Assessing Class Size Effect on Math Scores for 1st Grade

Contents

Team ID: 4

Min Kim

Kenneth Broadhead

Nanhao Chen

Koral Buch

Introduction

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 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 achivement from kindergarden through third grade. Additionally, teachers were randomly assigned to the classes they would teach. The three class types were as follows:

- [Control group] regular class (22 to 25 students per teacher)

- [Treatment group I] small class (13 to 17 students per teacher)
- [Treatment group II] 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. The effect of these variables on the test scores can be estimated as well.

Objective

In this project, we analyze the data from Project STAR, focusing on first grade math scores. The primary question of interest is whether class type affects the students' math scores in the SAT. In this project, we estimate the change in the math score for each treatment group given the control group. We do not examine class size effects on subset of students (male/female, or black/white etc). However, such analyses are possible and should be undertaken for the potential impact on policy changes.

Statistical Analysis

To answer our main question of interest, we fit a one-way ANOVA model and obtained basic properties. The dependent variable is the math test scores for first grade, and the independent variable is the class type. We formally investigate whether class sizes are associated with math performance. Furthermore, we investigate relative effect differences for each class size on math scores using Bonferroni's correction for multiple comparisons.

Results

Missing Values

The data contain 11,598 observations. We are interested only in those that have data for the relevant treatment, i.e. have values for first grade class sizes. We therefore trim the data set accordingly. Of the remaining 6829, there are 229 observations that do not have a first grade math score (**math1**). Due to the large number of observations, we simply delete these observations. We first investigate whether or not these observations can be considered as random, or if there is some problematic bias that might be introduced by deleting them. Since the number of missing values is relatively small compared with the remaining data, it is unlikely that deleting them would induce any noticable bias in our analyses. However, as a matter of formatlity, we investigate whether or not these missing values may be considered as random (not dependent on any variables we consider), or stem from some systematic reasons potentially inducing bias if more data points are missing. We investigate variables pertaining to gender, ethnicity, school location, whether or not the students qualified for free lunch (a possible economic indicator). All variables had a similar distribution between the dataset with the missing math scores, and the dataset with math scores, with the exception with the variable indicating whether the student qualifying for free lunch. A slightly higher proportion of students missing a math score qualify for free lunch then one might expect if the missing values were random. However, this difference is not extreme, and may be attributed to randomness after all, owing to the difference between the number of observations in each set. The data analysis is thus conducted on the remaining 6600 observations.

Descriptive Analysis

Figures 1 to 3 and Table 1 show summary statistics of 1st grade math scores and 1st grade class types. The distributions of each math score population seems to be approximately normal. According to the boxplot in Figure 3, the **math1** scores of different class types suggest that class types has an effect on the average of the mathematics scores, and the variations of each population appear to be similar. Since teachers were randomly assigned to classes, confounding variables, such as degree types and teaching experiences of the teachers, can be ignored in this analysis. Finally, a cell means one-way ANOVA model seems appropriate for the investigation of the effects of class type on mathematics scores.

