Class sizes and students' math test scores in first grade: STAR Project

Sirapat Watakajaturaphon

Februrary 17, 2023

You can find the R codes at the Code Appendix Section.

Abstract

In this project, we use the dataset from the Tennessee Student Teacher Achievement Ratio (STAR) study to assess the effect of class sizes on students' math scores in the first grade. Each teacher is treated as the basic unit of our analysis. Analyses show that the class types has a significant impact on math scaled scores and, to be specific, the small-sized class is statistically associated with the highest scores in the first grade.

Introduction

The Tennessee Student Teacher Achievement Ratio (STAR) experiment was conducted between 1985 and 1989 throughout many countries, such as the United States, Australia, Hong Kong, Sweden, and Great Britain. It was a randomized, longitudinal experiment with the goal of studying the effect of class sizes on students' academic performance in the early grades.

The dataset used in this project are from the AER package in R. It contains the key variables including class sizes (small, regular, and regular with a full time teacher's aide) and math test scores from kindergarten to third grade. In this study, the questions of interest will be focused on the first grade and their math scores. We will examine whether there are any differences in math scaled scores across class sizes, and if so, which class type is associated with the highest math scaled scores. The results will help schools improve students' performance by making more efficient policy related to class sizes. According to Achilles (2012) [1], "for school improvement, policies should rely on class size" and, in particular, there are indeed both short- and long-term impacts of small classes (about 15-17 students) on student achievement.

Background

The STAR dataset from the AER package contains 11,598 observations and 47 variables. Since the target in our study is the first graders, the key variables are those related to first grade:

1. `math1` total math scaled scores in first grade;
2. `star1` the class types in the first grade with three levels: *small* (about 15-17 students), *regular* (about 22-25 students), and *regular with a teacher's aide* (about 22-25 students);

3. `schoolid1` school IDs in first grade with 80 levels;
4. `experience1` years of teacher's total teaching experience in first grade;
5. `tethnicity1` teacher's ethnicity in first grade with two levels: *cauc* (Caucasian) and *afam* (African-American).

Each of 79 STAR schools were considered enrolled only if they had enough students to have at least one class of each type. Once the schools were enrolled, students were randomly assigned to the three types of classes, and one teacher was randomly assigned to each class. In our analysis, each teacher is treated as the basic unit. Even though there are no variables representing teacher IDs in the AER package dataset, we can uniquely identify teachers based on characteristics `experience1`, `tethnicity1`, `schoolid1`, and `star1`. The next issue is that because there are multiple students under each teacher, we need to choose one summary measure (for example, mean or median of `math1`) for each teacher. For more details, see Section 4.

There are many research [1][2] related to class size and academic achievement. Achilles (2012) [1] summarized the short- and long-term effects of small classes on students achievement in the early grades (kindergarten through third grade). Finn, Gerber, Boyd-Zaharias (2005) [2] focused on studying the long-term effects of early school experiences and its association to high school graduation. Findings showed that graduating was related to the early grades achievement and that attending small classes for 3 or more years increased the likelihood of graduating from high school.

Descriptive analysis

The most three relevant variables are:

1. `math1` total math scaled scores in first grade;
2. `star1` STAR class types in first grade (small, regular, and regular-with-aide. NA indicates that no STAR class was attended.); and
3. `schoolid1` school IDs in first grade.

Univariate descriptive statistics

We first see the distribution of each key variable. Since `math1` is quantitative, we can calculate its summary statistics shown in Table 1. For the categorical variable `star1` with 3 levels, the pie charts are drawn. In Figure 1, we can see a large portion of the missing values. Without those missing values, the percentages of each class size are displayed more clearly, see Figure 2. Because there are too many categories of `schoolid1`, we plot the count (bar) plot instead of a pie chart. The number of missing values of `schoolid1` is 4769 which is very large, so we plot Figure 3 without them just for an aesthetic purpose. The outstanding line is of `schoolid1=51` with 238 counts and there are four `schoolid1=6, 18, 42, and 76` with zero count.

```
library(dplyr)
library(ggplot2)
library(knitr)

# import data
library('AER')
data('STAR')

# only keep 1st grade
STAR.NA = STAR%>%dplyr::select(math1, school1, experience1, tethnicity1, schoolid1, star1)
```

