Bayesian Multi-level Analysis on Student Teacher Achievement Ratio (STAR) Experiment in Tennessee

Candidate Number: 23861

LSE ID: 202214446

London School of Economics and Political Science

ST308: Bayesian Inference

4 May 2024

I. Introduction

The report uses the Student Teacher Achievement Ratio experiment in Tennessee to study the relationship between class sizes and students' academic performance. Firstly, the report will provide a literature review of previous studies on class size effects and their implications. Secondly, the report will describe the data in use, including its structure, chosen variables, and institutional backgrounds. Finally, the report will conduct statistical analysis using competing methods at both the single and multi-levels to understand the effects of class size on students' performance.

II. Literature Review

Studies of class size on children's learning have been conducted throughout the twentieth century, involving nearly a million students. Early theoretical frameworks, such as the Effective Schools movement, suggested that "all children can learn and that the school controls the factors necessary to assure student mastery of the core curriculum" (Lezotte, 2001). In particular, smaller classes could enable more individualized attention and better classroom management, potentially enhancing academic achievement. Some programs have been implemented by governments around the world; for example, the CSR Program in California aims to improve students' performance by reducing class sizes throughout the state to roughly 20 students per class. However, while reducing class sizes may offer potential benefits, these programs can be costly, with an annual cost exceeding 10 billion dollars. Policymakers must carefully consider the cost-effectiveness of such initiatives and prioritize evidence-based approaches to educational improvement. Thus, it's out of interest to figure out whether the class size reduction is indeed associated with higher average student achievement.

Previous research conducted by Angrist and Lavy (1999), Hoxby (2000), and Rivkin, Hanushek, and Kain (2005) identified a positive effect of varying class size and student-teacher ratio on students' achievement by utilizing an exogeneous variation in class size, and studies found a significant effect of class size on achievements in early grades. Krueger (1999) and Krueger and Whitmore (2001) also find that smaller class sizes in kindergarten and first grade have a significant and lasting impact on academic achievement and educational attainment. Contextual factors such as school resources, teaching quality, and socio-economic status play a crucial role in shaping the impact of class size on student performance as well.

Our analysis extends previous work to look at the effect of class size using a Bayesian multi-level model. Multiple specifications in conventional and Bayesian fixed effects and a pooled model of single and multi-level provide a far richer set of estimates to comprehensively identify the potential correlations between class size and students' performance.

III. Data

The Tennessee STAR experiment is a large-scale randomized experiment in Tennessee to study the effect of smaller classes, costing $12 million. The study population was the cohort of kindergarteners in 1985–86, or about 11,600 children. The experiment lasted 4 years until the

kindergarteners were in grade 3, but it should be noted that the data is not panel since not all individuals are repeatedly recorded in all 4 grades. The average class size in Tennessee classes in 1985/86 was 22.3. Experimental assignments include small classes with 13–17 children, regular classes with a full-time teacher aside, and control regular classes with 22–25 children. The average difference between regular and small classes is 7.5 children. The outcome is measured in numerical and reading ability by the Stanford Achievement Test and the Tennessee Basic Skills First Test given in March of each school year.

The data can be viewed as hierarchical sampling by first sampling schools and then students within each school. For each student, we choose the following variables: 1) student ID; 2) school ID; 3) class type; 4) math score.

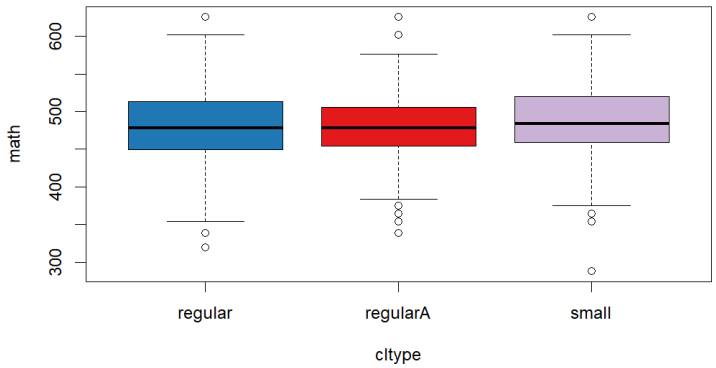
Based on the literature, the class size effects are more significant in lower grades. Given the large scale of the whole dataset, we subset the data to focus on the group of students in kindergarten and thus study the association between class size and students' performance among kindergarteners.

Overall, the dataset in use contains 6325 observations of four variables. Each observation represents a kindergartener in one of the 80 schools in Tennessee, randomly assigned to a small class, a regular class, or a regular class with a full-time teacher aside. Their performance is gauged by their numeracy ability.

