does not significantly predict changes in reading or math scores, fall general knowledge scores are a robust predictor of academic improvement. Notably, the increase in reading and math scores as general knowledge scores rise is more pronounced among students from the highest income group(Group 3).

These conclusions are drawn with caution due to the limitations presented by the assumption violations. The use of alternative statistical methods might be warranted for a more accurate analysis. Nonetheless, our findings suggest that initiatives aimed at enhancing early general knowledge could be beneficial, and may also need to consider socioeconomic disparities to maximize their impact on academic growth. This study underscores the multifaceted nature of education, where both economic background and foundational knowledge intertwine to shape the learning outcomes of young students.

Appendix

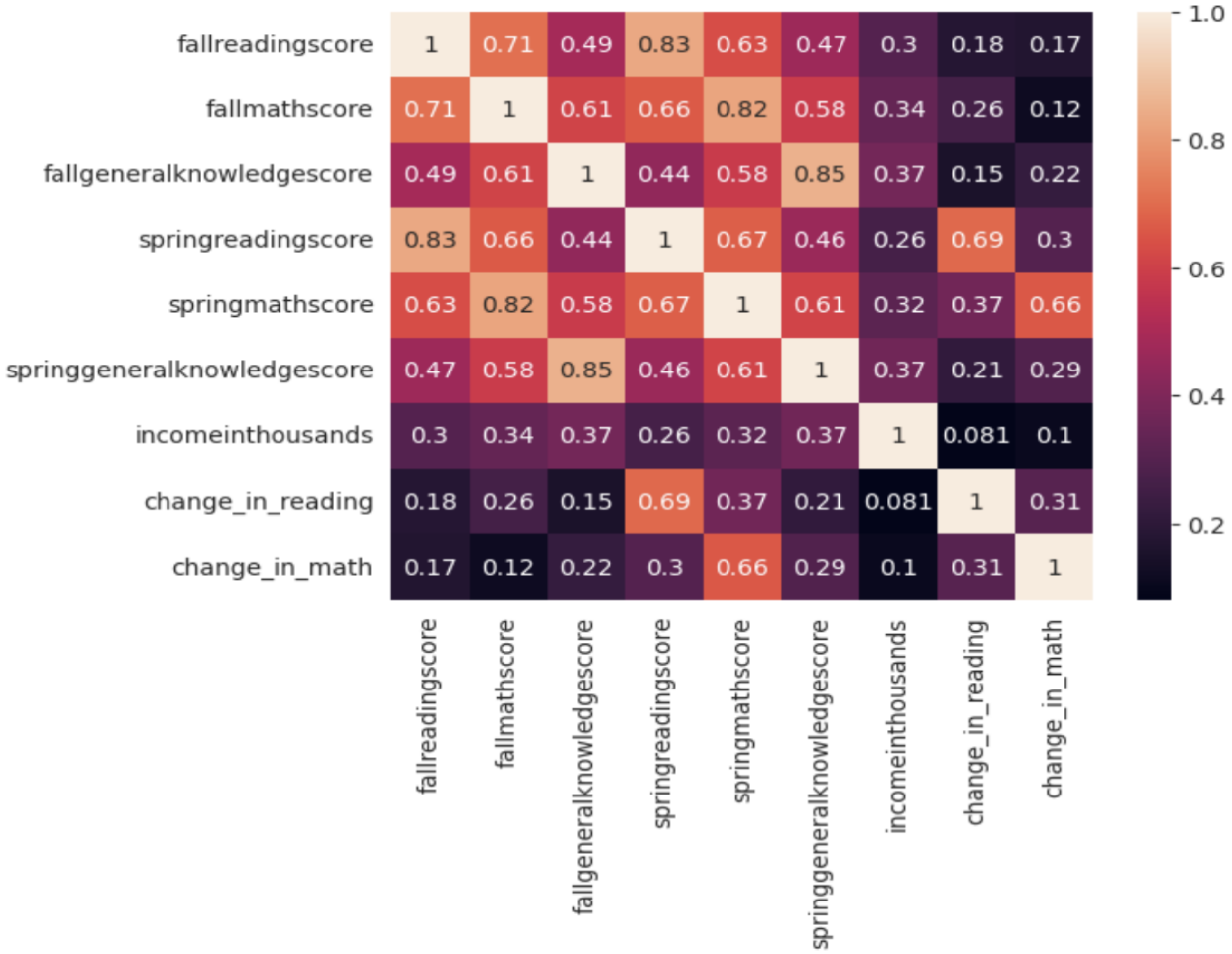


Figure 9 Correlation Matrix across variables