INF 2178 Technical Assignment 3
Xiaoxin Zhou 1003549742

Introduction:

The early years of childhood education are critical for cognitive development and academic achievement. Understanding the trajectories of learning across different domains such as reading, mathematics, and general knowledge is essential for developing educational interventions that support children's growth. This analysis draws on a subset of data from an early childhood longitudinal study, which tracks the academic performance of kindergarten students over several months, specifically from fall 1998 to spring 1999. The dataset includes scores in reading, math, and general knowledge, offering a unique opportunity to examine the factors influencing academic success during this pivotal stage.

Given the complex nature of educational achievement, which is influenced by a multitude of factors including socioeconomic status, educational environment, and individual differences, it is crucial to employ robust statistical methods to untangle these effects. One-way Analysis of Covariance provides a powerful tool for this purpose. By incorporating covariates into the analysis, ANCOVA allows for the adjustment of pre-existing differences among participants, thereby isolating the effect of the main variable of interest. This approach is particularly useful in educational research, where baseline differences between students can confound the results. This study aims to apply one-way ANCOVA to investigate the impact of various covariates on students' reading, math, and general knowledge scores. By doing so, we seek to identify significant predictors of academic performance and provide insights that can inform targeted educational strategies. The analysis will not only contribute to the existing literature on early childhood education but also offer practical implications for educators, policymakers, and researchers striving to enhance learning outcomes for young children.

Dataset for Child education from Fall 1998 - Spring 1999

This dataset has 11933 rows and 9 columns, and there are no null values for any feature in the data.

	fallreadingscore	fallmathscore	fallgeneralknowledgescore	springreadingscore	springmathscore	springgeneralknowledgescore	totalhouseholdincome	incomeinthousands	incomegroup
count	11933.00	11933.00	11933.00	11933.00	11933.00	11933.00	11933.00	11933.00	11933.00
mean	35.95	27.12	23.07	47.51	37.79	28.23	54317.19	54.31	1.89
std	10.47	9.12	7.39	14.32	12.02	7.57	36639.06	36.63	0.82
min	21.01	10.51	6.98	22.35	11.90	7.85	1.00	0.00	1.00
25%	29.34	20.68	17.38	38.95	29.27	22.80	27000.00	27.00	1.00
50%	34.06	25.68	22.95	45.32	36.41	28.58	47000.00	47.00	2.00

75%	39.89	31.59	28.30	51.77	44.22	33.78	72000.00	72.00	3.00
max	138.51	115.65	47.69	156.85	113.80	48.34	150000.00	150.00	3.00

Table 1 presents a summary of the dataset.

Based on Table 1, we can compare the fall reading scores with the spring reading scores, fall math scores with the spring math scores, and fall general knowledge scores with the spring general knowledge scores. All scores have improved from fall to spring, with the reading scores showing the most significant improvement. This leads us to consider what factors might influence these improvements. However, with the limited data available, we are unable to determine the reasons behind this trend. The data we do have includes family income, where 25% of families have an income around $27,000, 50% of families have an income around $47,000, and 75% of families have an income around $72,000. Based on this income range, we could formulate a research question: **Does family income affect a child's learning ability and their academic improvement from fall to spring?**

1. fallreadingscore: Reading scores of students in fall
2. fallmathscore: Math scores of students in fall
3. fallgeneralknowledgescore: General knowledge scores of students in fall
4. springreadingscore: Reading scores of students in spring.
5. springmathscore: Math scores of students in spring
6. springgeneralknowledgescore: General knowledge scores of students in spring
7. totalhouseholdincome: The total household income of the student's family.
8. incomeinthousands: The total household income in thousands
9. incomegroup: A categorical variable representing the income group of the student's family, with values ranging from 1 to 3.

The features totalhouseholdincome and incomeinthousands are overlapping, we will choose incomeinthousands for our analysis.

incomegroup	fallreading score	fallmath score	fallgeneral knowledge score	springrea dingscore	springma thscore	springgeneralkn owledgescore	totalhouseholdincome
1	32.78	23.92	19.94	43.66	33.88	25.06	22019.71
2	36.29	27.56	23.88	48.00	38.46	29.14	51742.75
3	39.89	31.01	26.45	52.20	2.41	31.56	100989.75

In Table 2, by grouping the data by 'incomegroup' and calculating the mean for each group, we found that higher income groups also have higher mean scores.

