A Class-level Analysis about the Effects of Class Type on Average Math Scores

Team ID: Team 2; Members: Kenneth Lee, Xialin Sang, Rong Duan, Chen Zhang

1. Introduction

1.1 Background

The effects of class sizes on student achievement is an important topic for policymakers in the American K-12 education system. To study the effects of class size on student achievement in the primary grades, the State Department of Education in Tennessee launched a four-year longitudinal class-size randomized study from 1985 to 1989 called The Student/Teacher Achievement Ratio (STAR). Over 7000 students in 79 schools participated in this project. We highlight the features of the experiment process in the study below.

- All participating schools had to agree to the random assignment of teachers and students to different class conditions: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide) [5].
- The assignments of various class types were initiated as the students entered school in kindergarten and continued through third grade [5].
- Each school must provide enough kindergarten students to be assigned to three numerous class types in order to participate in the project STAR [5].
- The student achievement is measured annually via Stanford Achievement Tests (SATs) during the spring term on testing dates specified by the Tennessee state.[5].
- Students moving from a school involved in STAR to another participating school were assigned to the same type of class as they had participated in previously. Also, it is possible that the size of a regular class can be as small as the small class type as students move out of the participating schools[5].
- Besides class size and teacher aides, there were no other experimental changes involved in the study [5].
- There were three schools resigned from the project STAR at the end of kindergarten, so that there were only left with 76 schools in the 1st-grade level [5].

Our primary scientific question of interest is whether there is a treatment effect of assigning various class types to the average math scaled scores in a 1st-grade class level. We implement exploratory data analysis, two-way ANOVA model, model diagnostics, hypothesis testing. In the end, we will discuss any causal statements that could possibly be made based on our analysis and assumptions and the differences between a student-level and a class-level analysis on this STAR dataset.

This work shows that the treatment effect of the class type does exist in a class level for this dataset. We also show that it is possible to make causal statements based on our analysis.

1.2 Exploratory Data Analysis

The original dataset has 11601 observations and 379 attributes. It describes the demographics of the students and teachers, class type assignment, the participating school and class identifiers, and test scores. We first examine if there are any missing values in the data. Then, we will explore the data with teachers as the unit for the 1st-grade students' math scaled scores. We summarize some the findings below:

- ***High level of missing values for teacher identifiers:*** Before we explore the data with teachers as the unit, we have found that there are 4772 observations that are missing teacher ID in the data. Above 40% of the total number of observations do not contain information related to teachers. We drop all the observations that do not have a teacher ID as we can hardly impute this identification information. Then, we are then left with 6829 observations.
- ***Missing class type information in some schools:*** We have also found that there are four schools (ID: 244728, 244796, 244736, 244839) that do not have at least one observation per class type, which contradicts with the experimental design. We then drop the observations from these schools to ensure we have at least one observation per class type in each school.
- ***Small class types have higher average math scores:*** After we drop the observations that do not contain math scaled scores in a student level, we then take the average of the math scaled scores based on the remaining 6334 observations. We are then left with 325 observations with teachers as the unit. The distribution of the averaged math scores by class types is shown by `Appendix II. figure 2` . We can see that the averaged math scores are higher in general.

2. Analysis Plan

2.1 Two-way ANOVA Model

To see whether there is a treatment effect of the class type assignment in a class level, we will use a two-way ANOVA model. Our two-way ANOVA model is an additive model as specified below:

$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ijk}$ where $i = 1, \ldots, 3$, $j = 1, \ldots, 72$ and $k = 1, \ldots, n_{ij}$

Explanation of the notation

- Y_{ijk} denotes the average math scaled score of the ith class type and the jth school for the kth teacher.
- $\mu_{..}$ denotes the overall average of math scaled scores in the population across class types and schools that we try to estimate and this is an unknown parameter.
- α_i denotes the main effect of the ith class type.
- β_j denotes the main effect of the jth school.
- ϵ_{ijk} denotes the random error in the ith class type, jth school for the kth teacher. This is an unobserved random variable.
- The index i represents the levels of class type: small ($i = 1$), regular ($i = 2$), regular with aide ($i = 3$).
- The index j represents the levels of the school indicator. Since we have only 72 schools in the data, we have $j = 1, \ldots, 72$.
- The index n_{ij} denotes the total number of teachers in the jth school corresponding to the ith class type.

