INF 2178-A3
Mingrui Fu 1010506551

# Analyze Kindergarten Scores

## 1. Introduction

In this report, we delve into the exploration and analysis of a dataset using one-way Analysis of Covariance (ANCOVA). The dataset under investigation originates from a subset of an early child longitudinal study conducted between 1998 and 1999. It encompasses data on reading, math, and general knowledge scores obtained during the fall and spring of 1998 and 1999, evaluating the academic progress of kindergarten students over several months. Additionally, the dataset includes information on household income categories, serving as the sole categorical variable.

The primary dataset utilized in this analysis is labeled "lNF2178-A3-data.csv". It comprises six continuous variables, representing individual students' fall and spring reading, math, and general knowledge scores. Furthermore, while household income is initially provided as a continuous variable, it is utilized to derive the income group variable, which categorizes individuals based on their income level.

The main objective of this assignment is to scrutinize kindergarten students' academic scores in relation to household income while accounting for their initial academic abilities. We aim to investigate the relationship between household income and academic performance over the school year, controlling for the students' fall scores. Additionally, we seek to determine if there exist significant disparities in academic growth among students from varying income groups, adjusting for their initial academic abilities:

1. **Research Question #1:** How does household income relate to the academic performance (reading, math adn general knowledge) of kindergarten students over the course of a school year, when controlling for initial academic abilities (fall scores)?
2. **Research Question #2:** Is there a significant difference in the academic growth (measured by changes in reading and math and general knowledge scores from fall to spring) among kindergarten students from different income groups, after adjusting for their initial academic abilities (fall scores)?

The analysis process involves comprehensive exploratory data analysis (EDA), model fitting using one-way ANCOVAs, creation of interaction plots to visualize relationships, and rigorous testing of assumptions necessary for running ANCOVAs. Through this process, we intend to derive meaningful insights into the interplay between household income, academic performance, and academic growth among kindergarten students.

## 2. Data Exploration and Wrangling

The initial phase of our analysis involves the preparation and exploration of the dataset. We begin by importing essential libraries for data manipulation, statistical analysis, and visualization. These libraries include tools for handling data, conducting statistical modeling, and creating visualizations. After library importation, we load the dataset from a CSV file. This dataset contains information pertinent to our analysis, including reading scores, math scores, general knowledge scores, total household income, income in thousands, and income group.

Subsequently, we embark on an exploratory data analysis (EDA) to gain insights into the dataset's characteristics. This involves computing summary statistics to understand the distribution and variability of numerical variables. Summary statistics such as mean, standard deviation, minimum, maximum, and quartiles are calculated for each numerical variable in the dataset. Below is the average scores for all tests in both fall and spring:

| | Fall reading | Fall math | Fall generalknowledge | Winter reading | Winter math | Winter generalknowledge |
|---|---|---|---|---|---|---|
| mean | 35.95 | 27.13 | 23.07 | 47.51 | 37.8 | 28.24 |

Table1: Average test scores

We also check for missing values across all columns to ensure data completeness and integrity. This step is crucial for identifying any gaps or inconsistencies in the dataset that may affect our analysis. Since there is no missing value in this dataset, it is very clean, so no further thing to do for dealing missing value.

Additionally, we create pair plots to visualize the relationships between variables. Pair plots offer a graphical representation of pairwise relationships, enabling us to identify patterns, trends, and potential correlations within the dataset.