Visualization

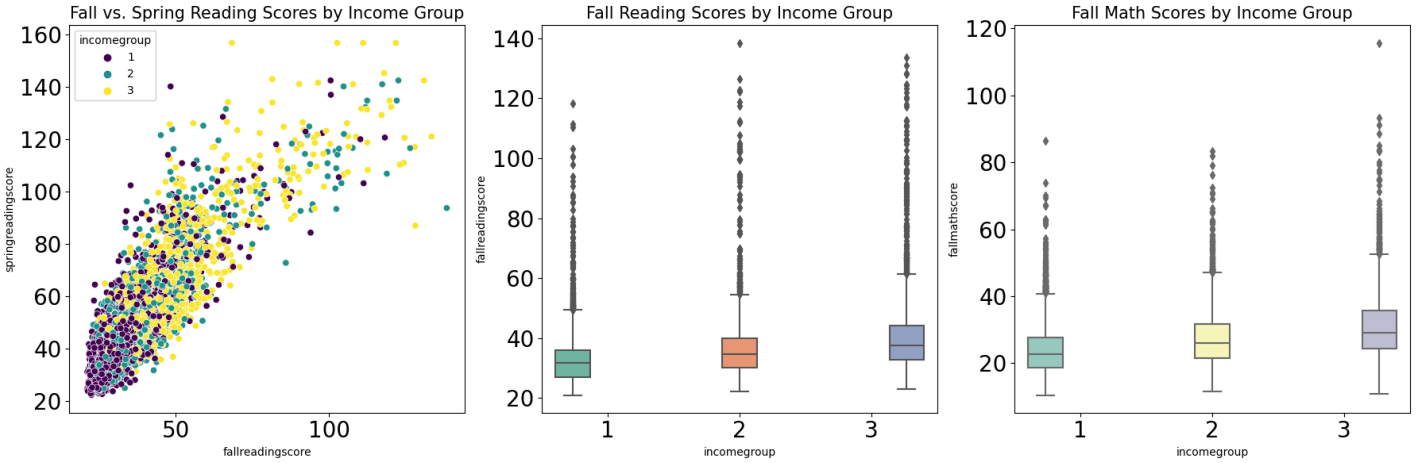

Figure 1 presents a comprehensive visualization of educational scores across different income groups. On the left, a scatter plot illustrates the relationship between fall reading scores and spring reading scores, with each point coloured according to its income group—lighter colours for lower income groups and darker for higher income groups. This visualization highlights any potential correlations between reading scores across the two time periods within each income group.

The center and right plots are box plots that display the distribution of fall reading scores and fall math scores, respectively, across the three income groups. These box plots reveal the central tendency and dispersion of scores within each income group, with outliers indicated as individual points. The colour-coding by income group in the box plots mirrors that of the scatter plot, providing a consistent visual cue across all three visualizations. The box plots show a clear gradation in median scores from the lower to higher income groups, suggesting that income may be a factor in educational performance. The variability within each income group is also visible, indicating the range of scores that exist within each category.

I've divided the income into seven brackets: '0-25k', '25k-50k', '50k-75k', '75k-100k', '100k-125k', '125k-150k', and '150k-175k'. Based on these income brackets, we will group each child's scores into the corresponding bracket.