The assumptions of the two-way ANOVA model

- The random errors are assumed to be identically and independently distributed from a normal distribution with mean 0 and variance σ^2.
- The outcomes are independent normal random variables with a common variance and with means equal to the overall average of math scaled scores across class types and schools.
- The interaction effects are absent.

Justification for the model choice*

We do not include the interaction term in our two-way ANOVA model because the interaction term is not of our primary interest. Another reason is that we are also concerned about the model complexity. Given that there is a limited number of observations within each combination of the class type and the school indicator via our exploratory data analysis, we may not have enough information to estimate the parameters. Also, we introduce a blocking factor β_j into our model for

controlling the source of varability; that is, we want to eliminate the effects due to variations among different schools while we are trying to determine the effects due to differences among class type assignments. Thus, this design of the experiment can give a greater accuracy of the model estimates [7].

2.2 Model Diagnostics

Next, our model diagnostics will confirm whether the assumptions of our two-way ANOVA model hold such that this model is appropriate for the problem setting.

- For the normality assumption, we will use a normal Q-Q plot and conduct a Shapiro–Wilk test to see if the residuals follow the normal distribution.
- For equal variance assumption, we will use a residual vs. fitted value plot to see if the residuals (the differences between the actual test scores and predicted test scores) have mean zero and equal variance.
- The independence assumption will not be tested. Since the design of this ANOVA model is randomized block design, not only students are randomly assigned, teachers within each school are also randomly assigned to a certain class type. As a result, independence is satisfied.

Furthermore, we would also like to confirm our assumption about absent or ignorable interaction effect between two factors by running the Tukey's test for additivity.

2.3 Hypothesis Testing

As our primary interest is to determine whether there exists a treatment effect of the class types on the math scaled scores in class-level, we plan to use the F-test to test whether there is any main effect across class types. For our hypothesis testing, the significant level is set as 0.05. Our null hypothesis for the F-test is that $H_0 : \alpha_i = 0$, no main effects across class types, with the alternative hypothesis H_a : not all a_i are zero, main effects exist based on some class types. Upon the rejection of the null hypothesis, we will then investigate the nature of the differences among the averaged test scores of the class types.

After knowing the overall F-test is significant for testing the main effect of class types, we may proceed to find out more specific information about where the difference should be accounted by comparing the averages of mean test scores across class types in 1st grade. Subsequently, we will use a multiple comparison analysis to determine where the differences among the averages of the mean scores on class types occur. Since the dataset does not have the same number of observations for each class type, it is suggested that we should use Tukey's procedure as it can give us a more precise estimation of the difference between the averages of the mean test scores from two different class types based on a narrower confidence interval [6].

3. Result

3.1 Two-way ANOVA model

`Table 1` shows a summary of our ANOVA model. The sum of squares of schools shows the variability among the averaged test scores across schools, which is 6.2. The more similar the average of the mean test scores of schools, the smaller this sum of squares tend to be. Across class types, we also see that the variation of the outcomes around their respective mean of the averaged test scores based on each class type is higher, which is about 22.6. The smaller this value is, the smaller error variance we have. Similar to the usage of F-value, the p-value helps us understand whether there is a difference among the main effects of each class type or the difference happens due to chance. We will leave the detailed discussion of hypothesis testing in section `3.3` , which also utilizes the information from `Table 1` .

Table 1: ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	p-value
Class Type	2	6559	6559	22.6	**3.35e-06**
School	71	128366	1808	6.2	**< 2e-16**
Error/Residual	251	73131	290.2		

3.2 Model Diagnostics

Equal Variance Assumption: According to `Appendix II. figure 3`, the residuals are scattered in between positive 50 (roughly) and negative 50 points. Visually, there seems to be more residuals lying in the positive extreme than negative extreme. Other patterns are hard to detect from this graph. Therefore, we conducted Levene's Test for Homogeneity of Variance. From the test output, we can see that the p-value is less than 0.05. It indicates there is strong evidence suggesting that the variance across groups is different. This result makes sense because the residuals are unique, rather than the average of a group. In the appendix II., `figure 4` and `7` in appendix plot the residual against each factor. From these plots, we can examine the equal variance assumption in terms of the class types and school id. The variance of residual is not constant in terms of both factors and seems to be affected by the factor level.

