STAR Data Analysis Final Report

Yehong Qiu

Mar. 17, 2025

Abstract

The STAR (Student/Teacher Achievement Ratio) project was a large-scale, randomized experiment conducted in the late 1980s to evaluate the impact of class size on student achievement. In this analysis, I focus on the math scores of 1st-grade students to determine whether different class types significantly affect academic performance. An individual-level linear mixed-effects model, is employed to assess differences in math scaled scores among class types. While prior research suggests that small class sizes positively influence long-term academic success, this study investigates whether these effects are evident as early as 1st grade. The findings indicate that students in small classes tend to achieve higher math scores; however, further investigation is needed to validate the causality of this association due to limitations in the study designs, model assumptions, and data structure.

Introduction

The Tennesses Student/Teacher Achievement Ratio study (a.k.a. Project STAR) was conducted in the late 1980s to evaluate the effect of class sizes on test scores. There are 3 class types in total: small classes, regular classes, and regular classes with a teacher’s aide. This is a well-known randomized, longitudinal experiment, following students from kindergarten to 3rd grade. Only the schools with adequate resources allowing at least one class of each type are included in the experiment. Students in different schools are randomly assigned to those classes, and teachers are also randomly assigned. The cognitive outcomes are measured by standardized tests.

In this analysis, only the math scores in 1st grade is examined. The primary question of interest is **whether there is any differences in math scaled scores in 1st grade across class types**, and if so, a secondary question of interest is **which class type is associated with the highest math scaled scores in 1st grade**.

Researchers have found that attending small classes in early grades (K-3) is good for students’ long term academic success. Thus, it is expected that students in small classes have better math scores than those who are not.

The analysis report is structured as following:

- In the Background session, the experimental designs of the STAR project are explained and criticized; then the causal effect is evaluated based on the principles in causal inference; besides, I revisit the initial analysis report I’ve done previously and comment on the caveats.
- In the Descriptive Analysis session, an EDA is conducted, the findings of which lies the bedrock for the designs of the statistical model.
- In the Inferential Analysis session, an individual-level linear mixed-effect model is proposed, along with the results of hypothesis testing and post-hoc multiple comparison. In this session, the two main questions of interest are answered with a statistical rigor.
- In the Sensitivity Analysis, the assumptions of the proposed model are examined. Besides, in order to check the robustness of the answers to the questions of interest, several alternate models are proposed and compared with the original model.

Background

In this session, I will talk about the experimental designs of STAR, comment on them from a perspective of causal inference, and revisit the initial analysis that was previously done.

(a) Experimental Design

The study designs feature in:

- The three different class arrangements are: a small class type (S) with 13-17 students, a regular class (R) of 22-25 students, and a regular class with a full-time teacher aide (RA).
- Each school participate in the program has at least one class of each type.
- The study follows the same group of students entering kindergarten in 1985 and those students were assigned at random to one of three class types. Once assigned to a class type, students were to remain in the assigned class type as long as possible.
- Teachers are also randomly assigned to each class within schools.

The issue in student reassignment is a key factor of randomization, and it is important to figure out if there are any non-random factors when students change their class types at the beginning of 1st grade due to any pre-assumed opinions such as “Small classes result in better performances”.

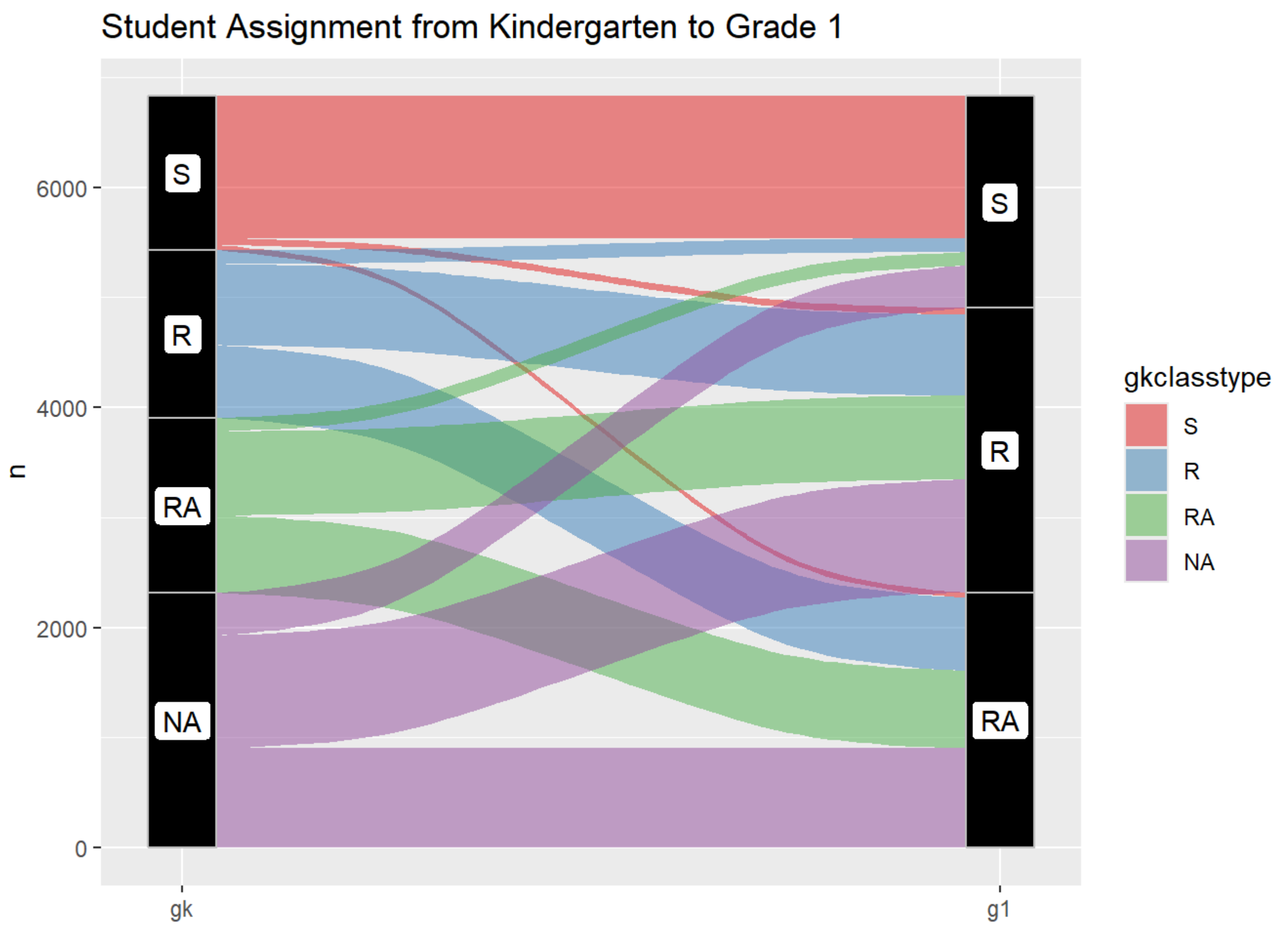
(a.1) Student Reassignment and Mobility

At the end of the kindergarten year, since they found that there had been no significant differences in the achievement of regular classes (R) and teacher-aide classes (RA) in the kindergarten year, nearly one-half of R students were randomly assigned to RA classes for their 1st grade and beyond. The majority of S students remain in the small class type when going into 1st grade.

What’s more, since kindergarten was not legally mandated in Tennessee at the time, a substantial number of students joined STAR as they entered 1st grade. These students were also randomly assigned at random to one of the three class types. Such a fact is confirmed in the following alluvial plot (See the students labelled as “NA” for their kindergarten).

Issue of the intentional mobility

The following figure shows the student assignment from kindergarten to 1st grade in STAR. The majority of students followed the randomized assignment policy as above, while there were more students in R or RA classes shifted to S classes compared to those did the opposite. There were 108 out of 1400 students shifted from S to either R or RS (roughly 7.7%), while there were 126 out of 1526 students (roughly 8.3%) shifted from R to S, and 122 out of 1589 students (roughly 7.9%) shifted from RA to S.



The fact that more students in R or RA shifted to S in their 1st grade might be driven by the non-random factors such as concerns from parents, say, they believed that their children could get better education if they were in small classes. If that so, it counterfeits the principle of randomization in experiment designs, and the causal effect of the analysis might not be convincing.

Issue of social-economic confounders

It is worthy to note that the student randomization only achieved within each school, i.e., within each class of each school, there was a cohort of students with mixed backgrounds (gender, race, and free-lunch status). But there are still substantial differences in the student families' social-economic status across schools of different regions. This informs me to incorporate the student demographic variables into the model.

(a.2) An overall remark on the Experimental Designs

The experimental designs of STAR are not strictly randomized, the artifacts mentioned above lead to the non-convincing causal effect, and the necessity of building an individual-level model involving social-economic covariates.

