Application of Multilevel Models in Education Research: Effect of Reduced Class Sizes for Student Achievement

Yuwei Sun, Cheng Lu, Yifan Li

May 8, 2018

1 Abstract

Project STAR (Student/Teacher Achievement Ratio) was a four-year educational reform experiment conducted from 1985-1989 by the state of Tennessee. It was intended to investigate the effect of reduced class sizes on student achievement. In this project, we analyzed a subset of the data from STAR project and addressed some specific questions related to impacting factors of student achievement as well as a longitudinal analysis on the effect magnitude of reduced class size. Due to dependency of each data point, we utilized multilevel modelling techniques to conduct our analysis. Multilevel models are fitted by lmerTest package in R as well as the probabilistic programming language Stan, after which we compared the results and made inference about questions of interest.

2 Introduction

In STAR project, the 80 participating elementary schools throughout the state randomly assigned students entering kindergarten to one of three class types: small (S) with 13-17 pupils; regular (R) with 22-26 pupils or regular with a full-time teaching aide (RA) with 22- 26 pupils. The study lasted for four consecutive years and students information were recorded each year from kindergarten to third grade. Participating schools represent different geographic regions and different communities (i.e. rural, urban, suburban, inner city). Teachers have various backgrounds with respect to levels of education degree/experience, race etc. Students also came from various demographic and socioeconomic backgrounds.

The STAR database is extremely huge and there exist many opportunities for different analyses using all or different portions of the data. We have to select variables among 14 independent variables to fit model efficiently and to determine which variable should be the random effect. In our project we used student level as well as school level data to study the effects of a reduced pupil-teacher ratio on students read and math test scores.

3 Research Questions

Our major interest includes two aspects: how students achievements are affected by reduced class size and how the effect varies among other factors (socioeconomic status etc.); what other factors are associated with students achievements. Specifally speaking, Our goal is to find the relation between mean of math scores for four years and the class type. That is, does the class type affect the students performance in math exam?

4 Data

We obtained the data from the mlmRev R package, which contains data and examples from a multilevel modelling software review as well as other well-known data sets from the multilevel modelling literature.

The dependent variables are math score and read score. The independent variables include student variable such as sex, race, socioeconomic status, teacher variable such as race, education level, race, school type and class type. The raw data contains 26,796 observations, with 80 schools and 11598 students in total. However, several variables have a tiny portion of missing values ($< 10\%$), in this scenario we can assume the data is missing at random, thus we removed observations that contain missing value. As a consequence, we have a total of $26,796 - 3981 = 22815$ data points. The first few lines from the data frame are shown in Table 1.

Table 1: Head of Data Frame star

id	sch	gr	cltype	hdeg	clad	exp	trace	read	...
100017	28	K	small	BS/BA	1	3	B	476	
100028	52	K	reg	MS/MA/MEd	1	12	W	410	
100045	41	1	small	BS/BA	1	20	W	507	
100045	41	2	small	MS/MA/MEd	APPR	15	B	575	
100045	41	3	small	BS/BA	1	5	W	610	

Here is a detailed explanation of all the variables:

1 **Outcome variables**:

- **read**: the students total reading scaled score
- **math**: the students total math scaled score

2 **Student variables**:

- **id**: a factor - student id number
- **ses**: socioeconomic status - a factor with levels F and N representing eligible for free lunches or not eligible
- **sx**: students sex - a factor with levels (M, F)
- **eth**: students ethnicity - a factor with the same levels as trace

- **birthq**: students birth quarter - an ordered factor with levels $(1977 : 1 < \cdots < 1982 : 2)$
- **birthy**: students birth year - an ordered factor with levels $(1977 < \cdots < 1982)$
- **yrs**: number of years of schooling for the student - a numeric version of the grade gr with Kindergarten represented as 0

3 **School variable**:

- **sch**: a factor - school id number gr grade - an ordered factor with levels $(K < 1 < 2 < 3)$
- **schtype**: school type - a factor with levels (inner, suburb, rural and urban)

4 **Class variable**:

- **cltype**: class type - a factor with levels small, reg and reg+A. The last level indicates a regular class size with a teachers aide.