Normality Assumption: In `Appendix II. figure 3`, the QQ plot shows a heavy tail on both ends, and the normal distribution assumption is questionable. To further investigate this matter, we conducted the Shapiro-Wilk normality test. Since the p-value is less than the chosen significance level 0.05, we should reject the null hypothesis and conclude that it is unlikely for the data to come from a normally distributed population.

In search for a remedy to the non-normality issue, we considered taking box-cox transformation of the response variable (math score). However, the transformed response variable still rejects the null hypothesis at the 0.05 significance level. Taking into account the fact that box-cox transformation undermines the interpretability of the data, we decided not to transform the response variable.

Interaction Effects: Lastly, we further conducted Tukey's test for additivity for the ignorable interaction assumption since it is more appropriate for a randomized block design. The final test produced a more satisfactory result. With p-value equals to 0.42, we can not reject the null hypothesis at the significance level 0.05. It is more likely that there is no interaction between the two factors in the ANOVA model.

3.3 Hypothesis Testing

To conduct a F-test for the main effects of class types, the first step is to state the null hypothesis H_0: $\alpha_1 = \alpha_2 = \alpha_3$, where each α_i represent the average of the mean test scores within each class type. The alternative hypothesis H_α: not all α_i's are equal. From `table 1`, we see that the pvalue is less than 0.05. Therefore, we reject the null hypothesis at the significance level 0.05 as the p-value suggests that the equality of the main effects of the class types happen less than 5% of the time. Therefore, we conclude that there is a main effect the class types.

Then, we use Tukey's procedure to find out more information where the difference may come from. In the `Appendix II. Figure 8`, it gives us an intuition of confidence intervals for the difference in the means for all pairs of class types, and all pairs of school indicators. For example, the confidence interval (the interval at the bottom of the `Appendix II. figure 8`) that compares regular classes with regular classes with aide. As the 95% confidence interval includes zero, there is no statistically significant difference between the treatments. Then, we observe from the other two confidence intervals, again in `Appendix II. figure 8`, that the main effect can be attributed to the differences between small class types and other class types.

4. Discussion

4.1 Possibility of making any causal statements

Now we would like to make a causal statement based on the above analysis. If the following assumptions are examined to be qualified in the study, then we could make a causal inference based on the analysis. The assumptions are summarized below.

- **Stable unit treatment value assumption (SUTVA):** SUTVA has two implications: 1) No interference. That means the class type assignment of one teacher does not affect the potential outcomes of others. A class-level average score in math is not dependent on whether other teachers are assigned to a certain class type. For example, if a teacher in the regular class is discouraged because of the assignment of class types, they would not work as harder as usual, which could interfere with the experiment outcome. 2) The treatments are consistent - there is no different version of each treatment level [1]. In STAR, this assumption means all teachers in the same class type should use

a similar way to teach, which makes sure the treatment is stable and the teaching quality is not accounted for the cause. This assumption excludes the unstable treatment cases for causal inference. However, it seems very difficult to implement the same teaching method in practice.

- **Randomization:** With the double randomization of the students and teachers, we can fairly assume that the teachers and students are independent with the 3 class types. With a large sample size, the variation of performance students and teachers would not like to interfere with our outcome. Also, with the randomization of student assignment, we could ignore the potential influences of other factors (like gender, sex, etc.).
- **Exchangeability:** Randomization experiment is expected to produce exchangeability [2]. Exchangeability means the potential outcome is independent with treatment. In our case, it means that the class type which a teacher assigned to is independent with the average math score in that class. In other words, because each teacher and student are distinctive individuals, the class type assignment does not dictate a different performance in the score. But potential outcome is not observed outcome. With exchangeability, we can conclude that the differences in scores among the 3 class types would exist in the whole population. Thus, we could use the association relationship to draw some causal inferences based on this assumption.
- **Positivity:** Positivity is the assumption that any individual has a positive probability of receiving all levels of the treatment. In STAR, we just simply check whether every school has all 3 types of classes, and drop the schools which don't have enough class types. This also corresponds to the randomized block design of our analysis.
- **Double-blind:** To satisfy the double-blind assumption, the teachers and students should not have any preconception that there is a treatment effect among different class types. Otherwise, this information would disturb their educational performance. In the same way, if the scientists believed the class types have treatment effects, their analysis would be interfered by this preconception.