(b) Evaluate the convincingness of causal effect

Due to the complexity of social sciences experiments, the causal effects of the analysis might not be convincing. In what follows a few artifacts against the principles of causal inference will be discussed.

(b.1) The treatment-potential outcomes independence questioned

From the alluvial plot above, it is observed that along with the randomized assignment, students intentionally choose small classes out of their own interests. Thus, the treatment assignment and the potential outcomes might be dependent, counterfeiting the key principle behind calculating ACF.

It is very likely that, **parents most concerned with their children education asked for shifting from R or RA to S. The class type S, in people's predisposed knowledge, leads to better academic outcomes.**

(b.2) SUTVA might not be satisfied

SUTVA requires that there is only one version of treatments across subjects. But **due to the teachers' different background and experience**, it is hard to say the same class type means exactly the same treatment.

Put together, there are a few counterfeits of causal effect principles in STAR. The causal effect of the analysis might not be convincing.

(c) Initial Analysis Revisited

In the previous initial analysis, I built up a two-way fixed-effect ANOVA model using each class's median math score as the basic unit.

The **assumptions** are:

- Each class's performance is independent with each other.
- The effect of the class types and schools are additive, i.e. the mean of the response variable is influenced additively by the two factors, without considering the interaction between the two factors.
- Homoskedasticity: there is a uniform variance in classes' performance across different class types and schools.

Specifically, the model:

- Class type `star1` and school id `schoolid` are defined at two fixed effect.
- the simple model answered the two questions of interest: First, the class type has significant effect on 1st grade math scores. Second, the small class type is associated with the best 1st grade math performance.

One **potential caveat** of the model is:

It aggregate students into a unit of class, failing to depict the impact of social-economic diversity on academic performance. It is widely known that the economic hardship of households directly associates with the students' academic achievement. If an **individual-level model** is built instead, we can incorporate more covariates such as gender, race, and free-lunch status into the model.

Descriptive analysis

In the session, an EDA is conducted to gain insights for the downstream analysis. The initial dataset comprises 11,601 observations across 379 variables, from which I select the variables related to students demographics and their 1st grade class types, schools, teachers, and math performances.

Variable Name	Meaning
stdntid	The unique ID of each student
gender	The gender of each student (male/female)
race	The ethnicity of each studnet (white/black/hispanic/asian/…)
g1freelunch	An indicator of students eligibility of having lunch for free
FLAGSG1	An indicator of if a student was in STAR for their 1st grade
flagg1	An indicator of if a student had grades for their 1st grade
g1tmathss	each student’s scaled math score in 1st grade
g1classtype	The STAR class type of grade 1
g1surban	A school urbanicity indicator
g1schid	The unique ID for each school
g1tchid	The unique ID for each teacher

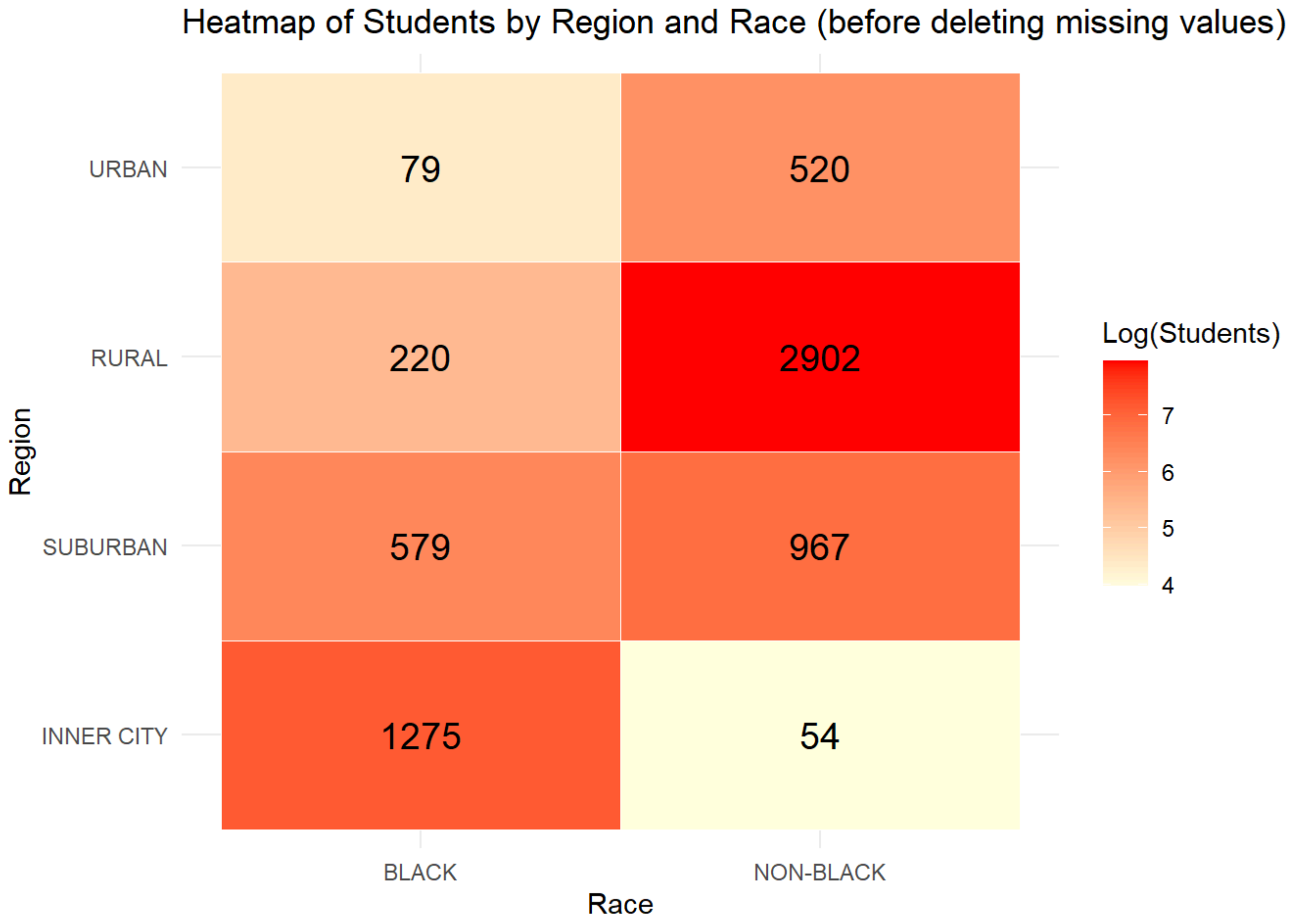
In the following, I first handle the missing data and then conduct an EDA around the relationship between students’ demographic variables and their 1st grade math scores.

