Lan Li
1005814326

Exploring Childhood Education

1. Introduction

This report delves into the field of early childhood education outcomes, employing a longitudinal dataset from a study conducted between 1998 and 1999. Through exploratory data analysis, interaction plots, ANCOVA test and assumption testing, we propose to interpret the complex relationships between academic performance and income levels, based on the dataset called 'INF2178_A3_data.csv'.

The paper will be around three research questions:

1. How do spring general knowledge scores vary by income group after controlling for fall general knowledge scores? Essentially, it's looking at whether the income group has a significant effect on spring scores when the fall scores are held constant.
2. Does the relationship between fall general knowledge scores and spring general knowledge scores differ among the various income groups? In other words, is the effect of fall general knowledge on spring scores moderated by income group?
3. What should be the result of the same questions as the previous two research questions, when we investigate math scores instead of general knowledge scores.

By addressing these questions, the paper would contribute valuable insights into the effectiveness of different agency types and funding mechanisms in providing child care services.

2. Data Wrangling

The dataset totally includes **9 columns** and **11,933 rows**, but this report only investigates part of the columns. The investigated columns all have **11,933** non-null, thus there is no data cleaning needed. The detailed information of target columns is listed below:

- fallreadingscore: reading scores of students assessed during the fall in 1998;
- fallmathscore: math scores of students evaluated during the fall in 1998;
- fallgeneralknowledgescore: broader assessment of students' knowledge in various areas during the fall in 1998;
- springreadingscore: reading scores of students assessed during the spring in 1999;
- springmathscore: math scores of students evaluated during the spring in 1999;
- springgeneralknowledgescore: broader assessment of students' knowledge in various areas during the spring in 1999;
- incomegroup: a categorical variable derived from the total household income data, which groups income into categories (like 1, 2, 3)

3. Exploratory Data Analysis (EDA)

The summary statistics of the dependent variable – spring general score – exhibits a trend that a child in an income group with a higher level has a higher general knowledge score. The boxplots in Figure 2 shows the distribution of scores for general knowledge, math, and reading, categorized by income group for a spring assessment. The consistency in median values across income groups might initially suggest that income level does not have a strong influence on spring scores. However, there is a slight trend showing that when the level of the income group increases, the spring scores also mildly increase.

In Figure 3, there appears to be a positive correlation between fall and spring general scores; as fall scores increase, so do the spring scores. This suggests that students who scored higher in the fall also tend to score higher in the spring. Although the distribution of the scores seems overlapped, we still can find the patterns that income group 3 have relatively higher spring and fall general knowledge scores than other groups.

Income Group	count	mean	std
1	4729	25.069	7.248
2	3726	29.144	6.965
3	3478	31.568	6.928

