An analysis of the impact of family income on academic achievement at the kindergarten level.

Chenwei Zhu
1006003619

Introduction

The correlation between socioeconomic status and the academic achievement of young children has emerged as an important area of research in evolving educational research. This analysis draws on data from an early childhood longitudinal study conducted during the 1998-99 school year that focused on a cohort of kindergarten students. The dataset titled "INF2178 A3 data.csv" provides rich empirical information that encapsulates the trajectories of students' academic abilities over several months as evidenced by their reading, math, and general knowledge scores. At its core, this investigation explores how socioeconomic factors, represented by income categories, interact with educational outcomes. Although income is a continuous variable in the dataset, it is divided into discrete income groups for clarity of analysis. This choice of methodology allows us to examine the impact of socioeconomic status on academic achievement. This will be divided into three research questions to explore the impact of socioeconomic status on academic achievement from different perspectives:

Research Question 1: whether there is an effect of family income group on children's reading achievement?

Research Question 2: whether there is an effect of family income group on children's achievement in math?

Research Question 3: whether there is an effect of family income group on children's overall achievement?

This study will examine the effect of family income group on children's spring grades using fall grades as a covariate using the One-Way ANCOVA model. The significance of this study lies not only in its contribution to the scholarly discourse on education and inequality, but also in its potential to inform policy and instructional strategies aimed at closing the education gap.

Data Cleaning and Data Wrangling

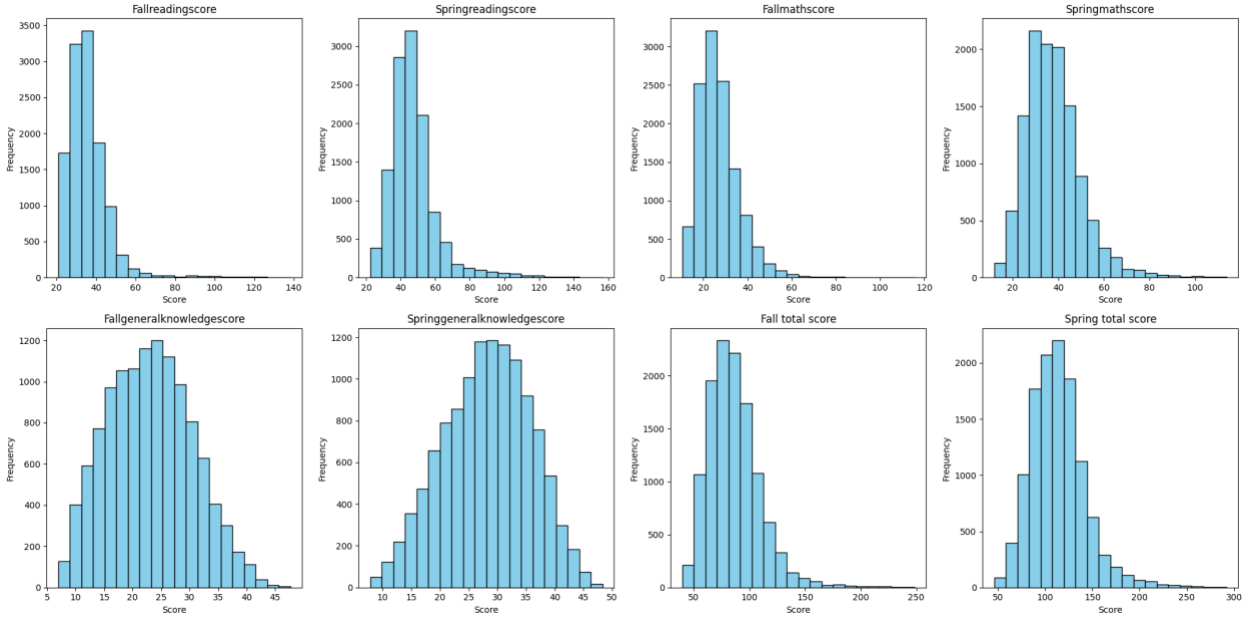
In our study of the dataset, we first carefully cleaned and organized the raw data to ensure the accuracy and validity of the analysis. The raw data were organized to ensure the accuracy and validity of the analysis, to identify and correct errors or inconsistencies in the data, and to lay a solid foundation for the data analysis. The raw data set contained 11933 rows of entities and 9 columns of variables. Our initial review indicated that the original dataset did not require extensive data cleaning within the scope of our analysis. However, we noted some observed differences and accordingly defined new features needed for future analysis. Here are the steps we took to organize the data.