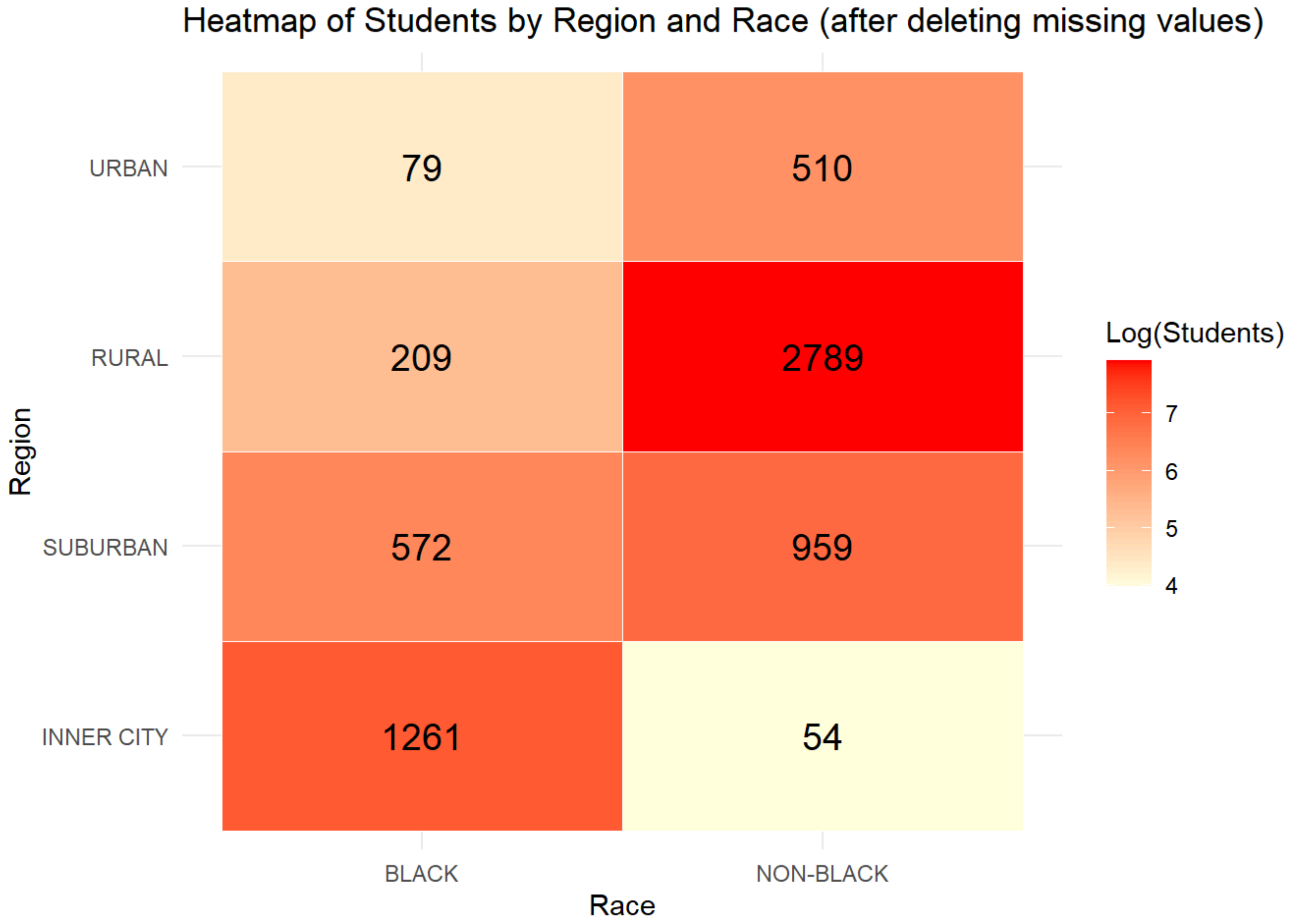
(a) Handling Missing Data

First, I directly delete the observations with missing 1st grade records or missing 1st grade math scores. Afterwards, a detailed inspection into the dataset shows that there are still 2 observations with missing “race”. I directly delete the 2 observations since they do not account for a large percentage in the dataset.

It is shown that there are 163 observations missing the state of “g1freelunch”. Is the missing values missing completely at random? Another question at hand is that since the free lunch status is closely related to the social-economic status of a student’s family, so inspecting the relationships of “g1freelunch”, “race” and “g1surban” is important.

For the first question, if the data is not missing completely at random, then we cannot simply delete those observations. But after inspection, directly deleting the missing values is acceptable. From the following two heatmaps, the demographic diversity of students do not change much after deleting the observations with missing values in “g1freelunch”.





For the second question, I will conduct an EDA around the “g1freelunch”, “race” and “g1surban” afterwards.

(b) the Dataset Checks after Deletion

I group students together based on their class type `g1classtype` , their school's ID `g1schid` , and their teachers' ID `g1tchid` .

After the aggregation process, I get 339 classes in total. It shows that each teacher is only responsible for one class. Besides, there are five schools without the RA class type, and their school IDs are `203452` , `244728` , `244736` , `244796` and `244839` . Considering that the 5 schools only account for a small percentage in all schools, I delete these 5 schools directly since they do not meet the requirement of having at least one class in each type.

Next I inspect if the current data set has a valid data proportion consistent with that in the official STARGuide.

1st Grade Class Type	Proportion after missing data handling (%)	Valid Data Proportion in Official Guide (%)
SMALL CLASS	28.2	28.3
REGULAR CLASS	36.7	37.8
REGULAR + AIDE CLASS	35.1	34.0

Student Gender	Proportion after missing data handling (%)	Valid Data Proportion in Official Guide (%)
MALE	51.9	52.9
FEMALE	48.1	47.1

Student Race	Proportion after missing data handling (%)	Valid Data Proportion in Official Guide (%)
NON-WHITE	31.2	37.2
WHITE	68.8	62.8

Student Free Lunch Status	Proportion after missing data handling (%)	Valid Data Proportion in Official Guide (%)
FREE LUNCH	49.9	51.6
NON-FREE LUNCH	50.1	48.4

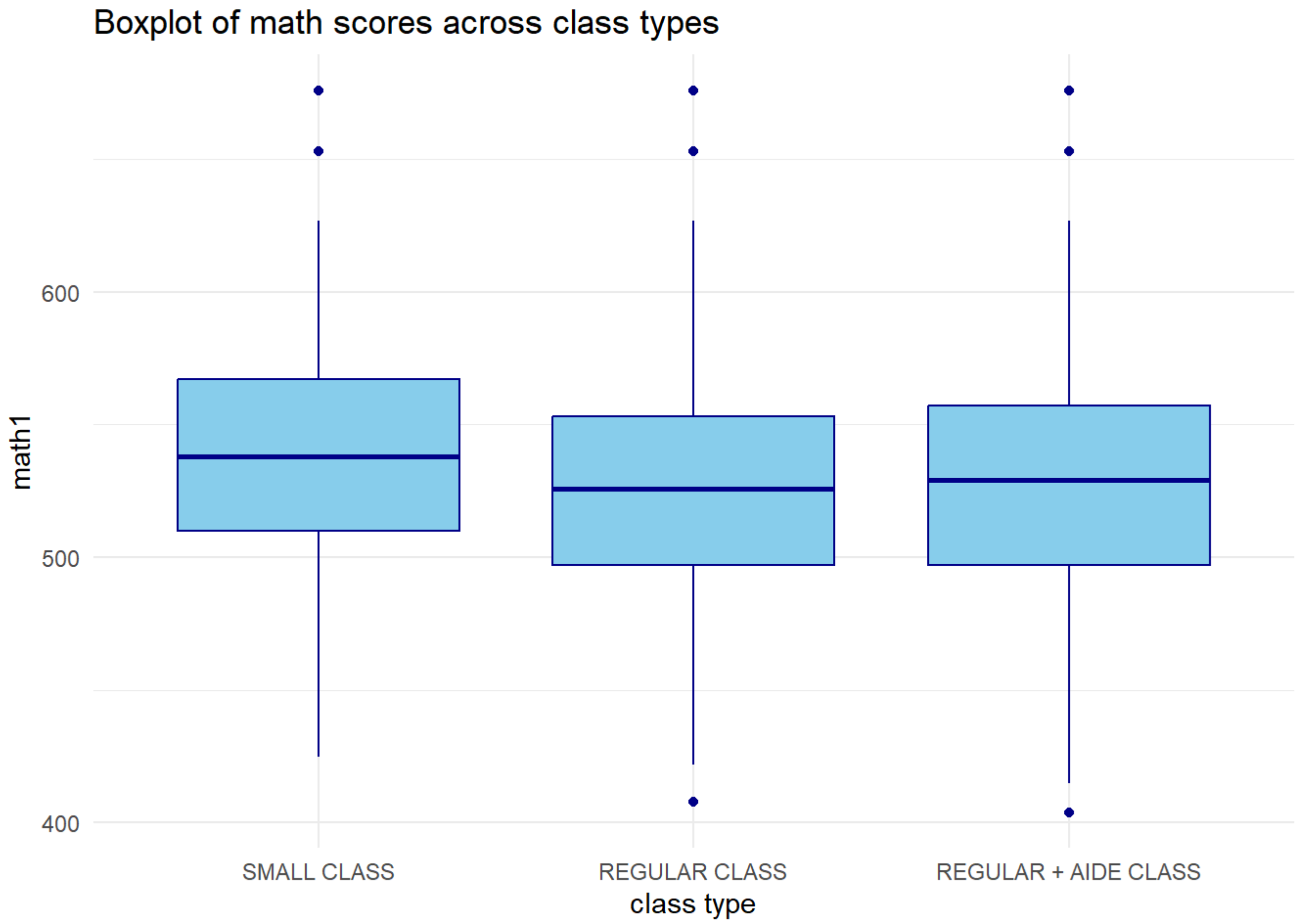
School Urbanicity	Proportion after missing data handling (%)	Valid Data Proportion in Official Guide (%)
INNER CITY	19.0	20.2
SUBURBAN	23.4	23.2
RURAL	47.9	47.4
URBAN	9.7	9.2

Put together, only the proportion of race shows an obvious difference from the official data given by the STARGuide. There are 6% less white students in my data set than that in the official valid data set. I consider this as a trivial deviation because in general there is still consistency of my data set with the official valid data set, given that the white students accounts for the majority in the sample.

