INF2178
Chi-shiun Yang
1009916897
Mar 20, 2024
Assignment 3

1. Introduction

In this analysis, we would like to learn if there are differences between scores of kindergarten students among different levels of income. To achieve the goal, we would like to explore the data collected from an early child longitudinal study.

There are four research questions in our exploration:

1. Do the three levels of income affect the reading, math, and general knowledge scores? If yes, how are they being influenced?
2. Do the three levels of income affect the total scores? If yes, how are they being influenced?
3. Do childrens' reading and math scores change from fall to spring influenced by the difference in levels of income?

2. Data Cleaning and Wrangling

The raw data contains 9 columns and 11933 rows. After reviewing the dataset, some columns are not used in this analysis while some columns are created for easier calculation. The newly created columns *falltotal* and *springtotal* are the total of students' scores in fall and spring calculated by adding the columns of reading, math, and general knowledge scores of the corresponding season. And the columns *readingdifference*, *mathdifference*, and *generaldifference* are the difference of the scores from spring to fall.

Beside the new create columns and the first six columns of single scores, the column *incomegroup* are used. There are three levels of income groups, level 1 is the income less than 40,000, level 2 is the income from 40,000 to 70,000, and level 3 is the income above 70,000.

3. Analysis of Covariance (ANCOVA)

For each research question, we would draw the scatter plot, the regression lines, and box plot for the column to visualize our dataset. After that, we would apply ANCOVA (analysis of covariance) to the data. We use these three values for our research questions:

1. The p-value (P>|t|) of our categorical factor (incomegroup)
2. Prob(Omnibus) is the measurement of the probability that residuals are normally distributed. 1 indicates a perfectly normal distribution.
3. Skew to measure the symmetry of our data. 0 is a perfect symmetry.

To make sure running ANCOVA in this case is reliable, we have to draw some plots to make sure that the dataset satisfied the assumptions. The assumptions of ANCOVAs are that each factor level has a normal population distribution (we use q-q plots, histograms, and Shapiro Wilk test to check residuals are normally distributed), the distributions have the same variance (we are using the Bartlett's test or the Levene's test), and the data are independent (this is true for this dataset). On the other hand, if the p value is less than our significance level (0.05 in this analysis), we reject the null hypothesis that the mean between groups are the same, and we would then use post-hoc tests with Tucky HSD to see where the difference is.

Research question 1

To discover whether income affects each score, we would like to use ANCOVA with corresponding spring scores as the dependent variable, income level as the categorical factor, and fall scores as the covariate.

For the reading score, we can see that the distributions are quite similar for each group. The p-value for this test is 0.006, which is much less than 0.05. Therefore, we reject the null hypothesis and say that the mean is different for each income group.