Steps taken to organize the data:

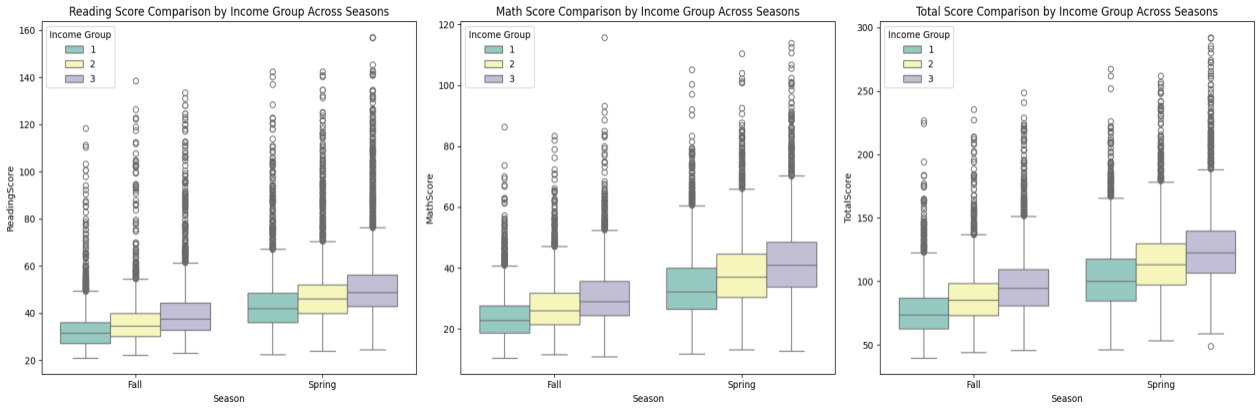
1. check for missing values: no null values in the dataset
2. Create additional characteristic variables needed for the study, "spring_total_score" and "fall_total_score", which are derived from summing the scores of each subject in the spring and fall, representing the total scores in the spring and fall, which are derived from the sum of the spring and fall subject scores, respectively, and represent the spring and fall total scores.

Exploratory Data Analysis (EDA)

First, descriptive analyses and histograms [Figure 1] were used to explore trends in the distribution of scores across subjects and total scores during the spring and fall seasons. The graph shows that in the fall 1998 and spring 1999 grades, the trend of the distribution of the subject grades and the total grades, except for the General Knowledge grades, was left-skewed and did not follow a normal distribution. And the spring grades were generally higher than the fall grades.



[Figure 1]



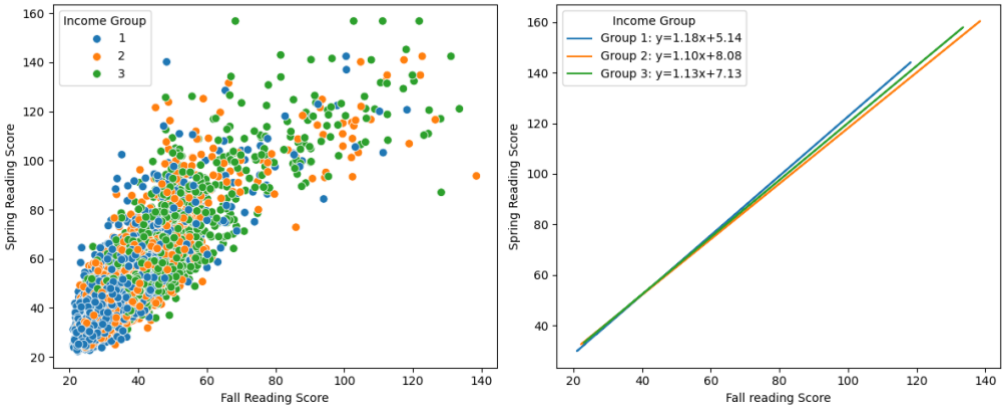
[Figure 2]

Box plots of students' spring and fall scores [Figure 2] were then plotted by family income group to explore whether there were differences in the scores of children in different income groups. In the fall, median reading scores were lower for all groups, with the lowest scores for the lowest income group. By spring, students in all groups showed significant improvement, especially in the highest income group. This suggests that improvements in reading proficiency are seasonal and may be influenced by educational programs or natural progression. Students with higher incomes performed better, suggesting a correlation between income and academic achievement. Math scores showed the same trend, with the highest income group maintaining the highest median scores in both seasons. Total scores reflected these trends, with all groups making progress from

fall to spring, with higher incomes associated with better overall academic performance. Despite the improvement in total scores, income-related achievement gaps remain significant.

