Name: Berke Derin Berktay

Exploring the Impact of Income Group on Kindergarten Academic Performance

1. Introduction

The relationship between household income and academic achievement has long been a subject of educational research. This study examines how the income group may influence the average reading and math performance of Kindergarten students over an academic year, controlling for their baseline general knowledge. This report presents a comprehensive analysis using one-way Analysis of Covariance (ANCOVA) to understand the influence of income group on the average yearly academic scores of Kindergarten students, accounting for the baseline scores in general knowledge. The data, sourced from an early child longitudinal study, features readings from fall 1998 to spring 1999, evaluating students over several months. This study will try to answer the following research questions:

Research Question 1: Does income group have a effect on the reading performance of students over the academic year when controlling for their baseline general knowledge? If so, how significant of an effect does it have?

Research Question 2: Does income group have an effect on the math performance of students over the academic year when controlling for their baseline general knowledge? If so, how significant of an effect does it have?

2. Data Cleaning and Data Wrangling/ Feature Engineering

After the loading and observation of the data, since there is no information given regarding the incomegroup, we have to understand what it is. It seems like it groups by income the top third middle third and bottom third of the entries. By computing the min and maxes of each group we confirm that this is the case. Then, we look at the amount of zero and null values for each entry, which would not make sense for any of the columns. After calling the function, we see that there are no entries that are either null or zero, so we can move on. Now, the dataset contains two types of variables for the same performance metric, one for fall and spring. Since the goal of this study is to examine the math and reading performance of the subjects irrespective of the season, three new columns are created: yearreadingscore, yearmathscore, and yeargeneralknowledgescore, which are computed by taking the average of their two corresponding fall and spring timeline metric values. These new variables give a clearer and more accurate picture of the scores of the subjects. Therefore, to answer the research questions, these 3 new variables will be used instead of the fall and spring variables. Hence, after this operation, the fall and spring columns for each metric are deleted.

The two one-way ANCOVA that'll be performed for question 1 is setup like below:
Dependent Variable (DV): yearreadingscore (average of fall and spring reading scores)
Covariate: Year general knowledge score (yeargeneralknowledgescore)
Factor: Income group (incomegroup)

The two one-way ANCOVA that'll be performed for question 12is setup like below:
Dependent Variable (DV): yearmathscore (average of fall and spring reading scores)
Covariate: Year general knowledge score (yeargeneralknowledgescore)
Factor: Income group (incomegroup)

Therefore, since the factor for the tests is income group, its form needs to be categorical, or in other words, grouped up. The incomegroup variable is a great candidate, and therefore since incomegroup is used for the ancova's and EDA later on, the incomeinthousands is not needed, and hence deleted.

3. Exploratory data Analysis (EDA)

Now, for the purposes of answering the research question, it is a good idea to graph the year reading score vs income group and year math group vs income group as a boxplot and swarmplot on top of each other is a great idea to see if the initial observations by just looking at the data supports the claim of an effect of class on the scores in both cases. Hence, the following two graphs were produced:

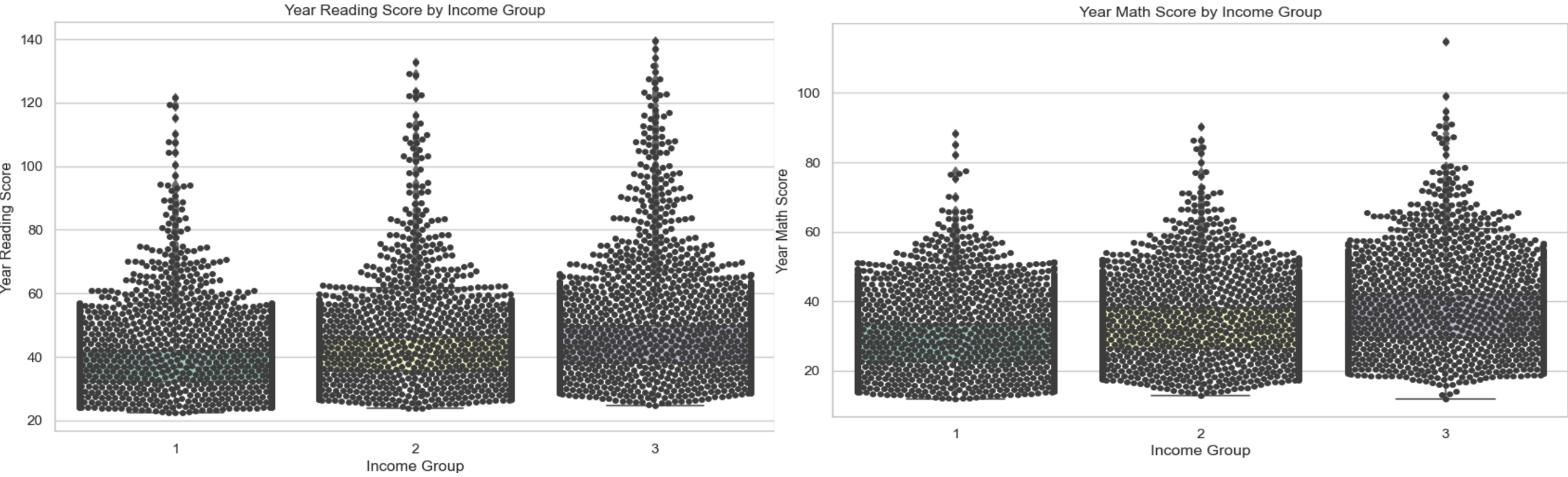

FIGURE 1: Year Reading Score vs Income Group (Left) & Year Reading Score vs Income Group (Right)

Both of these graphs suggest that the higher income group correlated with higher scores for both reading and math. It is also a good idea to perform the summarystatistics of all 3 groups in both terms of both reading and math scores to see if this observation is supported further. The results are in the below figure:

	Min	Max	Mean	25th %	Median	75th %	IQR
Group1 R	22.37	121.52	38.23	32.05	36.74	41.92	9.87
Group2 R	23.7	132.61	42.15	35.52	40.21	46	10.48
Group3 R	24.67	139.35	46.05	37.99	43.23	49.7	11.71
Group1 M	11.85	88.28	28.9	22.72	27.56	33.53	10.81
Group2 M	12.95	90.18	33.02	26.36	31.67	38.17	11.81
Group3 M	11.8	114.72	36.71	29.48	35.22	42.01	12.53

FIGURE 2: Summary Statistics of each Reading and Math Scores of the 3 Income Groups (Math=M, Reading=R)

All 6 of these summarystatistics for the groups back up this claim of corelation since the higher you go in terms of the income group levels, the higher the mins, maxs, and the averages get. Now is a good time to test the intuitions' and statistics' claims further by performing two one-way ANCOVAS, one for the reading scores, and one for the math scores.

4. Performing One-WAY ANCOVA's

The details of the two one-way ANCOVA's are stated at the end of Section 2. Now, we perform the first one-way ANCOVA. Below are the results:

	sum_sq	df	F	PR(>F)
C(incomegroup)	1.402498e+04	2	67.318143	8.469463e-30
yeargeneralknowledgescore	3.176658e+05	1	3049.511682	~0.0
Residual	1.242637e+06	11929.0	NaN	NaN

FIGURE 3: First ANCOVA Results

The sum of squares represents the variance in the dependent variable attributable to the differences between the income groups. The F value is the measure of the ratio of the variance explained by the income group to the unexplained variance within the groups. A higher F-value typically indicates a more significant effect, and a low p-value, below 0.05, suggests that would suggest strong evidence against the null hypothesis (which in this case is that there is no difference in the means across income groups). For the incomegroup, p-value is extremely low, indicating indicating a strong and significant difference between the reading scores values of each income group. The F-value is also 67.32, which is relatively high. For the yeargeneralknowledgescore, we even have a higher F-value and even a lower p-value, which indicates an even stronger effect. The residual sum of squares of 1242637 though suggests that there is a big amount of variation that is not explained by these two variables. In conclusion, the results suggest that both income group and year general knowledge score significantly affect the dependent variable (reading scores), which backs up the previous intuition. Though it is worthy to note that as logically expected, the effect of general scores are much higher on the reading scores than the income. Now, the following is the results of the second one-way ANCOVA, which observes the math scores of the students from the same perspective.

	sum_sq	df	F	PR(>F)
C(incomegroup)	8638.204010	2	73.633092	1.649436e-32
yeargeneralknowledgescore	395600.561330	1	6744.293720	~0.0
Residual	699720.280895	11929.0	NaN	NaN

FIGURE 4: Second ANCOVA Results

The results suggests the exact same conclusions as the previous test. For the incomegroup, the F-statistic is 73.6, which is considered high, and the p-value is below 0.05. Similarly for the general score, the f-statistic is a whooping 6744 and the p-value is almost zero.

