Technical Assignment 3:
Exploring Kindergarten Scores by Income

Devin W. de Silva
Faculty of Information
School of Graduate Studies, University of Toronto
INF2178: Experimental Design for Data Science
Professor Shion Guha
March 23, 2024

The dataset we are working with for this assignment is comes from an early childhood study over the period of a year. It provides us with reading, math, and general knowledge scores for kindergarten students during both fall and spring sessions. Crucially, it includes income category as a variable. For this analysis, I wish to explore how income influences early educational outcomes, which is why I would be using income group as my independent variable for all the analysis below. I also want to know whether students' scores in one subject can generally predict their achievements in another subject in the future. Therefore, I selected to use fall general knowledge as a baseline/covariate/control variables for all the analysis.

I wish to understand how a child's family income background influence early childhood education by learning how scores change over time among various income groups. Such an analysis, I hope, could potentially provide us insights into the lasting impact of family income on a child's future success. Information like these could be valuable in creating measures and policies on educational equity.

For this assignment, I would be focusing on three interrelated research questions:

a) **How does income influence the kindergarten students' general knowledge scores over time, when controlling for their initial scores?**

We use general knowledge scores in fall as a covariate identify the effect of income group (independent variable) on the spring general knowledge scores (dependent variable).

b) Compare changes in students' math scores over time by income group, using general knowledge scores as a baseline.

TECHNICAL ASSIGNMENT 3

We use general knowledge scores in fall as a covariate identify the effect of income group (independent variable) on the change in math scores (dependent variable).

c) Compare changes in students' reading scores over time by income group, using general knowledge scores as a baseline.

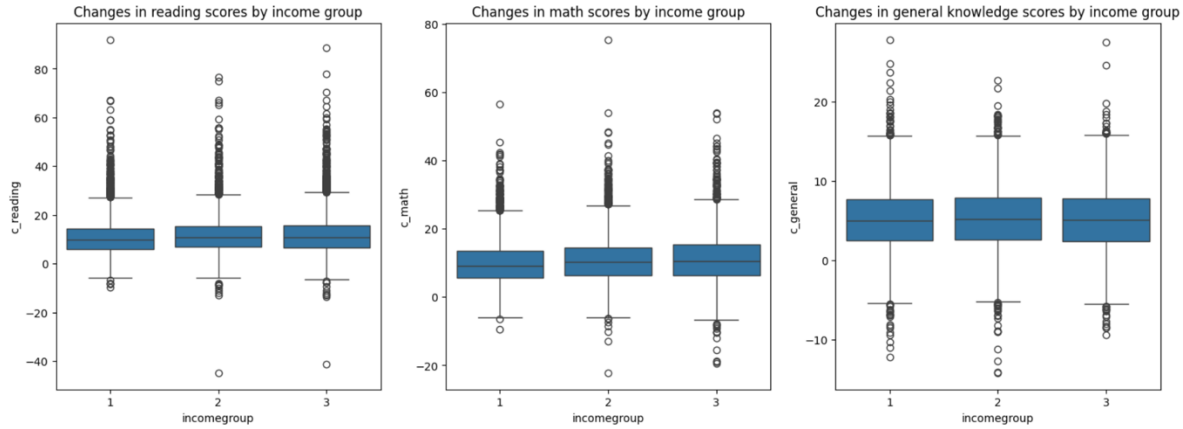
We use general knowledge scores in fall as a covariate identify the effect of income group (independent variable) on the change in math scores (dependent variable).

I begin by gathering some general information about the data using df.info(). The raw data contains 9 columns with a total of 11933 columns. After checking that there are no null values at all in this dataset, I proceeded to gather some key statistics of the dataset using df.describe().

Some key relevant information include:

- The average scores for reading, math, and general knowledge in the fall semester are roughly 35.95, 27.13, and 23.07.
- The average scores all increased in the following semester, to about 47.51 for reading, 37.80 for math, and 28.24 for general knowledge.
- The total household income ranges between $0 and $150,000, segmented into three income groups. For simplicity's sake, I would be using the categorical variable "income group" rather than the specific income for every household.

