Analyze the Effect of Class Type on First Grade Math Scores Using Two-way ANOVA

1 Introduction

This study is based on the Project Student-Teacher Achievement Ratio (STAR) public access data set, assessing the effect of class size on the performance of teachers. The full data set contains 11,601 observations on 379 variables. The Project STAR data set contains data on test scores, treatment groups, and student and teacher characteristics over the four years of the experiment, from the academic year 1985–1986 to the academic year 1988–1989. All students were randomly assigned to one of three class types, including small class, regular class, and regular-with-aide class, and all teachers and students were also randomly assigned to the classes. The questions we were interested in are:

- Whether there is an association between class types and teachers' teaching quality
- Whether we could make causal inference between class types and teachers' teaching quality

To study these problems, we first defined a measure of teachers' teaching quality. The measure we chose is the median math scores of all students taught by each teacher. Then, we analyzed the data by using two-way ANOVA. After the model assumptions justified, Tukey's test was applied and it showed that the differences between small class and regular class, small class and regular with aid calss were significant; but the difference between and regular calss and regular with aid calss was not significant. By further applying potential outcomes framewrok and Fish's Exact P-value to do causal inference, we draw the conclusion that class size does have an effect on first-grade teachers' teaching quality.

2 Descriptive Analysis

2.1 Data Preprocessing

As class types, teacher ID, school ID, and first-grade math scaled scores are the key variables we are interested in, we extract them from the full dataset. Other variables concerning teachers' characteristics have remained also. We dropped the observations if any one of the key variables were missing.

2.2 Data Summary

For the first grade, the dataset has three types of class and 339 different teachers. The math score varies from 404 to 676, with a mean score of 530.7 and a median of 529, as shown in Table 1.

Table 1: summary statistics for the variables of interest

star	teacher ID	math
regular :2507	Min : 11203804	Min. :404.0
small :1867	1st Qu.: 17029508	1st Qu.:500.0
regular+aide:2224	Median : 21252210	Median :529.0
	Mean : 20972685	Mean :530.5
	3rd Qu.: 24475512	3rd Qu.:557.0
	Max. : 26494510	Max. :676.0

2.3 Measuring Teaching Quality

To answer the questions of interest, we first defined measurement for teaching quality. We used the median math scores of all students taught by each teacher as indictor to represent the teaching quality. In project

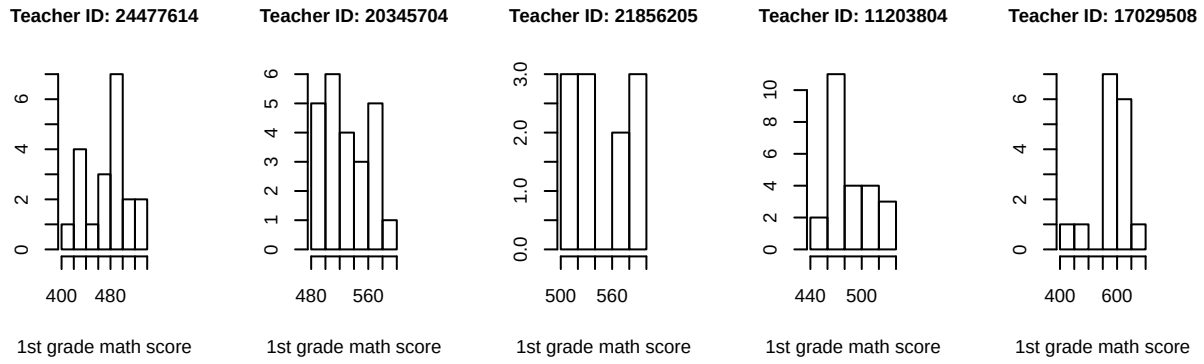

Figure 1: Math Score Distribution of Selected Teachers

star, each teacher only taught one class type, and each student was only taught by one teacher. Thus, this measure is well defined. We chose the median math scores because the plots of the distributions of students' math score under each teacher show the distributions are not normal and have some extreme values.