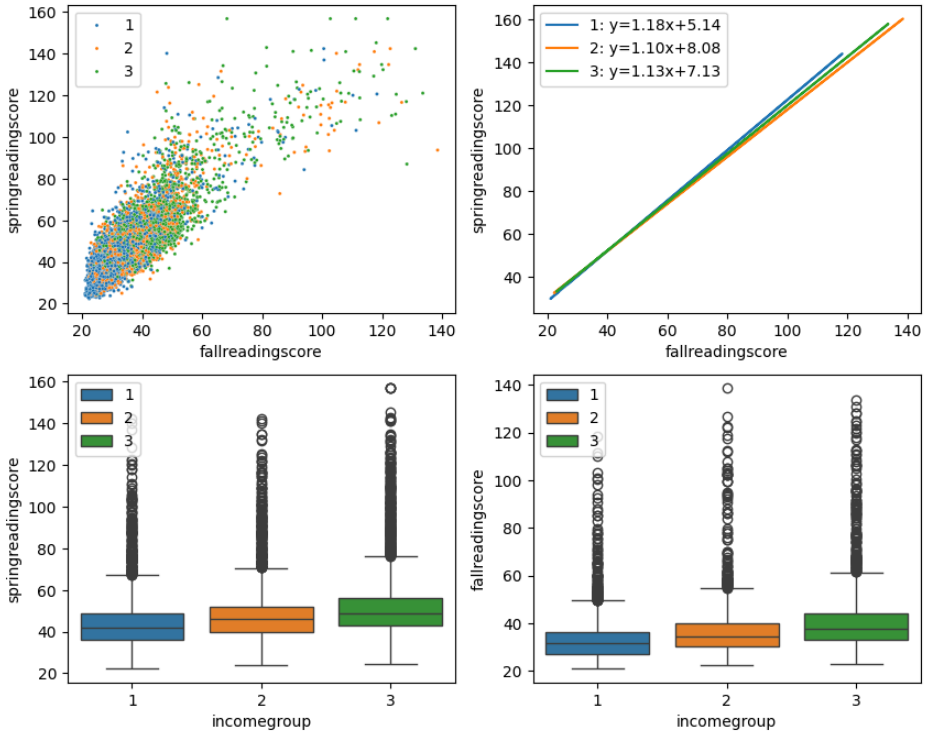

Figure 1: The scatter plot, regression line, and box-plots of reading scores grouped by incomes.

Since the means are different, we applied a post hoc test using Tukey's HSD. From the summary table, we can see that all p-values are lower than 0.05. This tells us that means among each group are different and families with higher income result in higher reading scores.

group 1	group 2	Mean Diff	Lower	Upper	p-value
1	2	4.34	3.63	5.06	<0.001
1	3	8.54	7.81	9.27	<0.001
2	3	4.20	3.43	4.97	<0.001

Table 1: Tukey summary of spring reading scores between different income groups.

From the q-q plot and the histogram, we can see that it is a bit skewed to the right, which gives the same result as the Skew value 1.371 in our ANCOVA result. And since the Shapiro Wilk test has a p-value <0.01, we use the Levene's test to see if the variances are homogeneous. The p-value is 7.57e-18 so from these results, the assumptions are not met so the reliability of the test is not very high but we can still somewhat say that the higher income result in higher reading score.

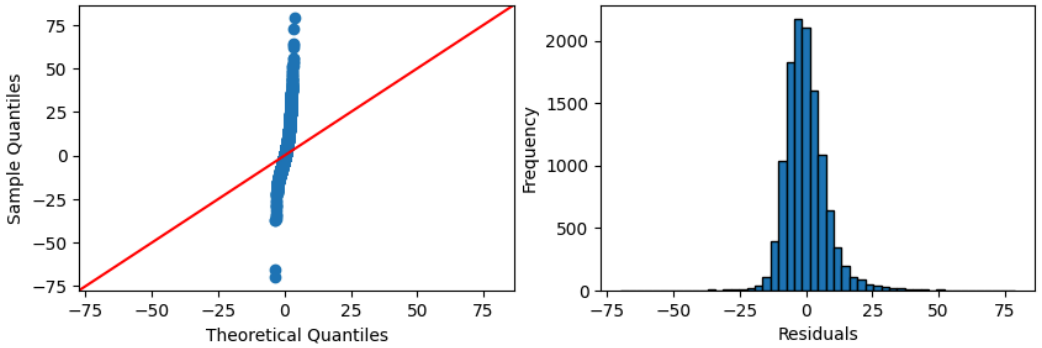

Figure 2: q-q plot and histogram of residuals.

For the math scores, we get a very similar result to the reading scores. The p-value of the ANCOVA test is less than 4.12e-09 < 0.05, the means among each group are different and families with higher income result in higher math scores. The samples are skewed to the right and it also does not meet the assumption that the variances are homogeneous.

For the general knowledge score, the distribution of the data is better. From the box-plots, we can see that there are less outliers and most of the outliers are lower scores instead of the higher ones in reading and math scores. The reason might be that reading and math scores can be improved more easily if the children get any training in the subjects but the general knowledge is accumulated from daily lives. The p-value of this ANCOVA test is 1.99e-24 < 0.05, so we said that the mean are not all the same. From the result of Tukey HSD, we can also see that the mean are different when comparing any two groups. And similar to the previous tests, although the histogram of the residuals are more symmetric, both the assumptions of normal distribution and same variance have not been met.