(c) Multivariate Descriptive Statistics

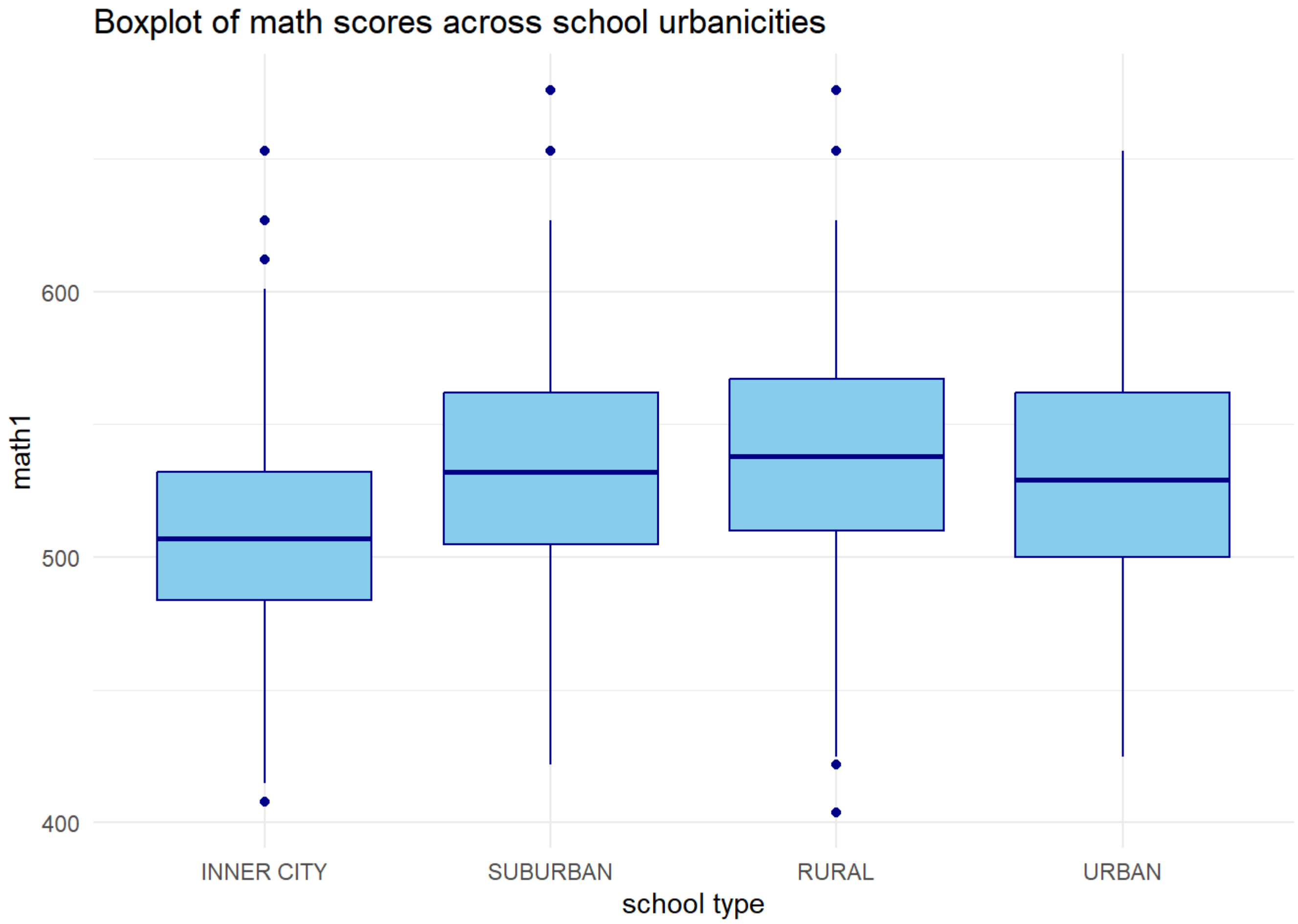
Next, I do multivariate descriptive statistics for the math score of each student with the class types and the school urbanicities, respectively. Besides, the relationships between the math score and student's race, and free lunch status are also visualized. This step aims to explore visually how these key factors impact the students' performance.

Outcome v.s. class types (3 types in total)



The key observation is that "small" classes (small) have the highest median math score out of the three class types, indicating that students in small classes outperform those in regular and regular+aide.

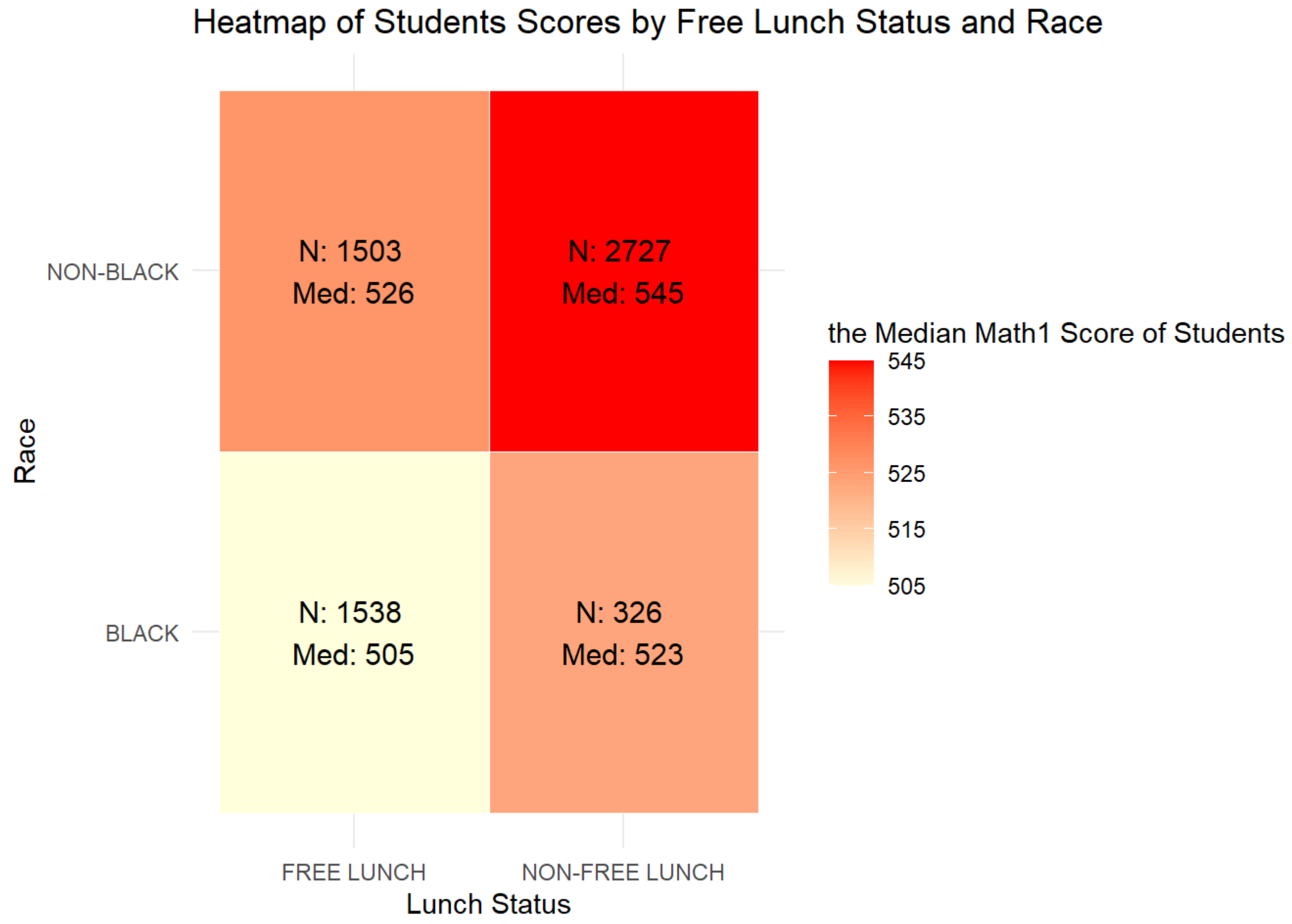
Outcome v.s. school urbanicity



The key observation is that students in “Inner City” schools have the lowest median math score out of the 4 school urbanicity types.

Outcome v.s. Students’ Free Lunch Status

The following heatmap shows that the students enjoying the free lunch have significantly lower math grades than those who don’t. Usually, the free lunch status reflects the economic hardship in the student’s family. Thus, the poorer the family, the lower the math grades of the student has on average.



What’s more, the student’s race is also associated with the 1st grade math score. From the above heatmap, we can learn that:

- The race and lunch status is interrelated: only 35.8% of non-black students are eligible for free lunches, while there are 74.6 % of black students eligible for free lunches.
- Students having free lunches score significantly lower than those who do not.
- Non-black students score significantly higher than black students.