The violin plot compares the median and means math scores of all students taught by each teacher between different class types. We observed that the median math scores have a lower variance with fewer outliers compares with the mean math scores by different class types. This comparison result gives us the intuition to use the median instead of mean as the measure of teacher level performance.

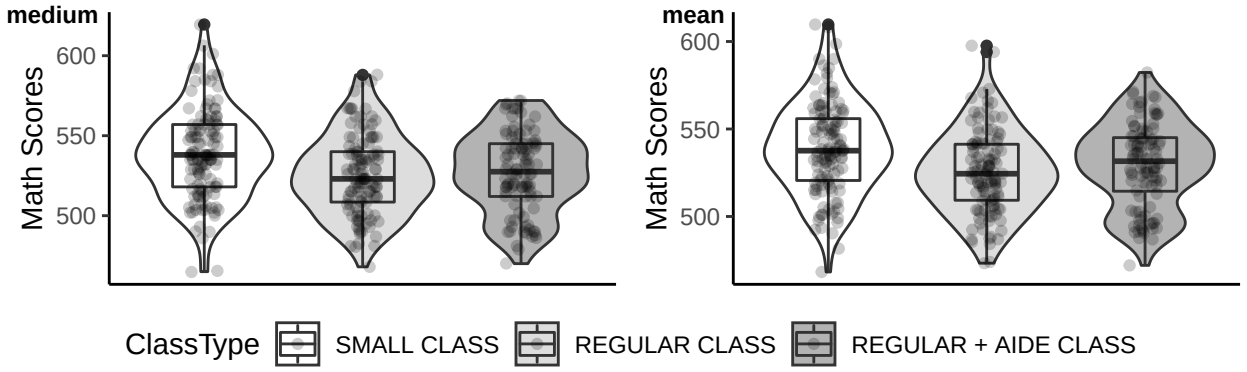

Figure 2: Violin Plot of Teaching Performance by Class Type

Besides, the median scores of all students taught by each teacher by different class types is shown in the boxplot of left panel in Figure 2. The difference in terms of the median score is not significant between regular calss and regular-with-aide class. Nevertheless, the differences between small class and regular class, small class and regular-with-aide class are significant.

3 Main Analysis

In this experiment, nearly all schools had at least one class of each type, and teachers were randomly assigned to classes, so it is a randomized block design. Class types are treatments; schools are blocks. The median math scores of all students taught by each teacher is treated as the response variable because we used it to measure the performance of teacher.

To analyse the question whether there is an association between class types and teachers' teaching quality, the most common analysis method is two-way ANOVA. There are two different kinds of two-way ANOVA model. One assumes that the effects on the outcome of a change in one variable may not depend on the level

of the other variable (additive model); another one assumes that it may depend on the level of the other variable (interaction model).

In this report, we mainly focus on the effects of class types, and there are 76 schools, it is more reasonable to implement the additive model. However, if the interaction terms do have a significant impact on the first-grade math scaled scores, it may cause some problems concerning model diagnostics and hypothesis testings. Thus, we will also test whether interaction terms should be included in the model. At the end of this part, we will do model diagnostics and hypothesis testing.

3.1 Two-way ANOVA Model

Interaction model:

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \epsilon_{i,j,k}$$

Additive model:

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \epsilon_{i,j,k}$$

i denotes the index of class types. 1 denotes small class; 2 denotes regular class; 3 denotes regular class with aide.j denotes the index of school. $j = 1, 2 \cdots, 76$.
k denotes the index over experimetnal units in the treatment group (i, j). $k = 1, 2, \cdots, n_{i,j}$.
$Y_{i,j,k}$ denotes the outcome of the kth experimental unit in the treatment group (i, j)
μ denotes the overall mean.
α_i denotes an adjustment for level i of class types. β_j denotes an adjustment for level j of schools.
$\gamma_{i,j}$ denotes an additional adjustment that takes into account both i and j.
$\varepsilon_{i,j,k}$ denotes random errors.

3.2 Model Assumptions

- Independence assumption: error terms are independent with each other. In this experiment, we assume that first-grade math scaled scores of students taught by one teacher will not be affected by other teachers.
- Normality assumption: error terms are normally distributed.
- Equal variance assumption: variances of error terms are all equal. σ^2 denotes variances of error terms.

