

STAR Data Analysis Initial Report

Yehong Qiu

Feb. 14, 2025

Descriptive analysis

The STAR dataset downloaded from the “AER” library has 11598 rows and 47 variables. In this part, I will introduce my way of data cleaning, the students’ grouping into classes, and the EDA methods used on the dataset.

Data Cleaning

Variable Selection

First, I select “star1”, “math1”, “school1”, “schoolid1”, “ladder1”, “degree1”, “experience1”, and “tethnicity1” so that they carry necessary message about the 1st-grade pupils’ math scores and the teachers’ information.

Missing Data Handling

Then, I handle missing values by simply deleting the lines with all entries being NA, and the observations with missing 1st grade math scores. It is assumed that the data is missing completely at random (MCAT) so that the direct deletion won’t bring in any bias.

The below shows the univariate descriptive statistics for the selected variables.

Data summary



Name	star_sub
Number of rows	6600
Number of columns	8
Column type frequency:	
factor	6
numeric	2
Group variables	
None	

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
star1	0	1.00	FALSE	3	reg: 2507, reg: 2225, sma: 1868
school1	0	1.00	FALSE	4	rur: 3124, sub: 1547, inn: 1330, urb: 599
ladder1	33	1.00	FALSE	6	lev: 4354, app: 696, pro: 640, not: 492
degree1	12	1.00	FALSE	4	bac: 4305, mas: 2224, spe: 37, phd: 22

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
tethnicity1	42	0.99	FALSE	2	cau: 5420, afa: 1138
schoolid1	0	1.00	FALSE	76	51: 237, 63: 143, 27: 139, 9: 130

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
math1	0	1	530.53	43.10	404	500	529	557	676	
experience1	12	1	11.63	8.92	0	4	10	17	42	

Extra details about missing data handling

Even after some missing data processing, there are some missing values respect to the teachers' information. Next I will briefly explain how I handle them.

The 12 pupils whose teacher lacking all the information:

Only 12 out of 6600 observations have all teacher information missing, meaning that these pupils' teachers are completely unknown. Upon examining these pupils, all are found to belong to a small class at School 70, a rural school. Further inspection of other small classes within the school reveals another group of 17 pupils whose teachers share identical information, indicating they were taught by the same person. According to the STAR project guidelines, small classes typically consist of 15–17 pupils, suggesting that the 12 pupils with missing teacher information were not part of this same class. For simplicity, I assign these 12 pupils to a new class.

The 33 students whose teachers lacking the "ladder1"

Among the 33 students whose teachers lack "ladder1" information, they either belong to the small-size class in School 70, where the teacher's information is entirely missing (as previously mentioned), or to a regular-size class in School 19. For the 21 students in School 19, all available teacher information is identical, and their class size aligns with the STAR project guidelines, making it reasonable to group them into a single class.

The 42 students whose teachers lacking the "tethnicity1"

Among the 42 students whose teachers lack "tethnicity1" information, most come from the small-size class at School 70, where the teacher's information is entirely missing. The remaining students are from various classes across different schools. Given the difficulty of comparing these students' data with others to identify the missing information, and because removing 1–3 students from their classes is unlikely to significantly impact the results, I have decided to delete these students from the dataset.

Students' Performance Aggregation

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' performance.

the Median Scores used as the summary measure

Specifically, I select the median scores of students within each class, as the median is more robust to outliers. My reasoning is that, in reality, the performance of exceptionally high- or low-achieving students is influenced more by individual factors than by the pedagogical methods used.

Summary Measures for each teacher/class

I group students together based on their class type "**star1**", **their teachers' experience, career ladder, degree, ethnicity, and their schools**. Here it shows the first 5 and last 5 classes in the grouped data set rearranged according to the ascending order of the class size. It is noticed that for the first observation, there is only one student

in the class. Since it makes no sense of having only 1 students in a class, and that school 22 has multiple classes of type “Regular+aide”, I delete the teacher whose id is 327 from the grouped data set.