5 **Teacher variable**:

- **tch**: a factor - teacher id number
- **hdeg**: highest degree obtained by the teacher - an ordered factor with levels $(ASSOC < BS/BA < MS/MA/MEd < MA+ < Ed.S < Ed.D/Ph.D)$
- **clad**: areer ladder position of the teacher - a factor with levels (NOT, APPR, PROB, PEND, 1, 2, 3)
- **exp**: a numeric vector - the total number of years of experience of the teacher
- **trace**: teachers race - a factor with levels (W, B, A, H, I, O) representing (white, black, Asian, Hispanic, Indian (Native American), other)

Summary tables for numerical variables and some important categorical variables are listed below.

Table 2: summary table for numerical variables

Numeric	exp	read	math	yrs
Min	0.00	315.0	320.0	0.000
1st Qu	6.00	470.0	506.0	1.000
Median	11.00	553.0	558.0	1.000
Mean	12.13	541.5	555.3	1.521
3rd Qu	17.00	604.0	603.0	3.000
Max	42.00	775.0	774.0	3.000

5 Data Preparation and Visualization

The whole data is complicated. The total consecutive years students were followed vary from 1 to 4 years. Some students change class type even change school during different grade, with different teachers from various backgrounds. Thus, the variance source can be complicated and its hard to consider all the potential interactions. In addition, we have student level

Table 3: summary table for some categorical variables

sch	cltype	schtype	eth	ses	sx
55: 744	small: 6899	inner: 4791	W: 15150	F: 11120	M: 11763
22: 492	reg: 7786	suburb: 547	B: 7539	N:11695	F: 11052
63: 480	reg+A: 8130	rural: 10759	A: 59		
9: 461		urban: 1791	H: 26		
27: 453			I: 8		
7: 445			O: 33		
Other: 19740					

and school level as random effects, where student is nested within each school.

In order to reduce the magnitude of statistical computations, for this part we took a subset of observations where students stayed in the same class type and in the same school for four consecutive years (K-1-2-3), after which data were aggregated to the level of individual student means of 4 years test score. Due to aggregation, we only kept independent variables which are constant for each student among the four years: ses, sx, eth, birthq, birthy, cltype, sch, schtype. This subset contains 1124 observations, with 69 schools in total.

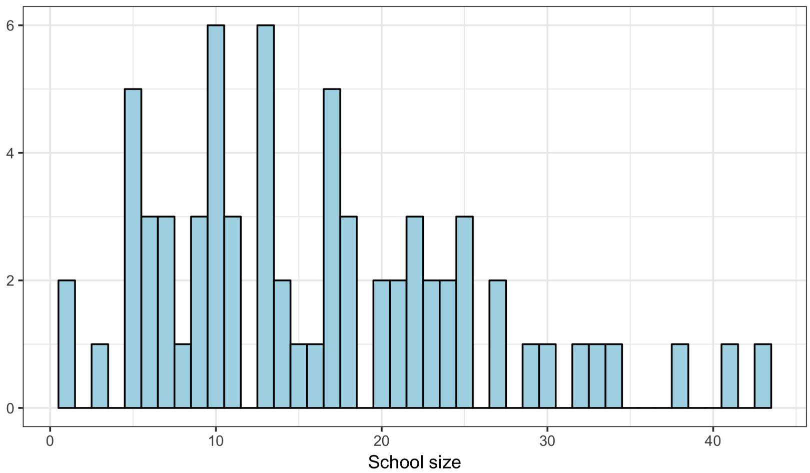

Figure 1: School Size

From Fiture 1, the total number of students in different schools is imbalanced. Notice that there exists two schools containing only one student, taking variable 'sch' as a random effect is a better choice.