Thus, error terms are independent and identically distributed random variables and are distributed as $Normal(0, \sigma^2)$.

Since the experiment is a stratified randomized experiment, the independence assumption is reasonable. The normality assumption and equal variances assumption will be tested in the model diagnostics part.

3.3 Fitted Model

Since we are mainly interested in the effects of class types, we only report the fitted value of μ, α_1, α_2, and α_3. $\hat{\mu} = 531.58$, $\hat{\alpha}_1 = 7.40$, $\hat{\alpha}_2 = -6.18$, and $\hat{\alpha}_3 = -2.08$. Other estimators are listed in Appendix.

3.4 Interactions Terms

In this part, we will employ F-test to analyze whether interaction terms should be included in the model.For these test, the null hypothesis is,

$$H_0 : \text{In interaction model, } \gamma_{i,j} = 0, \text{ for } i = 1, 2, 3; j = 1, 2, \cdots, 76,$$

against the alternative hypothesis H_a : In interaction model, interaction terms are not all equal to zero. The test statistics is F ratio:

$$F^* = \frac{\frac{SSE(A) - SSE(I)}{df_A - df_I}}{\frac{SSE(I)}{df_I}}$$

$SSE(A)$ denotes the error sum of squares(SSE) of the intreaction model and SSE(I) denotes SSE of additive model; df_A denotes the degrees of freedom of $SSE(A)$ and df_I denotes the degrees of freedom of $SSE(I)$. At significant level α, under H_0, $F^* \sim F(df_A - df_I, df_I)$. Thus, if $P(F(df_A - df_I, df_I) > F^*) < \alpha$, the null hypothesis is rejected at level α. In the project, $P(F(df_A - df_I, df_I) > F^*) = 0.7056$ and H_0 is rejected at significant level 0.05. Therefore, it is reasonable to use additive model.

3.5 Model Diagnostics

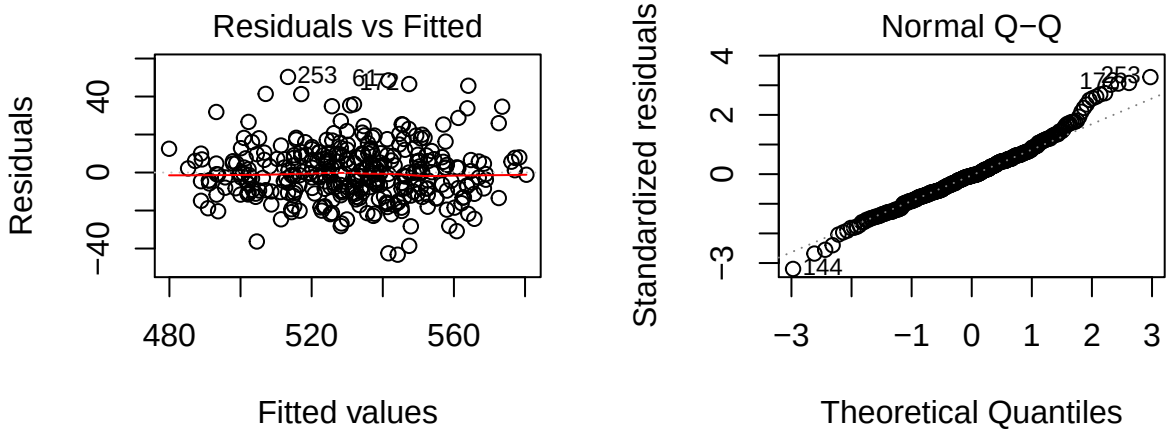

Figure 3: Left panel: Residual versus Fitted Values. Right panel: Q-Q Plot with Residuals

According to residual versus fitted values, there should be no relationship between the size of the residuals and the fitted values. Equal variance assumption holds. According to the Q-Q plot, there is no severe indication of non-normality.

