Early Child Longitudinal Study Analysis

INF2178: Experimental Design for Data Science

Instructor: Professor Shion Guha

Yuanyuan Pan 1003980150

Assignment Introduction

This assignment aims to conduct a statistical analysis on the academic performance of kindergarten students over a span of several months by examining the data collected from an early child longitudinal study from 1998 to 1999. Motivated by the discrepancy found on the academic performance of early childhood education, the primary goal of this study is to investigate the factor, particularly in the context of income disparities, that may influence 1999 spring academic scores while controlling for the 1998 fall academic performance. The study employs statistical techniques, including the exploratory data analysis (EDA) and analysis of covariance (ANCOVA), to gain insights into how income groups might affect child's spring academic scores in 1999 after accounting for 1998 fall scores as covariates.

The study addresses two fundamental research questions:

1. How does income group influence spring math scores after controlling for fall math scores?
2. How does income group influence spring general knowledge scores after controlling for fall general knowledge scores?

Data Pre-processing and Cleaning

In the data pre-processing step, the data named "INF2178_A3_data.csv" has been read into the data frame for analysis. This original dataset contains nine variables, including reading, math, and general knowledge scores for both fall and spring term, total household income, income in thousands, and income group which categorizes household income into three groups from low to high (1, 2, 3). However, only five variables have been kept for use in the later analysis, which are math and general knowledge scores for both fall and spring terms and the income group.

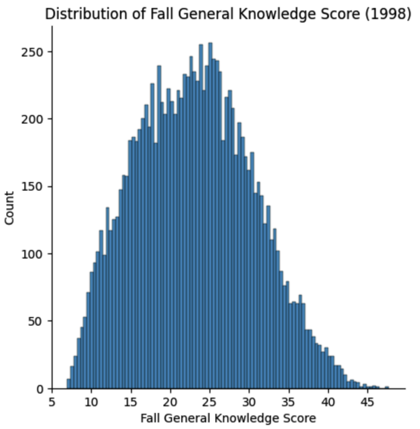
Exploratory Data Analysis

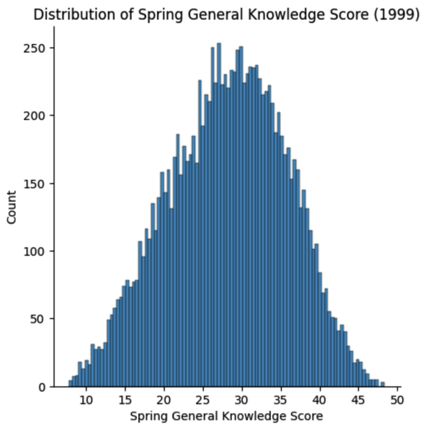
Summary statistics table:

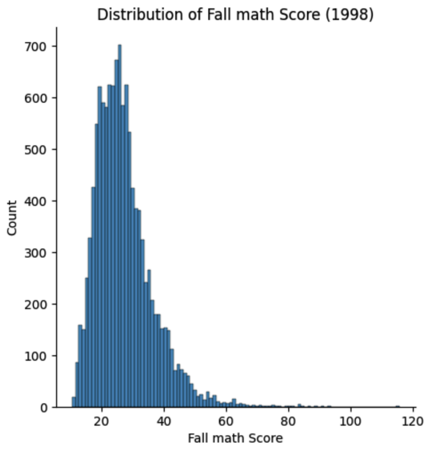
	Fall math	Fall general knowledge	Spring math	Spring general knowledge
Min	10.51	6.98	11.9	7.86
Mean	27.13	23.07	37.8	28.24
Max	115.65	47.69	113.8	48.34
Median	25.68	22.95	36.41	28.58
IQR	10.91	10.91	14.95	10.98
sd	9.12	7.4	12.03	7.58