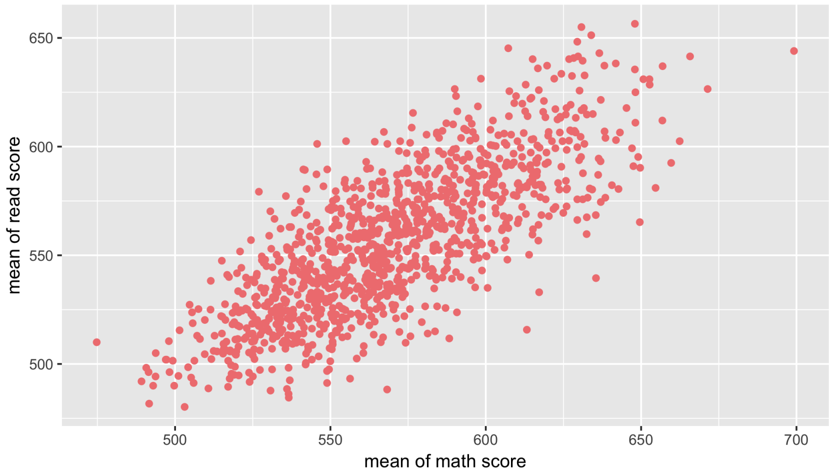

Figure 2: math vs read

From Fiture 2, we can see that there is a obvious linear relation between the mean of math score and the mean of read score, which means we can choose either of the two to be our dependent variable. We just choose the mean of math score to be our response.

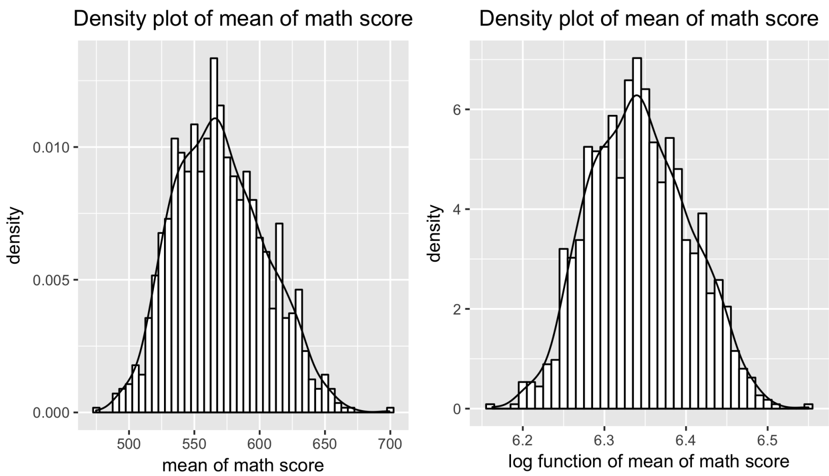

Figure 3: Distribution of math

From Fiture 3, the distribution of math score is close to normal distribution. And after log transformation, the assumption of normal distribution does not improve much. So math is taken as dependent variable without transformation.

Next, we draw several plots to explore the relationship between the mean of math score and the other variables.