Students Math Scores in the 5 Smallest Classes

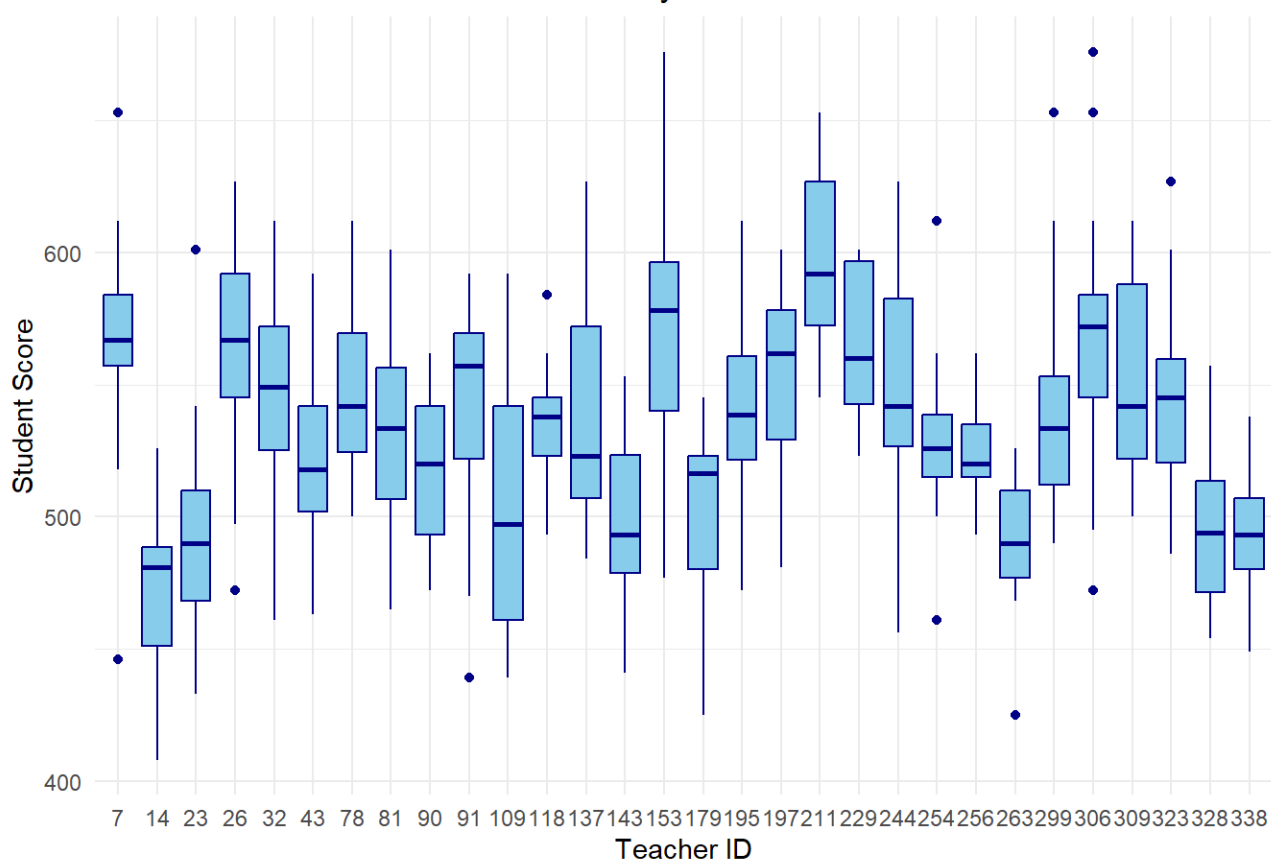
tid1	class_size	star1	school1	schoolid1	sd_math1	mean_math1	p0	p25	p50	p75	p100
327	1	regular+aide	inner-city	22	NA	481.0000	481	481.00	481	481.0	481
145	11	small	urban	59	31.49690	550.3636	515	523.00	538	581.0	592
117	12	small	rural	47	39.81196	551.4167	481	531.50	551	579.5	627
134	12	small	suburban	20	22.37423	532.6667	507	515.00	526	553.0	578
135	12	small	rural	78	30.61330	549.4167	493	528.25	549	573.5	592

Students Math Scores in the 5 Largest Classes

tid1	class_size	star1	school1	schoolid1	sd_math1	mean_math1	p0	p25	p50	p75	p100
36	27	regular	urban	2	30.87753	496.0370	449	476.50	490.0	521.5	562
270	27	regular+aide	urban	48	40.15164	513.0000	446	482.50	512.0	540.5	584
288	27	regular+aide	rural	1	49.54402	545.9259	458	501.00	545.0	584.0	653
50	28	regular	suburban	44	37.06813	506.2500	430	483.25	509.5	524.5	612
247	29	regular+aide	suburban	44	35.66918	517.8276	446	497.00	518.0	542.0	584

After the above pre-processing process, I get 339 teachers in total. The below side-by-side boxplot shows the distribution of the math scores of students taught by random selected 30 teacher. In the 30 classes, we can easily spot outliers, which means using median here is more likely to help us derive robust outcomes.