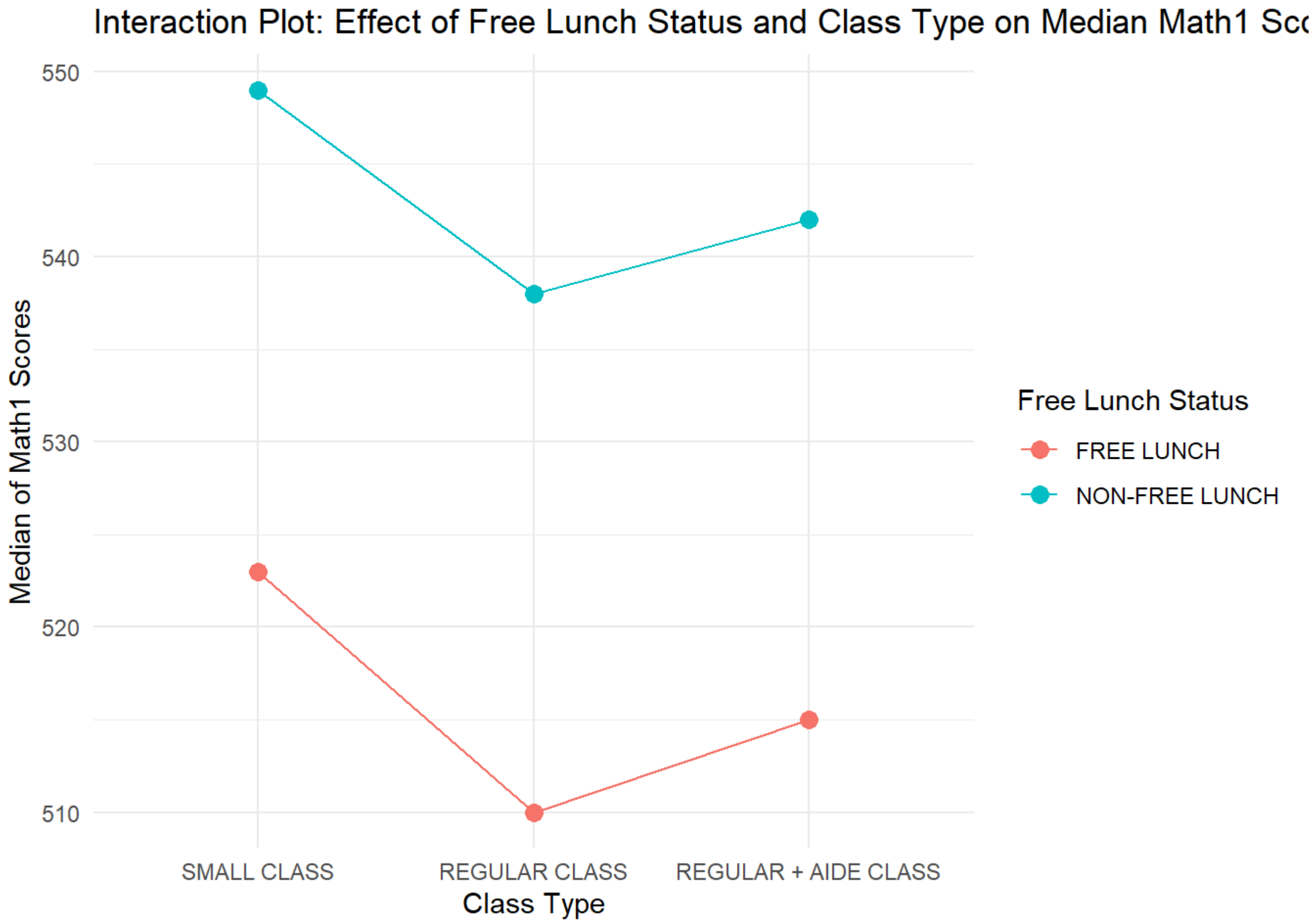
Conclusion:

Put together, apart from the class type and school urbanicity, the student’s race and free lunch status are two key demographic factors behind their 1st grade math performance.

(d) Visualization Checks for Linear Mixed Modeling

(d.1) The pair-wise intereaction among free lunch status, class type, school urbanicity, and race:

From the below interaction plot, there is no interaction between free lunch status and class type. Actually there is also no significant interaction between any other two of the four effects.

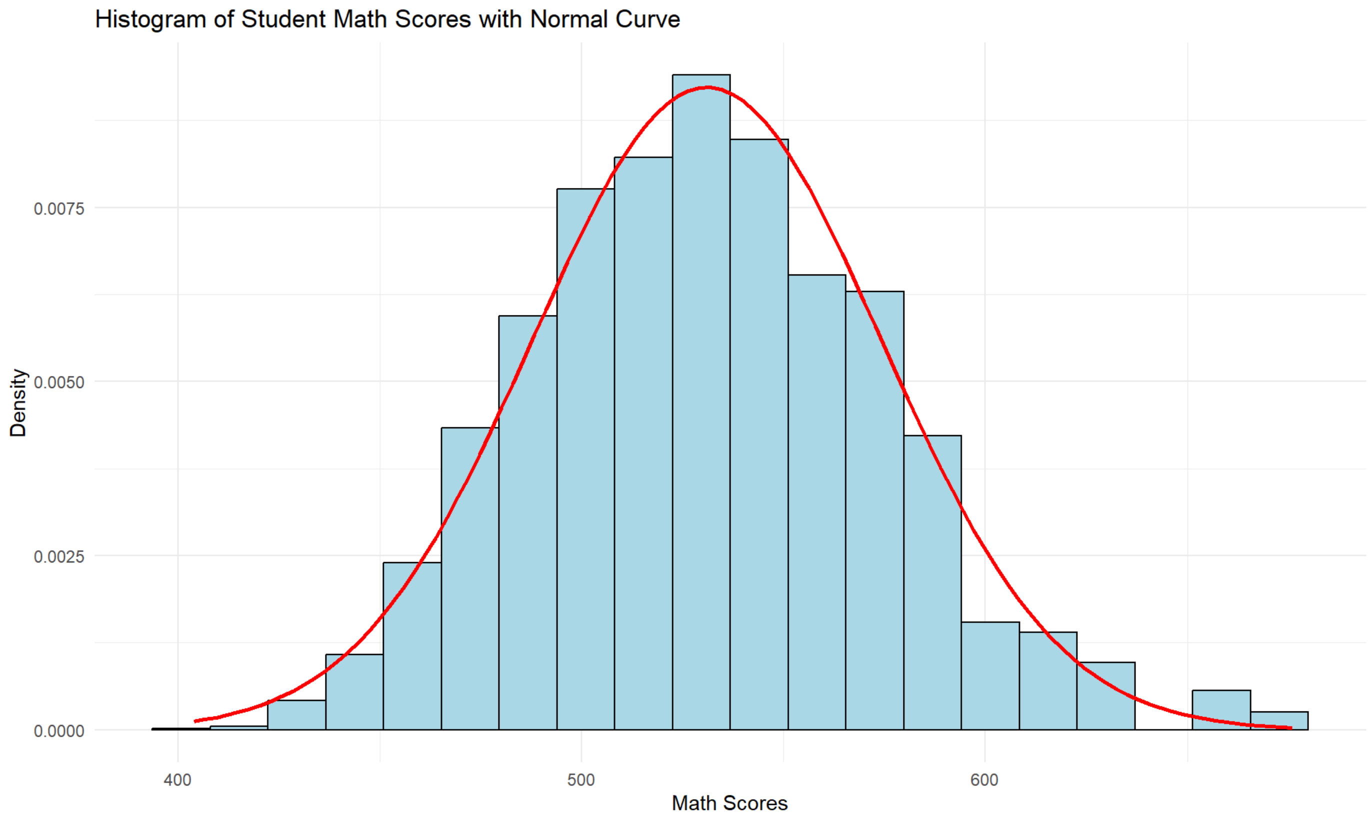


Why no interaction between the class type and the school urbanicity?

- From the initial analysis, I found that the interaction between the class type and the school urbanicity is not significant.
- The key point for the analysis is about the impact of the different class types on students' academic performances. The interaction between the class type and other covariates is not the main focus.

(d.2) Normality Check of students' math scores distribution

Next, I check the distribution of students' math scores. It is found that it is roughly normally distributed, lying the bedrock for building a linear mixed-effect model with the normal error assumptions.



(e) Gained Insights

After the above analysis, the following insights are gained and these insights help to build an appropriate model in the next session:

- **Class Type Matters:** Small class teaching is associated with higher math scores.
  - Model class type as a fixed effect.
- **Urbanicity Counts:** Students in schools located in the inner city regions perform systematically lower than others in math scores.
  - Model school urbanicity as a fixed effect.
- **Race and Free-Lunch Status:** Black students or those students eligible for free lunches do poorer in math than those who are not black or do not need free lunches.
  - Model student's race and free-lunch status as fixed effects.
- **No Obvious Interactions:** There are no obvious interactions between any two of these covariates.
  - Model an additive model without interaction terms.
- **Teacher's Diversities:** Teachers are different in their experience and career stages. Although there are only three class types, there are actually far more "treatments" — When it comes to education, it is hard to standardized everything.
  - Model teacher's id as a random effect.
- **Normality:** The distribution of each student's math score is roughly normal.
  - Use the normal error assumption for the model.