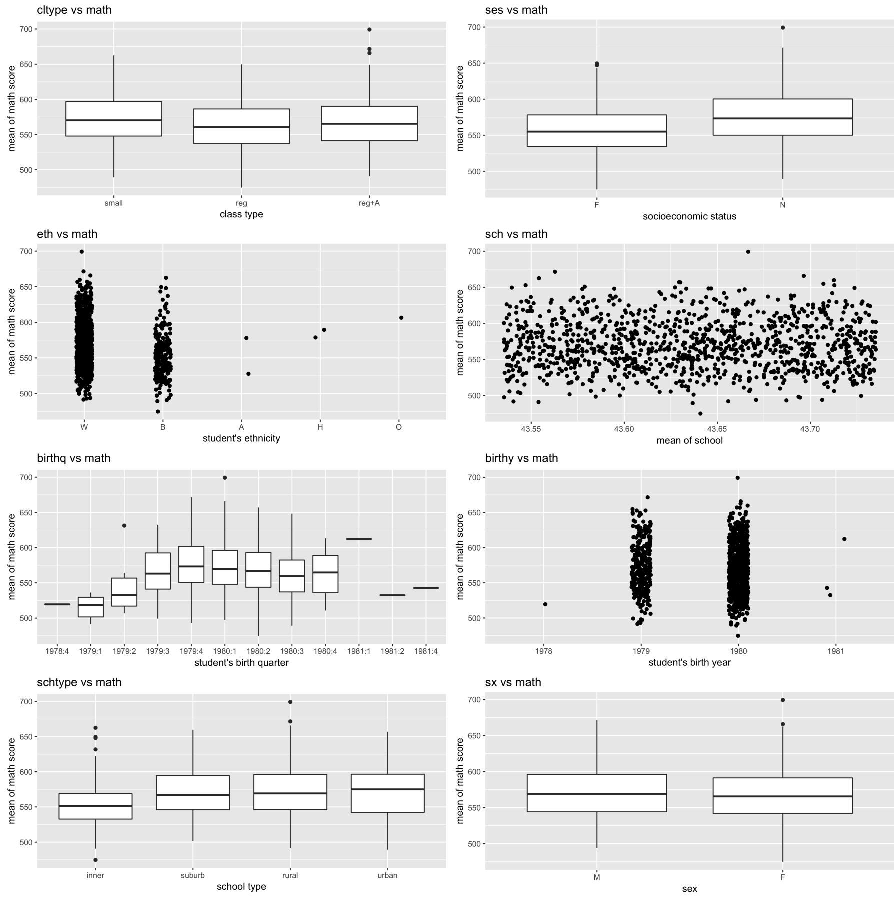

Figure 4: Effect of Each Variable

Variables like 'ses', 'eth', 'schtype' obviously affect the mean of math score while 'sex' and 'birthy' doesnt affect much. Situation of birthq is more complicated. If quarterly has an effect on score, its influence should be periodic. However, the plot between 'birthq' and 'math' shows that this effect seem to be random. So we don't include 'birthq' at the beginning. Other variables should be examined by model using F-test which can hardly be judged by eyes.

6 Modeling

The fixed effects model is inappropriate for this task because it does not take into account the fact that the students are grouped by schools. It leads to a violation of the independence of errors assumption. In linear mixed models, we take this by-school variability into account by adding adjustment terms μ_{0j} which adjust β_0 for school j. In our analysis a series of random intercept mixed effects models are fitted, with school as random effect and other predictor variables as fixed effects.

6.1 Model 0

First, we build a basic model. Only variable 'sch' is considered as a random effect and all other variables including 'ses', 'eth', 'cltype', 'sx', 'schtype' as fixed effects. The basic model is in this form:

$$
\begin{aligned}
math_i &= \beta_0 + \beta_1 * I_{cltype_i='reg'} + \beta_2 * I_{cltype_i='reg+A'} + \beta_3 * I_{ses_i='N'} \\
&\quad + \beta_4 * I_{eth_i='B'} + \beta_5 * I_{eth_i='A'} + \beta_6 * I_{eth_i='H'} + \beta_7 * I_{eth_i='O'} \\
&\quad + \beta_8 * I_{se_i='F'} + \beta_9 * I_{schtype_i='suburb'} + \beta_{10} * I_{schtype_i='rural'} + \beta_{11} * I_{schtype_i='urban'} \\
&\quad + sch_{j[i]} + \sigma_e \\
sch_j &\sim N(0, \sigma_{sch})
\end{aligned}
$$

6.2 Model 1

Since the basic model contains all possible important variables, the next step is to drop the variables that are not significant. Backward elimination method here is used to delete variables one by one and the criterion for deleting variables is the F-test. Through 'step' function in 'lmerTest' R-package, three variables 'sx', 'schtype' are considered as insignificant, and the new model looks like:

$$
\begin{aligned}
math_i &= \beta_0 + \beta_1 * I_{cltype_i='reg'} + \beta_2 * I_{cltype_i='reg+A'} + \beta_3 * I_{ses_i='N'} \\
&\quad + \beta_4 * I_{eth_i='B'} + \beta_5 * I_{eth_i='A'} + \beta_6 * I_{eth_i='H'} + \beta_7 * I_{eth_i='O'} \\
&\quad + sch_{j[i]} + \sigma_e \\
sch_j &\sim N(0, \sigma_{sch})
\end{aligned}
$$