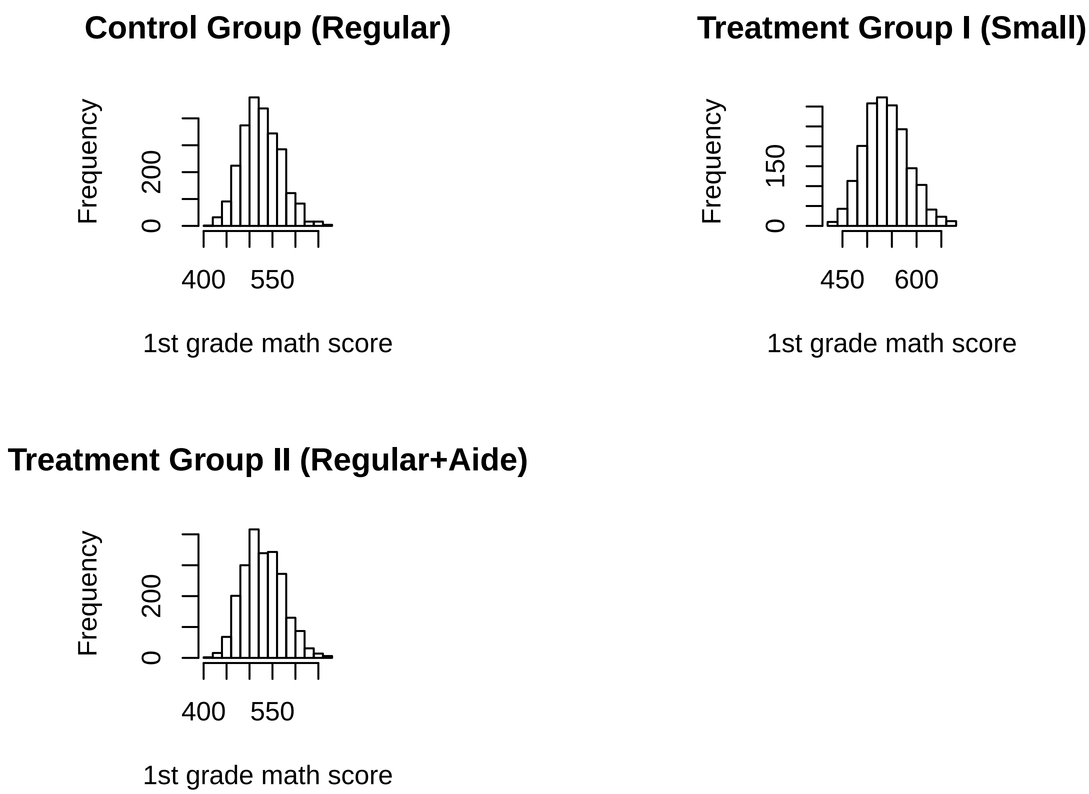

Figure 1: Figure 1: Histograms of 1st grade math scores by different class types.

Table 1. The summary statistics of the 1st grade math scores by class types.

Class Type	Count	Min.	1st Qu.	Mean	Median	3rd Qu.	Max.
regular	2507	408	495.00	525.2744	523	495.00	676
small	1868	425	509.25	538.6777	535	509.25	676
regular+aide	2225	404	497.00	529.6252	529	497.00	676

Inferential Analyses and Model Diagnoses

Here is the one-way ANOVA model we employed for the analysis: $Y_{ij} = \mu_i + \varepsilon_{ij}$, $j = 1, ..., n_i; i = 1, ..., r$

where:

Y_{ij} is the value of the response variabl in the jth observation for the ith Class Type;

μ_i is the means of the ith Class Type;

ε_{ij}'s are random errors, normally distributed with cosntant variance;

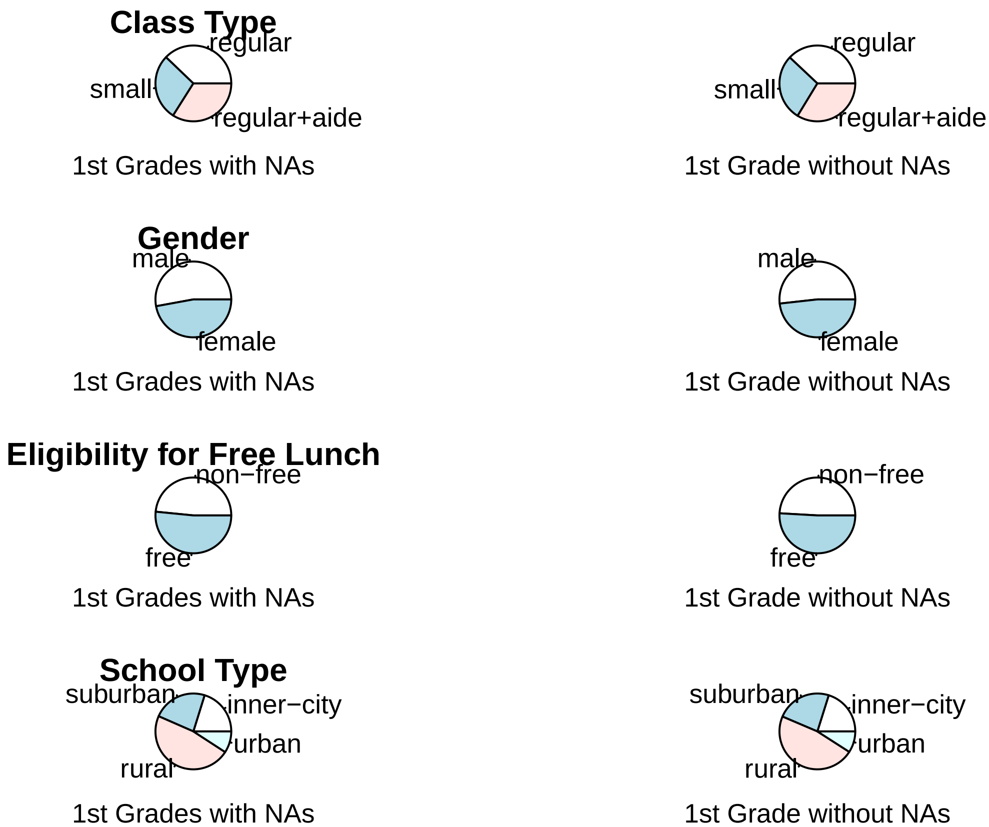

Figure 2: Figure 2. Pie charts of the class types, gender, eligibility for free lunch, and school type.

