Project 2: Investigation of Class Size and School ID on First Grade Math Scores with Teacher as a unit from Project STAR

1/31/2020

Introduction

Project STAR (Student/Teacher Achievement Ratio) was a longitudinal, randomized study to investigate how class-size affected student math and reading performance in the 1980s. The study included 79 schools and over 7,000 students in Tennessee. There were three treatment groups: small class (13-17 students), regular class (22-25 students), and regular class with aide. Teachers and students were randomized to treatment groups to limit bias and confounding of the treatment effect. Students were followed for grades kindergarten through 3rd grade.

The collected data include records of 11,598 students. The variables associated with each student include demographics, school information, teacher information, treatment group (class type), and test scores for each grade. The data was retrieved from the "AER" R package. In this analysis, the teachers will be considered as the unit of study with the mean class math score as the outcome. We will limit this analysis to first-grade alone. There were 340 first grade teachers in the study.

The question of interest is whether class type affects math scores. To answer this, we will perform two-way ANOVA with class mean math score as the outcome and class type and school ID as factors. This allows us to adjust for the different schools that teachers are in.

Exploratory Data Analysis

The variables associated with each teacher include: mean math score, median math score, standard deviation of maths scores, number of students in class, class type, school ID, school setting, school system ID, teacher degree level, teacher career ladder level, teacher years of experience, and teacher race. The distributions of these variables are similar across treatment groups, as expected from randomization.

In our prior project, we found that first-grade math scores were associated with a number of the variables in the datasest. For the variables that are specific to the class, all of these same associations (or lack thereof) hold in the aggregated data. The school setting is associated with math scores (Fig. 2); inner-city classes, in particular, performed worse than suburban, rural, and urban classes. 47% of classes were in a rural setting. The years of teacher experience does not appear to be associated with math scores (Fig. 1). Black teachers tended to have lower mean class scores than white teachers (Fig. 2). Teachers with master's degrees tended to have slightly higher mean math scores than those without them (Fig. 2). And finally, the class type is associated with average math score (Fig. 2); the small classes tended to have higher mean scores than the other two treatment groups.

How to aggregate math scores over the classes?

When using teachers as the unit of analysis, the first question must be how to aggregate the math scores of students in a class. The two natural options are the mean and median. The median would be more appropriate if the scores were highly skewed or had extreme outliers. In fact, the scores are fairly normally distributed and do not have many extreme outliers. Thus, for most classes, the mean and median are fairly close.

A concern with using ANOVA to analyze this data is that the variance of the mean and median of math scores is theoretically associated with the class size. The smaller the class, the greater the variance of the mean and median of the math scores. Would the mean or median do more to reduce the difference in variance across class sizes? Below are the standard deviations of the class mean and median math scores. We can see the difference in standard deviation between small and regular classes is the same for the mean and median.

Class Type	Std Dev of Mean	Std Dev of Median
Regular + aide	24.02	24.88
Regular class	23.86	25.19
Small class	26.60	28.10

More theoretically, in a simulation of normally-distributed scores in classes of size 15 and 22, the ratio of the variance of the median between the regular and small class sizes was roughly identical to the ratio of the variance of the mean (0.833 and 0.826). This suggests that neither median nor mean would be better to address the potential issue of heteroscedasticity. Because the mean is the more common measure of central tendency, we opt to use the mean math score as the outcome of interest.

Two-way ANOVA model

We are interested in examining the influence of class size and school id on average (mean) 1st grade math scores per teacher.

The following model was chosen:

$\mathsf{y}_{\mathsf{ijk}} = \mu + \alpha_{\mathsf{i}} + \beta_{\mathsf{j}} + \epsilon_{\mathsf{ijk}}$

where:

$\mathsf{y}_{\mathsf{ijk}}$ = mean teacher math scores for the 1st graders in the STAR study with i and j as the factors (class size and school id and k as each individual data point)

μ = baseline score

α_{i} = the additive effect of factor 1 (class size) on the baseline score

β_{j} = the additive effect of factor 2 (school id) on the baseline score

ϵ_{ijk} = error term

The validity of the assumptions for a two-way ANOVA may be strained here. A priori, the high number of schools (76 schools, each treated as an individual factor) does not lend to a precise conclusion from a significant result. This is simply due to the lack of data available to construct estimates for each class type – school combination (mean of 4.4 teachers per school, range: 3-12). It would be inappropriate to assume normality with so few observations within each class type – school combination. Adding to this, stratifying the math scores into the 76 schools graphically, we see there are also alarming departures from normality within many schools (Figure 4).

