Bridging the Gap: ANCOVA Study on Income and Academic Performance

INF2178 - Technical Assignment 3

Fatima Ashfaq - 1010784732

Professor Shion Guha

March 23, 2023

Link to ipynb code file:
https://colab.research.google.com/drive/13B_gvVuYZtaxJPMNlzd_iE06Bqss4kIP?usp=sharing

1. Introduction

This report presents an analysis of kindergarten students' academic performance specifically in reading and math scores during the Fall and Spring terms. The analysis focuses on examining whether income group levels have a significant impact on students' academic performance, while controlling for the covariate variable of general knowledge scores.

2. Research Questions

- Is the income group level a predictor for kindergarten students' academic performance in Fall and Spring terms?
- Is there a significant difference in kindergarten students' reading and math scores between the Fall and Spring terms across the three different income groups using general knowledge as a baseline variable?

3. Exploratory Data Analysis (EDA)

In any statistical analysis, a crucial first step is to thoroughly explore the dataset to gain insights into its structure, distribution, and potential patterns. In this analysis, both non-graphical and graphical exploratory analysis was conducted to understand the dataset's characteristics fully.

3.1 Non-graphical EDA

The summary descriptive statistics computed for the dependent (math and reading scores), and covariate variables (general knowledge scores) across the different income groups (independent variable) for the Fall and Spring terms, table 3.1.1, revealed notable differences in academic performance in both terms across income groups, indicating potential impacts of income disparities on kindergarten students' educational outcomes.

	Fall Reading Scores	Fall Math Scores	Spring Reading Scores	Spring Math Scores	Fall General Knowledge Scores	Spring General Knowledge Scores
Income Group 1	Mean = 32.79	Mean = 23.92	Mean = 43.67	Mean = 33.88	Mean = 19.95	Mean = 25.07
	Std = 8.09	Std = 7.64	Std = 12.00	Std = 10.73	Std = 6.72	Std = 7.25
	Min = 21.01	Min = 10.51	Min = 22.35	Min = 11.90	Min = 6.98	Min = 8.12
	Max = 118.29	Max = 86.33	Max = 142.49	Max = 105.06	Max = 43.51	Max = 45.58
Income Group 2	Mean = 36.29	Mean = 27.57	Mean = 48.01	Mean = 38.46	Mean = 23.89	Mean = 29.14
	Std = 9.99	Std = 8.54	Std = 13.51	Std = 11.36	Std = 6.87	Std = 6.97
	Min = 22.19	Min = 11.59	Min = 23.93	Min = 13.14	Min = 7.12	Min = 7.86
	Max = 138.51	Max = 83.42	Max = 142.49	Max = 110.33	Max = 46.12	Max = 48.05
Income Group 3	Mean = 39.90	Mean = 31.01	Mean = 52.21	Mean = 42.41	Mean = 26.45	Mean = 31.57
	Std = 12.29	Std = 9.93	Std = 16.45	Std = 12.61	Std = 7.10	Std = 6.93
	Min = 23.01	Min = 10.90	Min = 24.54	Min = 12.70	Min = 7.50	Min = 9.51
	Max = 133.56	Max = 115.65	Max = 156.85	Max = 113.80	Max = 47.69	Max = 48.34

Table 3.1.1 - Descriptive Statistics for All Variables Across Income Groups

3.2 Graphical EDA

To delve deeper into the dataset and extract meaningful insights, non-graphical Exploratory Data Analysis (EDA) techniques were translated into visual representations using boxplots, depicted in Figures 3.2.2 to 3.2.7. Additionally, Figure 3.2.1 highlights unequal sample sizes across income groups, suggesting a potential correlation between family income and early education enrollment. This observation aligns with existing research indicating that higher-income families are more likely to enroll their children in early education programs such as kindergarten (Bainbridge et al., 2005). These disparities persist even after adjusting for a wide range of variables like as race/ethnicity, mother employment, family structure, and parental education. Furthermore, a correlation heatmap was also generated to further understand the relation between the variables, which portrayed a strong correlation between the fall and spring scores of the students' and a weaker correlation between other test scores of the other term.