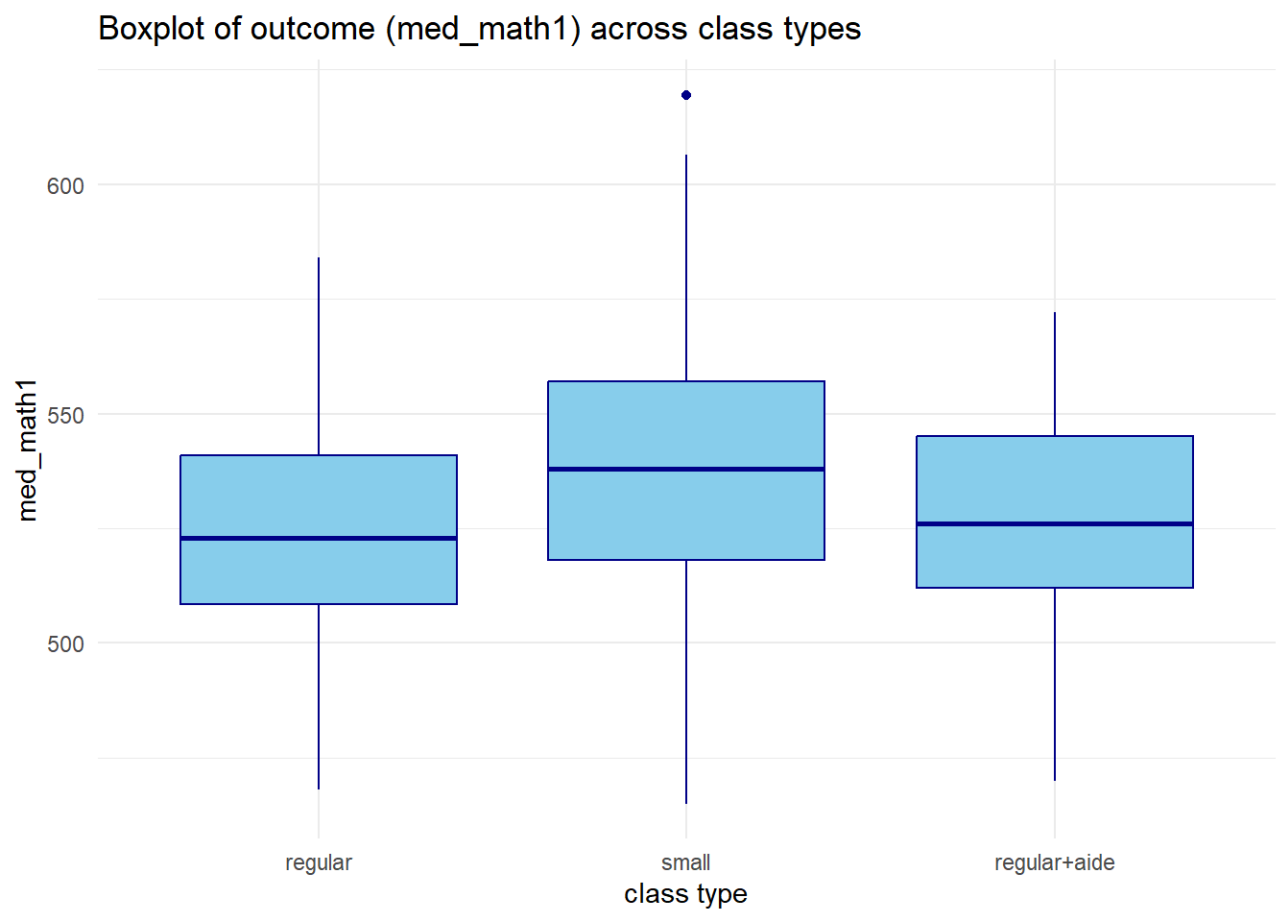
Distribution of Student Math1 Scores by the 30 Random-Selected Teacher



Multivariate Descriptive Statistics

Next, I do multivariate descriptive statistics for the median math score of each class with the class types and the school IDs, respectively. 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 IDs (76 schools in total)

Since there are many schools and a handful of teachers/classes per school, I calculate the summary statistics (sample mean, standard deviation, median, and other quantiles) of the outcome (med_math1) and also the class number in each school. The following lists the summary statistics for the top 5 and last 5 schools ranking according to the medians of their classes' median scores. In total there are 76 schools.

Top 5 Math Scores Comparison by School ID

schoolid1	school1	class_number	p0	p25	p50	p75	p100
73	rural	4	535.0	559.000	569.5	569.5	569.5
10	urban	3	567.0	567.000	567.0	567.0	567.0
74	rural	4	564.5	566.375	567.0	567.0	567.0
72	rural	6	538.0	549.500	566.0	566.0	566.0
3	rural	6	538.0	545.500	564.5	564.5	564.5

Last 5 Math Scores Comparison by School ID

schoolid1	school1	class_number	p0	p25	p50	p75	p100
19	inner-city	6	468.0	472.625	490.25	490.25	490.25
2	urban	3	479.0	484.500	490.00	490.00	490.00
28	inner-city	6	481.0	487.000	490.00	490.00	490.00
32	inner-city	6	465.5	483.250	490.00	490.00	490.00
26	inner-city	3	465.0	475.500	486.00	486.00	486.00

The key observations are:

- Among the 5 top schools, 4 are from “rural” and 1 from “urban”, indicating that the schools with the best student performance are mostly in the rural regions.
- Among the 5 last schools, 4 are from “inner-city” and 1 from “urban”, indicating that the schools with the poorest student performance are mostly in the inner-city areas.