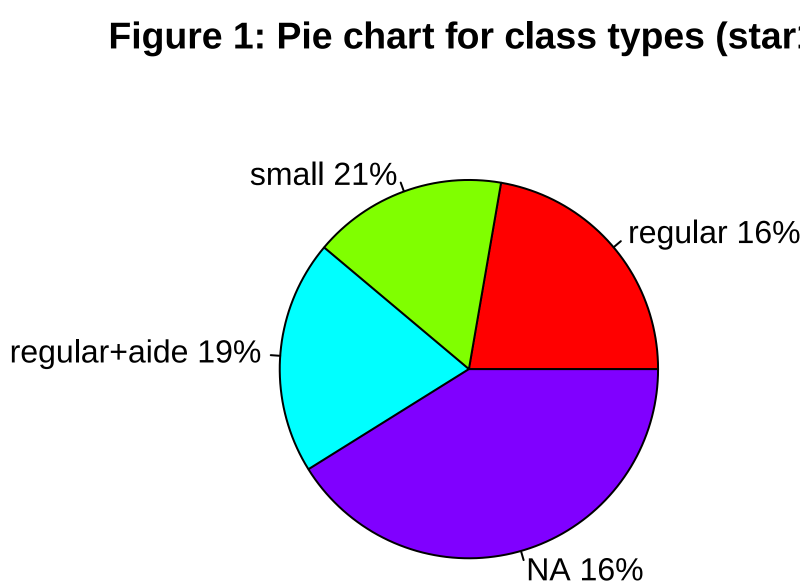
```r
# 1st grade and remove rows with NA value in any column
STAR.dat = STAR.NA %>% na.omit()

# check types of variables
# sapply(STAR.dat, class)

# i=1 small ; i=2 regular; i=3 regular+aide
STAR.dat$star1 = factor(STAR.dat$star1, levels = c('small','regular','regular+aide'))
# levels(STAR.dat$star1)
```

```r
# descriptive stats: math scores
desc.math1 = STAR.dat %>%
  summarise(Min = min(math1), '1st Qu.' = quantile(math1, prob=c(.25)),
            Median = median(math1), Mean = mean(math1),
            '3rd Qu.' = quantile(math1, prob=c(.75)),
            Max = max(math1), sd = sd(math1), 'NAs'=summary(STAR.NA$math1)[7])
kable(desc.math1, caption = 'Table 1: Desciptive statistics for math scores in 1st grade (math1)')
```

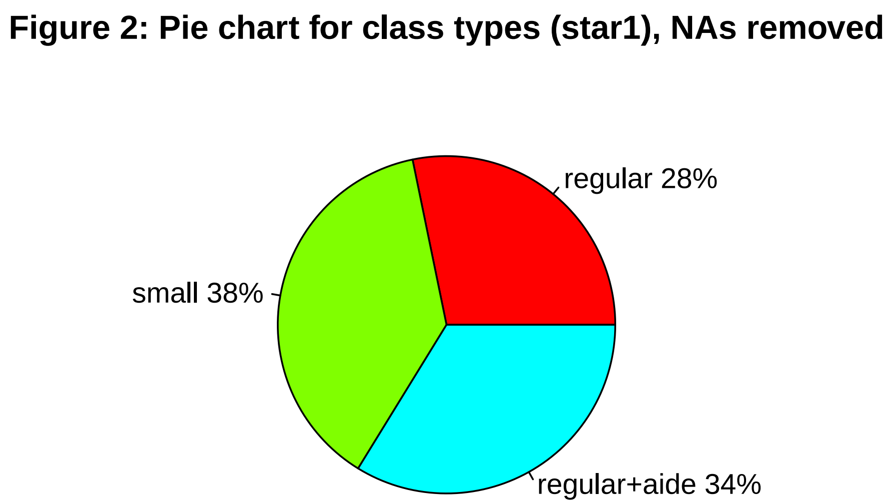
Table 1: Table 1: Desciptive statistics for math scores in 1st grade (math1)

Min	1st Qu.	Median	Mean	3rd Qu.	Max	sd	NAs
404	500	529	530.5541	557	676	43.11925	4998