6.3 Model 2

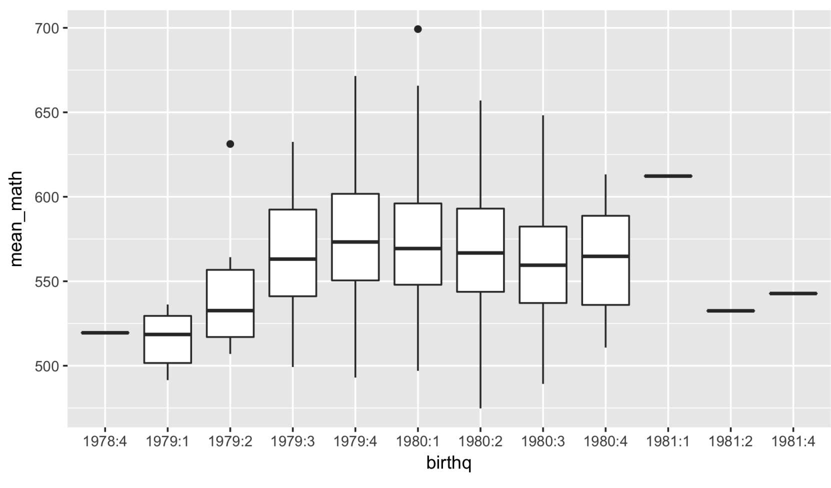

Figure 5: birthq vs math

Form Figure 5, we find that variable 'birthq' has effect on the math score. However, if 'birthq' is a fixed effect, its impact on the math score should be cyclical. There is no reason to believe that children born in the first quarter are smarter than children born in the second quarter of 1980, while the situation in 1979 was the opposite. Because of this, variable 'birthq' is added in the model as a random effect. And our model becomes:

$$
\begin{aligned}
math_i &= \beta_0 + \beta_1 * I_{cltype_i='reg'} + \beta_2 * I_{cltype_i='reg+A'} + \beta_3 * I_{ses_i='N'} \\
&\quad + \beta_4 * I_{eth_i='B'} + \beta_5 * I_{eth_i='A'} + \beta_6 * I_{eth_i='H'} + \beta_7 * I_{eth_i='O'} \\
&\quad + sch_{j[i]} + birthq_{k[i]} + \sigma_e \\
sch_j &\sim N(0, \sigma_{sch}) \\
birthq_k &\sim N(0, \sigma_{birthq})
\end{aligned}
$$

6.4 Model 3

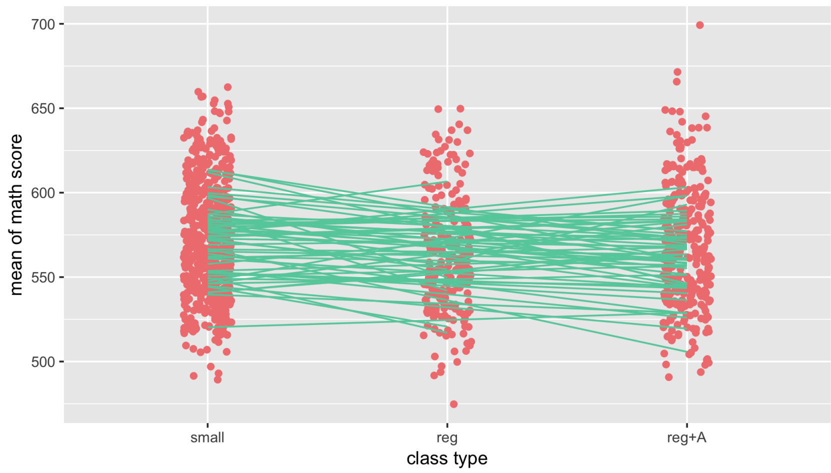

Figure 6: The influence of schools on the relationship between cltype and math

From the plot we can see that the slopes of straight lines vary among different schools. Class type affects the mean of math score, but the degrees of the effect are different among schools. Thus, take school as random slope may be a good choice. The model is shown below:

$$
\begin{aligned}
math_i &= \beta_{0i} + \beta_{1i} * I_{cltype_i='reg'} + \beta_{2i} * I_{cltype_i='reg+A'} + \beta_3 * I_{ses_i='N'} \\
&\quad + \beta_4 * I_{eth_i='B'} + \beta_5 * I_{eth_i='A'} + \beta_6 * I_{eth_i='H'} + \beta_7 * I_{eth_i='O'} + \sigma_e \\
\beta_{0i} &\sim N(sch_{j[i]}, \sigma_2) \\
\beta_{1i} &\sim N(sch_{j[i]}, \sigma_2) \\
\beta_{2i} &\sim N(sch_{j[i]}, \sigma_2)
\end{aligned}
$$

6.5 Model Selection

After building all four models with stan, we find that all estimations converge. Then we use BIC method to select the final model. The result is Model 2 $<$ 3 $<<$ 1 $<$ 0, and since Model 2 is simpler than Model 3, we choose Model 2 as our final model.

