University of Toronto

INF2178

Experimental Design for Data Science

Technical Assignment 3

Instructor

Professor Shion Guha

Submitted by

Lam Hong Kevin Ching

1009243043

2024/03/23

1. **Introduction**

   This analysis leverages a meticulously curated subset of data from the early child longitudinal study conducted in the academic year of 1998-99, focusing on a critical developmental stage in children's education—Kindergarten. The chosen dataset provides a valuable window into the academic progress of Kindergarten students over a significant span of several months, capturing their growth across key educational areas. Specifically, the dataset includes measurements of reading, math, and general knowledge scores at two critical points: the beginning and end of the school year (fall 1998 and spring 1999). These continuous variables are complemented by a categorical variable that classifies students into different income groups, providing insights into how socioeconomic factors might influence early learning outcomes.

   The research will address three fundamental research questions:

   - **Research Question 1:** Does income category affect the improvement in reading scores from Fall 1998 to Spring 1999 when controlling for the initial reading level?
   - **Research Question 2:** Is there a difference in math score improvements across income categories, after adjusting for initial math skills?

   By examining these scores, this analysis aims to uncover patterns and insights into the developmental trajectories of young learners, with a particular focus on the interplay between academic achievement and socioeconomic status. This dataset serves as a crucial tool for educators, policymakers, and researchers striving to foster environments that support all children's learning and development, regardless of their socio-economic background.

2. **Data Cleaning and Data Wrangling**

   The dataset under consideration comprises 11,933 entries across 9 columns. A preliminary examination revealed the absence of missing values, indicating minimal necessity for data cleaning within the analysis's intended scope. Nevertheless, an adjustment was made to the data structure: the 'incomegroup' variable was originally formatted as an integer; this has been altered to a categorical data type to accurately reflect its intrinsic nature as the sole non-continuous variable within the dataset. The following points outline the observations:

   - fallreadingscore: kindergarten students' reading score on fall 1998;
   - fallmathscore: kindergarten students' math score on fall 1998;
   - fallgeneralknowledgesoce: kindergarten students' general knowledge score on fall 1998;
   - springreadingscore: kindergarten students' reading score on spring 1999;
   - springmathscore: kindergarten students' math score on spring 1999;
   - springgeneralknowledgesoce: kindergarten students' general knowledge score on spring 1999;
   - totalhouseholdincome: total household income;
   - incomeinthousands: total household income in thousands;
   - incomegroup: classified by total household income.

3. **Exploratory Data Analysis (EDA)**

   Before we start to answer the research questions, we proceeded with a comprehensive EDA. The summary statistics (Table 1) from the fall period reveal that students' performance in reading, math, and general knowledge displayed variability, with reading scores exhibiting the widest range. Math scores clustered more closely around a lower average, indicating more

consistency among the students, while general knowledge scores were the most consistent of the three, showing the least variation. By spring, there was a noticeable increase in the average scores for all subjects, signaling academic improvement throughout the school year. The rise in reading and math scores came with greater dispersion, suggesting a wider spread in students' abilities. General knowledge scores also improved, maintaining a relatively consistent variation similar to the fall period. These trends indicate not only overall growth but also an expansion in the range of students' scores as the year progressed. Meanwhile, the minimum and maximum scores also reveal a broad range of outcomes, with the maximum scores for both reading and math showing a significant jump from fall to spring. This range suggests that while some students might have excelled exceptionally, others remained with scores closer to the minimum observed. The percentiles further delineate this growth and distribution, with the 50th percentile (median) scores for reading, math, and general knowledge also exhibiting an upward trend from fall to spring.

	FALLREADING	FALLMATH	FALLGENERALKNOWLEDGE	SPRINGREADING	SPRINGMATH	SPRINGGENERALKNOWLEDGE
MEAN	35.95	27.12	23.07	47.51	37.79	28.23
STD	10.47	9.12	7.39	14.32	12.02	7.57
MIN	21.01	10.51	6.98	22.35	11.90	7.85
25%	29.34	20.68	17.38	38.95	29.27	22.80
50%	34.06	25.68	22.95	45.32	36.41	28.58
75%	39.89	31.59	28.30	51.77	44.22	33.78
MAX	138.51	115.65	47.69	156.85	113.8	48.34

