JIANHENG-CHEN-A3

Analyze the score difference between 1998 Fall and 1999 Spring considering the factor of Income categories

1. Introduction

In recent years, more and more families are paying attention to the importance of preschool education and hope their children can learn more basic knowledge before they go to school. However, there are many factors that could impact the effectiveness of preschool education. The children can be significantly varying in intelligence when they are very little, they may not be able to focus on learning as they have shorter attention span compared to adults. What is more, the income of the family can also be a unignorable factor to affect how they perform in kindergarten study.

This analysis will focus on the problems how the scores of kindergarten children change over time considering the different income categories. The methodology of choice will be ANCOVA, which helps us to adjust the outcome for difference in covariates. To dig into the patterns, our study will use the dataset named INF2178_A3_data regarding to early child longitudinal study (1998-1999). To be specific, not every column of variables will be studied, and our research will be focusing on three main research questions:

Research question1: Is there a significant difference in the reading score from fall 1998 to spring 1999 by different income groups?

Research question2: Can we observe statistically significant improvements in math score from fall 1998 to spring 1999 across different income groups?

Research question3: How do income groups affect the general knowledge score from fall 1998 to spring 1999?

By addressing these questions, we will be able to see how different income can be possibly affect the scores of kindergarten children and further understand the preschool child education situations.

2. Data Cleaning and Data Wrangling

In this dataset, we can observe **11933** rows and **9** columns. Specifically, all the data are numerical data in this dataset. By viewing the data info and descriptive data, we can see there are no missing data so we will not have to fill any space.

To perform further study on the research questions, we will not use all the columns, here are the variables of our choice:

- **fallreadingscore:** The kindergarten children's reading score in fall 1998

- **fallmathscore:** The kindergarten children's math score in fall 1998
- **fallgeneralknowledgescore:** The kindergarten children's general knowledge score in fall 1998
- **springreadingscore:** The kindergarten children's reading score in spring 1999
- **springmathscore:** The kindergarten children's math score in spring 1999
- **springgeneralknowledgescore:** The kindergarten children's general knowledge score in spring 199
- **incomegroup:** The income groups of the families, denoted as 1, 2 and 3

To be specific, the covariate income groups we use here is separating households with different total income into 3 categories, this variable will be used as a categorical variable in the further study.

3. Exploratory Data Analysis

	Fallreading	Fallmath	Fallgengeral knowledge	Springreading	springmath	Springgengeral knowledge
Mean	35.95	27.13	23.07	47.51	37.80	28.24
Std	10.47	9.12	7.40	14.32	12.03	7.58
min	21.01	10.51	6.99	22.35	11.90	7.86
25%	29.34	20.68	17.39	38.95	29.27	22.80
50%	34.06	25.68	22.95	45.32	36.41	28.58
75%	39.89	31.59	28.31	51.77	44.22	33.78
max	138.51	115.65	47.69	156.85	113.80	48.35

Table 1: Descriptive data of interested variables

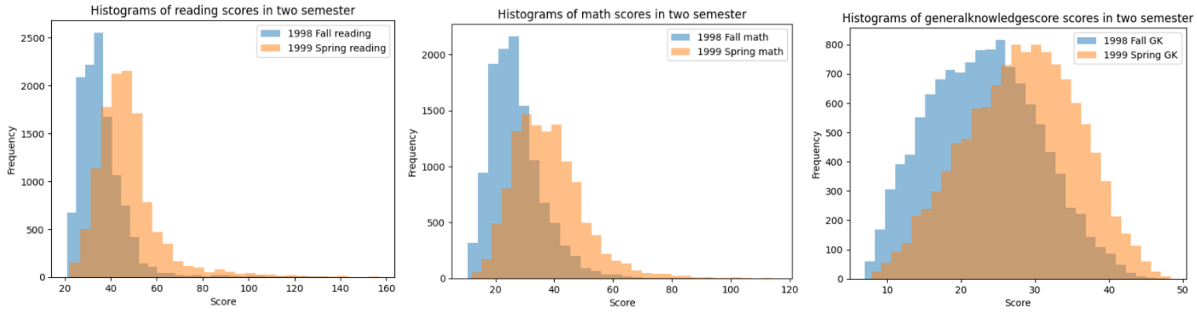

Figure 1: Histograms of reading, math, and general knowledge score in two semesters

In figure 1 and table 1, the first thing we can conclude that is all grades are right skewed, most of the children are not getting ideal grades. In addition, we can observe similar patterns change in the reading and math scores – the kurtoses are decreasing, and the means are increasing. We will say that the children are getting better reading and math grades from fall 1998 to spring 1999 based on these histograms. Moreover, a conclusion can be drawn that the general knowledge scores are less fat tail compared with the other two scores. The whole distribution is

moving right while there are more students getting a higher grade (however, the mean of general knowledge score is relatively low compared to the other scores).

