Assessing Class Type Effects on First Grade Math Scores

Contents

Team ID: 4

Kenneth Broadhead

Koral Buch

Min Kim

Nanhao Chen

1 Introduction

1.1 Background

The Student/Teacher Achievement Ratio (STAR) Project was a four-year study in the late 1980s in Tennessee, which assessed the effect of class size and type on the students' academic performance (math and reading scores) of the Stanford Achievement Test (SAT). The longitudinal study randomly assigned students to one of three class types and tracked their achievement from kindergarten through third grade. Additionally, teachers were randomly assigned to the classes they would teach. The three class types were as follows: Regular class (22 to 25 students per teacher), small class (13 to 17 students per teacher), regular-with-aide class (22 to 25 students with a full-time teacher's aide). The study provides additional information about the students, such as gender and ethnicity, and the teachers, such as a number of years of experience and level of education.

1.2 Objective

In this project, we analyze data from Project STAR, focusing on first graders' SAT math scores and class type. Our primary goal is to establish whether class type has a significant effect on students' SAT math scores. This question is analyzed by fitting a two-way ANOVA model with two factors from the dataset: class type and school ID. A secondary aim is to provide a causal interpretation for the results of our analysis. To this end, our analysis employs teachers as the unit of statistical analysis. Finally, we seek to determine whether this analysis plan yields results that conflicts with the results of our previous investigation (Project 1).

1.3 Statistical Reasoning and Analysis

To investigate whether class type has a significant effect on students' mathematics performance, we fit a two-way ANOVA model to our data, blocking by school ID and using class type as the primary factor of interest. Our response variable will be carefully chosen so that teachers will be considered as the statistical unit of analysis. The choice of a teacher as a statistical unit, how our response variable is then defined, as well as our choice of blocking variable is further explained below.

The Choice of Teachers as an Analysis Unit

Using teachers as a unit, rather than students, is motivated primarily by causal inference concerns and the potential implications of this study for education policy and legislation. With students as the statistical unit, causal inferences concerning class type and any direct causal effect it has on students is difficult discern using the design of this experiment. However, an analysis of the effect of class size on an instructor's teaching ability is facilitated by this experiment. See the discussion of causal inference below for more.

The Response Variable

To perform this analysis using teachers a the statistical unit, we first define a measure of a teacher's performance in math. Our measure of performance is found by taking the median mathematics score for all students under a given teacher. Each teacher was assigned to only one class type, and each student had only one teacher. Thus, this measure of a teacher's performance is well defined. We use the median for this measure for its robustness against potential outliers and the possible skewness of the data distribution; it thus provides a more accurate reflection of a teacher's teaching ability for a given class type.

Blocking by School

In project STAR, students and teachers were randomly assigned to a class type; however, this randomization was done only within each school participating in project STAR. This makes it necessary to block by school (using school ID) in our analysis, for this was part of the inherent design of project STAR.

A specialized analysis plan is required to analyze the results of completely randomized experiments (CRE) with only one observation per cell. Furthermore, within a stratified experiment (or randomized block design) one essentially conducts a CRE within each stratum, or block. Because of this, for the purpose of our analysis, we consider only those schools that have at least two classes for each class size (1,2).

2 Results

2.1 Descriptive Analysis

Figure 1 provides summary statistics of teaching performance grouped by class type in the form of violin plots with inset boxplots. Each plot provides a summary of the distribution of each treatment population (similar to a histogram), from which one can see relative spread, and measures of center. Furthermore, summary statistics in the form of quartiles, minimum and maximum values, and the median for each treatment population are provide by the inset boxplot.

We highlight a few notable observations. First, we note it is easy to see that teachers with a smaller class size were generally performed better than those with larger classes. Furthermore, teachers in regular classes with aide performed slightly better, on average, than teachers without aide with a similar class size. We also note that there is only one outlier and that it is therefore unlikely to pose a problem for the subsequent analysis. These initial findings suggest that a more formal study of the effects of class type on teaching ability in mathematics is warranted.