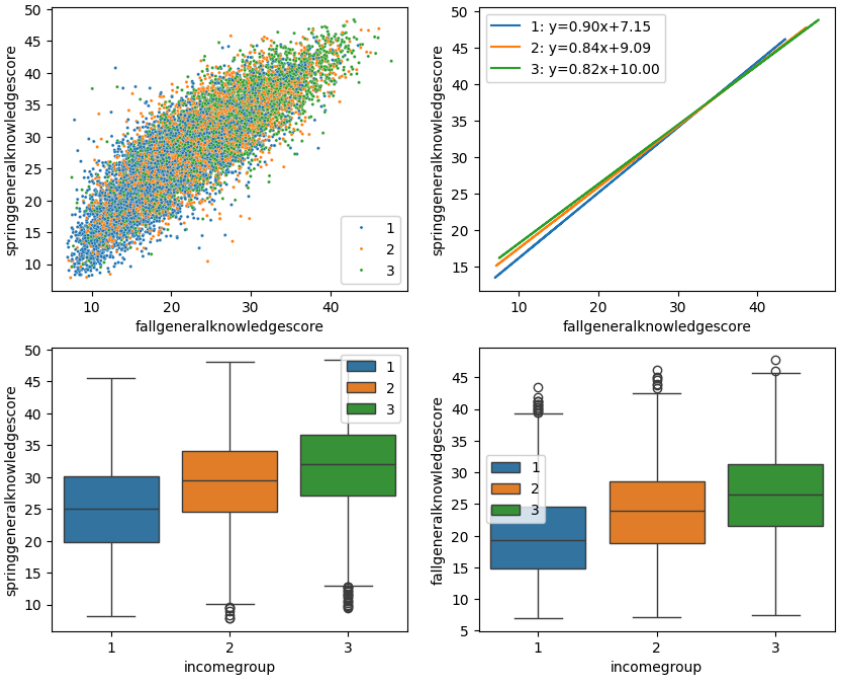

Figure 3: The scatter plot, regression line, and box-plots of general knowledge grouped by incomes.

Research question 2

For this research question, we would like to discover whether income affects the total score. We would also use ANCOVA with spring total score as the dependent variable, income level as the categorical factor, and fall total score as the covariate. We draw the same plots and run the same tests as we did in research question 1 just with different columns of the value we want to explore. From the ANCOVA summaries, the p-value is 0.174 > 0.05, so we do not reject the null hypothesis saying that the means of total score for each income group are equal.

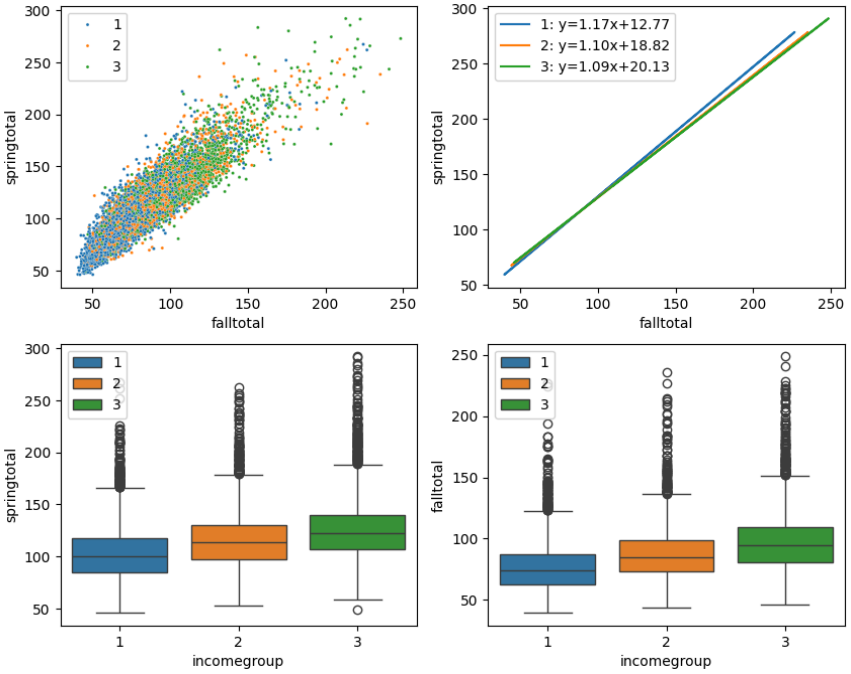

Figure 4: The scatter plot, regression line, and box-plots of total scores grouped by incomes.