Research Question 1

In this study, we will use the most recent grades (i.e., Spring 1999 grades) as a measure of the child's learning outcomes. First we will plot a scatter plot and regression line [Figure 3] for the dependent variable and potential covariates to explore whether there is a potential correlation between them. The graph shows that there may be some correlation between the dependent variable (spring reading score) and the potential covariate (fall reading score), i.e., the previous level of learning will have an impact on this achievement. Therefore, when we analyze the effect of income group on children's reading score, we need to use the ANOCA model and set the fall achievement as a covariate to control the effect of potential confounding variables.



[Figure 3]

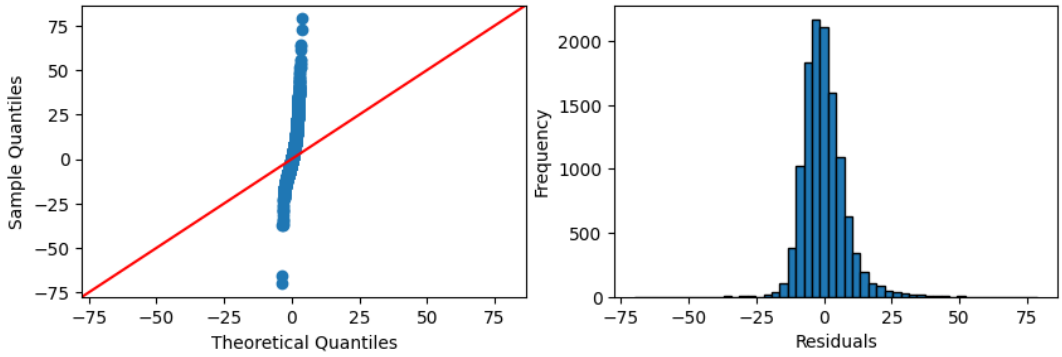
A One-Way ANCOVA model was developed to analyze spring reading scores as the dependent variable, family income group as the independent variable, and fall reading scores as the covariate. Based on the results of the ANCOVA analysis [Table 1] provided, we can draw the following conclusions: The categorical dependent variable for different income groups has a corresponding F-statistic of 4.055660, while the corresponding p-value (0.017348) is less than 0.05, which indicates that there is a statistically significant difference between the different income groups on the dependent variable. Fall reading scores as a covariate has a very high corresponding F-statistic (24455.397576) while the corresponding p-value less than 0.0001. Therefore, fall reading scores have a very strong predictive power on the dependent variables of our study. and that this effect is statistically extremely significant. In summary, this ANCOVA model suggests that when controlling for the effect of fall reading scores, students in different income groups exhibited significant differences in the dependent variables of the study.

	sum_sq	df	F	PR(>F)
C(incomegroup)	5.131201e+02	2.0	4.055660	0.017348
fallreadingscore	1.547042e+06	1.0	24455.397576	0.000000
Residual	7.546256e+05	11929.0	NaN	NaN

[Table 1]

To ensure the accuracy of the results of the study, we will also test whether ANCOVA's Assumption was met in this study. Assumption 1: There should be a linear relationship between the covariates and the dependent variable. This is true in the previous figure [Figure 3] Assumption 2: The residuals should approximately follow a normal distribution. The results from the Q-Q Normal Plot [Figure 4] and the Shapiro-Wilk test (p - value < 0.0001) show that the residuals do

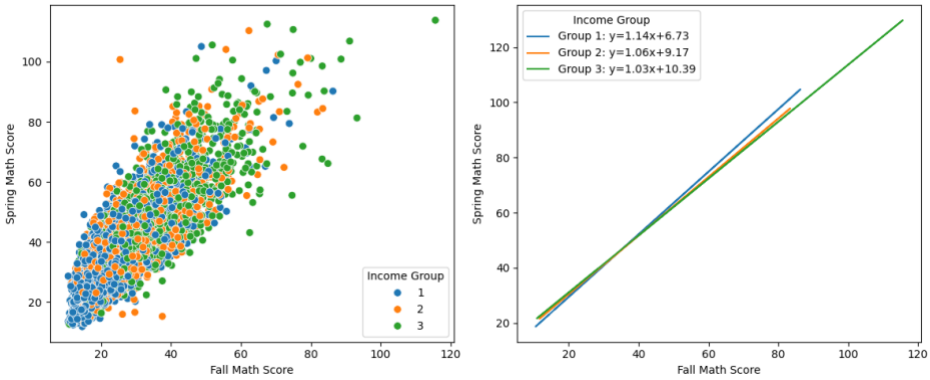
not follow a normal distribution. Assumption2 is not true. Assumption 3: Homoskedasticity: combinations of independent variables and covariates at different levels should have equal variance between them. The result of Levene's test shows p-value (7.5976e-11) < 0.05, which indicates that there is a difference between combinations of independent variables and covariates at different levels of the hierarchy. Assumption3 is not valid.