3.6 Hypothesis Testing

3.6.1 F-test for Factor Effects

For a simple explanation, $SSTR$ denotes the sum of squares of variance of class type and $MSTR$ denotes mean of the sum of squares of the variance of class type; Similarly, $SSBL$ and $MSBL$ denotes the sum of squares of variance of school Id and mean of the sum of squares of the variance of school Id respectively. Firstly, We want to explore whether there are main effects for class type and school Id.

Test the class type main effect
We test the null hypothesis.

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$

against the alternative H_a : Not all α_i's equal zero
The test statistics is $F^* = \frac{MSTR}{MSE}$. Under H_0, $F^* \sim F(0.95, 2, 150)$. $F^* = 21.72$, $P_{value} = 1.87 * 10^{-09}$. Thus, at significance level $\alpha = 0.05$, H_0 is rejected. It is likely that class types affect the math scores in first-grade.

Test the school ID main effect

We test the null hypothesis

$$H_0 : \beta_1 = ... = \beta_{76} = 0$$

against the alternative H_a : Not all β_j's equal zero
The test statistics is $F^* = \frac{MSBL}{MSE}$. Under H_0, $F^* \sim F(0.95, 75, 216)$. $F^* = 6.59$; $P_{value} = 1.17 * 10^{-30}$. Thus, at significance level $\alpha = 0.05$, H_0 is rejected, which means it is likely that school Id affects the math scores in first-grade.

3.6.2 Pairwise Comparison

We further construct simultaneous confidence intervals for all pairwise differences and run the simultaneous testing for difference among the means of class types. Tukey's test compares all possible pairs of means simultaneously, which suits our purpose in this project.
The null hypothesis is

$$H_{ii',0} : D_{ii'} = \mu_i - \mu_{i'} = 0$$

against the alternative$H_{ii',a} : D_{ii'} = \mu_i - \mu_{i'} \neq 0$
This null hypothesis could be rejected if 0 is not included in the confidence interval of $D_{ii'}$.

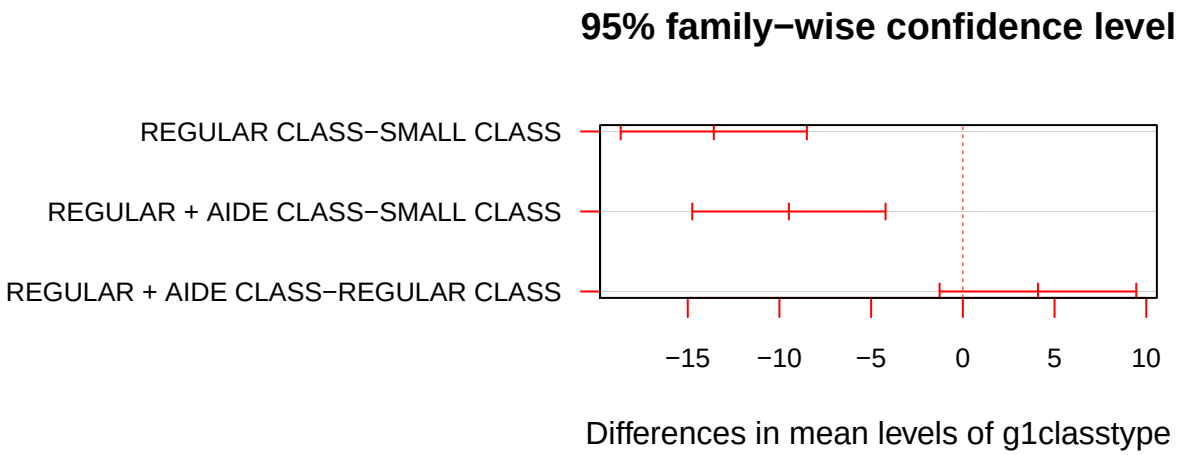

Figure 4: Tuckey's pairwise comparison

As we could see from Figure 4, one of the three confidence intervals contains zero; it's regular-with-aide class compared to the regular class. The other two confidence intervals don't contain zero. Therefore, at significance level 0.05, we could reject the hypothesis and draw the conclusion that the differences between small class and regular class, small class and regular-with-aid class were significant; but the difference between and regular calss and regular-with-aid class was not significant.