Inferential analysis

In this session, a linear mixed-effect model is proposed based on the insights gained in the process of EDA. The hypothesis testing and pairwise comparison are conducted to answer the two main questions of interest:

- **Hypothesis Testing:** Whether there is any differences in math scaled scores in 1st grade across class types?

- **Pairwise Comparison:** If the answer to the previous question is yst, which class type is associated with the highest math scaled scores in 1st grade?

(a) Model I

After the EDA, Model I is defined as follows:

$$Y_n = \mu + \text{ClassType}_i + \text{SchUrbanicity}_j + \text{StdntRace}_k + \text{StdntLunch}_l + \text{Teacher}_m + \epsilon_n$$

- Y_n is the n^{th} student's 1st grade math score.
- μ is the overall mean of the cohort's math scores.
- ClassType_i is the fixed effect of the i^{th} class type ($i = 1, 2, 3$). The three types are Small (S), Regular (R), and Regular + Aide (RA).
- SchUrbanicity_j is the fixed effect of the j^{th} school urbanicity ($i = 1, 2, 3, 4$). The four types are Inner City, Urban, Rural, and Suburban.
- StdntRace_k is the fixed effect of the k^{th} race of students ($i = 1, 2$). The two types are Black and Non-Black.
- StdntLunch_l is the fixed effect of the l^{th} lunch status of students ($i = 1, 2$). The two types are Free and Non-Free.
- Teacher_m is the random effect of the m^{th} teacher, modeled as a i.i.d. random variable of a zero-mean and a variance of σ_r^2.
- ϵ_n is the i.i.d. error term under the normal assumption, $\epsilon \sim N(0, \sigma^2)$.

(b) Model I Fitting

```
model_I <- lmer(g1tmathss ~ g1classtype + race + g1surban + g1freelunch + (1|g1tchid), data=sta
r_sub)
```

Fixed Effects	Estimate	Std. Error
(Intercept)	514.5816	3.2421
Class Type:REGULAR CLASS	-11.5798	2.9216
Class Type:REGULAR + AIDE CLASS	-8.9365	2.9693
Stdnt Race: NON-BLACK	18.6108	1.7166
School: SUBURBAN	0.8863	3.8967
School: RURAL	3.3845	3.5833
School: URBAN	0.8216	4.9562
Free Lunch: NON-FREE LUNCH	18.3216	1.0901

Interpretation on a subset of estimated parameters:

- **Esitmateion of μ**: The estimated intercept is `514.5816` , which is the overall mean of the 1st grade math scores of the cohort. It can be interpreted as the expected math score of a balck student in a small class in an inner city school, eligible for free lunch, taught by the baseline teacher.

- **Parameters related to $\mathbf{ClassType_i}$**: The interpretation based on the output is slightly different from that based on the model's factor-effect form of formula.
    - Holding other factors the same, students in regular classes have 11.5798 less on math scores on average than those in small classes.
    - Holding other factors the same, students in regular+aide classes have 8.9365 less on math scores on average than those in small classes.
- **Parameters related to $\mathbf{StdntRace_k}$**:Holding other factors the same, non-black students have 18.6108 more on math scores on average than those who are black.

(c) Hypothesis Testing

To answer the first question of interest: Whether there is any differences in math scaled scores in 1st grade across class types? I use the likelihood ratio test (LRT) on the following hypotheses:

- **Null Hypothesis (H0):** $\text{ClassType}_i = 0, i = 1, 2, 3.$
- **Alternative Hypothesis (H1):** $\exists\ \text{ClassType}_i \neq 0.$

The test statistic and its null distribution is:

$$-2 \times (\text{log-likelihood}_{\text{reduced}} - \text{log-likelihood}_{\text{full}}) \sim \chi^2_{df=2}$$

.

	n...	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
model_null	8	61194.36	61248.08	-30589.18	61178.36	NA	NA	NA
model_I	10	61181.17	61248.33	-30580.59	61161.17	17.18373	2	0.0001856095

2 rows

The test statistic value is 17.184, with a p-value of 0.0001856, indicating that the class type has a significant effect on students' math scores.

So, for the primary question, the answer is: there are significant differences in the math scores of among students in different class types.

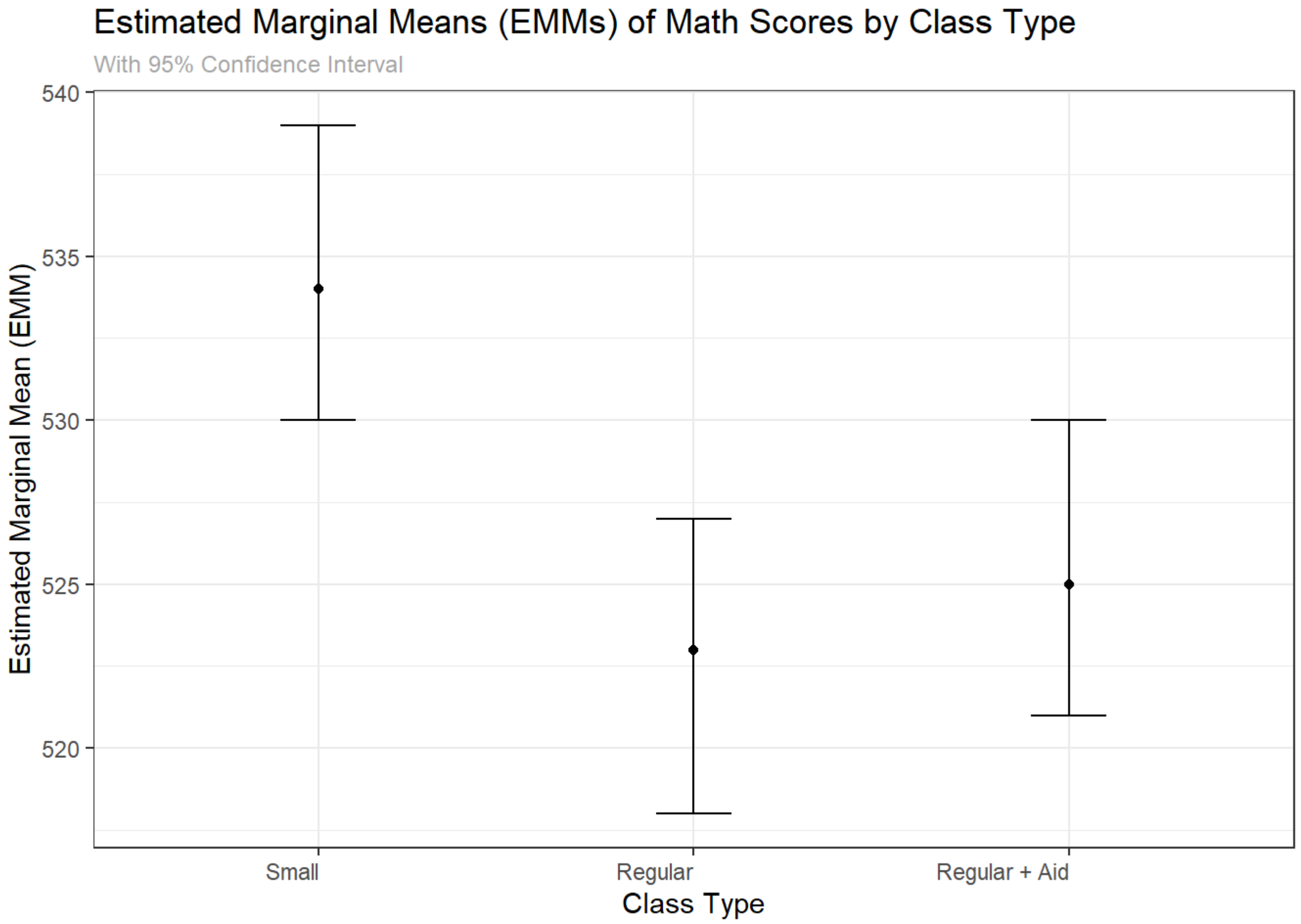
(d) Pairwise Comparison

Since there are significant differences in the math scores of students from different class types, it is necessary to do a post-hoc multiple comparison.

The second question of interest is: Which class type is associated with the highest math scaled scores in 1st grade?

I use the Tukey's Range test to compare the estimated marginal mean differences of any two out of the three class types.

ClassType	EMean	SE	LCL	UCL
Small	534	2.18	530	539
Regular	523	2.23	518	527
Regular + Aid	525	2.29	521	530



contrast	estimate	SE	z.ratio	p.value
SMALL CLASS - REGULAR CLASS	11.579755	2.921626	3.9634624	0.0002180
SMALL CLASS - (REGULAR + AIDE CLASS)	8.936509	2.969349	3.0095853	0.0073776
REGULAR CLASS - (REGULAR + AIDE CLASS)	-2.643246	3.004018	-0.8799036	0.6530093

So, for the secondary question, the answer is: under 5% significance level, students in the small classes achieve on average significantly better than those students in the regular or regular+aide classes.