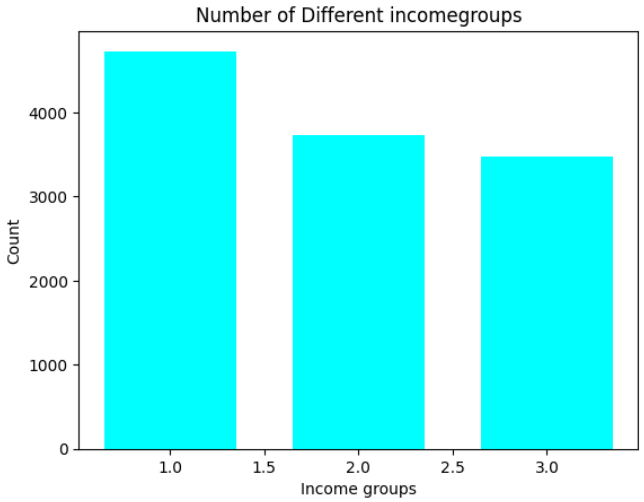

Figure 3: Bar plot of the number of households in different income categories

As for the income groups, a stepped distribution can be seen from the lowest income group to the highest income group.

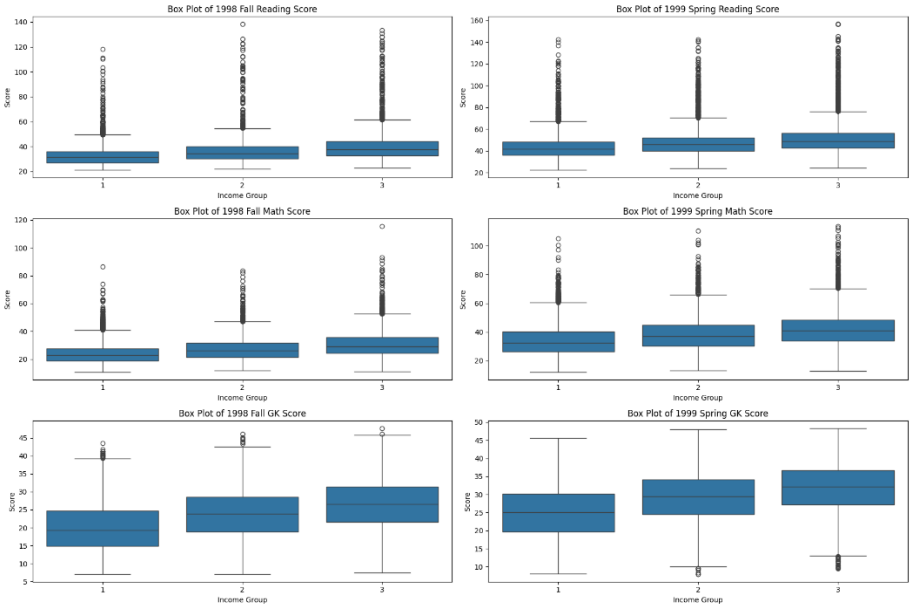

Figure 4: Box plots for different scores considering different income groups

According to the box plots, we will see the reading scores have most outliers in either year. Income group 3 has the highest median and smallest variation. The boxplot of math scores shows similar pattern, with less outliers and larger interquartile range. However, while fall 1998 general knowledge score has few outliers above the upper whiskers, there are outliers below

the lower whiskers in spring 1999.