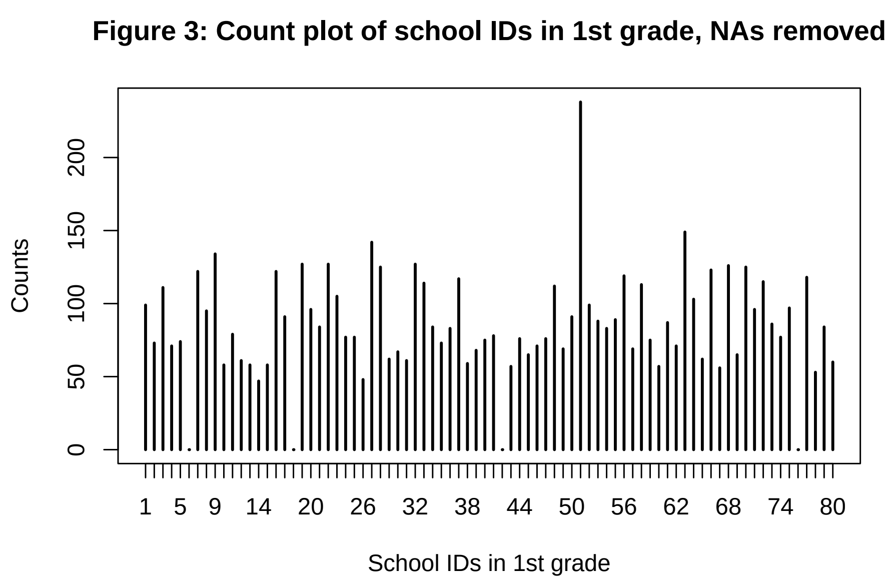
```r
# descriptive stats: class types
n = nrow(STAR.NA)
lbls = c('regular','small', 'regular+aide', 'NA')
pct = round(100*summary(STAR.dat$star1)/n)
lab = paste(lbls,pct)
lab = paste(lab,'%',sep='')
pie(summary(STAR.NA$star1), labels=lab, col=rainbow(4),
    main='Figure 1: Pie chart for class types (star1)')
```



```r
n = nrow(STAR.dat)
lbls = c('regular','small', 'regular+aide')
pct = round(100*table(STAR.dat$star1)/n)
lab = paste(lbls,pct)
lab = paste(lab,'%',sep='')
pie(table(STAR.dat$star1), labels=lab, col=rainbow(4),
    main='Figure 2: Pie chart for class types (star1), NAs removed')
```



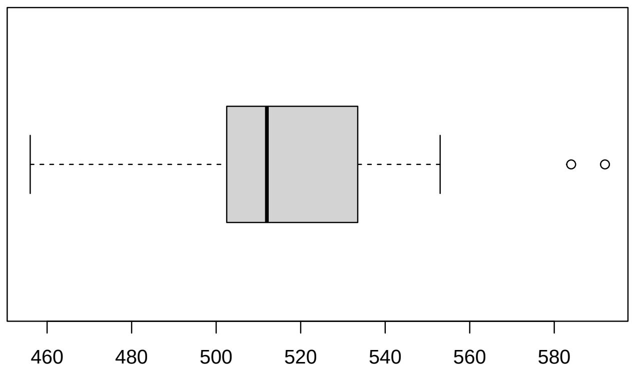

```r
# descriptive stats: school ids
plot(table(STAR.NA$schoolid1), ylab='Counts', xlab='School IDs in 1st grade',
     main='Figure 3: Count plot of school IDs in 1st grade, NAs removed')
```



From the data set, we can easily notice that various number of students are assigned to each teacher. In order to obtain one summary measure with teacher as the unit, we need to aggregate students' math scores. Note that we can uniquely identify teachers based on the following characteristics: `experience1`, `tethnicity1`, `schoolid1`, and `star1`.

As an example, we draw the boxplot (see Figure below) of `math1` for the teacher who has the following characters `experience1=0`, `tethnicity1=cauc`, `schoolid1=5`, and `star1=regular`. Since the outliers are present, the median might be a more appropriate choice because it is not as easily influenced by those extreme values as the mean is.

```
teacher = STAR.dat %>% filter(experience1=='0', tethnicity1=='cauc',
                              schoolid1=='5', star1=='regular')
boxplot(teacher$math1, horizontal=T)
```