4 Causaul Inference

As shown in Figure 4, the confidence interval of regular-with-aide – regular class contains zero, which means their median math score difference is not significant, and the differences are significant for the other two pairs. Since the math score difference is not significant, we combine regular-with-aide class and regular class together as new regular class. We treat regular class as control and small class as treatment to make the causal inference.

4.1 Potential Outcomes Framework

In this project, we treat the experiment as a randomized block design and analyze the impact of school and class types on first-grade math scaled scores. In a randomized block design, it employs blocking to systematically eliminate the effect of a variable on the statistical comparisons among treatments. Randomized block design could better ensure the balance of treatment groups concerning various combinations of prognostic variables. We could apply the potential outcome framework to make a causal inference since the SUTVA holds in this circumstance:

- No interference: in the STAR project, each teacher only taught one class, and one teacher only taught each student. Thus the performance of the teacher is not affected by other teachers. Thus no interference assumption holds.

- Single version of each treatment level: consider the design of the STAR project, class types were defined under the same criteria across all schools, so treatments are stable.
- Ignorability: The design of this experiment is a stratified randomized experiment; teachers and students were radomly assigned. The ignorability assumption holds.

Based on the results of the new model, class types affect the performance of teachers.

4.2 Fisher's exact p-value(FEP)

Another method we introduce is Fisher's exact p-value(FEP). In the FEP framework, the potential outcomes are considered fixed, and the randomness only comes from the assignment mechanism. The sharp null hypothesis of this method is that there is no individual treatment effect. Under the null hypothesis, SUTVA no interference assumption automatically holds, and all potential outcomes are known. In this project, as mentioned above, we combined regular class and regular+aide as a new regular class, set as the control(0), and small class as treatment(1). To be specific, there were 76 schools in total and 339 teachers, among which 124 were for small classes and 215 for a new regular class. The outcomes we focused on were the median math scores of all students taught by each teacher. We have the following Fisher's sharp null hypothesis:

$$H_0 : Y_K(0) = Y_K(1), \ K = 1, 2, \dots 339$$

where $Y_K(0)$ and $Y_K(1)$ being the median math score of all students under the teacher K of new regular and small class respectively.

The statistic we chose in this approach is the weighted-average of 76 within-school average differences between small and regular median math scaled score.

$$T^{obs} = \left| \sum_{m=1}^{M} \frac{N(m)}{N} \left(\bar{Y}_1^{obs}(m) - \bar{Y}_0^{obs}(m) \right) \right|$$

M denote the number of strata, in our case, the number of schools which is 76, $m = 1, 2, \dots M$.
$N(m)$ denotes the number of classes in school m. N denotes the total number of classes, $N = \sum_{m=1}^{M} N(m)$.
$\bar{Y}_1^{obs}m$ and $\bar{Y}_2^{obs}m$ denote the average of observed median math score for small and regular classes in school m respectively.