IV. Explanatory Data Analysis

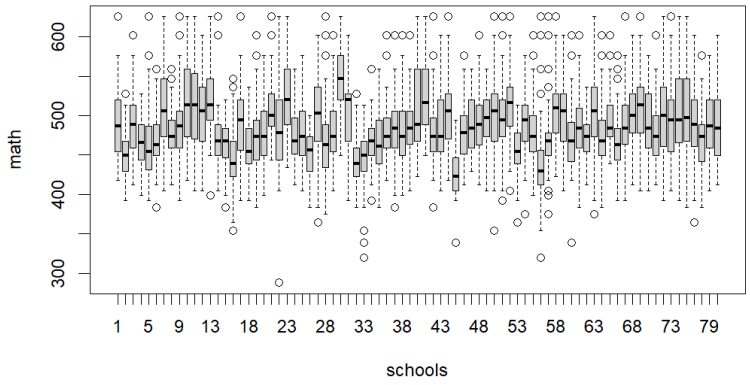
In this section, we perform the exploratory data analysis into descriptive statistics and data visualization to lay the groundwork for a comprehensive understanding of the factors influencing student outcomes.

Figure 1 – Distribution of math scores across class types



The boxplot shows how students' math scores are distributed across different class sizes. For both math score, no significant differences are observed in different class types. However, since students are sampled from different schools. Some intrinsic school characteristics can potentially counteract each other when pooling them together. If looking at the distribution of math score in different schools, as shown in Figure 2, the statistics presents some degree of variability across schools. Thus, levels in schools should be considered when conducting the analysis.

Figure 2 – Distribution of math scores across schools



V. Linear Regression Models

We first fit the simple pooled linear regression models using frequentist approaches based on MLE to examine whether class size is associated with performance.

We run the regressions of the form below as Model 1

$$y_{ij} = a + bX_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma_y^2)$$

where y_{ij} captures the students' math scores and X_{ij} is a categorical variable of class type yielding small, regular, and regular with a full-time teacher aside.

From column (1) in Table 1, the constant term indicates that the average math score for students in regular class size is 483.1993. The coefficient for small class size is 7.732, suggesting that the average difference in math score between a student in regular class size and small class size is 7.732, i.e., the math scores for students in small class size are average 7.732 points higher than those in regular class size. The estimate is statistically significant at 1% significance level. In contrast, the coefficient for regular classes with a teacher aside is -0.403, which not statistically significant at any conventional levels, suggesting that we fail to reject the null that there is no difference in math scores between regular-size classes with and without a teacher aside.

However, this naïve regression does not take the variability across schools into account, which could be a huge problem based on the boxplots of score distributions in schools in section IV. Thus, a fixed-effect model may be more suitable in this context.

We run the regressions of the form below as Model 2

$$y_{ij} = a_i + bX_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma_y^2)$$

where y_{ij} captures the students' math scores, X_{ij} is a categorical variable of class type yielding small, regular, and regular with a full-time teacher aside, and a separate intercept for different schools.

The results are quite consistent with the pooled model. Being in a small class had a statistically significant positive effect on math scores. Specifically, students in small-size classes score approximately 8.93 points higher in math compared to students in regular-size classes in the same school. The treatment of regular classes with a teacher aside still shows no significant effect on students' performance.

Table 1: Linear Regression: key results

	Dependent variable:	
	math	
	(1)	(2)
cltypereg+A	−0.403	0.276
	(1.484)	(1.350)
cltypesmall	7.732***	8.932***
	(1.548)	(1.407)
Constant	483.199***	490.434***
	(1.055)	(5.270)
Fixed Effect	No	Yes
Observations	5,871	5,871
R^2	0.006	0.215
Adjusted R^2	0.005	0.204
Residual Std. Error	47.567 (df = 5868)	42.556 (df = 5790)
F Statistic	17.200*** (df = 2; 5868)	19.805*** (df = 80; 5790)

*p<0.1; **p<0.05; ***p<0.01

VI. Bayesian Linear Regression Models – Single Level

In previous frequentist fixed-effect model, there are some school categories where we don't have enough data points to make valid inferences. In this case, Bayesian linear regression offers several advantages for modelling relationships. It allows for the incorporation of prior knowledge and can accommodate hierarchical data structures in our dataset, which enables us to obtain more reliable estimates.

Prior beliefs refer in generality to the information content of the prior distribution. Suppose we believe, prior to seeing the data, that a and b are probably close to zero, and as likely to be positive as they are to be negative but have a small chance of being quite far from zero. These beliefs can be represented by normal distributions with a proper variance in order to produce moderately heavy tails. Previous research, both on other programs and some on STAR experiments, show a magnitude of effects ranging from 0 to 15 (Angrist and Lavy, 1999; Hoxby, 2000; Rivkin, Hanushek, and Kain, 2005). As we cannot rule out the possibly of having insignificant and negative effects, it would be sensible to set the prior distribution of the coefficient as $N(0, 15^2)$. Since we have no reliable information on the range of the intercept, we will use a non-informative large variance prior distribution of $N(0, 150^2)$ after centring. The standard deviation of this prior distribution is three times as large as the standard deviation

of the response, and thus should be a close approximation to a non-informative prior over the range supported by the likelihood, which should give similar inferences to those obtained by MLE if similarly non-informative priors are used for the other parameters.

