Student ID: 1010737874
Student name: Han Zheng

Exploring Kindergarten Scores

1. Introduction

In this analytical study, I delve into the educational achievement across different socioeconomic strata, measured by the variable 'incomegroup.' The dataset encapsulates an array of student scores in reading, math, and general knowledge, recorded in two different sessions—Fall 1998 and Spring 1999. The primary objective of the research is to dissect and comprehend the variation in academic performance, specifically focusing on the reading and math score differences over time.

I am particularly interested in understanding whether the 'incomegroup'—categorized into three levels—serves as a significant determinant of academic progress, as observed through the lens of reading score improvements from fall to spring. I utilize the difference of general knowledge score between two terms as a baseline, acknowledging its potential role in academic advancement.

Employing a methodological framework rooted in Analysis of Covariance (ANCOVA), I aim to control for initial disparities in general knowledge while evaluating the significance of the 'incomegroup' on reading score improvements. Prior to running our ANCOVA, I undertake rigorous data preparation, which includes cleaning the dataset by identifying and excluding outliers—a critical step to ensure the robustness of our findings. Subsequently, I conduct diagnostic checks to verify the assumptions intrinsic to linear regression: linearity and normality of residuals.

Our exploration will address three fundamental research questions:

- **Research Question 1**: Does household income group influence the improvement in reading scores from Fall 1998 to Spring 1999?
- **Research Question 2**: Does household income group influence the improvement in math scores from Fall 1998 to Spring 1999?

By addressing these questions, we aim to contribute insights into the scores difference in reading, and math from Fall 1998 to Spring 1999.

2. Data Cleaning and Data Wrapping

The raw data has **9 columns** with **11934** entities (**rows**). After initial review of the datasets we were confident that not much data cleaning was deemed necessary for the analysis since no variables have missing values.

Observations and Considerations:

Since our analysis only focused on a few aspects, we want to work on the following columns. Below we provide a short description of each column:

- fallreadingscore: score of reading in Fall 1998
- fallmathscore: score of math in Fall 1998
- fallgeneralknowledgescore: score of general knowledge in Fall 1998
- spirngreadingscore: score of reading in Spring 1999

- spirngmathscore: score of math in Spring 1999
- spirnggeneralknowledgescore: score of general knowledge in Spring 1999
- incomegroup: group of income which divided by income, and stated as 1, 2, and 3

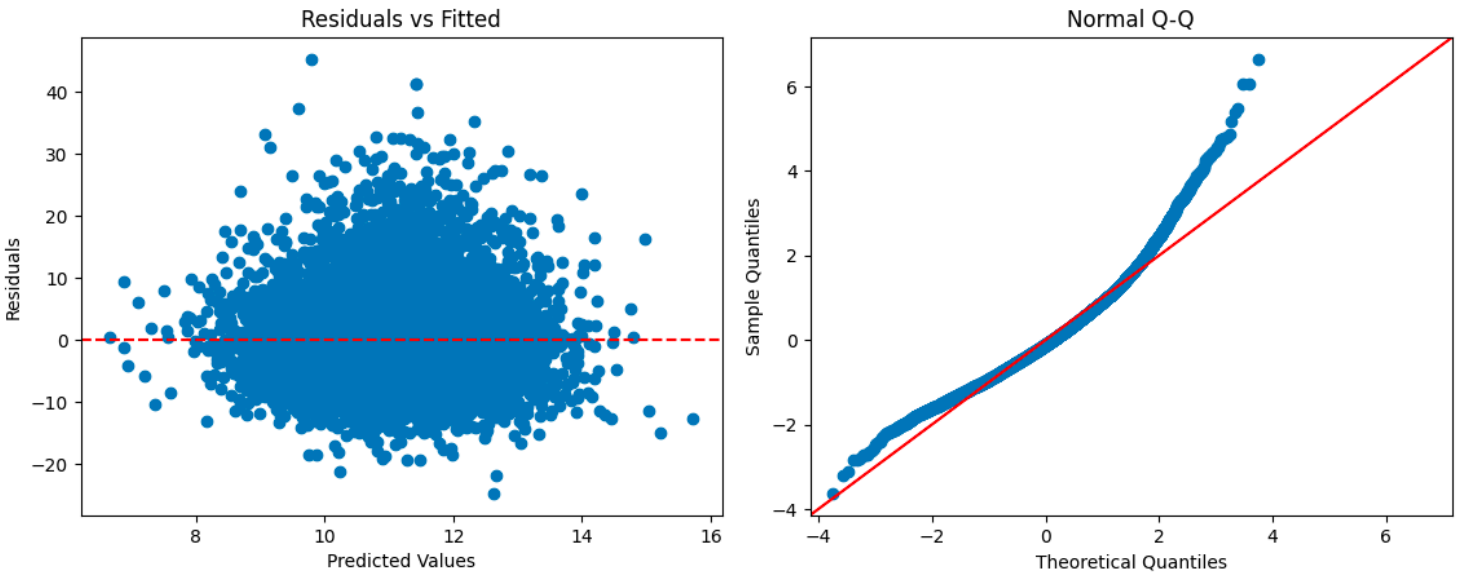
3. Exploratory Data Analysis (EDA)