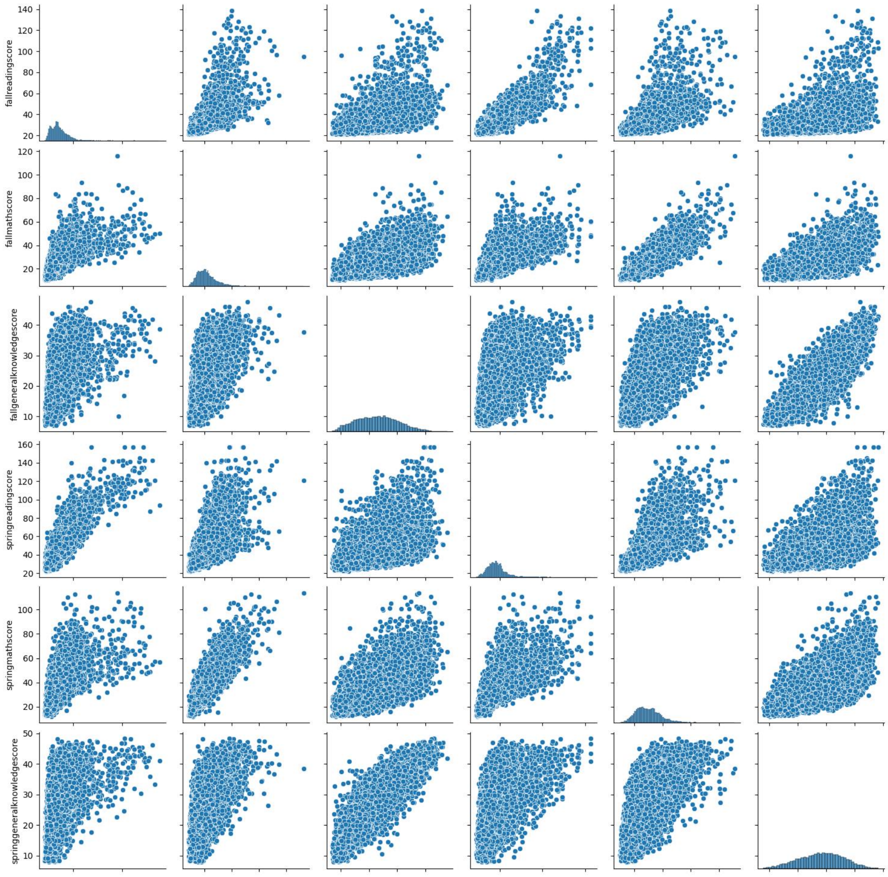

Figure1: Pair plots for 6 test scores

## 3. Quantitative Analysis Using One-way ANCOVAs

3.1 **Research Question #1:** How does household income relate to the academic performance (reading, math, and general knowledge) of kindergarten students over the course of a school year, when controlling for initial academic abilities (fall scores)?

To investigate this question, we conducted ANCOVAs for each academic performance measure (reading, math, and general knowledge scores) with household income as the independent variable and fall scores as covariates.

| | R-squared | Adj. R-squared | F | Prob(F) |
|---|---|---|---|---|
| Reading model | 0.692 | 0.692 | 8929 | < 0.001 |
| Math model | 0.68 | 0.68 | 8469 | < 0.001 |
| General knowledge model | 0.73 | 0.73 | 10820 | < 0.001 |

Table2: ANCOVA result for reading, math and general knowledge scores

For reading scores, the ANCOVA model revealed a significant relationship between household income and spring reading scores (F(2, 11929) = 8929.00, p < 0.001). The model's R-squared value indicates that approximately 69.2% of the variance in spring reading scores can be explained by the variables included in the model. Furthermore, the coefficient for fall reading scores was statistically significant ($p < 0.001$), suggesting that initial academic abilities significantly influence spring reading scores.

The interaction plot for reading scores illustrated the relationship between fall and spring reading scores across different income groups. The plot showed a positive trend, indicating that as fall reading scores increase, spring reading scores also tend to increase, with slight variations across income groups.

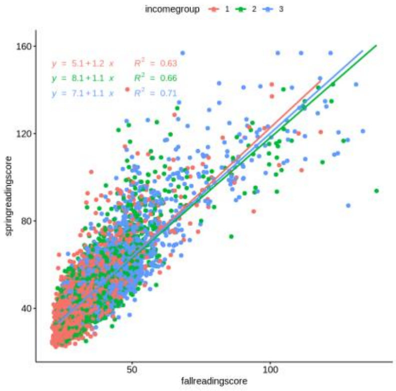

Figure2: Interaction plot for reading scores

For math scores, the ANCOVA model similarly demonstrated a significant association between household income and spring math scores (F(2, 11929) = 8469.00, $p < 0.001$). The model's R-squared value indicated that approximately 68.1% of the variance in spring math scores can be attributed to the variables included in the model. Additionally, the coefficient for fall math scores was statistically significant ($p < 0.001$), highlighting the impact of initial academic abilities on spring math scores.

The interaction plot for math scores exhibited a similar positive relationship between fall and spring math scores across income groups, indicating that higher fall math scores correspond to higher spring math scores, with slight variations based on household income, group2 and 3 are very close but group1 has little difference with them.