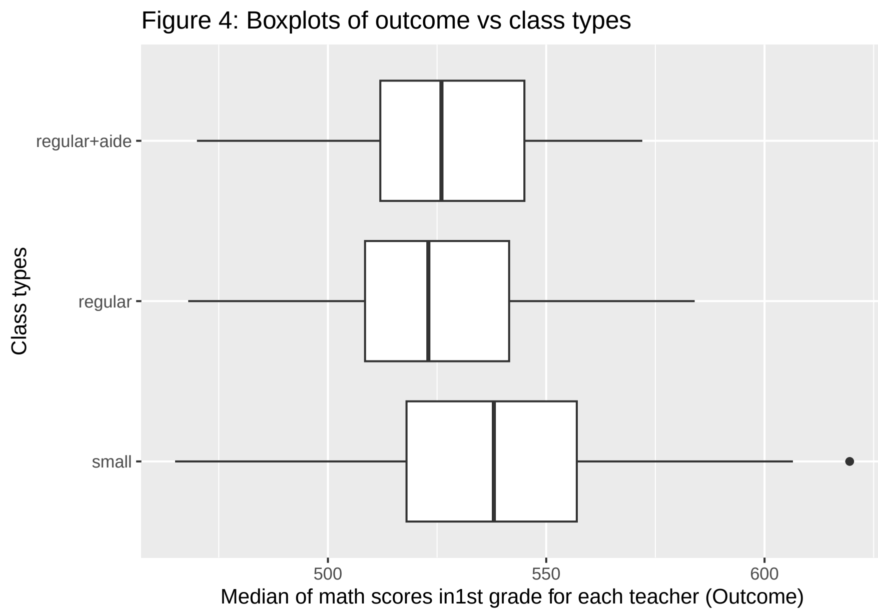
Then we apply the median measure for all teachers.

```
df = STAR.dat %>% group_by(experience1, tethnicity1, schoolid1, star1) %>%
  summarise(med_mathscore = median(math1))
```

Multivariate descriptive statistics

- Boxplots for the median of math score in first grade for each teacher v.s. class types: The distributions of `regular` and `regular+aide` are quite similar. The distribution of `small` is different, the spread is wider, and the median and mean are higher. So, based on Figure 4, it is possible that there is a difference in math scaled scores across class sizes.
- Summary statistics for the median of math score in first grade for each teacher v.s. school IDs: From Table 2, There appears to be differences in the outcome across school IDs. Thus, it is possible that this variable `schoolid1` is also significant in the model.

```
# boxplot: outcome vs class types
ggplot(df, aes(med_mathscore,star1)) + geom_boxplot() +
  xlab('Median of math scores in1st grade for each teacher (Outcome)') + ylab('Class types') +
  labs(title='Figure 4: Boxplots of outcome vs class types')
```




```
# median: outcome vs school IDs
med_outcome_by_schoolid = aggregate(med_mathscore ~ schoolid1, FUN=median, data = df)

# descriptive stats: math scores
desc.med.math1 = med_outcome_by_schoolid %>%
  summarise(Min = min(med_mathscore), '1st Qu.' = quantile(med_mathscore, prob=c(.25)),
            Median = median(med_mathscore), Mean = mean(med_mathscore),
            '3rd Qu.' = quantile(med_mathscore,prob=c(.75)),
            Max = max(med_mathscore), sd = sd(med_mathscore))
kable(desc.med.math1, caption = 'Table 2: Summary statistics for outcome vs school IDs')
```

Table 2: Table 2: Summary statistics for outcome vs school IDs

Min	1st Qu.	Median	Mean	3rd Qu.	Max	sd
486	515	531.375	529.2138	545	569.5	21.44204

Sine there are too many schools, it is not practical to draw the interaction plot in order to observe the interaction effects. We will use the inference in Section 5 to decide whether or not the interaction terms are significant.

Inferential analysis

Define the index i represents the class type: small ($i = 1$), regular ($i = 2$), regular with aide ($i = 3$), and the index j represents the school indicator. The outcome Y_{ijk} is the median of math scores in the first grade of the kth sample in the ith class type and jth school ID. The overall mean $\mu_{\cdot\cdot}$ is the total sum of all cell means $\{\mu_{ij} : i = 1, 2, 3, j = 1, ..., 80\}$.

We tried fitting models with and without the missing values and found that the results are similar. So, we choose to perform the following inferences with the missing values removed.

We look at the numbers of observations in cells $i = 1, 2, 3$ and $j = 1, ..., 5$ and see that those numbers vary across cells. This indicates that it is an imbalanced design.

```r
# see no. of observation in each cell
table(df$star1, df$schoolid1)[1:3,1:5]
```

```
##
##                  1 2 3 4 5
##   small          1 1 2 2 2
##   regular        1 1 2 1 1
##   regular+aide 1 1 2 1 1
```

First, we quickly test for the interaction terms to see whether those terms should be included in our analysis further or not. So, here we have

- Full model: $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$
- Reduced model: $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ijk}$

```r
# test for interactive effect
full.model = aov(med_mathscore ~ star1 * schoolid1, data = df)
red.model  = aov(med_mathscore ~ star1 + schoolid1, data = df)

# anova(red.model, full.model) # fail to reject the null model, so model1 is preferred!
```

Putting those two models above into the `anova()` function, the resulting p-value is computed to be 0.428. Hence, we decide that the interaction terms can be dropped at the significance level $\alpha = 0.05$.