In the process of EDA, I first ensured that there were no missing values for each variable, and then removed outliers in the dataset to reduce the impact of unusual data and improve the accuracy of the analysis. Through EDA, I establish a foundation for building a robust model that aims to discern the influence of income group on score differences, using general score as a baseline for comparison.

4. Reading Scores Over Time

Research Question #1: Does household income group influence the improvement in reading scores from Fall 1998 to Spring 1999?

For the first research question, the analysis aims to explore the impact of income, represented by different groups, on the reading scores of kindergarten over time (Fall 1998 and Spring 1999). To check the assumptions of ANCOVA, I firstly made two plots shown below as Graph 1. The first one is a Residuals vs Fitted plot, which is used to check the linearity and homoscedasticity. Ideally, we would expect to see a random scatter of points around the horizontal line at zero, with no distinct patterns. While the first plot is not so ideal, the second QQ plot basically matches the pattern I expect.



(Graph 1)

I try to fit ANCOVA without interaction and get the Table 1 below. The OLS regression results present a model examining the influence of income group and general knowledge score difference on the difference in reading scores. The model includes 11,490 observations and features an R-squared value of 0.021, indicating that about 2.1% of the variance in the reading score difference is explained by the model, which is a relatively small amount. The coefficients for both income groups 2 and 3 are significant ($p < 0.0001$), suggesting that being in these income groups is associated with an increase in the reading score

difference compared to the reference income group (likely income group 1), with income group 3 having a slightly larger effect than group 2.

OLS Regression Results

Dep. Variable:	reading_score_diff	R-squared:	0.021
Model:	OLS	Adj. R-squared:	0.021
Method:	Least Squares	F-statistic:	81.78
Date:	Mon, 25 Mar 2024	Prob (F-statistic):	2.38e-52
Time:	00:16:58	Log-Likelihood:	-38333.
No. Observations:	11490	AIC:	7.667e+04
Df Residuals:	11486	BIC:	7.670e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.4291	0.128	73.669	0.000	9.178	9.680
C(incomegroup)[T.2]	0.6531	0.151	4.323	0.000	0.357	0.949
C(incomegroup)[T.3]	0.8605	0.156	5.528	0.000	0.555	1.166
gk_score_diff	0.2264	0.016	14.441	0.000	0.196	0.257

Omnibus:	1765.075	Durbin-Watson:	1.700
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3666.109
Skew:	0.928	Prob(JB):	0.00
Kurtosis:	5.052	Cond. No.	21.2

(Table 1)

Then I try to fit ANCOVA with interaction and get the Table 2 below. The coefficients for income groups 2 and 3 remain significant, implying that these income groups have a notable impact on reading score differences when compared to the baseline group (income group 1). The general knowledge score difference continues to be a significant predictor. However, when interaction terms are introduced, they do not appear to be significant (p-values for the interaction terms are 0.108 and 0.939 for income groups 2 and 3, respectively), suggesting that the effect of general knowledge score difference on reading score difference does not significantly differ across income groups.