Further analysis of the correlation patterns among various academic test scores through the use of a correlation heatmap as shown in figure 3.2.8 revealed a notable trend. Across different terms, there was a consistent and strong correlation observed between scores of the same subject. For instance, spring math scores were found to have a higher correlation with fall math scores than with fall reading scores or spring reading scores. This pattern was consistent across all subjects, indicating that students tend to exhibit more consistent performance within a specific academic domain across different terms rather than across different subjects within the same or different terms.

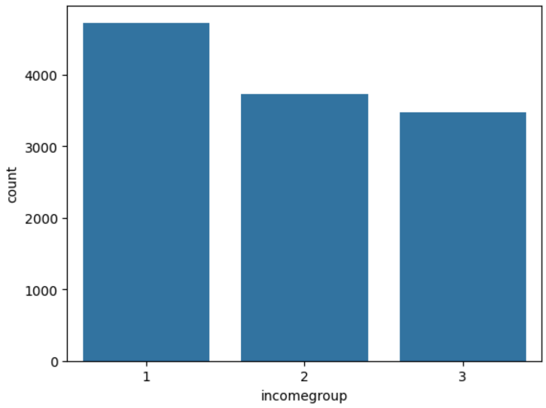

Figure 3.2.1 - Countplot of Income Group Levels

Figure 3.2.2 - Boxplot of Fall Reading Scores per Income Group

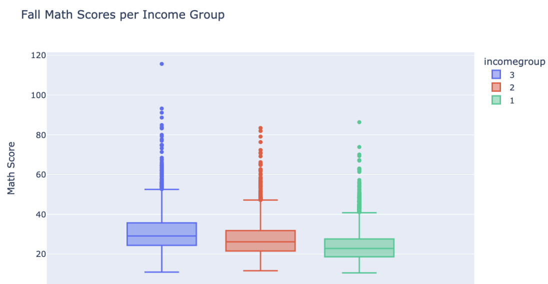

Figure 3.2.3 - Boxplot of Fall Math Scores per Income Group

Figure 3.2.4 - Boxplot of Spring Reading Scores per Income Group

Figure 3.2.5 - Boxplot of Spring Math Scores per Income Group

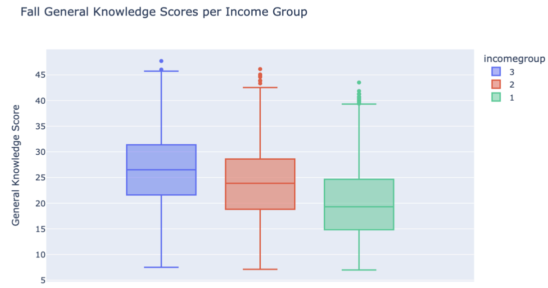

Figure 3.2.6 - Boxplot of Fall General Knowledge Scores (GKS) per Income Group

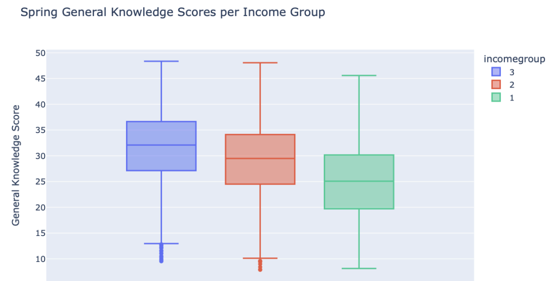

Figure 3.2.7 - Boxplot of Spring General Knowledge Scores (GKS) per Income Group

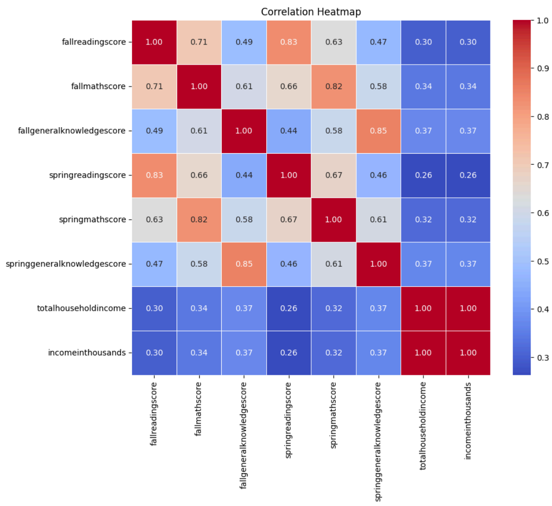

Figure 3.2.8 - Correlation Heatplot

4. One-Way Analysis of Covariance (ANCOVA)

ANCOVA was selected as the analytics framework in this research due to its ability to simultaneously examine the impact of categorical factors (such as income groups) and continuous variables (like general knowledge scores) on the academic performance of kindergarten students. The hypotheses formulated for this analysis are as follows:

- ***Null Hypothesis (H0):*** There is no difference in kindergarten students' reading and math scores between Fall and Spring terms across different income groups after correcting for the covariate variable, the general knowledge scores.
- ***Alternative Hypothesis (H1):*** There is a significant difference in kindergarten students' reading and math scores between Fall and Spring terms across different income groups.

Several one-way ANCOVA models were conducted to analyze how income group levels impact students' academic scores, with general knowledge scores as a covariate (see tables 4.1 to 4.4). Across all models, significant effects of income group levels were observed for both Fall and Spring terms, with p-values below 0.001 ($p < 0.001$). We also employed the StatsModels method for confirmation and additional insights.