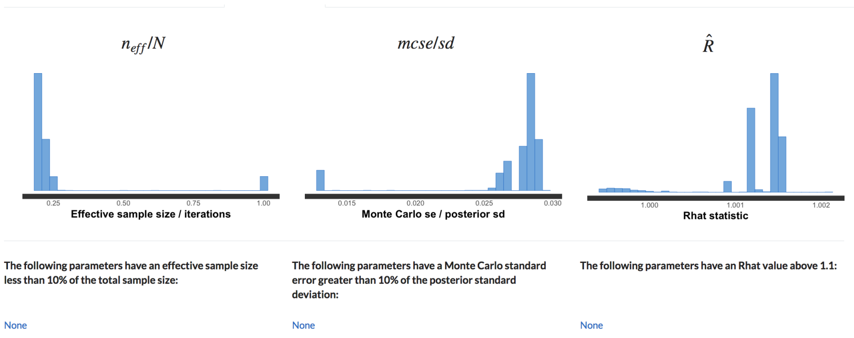

Figure 7: Diagnostic of Model 2

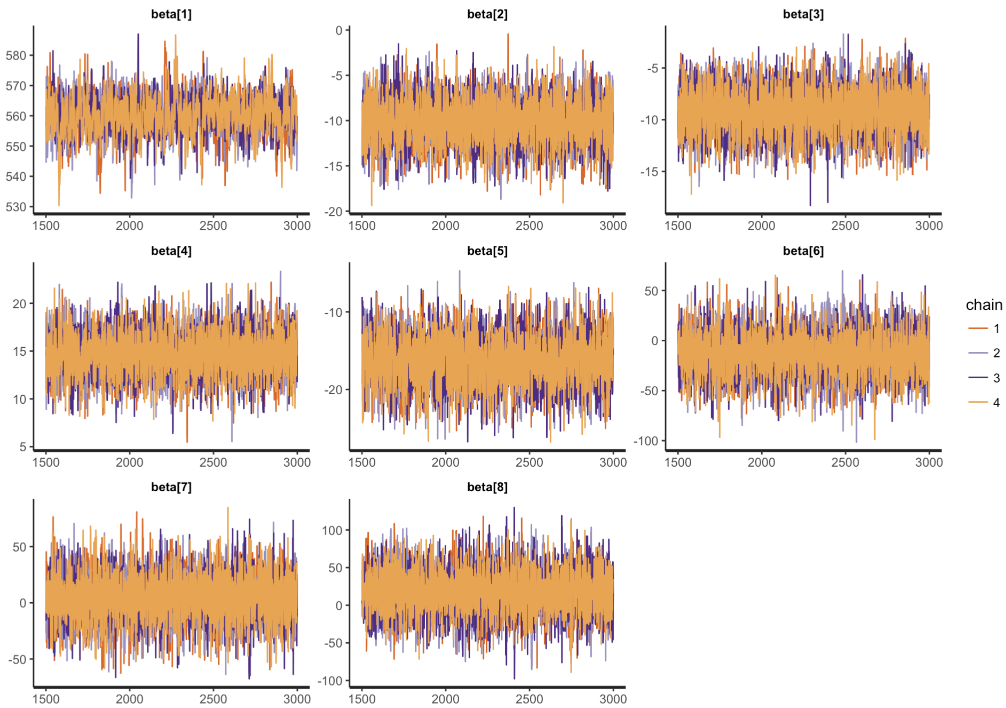

Figure 8: Traceplot of Model 2

From Figure 7 and Figure 8, $\hat{R}$ and traceplot shows that our model converges. The posterior distributions of each of the parameters is summarized in Table 4. Since the number of levels of the random effect 'sch' exceeds 60, only the first two schools' estimates are listed in the table.