Conducting Bayesian analysis using a pooled model (Model 1 in section V) in these priors gives results shown in Table 2. The posterior mean of small class size is statistically significant from 0 as the 95% credible interval does not contain 0, suggesting that being in a small class is associated with 7.7 score high in math compared to being in a regular class. On the other hand, the posterior mean of regular class with a full-time teacher aside is still not significant at 5% level. This suggests that there is no statistically significant difference in math scores between students in regular classes and those in regular classes with a full-time teacher aside.

The model is valid in terms of Rhat, which is a metric that measures whether the Markov chains are stable and converged to a value during the total progress of the sampling procedure. The Rhat statistic equals one for all parameters, which signals that all chains have converged to an equilibrium distribution. Using MCMC sampling implies that we get dependent samples from the posterior distribution. Inspecting the autocorrelation plot suggest dependence in previous variables but then goes to zero, which allows us to drop the burn-in period and use data after to get independent samples.

Table 2: Posterior Summary: pooled model

Parameter	Rhat	mean	sd	2.5%	97.5%
(Intercept)	1.0	483.2	1.1	481.1	485.3
cltypereg+A	1.0	-0.4	1.5	-3.4	2.4
cltypesmall	1.0	7.7	1.5	4.6	10.7
sigma	1.0	47.6	0.4	46.8	48.4

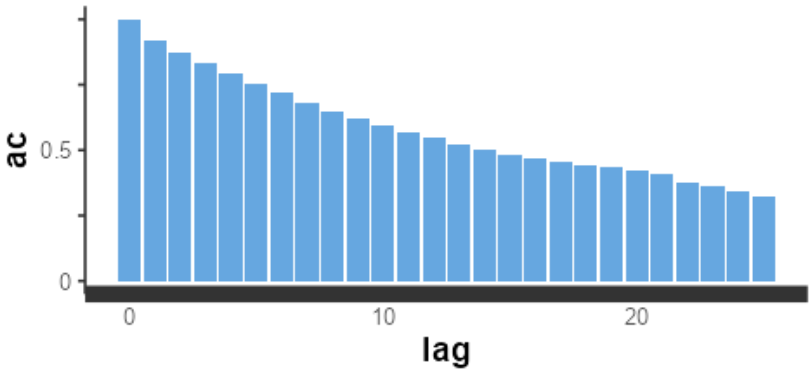
In contrast, if we perform the same analysis on the fixed-effect model (Model 1 in section V), the model turns to be not as efficient as the pooled one. Since we add fixed effects into the model, the prior distribution for coefficient as $N(0, 15^2)$ is no longer appropriate. Thus, we use non-informative prior distributions $N(0, 150^2)$ specified above for both intercept and coefficient.

The results in Table 3 shows that effective samples size of 134 is too low, indicating posterior means and medians may be unreliable. Further, the autocorrelation plot (Figure 3) signals the presence of autocorrelation and the chains have not mixed with some convergence issues. This is reasonable since the model has too many parameters. Thus, we need to find a model that could better analyse our data.

Table 3: Posterior Summary: fixed-effect model

Parameter	n_eff	mean	sd	2.5%	97.5%
(Intercept)	134	489.8	4.8	481.0	498.9
cltypereg+A	3403	0.3	1.3	-2.4	2.9
cltypesmall	3801	8.9	1.4	6.1	11.7
sigma	5969	42.6	0.4	41.8	43.4

Figure 3: Autocorrelation plot



VII. Bayesian Linear Regression Models – Multi Level

Since the data exhibits a hierarchical arrangement where units of analysis (students) are nested within higher organizational clusters (schools), it's natural to suspect this hierarchical setup introduces dependence among the observed responses of units within the same cluster, i.e., students within the same school exhibit more similarity in their academic attributes compared to students selected randomly from the broader population.

Multi-level models are helpful to model such within-cluster dependence. In our context, a two-level model that allows for grouping of student outcomes within schools would include residuals at both the student and school level, which effectively partitions the residual variance into the between-school variability, reflecting the variance of school-level residuals, and the within-school variability, reflecting the variance of student-level residuals.