In all one-way ANCOVA models, the income group variable significantly affected reading and math scores in both Fall and Spring terms, with very low p-values ($p < 0.001$), indicating notable differences among income groups. However, the relatively small effect size of the income group suggests that it explains only a small portion of the score variances. General knowledge scores in Fall and Spring terms also had a significant effect on reading and math scores, with extremely low p-values ($p < 0.001$), highlighting a substantial relationship. The effect sizes for general knowledge scores were larger than those for income groups, suggesting they explain a significant portion of the score variances. Furthermore, the regression coefficients for general knowledge scores demonstrated a significant increase in reading and math scores with each one-unit rise in the general knowledge score. Additionally, the coefficients for income groups 2 and 3 indicated differences in test scores compared to students in income group 1.

Overall, given that all the coefficients, including the intercepts, and the overall models, as indicated by the F-statistics, are statistically significant ($p < 0.001$), there is significant evidence to reject the null hypothesis. This suggests that there are significant differences in the Fall and Spring reading and math scores among different income groups and that the general knowledge scores is a significant predictor of the terms' academic test scores - both income group and General Knowledge Score (GKS) have significant effect on the test scores.

Source	Degree of Freedom	Sum of Squares	F-Statistic	Uncorrected P-Value	Effect Size		Regression Coefficients
Income Group	2	15987.076426	97.146849	1.411113e-42	0.016026	**Intercept**	20.1862
Fall GKS	1	225245.650280	2737.449236	0.000000e+00	0.186647	**Income Group 2**	1.0168
Residual	11929	981554.407140	NaN	NaN	NaN	**Income Group 3**	3.0031
						Fall GKS	0.6317

Table 4.1 - One-Way ANCOVA Results for Fall Reading Scores

Source	Degree of Freedom	Sum of Squares	F-Statistic	Uncorrected P-Value	Effect Size		Regression Coefficients
Income Group	2	11585.864980	111.928370	6.927877e-49	0.018420	**Intercept**	10.0418
Fall GKS	1	273412.600276	5282.752173	0.000000e+00	0.306927	**Income Group 2**	0.9018
Residual	11929	617393.889005	NaN	NaN	NaN	**Income Group 3**	2.5616
						Fall GKS	0.6960