Sensitivity Analysis

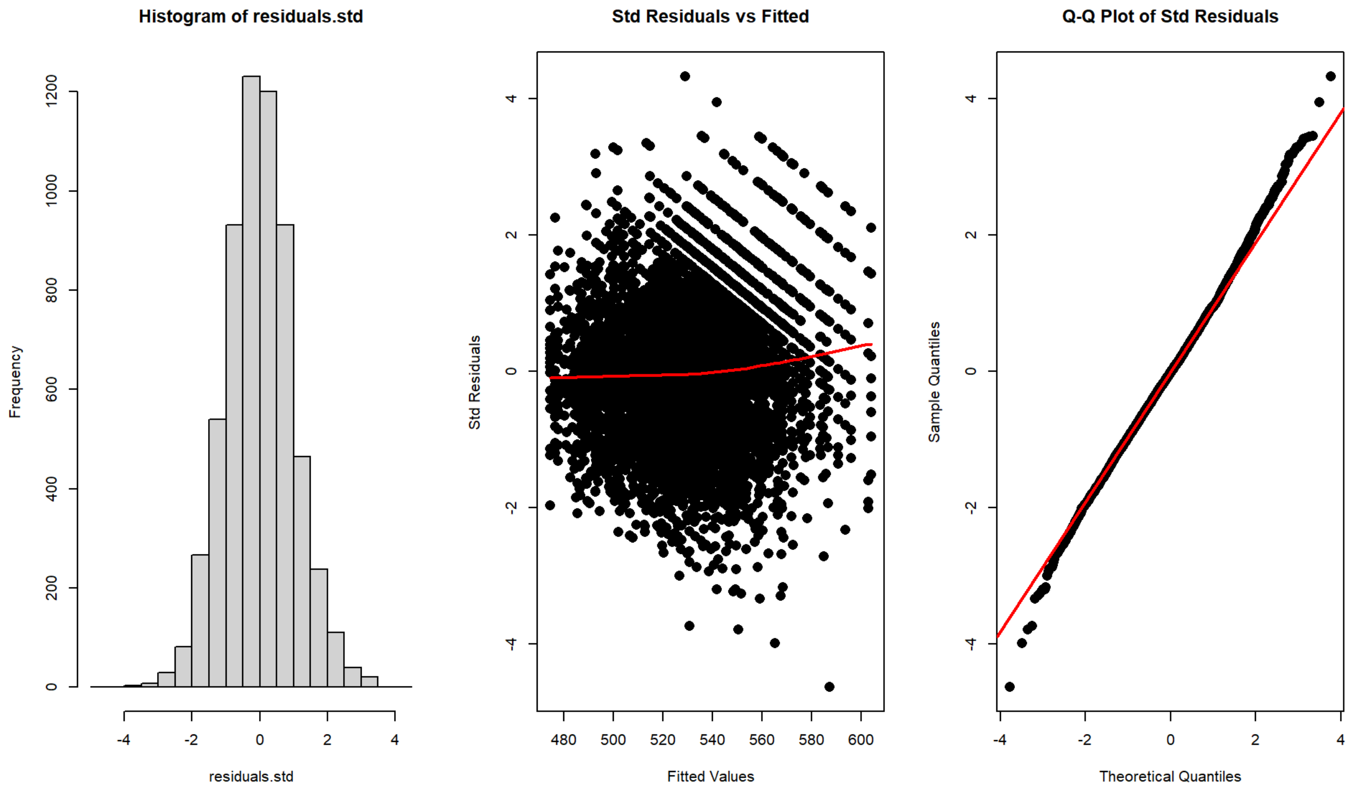
In this session, I verify the assumptions of the `Model I` proposed above. What's more, some alternate models are proposed and check if the results are consistent among slightly different models.

(a) Diagnostics of Model I

(a.1) Residual Diagnostics

I examine the residual plots of the fitted model `model I`. From the Standardized Residuals v.s. Fitted plot and the Standardized Residuals' Q-Q plot, it is found that:

1. As the fitted math score increases, the mean of the std. residuals begins to deviate from 0, indicating a systemic deviation of the data from the proposed model.
2. The standardized residuals are heavily-tailed distributed.