Figure 1: Summary of Dependent Variable – Spring General Score

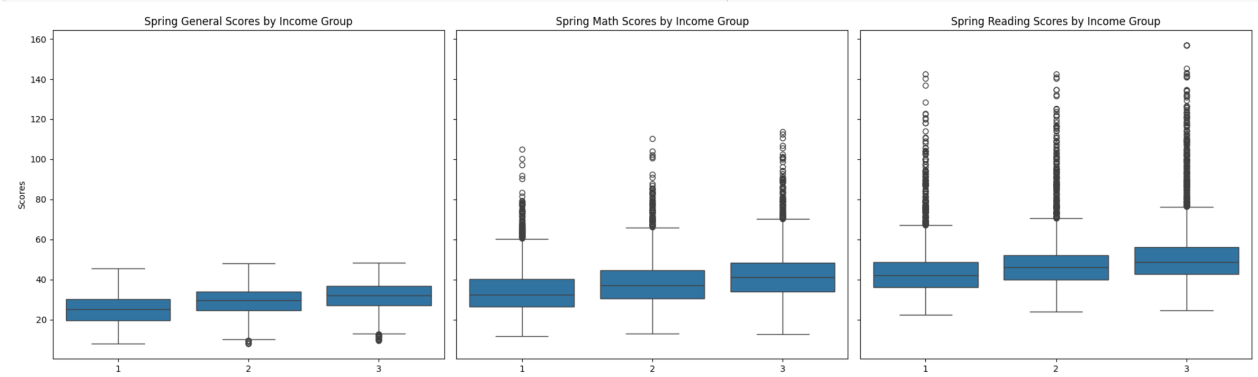

Figure 2: Boxplots of Different Scores in Spring

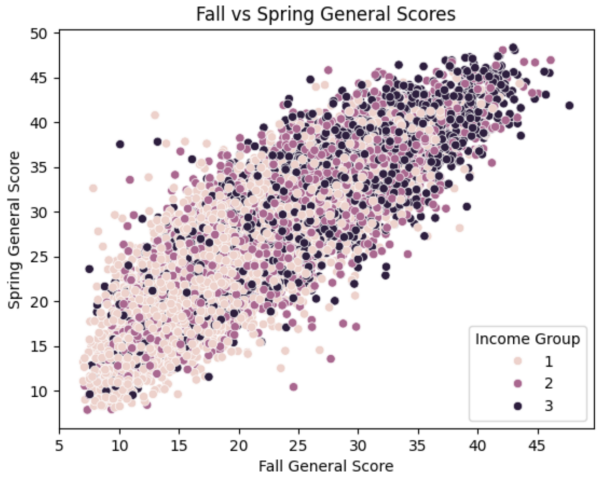

Figure 3: Scatterplot of Spring General Score vs. Fall General Score

4. General Knowledge Scores ANCOVA

ANCOVA would help answer questions about the impact of income group on spring general knowledge scores after controlling for the baseline scores in the fall. It adjusts for the initial knowledge level to more accurately measure the effect of income group on the spring scores, potentially leading to insights into how socioeconomic factors influence educational outcomes over and above the students' starting knowledge levels.

Research Question 1: How do spring general knowledge scores vary by income group after controlling for fall general knowledge scores?

One-Way ANCOVA without Interaction
The p-value in C(incomegroup)[T.2] suggests that there is a statistically significant difference in spring general knowledge scores between income group 1 and income group 2. The p-value in C(incomegroup)[T.3] indicates the difference in spring general knowledge scores between income group 1 and income group 3 is statistically significant. The p-value for the fall general knowledge score shows that this covariate is a statistically significant predictor of the spring general knowledge score.

These p-values provide evidence that both the income group and the fall general knowledge score have significant impacts on the spring general knowledge score. The analysis suggests that after adjusting for students' initial knowledge as measured in the fall, their income group is still a significant factor in their knowledge scores in the spring.

Variable	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.0303	0.119	67.519	<0.001	7.797	8.263
C(incomegroup)[T.2]	0.7084	0.088	8.005	<0.001	0.535	0.882
C(incomegroup)[T.3]	0.9424	0.094	10.013	<0.001	0.758	1.127
fallgeneralknowledgescore	0.8542	0.005	163.347	<0.001	0.844	0.864

Figure 4: ANCOVA Result (General Score without interaction)

Research Question 2: Does the relationship between fall general knowledge scores and spring general knowledge scores differ among the various income groups?

One-Way ANCOVA with Interaction
For C(incomegroup)[T.2]: fallgeneralknowledgescore, the p-value for the interaction term between income group 2 and fall general knowledge score is <0.001, suggesting that the interaction effect is significant. This means the relationship between fall general knowledge score and the dependent variable is significantly different for income group 2 compared to the reference group. The p-value of C(incomegroup)[T.3]: fallgeneralknowledgescore also indicates a significant interaction effect, meaning the relationship between fall general knowledge score and the dependent variable is significantly different for income group 3 compared to the reference group. The rest variables can be explained by the same way as the explanation in “One-Way ANCOVA without Interaction”.