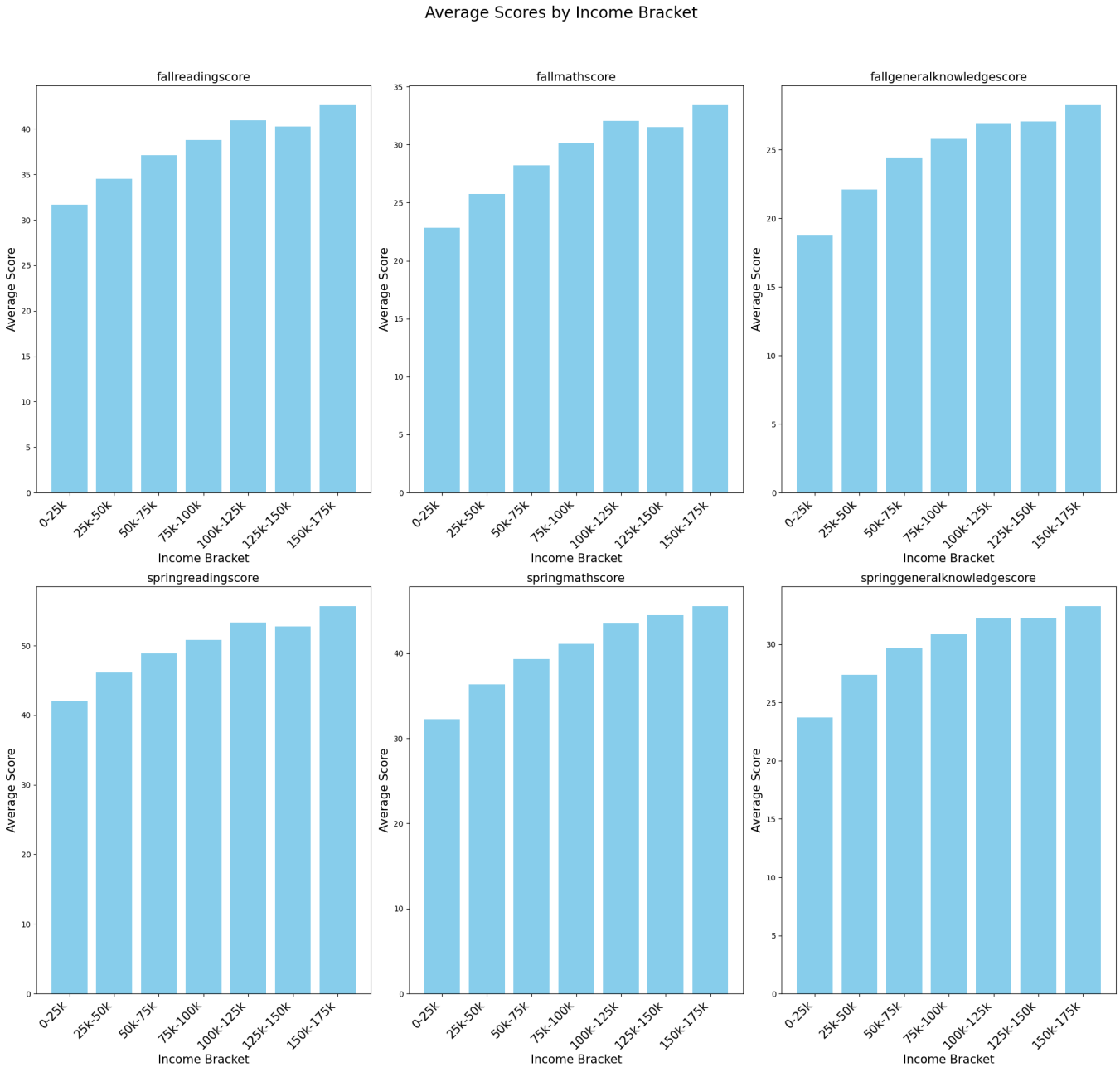

Figure 1, In the figure, we can visually observe the different income brackets and their mean scores for each category. Although we have divided the data into seven income brackets, the gap between different brackets is still easily obvious. There must be a linear relationship between the income group and learning ability.

ANCOVA

I will consider the average fall and spring scores as measures of initial learning ability and academic improvement, respectively, along with the incomeinthousands as our independent variable. The incomegroup can be useful for categorizing incomes, but we'll focus on the actual income values for a more precise analysis.

Data Exploration and Assumption Checks
Lets calculate the average fall and spring scores to use as the dependent variable and covariate, respectively, and explore the relationship between these scores and family income to ensure the data meets ANCOVA assumptions.

In this section, we will preprocess the data before conducting the ANCOVA analysis. We aim to compare the fall average scores and spring average scores to create a new feature, 'academic_improvement,' which will help us understand how students have improved academically from fall to spring.

income_group	avg_fall_score	academic_improvement
0-25k	24.562636	8.326800
25k-50k	27.956204	9.175416
50k-75k	30.153048	9.336090
75k-100k	32.008576	9.515880
100k-125k	33.535830	9.666218
125k-150k	34.468237	10.171485

Table 3. The first five rows of data, representing income, fall score average, and academic improvement.

Before proceed to perform ANCOVA, let's check for assumptions:
Linearity: The relationship between the covariate (initial learning ability) and the dependent variable (academic improvement) is linear from Table 3.

Homogeneity of regression slopes: The relationship between initial learning ability and the academic improvement variable is consistent across groups formed by the independent variable, income groups. Table 3 calculates the mean for each income group.

Homogeneity of variances: The variances of the dependent variable should be equal across groups.However, the result of Levene's test, with a statistic of approximately 4.545 and a p-value of roughly 0.000384, indicates a significant violation of the homogeneity of variances assumption. In the context of ANCOVA analysis, this means the variance in academic improvement across different income groups is not consistent.
Levene's test statistic: 4.544679042165699, p-value: 0.00038439619071720376

Luckily, I realize that the ANCOVA required continuous data instead of income group.
The value of 1.052 is a measure of how much the group variances differ from the overall variance. A p-value of 0.295 is above the common alpha level of 0.05, suggesting that the observed differences in variances among the income groups are not statistically significant. The

result supports the assumption required for ANCOVA that the variances in the dependent variable (academic improvement) are similar across the groups defined by the independent variable (income groups). This validates the use of ANCOVA for our analysis, as one of the key assumptions has been met! !
Levene's test statistic: 1.0523787120020935, p-value: 0.294886085199163

The ANCOVA results

	sum_sq	df	F	PR(>F)
incomeinthousands	131.543695	1.0	6.940442	8.437712e-03
avg_fall_score	8298.837035	1.0	437.859060	1.627401e-95
Residual	226111.858389	11930.0	NaN	NaN

Table 4, The ANCOVA results offer statistical insight into the impact of family income (as a continuous variable very important for ANCOVA) and initial learning ability on the dependent variable, presumably academic improvement.