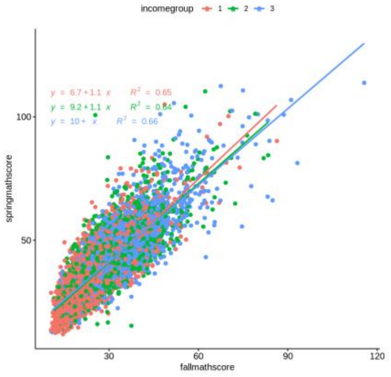

Figure3: Interaction plot for math scores

Finally, for general knowledge scores, the ANCOVA model revealed a significant relationship between household income and spring general knowledge scores (F(2, 11929) = 10820.00, p < 0.001). The model's R-squared value suggested that approximately 73.1% of the variance in spring general knowledge scores can be explained by the variables included in the model. Similar to reading and math scores, the coefficient for fall general knowledge scores was statistically significant (p < 0.001), emphasizing the influence of initial academic abilities on spring general knowledge scores.

In summary, our analysis indicates that household income is significantly associated with academic performance among kindergarten students, even when controlling for initial academic abilities. Higher household income tends to be linked with higher spring academic scores across reading, math, and general knowledge domains. To ensure the validity of our one-way ANCOVA results, we tested two key assumptions: normality of residuals and homogeneity of variances. Normality of Residuals shows the Shapiro-Wilk test for normality of residuals yielded a significant p-value (p < 0.001), indicating that the residuals are not normally distributed. Homogeneity of Variances shows we conducted Levene's test for homogeneity of variances, resulting in a significant p-value (p < 0.001). This indicates that the variances across groups are not equal, violating the assumption of homogeneity of variances. Alternative approaches, such as non-parametric tests or transformations of the data, may be considered to address these violations and ensure the validity of the analysis. Further exploration and refinement of the statistical approach may be necessary to obtain reliable insights

**3.2 Research Question #2:** Is there a significant difference in the academic growth (measured by changes in reading and math and general knowledge scores from fall to spring) among kindergarten students from different income groups, after adjusting for their initial academic abilities (fall scores)?

| | R-squared | Adj. R-squared | F | Prob(F) |
|---|---|---|---|---|
| Reading change model | 0.032 | 0.032 | 133.5 | 4.50e-85 |
| Math change model | 0.016 | 0.016 | 66.62 | 1.02e-42 |
| General knowledge change model | 0.061 | 0.061 | 260.2 | 1.43e-163 |

Table3: ANCOVA results for change of reading, math and general knowledge scores

For changes in reading scores, the ANCOVA model yielded a statistically significant result (F(2, 11929) = 133.5, p < 0.001). The model explained approximately 3.2% of the variance in reading score changes. Additionally, the coefficient for fall reading scores was statistically significant ($p < 0.001$), indicating the influence of initial academic abilities on changes in reading scores over the school year.

Similarly, for changes in math scores, the ANCOVA model revealed a significant result ($F(2, 11929) = 66.62$, $p < 0.001$). The model explained approximately 1.6% of the variance in math score changes. The coefficient for fall math scores was statistically significant ($p < 0.001$), indicating the impact of initial academic abilities on changes in math scores over the school year.

Lastly, for changes in general knowledge scores, the ANCOVA model produced a significant result ($F(2, 11929) = 260.2$, $p < 0.001$). The model explained approximately 6.1% of the variance in general knowledge score changes. Similar to reading and math scores, the coefficient for fall general knowledge scores was statistically significant ($p < 0.001$), indicating the influence of initial academic abilities on changes in general knowledge scores over the school year.

The trends of interaction plots for changes in reading and math scores are similar with trends of interactions plots for reading and math scores respectively in questions 1. However, on interaction plot for changes in general knowledge scores, both group2 and group3 have negative changes, only group1 has increasing trend.

Overall, these findings suggest that while there are significant differences in academic growth among kindergarten students from different income groups, even after adjusting for their initial academic abilities, the magnitude of these differences may vary across academic domains. However, just like research question #1, two assumption checks are broke here as well. Caution is warranted due to violations of ANCOVA assumptions, suggesting the need for alternative statistical approaches for more reliable insights.

## 4. Conclusion

In conclusion, our analysis highlights the significant influence of household income on kindergarten students' academic performance and growth. We found that higher household income is consistently associated with higher academic scores in reading, math, and general knowledge, even after accounting for initial abilities. Additionally, while there are significant differences in academic growth among income groups, the extent of these differences varies across subjects. Reading and general knowledge scores show more pronounced variations, emphasizing the need for targeted interventions to support students from lower-income backgrounds. These findings underscore the importance of addressing socioeconomic disparities in education to ensure equitable opportunities for all students, guiding policymakers and educators in implementing effective interventions for student success.