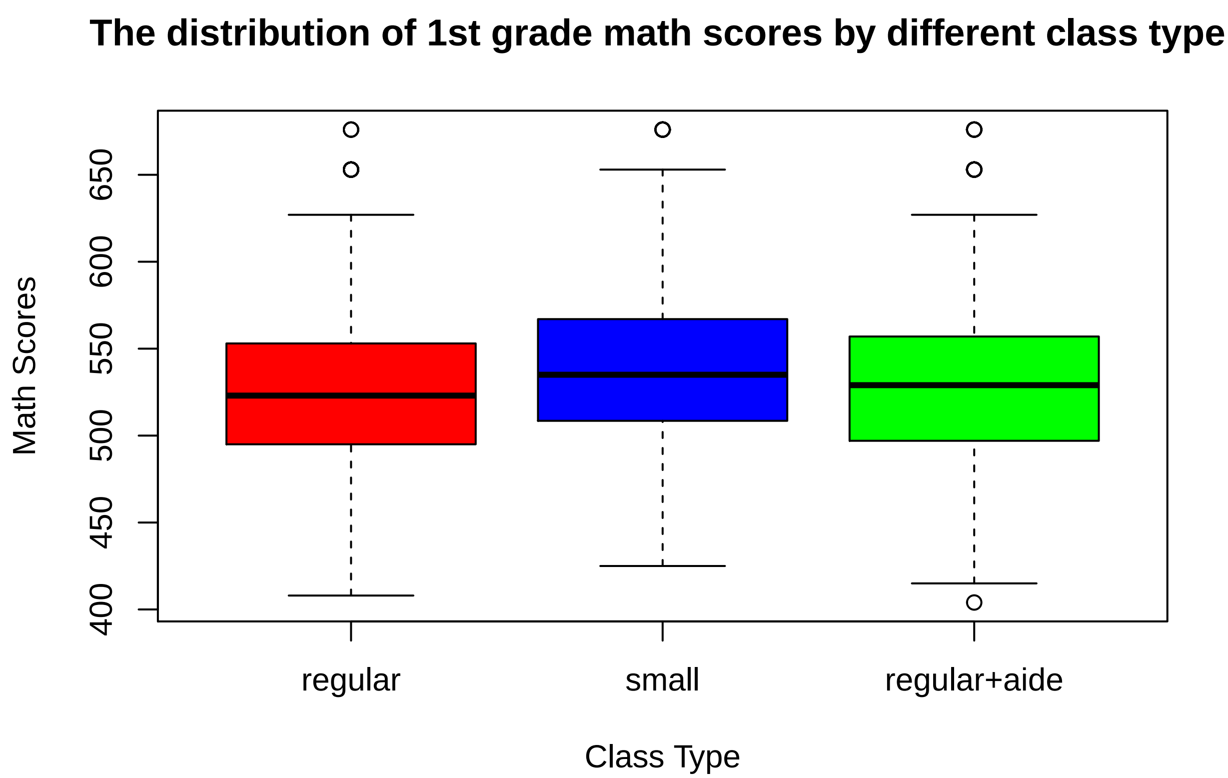

Figure 3: Figure 3. Box plot of the distribution of 1st grade math scores by different class type.

n_i is the number of observations for the ith Class Type and r is the numbers of the Class Types.

Based on the one-way ANOVA results (Table 2), the mean square treatment (MSTR) is 97538 and the mean square error (MSE) is 1829. The fitted values for each treatment are 525.2744 for regular class size, 538.6777 for small class size, and 529.6251 for regular class size with aide. The F-test statistic value is 53.33 and the corresponding P-value is close to 0, indicating there is a significant difference between means of 1st grade math scores in different class sizes. Therefore, we conclude that there is a significant relationship between class sizes and 1st grade students' math scores. The residual plot in Figure 4 shows that the resdiuals of different class types are dispersed equally around zero, which indicates the equal variance of the residuals at different factor levels. To quantitatively describe the constancy of the variance at different factor level, the Levene's test is applied and the results are listed in Table 3. Accordingly, the p-value of Levene's test is slightly larger than 0.05 (0.0517), and therefore we conclude the residual variances are not significant different between factor levels. At the same time, the histograms of the residuals in the different class types in Figure 5 support the approximate normality of the residuals at different factor levels. The residual Q-Q plot shows that the residuals deviate only slightly from normality; however, the F-test used above is robust against slight deviation from normality. Thus, the figures and discussion above demonstrate the appropriateness of the one-way ANOVA model.