As long as the above assumptions hold and our analysis confirms that the main effects of different class types are significant, we can quantify the average causal effects due to class types. However, we recognize that there are many potential aspects of the experiment that can undermine the above assumptions. For example, missing values would invalidate the randomization. If the missing values are not randomly missing but because of some certain purpose, the exchangeability assumptions would not hold, which is difficult for us to draw a causal statement.

4.2 Differences between student-level and class-level analysis

In the project 1, we used one-way ANOVA for our model to test whether there exists a treatment effect of the class type on the math scaled scores based a student-level analysis. The causal statement was made possible assuming the implementation of completely randomized design. In fact, the Stable Unit Treatment Value Assumption (SUTVA) is not plausible in the student-level study, because students are prone to peer influences. Violations of SUTVA complicate the regression and imputation approaches considerably, and we therefore primarily focus on teacher-unit in order to draw some causal inferences [3]. In this project, with teachers as the unit, the observed values are the averages of test scores within each class. It is ensured that the class type assignment of one teacher does not affect the potential outcomes of others.

In addition, we implemented a randomized block design in project 2 by adding the school identity as another factor. This is necessary because different schools would have pre-treatment effects on different levels, which interfere with the scores of each class type. Under the randomized block design setup, subjects within each block are randomly assigned to different treatments. Compared to the completely randomized design in project 1, this design reduces within-class type variability and potential confounding, producing better estimates of the treatment effects [4]. As a part of the two-way ANOVA model, we account for an additional source of variability coming from schools as the blocking factor. The previous within treatment variability is split into two sources: the reduced error variability and block variability. As a result, the treatment variability becomes larger compared to the reduced error variability. Thus, the class types can provide a better explanation to the differences in scores.

Appendix I. Reference

[1] Marin Vlastelica Pogančić, 2019, "Causal vs. Statistical Inference", https://towardsdatascience.com/causal-vs-statistical-inference-3f2c3e617220 (https://towardsdatascience.com/causal-vs-statistical-inference-3f2c3e617220), Max Planck Institute for Intelligent Systems

[2]Miguel A. Hernán, James M. Robins(2018).Causal Inference: What If.1420076167

[3]Guido W. Imbens & Donald B. Rubin, CAUSAL INFERENCE for Statistics, Social, and Biomedical Sciences An Introduction, chapter 9, ISBN 978-0-521-88588-1.

[4]Hanushek, Eric A. “Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects.” Educational Evaluation and Policy Analysis 21.2 (1999): 143-163.

[5] C.M. Achilles; Helen Pate Bain; Fred Bellott; Jayne Boyd-Zaharias; Jeremy Finn; John Folger; John Johnston; Elizabeth Word, 2008, “Tennessee’s Student Teacher Achievement Ratio (STAR) project”, https://doi.org/10.7910/DVN/SIWH9F (https://doi.org/10.7910/DVN/SIWH9F), Harvard Dataverse, V1, UNF:3:Ji2Q+9HCCZAbw3csOdMNdA== [fileUNF]

[6] Kutner, Michael H., et al. Applied linear statistical models. Vol. 5. New York: McGraw-Hill Irwin, 2005.

[7] (2008) Randomized Block Design. In: The Concise Encyclopedia of Statistics. Springer, New York, NY