Variable	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.1532	0.179	40.027	<0.001	6.803	7.504
C(incomegroup)[T.2]	1.9328	0.293	6.591	<0.001	1.358	2.508
C(incomegroup)[T.3]	2.8491	0.313	9.114	<0.001	2.236	3.462
fallgeneralknowledgescore	0.8982	0.008	105.783	<0.001	0.882	0.915
C(incomegroup)[T.2]: fallgeneralknowledgescore	-0.0585	0.013	-4.632	<0.001	-0.083	-0.034
C(incomegroup)[T.3]: fallgeneralknowledgescore	-0.0829	0.013	-6.557	<0.001	-0.108	-0.058

Figure 5: ANCOVA Result (General Score with interaction)

The interaction plot shows the lines are not parallel (though they are close), suggesting that the increase in the spring score with respect to the fall score is not uniform across income groups. This indicates an interaction effect, where the relationship between the fall and spring scores depends on the income group. This is consistent with the negative interaction terms found in the ANCOVA results, indicating that the impact of fall scores on spring scores is slightly reduced for higher income groups. The result indicates that the benefit of higher fall scores on spring scores is somewhat less for higher income groups.

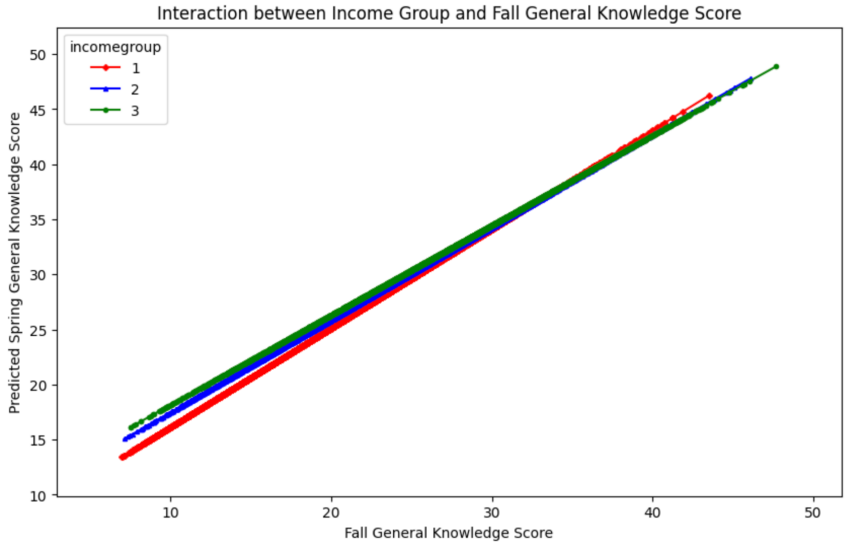

Figure 6: Interaction Plot of General Knowledge Score

Limitation

Shapiro-Wilk Test for Normality: The p-value which is (< 0.001) leads us to reject the null hypothesis that the data is normally distributed. Hence, in this case, the test indicates that the residuals are not normally distributed.

Levene's Test for Equal Variances: This result ($p < 0.001$) suggests that the variances are not equal across groups, and therefore, the assumption of homogeneity of variances is violated.

The assumptions are violated, it may affect the validity of the ANCOVA results, and alternative methods or transformations of data might be necessary to proceed with the future improved analysis.

5. Math Scores ANCOVA

Research Question 3: What should be the result of the same questions as the previous two research questions, when we investigate math scores instead of general knowledge scores.

One-Way ANCOVA without Interaction
The ANCOVA model results suggest that both income group and previous math score are significant predictors of the spring math score, with a very strong relationship between fall and spring math scores. The model suggests that as students' fall math scores increase, their spring scores also increase, and that being in a higher income group is associated with a higher spring math score after controlling for fall math scores.

Variable	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.2011	0.199	41.273	<0.001	7.812	8.591
C(incomegroup)[T.2]	0.6700	0.151	4.430	<0.001	0.374	0.966
C(incomegroup)[T.3]	0.9199	0.160	5.741	<0.001	0.606	1.234
fallmathscore	1.0735	0.007	149.007	<0.001	1.059	1.088

Figure 7: ANCOVA Result (Math Score without interaction)

One-Way ANCOVA with Interaction
The significant p-values for the interaction terms suggest that the slopes of the relationship between fall math score and spring math score differ between the reference income group and groups 2 and 3, and these differences are unlikely due to random chance. This implies that the benefit of higher fall math score is not uniform across income groups; instead, it's somewhat attenuated in higher income groups.