Therefore a model omitting the interaction term between class type and school was chosen. For this model, (1) there was no alarming departure from normality seen in a histogram of the model residuals nor drastic deviations seen in the QQ-plot (FIGUREs 5 and 6), (2) a scatter plot of the residuals versus the fitted values give no indication of heteroscedasticity (FIGURE 7), and (3) it is reasonable to assume one teacher's mean score is independent from another's. This is primarily due to the independence assumption of individual students.

Table 2: ANOVA Table

	df	Sumsquares	MeanSquares	F_value	percent
class type	2	11617	5809	20.991	5
school id	75	136833	1824	6.593	62
residuals	261	72225	277	NA	NA

We can see that school accounts for 62% of the variation in teachers' average math score. Class type accounts for 5% of the variation in average math scores.

Tukey's Post-Hoc Analysis

We now investigate the hypothesis that all class types have the same effect on the average math scaled scores for grade 1 teachers. We define our hypothesis as

H_0: All class types have equal math scaled scores mean for grade 1 teachers VS H_a: Not all class types have equal math scaled scores mean for grade 1 teachers.

From the result of our analysis of variance of the math scaled scores for the grade 1 teachers, we decide our data does not provide evidence towards the null hypothesis. Thus, we reject the null hypothesis. This implies there are significant differences in the average math scaled scores of teachers across the class types. However, which class type can be said to be different and which class types are not distinguishable? To understand this, we perform a pairwise comparison of the means of the class types. We obtain Tukey's family-wise significance level of 0.05. We have chosen Tukey's method over the Scheffe's method because it provided a more conservative confidence interval in this case. We present the results from the Scheffe's method in the table below.

Table 2: Table 3: A table showing the difference in mean of teachers' average math scaled score between class types, confidence interval fro the mean difference, and the corresponding adjusted p-values

	mean difference	confidence interval lower bound	confidence interval upper bound	Adj. P-val
Small - Regular	13.58	8.33	18.82	<2e-16
Small - Regular/aide	9.49	4.04	14.93	0.00009
Regular/aide - Regular	4.1	-1.45	9.63	0.17

From the table, we observe that there is a significant difference between the class type means. The teachers in small class types had better average math scaled scores when compared to those in the regular class type with or without an aide. Also, this analysis suggests that having teaching aides in regular classes does not provide any significant improvement in the average math scores for the teachers.

To go beyond the vast evidence that suggests teachers in the small class type do have a better math scaled score than other class types and having aides does not improve teachers' average maths scaled scores, we explore the possibility of making causal statements.

Rubin Causal Model Assumptions

1. Stable unit treatment value assumption (SUTVA): Firstly, by design, the assignment of a teacher to a particular class type does not affect the potential outcomes of the other teachers. Secondly, the average math score of a teacher does not affect the average math score of other teachers. These two observations imply that potential outcomes are invariant to the random assignment of others. Therefore, with pure random assignment of teachers in project STAR, the differences in means identify the average treatment effect (ATE) and the average treatment effect on the treated (ATET).

2. Random Assignment Conditional on Observables (Strong ignorability): One key structure in project star is the randomization of students and teachers to class types. However, it was not designed to evaluate so many other relationships between teachers and student achievement. Relationships such as race and gender could prove to confound and may have affected the results. Thus, it is quite hard to sufficiently verify there is ignorability of unconfoundedness and selection on observables. Nonetheless, the design provides a potentially compelling opportunity to take this for granted since random assignments of students and teachers should circumvent the confoundedness inherent in data on student achievements.

Considering some of the above assumptions could not be verified by design but are believed to be satisfied by project STAR, we could say potentially isolate the effects of the class types as the cause of the difference in teachers' average math scaled score performance.

Discussion

In this extension of the STAR study, we found that both class type and school indicator had a significant effect on students' first grade math scaled scores averaged at the teacher level.

Similar to the results seen from previous analysis in Project 1, the results here suggest that teachers randomly assigned to teach small class sizes had higher class-averaged scores than both the regular size and regular size + aid. Implying again that small class size is associated with higher math scores. However, this analysis also attempted to account for variation of math score due to the students' school. Surprisingly, the school indicator accounted for 62% of the variation in teachers' averaged math scores. Where the class type only accounted for around 5% of the variation in averaged math scores.

Since the STAR study was conducted across Tennessee, it is likely that the school indicator is simply a proxy for community, and thereby, for socioeconomic level. Allowing this interpretation of the school indicator, it can be suggested that a given school can have a substantially larger impact on teacher averaged math scores than a given class size.