Since we do not reject the null hypothesis, Tukey HSD does not need to be run, so we continue to test the assumptions. The p-value of the Shapiro Wilk test is 3.43e-43 < 0.05, and that of the Levene's test is 2.06e-06 < 0.05. The sample is not normal nor has the same variance among different income groups.

Research question 3

To discover whether the score changes are influenced by the difference in levels of income, we would like to apply ANCOVAs with reading/math differences as the dependent variable, income level as the categorical factor, and general knowledge differences as the covariate. From the plots, the scatter plot dependent variable and covariate does not have linear relations like the previous ones, the regression lines are more parallel especially in group 1 and 3, and the outliers in boxplots are more evenly distributed. The p-value is 6.89e-16 < 0.05, so we reject the null hypothesis and say the mean are not equal.

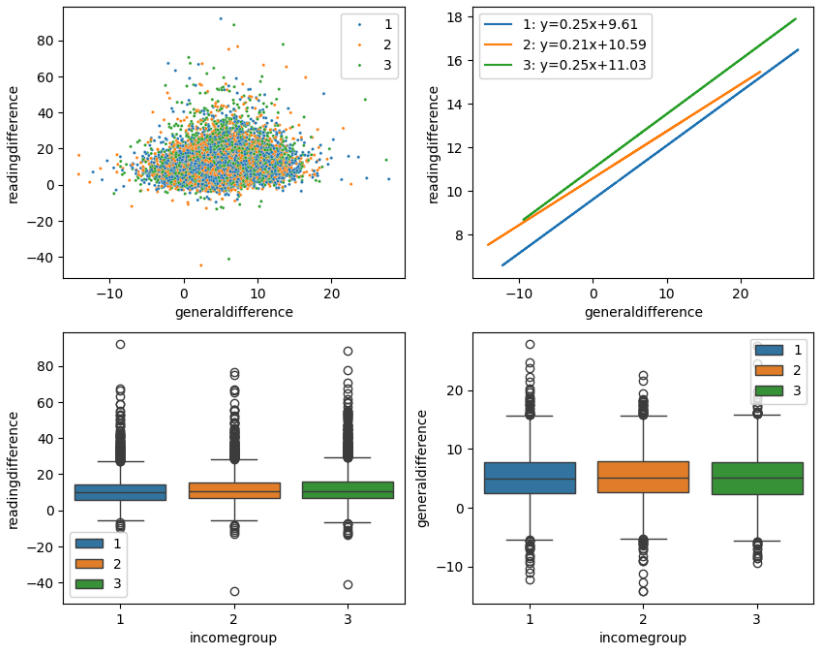

Figure 5: The scatter plot, regression line, and box-plots of reading difference grouped by incomes.

From the Tukey HSD, we can see that the difference between group 2 and 3 is slightly smaller. And similar from previous analysis, the assumptions are not met. And for math scores, the results are identical to the reading scores.

group 1	group 2	Mean Diff	Lower	Upper	p-value
1	2	0.84	0.42	1.25	<0.001
1	3	1.43	1.01	1.85	<0.001
2	3	0.59	0.14	1.04	0.005

Table 2: Tukey summary of reading score difference between different income groups.

4. Conclusion

From the results of the three research questions we have answered, the first conclusion is that the mean of reading, math, and general knowledge scores is different when grouped by income of the family. Children from higher income families have higher scores, and the difference in reading and math scores are larger than that of general knowledge. But for the total score, we do not reject the null hypothesis so we cannot say that the total scores are affected by the income group. Last but not least, from the third research question, we have that the mean of improvement from fall to spring for reading and math scores are all different. Children from higher income families also improve more in scores. A limitation of this analysis is that in most cases, the distributions are not normal and the variance are not homogeneous, which lowers the reliability of the ANCOVAs' results. Combining other methods to solve this problem will give us a more accurate analysis.