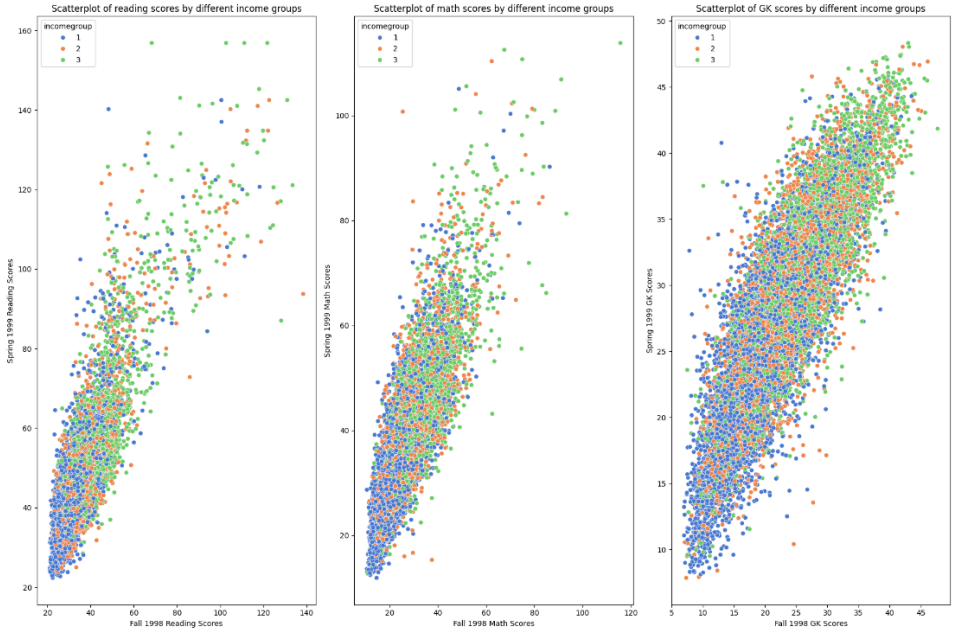

Figure 5: Scatter plots for different scores considering different income groups

Regarding to figure 5, students who scored higher in the fall also tended to score higher in the spring for reading scores. All income groups appear to follow this trend. Math scores have very similar shape, while income group 1 and 2 are more spread out with group 1 having a wider range of lower scores. The pattern for general knowledge scores is consistent with the previous subjects, with income group 3 showing higher scores overall.

4. One-Way ANCOVA

To address research question 1, we can fit a ANCOVA model while spring reading score is the dependent variable. The result is shown in the following table 2.

	coef	**Std err**	**t**	**p>|t|**
Intercept	6.54	0.264	24.78	<0.001
C(incomegroup)[T.2]	0.375	0.176	2.13	0.033
C(incomegroup)[T.3]	0.490	0.185	2.65	0.008
fallreadingscore	1.132	0.007	156.38	<0.001

R-squared	0.692
Adj. R-squared	0.692
F-statistic	8929
Prob (F-statistic)	<0.001
Log-Likelihood	-41675

Table 2: ANCOVA table for reading score

According to the table 2, we can draw a conclusion that the model has a moderate to strong explanatory power as approximately 69.2% of the variance in the dependent variable can be explained by the independent variables. The reading score in income group 2 and 3 both show a significant positive difference in the dependent variable compared to the baseline Income group 1 referring to the p-value.

The same step will be performed on math scores and general knowledge scores to address research questions 2 and 3.

	coef	Std err	t	p>|t|
Intercept	8.201	0.199	41.27	<0.001
C(incomegroup)[T.2]	0.670	0.151	4.43	<0.001
C(incomegroup)[T.3]	0.920	0.160	5.74	<0.001
fallreadingscore	1.074	0.007	149.01	<0.001

R-squared	0.681
Adj. R-squared	0.680
F-statistic	8469
Prob (F-statistic)	<0.001
Log-Likelihood	-39804

Table 3: ANCOVA table for math score

	coef	Std err	t	p>|t|
Intercept	8.030	0.119	64.519	<0.001
C(incomegroup)[T.2]	0.708	0.088	8.005	<0.001
C(incomegroup)[T.3]	0.942	0.094	10.013	<0.001
fallreadingscore	0.854	0.005	163.35	<0.001

R-squared	0.731
Adj. R-squared	0.731
F-statistic	1.082e+04
Prob (F-statistic)	<0.001
Log-Likelihood	-33259

Table 4: ANCOVA table for general knowledge score

With what we can see in table 3 and 4, the R-squared of 0.681 and 0.731 can prove the validity of the models. Same patterns in income group 2 and 3 showing significant increases for scores given the p-value smaller than 0.05.

5. Assumption Check

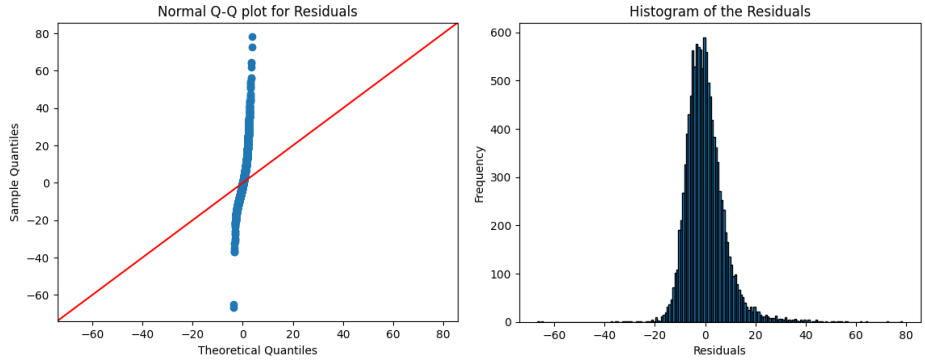

Figure 6: Q-Q & Distribution plot of residual (reading)