OLS Regression Results

Dep. Variable:	reading_score_diff	R-squared:	0.021
Model:	OLS	Adj. R-squared:	0.021
Method:	Least Squares	F-statistic:	49.68
Date:	Sun, 24 Mar 2024	Prob (F-statistic):	4.45e-52
Time:	23:39:00	Log-Likelihood:	-38331.
No. Observations:	11490	AIC:	7.667e+04
Df Residuals:	11486	BIC:	7.672e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.3292	0.160	58.283	0.000	9.015	9.643
C(incomegroup)[T.2]	0.9648	0.245	3.931	0.000	0.484	1.446
C(incomegroup)[T.3]	0.8753	0.252	3.478	0.001	0.382	1.369
gk_score_diff	0.2460	0.024	10.047	0.000	0.198	0.294
C(incomegroup)[T.2]:gk_score_diff	-0.0598	0.037	-1.607	0.108	-0.133	0.013
C(incomegroup)[T.3]:gk_score_diff	-0.0030	0.039	-0.077	0.939	-0.079	0.073

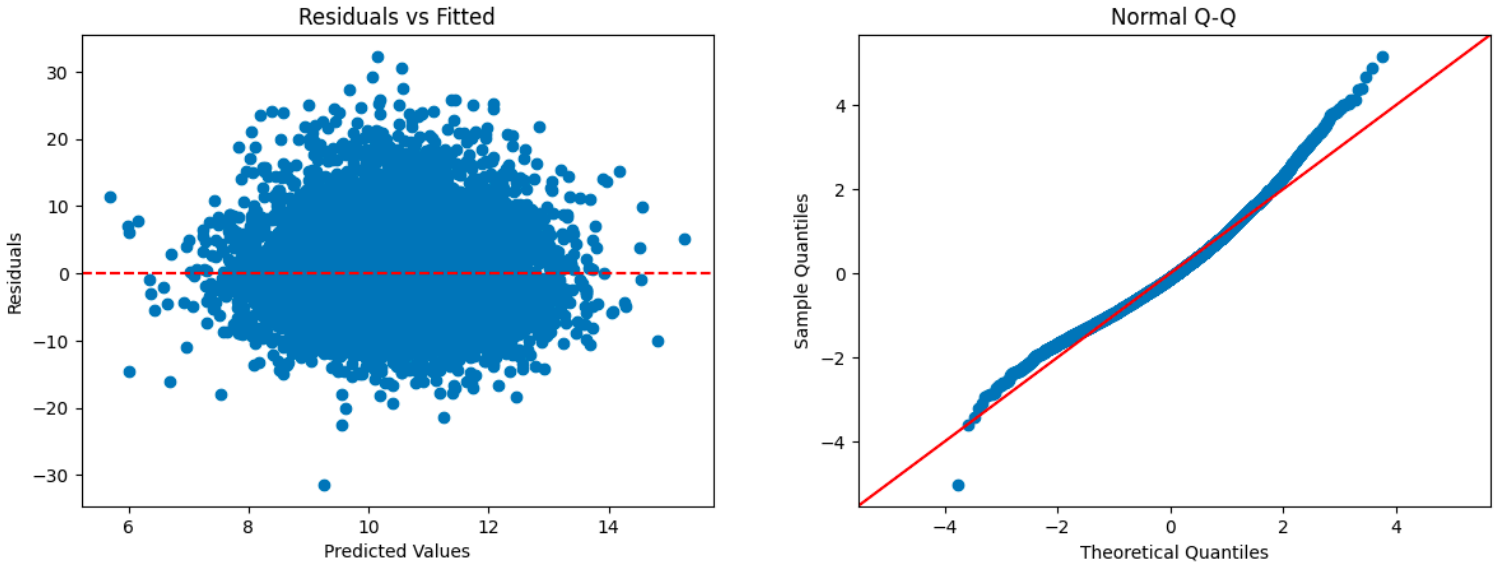
Omnibus:	1762.903	Durbin-Watson:	1.700
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3662.248
Skew:	0.927	Prob(JB):	0.00
Kurtosis:	5.052	Cond. No.	37.1

(Table 2)

5. Math Scores Over Time

Research Question #2: Does household income group influence the improvement in math scores from Fall 1998 to Spring 1999?

For the second research question, I want to explore the similar question but only changed the main character from reading score to math score. I also firstly check the assumptions of ANCOVA and get 2 plots shown below as Graph 2.