So, our two-way ANOVA model is expressed as follows:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ijk}, \quad k = 1, ..., n_{ij}, \quad j = 1, ..., 80, \quad i = 1, 2, 3.$$

where the random errors ϵ_{ijk} are i.i.d. $N(0, \sigma^2)$ and the constraints on the factor effects are

$$\sum_{i=1}^{3} \alpha_i = \sum_{j=1}^{80} \beta_j = 0.$$

We then fit the model on the Project STAR data. The test results for the main effects remain the same regardless of the orders. So, we decided to continue the analysis with the Type I ANOVA model even though it is an imbalanced design.

```r
model1 = red.model
summary(model1)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## star1         2  10935    5467  17.501 7.43e-08 ***
## schoolid1    75 149887    1998   6.397  < 2e-16 ***
## Residuals   259  80912     312                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(med_mathscore ~ schoolid1 + star1, data = df))
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## schoolid1      75 148750    1983   6.349  < 2e-16 ***
## star1           2  12072    6036  19.321 1.51e-08 ***
## Residuals     259  80912     312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Anova(lm(med_mathscore ~ schoolid1 + star1, data=df), type = 'II')
```

Since there are 80 schools, reporting the estimated coefficients for school IDs does not contribution information to readers because it is difficult to interpret and takes a lot of spaces. Hence, we will only report the estimated coefficients for `(intercept)` which is 551.09, `star1 regular` -12.10, and `star1 regular+aide` -13.16.

```
# model1$coefficients
```

Next, we will investigate our primary question of interest "whether there are any differences in math scaled scores in the first grade across class types". The hypotheses are

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad \text{vs} \quad H_1 : \text{at least one of } \alpha_i \text{ in } H_0 \text{ is not zero}$$

From the inferences above, the p-value for `star1` is smaller than 0.05 (for both orders). This means, we have enough evidence to reject H_0 and conclude that there are differences in math scaled scores in the first grade across class types at the significance level $\alpha = 0.05$.

Since the differences in the median of math scores across class sizes exist, we would like to examine further which class size is associated with the highest score. Recall, from Figure 4, we suspect that it is class size `small`. To carry out the formal test, Tukey-Kramer method is used with 95% family-wise confidence level. We only need to focus on the difference of the two largest means which are `small` and `regular+aide` (See table 3).

```
cell.mean = aggregate(med_mathscore ~ star1, FUN=mean, data = df)
kable(cell.mean, caption='Table 3: Means of output across class types')
```

Table 3: Table 3: Means of output across class types

star1	med_mathscore
small	537.9512
regular	525.6535
regular+aide	526.7200

Given the resulting p-value on the second row of Tukey multiple comparisons of means, the difference exists. Furthermore, the interval of the difference of `regular+aide` and `small` has the lower bound (lwr=-16.84) and the upper bound (upr=-5.62). So, the class size `small` is most likely to be associated with the highest score at the overall significance level 0.05.

```
TukeyHSD(model1, conf.level = 0.95, which = 'star1')
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = med_mathscore ~ star1 + schoolid1, data = df)
##
## $star1
##                            diff        lwr       upr     p adj
## regular-small        -12.297711 -17.714410 -6.881012 0.0000006
## regular+aide-small   -11.231220 -16.841252 -5.621187 0.0000116
## regular+aide-regular   1.066491  -4.641982  6.774965 0.8986669
```

Sensitivity analysis

```r
par(mfrow=c(2,2), mar=c(3,3,2,2), mgp=c(1.7,.7,0))
plot(model1)
```