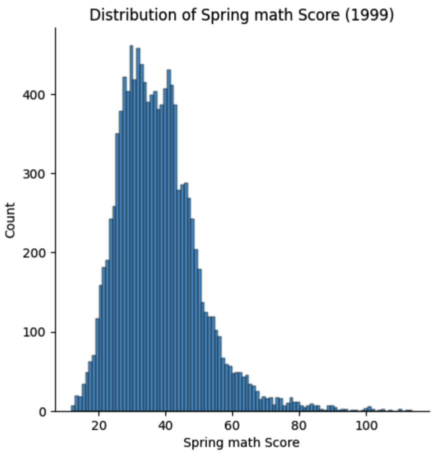
Above table records the summary statistics of math and general knowledge scores of kindergarten students in fall and spring terms. Regardless of the income group, in general, there is a clear improvement in scores on both math and general knowledge from fall to spring. By examining the above statistics, the mean math score increased around 10 units while the mean general knowledge score increased around 5 units ; the median has increased about a similar amount compared to the mean. The interquartile range (IQR) are the same for both math and general knowledge scores in fall, however, IQR of math score increased about 4 units while general knowledge score almost remained the same in the spring. This shows that the spread in spring math score has widened, indicating a larger dispersion of data points around the median, which could also be detected by examining the above standard deviation (sd). While the sd of general knowledge almost remains the same from fall to winter, math score shows a clear increase, which indicates an increase in variation of the data.

Histogram:


Distribution of Fall General Knowledge Score (1998)


Distribution of Spring General Knowledge Score (1999)


Distribution of Fall math Score (1998)

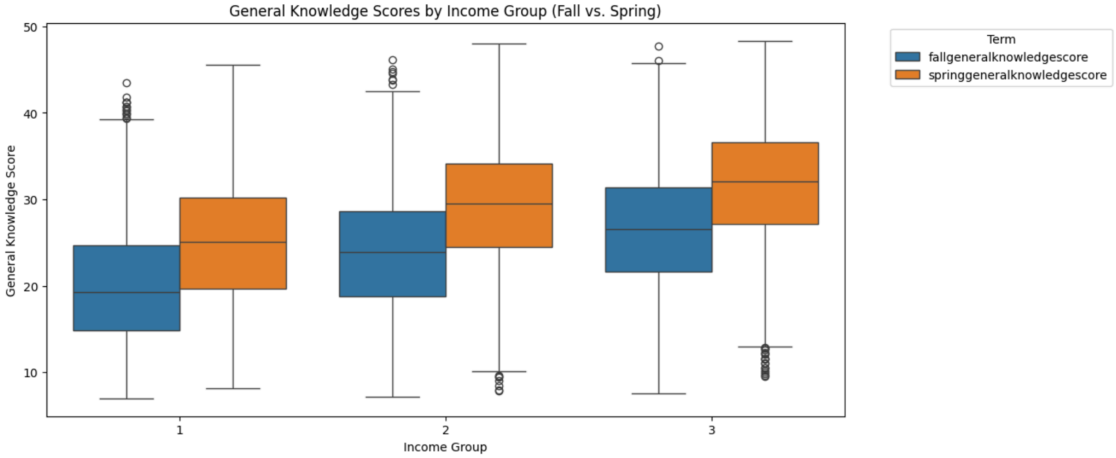

Distribution of Spring math Score (1999)

By examining the above upper two histograms of general knowledge in both fall and spring terms, there is a shift of the middle 50% of the data to the right, indicating an increase in score from fall to winter regardless of the outliers. Both histograms are unimodal and the shapes are almost symmetric, closed to a normal distribution. In the case of measure of central tendency, both mean and median could be used as a measurement since they are almost normal distribution.

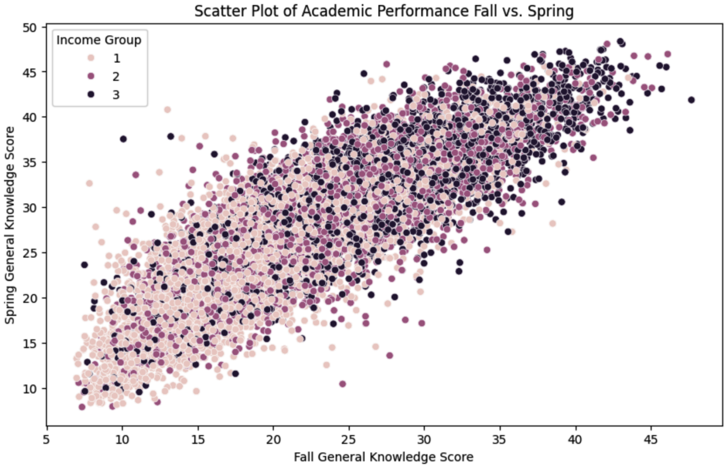
By examining the lower two histograms of math score in both fall and spring terms above, two clear unimodal histograms have been shown with both positively skewed (right skewed). Spring math scores have a wider spread compared to fall term and it also has higher scores in the middle 50% of the data. Median is more suitable when measuring the central tendency since it is more robust to outliers.