Table 2. The summary of the one-way ANOVA.

```
##               Df   Sum Sq Mean Sq F value Pr(>F)
## star1          2   195075   97538   53.33 <2e-16 ***
## Residuals   6597 12065523    1829
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3. The summary of the Levene's test for the constancy of the variance.

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## star1          2    3900    1950   2.964 0.0517 .
## Residuals   6597 4340562     658
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In order to perform a comparison of the means of the math scores in different class type, Three different cutoff-procedures are considered, namely Bonferroni, Tukey, and Scheffe. Since we are only interested in three pairwise comparisons between each of the three different class sizes, we only consider the Bonferroni's procedure. Though applicable, Tukey's procedure gives wider confidence intervals than Bonferroni's procedure, we therefore only use Bonferroni's procedure. Here the 99% confidence intervals of the mean comparisons between regular class type and small class type, regular and regular with aide class types, and small and regular with aide class types are listed in Table 4. We note that none of the intervals contain 0, and thus, indicating a significant difference between 1st grade mean math scores for each factor level.

Table 4. The 99% Bonferroni confidence interval of mean comparisons

	Lower	Upper
Regular vs Small	9.5651733	17.241424
Regular vs Aide	0.6933079	8.008166
Small vs Aide	5.1119544	12.993169

Conclusion

In conclusion, we find a significant difference in 1st grade students' mathematics performance on the SAT between class size, with students in smaller classes outperforming students in larger classes on average.