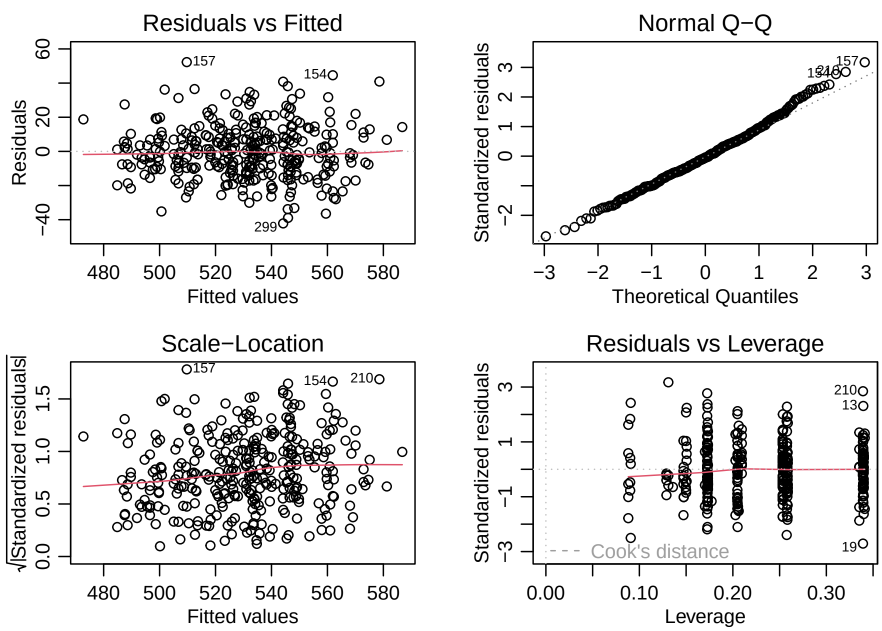


Figure 1: Figure 5: Diagnostics plots

- There is a presence of outliers. But since they seem not be very severe, we may need further investigations to decide whether to remove them.
- From the Residuals vs Fitted plot and Scale-Location plot, the points appear to have an equal spread along the X-axis. So, the constant variance assumption is satisfied.
- Normal Q-Q plot shows a straight line pattern, so the normality assumption seems to hold. Moreover, from Shapiro-Wilk normality test, the p-value is larger than 0.05, hence we conclude that the errors are normally distributed at $\alpha = 0.05$.

```r
# Shapiro-Wilk normality test
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.99305, p-value = 0.1208
```

It seems like our model assumptions are not violated. We try to carry out the rank test (nonparametric approach) to check if the answers change. We found that all of our answers to the questions of interest remain the same at $\alpha = 0.05$.

```
# The rank test
df$rank = rank(df$med_mathscore)
summary(aov(rank~star1 + schoolid1, data=df))
```

```
##               Df  Sum Sq Mean Sq F value   Pr(>F)
## star1          2  123416   61708  14.621 9.63e-07 ***
## schoolid1     75 1970543   26274   6.225  < 2e-16 ***
## Residuals    259 1093090    4220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(aov(rank~star1 + schoolid1, data=df), conf.level = 0.95, which = 'star1')
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = rank ~ star1 + schoolid1, data = df)
##
## $star1
##                             diff       lwr       upr     p adj
## regular-small         -42.696834 -62.60614 -22.78752 0.0000024
## regular+aide-small    -35.411220 -56.03114 -14.79130 0.0002009
## regular+aide-regular    7.285614 -13.69613  28.26735 0.6918830
```

Discussion

We would like to answer two questions of interest about the relationships of class sizes and students' academic performance in first grade where each teacher is treated as the basic unit. Since the STAR dataset in the AER package does not provide the teacher IDs variable, we need to aggregate the math score in first grade using the median as a summary statistics. So, right now our outcome Y_{ijk} is the median of math scores in first grade for each teacher. We examine further and find that the interaction effects $(\alpha\beta)_{ij}$ of class sizes and school IDs are not significant. Hence, in this project, we use the imbalanced additive two-way ANOVA model $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ijk}$. According to model diagnostics, our model follows the model assumptions. So, the inference results are reliable.

The analysis revealed a significant association between class sizes and math scaled scores in first grade, i.e., there are differences in math scaled scores across class sizes in first grade. Moreover, small classes (about 15-17 students) in first grade are statistically found to be beneficial to student achievement (in this case, highest math scaled scores). These findings can be used as an evidence to support the class-size reduction policy for school improvement.

Acknowledgement

I am grateful to all of those with whom I have discussed this project, Matthew Chen and Jasper Tsai.

Reference

[1] Achilles, C. M., et al. (2012). Class-Size Policy: The STAR Experiment and Related Class-Size Studies. NCPEA Policy Brief. Vol. 1, No. 2.

[2] Finn, J. D., Gerber, S. B. & Boyd-Zaharias, J. (2005). Small classes in the early grades, academic achievement, and graduating from high school. Journal of Educational Psychology, 97(2), 214-223.

Session info

```
# sessionInfo()
```