Table 4.2 - One-Way ANCOVA Results for Fall Math Scores

Source	Degree of Freedom	Sum of Squares	F-Statistic	Uncorrected P-Value	Effect Size		Regression Coefficients
Income Group	2	1.866144e+04	58.532725	5.052222e-26	0.009718	**Intercept**	23.1260
Spring GKS	1	4.000620e+05	2509.636510	0.000000e+00	0.173814	**Income Group 2**	1.0065
Residual	11929	1.901606e+06	NaN	NaN	NaN	**Income Group 3**	3.2179
						Spring GKS	0.8193

Table 4.3 - One-Way ANCOVA Results for Spring Reading Scores

Source	Degree of Freedom	Sum of Squares	F-Statistic	Uncorrected P-Value	Effect Size		Regression Coefficients
Income Group	2	1.166991e+04	64.830308	9.919072e-29	0.010752	**Intercept**	10.8221
Spring GKS	1	5.043363e+05	5603.516818	0.000000e+00	0.319607	**Income Group 2**	0.8339
Residual	11929	1.073652e+06	NaN	NaN	NaN	**Income Group 3**	2.5513
						Spring GKS	0.9199

Table 4.4 - One-Way ANCOVA Results for Spring Math Scores

Tukey's HSD Tests were conducted to further examine the pairwise mean differences in reading and math scores between the income group levels for both Fall and Spring terms, and the results, as summarised in tables 4.5 to 4.8, showed significant differences in the mean scores between income group levels with extremely low p-values ($p < 0.001$), supporting the ANCOVA findings.

Moreover, to gain deeper insights into this phenomenon, additional secondary research was conducted. It was found that a $1000 increase in income could lead to a 2.1% rise in kindergarten students' math and reading scores, with a standard deviation of 3.6% (Reardon & Portilla, 2016). This finding aligns with the broader trend observed in parental behavior: as income levels rise, parents are inclined to allocate more resources, both in terms of finances and time, towards their children's education and overall development (Dahl & Lochner, 2005). Such investments encompass various educational and developmental inputs, including child care services, cognitive enrichment activities, and resources. Research has consistently shown that these investments significantly impact children's developmental outcomes (Becker, 2009).

Group 1	Group 2	Mean Difference	Lower Bound	Upper Bound	Adjusted P-Value	Reject
1	2	3.5057	2.9893	4.0221	0.0	True
1	3	7.1117	6.5851	7.6383	0.0	True
2	3	3.606	3.0501	4.1618	0.0	True

Figure 4.5 - Tukey's HSD Post Hoc Test for Fall Reading Scores

Group 1	Group 2	Mean Difference	Lower Bound	Upper Bound	Adjusted P-Value	Reject
1	2	3.644	3.2003	4.0877	0.0	True
1	3	7.0882	6.6358	7.5407	0.0	True
2	3	3.4443	2.9667	3.9218	0.0	True

Figure 4.6 - Tukey's HSD Post Hoc Test for Fall Math Scores

Group 1	Group 2	Mean Difference	Lower Bound	Upper Bound	Adjusted P-Value	Reject
1	2	4.3444	3.6312	5.0576	0.0	True
1	3	8.5418	7.8145	9.2691	0.0	True
2	3	4.1974	3.4298	4.9651	0.0	True

Figure 4.7 - Tukey's HSD Post Hoc Test for Spring Reading Scores

Group 1	Group 2	Mean Difference	Lower Bound	Upper Bound	Adjusted P-Value	Reject
1	2	4.5816	3.9911	5.1722	0.0	True
1	3	8.5288	7.9267	9.131	0.0	True
2	3	3.9472	3.3116	4.5828	0.0	True

Figure 4.8 - Tukey's HSD Post Hoc Test for Spring Math Scores

Furthermore, the model diagnostics were checked to see if the ANCOVA assumptions were satisfied. The first assumption - linear relation between the covariate independent variables and the continuous numerical dependent variable - was visually inspected and assessed by plotting scatter plots accompanied with regression lines for Fall and Spring reading and math scores against Fall and Spring general knowledge scores as shown in figures 4.9 to 4.10. In addition, a linear (OLS) regression model, tables 4.11 to 4.14, was also utilised to confirm linear relations between the variables which proved a statistically significant result ($p < 0.001$) hence confirming a linear relation between the variables.