Box plot:

Below is a combined box plot comparing fall and spring general knowledge scores based on different income groups. Income group 1, 2, 3 indicating an increase in household income, which could be seen as low, median, and high income. There is a clear increase in middle 50% data of general knowledge scores from 1 to 3 income groups regardless of the academic term, however, spring term score shows a high increase in middle 50% data when comparing to fall term in all income groups, shifted almost 5 units higher in scores. All outliers are detected in higher scores in fall term; the lower the income, the more the outliers. However, outliers in spring term are all detected in lower scores, and the higher the income, the more the outliers. This shows that from fall to the spring, students' academic performance has a high improvement in all income group, yet higher income group tends to have higher scores.



Scatter plot:



The scatter plot on the right shows the relationship between fall and spring general knowledge scores, marked by different income groups. There is a clear positive linear relationship detected in the plot, indicating a high strength of correlation between fall and spring general knowledge scores. Other than that, most of the light pink data points gathered before score 20 while most of the black data points gathered above 30. This shows that under the whole trend of increasing in the score of general knowledge from fall to winter, low income group students tend to have lower scores compared to high income group students.

One-way ANCOVA

According to what has been specified in research questions, the following one-way ANCOVAs will focus on examining whether there are significant differences in the means of the spring math/general knowledge scores across different levels of income group, while also accounting for the effects of the covariate, fall math/general knowledge score. In the following ANCOVAs, the independent variable is the income group, the dependent variable is spring math/general knowledge scores, and the covariate is fall math/general knowledge scores.

1. Math Score ANCOVA:

Null Hypothesis (H0): There is no significant difference in the mean spring math scores across different income groups after controlling the effects of fall math scores.
Alternative Hypothesis (H1): There is a significant difference in the mean spring math scores across different income groups after controlling the effects of fall math scores.

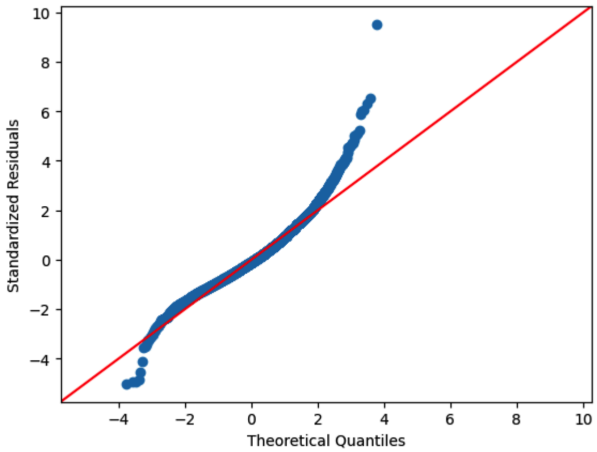
OLS Regression Results:

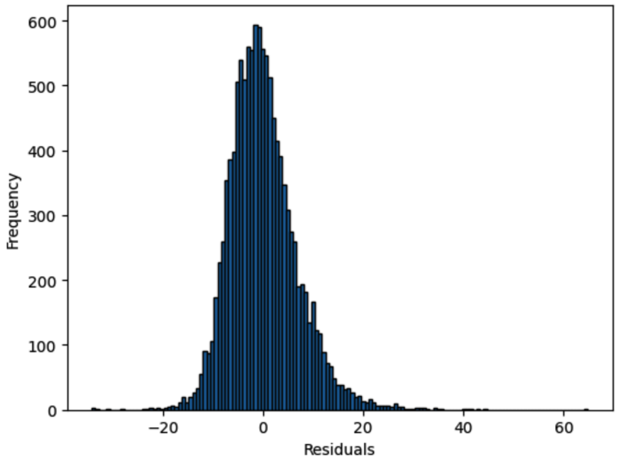
	Coefficient	P>|t|	Standard error		R-squared	F-stats	PR(>F)
Intercept	7.79	<0.001	0.22		\	\	\
Income group	0.47	<0.001	0.08		\	\	\
Fall math score	1.07	<0.001	0.01		\	\	\
Model fit	\	\			0.68	1.27e+04	<0.001