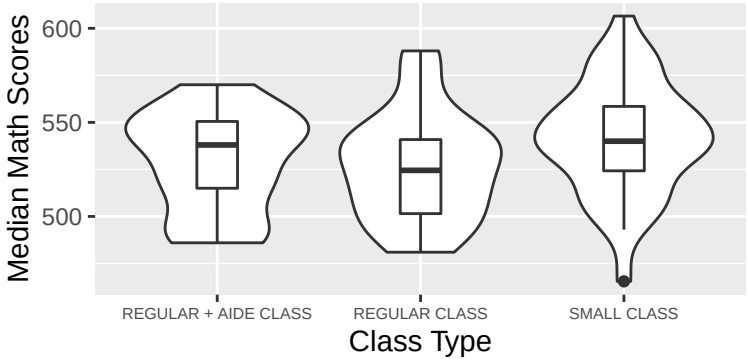

Figure 1: Violin plot and boxplot of teaching performance by class type.

2.2 Two-Way Anova Model

Model Description

(1) $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, $i = 1, 2, 3; j = 1, ..., 16; k = 1, ..., n_{ij}$

where:

Y_{ijk} is the median of 1st grade math score of the kth teacher for the ith Class Type in jth School ID;

$\mu_{..}$ is a constant equal to the means of the total math scores;

α_i are the main effects for the class type at the ith level, which are constants subject to the restriction $\sum \alpha_i = 0$

β_j are the main effects for the School ID at the jth level, which are constants subject to the restriction $\sum \beta_j = 0$

$(\alpha\beta)_{ij}$ are the interaction effect when the Class Type is at the ith level and the School ID is at the jth level. They are constants subject to the restrictions:

$\sum_i (\alpha\beta)_{ij} = 0 \ i = 1, 2, 3 \ \sum_j (\alpha\beta)_{ij} = 0 \ j = 1, ..., 16$

ε_{ijk}'s are random errors, normally distributed with constant variance; n_{ij} is the number of observations for the ith Class Type in the jth School; n_i is the number of level in the second facror (School ID), which is 16.

The number of observations at different levels (see Table 4 in the appendix) differe from each other, indicating an unbalanced design. The regular partitioning of the sum of square treatment regression (SSTR) into the sum of squares of Factor A (SSA), the sum of squares of Factor B (SSB), and the interaction (SSBA) is no longer guaranteed. We therefore fit an ANOVA model using a regression approach. The full model (1) and three reduced models (2, 3, and 4) have been constructed to study the importance of the interaction and the main effects of class type and school ID.

(2) $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ijk}$,
(3) $Y_{ijk} = \mu_{..} + \alpha_i + \varepsilon_{ijk}$,
(4) $Y_{ijk} = \mu_{..} + \beta_j + \varepsilon_{ijk}$,

Though interaction effects are unimportant here, we formally test whether we may exclude them. The two-way ANOVA model is set up by the full model (1) and the reduced model (2) with:

null hypothesis H_0 : all $(\alpha\beta)_{ij} = 0$

alternative hypothesis H_a : not all $(\alpha\beta)_{ij} = 0$.

The P-value of this model is as large as 0.15, leading to the failure of rejecting the null hypothesis at 0.05 significant level. Therefore, we can conclude that there is no interaction between class type and school ID. We therefore fit the following ANOVA model: $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ijk}$

Model Fitting

After fitting the model, we can obtain the following:

Table 1: ANOVA Model Summary

	DF	Sum of Squares	Mean Squares	F-Value	Pr(>F)
Class Type	2	4437.55	2218.77	6.28	2.82e-03
School ID	15	50605.2	3373.68	9.56	8.26e-13
Residuals	87	30717.11	353.07		
Total	104	85759.85			

- Number of regular classes with aide is $n_1 = 34$. Number of regular classes is $n_2 = 36$. Number of small classes is $n_3 = 35$. Total sample size is $n_T = 105$. Table 4 in the Appendix shows the full distribution of sample size of each treatment group.
- The mean of regular classes with aide is $\bar{Y}_{1.} = 2.125$. The mean of regular classes is $\bar{Y}_{2.} = 2.25$. The mean of small classes is $\bar{Y}_{3.} = 35$. Total sample mean is $\bar{Y}_{..} = 532.25$. Table 5 in the Appendix shows the full distribution of sample means of each treatment group.