Further analysis should explore a different, more simplistic, measure of socioeconomic class in the model in place of school indicator, such as the frequency of free lunches. This may be more informational for measuring the effect of a communities socioeconomic class has on teacher averaged math scores.

Limitations

It should also be noted that the data under analysis here may be cluster-correlated, or hierarchical in nature. That is, within each school exist numerous teachers. Since a teacher's ability or instruction style may be linked to a given school's policies at which they teach, other teachers within the same school may have correlated average math scores. Though accounting for this cluster correlation is likely negligible to overall conclusion from the results.

Appendix

Introduction

Project STAR (Student/Teacher Achievement Ratio) was a longitudinal, randomized study to investigate how class-size affected student math and reading performance in the 1980s. The study included 79 schools and over 7,000 students in Tennessee. There were three treatment groups: small class (13-17 students), regular class (22-25 students), and regular class with aide. Teachers and students were randomized to treatment groups to limit bias and confounding of the treatment effect. Students were followed for grades kindergarten through 3rd grade.

The collected data include records of 11,598 students. The variables associated with each student include demographics, school information, teacher information, treatment group (class type), and test scores for each grade. The data was retrieved from the "AER" R package. In this analysis, the teachers will be considered as the unit of study with the mean class math score as the outcome. We will limit this analysis to first-grade alone. There were 340 first grade teachers in the study.

The question of interest is whether class type affects math scores. To answer this, we will perform two-way ANOVA with class mean math score as the outcome and class type and school ID as factors. This allows us to adjust for the different schools that teachers are in.

Exploratory Data Analysis

The variables associated with each teacher include: mean math score, median math score, standard deviation of maths scores, number of students in class, class type, school ID, school setting, school system ID, teacher degree level, teacher career ladder level, teacher years of experience, and teacher race. The distributions of these variables are similar across treatment groups, as expected from randomization.

In our prior project, we found that first-grade math scores were associated with a number of the variables in the datasest. For the variables that are specific to the class, all of these same associations (or lack thereof) hold in the aggregated data. The school setting is associated with math scores (Fig. 2); inner-city classes, in particular, performed worse than suburban, rural, and urban classes. 47% of classes were in a rural setting. The years of teacher experience does not appear to be associated with math scores (Fig. 1). Black teachers tended to have lower mean class scores than white teachers (Fig. 2). Teachers with master's degrees tended to have slightly higher mean math scores than those without them (Fig. 2). And finally, the class type is associated with average math score (Fig. 2); the small classes tended to have higher mean scores than the other two treatment groups.