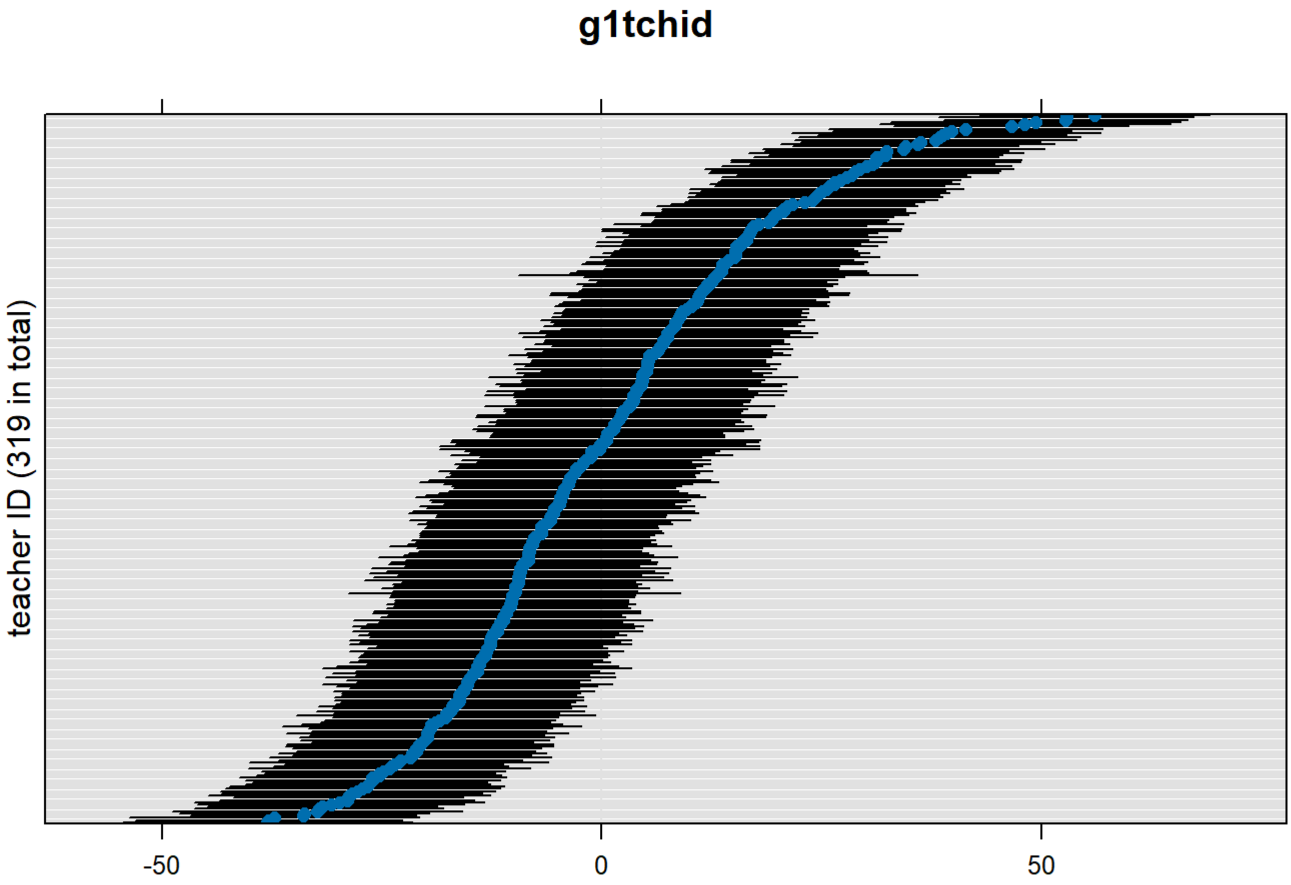


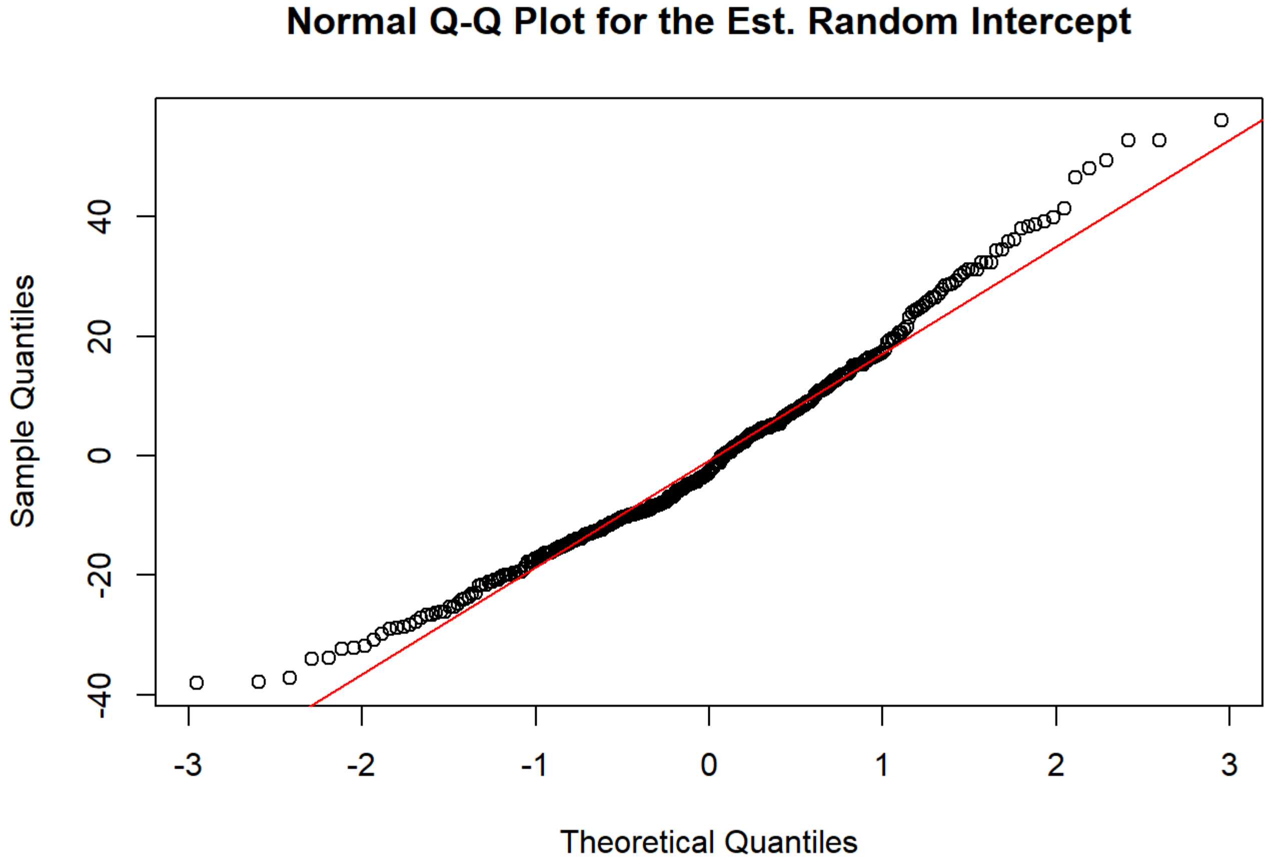
Additionally, the standardized residuals do not pass the normality check. The p-value of the Anderson-Darling test is `1.003e-06` , indicating significant deviation of the standardized residuals from a normal distribution.

(a.2) Random Effect Diagnostics

In what follows I check the validity of the normal assumption of the random effect `gltchid` in Model I. From the plots below, the estimated random effects are right-skewed, deviating from the normal distribution as well. The p-value of the Anderson-Darling test is `0.0001397` .

```
## $gltchid
```





Conclusion: The data does not satisfy the assumptions of Model I. The results of the statistical inference are only for reference.

(b) Alternate Models

From the above analysis, the data does not strictly follow the assumptions in Model I. Thus, several other models are proposed and the same inferential analysis & Diagnostics are performed so as to ensure the correctness of the results.

(b.1) Model II: Replace the fixed effect school urbanicity with schoolid

In Model II, one of the fixed effects in Model I — the school urbanicity, is replaced with the school ID. The other effects remain the same.

$$Y_n = \mu + \text{ClassType}_i + \text{School}_j + \text{StdntRace}_k + \text{StdntLunch}_l + \text{Teacher}_m + \epsilon_n$$

```
model_II <- lmer(g1tmathss ~ g1classtype + race + g1schid + g1freelunch + (1|g1tchid), data=sta
r_sub)
```

Characteristics of Model II:

- **Advantage of Model II:** Depict the differences of math scores of students in different schools in a more fine-grained way. Besides, the AIC of Model II is `61066` , smaller than that of Model I, which is `61181` .
- **Disadvantage of Model II:** Significantly increase the size of the model parameters, leading to an increase in the std. error of the estimated parameters for each school.
- **Inferential Analysis:** The inference conclusions of Model II are the same to Model I.
- **Diagnostics:** The residual diagnostics of Model II still indicate a heavy-tailed standardized residuals. Besides, the estimated intercepts of the random effect also demonstrate a heavy-tailed distribution, deviating from the normal assumption.

(b.2) Model III: Turn the model into a fixed effect model

The only difference between Model II and Model III is that Model III turns the random effect `g1tchid` into a fixed effect. The other effects remain the same.

```
model_III <- lm(g1tmathss ~ g1classtype + race + g1schid + g1freelunch + g1tchid, data=star_su
b)
```

Characteristics of Model III:

- **Advantage of Model III:** Conceptually simple, avoid any random effect.
- **Disadvantage of Model III:** Significantly increase the size of the model parameters, leading to an increase in the std. error of the estimated parameters for each school.
- **Inferential Analysis:** The inference conclusions of Model III are the same to Model I and Model II.
- **Diagnostics:** The residual diagnostics of Model III still indicate a heavy-tailed standardized residuals.

(b.3) Comparison among Model I~III

Considering that all Models proposed are deviated from the respective model assumptions, I use AIC and BIC to select one model that is relatively the best. According to the following chart, Model II is the best.

Model	AIC	BIC
Model I	61153.58	61220.73
Model II	**60640.15**	**61157.21**
Model III	60852.69	63014.94

Cast Doubt against Small Class Superiority: Differ in Difference Model (DiD)

The differ in difference model is used because from the previous analysis, we cannot exclude the possibility that the students in the small classes are intellectually better than students in other types of classes, considering the fact that from the alluvial plot, the majority of students in small classes in kindergarten go directly into small classes in 1st grade. Thus, the response variable in the model here should be the improvements of the math grades from kindergarten and 1st grade.

The DiD model is of the same form with Model II, with the only difference being the response variable now is the difference between the 1st grade math score and the kindergarten math score.

$$\text{Math Score in 1st Grade}_n - \text{Math Score in Kindergarten}_n = \\ \mu + \text{ClassType}_i + \text{School}_j + \text{StdntRace}_k + \text{StdntLunch}_l + \text{Teacher}_m + \epsilon_n$$

```
Model_did <- lmer((g1tmathss - gktmathss)  ~ g1classtype + race + g1schid + g1freelunch + (1|g1
tchid), data=star_sub_did)
```

	n…	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
model_null	80	40403.31	40908.30	-20121.65	40243.31	*NA*	*NA*	*NA*
Model_did	82	40402.98	40920.59	-20119.49	40238.98	4.329161	2	0.1147981

2 rows

From the likelihood-ratio test, the p-value is `0.1148`, indicating that we fail to reject the null hypothesis: $\text{ClassType}_i = 0, i = 1, 2, 3$. Thus, the small class type is not necessarily associated with the significant academic improvement in math of students.

Although students in small classes have higher math scores, this does not mean being in small classes have a positive impact on students' 1st grade academic performance.

- It is worthy to further inspect the association between the class type and the student's math scores in **kindergarten**.
- The class types are not associated with students' math scores **improvements** from kindergarten to 1st grade.

Conclusion

This study finds that while students in small classes exhibit higher math scores in 1st grade, the data does not fully meet the assumptions required for the proposed statistical model, implying that the inferential analysis results should be interpreted with caution. What's more, the statistical analysis does not conclusively establish a causal relationship between class type and academic improvement. Future research should explore alternative modeling approaches and examine the impact of class size from a longitudinal perspective to capture its long-term effects.

Acknowledgement

The author gratefully expresses her thanks to Prof. Shizhe Chen and Yanhao Jin for providing insights and instructions throughout the project. Besides, she also discussed with her classmate Ziyue Yang (zyzyang@ucdavis.edu (mailto:zyzyang@ucdavis.edu)). She uses generative tools such as ChatGPT only for grammar checking.

References

1. Imbens, G., & Rubin, D. (2015). Stratified Randomized Experiments. In Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction (pp. 187-218). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139025751.010 (doi:10.1017/CBO9781139025751.010)

2. STA 207 Course Notes. Available at: https://nbviewer.org/github/ChenShizhe/StatDataScience/tree/master/Notes/ (https://nbviewer.org/github/ChenShizhe/StatDataScience/tree/master/Notes/), Mar.17. 2025

3. Achilles, Charles M. (2012). Class-Size Policy: The STAR Experiment and Related Class-Size Studies. NCPEA Policy Brief. Volume 1, Number 2. Available at: https://eric.ed.gov/?id=ED540485 (https://eric.ed.gov/?id=ED540485), Mar.17, 2025