The 95% credible interval for the estimate of 'cltype: reg' is below zero, thus there is strong evidence that students in small class type significantly have better performance in math than those in regular class. The estimate of 'cltype: reg+A' is -9.2, slightly larger than -10.1. And this tells us that teachers aide slightly improve students math performance, but small class type is still much better than it.

The estimate of 'ses: N' is 14.7 and this positive estimation means that students who are not eligible for free lunches have better performance. This is in line with our intuition because those students usually come from wealthy families and their parents can provide them with

Table 4: Estimations of parameters in stan

Variable	**n_reff**	**Rhat**	**mean**	**mcse**	**sd**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
Intercept	1249	1	560.2	0.2	6.3	546.9	556.4	560.4	564.1	572.1
cltype: reg	6000	1	-10.1	0	2.5	-15.1	-11.8	-10.1	-8.4	-5.2
cltype: reg+A	6000	1	-9.2	0	2.2	-13.6	-10.7	-9.2	-7.6	-4.9
ses: N	6000	1	14.7	0	2.3	10.1	13.2	14.7	16.2	19.2
eth: B	5061	1	-16.3	0	3.2	-22.8	-18.5	-16.3	-14.2	-10
eth: A	6000	1	-13.4	0.3	22.4	-57.5	-28.7	-13.2	1.8	30.4
eth: H	6000	1	4.4	0.3	21.7	-37.6	-10.7	4.6	19.1	46
eth: O	6000	1	18.3	0.4	30.7	-42	-2	18.4	38.9	77.6
sch: 2	6000	1	6.3	0.1	7.1	-7.4	1.5	6.2	11	20.6
sch: 3	6000	1	-12.2	0.1	9.1	-30.6	-18.5	-12.1	-5.9	5

better learning resources.

The estimate of different 'eth' indicates that students with different trace tend to have different performance.

Table 5 is the result of 'lmer' function. The result is similar with those in stan.

Table 5: Estimations of parameters in lmer

Variable	**mean**	**sd**	**2.5%**	**97.5%**
Intercept	559.225	5.717	546.567096	570.300773
cltypereg	-10.059	2.502	-14.976355	-5.177628
cltypereg+A	-9.247	2.227	-13.590642	-4.875379
sesN	14.659	2.292	10.181001	19.164992
ethB	-16.306	3.277	-22.751790	-9.868099
ethA	-13.040	22.465	-56.832600	31.097515
ethH	4.519	21.869	-38.263411	47.301601
ethO	18.307	30.694	-41.706789	78.394592

We also compare our model with regular linear model $math \sim cltype + ses + eth$.

Table 6: Comparison between linear model and multilevel model

Model	Intercept	reg	reg+A	ses: N	eth: B	eth: A	eth: H	eth: O
Linear Model	576.9	-9.5	-8,9	16.4	-20.5	-5.6	4.9	22.5
Multilevel Model	560.2	-10.1	-9.2	14.7	-16.3	-13.4	4.4	18.3

From Table 6, we can see that the estimates of coefficients in both models are slightly different except 'eth: B' and 'eth: A'. After calculating BIC, Multilevel Model is better than Linear Model.

7 Summary

The final model answers research question:

- Class size significantly affects performance of students on math.
- 'ses' along with 'eth' also have influence on performance of students on math.

8 Future Work

All our analysis above focuses on factors that affect math scores, if given more time we can take a look at the factors which affect the read scores. It is interesting to see whether two sets of factors are the same or not.

If time permits, more interesting questions can be discussed. Based on the whole dataset, we are curious to find if class type affect the test scores when the grade goes up. Or what differences will be made on the test scores if a student change the class type even change the school during the four consecutive years. We can both discuss the student level and school level which are more complex.

9 Reflection

1. We learned strength and weakness of multilevel model and how to apply it to educational research, which can also generalize to other fields such as linguistic, biology etc.
2. The whole process of data analysis, from data collection, data cleaning to statistical modeling and insightful inference.
3. The best way to learn is doing a project by ourselves from start to end. Its important to understand the theoretical knowledge well, but more important is knowing how to apply statistical methods to answer questions of interest.