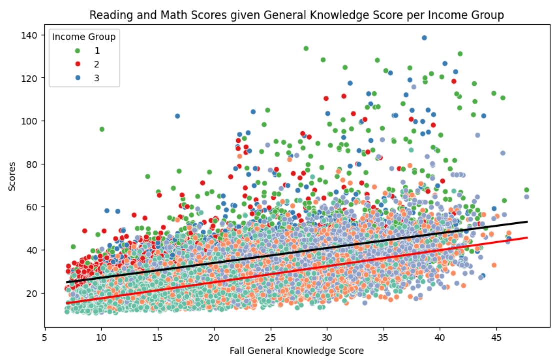

Figure 4.9 - Regression Scatter Plot to Check for Linearity and Parallel Slopes (Fall)

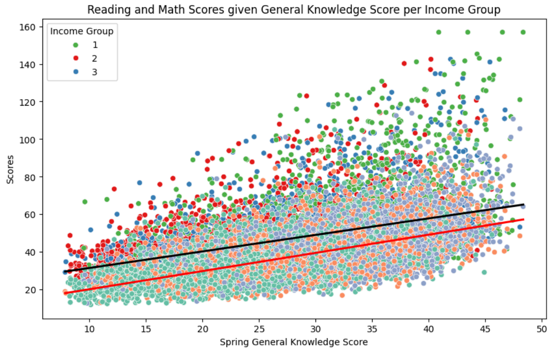

Figure 4.10 - Regression Scatter Plot to Check for Linearity and Parallel Slopes (Spring)

	Regression Coefficient	P-Value
Intercept	20.0229	0.000
Fall General Knowledge Score (GKS)	0.6905	0.000

Table 4.11 - Liner Regression Model Results for Fall Reading Scores

	Regression Coefficient	P-Value
Intercept	9.9096	0.000
Fall General Knowledge Score (GKS)	0.7462	0.000

Table 4.12 - Liner Regression Model Results for Fall Math Scores

	Regression Coefficient	P-Value
Intercept	22.7009	0.000
Spring General Knowledge Score (GKS)	0.8787	0.000

Table 4.13 - Liner Regression Model Results for Spring Reading Scores

	Regression Coefficient	P-Value
Intercept	10.4913	0.000
Spring General Knowledge Score (GKS)	0.9672	0.000

Table 4.14 - Liner Regression Model Results for Spring Math Scores

Homogeneity of regression slopes, the second assumption, was checked by assessing the regression scatter plots generated for the first assumption to confirm that the regression slopes are parallel across different levels of income group. To further investigate this assumption, a two-way ANOVA model was also constructed (tables 4.15 to 4.18), which unlike the visual check, provided a statistically significant outcome with a small p-value ($p < 0.001$) concluding that the effect of general knowledge scores on Reading and Math scores for the Fall and Spring terms vary depending on the income group level, suggesting that the relationship between these variables is not consistent across all income groups.

	Degree of Freedom	Sum of Squares	Mean Squares	F-Statistic	P-Value
Income Group	2	101978.658155	50989.329077	620.585977	4.014876e-257
Fall GKS	1	225245.650280	225245.650280	2741.442070	0.000000e+00
Fall GKS : Income Group	2	1593.933092	796.966546	9.699799	6.178071e-05
Residuals	11927	979960.474049	82.163199	NaN	NaN

Table 4.15 - Two-Way ANOVA of Fall Reading Score to Assess Homogeneity of Regression Slope Assumption

	Degree of Freedom	Sum of Squares	Mean Squares	F-Statistic	P-Value
Income Group	2	101978.658155	50989.329077	620.585977	4.014876e-257
Fall GKS	1	225245.650280	225245.650280	2741.442070	0.000000e+00
Fall GKS : Income Group	2	1593.933092	796.966546	9.699799	6.178071e-05
Residuals	11927	979960.474049	82.163199	NaN	NaN