[Figure 4]

Research Question 2

The research steps in this question are roughly the same as in the first research question. First, we will plot a scatter plot and regression line [Figure 5] for the dependent variable and potential covariates to explore whether there is a potential correlation between them. The graph shows that there may be some correlation between the dependent variable (spring math score) and the potential covariate (fall math score). Therefore, a One-Way ANCOVA model was developed with spring math scores as the dependent variable, fall math scores as the covariate, and family income groups as the independent variables.



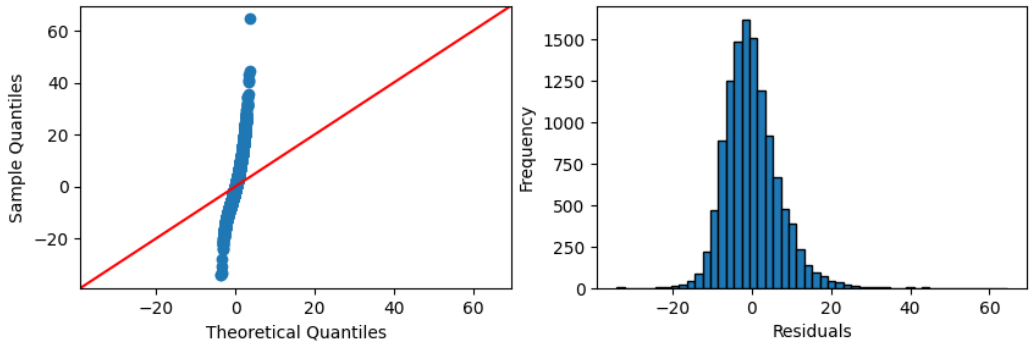
[Figure 5]

The result of ANOCVA [Table 2] shows that: The different income groups and corresponds to a p-value (9.28e-09) < 0.05. this indicates that there is a statistically significant difference in spring reading scores between the different income groups after controlling for the effect of fall math scores. The fall math score as a covariate. the p-value is less than 0.0001, which implies that fall math score has a very strong effect on the dependent variable of our study (spring reading score), and that this effect is extremely statistically significant. In summary, this ANCOVA analysis indicates that there is a significant difference in spring reading scores among students in different income groups after controlling for the effect of fall math scores.

	sum_sq	df	F	PR(>F)
C(incomegroup)	1.712758e+03	2.0	18.523585	9.284861e-09
fallmathscore	1.026489e+06	1.0	22203.081238	0.000000
Residual	5.514994e+05	11929.0	NaN	NaN

[Table 2]

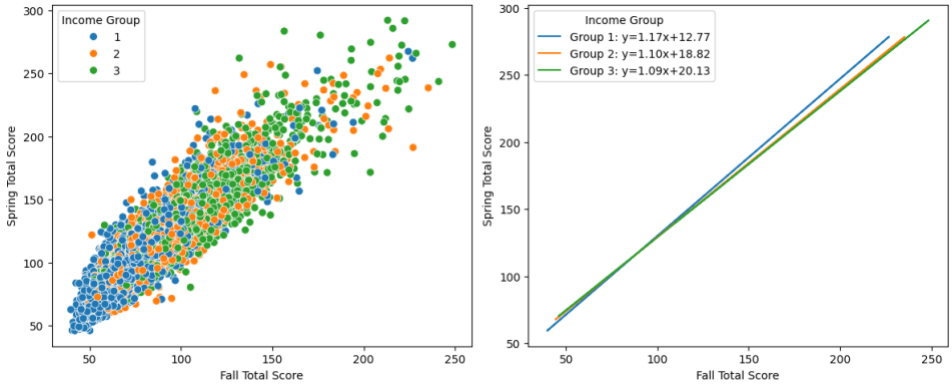
To ensure the accuracy of the results of the study, we will also test whether ANCOVA's Assumption was met in this study. Figure 5 confirms the existence of a linear relationship between the dependent variable and the covariates. The results of the Q-Q normality plot [Figure 6] and the Shapiro-Wilk test (with a p-value less than 0.0001) indicate that the residuals do not follow a normal distribution. The results of the Levene's test show a p-value of (2.1034e-12) < 0.05, which suggests that there is a difference between the combinations of independent and covariates at different levels of the hierarchy Assumption2 and 3 are not valid.