Table 1: Dataset Quantitative Data Statistics

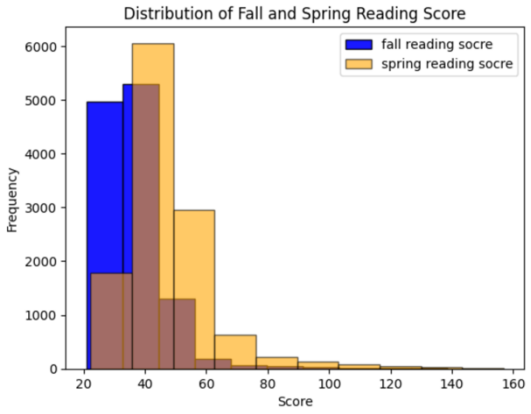

Figure 1: Distribution of Fall and Spring Reading Score

In the reading scores histogram (Figure 1), the distribution for Spring 1999 is shifted to the right compared to Fall 1998, suggesting a general improvement in reading scores. This shift can provide insights relevant to Research Question 1. If the income category is overlaid on this distribution, one could analyze whether the shift in the mean score differs by income category, especially after controlling for the initial reading level. A larger rightward shift in higher income categories could indicate a more significant improvement in reading scores related to socioeconomic factors.

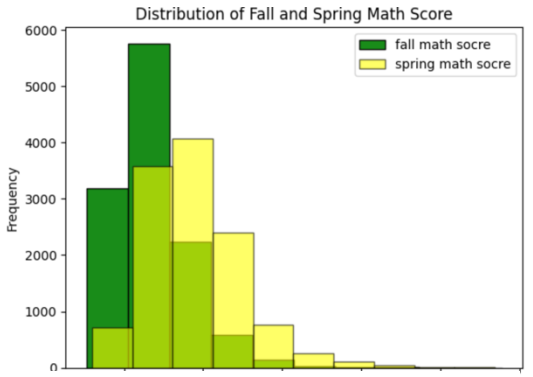

Figure 2: Distribution of Fall and Spring Math Score

The math scores histogram (Figure 2) also shows a rightward shift from Fall to Spring, indicative of overall improvement. When considering Research Question 2, this visual cue sets the stage for ANCOVA analysis to determine if the extent of improvement varies across different income categories, after factoring in the initial math skills. If certain income groups demonstrate a disproportionately larger shift, this may imply that income impacts the rate of improvement in math skills.

4. One-way ANCOVA Test

Research Question 1: Does income category affect the improvement in reading scores from Fall 1998 to Spring 1999 when controlling for the initial reading level?

The research question investigates whether the income category of students has an impact on the improvement of their reading scores from Fall 1998 to Spring 1999, while controlling for their initial reading levels. The interest lies in understanding the extent to which socioeconomic status, as reflected by the income category, influences the rate of progress in reading skills over the course of the academic year. The question aims to discern if there is a differential effect of income on educational advancement in reading when the starting point, or baseline reading ability, is taken into consideration.

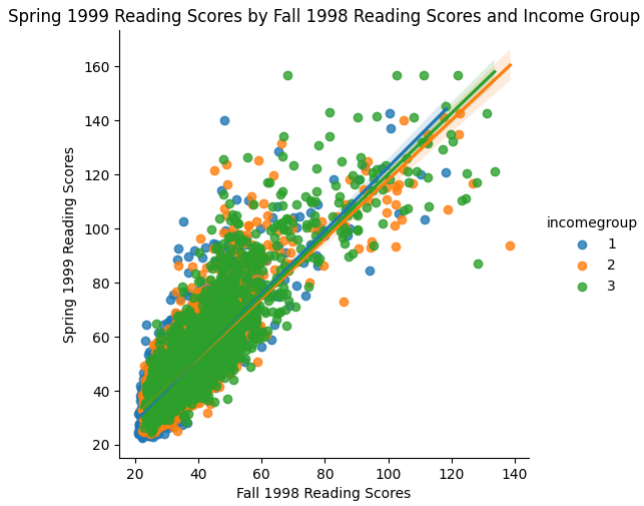

Figure 3: Spring 1999 Reading Scores by Fall 1998 Reading Scores and Income Group

From the scatterplot (Figure 4), it's observable that as Fall 1998 reading scores increase, Spring 1999 reading scores also tend to increase, as shown by the positive slope of the lines representing each income group. The different slopes for each income group suggest that the rate of improvement in reading scores may differ by income category. Specifically, it appears that the highest income group (3) has the steepest slope, followed by the middle income group (2), and then the lowest income group (1). This suggests that students from higher-income families may have experienced a greater improvement in their reading scores over the school year when compared to their lower-income peers.