Appendix II. Figures

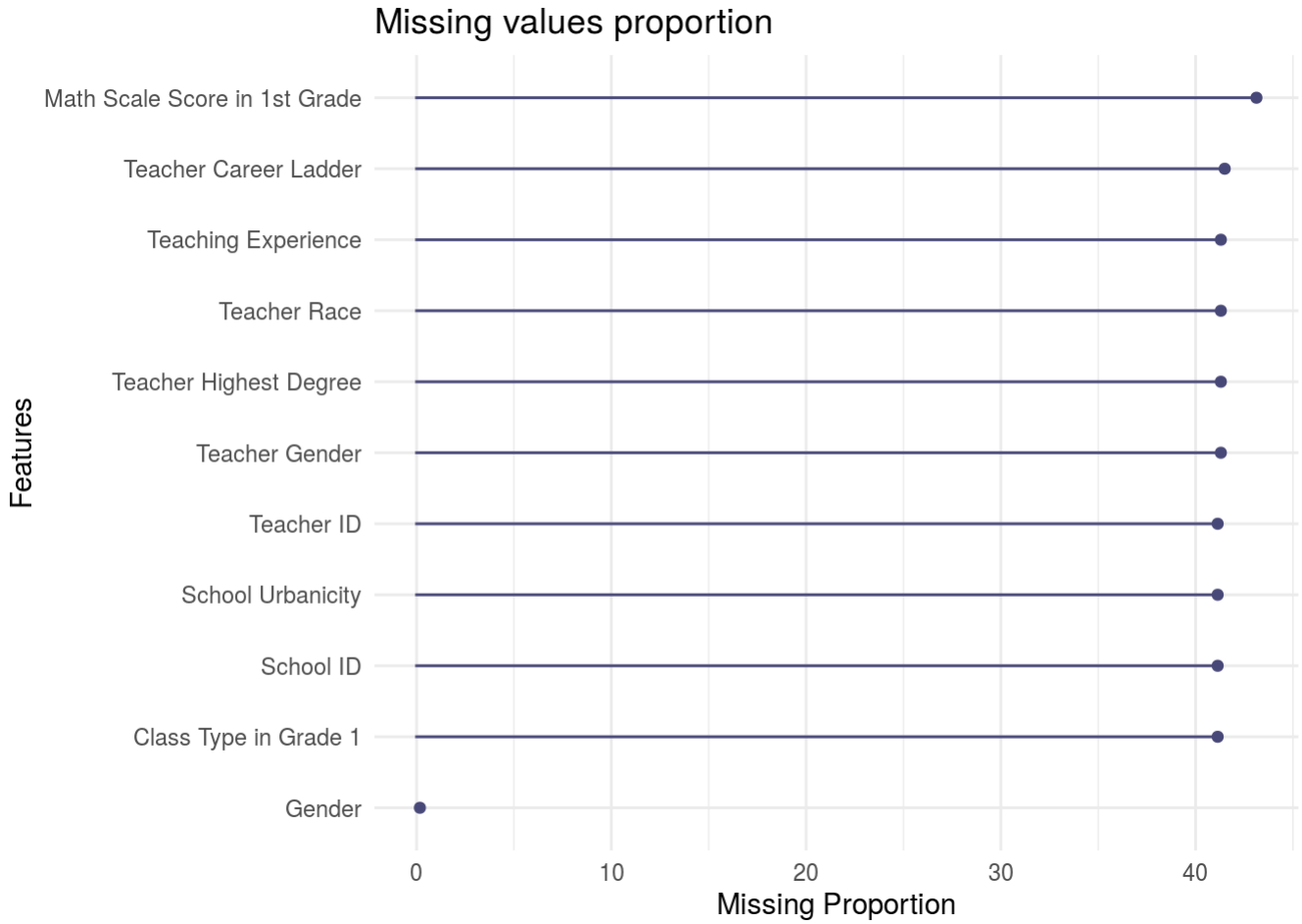

Figure 1: *Missing Proportion Plot*: It shows the variables that have missing values in descending order. The x-axis shows the percentage of the missing values based on the overall number of observations.