[Figure 6]

Research Question 3

The research steps in this question are roughly the same as in the previous research question. First, we will plot a scatter plot and regression line [Figure 7] for the dependent variable and potential covariates to explore whether there is a potential correlation between them. The graph shows that there may be some correlation between the dependent variable (spring total scores) and the potential covariate (fall total scores). Therefore, a One-Way ANCOVA model was developed with spring total scores as the dependent variable, fall total scores as the covariate, and family income groups as the independent variables.



[Figure 7]

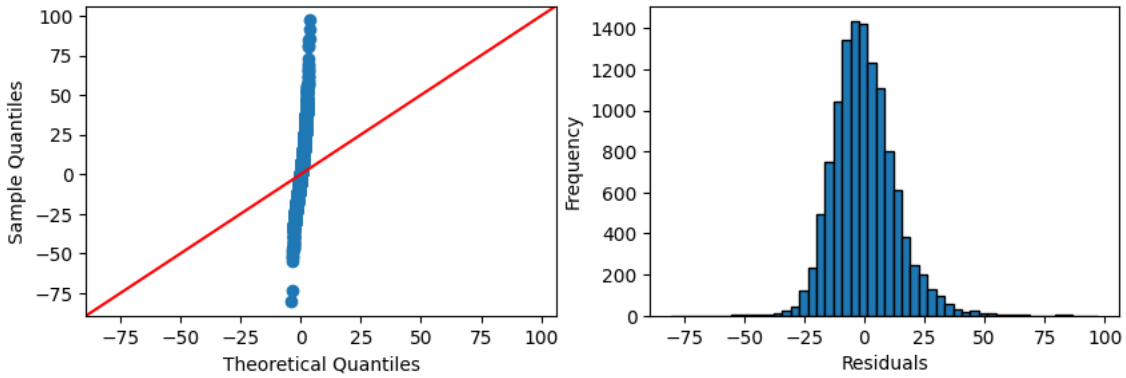
The result of ANOCVA [Table 3] shows that: The different income groups and corresponds to a p-value (0.128156) > 0.05. this indicates that there is not statistically significant difference in spring reading scores between the different income groups after controlling for the effect of fall

math scores. The fall math score as a covariate. the p-value is less than 0.0001, which implies that fall math score has a very strong effect on the dependent variable of our study (spring reading score), and that this effect is extremely statistically significant. In summary, this ANCOVA analysis indicates that Total fall scores were significant predictors of spring grades, while income groupings had a non-significant effect on spring grades after controlling for total fall scores. This may indicate that income level has a limited additional explanatory role in spring academic achievement when overall fall academic achievement is considered.

	sum_sq	df	F	PR(>F)
C(incomegroup)	7.012567e+02	2.0	2.054860	0.128156
fall_total_score	6.976113e+06	1.0	40883.557448	0.000000
Residual	2.035489e+06	11929.0	NaN	NaN

[Table 3]

To ensure the accuracy of the results of the study, we will also test whether ANCOVA's Assumption was met in this study. Figure 7 confirms the existence of a linear relationship between the dependent variable and the covariates. The results of the Q-Q normality plot [Figure 8] and the Shapiro-Wilk test (P- value = 3.1809e-43) indicate that the residuals do not follow a normal distribution. The results of the Levene's test show a p-value of (1.0069e-05) < 0.05, which suggests that there is a difference between the combinations of independent and covariates at different levels of the hierarchy Assumption2 and 3 are not valid.



[Figure 8]

Conclusion

An investigation of the relationship between socioeconomic status, as represented by income groups, and academic achievement yielded insights. Research Question 1 examined the effects of family income on children's reading achievement. ANOVA results showed significant differences between income groups after controlling for fall reading achievement, suggesting that income may have an impact on reading achievement. Research Question 2 focused on math achievement and the results also showed significant differences between income groups. No significant differences were found between income groups for research question 3, which dealt with overall academic achievement. This could mean that family income in some will have some impact on student achievement in some subjects, but that this impact is not necessarily equally present in overall scores. This phenomenon may be due to Simpson's Paradox and deserves subsequent in-depth study. However, it is important to note that some of the assumptions of the ANOVA were not met, particularly the normality of the residuals and the homogeneity of the variance across groups. These violations may affect the robustness of the findings. Future research could explore alternative statistical methods or data transformations to address these issues.