The ANCOVA table (Table 2) complements these findings. With a significant F-statistic and a p-value less than 0.01, the model is statistically significant, indicating that there are indeed differences in reading score improvements among the income groups when controlling for initial

reading levels. The coefficients for income groups 2 and 3, when compared to the reference group (income group 1), are positive and significant ($p < 0.05$), suggesting that these groups have a higher expected increase in reading scores than the lowest income group. The coefficient for the fall reading score is also positive and highly significant ($p < 0.01$), confirming the importance of the initial reading level in predicting Spring reading scores.

F-statistic	**8929**
Prob(F-statistic)	0.00

	Coef.	**P>|t|**
Intercept	6.5430	0.000
C(incomegroup) [T.2]	0.3751	0.033
C(incomegroup) [T.3]	0.4898	0.008
fallreadingscore	1.1322	0.000

Table 2: ANCOVA Test Results for Research Question 1

These results suggest that income category does have an effect on the improvement of reading scores from fall to spring when initial reading ability is accounted for. Students from higher income categories tend to show greater improvements, highlighting a potential socioeconomic impact on educational progress in reading during the kindergarten year.

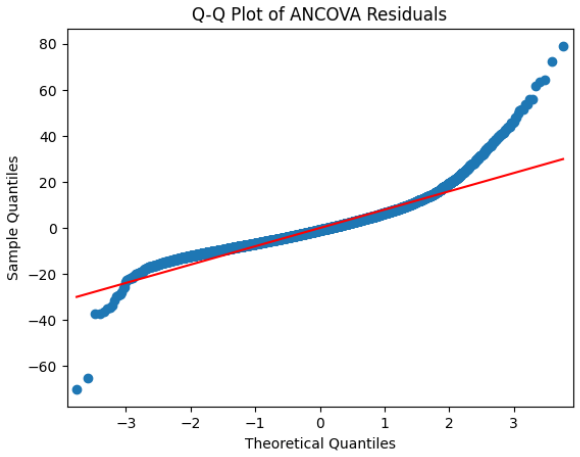

Figure 4: Q-Q Plot of ANVCOVA Residuals for Research Question 1

The assumption checks for ANCOVA applied to Research Question 1 reveal challenges with the data that may impact the validity of the statistical results. The Q-Q plot (Figure 5) of ANCOVA residuals shows divergence from the expected normal distribution, particularly at the tails, indicating potential non-normality in the residual distribution. This is partly confirmed by the Shapiro Test, which returned a statistic of 0.9119 and a p-value smaller than 0.01, suggesting deviations from normality. Compounding this issue is the Levene Test result, with a statistic of 39.5528 and a p-value less than 0.01, clearly indicating that the assumption of homogeneity of variances is not met across the income groups. These statistical tests suggest significant departures from the assumptions underlying ANCOVA, implying that the conclusions drawn from this model about the effect of income on reading score improvements might be less reliable and would benefit from further scrutiny or methodological adjustments.

	STATISTIC	**P-VALUE**
SHAPIRO TEST	0.9119	0.00
LEVENE TEST	39.5528	0.00

Table 3: Assumptions Check Table for Research Question 1

Research Question 2: Is there a difference in math score improvements across income categories, after adjusting for initial math skills?

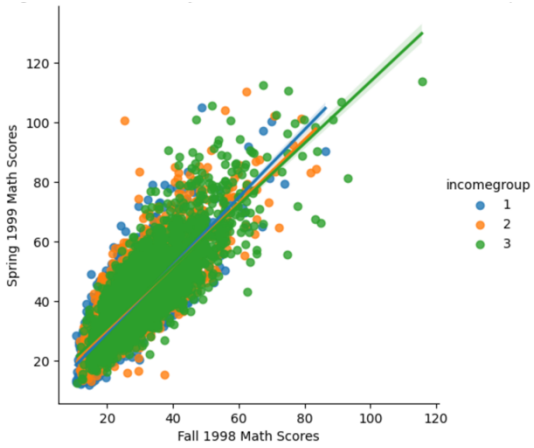

Figure 5:Spring 1999 Math Scores by Fall 1998 Math Scores and Income Group

This research question seeks to explore the influence of income category on the improvement of math scores from Fall 1998 to Spring 1999, with a particular focus on the change relative to initial math abilities. The scatterplot (Figure 6) displays a positive correlation between the scores from Fall to Spring across all income groups. The steeper slopes for higher income groups suggest that students from these groups may have experienced greater improvements in their math scores over the academic year.