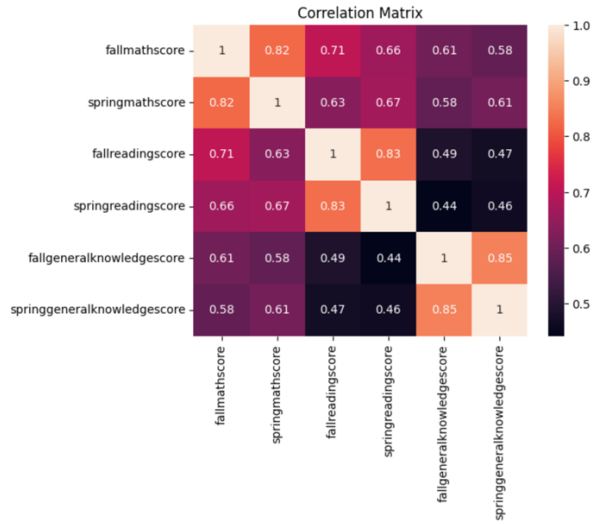
I then added three additional columns to the dataset to identify the changes in math, reading and general knowledge scores over the two semesters, by subtracting the fall scores from the spring scores. The changes are visualized below.



Interestingly, the median change in reading and math scores seem to increase marginally in higher income groups. Higher income groups also show greater variability in score changes. The change in general knowledge scores seems slightly to be more consistent irrespective of income groups.

TECHNICAL ASSIGNMENT 3

To gain a general idea about the covariates' influences on one another, I created a correlation matrix. There are some very clear correlation between fall and spring scores, both within the same subjects and between subjects.



I then proceeded to use ANCOVA to answer the three questions I set out to explore:

Research Question 1: How does income influence the kindergarten students' general knowledge scores over time, when controlling for their initial scores?

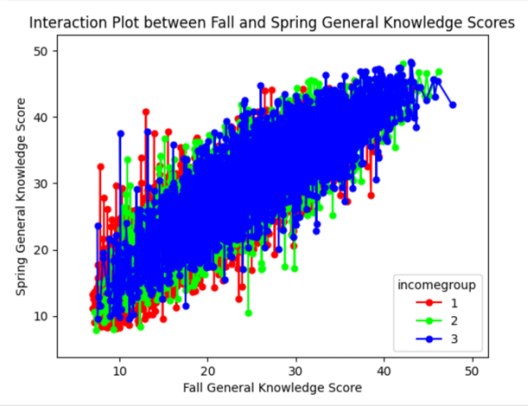
- **Null Hypothesis (H0):** Income group has no effect on the kindergarten students' spring general knowledge scores when controlling for their fall scores.
- **Alternative Hypothesis (H1):** Income group has an effect on the kindergarten students' spring general knowledge scores when controlling for their fall scores.

The one-way ANCOVA test provided **a F-statistic of 56.9 and an extremely small p value of < 0.001** for income groups, much smaller than the alpha value of 0.05, showing that the null hypothesis is rejected, as differences in general knowledge scores across income groups are statistically significant when fall general knowledge scores are controlled.

For the covariate fallgeneralknowledgescore, **the F-statistic is around 26682 and the p value is also extremely small at < 0.001**, meaning that fall general knowledge scores are a very good predictor of spring general knowledge scores.

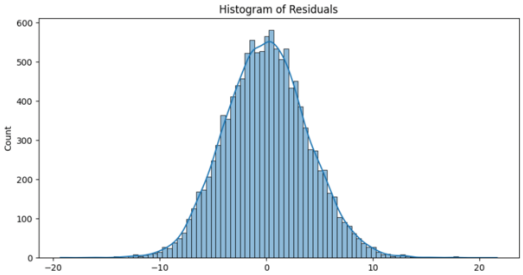
To ascertain these influences, I plotted an interaction plot between fall and spring general knowledge scores, based on income groups, and a clear, generally linear relationship can be seen across income groups.