Fig. 1

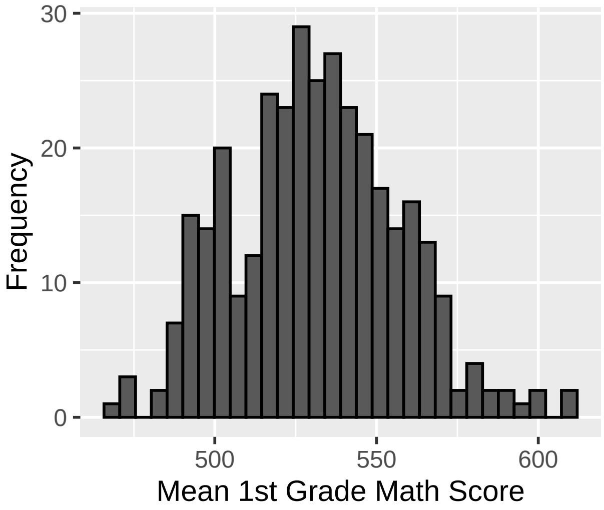


Fig. 2

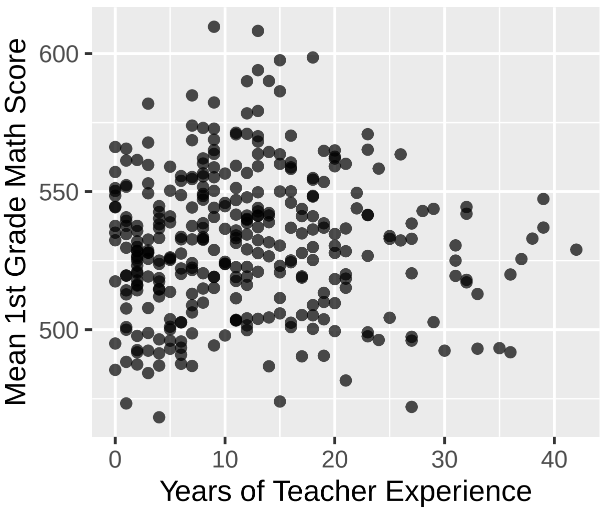


Fig. 3: Distribution of Mean 1st Grade Math Scores by Covariates

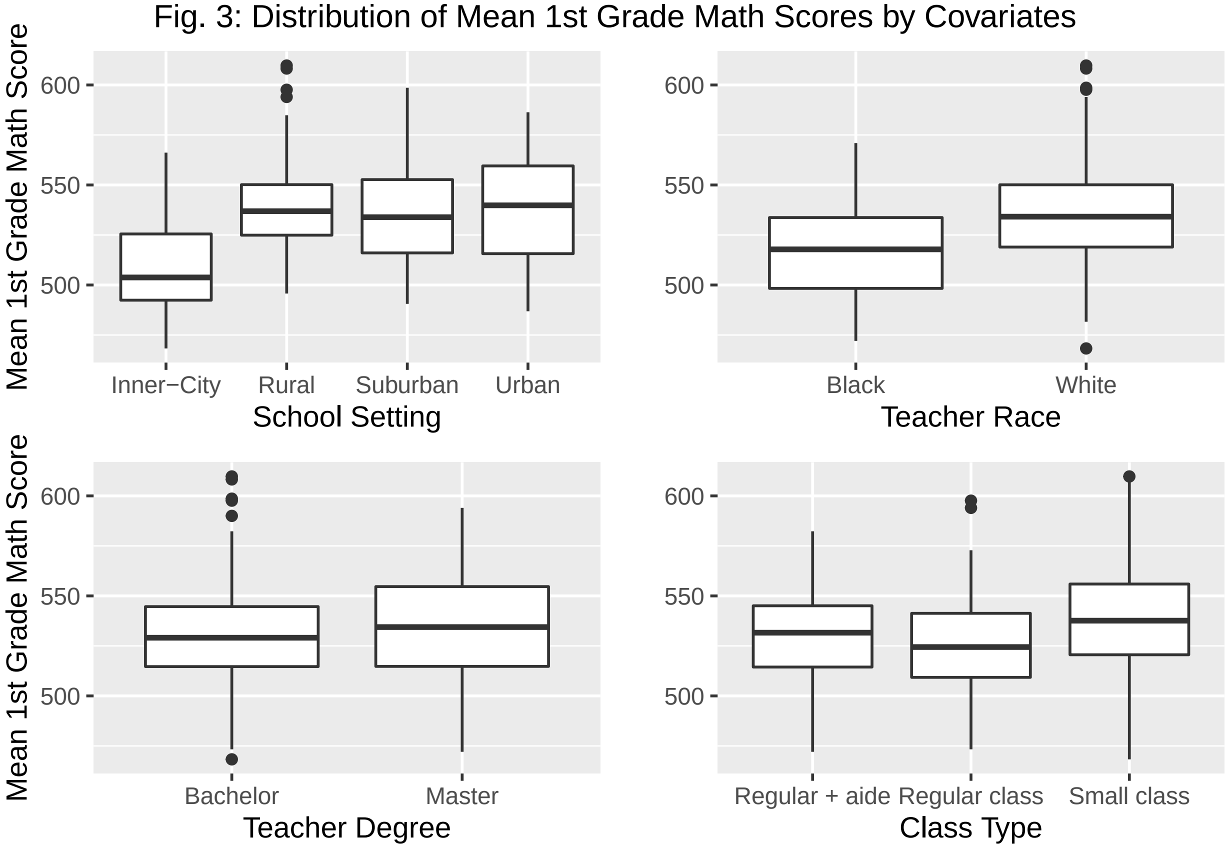
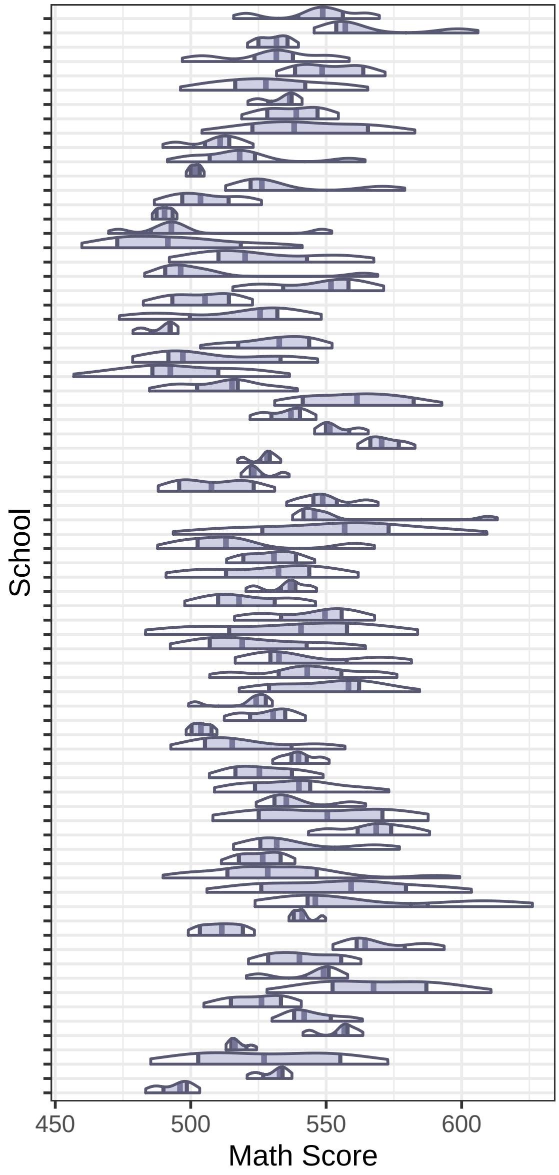