Inferential analysis

Two-Way Anova Model without interaction

Parameters Explanation

We can define a two-way ANOVA model as follows $Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \epsilon_{ijk}$, where 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. Y_{ijk} denotes the median math score of the k th class of the type i and within the school j .

Model Assumptions

1. The model assumes that 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.
2. ϵ_{ijk} is the random error term associated with the k th class of the type i and within the school j . As the random variables, ϵ_{ijk} are independently and identically distributed (iid).

Comment: Here is a Potential Caveat

This “iid” assumption means that each subject is independent with each other. Such an assumption indicates whether the classes come from the same school or not, or are of the same class type or not, the median score of each class is all independent with each other. The students performance within the same school should not be independent, so the “iid” assumption is not realistic.

Constraints on Parameters

In the above two-way ANOVA model of its factor-effect form, there are two constraints on the parameters:

$$\sum_{i=1}^3 \alpha_i = \sum_{j=1}^{76} \beta_j = 0.$$

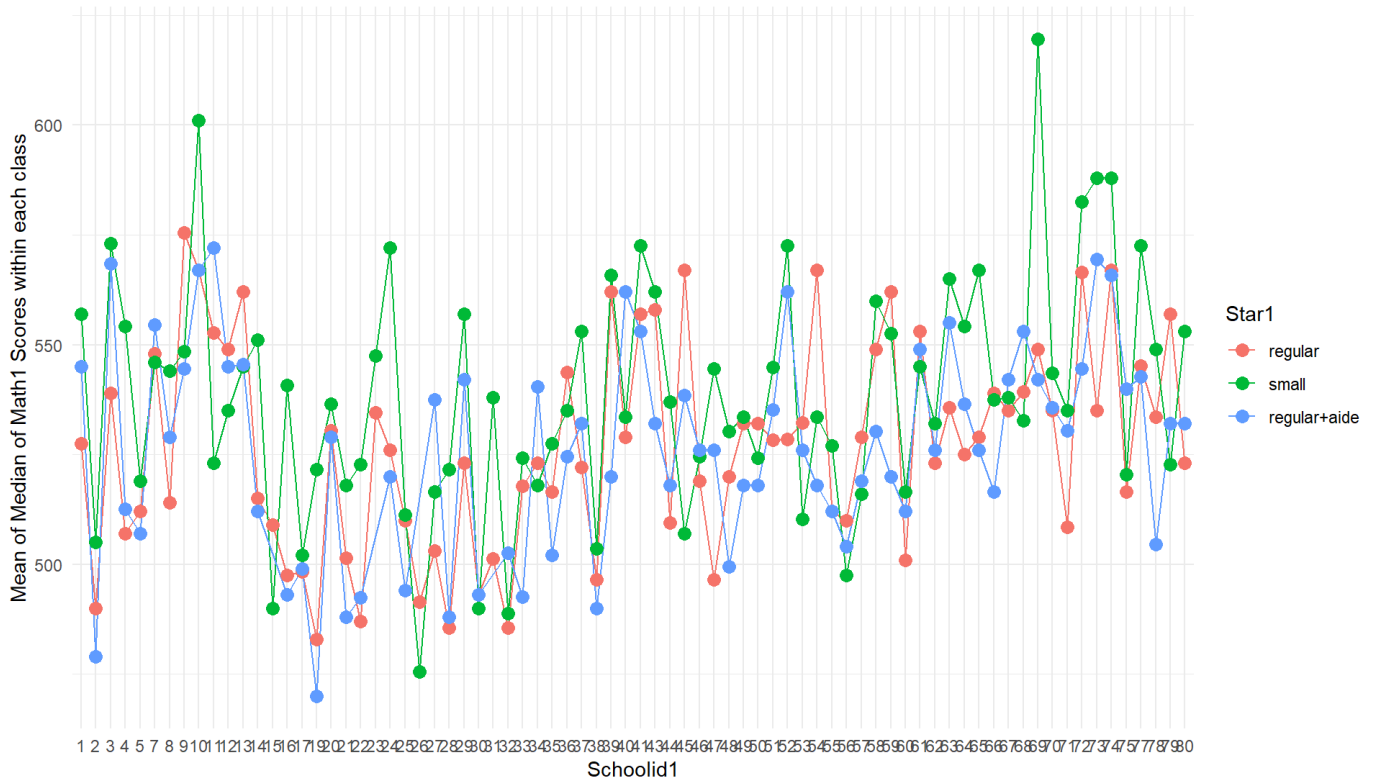
Having these two evenly weighted constraints, instead of the weighted constraints based on the number of observations per cell, is because these constraints simplify the model fitting process.