2.3 Model Diagnostics

We note that model diagnostics confirm that most model assumptions have been fulfilled. As shown in Figure 2, the residuals of all the treatments are distributed around 0, indicating it is unrelated to the response variable and the groups. Additionaly, the Q-Q plot shows that the residuals are normally distributed.

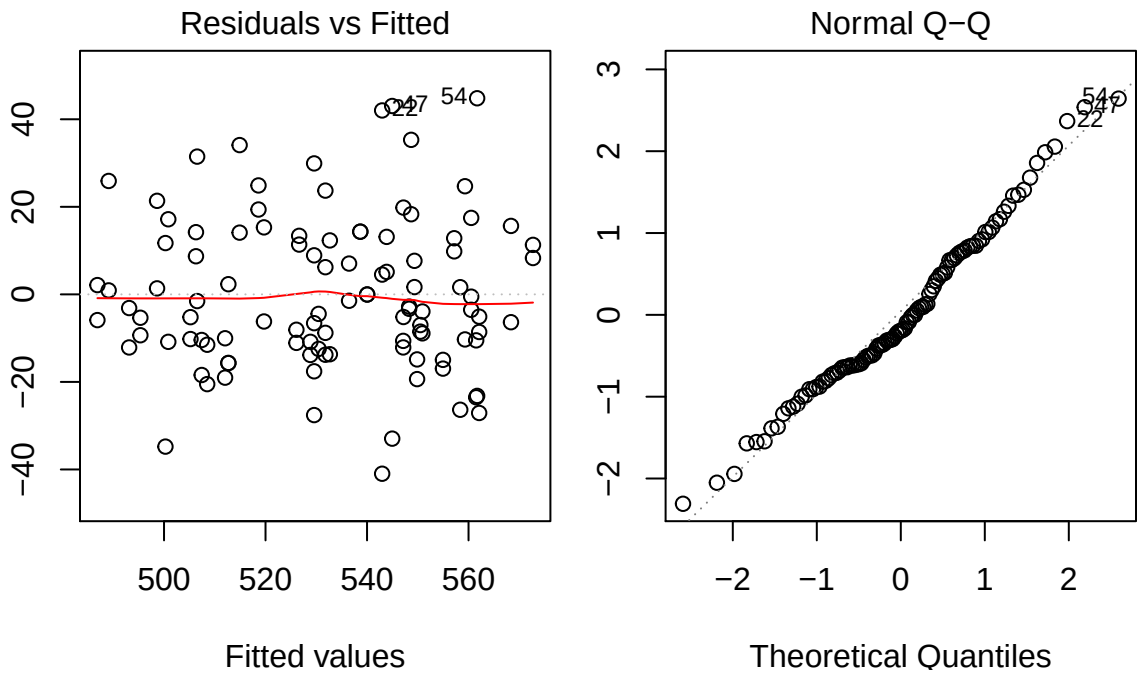

Figure 2: The residuals vs fitted values plot and residuals Q-Q plot of the ANOVA model.

As for the constancy of the residuals, since the number of samples in every treatment is small (Table 4 in the Appendix), some treatments have zero variance, leading to the failure of both Hartley and Bartlett's test. Therefore, Levene's test is employed at the 0.05 level of significance. With a p-value close to zero (6.135e-06, see Table 6 in the Appendix) the test suggests that the null hypothesis of constant error variance should be rejected, and that not all $(\sigma)_{ij}^2$s are equal. This unequal variance might be caused by the very small number of samples in each block. Sections 1.3 and 2.1 respectively address why missing values and outliers do not pose a problem for our analysis.