Variable	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.7261	0.325	20.727	<0.001	6.090	7.362
C(incomegroup)[T.2]	2.4488	0.497	4.931	<0.001	1.475	3.422
C(incomegroup)[T.3]	3.6637	0.498	7.359	<0.001	2.688	4.640
fallmathscore	1.1351	0.013	87.852	<0.001	1.110	1.160
C(incomegroup)[T.2]: fallmathscore	-0.0727	0.018	-3.961	<0.001	-0.109	-0.037
C(incomegroup)[T.3]: fallmathscore	-0.1026	0.017	-5.908	<0.001	-0.137	-0.069

Figure 8: ANCOVA Result (Math Score with interaction)

This interaction plot suggests that while all students improve, the rate of improvement as measured from fall to spring differs by income group. Specifically, income group 1 sees the most substantial gain in spring math scores for each unit increase in fall math scores, with groups 2 and 3 experiencing slightly less gain. This visualization supports the conclusion that both fall math scores and income group are important factors in predicting spring math scores, and their interaction should be considered when interpreting the results.

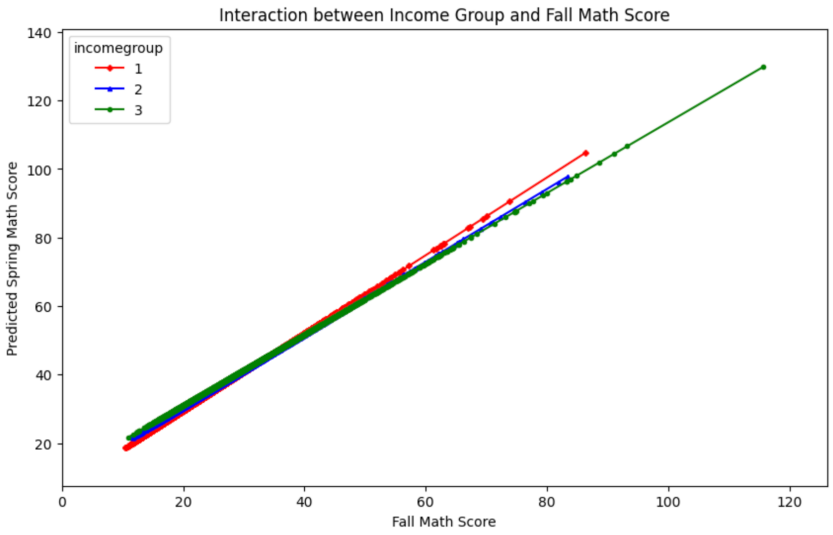

Figure 9: Interaction Plot of Math Score

Limitation

Shapiro-Wilk Test for Normality: The p-value which is (< 0.001) indicates that the residuals of the model do not follow a normal distribution, violating the assumption of normality required for ANCOVA.

Levene's Test for Equal Variances: This result ($p < 0.001$) suggests that the assumption of homogeneity of variances is violated, meaning the variances of the residuals for the different income groups are not equal.

These violations could impact the validity of any conclusions drawn from the ANCOVA, and alternative methods or modifications to the assumptions might be required for future study.

6. Conclusion

Using a methodology that included exploratory data analysis, interaction plots, and ANCOVA tests, we showed that income grouping and prior knowledge significantly affected students' spring general knowledge and math scores. Notably, income grouping affected spring general knowledge and math scores even after considering initial fall scores. This is further illustrated by the interaction effect between income group and fall scores, whereby the effect of higher initial scores on spring achievement diminishes slightly as income levels increase. Despite the limitations, the study offers a compelling narrative about the persistent influence of socioeconomic factors on educational achievement, and it underscores the importance of considering such factors in educational policy and practice.