Why no interaction terms

Reason 1: similar patterns in the interaction plot

One reason of not including interaction terms in the model is because the interaction plot below shows that the math scores within the three class types share a similar pattern across all schools.

Interaction Plot: Effect of Star1 and Schoolid1 on Mean Math1 Scores

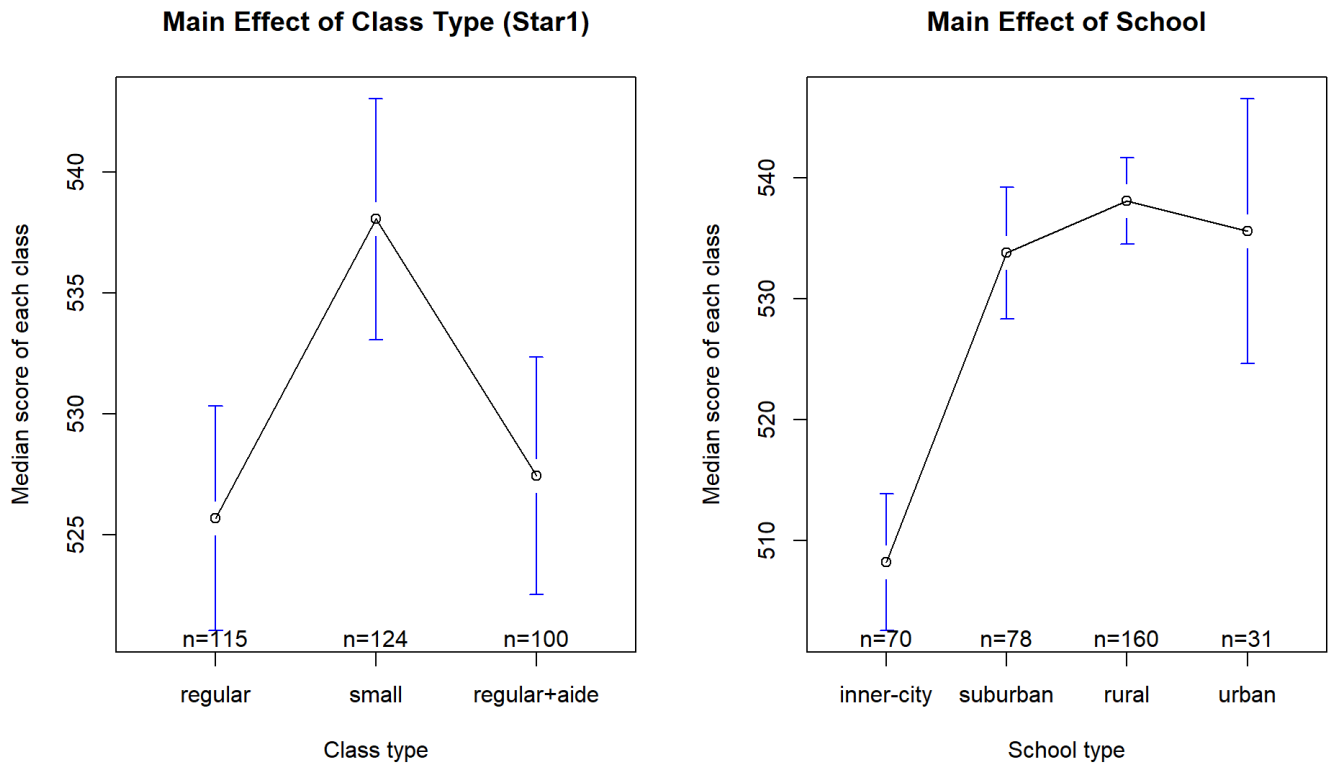


Reason 2: missing cells

Another reason of not including the interaction terms is that the observations do not cover all the combinations of the levels in each factor. Say, school 15 only has two “regular” class and one “small” class, without the class type “regular+aide”. Since the data set cannot guarantee there is at least one observation for all combinations of the two factors, introducing interaction terms into the two-way anova model would result in singularity problem when estimating the model’s parameters.

Visualization on the Main Effects

The main effects of the two factors are shown below. To make the plots clearer, I group the 76 schools based on their type/location.



The information shown by the two main effect plots align with that derived from the previous part.

- **Class type:** Students in small classes outperform those in regular and regular+aide.
- **School type:** Students in rural areas perform the best, while those in inner-city perform the poorest.

Model Fitting

Next, I fit the model on the data and below are the fitted results. To report the estimated coefficients for class types and school IDs, I change the encoding of the indicator variable to be: for all except the one corresponding to the last level, the encoded indicator variable is 1 if the case is from the current level, -1 if the case comes from the last level, and 0 otherwise.

```
# fit a linear regression model using lm. But the encoding of the factor levels should be changed
contrasts(new_star_sub$star1) <- contr.sum(levels(new_star_sub$star1))
contrasts(new_star_sub$schoolid1) <- contr.sum(levels(new_star_sub$schoolid1))
# Fit the linear model
model_1 <- lm(p50 ~ star1 + schoolid1, data = new_star_sub)
# View model summary
##summary(model_1)
```

Hypothesis Testing

Here are two hypotheses to test, corresponding to the main effect of the class type and the school id, respectively.