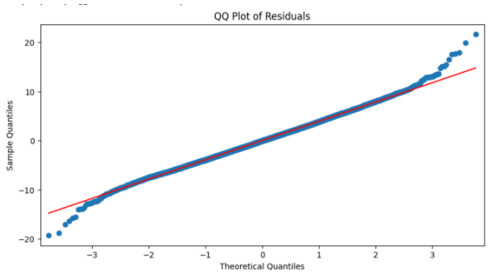
TECHNICAL ASSIGNMENT 3



I then tested the two main conditions for this ANCOVA, namely normality of residuals and homogeneity of variances, using the Shapiro-Wilks test and the Levene's test respectively.

The Shapiro-Wilk test provided a statistic of approximately 0.998 with a p-value of < 0.001. This shows that while the normality assumption might not hold, despite the histogram of residuals graph seemingly relatively normally distributed and the qq plot only diverging from regression line near the ends.





The Levene's test across income groups also provided a p-value less than 0.001. This means that the homogeneity of variances assumption is also violated, hence making our results slightly less robust.

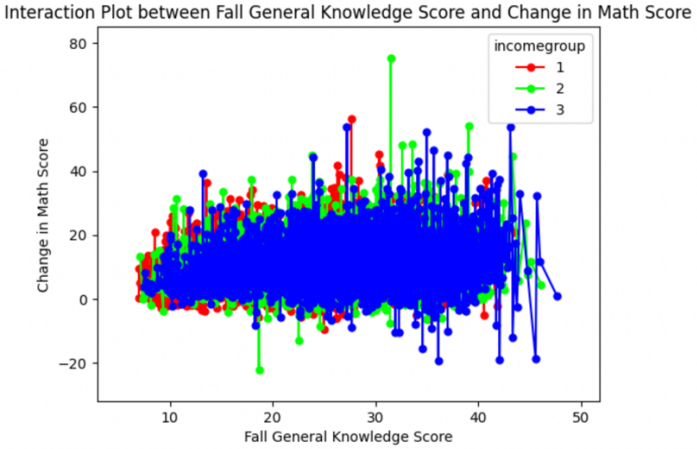
Research Question 2: How students' math scores change over time by income group, using fall general knowledge scores as a baseline.

- **Null Hypothesis (H0):** Changes in math scores over time are not influenced by the income group when controlling for fall general knowledge scores.
- **Alternative Hypothesis (H1):** Changes in math scores over time are influenced by the income group when controlling for fall general knowledge scores.

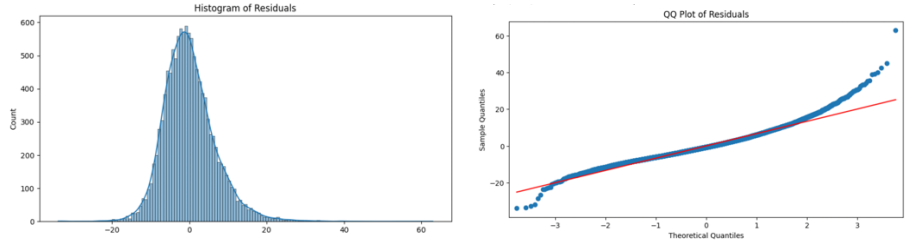
At **F = 0.624 and p = 0.536** > 0.05, the ANCOVA shows that the F-statistic and p-value for the income group were not statistically significant. This shows that the null hypothesis stands, and that the income group does not significantly impact changes in math scores when controlling for fall general knowledge scores.

With that said, as with the first research question, at **F = 501.084, p < 0.001**, we can see that the fall general knowledge score is a highly significant predictor of changes in math scores, which confirms the 0.58 correlation shown on the correlation matrix earlier.