Sum of Squares: 131.543695 indicates the variation in academic improvement attributable to differences in family income.
Income:
Df: 1. Income is considered as a single continuous variable.
F-value: 6.940442, which measures the ratio of the variance explained by income to the unexplained variance (residual). An F-value greater than 1 suggests that income explains a significant portion of the variance in academic improvement.
P-Value: 8.437712e-03, indicating the probability of observing such an F-statistic if income had no effect on academic improvement. Since this p-value is less than 0.05, we conclude that family income has a statistically significant effect on academic improvement.
Average score:
Sum of Squares: 8298.837035 indicates the variation in academic improvement due to initial learning ability.
Df: 1, as this variable is also considered individually.
F-value: 437.859060, a very high value indicating that initial learning ability strongly predicts academic improvement.
P-Value: 1.627401e-95, virtually zero, showing that the effect of initial learning ability on academic improvement is extremely statistically significant.

Significant Impact of Income: The results suggest that differences in family income significantly affect academic improvement, controlling for initial learning ability. This implies that, on average, income level is a predictor of how much a student's academic performance might improve.

Stronger Impact of Initial Learning Ability: The very high F-statistic and practically zero p-value for the initial learning ability indicate that this factor has a much stronger influence on academic

improvement than income. It suggests that students' baseline academic abilities play a crucial role in their subsequent academic progress.

Residual Variance: A large portion of the variance in academic improvement remains unexplained by these two factors alone, suggesting that other variables not included in the model may also influence academic outcomes.

OLS Regression Results (more detail for ANCOVA)

OLS Regression Results

Dep. Variable:	academic_improvement	**R-squared:**	0.046
Model:	OLS	**Adj. R-squared:**	0.046
Method:	Least Squares	**F-statistic:**	286.4
Date:	Wed, 20 Mar 2024	**Prob (F-statistic):**	3.30e-122
Time:	21:47:30	**Log-Likelihood:**	-34484.
No. Observations:	11933	**AIC:**	6.897e+04
Df Residuals:	11930	**BIC:**	6.900e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.6081	0.153	36.628	0.000	5.308	5.908
incomeinthousands	0.0031	0.001	2.634	0.008	0.001	0.005
avg_fall_score	0.1168	0.006	20.925	0.000	0.106	0.128

Omnibus:	1465.101	**Durbin-Watson:**	1.676
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3946.115
Skew:	0.684	**Prob(JB):**	0.00
Kurtosis:	5.463	**Cond. No.**	271.

Figure 2, The OLS regression results provided summarize the findings of a statistical analysis investigating the relationship between academic improvement (the dependent variable) and two predictor variables: income in thousands and average fall score.

R-squared: With a value of 0.046, this indicates that approximately 4.6% of the variance in academic improvement is explained by the model. While this is a positive number, it's relatively low, suggesting that there are other factors not included in the model that also affect academic improvement.
F-statistic: A value of 286.4 is quite high and, along with the very low Prob (F-statistic), suggests that the model is statistically significant and the explanatory variables collectively have a significant effect on the dependent variable.
Coefficients:
The Intercept is 5.6081, meaning if income and the average fall score were zero, the expected academic improvement would be 5.6081 points. However, this is a theoretical interpretation because an income of zero is not practical in this context.
incomeinthousands has a coefficient of 0.0031, suggesting a small positive relationship between income and academic improvement. For every thousand dollars increase in income, academic improvement increases by 0.0031 points, on average.
avg_fall_score has a coefficient of 0.1168, which means for every point increase in the average fall score, there is an average increase of 0.1168 points in academic improvement.
It just the same conclusion as the previous result shows that both family income and initial learning ability (as measured by the average fall score) have statistically significant effects on academic improvement. Initial learning ability appears to have a stronger effect than family income based on the F-statistic and corresponding p-values.

Conclusion:
This analysis underscores the importance of considering both socioeconomic factors and individual academic abilities when assessing educational outcomes. If one day your parents blame you for getting a B instead of an A+, show them this work as a plus but don’t explain the residual parts. :P