(Graph 2)

The two diagnostic plots for the math score model above show that although there are still some concerns for the Residuals vs Fitted plot, the QQ plot is much better than the reading scores' one. Similar to the previous OLS Regression Results, the model includes 11,490 observations and features an R-squared value of 0.021, indicating that about 2.1% of the variance in the reading score difference is explained by the model, which is a relatively small amount. Besides, the diagnostic tests raise concerns: the Omnibus test yields a p-value of 0.000, indicating that the residuals are not normally distributed, which is confirmed by the Jarque-Bera test. The Durbin-Watson statistic of 1.813 suggests there may be a modest amount of autocorrelation. Finally, the Condition Number is relatively high, signaling potential multicollinearity issues which could affect the stability and interpretation of the model coefficients.

OLS Regression Results

Dep. Variable:	math_score_diff	R-squared:	0.029
Model:	OLS	Adj. R-squared:	0.029
Method:	Least Squares	F-statistic:	116.4
Date:	Sun, 23 Mar 2024	Prob (F-statistic):	3.14e-74
Time:	23:40:12	Log-Likelihood:	-37365.
No. Observations:	11490	AIC:	7.474e+04
Df Residuals:	11486	BIC:	7.477e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.6180	0.118	73.250	0.000	8.387	8.849

C(incomegroup)[T.2]	0.7749	0.139	5.579	0.000	0.503	1.047
C(incomegroup)[T.3]	1.1675	0.143	8.159	0.000	0.887	1.448
gk_score_diff	0.2390	0.014	16.583	0.000	0.211	0.267

Omnibus:	801.564	Durbin-Watson:	1.813
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1076.555
Skew:	0.618	Prob(JB):	1.69e-234
Kurtosis:	3.848	Cond. No.	21.2

(Table 3)

And then, for the ANCOVA with interaction, there is Table 4 shown below. The table presents an OLS regression model for math score differences that now includes interaction terms between income groups and the general knowledge score difference. The model's R-squared value is slightly higher at 0.030 compared to the previous model without interactions, meaning it explains 3% of the variance in math score differences. The coefficients for income groups 2 and 3 are both significant. The interaction terms, however, show a different picture. The interaction between income group 2 and `gk_score_diff` is not statistically significant (p-value of 0.221), nor is the interaction between income group 3 and `gk_score_diff` (p-value of 0.197). This suggests that the effect of general knowledge score difference on math score differences does not vary significantly across income groups 2 and 3.

OLS Regression Results

Dep. Variable:	math_score_diff	R-squared:	0.030
Model:	OLS	Adj. R-squared:	0.029
Method:	Least Squares	F-statistic:	70.26
Date:	Sun, 24 Mar 2024	Prob (F-statistic):	1.21e-72
Time:	23:40:06	Log-Likelihood:	-37364.
No. Observations:	11490	AIC:	7.474e+04
Df Residuals:	11486	BIC:	7.478e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.4862	0.147	57.674	0.000	8.198	8.775
C(incomegroup)[T.2]	0.9915	0.226	4.395	0.000	0.549	1.434
C(incomegroup)[T.3]	1.4018	0.231	6.060	0.000	0.948	1.855
gk_score_diff	0.2648	0.023	11.766	0.000	0.221	0.309
C(incomegroup)[T.2]:gk_score_diff	-0.0419	0.034	-1.225	0.221	-0.109	0.025
C(incomegroup)[T.3]:gk_score_diff	-0.0457	0.035	-1.290	0.197	-0.115	0.024

Omnibus:	801.876	Durbin-Watson:	1.812
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1078.168
Skew:	0.618	Prob(JB):	7.57e-235
Kurtosis:	3.851	Cond. No.	37.1

(Table 4)

6. Insights and Conclusion

The entire analysis has encompassed a comprehensive exploration of how students' reading and math score differences can be explained by their income group classification and general knowledge score differences. Through the application of OLS regression and ANCOVA, with and without interaction terms, I've gleaned several insights:

1. Significant Main Effects: Both income group and general knowledge score differences have significant main effects on reading and math score differences. Students from higher income groups tend to have larger score improvements.

2. Non-significant Interactions: The interaction terms between income group and general knowledge score differences were generally found to be non-significant. This suggests that the relationship between general knowledge score differences and score improvements does not differ markedly across income groups.

In conclusion, while income group status and general knowledge gains are important factors, their predictive power for score improvements is limited. The interaction between these variables does not significantly alter the outcome, indicating that the effects of income group and knowledge gains are additive rather than multiplicative.