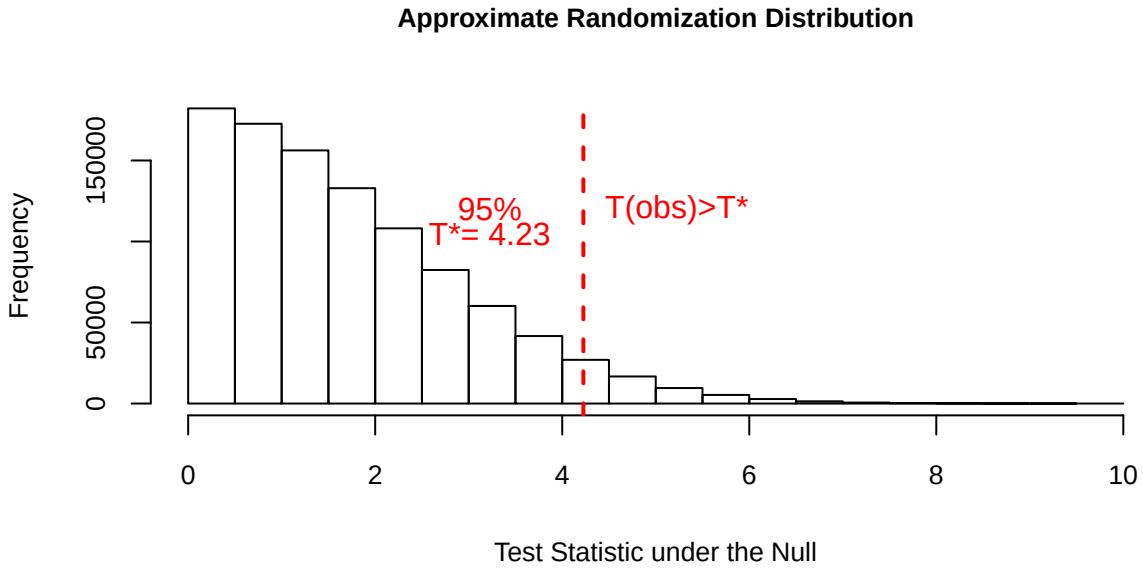

Figure 5: Approximate Randomization Distribution

The realized value of the test statistics is 12.66.
By exhausting all possible assignments of teachers, the distribution of T arises. The exact p-value is the

proportion of test statistics in this randomization distribution that are as extreme as T^{obs}.
However, in our case, the number of possible assignments is very large. Enumerating every possible assignment is computationally challenging, thus, we have to use numerical methods to approximate the p-value for the FEP approach. With 1000000 simulate random assignments, we have distribution in Figure 5. The approximate p-value is 0, which is the probability of finding the value of observed statistics under randomization distribution above, thereby suggesting that teachers with small classes had different performance than teachers with other types of classes.

5 Appendix

5.1 Appendix 1

The scatter plot shows the scatter plots for all the variables. The class types assigned to teachers are even. The second type, a.k.a. the small type class students are more unlikely to obtain lower math scores. In general, only several students, of course, and their teachers obtain scores larger than 600. The scores mainly lie in the range from 420 to 550.