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0, H_a : \text{not all } \alpha_i = 0, i = 1, 2, 3$$

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{76} = 0, H_a : \text{not all } \beta_i = 0, i = 1, \dots, 76$$

Apart from the assumptions listed in the above model assumption part, in order to do the F tests, the required additional assumption is the normal error assumption, i.e. the random error term

$$\epsilon_{ijk}(\text{iid}) \sim N(0, \sigma^2), i = 1, 2, 3, j = 1, \dots, 76, k = 1, \dots, n_{ij}.$$

I specify the significance level as $\alpha = 0.01$, and according to the F tests' outcomes below, the p-values corresponding to the F values of the class type (star1) and the school id are both far below 0.01, showing that both the two factors are significant.

```
test_1 = anova(model_1)
test_1
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
star1	2	10638.80	5319.4017	17.11309	1.038427e-07
schoolid1	75	148459.34	1979.4578	6.36813	1.440354e-29
Residuals	261	81128.77	310.8382	NA	NA
3 rows					

- star1 has 2 degrees of freedom (DF) and is statistically significant ($p < 0.001$). The F-value is 17.1131 with $p = 1.038e-07$, meaning there are significant differences in 1st-grade math scores across class types.
- schoolid1 has 75 degrees of freedom and is also statistically significant ($p < 0.001$), meaning math scores vary significantly across schools. The F-statistic is 6.3681 with $p < 2.2e-16$, meaning significant difference across schools.
- The residual sum of squares is large, suggesting unexplained variability.

So, for the primary question, the answer is: there are significant differences in math scaled scores in 1st grade across all the 3 class types.

Notes: The alternative R function "aov"

Although "aov" is best used for balanced design, when the data is near balanced using "aov" can give us a rough picture on the significance of the factors involved. From the outputs below there is hardly any difference from those of using "anova". In the following report, I use the outputs from "aov" ("test_2") to do post-hoc tests such as TukeyHSD() and any further sensitivity analyses.

```
test_2 = aov(p50 ~ star1 + schoolid1, data = new_star_sub)
test_2
```

```
## Call:
##   aov(formula = p50 ~ star1 + schoolid1, data = new_star_sub)
##
## Terms:
##               star1 schoolid1 Residuals
## Sum of Squares  10638.80 148459.34  81128.77
## Deg. of Freedom      2         75      261
##
## Residual standard error: 17.6306
## Estimated effects may be unbalanced
```

Pairwise Comparison

For the secondary question of interest, which class type is associated with the highest math scaled scores in 1st grade, I use Tukey's Range Test to do the pairwise comparison.

The TukeyHSD method requires the assumptions that:

1. The error terms within the model are i.i.d random variables under $N(0, \sigma^2)$

2. The sample sizes within each cell are equal. But since the design of the data can be treated as slightly unbalanced, we can still use the method for pairwise comparison.

The null hypothesis and the alternative hypothesis are:

H_0 : Any two means from different cells are equal, H_1 : Not any two different means are equal

```
tukey_results <- TukeyHSD(test_2, conf.level=0.99)
print(tukey_results$star1)
```