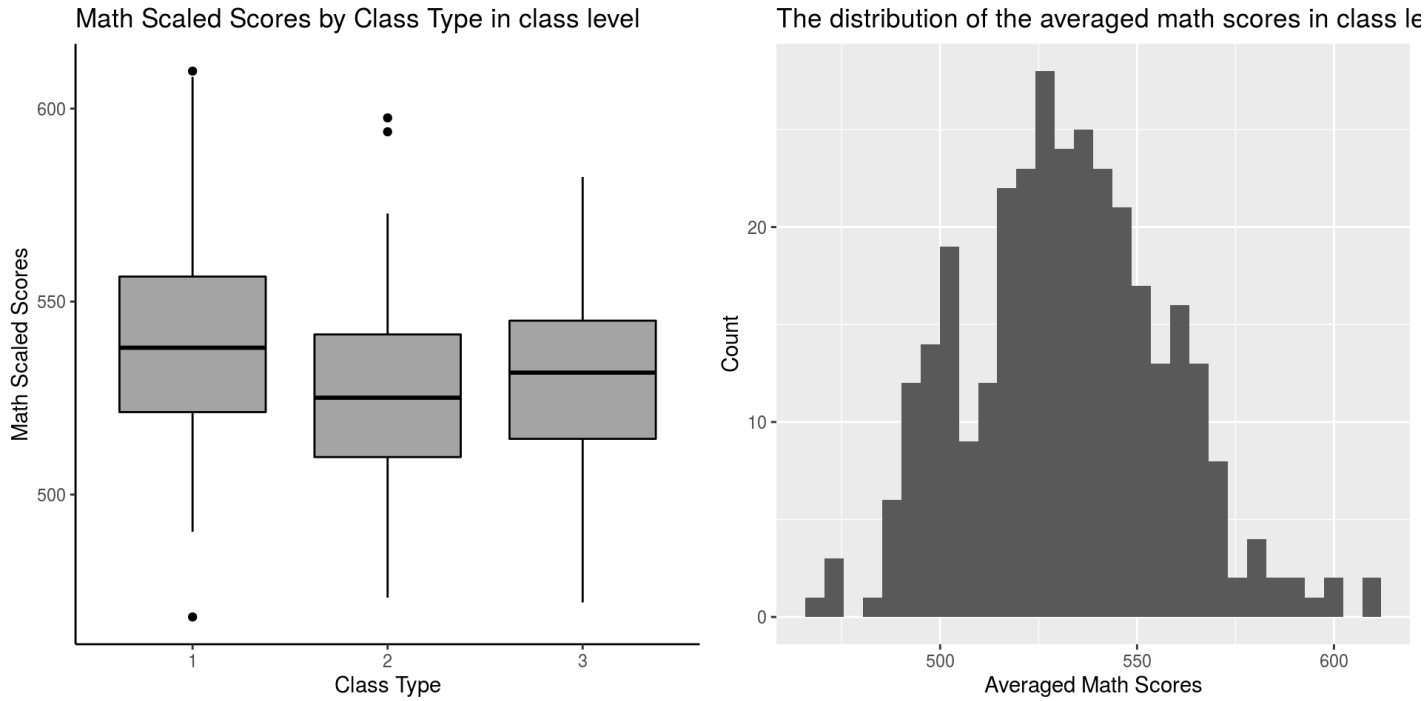

Figure 2: *Distribution plots.* Left panel: The distribution of average math scaled scores by class type. The class type is described on the x-axis (1: small class; 2: regular class; 3: regular class with aide). Both the highest averaged math score and the lowest math score belong to the small class type. Right panel: The distribution of the averaged math scaled scores. The distribution of the averaged math scores seem to be roughly normal.