This could lead us to a multi-level model with following specifications

$$y_{ij} \sim \mathrm{N}(a_i + bX_{ij}, \sigma_y^2)$$

$$a_i \sim N(\mu_a, \sigma_a^2)$$

where y_{ij} captures the students' math scores, X_{ij} is a categorical variable of class type yielding small, regular, and regular with a full-time teacher aside. The intercepts come from the normal distribution.

In full Bayesian inference, all the hyperparameters (μ_a, σ_y, σ_a) also need a prior distribution. We keep our prior for μ_a as $N(0, 150^2)$ and for b as $N(0, 15^2)$. For the rest of the hyperparameters, we keep it as the default options of rstanarm to get a reasonable large variance prior.

This model is superior because it not only estimates a specific intercept for a school but will also consider the connections among schools, so a common mean could be used when not enough information is available for a specific school. It also provides information on the variability across schools.

Table 4: Posterior Summary: multi-level model

Parameter	Rhat	mean	sd	2.5%	97.5%
(Intercept)	1.0	482.7	2.6	477.5	487.7
cltypereg+A	1.0	0.2	1.3	-2.4	2.9
cltypesmall	1.0	8.8	1.4	6.0	11.5
sigma	1.0	42.6	0.4	41.8	43.4
Sigma[sch:(Intercept),(Intercept)]	1.0	478.8	83.4	340.2	667.7

The results for the Bayesian multi-level model are presented in Table 4. The point estimate for μ_a is 482.7 and the point estimates for the vector of coefficient μ_{b} are 0.2 and 8.8 respectively. In the context, the difference between students' math scores in small class and regular class is statistically significant from 0 since the 95% credible interval excludes 0, with a mean value increase in students' math score of 8.9 points, but the difference in math scores between regular class with and without a full-time teacher aside is still not significant.

Sigma and Sigma[sch:(Intercept),(Intercept)] correspond to σ_{y} and σ_{a}, which represent within-school and between-school variability respectively. The mean of the variability across intercepts is 478.8 points, which signals quite varying distribution for students' performance in regular class (original control group without treatment) across schools, justifying our application of a multi-level model.

Additionally, the prior choice is robust to sensitivity checks. Table 5 gives a posterior summary of the same multi-level but uses a larger variance prior $N(0, 500^2)$, and the results turn to pretty much resemble the one given above.

Table 5: Posterior Summary: multi-level model (Sensitivity Checks)

Parameter	Rhat	mean	sd	2.5%	97.5%
(Intercept)	1.0	482.6	2.8	477.2	487.9
cltypereg+A	1.0	0.3	1.3	-2.4	3.0
cltypesmall	1.0	8.9	1.4	6.0	11.6
sigma	1.0	42.6	0.4	41.8	43.4
Sigma[sch:(Intercept),(Intercept)]	1.0	476.6	83.7	338.9	662.2

We may also interest in considering a multi-level model with varying intercepts and slopes. This could be achieved through the model specifications

$$y_{ij} \sim \mathrm{N}(a_i + b_{\mathrm{i}} X_{ij}, \sigma_y^2)$$

$$a_i \sim N(\mu_a, \sigma_a^2)$$

$$b_{\mathrm{i}} \sim N(\mu_b, \sigma_b^2)$$

where y_{ij} captures the students' math scores, X_{ij} is a categorical variable of class type yielding small, regular, and regular with a full-time teacher aside. The intercept and slopes come from normal distributions, so we have a population intercept and a population slope and at the same time estimate the variability both in terms of the intercept and the slopes

Table 6: Posterior Summary: multi-level model (varying slopes and intercepts)

Parameter	Rhat	mean	sd	2.5%	97.5%
(Intercept)	1.0	483.0	3.1	477.0	489.3
cltypereg+A	1.0	-0.0	2.6	-5.0	5.0
cltypesmall	1.0	8.4	2.8	3.1	13.7
sigma	1.0	41.1	0.4	40.4	41.9

The results shown in Table 6 are still pretty similar to the model with only varying intercepts. The posterior mean for regular class size with a full-time teacher aside is slightly different, but it's sensible considering its insignificance nature.

In fact, if we compare the results for all models in this paper, the results don't vary much, even when comparing those results from the very basic linear regression model and those from the Bayesian multi-level model. All of them yield a significant difference between students' math performance in a small class and a regular class and no significant difference between those in a regular class with and without a full-time teacher aside. This suggests that the data is strong and dominates, so it quickly overwhelms the prior and gives a similar result every time, which is sensible given the nature of this dataset being a random control trial.

VII. Concluding Remarks

This paper utilises linear regression models, Bayesian linear regression models, and Bayesian multi-level models to systematically test the association between class size and students' math scores. A significant difference between small class size and regular class size is identified, as opposed to the insignificance difference for math scores between regular class size with and without a full-time teacher aside. The results are similar across all models, highly likely due to the domination of the data considering its nature of being a random control trial.