The ANCOVA analysis result (Table 4) indicate that the model is statistically significant, with an F-statistic approaching 8469 and a p-value less than 0.01. The coefficients for income groups 2 and 3 are positive and statistically significant ($p < 0.01$) when compared to the reference group (income group 1), implying that these groups have a higher expected increase in math scores, after controlling for initial math abilities. Specifically, the coefficient for income group 2 is 0.6700 and for income group 3 is 0.9199, which can be interpreted as the expected additional increase in Spring 1999 math scores associated with being in these income groups, beyond what would be predicted by Fall 1998 math scores alone. The coefficient for the fall math scores is 1.0735, with a p-value smaller than 0.01, underscoring that initial math scores are a strong predictor of spring scores.

F-statistic	**8469**
Prob(F-statistic)	0.00

	Coef.	**P>|t|**
Intercept	8.2011	0.000
C(incomegroup) [T.2]	0.6700	0.000
C(incomegroup) [T.3]	0.9199	0.000
fallreadingscore	1.0735	0.000

Table 4: ANCOVA Test Results for Research Question 2

The results indicate that students from higher-income categories show greater improvements in math scores from Fall to Spring compared to their peers from the lowest income category, after adjusting for initial math skills. This highlights the possible impact of socioeconomic factors on the progression in math achievement during the kindergarten year.

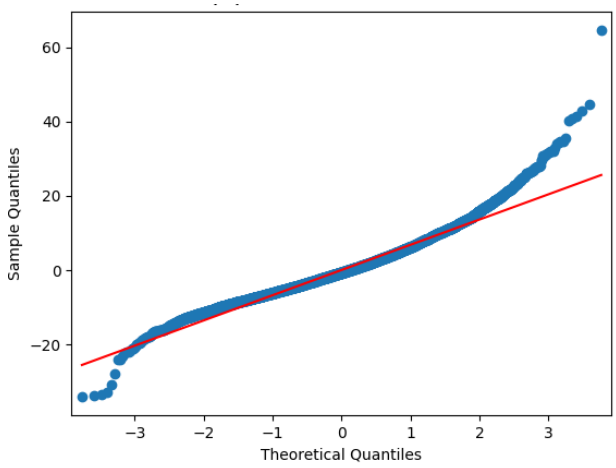

Figure 6: Q-Q Plot of ANVCOVA Residuals for Research Question 1

The assumption checks for the ANCOVA applied to Research Question 2 indicate significant departures from the model prerequisites. The Q-Q Plot (Figure 7) demonstrates notable deviations from the expected line, especially at the tails, suggesting that the residuals do not conform to a normal distribution. This is confirmed by the Shapiro-Wilk test (Table 5), which presents a statistic of 0.9119, and a decisive p-value, which smaller than 0.01, clearly indicating non-normality. Furthermore, the Levene Test for homogeneity of variances returns a statistic of 39.5528 with a p-value less than 0.01, rejecting the assumption of equal variances across groups. These results—showing violation of both the normality of residuals and homogeneity of variance—call into question the reliability of the ANCOVA findings for assessing the influence of income on the improvement of math scores when adjusted for initial math ability.

	STATISTIC	P-VALUE
SHAPIRO TEST	0.9119	0.00
LEVENE TEST	39.5528	0.00

Table 5: Assumptions Check Table for Research Question 2

5. Conclusion

In conclusion, this analysis presents compelling evidence that socioeconomic factors, as categorized by household income, play a significant role in the academic progress of Kindergarten students, particularly in reading and mathematics. The study's results demonstrate that students from higher-income groups show statistically significant greater improvements in both reading and math scores over the academic year, controlling for their initial abilities. This finding aligns with the theory that socioeconomic status can influence educational outcomes, even at the outset of formal schooling. While the ANCOVA tests confirm the effect of income on academic improvement, assumptions checks for both Research Questions 1 and 2 highlight potential violations of normality and homogeneity of variances. These statistical considerations suggest caution in the interpretation of the results and may prompt further investigation using alternative methodologies. Nonetheless, the analysis underscores the critical interplay between income and educational advancement, emphasizing the need for policies that address these disparities from early childhood education to support equitable learning opportunities for all children.