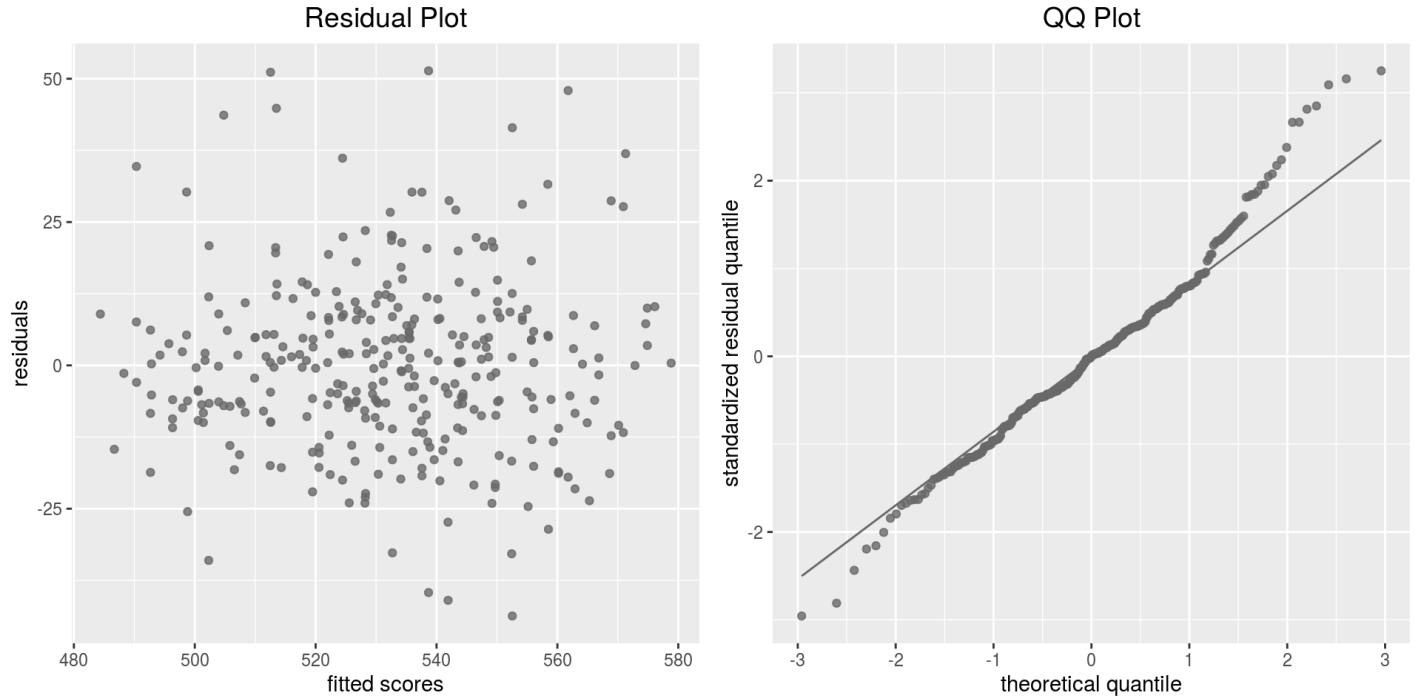

Figure 3: *Model diagnostic plots.* Left panel: residual versus fitted values. Right panel: QQ plot with residuals.