These results suggest that just like their effects on the reading scores, both income group and year general knowledge score significantly affect the dependent variable (the math score in this case), which again backs up the previous intuition. The effect of the general score of the subject is much higher than the income on the math scores. It can be seen that the F-statistic for the income group is higher for math scores compared to reading scores, indicating a slightly stronger effect of income group on math scores. The p-values in both cases are extremely small, indicating significant effects in both scenarios. Additionally, the F-statistic for the general knowledge scores is substantially higher for math scores (6744.29) than for reading scores (3049.51). This indicates a much stronger effect of general knowledge on math scores compared to reading scores. The p-values for both are essentially zero, indicating very significant effects. Therefore, from these observations it could be stated that the general knowledge of the children has distinctly much more effect on the math scores of the children than it has on the reading scores of the children. This suggests that general cognitive knowledge and learning overlaps more with learning math than it does with reading. Similarly, though not to the same extent, the income factor has a bit more of an effect on the math scores of the children than it has on the reading scores. This could be explained by the fact that learning how to read requires less funding than developing the children’s math skills, which might require the family to hire a tutor. Whereas, simply by reading books, children can improve their reading skills, which is a cheaper form of learning than hiring a tutor. Of course, it is also important to note that the assumptions for performing ANCOVA has to be tested as well. If they are violated we should take the results with a pinch of salt, or in the future, before performing the tests, some sets of transitions can be applied to the data. For the linearity assumption, predicted vs. residuals plots are created from the models’ results. For the normality assumption, a qqplot is created and a shapiro-wilks test is performed. Lastly, for the homogeneity of variances, a breusch-pagan test is performed. The results of the tests for the first ANCOVA are listed below:

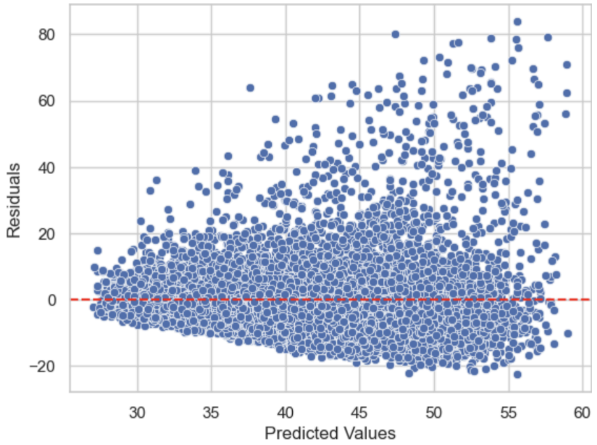

FIGURE 5: Residual vs Predicted Values for the First ANCOVA Model

Statistic	P-value
0.8038636445999146	~0.0

FIGURE 6: Shapiro-Wilk Test Results for the First ANCOVA Model

Name: Berke Derin Berktay

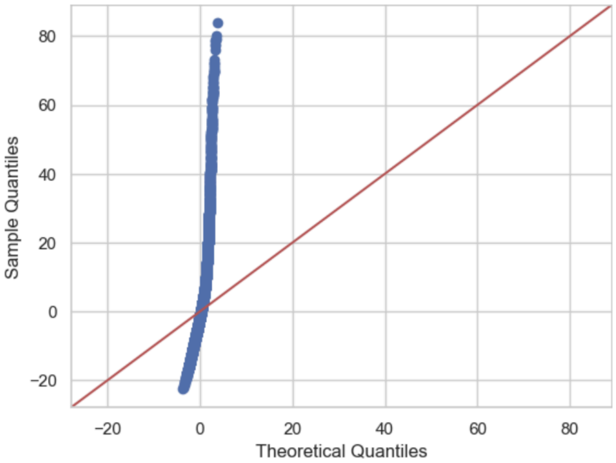

FIGURE 7: Sample Outliers vs Theoretical Quantities for the First ANCOVA Model

Lagrange Multiplier Statistic	P-value	f-value	f p-value
447.71280083416576	1.0204862971468354e-96	155.00311858515988	1.4832521033147978e-98