Table 4.16 - Two-Way ANOVA of Fall Math Score to Assess Homogeneity of Regression Slope Assumption

	Degree of Freedom	Sum of Squares	Mean Squares	F-Statistic	P-Value
Income Group	2	1.475673e+05	73783.672906	463.075992	2.065001e-194
Spring GKS	1	4.000620e+05	400061.995883	2510.841469	0.000000e+00
Spring GKS : Income Group	2	1.231253e+03	615.626573	3.863753	2.101536e-02
Residuals	11927	1.900375e+06	159.333833	NaN	NaN

Table 4.17 - Two-Way ANOVA of Spring Reading Score to Assess Homogeneity of Regression Slope Assumption

	Degree of Freedom	Sum of Squares	Mean Squares	F-Statistic	P-Value
Income Group	2	1.481765e+05	74088.242248	823.953680	0.000000
Spring GKS	1	5.043363e+05	504336.296614	5608.848784	0.000000
Spring GKS : Income Group	2	1.200487e+03	600.243474	6.675456	0.001266
Residuals	11927	1.072452e+06	89.917970	NaN	NaN

Table 4.18 - Two-Way ANOVA of Spring Math Score to Assess Homogeneity of Regression Slope Assumption

The third assumption - normality of residuals - was first checked by plotting a histogram as depicted in figures 4.19 to 4.22. All histograms follow a similar unimodal, right-skewed pattern. This assumption was further checked statistically with the use of a Shapiro-Wilk test which also provided statistically significant results with an extremely low p-value ($p < 0.001$) as summarised in table 4.23, thus pointing to the rejection of residuals following a normal distribution.

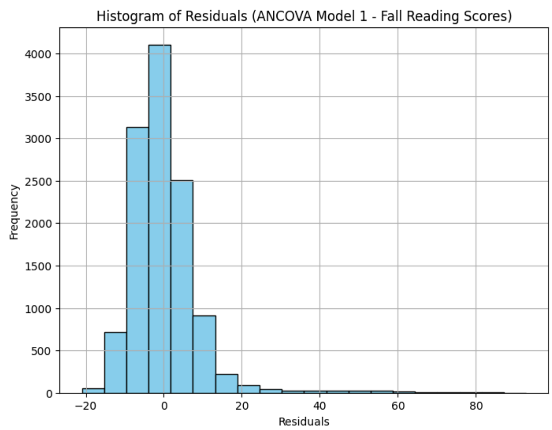

Figure 4.19 - Histogram of Residuals (Fall Reading Scores)

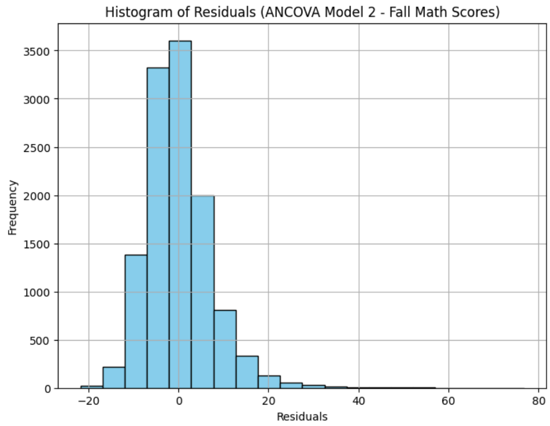

Figure 4.20 - Histogram of Residuals (Fall Math Scores)

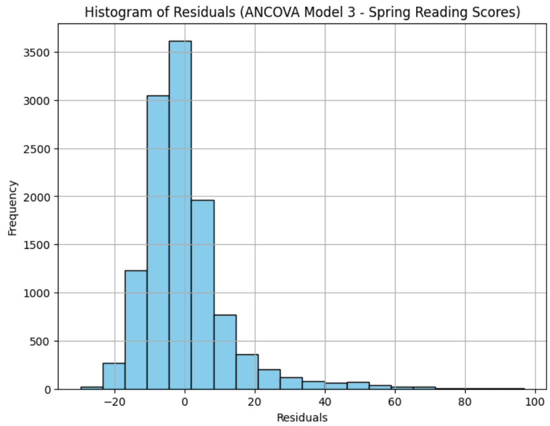

Figure 4.21 - Histogram of Residuals (Spring Reading Scores)