The interaction plot below confirms how income group do not significantly differentiate how fall general knowledge score predicts math score changes.



The next step, again, is to test the assumptions of the ANCOVA model. The Shapiro-Wilk test statistic was approximately 0.966 with a p-value of < 0.001. This shows that the normality of residuals assumption is not met. The skewed histogram and qq plot confirm these results.



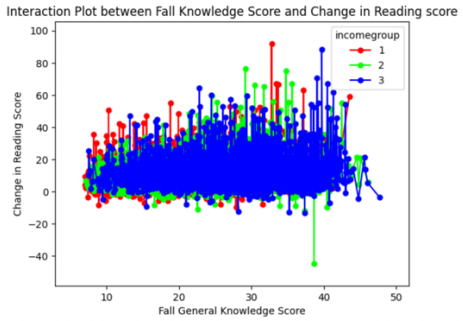
Similarly, the Homogeneity of Variances assumption is not met either, as the statistic was approximately 22.22 with a p-value < 0.001.

Research Question 3: How students' reading scores change over time by income group, using fall general knowledge scores as a baseline.

- **Null Hypothesis (H0):** Changes in reading scores over time are not influenced by the income group when controlling for fall general knowledge scores.
- **Alternative Hypothesis (H1):** Changes in reading scores over time are influenced by the income group when controlling for fall general knowledge scores.

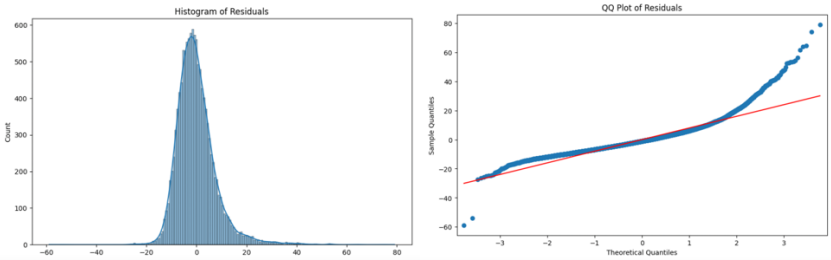
The ANCOVA results are very similar to the last test. The F-statistic for the income group was not significant at **F = 2.251 and p = 0.105** > 0.05, meaning that the null hypothesis stands, as income group does not significantly influence changes in reading scores either, when fall general knowledge is controlled. Like the last test, the fall general knowledge score, however, significantly affected reading score changes (F = 220.110, p < 0.001), meaning that general

knowledge level is more of a predictor of changes in reading scores than the income group, confirming the 0.44 correlation shown on the correlation matrix earlier.



This relatively parallel interaction plot also confirms the statistics, as no clear pattern can be seen between income group and initial knowledge level in relation to the improvement in reading scores.

As with the earlier research questions, I also conducted Shapiro-Wilk and Levene's tests to confirm the results of the ANCOVA. With a Shapiro-Wilk test statistic around 0.90 and a p-value of < 0.001, the normal distribution of residuals assumption is not met. The QQ plot and a skewed histogram of residuals confirm these results.



Similarly, the Levene's test statistic was approximately 19.73 with a p-value of < 0.001, also showing that the homogeneity of variances assumption across income groups for changes in reading scores is not met.

Conclusion

In conclusion, our analysis of kindergarten students' scores over time highlights that family income do significantly impact student scores. In general knowledge scores, when fall scores are considered, the trajectory of student scores over time is evidently different. Students' scores in one subject also seem to be a predictor of scores in other subjects in the future. However, income's influence doesn't necessarily affect every single element of the child's grades. Income's influence on changes in math and reading scores was not significant when considering fall general scores. It is also worth noting that these results may not be completely robust, as some of the main assumptions of ANCOVA are not met in these studies. Despite these limitations, it remains clear that we need to understand the systematic ways that income influence childhood educational attainment since a very young age, and it is important that we devise and implement more equitable educational policies to address them.