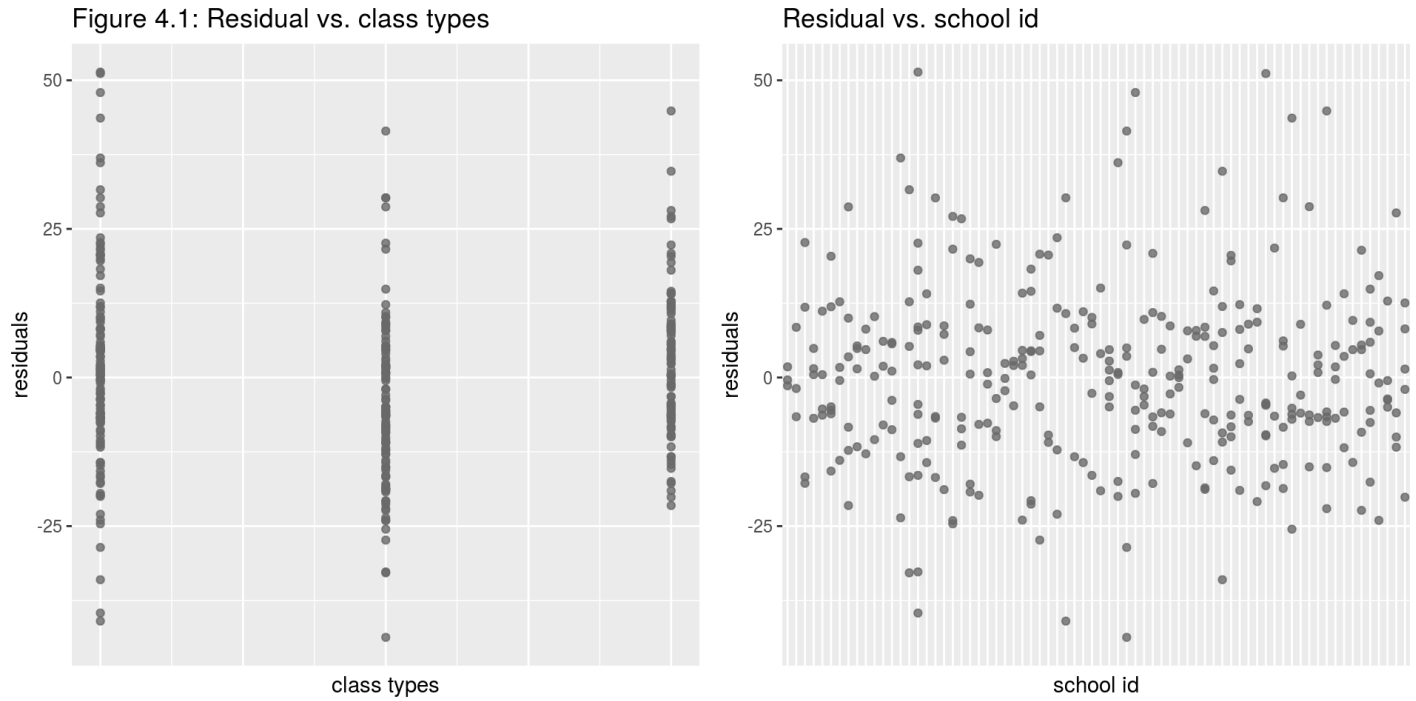

Figure 4: *More residual plots.* Left panel: esidual vs. class types Right panel: Residual vs. school id.

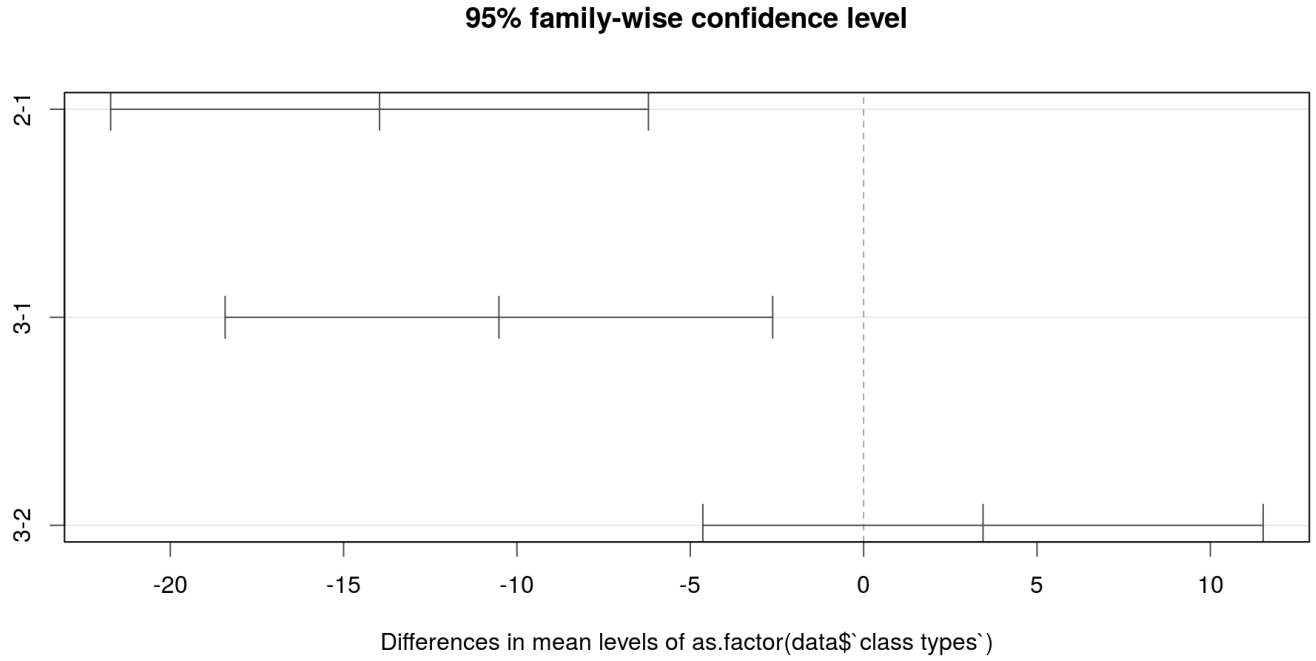

Figure 5: *The Confidence Interval of Multiple Comparisons of Means.* The confidence interval of Multiple Comparisons of the averages of the class types (1:small class, 2: regular class, 3: regular class + aide).

Github Repository

https://github.com/kenneth-lee-ch/STA-207/tree/master/project2 (https://github.com/kenneth-lee-ch/STA-207/tree/master/project2)