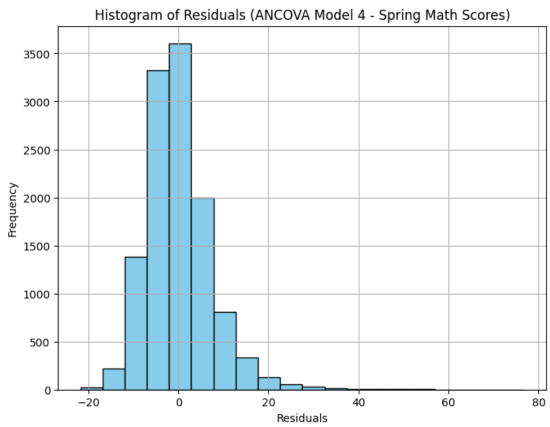

Figure 4.22 - Histogram of Residuals (Spring Math Scores)

	Statistic	P-Value
Fall Reading Scores	0.7827902436256409	0.0
Fall Math Scores	0.936098575592041	0.0
Spring Reading Scores	0.8265702724456787	0.0
Spring Math Scores	0.936098575592041	0.0

Table 4.23 - Shapiro-Wilk Test Results

Finally, a Levene's test, table 4.24, was conducted to check the fourth assumption of homogeneity of variances across the residuals from the different ANCOVA models, which provided a statistically significant result with a p-value of less than 0.001 ($p < 0.001$) hence leading to the rejection of the null hypothesis as the variances have been proven to be unequal.

	Statistic	P-Value
Levene's Test	457.07528101817735	8.014073079026357e-293

Table 4.24 - Levene's Test Results

5. Limitations and Recommendations

The limitations stemming from the non-satisfaction of ANCOVA assumptions in this research underscore the importance of further investigation and refinement in future studies. Addressing the assumption of linearity could involve employing more sophisticated modeling techniques, such as polynomial regression models, to capture potential non-linear relationships between variables more accurately. Moreover, exploring alternative statistical methods or robust regression approaches that are less sensitive to violations of ANCOVA assumptions could provide more robust and reliable results. Furthermore, conducting sensitivity analyses to assess the impact of assumption violations on study outcomes and implementing strategies to mitigate these limitations could enhance the validity and generalizability of the findings. It's important to note that the analysis is based on observational data, and as such, causal relationships cannot be inferred. Future studies could incorporate additional variables and explore longitudinal data to further understand the factors influencing students' academic performance over time and establish more robust causal relationships.

6. References

Bainbridge, J., Meyers, M. K., Tanaka, S., & Waldfogel, J. (2005, August 3). *Who Gets an Early Education? Family Income and the Enrollment of Three‐ to Five‐Year‐Olds from 1968 to 2000**. Social Science Quarterly. https://doi.org/10.1111/j.0038-4941.2005.00326.x

Reardon, S. F., & Portilla, X. A. (2016). *Recent Trends in Income, Racial, and Ethnic School Readiness Gaps at Kindergarten Entry.* AERA Open, 2(3). https://doi.org/10.1177/2332858416657343

Dahl, G. B., & Lochner, L. (2005, April 1). *The Impact of Family Income on Child Achievement*. https://doi.org/10.3386/w11279

Becker G. S. (2009). *Human capital: A theoretical and empirical analysis, with special reference to education*. Chicago, IL: University of Chicago Press.

G. (2022, February 22). *How to Perform an ANCOVA in Python*. GeeksforGeeks. https://www.geeksforgeeks.org/how-to-perform-an-ancova-in-python/

Analysis of covariance using Python