Above table displaces the Ordinary Least Squares (OLS) regression model results. The model coefficient represents the estimated effects of each variable on the dependent variable, spring meth score. The intercept coefficient predicts the spring math score when all independent variables are 0. The coefficient 0.47 of income group indicates the estimated change in the spring math score for a one-unit increase in income group when holding fall math score constant. The coefficient 1.07 represents the estimated change in the spring math score for a one-unit increase in fall math score while holding the income group constant. Since the p-value for all three coefficients are less than 0.001, there is strong evidence suggesting that they are all statistically significant predictors of the spring math score. The standard error for all the coefficients are low, suggesting a relatively precise estimation.
By examining the model fit, R-squared indicates that about 68.0% of the variability in the spring math scores can be explained by the independent variables in the model, which could be considered as a relatively good value. The p-value associated with the F-statistics (PR(>F)) is less than 0.001, indicating the overall model is statistically significant, hence there is strong evidence to reject the null hypothesis.

Checking model diagnostics:

After conducting the ANCOVA above and fitted the OLS regression model, assumptions need to be checked since it could help to ensure the validity and interpretability of the results.





Normality Assumption:

The plot on the left above is a **Quantile-Quantile (Q-Q) Plot** which provides a visual assessment of the normality assumption. By examining the plot, while the central region almost aligns with the theoretical quantiles, the sudden curve on the two sides, especially the upper part, suggest that the residuals may not perfectly follow normal distributions. The histogram on the right above shows the **Residual Distribution**; a perfect symmetric distribution suggests that the residuals are unbiased and random. However, it almost follows a normal distribution, but still not a perfect symmetric shape. Overall, the normality assumption could be seen as somewhat violated, which may be due to non-constant variance or outliers.

Homogeneity Assumption:

Levene's test result: Test statistics (W) = 18.90; **p-value** < 0.001

According to the histogram shown in EDA, the sample distribution for spring math score is not normally distributed. Hence, to check the homogeneity assumption of the model, a Levene's test has been conducted. Since the p-value is less than 0.001, it indicates that there is strong evidence to reject the null hypothesis of equal variances.

2. General Knowledge Score ANCOVA:

Null Hypothesis (H0): There is no significant difference in the mean spring general knowledge scores across different income groups after controlling the effects of fall general knowledge scores.

Alternative Hypothesis (H1): There is a significant difference in the mean spring general knowledge scores across different income groups after controlling the effects of fall general knowledge scores.

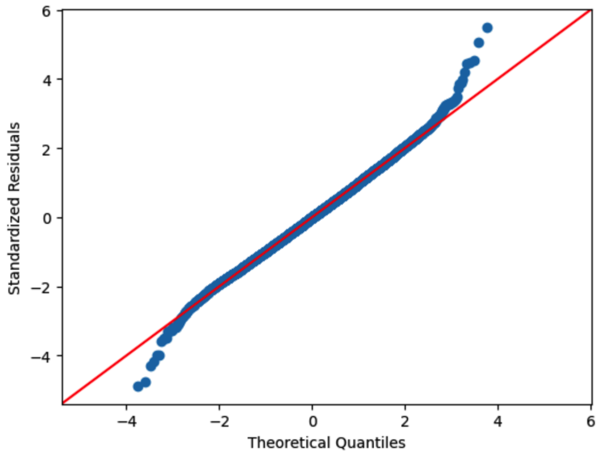
OLS Regression Results:

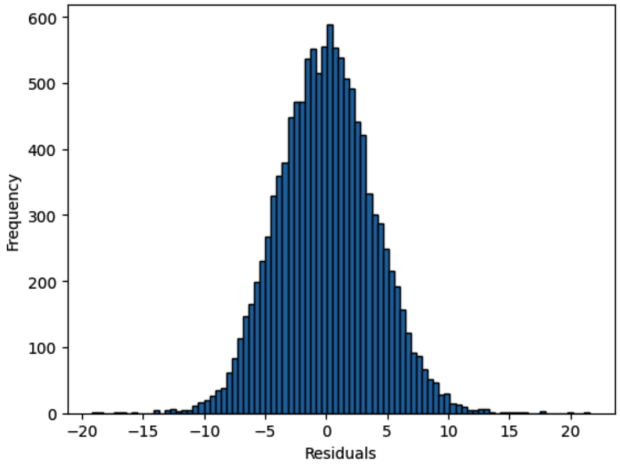
	Coefficient	**P>|t|**	**Standard error**	**R-squared**	**F-stats**	**PR(>F)**
Intercept	7.60	<0.001	0.13	\	\	\
Income group	0.48	<0.001	0.05	\	\	\
Fall general knowledge score	0.85	<0.001	0.01	\	\	\
Model fit	\	\		0.73	1.62e+04	<0.001

Above table displaces the Ordinary Least Squares (OLS) regression model results. The intercept coefficient predicts the spring general knowledge score when all independent variables are 0. The coefficient 0.48 of income group indicates the estimated change in the spring general knowledge score for a one-unit increase in income group when holding fall general knowledge score constant. The coefficient 0.85 represents the estimated change in the spring general knowledge score for a one-unit increase in fall general knowledge score while holding income group constant. Since the p-value for all three coefficients are less than 0.001, there is strong evidence suggesting that they are all statistically significant predictors of the spring general knowledge score. The standard error for all the coefficients are low, suggesting a relatively precise estimation.

By examining the model fit, R-squared indicates that about 73.0% of the variability in the spring math scores can be explained by the independent variables in the model, which could be considered relatively high. The p-value associated with the F-statistics (PR(>F)) is less than 0.001, indicating the overall model is statistically significant, hence there is strong evidence to reject the null hypothesis.

Checking model diagnostics:





Normality Assumption:

The plot on the left above is a **Quantile-Quantile (Q-Q) Plot** which provides a visual assessment of the normality assumption. By examining the plot, almost all residuals align with the theoretical quantiles although there are some exceptions on the two sides. The histogram on the right above shows the **Residual Distribution**; it shows an almost perfect symmetric distribution, which suggests that the residuals are unbiased and random. Hence the normality assumption is met.

Homogeneity Assumption:

Barlett's test result: Test statistics (T) = 10.41; **p-value** ≈ 0.006

According to the histogram shown in EDA, the sample distribution for spring general knowledge score is normally distributed. Hence, to check the homogeneity assumption of the model, a Barlett's test has been conducted. Since the p-value is less than the significance level, it indicates that there is strong evidence to reject the null hypothesis of equal variances. This violation of the assumption of homogeneity indicates that at least two income groups show significant difference in variability of the spring math scores.

Conclusion

According to the EDA, by considering general knowledge as the academic performance baseline, there is a positive linear correlation between the fall scores and spring scores of kindergarten students. Under this increasing trend, students with higher household income tend to score higher among others. By examining the two ANCOVAs test results, there is strong evidence suggesting a significant difference detected in the mean spring math/general knowledge scores across different income groups after controlling the effects of covariate, fall math/general knowledge scores. As a whole, household income could be a crucial factor in affecting a child's performance in the context of education, and the past performance is also correlated to a child's performance in the later terms, meaning students who used to perform well will tend to achieve higher score in the future terms compared to those who did not perform well in the past. However, both models have violated some assumptions, indicating the findings from the test results may be unreliable and lack generalization. Drawbacks could be found in this study since it may not consider other confounding variables such as individual differences and culture background. In the future studies, interaction effects and other covariates may be considered to be included when exploring the features that might influence the academic performance of kindergarten students.