Regarding to Assumption 1: residuals are normally distributed; we can see the Q-Q plot points not fall along the reference line and there are extreme values existing. Histogram shows slightly right skewed. Combined with Shapiro-Wilk test, from which we have Shapiro statistic (**w**) equal to **0.91243** and p-value **<0.001**, we will reach a conclusion that the residuals for reading scores are not normally distributed. We will them perform Levene's test if the sample data is not normal distribution to check Assumption 2: variances are homogeneous. As a result, the Levene statistic (**w**) is **39.55** with p-value **<0.001**, which means the assumption 2 is also failed to meet.

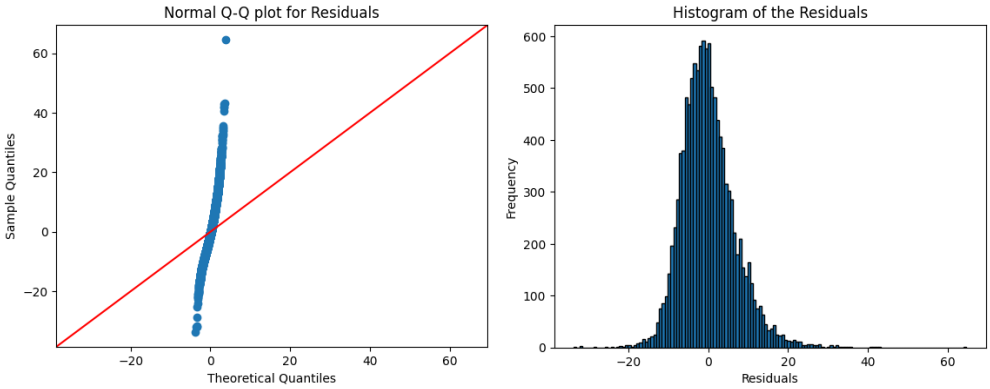

Figure 7: Q-Q & Distribution plot of residual (math)

In figure 7, we can have the same conclusion regarding to the Q-Q plot as we did with the reading score. The Shapiro statistic (**w**) equals to **0.9651** and p-value **<0.001**, we will say it fail to meet assumption 1. For the Levene statistic (**w**), **18.90** with p-value **<0.001**, which means the assumption 2 is also failed to meet.

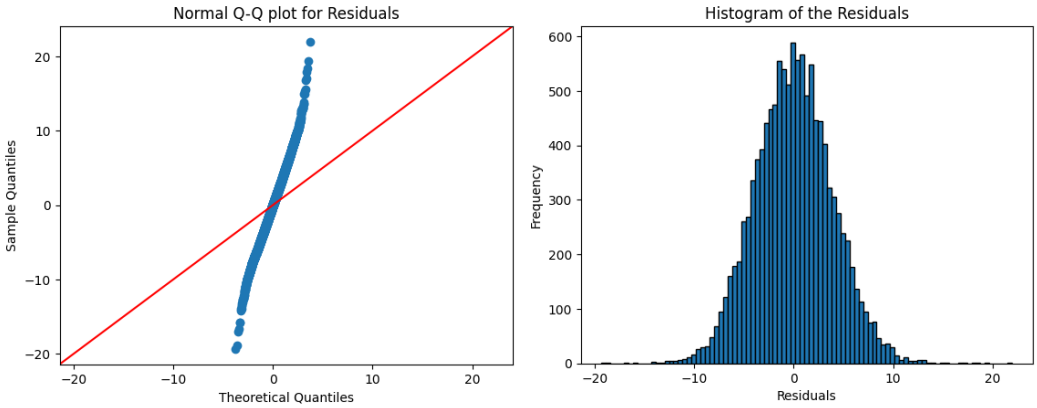

Figure 8: Q-Q & Distribution plot of residual (general knowledge)

The histogram is closed to normal and Shapiro statistic (**w**) equals to **0.998** and p-value **<0.001**. In this case, while the data is large enough the deviation from normality may not be of practical concern despite the statistically significant p-value. We perform the Bartlett statistic (**w**), 10.41 with p-value **= 0.0055**. We still fail to meet assumption 2.

6. Conclusion

Performing the EDA and ANCOVA analysis, we will reach a conclusion that kindergarten children born in households with higher income will receive greater improvements in scores in reading, math, and general knowledge. Assumptions checking is a problem regarding to the models and still require further study and modification to fit.