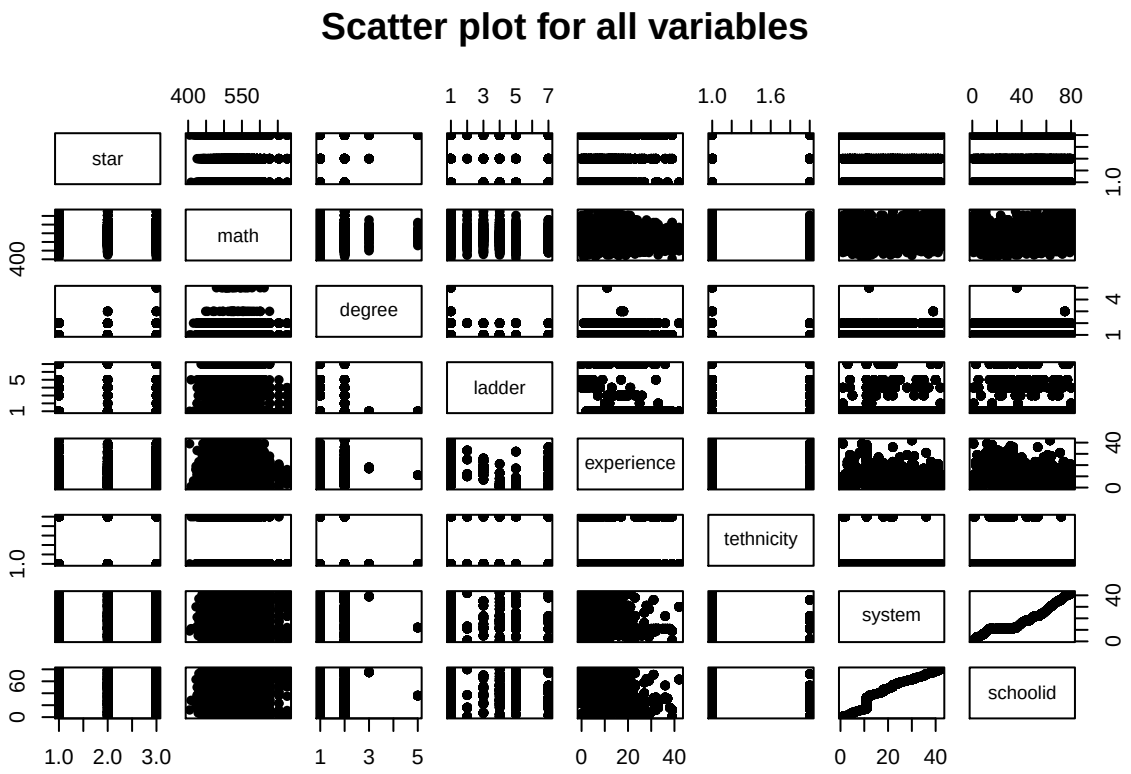

Figure 6: Scatter plot for all variables

5.2 Appendix 2: Fitted two-way ANOVA model

```
##
## Call:
## lm(formula = g1tmathss ~ g1classtype + g1schid, data = data_anova)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.253  -9.385  -0.803   7.896  50.319
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     502.5007     9.6879  51.869  < 2e-16 ***
## g1classtypeREGULAR CLASS        -13.3699     2.1913  -6.101 3.78e-09 ***
## g1classtypeREGULAR + AIDE CLASS -11.3967     2.2774  -5.004 1.03e-06 ***
## g1schid123056                    36.0914    13.5824   2.657 0.008365 **
## g1schid128076                    31.7177    12.7091   2.496 0.013190 *
## g1schid128079                    21.1921    12.7091   1.667 0.096620 .
## g1schid130085                    61.1310    12.7097   4.810 2.56e-06 ***
```

```
## g1schid159171   50.1415   11.7627   4.263 2.82e-05 ***
## g1schid161176   31.0192   12.7097   2.441 0.015329 *
## g1schid161183   74.6535   11.7627   6.347 9.68e-10 ***
## g1schid162184   47.8227   12.7091   3.763 0.000207 ***
## g1schid164198   47.1742   13.5824   3.473 0.000602 ***
## g1schid165199   75.8890   13.5824   5.587 5.79e-08 ***
## g1schid166203   17.1265   13.5824   1.261 0.208459
## g1schid168211   46.1819   12.1513   3.801 0.000180 ***
## g1schid168214   71.0581   13.5824   5.232 3.45e-07 ***
## g1schid169219   58.7557   12.1515   4.835 2.27e-06 ***
## g1schid169229   39.0727   10.7394   3.638 0.000331 ***
## g1schid169231   30.9424   12.1515   2.546 0.011460 *
## g1schid169280   42.7722   12.7091   3.365 0.000879 ***
## g1schid170295   73.6803   12.7097   5.797 1.94e-08 ***
## g1schid173312   56.2036   12.7093   4.422 1.43e-05 ***
## g1schid176329   43.5139   12.7091   3.424 0.000717 ***
## g1schid180344   43.3460   11.7627   3.685 0.000278 ***
## g1schid189378   33.1502   12.7097   2.608 0.009625 **
## g1schid189382   47.2680   12.7093   3.719 0.000245 ***
## g1schid189396   23.7389   12.7091   1.868 0.062902 .
## g1schid191411    9.6802   13.5824   0.713 0.476669
## g1schid193422   34.1487   13.5824   2.514 0.012533 *
## g1schid193423   25.9019   12.1515   2.132 0.033975 *
## g1schid201449   55.1484   11.4807   4.804 2.63e-06 ***
## g1schid203452   48.2226   12.1515   3.968 9.35e-05 ***
## g1schid203457   49.2484   13.5824   3.626 0.000346 ***
## g1schid205488   27.4200   12.7091   2.158 0.031879 *
## g1schid205490   41.6935   13.5824   3.070 0.002369 **
## g1schid205491   50.2866   13.5824   3.702 0.000261 ***
## g1schid205492   26.3177   13.5824   1.938 0.053747 .
## g1schid208501   38.8152   12.7091   3.054 0.002491 **
## g1schid208503   34.1437   13.5824   2.514 0.012546 *
## g1schid209510   35.3795   11.7627   3.008 0.002889 **
## g1schid212522   23.1986   12.1513   1.909 0.057339 .