```
##                diff      lwr      upr      p adj
## small-regular    12.357714   5.649148 19.066280 4.178050e-07
## regular+aide-regular  1.757391 -5.327951  8.842733 7.465094e-01
## regular+aide-small -10.600323 -17.565043 -3.635603 3.415227e-05
```

I specify the significance level as $\alpha = 0.01$.

key observations are:

- *Regular vs. Small* ($p < 0.01$) Students in small classes perform ~12 points higher than those in regular classes. This difference is statistically significant.

- *Regular vs. Regular + Aide* ($p = 0.7465$) There is no significant difference between Regular and Regular + Aide groups.

- *Small vs. Regular + Aide* ($p < 0.01$) Small classes significantly outperform Regular + Aide by ~10 points. This difference is statistically significant.

In conclusion, based on the pairwise comparison test, the “small” class type is associated with the highest math scaled scores in 1st grade.

Sensitivity analysis

Visualized Residual Diagnostics

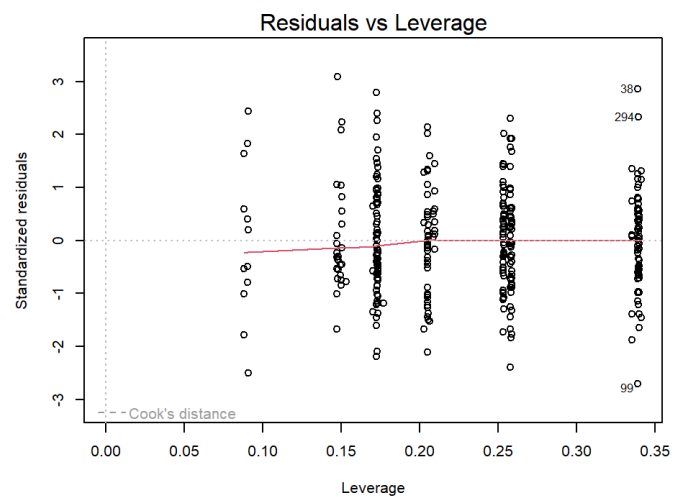
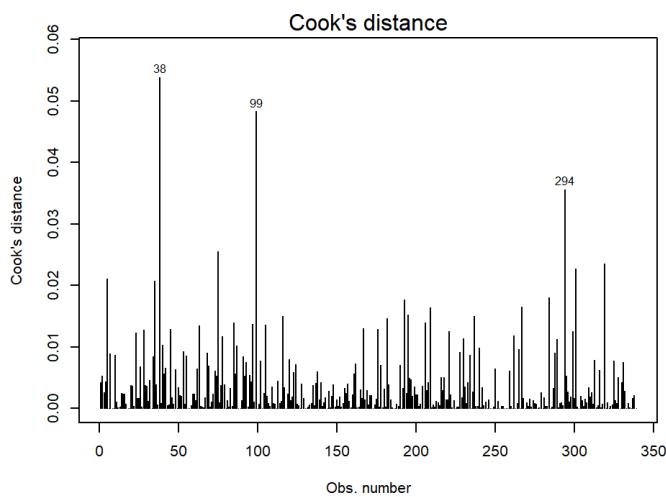
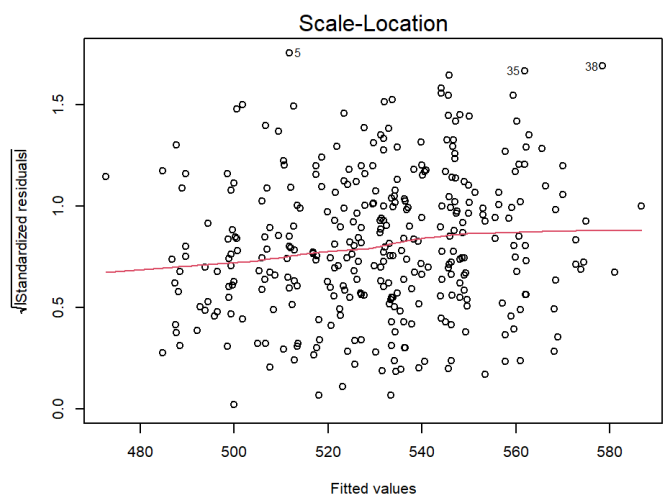
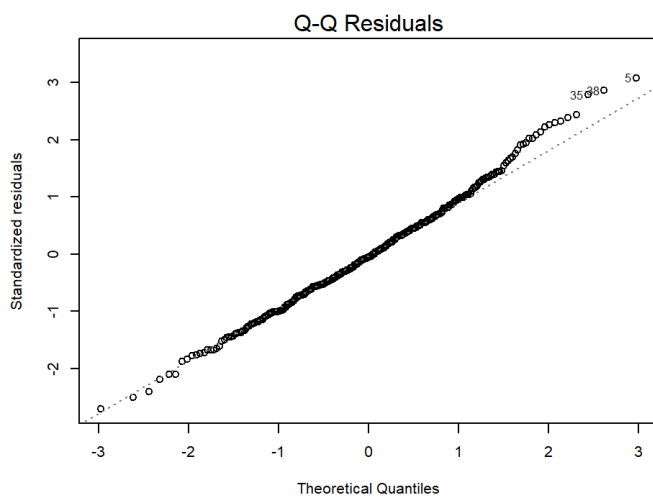
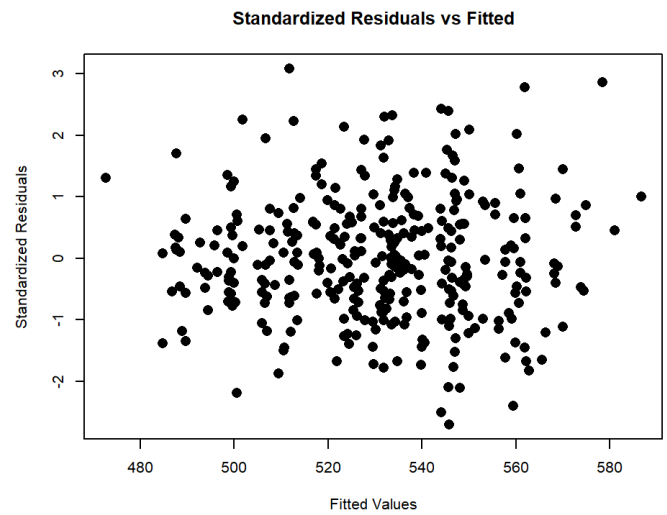
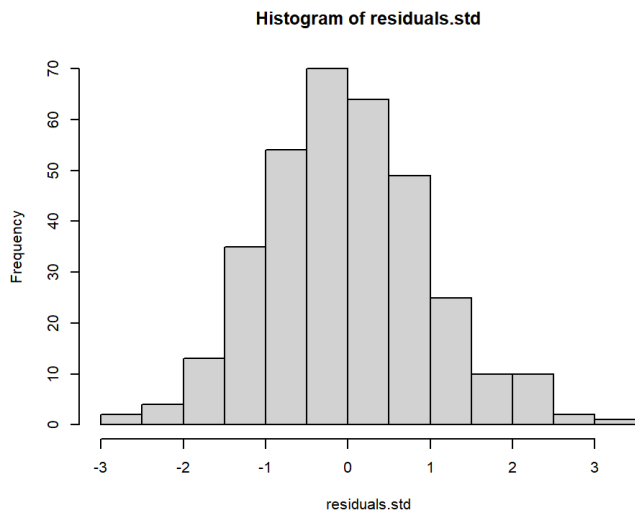
I examine the residual plots of the fitted model “test_2”. From the Standardized Residuals v.s. Fitted plot and the Standardized Residuals’ Q-Q plot, it is found that:

1. there is no obvious pattern of different fitted values impacting the dispersion of the standardized residuals in the scatter plot.
2. the standardized residuals are slightly heavily-tailed distributed.

These visualizations suggest that the model assumptions roughly hold:

- The assumption that “the two factors’ effects are additive” is satisfied.
- The assumption that “error terms are iid random variables” also holds, though their distribution may deviate from normality, which requires further inspection.

Regarding the outliers, only a few observations deviate from the model. Leverage is less meaningful in this context because, in ANOVA, the design matrix X is encoded based on the experimental design rather than directly measured values, as in conventional linear regression models.



The visualizations show that the diagnostics on the model assumptions require further Hypothesis Testing with respect to the normality and the equal variance of the error terms, which is the focus of the following part.

Hypothesis Testing on the Error Terms