Figure 4. Density of teacher mean across schools



Each density is marked with a mean and inner two standard deviations

Figure 5. Histogram of model re

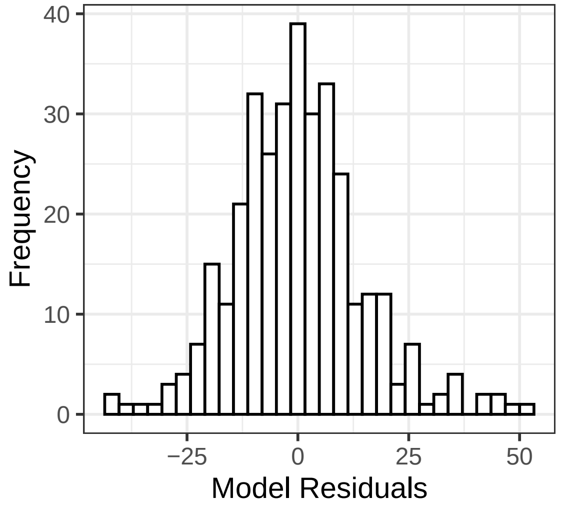


Figure 6. Normal Quantile−Quar

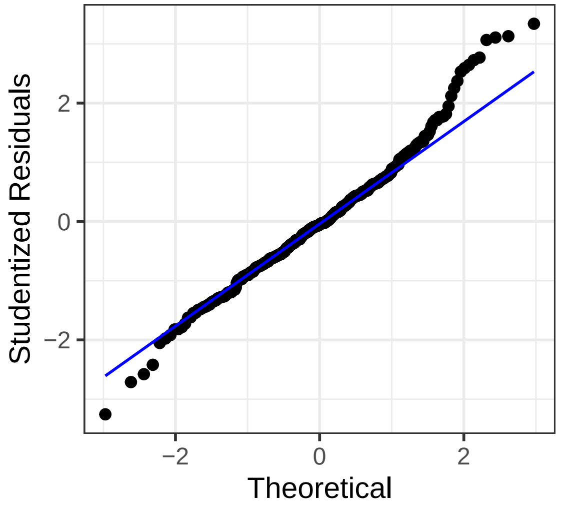
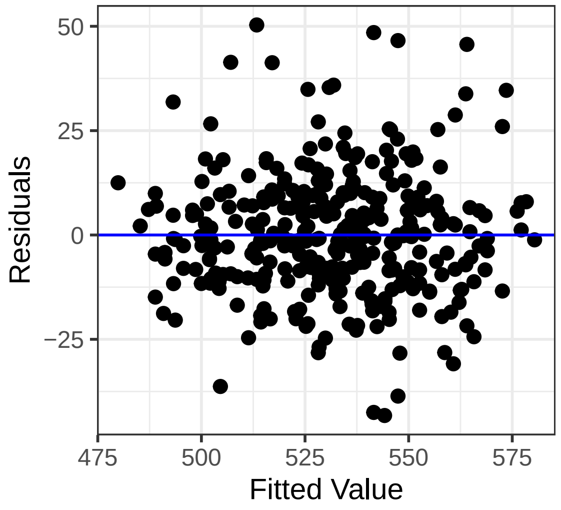


Figure 7. Residuals Versus Fit


Residuals
Fitted Value