## g1schid215533   58.2960   11.7627   4.956 1.30e-06 ***
## g1schid216537   61.5502   11.7627   5.233 3.44e-07 ***
## g1schid218562   54.2052   12.7091   4.265 2.80e-05 ***
## g1schid221571   14.0999   11.7627   1.199 0.231732
## g1schid221574   29.0660   12.7091   2.287 0.022996 *
## g1schid225585   30.4132   12.7091   2.393 0.017418 *
## g1schid228606   77.8833   12.7097   6.128 3.27e-09 ***
## g1schid230612   59.9567   13.5824   4.414 1.48e-05 ***
## g1schid231616   40.8215   13.5824   3.005 0.002910 **
## g1schid234628   65.9387   11.7627   5.606 5.27e-08 ***
## g1schid244697   17.5500   11.7627   1.492 0.136907
## g1schid244708    2.0793   11.7627   0.177 0.859824
## g1schid244723   13.1557   11.7627   1.118 0.264413
## g1schid244727   36.6942   12.1515   3.020 0.002781 **
## g1schid244728   -9.2304   13.6036  -0.679 0.498041
## g1schid244736   22.2441   13.6046   1.635 0.103246
## g1schid244745   11.0033   12.7093   0.866 0.387412
## g1schid244746   51.9361   13.5824   3.824 0.000164 ***
## g1schid244755   10.8470   11.4853   0.944 0.345826
## g1schid244764   32.3267   13.5824   2.380 0.018029 *
```

```
## g1schid244774                    -0.2481    11.7691  -0.021 0.983199
## g1schid244776                     4.5687    11.7627   0.388 0.698034
## g1schid244780                    -3.8831    13.5824  -0.286 0.775188
## g1schid244796                    11.8446    13.6046   0.871 0.384755
## g1schid244799                    43.1300    12.7093   3.394 0.000797 ***
## g1schid244801                     8.7337    12.7093   0.687 0.492577
## g1schid244806                    25.9719    11.4807   2.262 0.024507 *
## g1schid244831                    12.3135    12.7091   0.969 0.333506
## g1schid244839                    48.5523    12.1696   3.990 8.60e-05 ***
## g1schid252885                    44.8531    12.7093   3.529 0.000493 ***
## g1schid253888                    38.6795    13.5824   2.848 0.004753 **
## g1schid257899                    34.5776    12.1515   2.846 0.004785 **
## g1schid257905                    55.5184    11.4807   4.836 2.27e-06 ***
## g1schid259915                    33.3439    12.7091   2.624 0.009212 **
## g1schid261927                    35.7401    12.1515   2.941 0.003563 **
## g1schid262937                    70.1170    12.7091   5.517 8.30e-08 ***
## g1schid264945                    51.2676    12.1513   4.219 3.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.63 on 261 degrees of freedom
## Multiple R-squared:  0.6727, Adjusted R-squared:  0.5762
## F-statistic: 6.967 on 77 and 261 DF,  p-value: < 2.2e-16
```

5.3 Appendix 3: Tukey's confidence intervals for pairwise comparisons

Table 2: Tukey's confidence intervals for pairwise comparisons

	Regular - Small	Regular + AIDE - Small	Regular + AIDE - Regular
Confidence Interval	(-18.66, -8.50)	(-14.75, -4.21)	(-1.26, 9.45)

6 Reference

1. Tennessee's Student Teacher Achievement Ratio (STAR) project https://doi.org/10.7910/DVN/SIWH9F
2. Causaul Inference for Statistics Social and Biomedical Sciences An Introduction Chapter 9
3. https://www2.stat.duke.edu/courses/Spring14/sta320.01/Class5.pdf

---

Team ID: Course project group 13

Name (responsibilities): Zheng Gu (Background, Descriptive Analysis)

Name (responsibilities): Jieyun Wang (Causal Inference, Polish Report)

Name (responsibilities): Siyao Wang (Model Fitting, Remedial Measures for Nonnormality)

Name (responsibilities): Zhi Zhang (Hypothesis Testing,Model Diagnostics)

Github: https://github.com/jynwang/STA207Project.git