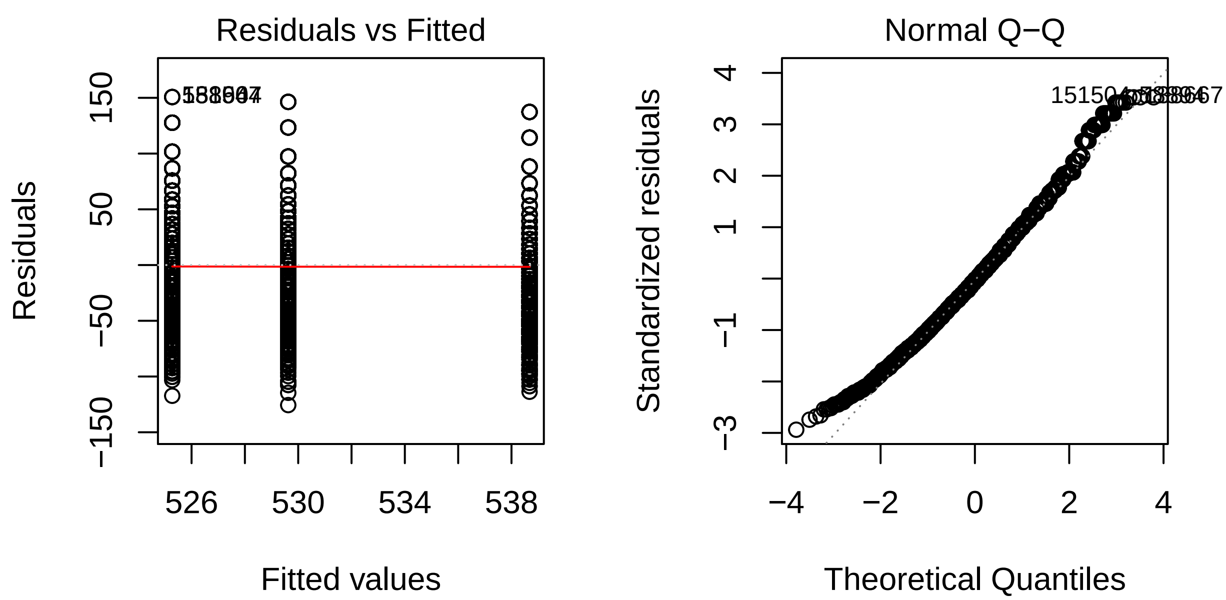

Figure 4: Figure 4. Residuals vs Fitted and Normal Q-Q plots.

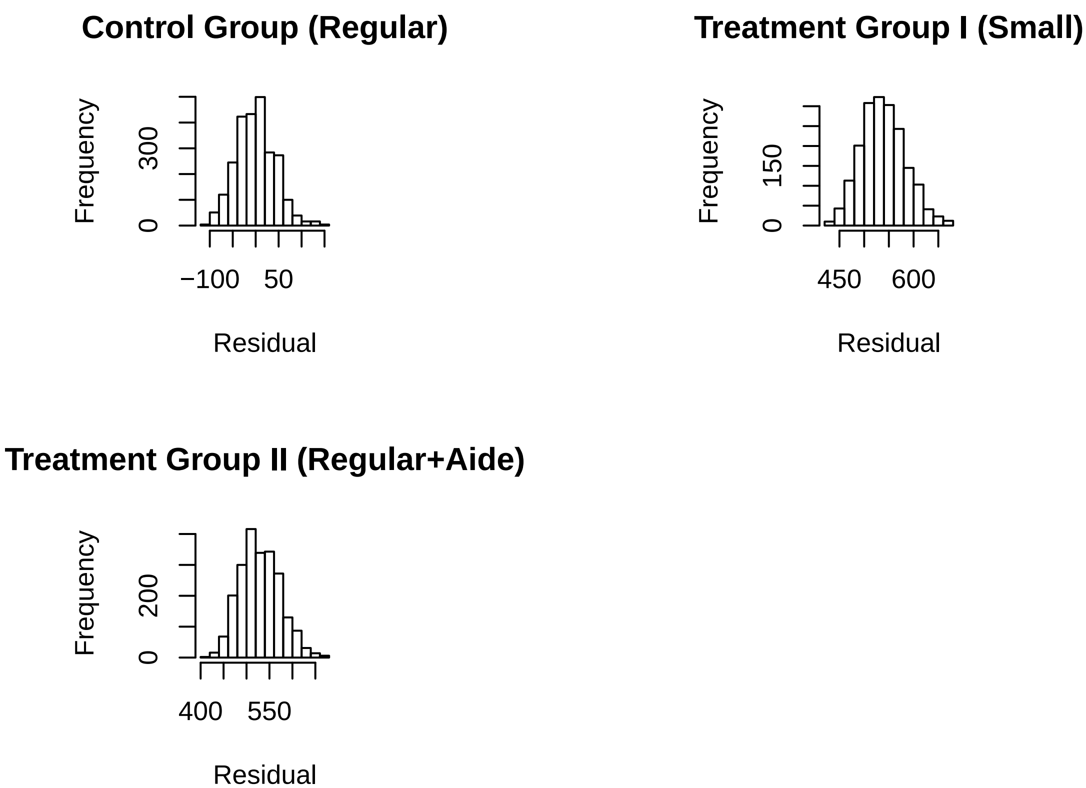

Figure 5: Figure 5: Histograms of residuals of 1st grade math scores by different class types.

Furthermore, according to the design of this experiment, the students and teachers are randomly assigned to different levels (class types), thus ensuring that it is the size of the class and not the teachers' ability affecting students' performance. Therefore, some causal inferences can be made. Comparing math scores between 1st grade students in the regular class type and small class type, one can conclude that the smaller class sizes lead to improved mathematics performance. Due to lack of information on the aide in the regular+aide group, we are unable to make any causal inferences.

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] forcats_0.4.0    stringr_1.4.0    dplyr_0.8.4      purrr_0.3.3
##  [5] readr_1.3.1      tidyr_1.0.2      tibble_2.1.3     ggplot2_3.2.1
##  [9] tidyverse_1.3.0  kableExtra_1.1.0 AER_1.2-9        survival_3.1-8
## [13] sandwich_2.5-1   lmtest_0.9-37    zoo_1.8-7        car_3.0-6
## [17] 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        generics_0.0.2    withr_2.1.2
## [29] lazyeval_0.2.2    cli_2.0.1         magrittr_1.5      crayon_1.3.4
## [33] readxl_1.3.1      evaluate_0.14     fs_1.3.1          fansi_0.4.1
## [37] nlme_3.1-142      xml2_1.2.2        foreign_0.8-72    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   rmarkdown_2.5     gtable_0.3.0
## [53] abind_1.4-5       DBI_1.1.0         curl_4.3          R6_2.4.1
## [57] lubridate_1.7.4   knitr_1.28        stringi_1.4.4     Rcpp_1.0.3
## [61] vctrs_0.2.2       dbplyr_1.4.2      tidyselect_1.0.0  xfun_0.18
```