FIGURE 8: Breusch-pagan's Test Results for the First ANCOVA Model

The predicted vs residual plot of the model suggest that there is indeed linearity since the data points are relatively evenly distributed around the horizontal line of x=0. The Shapiro-Wilk test statistic of 0.8039 and a p-value of 0.0 indicate that the residuals of the model do not follow a normal distribution. The qq-plot does not follow the red line, which again suggests no normality strongly. Lastly, The Breusch-Pagan test results with a Lagrange multiplier statistic of 447.71 and extremely low p-values indicate strong evidence of heteroscedasticity, meaning that the variance of the residuals is not constant across the predicted values, which violates the assumption of homogeneity of variances. Now, let us look at the results of the tests and graphs for the assumptions of the second ANCOVA performed below:

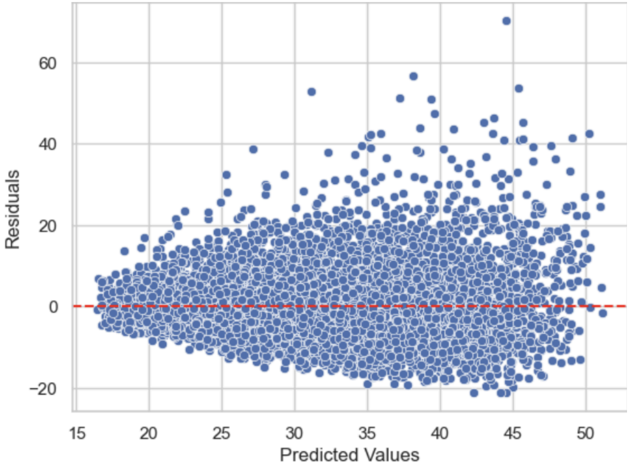

FIGURE 9: Residual vs Predicted Values for the Second ANCOVA Model

Statistic	P-value
0.9446665048599243	~0.0

FIGURE 10: Shapiro-Wilk Test Results for the Second ANCOVA Model

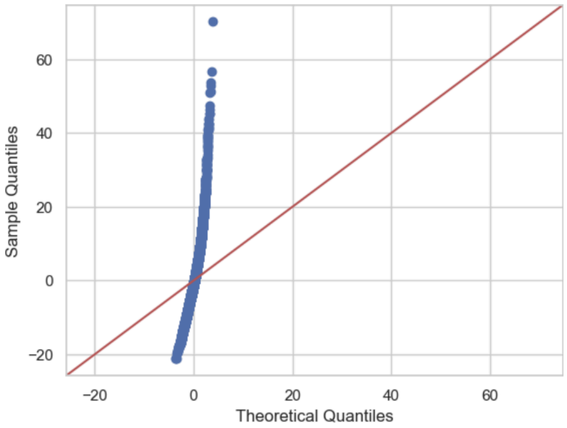

FIGURE 11:Sample Outliers vs Theoretical Quantities for the Second ANCOVA Model

Lagrange Multiplier Statistic	P-value	f-value	f p-value
463.4460029432254	3.9797230162781984e-100	160.6702222402139	4.244640226057436e-102

FIGURE 12: Breusch-pagan's Test Results for the Second ANCOVA Model

Similarly, the predicted vs residual plot of the model suggest that there is indeed linearity since the data points are relatively evenly distributed around the horizontal line of x=0. The Shapiro-Wilk test statistic of 0.9447 and a p-value of 0.0 indicate that the residuals of the model do not follow a normal distribution. The qq-plot does not follow the red line, which again suggests no normality strongly. Lastly, The Breusch-Pagan test results with a Lagrange multiplier statistic of 463.446 and extremely low p-values indicate strong evidence of heteroscedasticity, meaning that the variance of the residuals is not constant across the predicted values, which violates the assumption of homogeneity of variances. For both tests, we should take the results with a little pinch of salt due to the assumption violations.

5. Conclusion

This study explored the impact of income groups on Kindergarten academic performance, revealing that income significantly affects both reading and math scores even when the effect of the baseline general knowledge is taken into account, with a more pronounced effect on math. General knowledge showed an even stronger influence, particularly on math, suggesting that early cognitive abilities play a crucial role in educational outcomes. While the findings are compelling, assumptions violations in the ANCOVA tests warrant cautious interpretation. Overall, the study highlights the importance of socio-economic factors and foundational cognitive skills in early education, underscoring the need for supportive interventions to bridge disparities. In the future, transformations to the models can be applied for the purpose of not violating the ANCOVA assumptions.