2.4 Inferential Analysis

Due to the small sample size in each treatment cell, tranformations of the data are unlikley to resolve the issue. We use a nonparametric test, the F rank test, to determine the roubustness of our ANOVA model, and compare the pairwise mean comparisons made under our ANOVA model with those made under the rank F test. Since it is not meaningful to compare the teachers' performance between different schools, only the class type pairwise mean comparisons are carried out.

Two procedures are taken to calculate the confidence interval, namely Bonferroni and Tukey's procedures, yield multipliers of 2.44 and 2.38 respectively. Since the Tukey procedure has a smaller multiplier, indicating a narrower confidence interval (CI), it has been used for the following comparisons. Herein, the Tukey 95% CIs of the class type pairwise mean comparison of the two-way ANOVA model and the rank F test are listed in Table 2 and Table 3.

Table 2: Tukey 95 Percent Confidence Interval for ANOVA Model

	Diff	Lower	Upper	P-value
REGULAR CLASS-REGULAR + AIDE CLASS	-2.54	-13.25	8.18	0.84
SMALL CLASS-REGULAR + AIDE CLASS	12.31	1.52	23.1	0.02
SMALL CLASS-REGULAR CLASS	14.85	4.21	25.48	0

Table 3: Tukey 95 Percent Confidence Interval for Rank F Model

	Diff	Lower	Upper	P-value
REGULAR CLASS-REGULAR + AIDE CLASS	-5.21	-16.53	6.12	0.52
SMALL CLASS-REGULAR + AIDE CLASS	10.78	-0.63	22.19	0.07
SMALL CLASS-REGULAR CLASS	15.99	4.74	27.23	0

The mean differences between the regular class and the small class are significant in both CI sets, because the 95% CIs do not include zero. As for the comparison between the regular+aide class and the small class, and the regular+aide class and the regular class, the nonparametric CIs suggest that the differences of these means are not significant.

In sum, teachers perform better in the small class type than in the regular class type, while the difference of teachers' performances is not significant between the regular+aide and the small class types, as well as the regular+aide and regular class types. Both models obtain similar results for the effects of small classes. Thus, our ANOVA model is relatively robust against the departures from constancy of error variance.

3 Causal Inference

The underlying experimental design of project STAR and fact that our analysis treats teachers as the unit of statistical interest, allows causal inference to be made. We note that the assumptions for causal inference are satisfied:

Firstly, the randomization within each school employed in project STAR (for both teachers and students) ensures that the ignorability condition is satisfied. Secondly, the SUTVA assumption is satisfied for the following reasons. The treatment is the same between teachers because each class size (small, regular, regular+aide) is kept at fixed levels between teachers, and each teacher is required to teach the same curriculum. There is no spillover effect because each student is only assigned one teacher, thus ensuring that a teacher's impact on a student in one class size is not carried over to any other teacher's classroom. Finally, if teachers taught different curriculums in different class sizes, a teacher in one class type might not be able to teach their class as effectively compared to a teacher in another class type because of the difference in curriculum between class types. However, since all teachers teach the same curriculum, and do not share students, the potential outcome of one teacher does not affect another.

Thus, all assumptions necessary for causal inference on average causal effects are satisfied, and we are able to perform causal inference. We conclude that class size does have an effect on first-grade teachers' mathematics performance; in particular, smaller class sizes allow teachers to more effectively instruct their students when compared to teachers with larger class sizes.

4 Conclusions

In this report, we investigated the effects of class type on first grade mathematics performance using teachers as the unit of statistical analysis by fitting a two-way ANOVA model for a randomized block design. Our results show that there is a significant treatment effect: teachers in smaller class sizes perform better, on average, than teachers with a regular class size. Furthermore, we determined that causal inference can be made here, and determined that smaller class sizes facilitate more effective instruction. We note that while the results of this project agree with our non-causal inferences in Project 1, namely that there is a significant difference between treatment means, treating the teacher as a unit allows causal inference to be done in this project, unlike the previous one.

Future analysis of this data should include school location as a third factor. Children in different locations (eg. rural, or urban) may have different needs, and respond to differences in class size treatments differently. These differences could hold important implications for policy makers.

5 References

1. G.W. Imbens, and D.B. Rubin. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction. Cambridge University Press.
2. M. H. Kutner, C.J. Nachtsheim, J. Neter, and W Li. 2005. Applied Linear Statistical Models. Fifth Edition.

6 Appendix

Table 4: Sample Size of Each Treatment Group

School ID	REGULAR + AIDE CLASS	REGULAR CLASS	SMALL CLASS
159171	2	2	2
161183	2	2	2
169229	4	5	3
180344	2	2	2
201449	2	2	3
209510	2	2	2
215533	2	2	2
216537	2	2	2
221571	2	2	2
234628	2	2	2
244697	2	2	2
244708	2	2	2
244723	2	2	2
244776	2	2	2
244806	2	3	2
257905	2	2	3

Table 5: Sample Mean of Each Treatment Group

School ID	REGULAR + AIDE CLASS	REGULAR CLASS	SMALL CLASS
159171	553.00	539.25	532.75
161183	544.50	566.50	582.50
169229	533.62	527.00	544.83
180344	532.00	522.00	553.00
201449	554.00	535.75	565.00
209510	516.50	539.00	540.00
215533	554.50	550.00	546.00
216537	542.75	545.25	572.50
221571	504.00	510.00	497.50
234628	568.50	539.00	573.00
244697	492.50	517.75	524.25
244708	502.50	485.00	488.75
244723	493.00	497.50	540.75
244776	488.00	485.50	521.50
244806	539.00	503.00	516.50
257905	544.50	575.50	548.50

Table 6: Leven's Test Results

	Df	F value	Pr(>F)
group	47	3.44	0
	57		

7 Session Information

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18363)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] kableExtra_1.1.0  SuppDists_1.1-9.5 foreign_0.8-72    forcats_0.4.0
##  [5] stringr_1.4.0     dplyr_0.8.4       purrr_0.3.3       readr_1.3.1
##  [9] tidyr_1.0.2       tibble_2.1.3      ggplot2_3.2.1     tidyverse_1.3.0
## [13] AER_1.2-9         survival_3.1-8    sandwich_2.5-1    lmtest_0.9-37
## [17] zoo_1.8-7         car_3.0-6         carData_3.0-3
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.1        jsonlite_1.6.1    viridisLite_0.3.0 splines_3.6.2
##  [5] modelr_0.1.5      Formula_1.2-3     assertthat_0.2.1  highr_0.8
##  [9] cellranger_1.1.0  yaml_2.2.1        pillar_1.4.3      backports_1.1.5
## [13] lattice_0.20-38   glue_1.3.1        digest_0.6.23     rvest_0.3.5
## [17] colorspace_1.4-1  htmltools_0.4.0   Matrix_1.2-18     pkgconfig_2.0.3
## [21] broom_0.5.4       haven_2.2.0       scales_1.1.0      webshot_0.5.2
## [25] openxlsx_4.1.4    rio_0.5.16        farver_2.0.3      generics_0.0.2
## [29] withr_2.1.2       lazyeval_0.2.2    cli_2.0.1         magrittr_1.5
## [33] crayon_1.3.4      readxl_1.3.1      evaluate_0.14     fs_1.3.1
## [37] fansi_0.4.1       nlme_3.1-142      xml2_1.2.2        tools_3.6.2
## [41] data.table_1.12.8 hms_0.5.3         lifecycle_0.1.0   munsell_0.5.0
## [45] reprex_0.3.0      zip_2.0.4         compiler_3.6.2    rlang_0.4.4
## [49] grid_3.6.2        rstudioapi_0.11   labeling_0.3      rmarkdown_2.5
## [53] gtable_0.3.0      abind_1.4-5       DBI_1.1.0         curl_4.3
## [57] R6_2.4.1          lubridate_1.7.4   knitr_1.28        stringi_1.4.4
## [61] Rcpp_1.0.3        vctrs_0.2.2       dbplyr_1.4.2      tidyselect_1.0.0
## [65] xfun_0.18
```