Next, I perform Hypothesis Testing on the error terms, with respect to their normality and equal variance across different cells. The spoiler is both of the assumptions hold under a significance level of 0.05.

the Normality Check

I use the Shapiro-Wilk test, the Kolmogorov-Smirnov (K-S) Test, and the Anderson-Darling Test to perform the normality check. The null and alternative hypotheses are:

$$H_0 : \epsilon_{ijk} \sim N, H_1 : \epsilon_{ijk} \text{ not } \sim N$$

The boiler is that the error terms pass all three normality tests under a significance level of 0.05, indicating that although the distribution is lightly heavily-tailed, it is roughly normal.

In specific, the p-value for the Shapiro-Wilk Test is 0.2384, and that for the Kolmogorov-Smirnov Test is 0.9202, and that for the Anderson-Darling Test is 0.2471, all above 0.05.

Conclusion

The outcomes of all three tests indicate that the normality of error terms in the original model holds under the significance level of 0.05.

the Equal Variance Check

In the following I use the Levene test to validate if the equal variance assumption on the error terms in the model really holds. The boiler is the equal variance null hypothesis should not be rejected under a significance level of 0.05.

Levene Test

Choosing the “Levene Test” is because it is the most robust method for the unbalanced two-way ANOVA model. It performs the F-test on a new one-way ANOVA model with the absolute values of the residuals of the original model as the response and each of the original cells as the factor.

Here I choose the standardized residuals rather than the raw residuals because theoretically only the standardized residuals can be compared with each other on their variance.

The null and alternative hypotheses are: H_0 : The means of the absolute standardized residuals within each cell are all the same; H_a : The means of the absolute standardized residuals within each cell are not the same.

```
new_star_sub$Group <- interaction(new_star_sub$star1, new_star_sub$schoolid1)
new_star_sub$resiabs = abs(residuals.std)
summary(aov(resiabs~Group, data=new_star_sub))
```

##		Df	Sum Sq	Mean Sq	F value	Pr(>F)
##	Group	223	82.70	0.3709	0.982	0.55
##	Residuals	115	43.41	0.3775		

I set the significance level as $\alpha = 0.05$. From the result, the p value is 0.55 (> 0.05), indicating that the null hypothesis should not be rejected.

Conclusion

The equal variance assumption on the error terms in the model holds under 0.05 significance level.

Alternative Methods

In this section, I use other methods to test the robustness of the two-way ANOVA model proposed above.

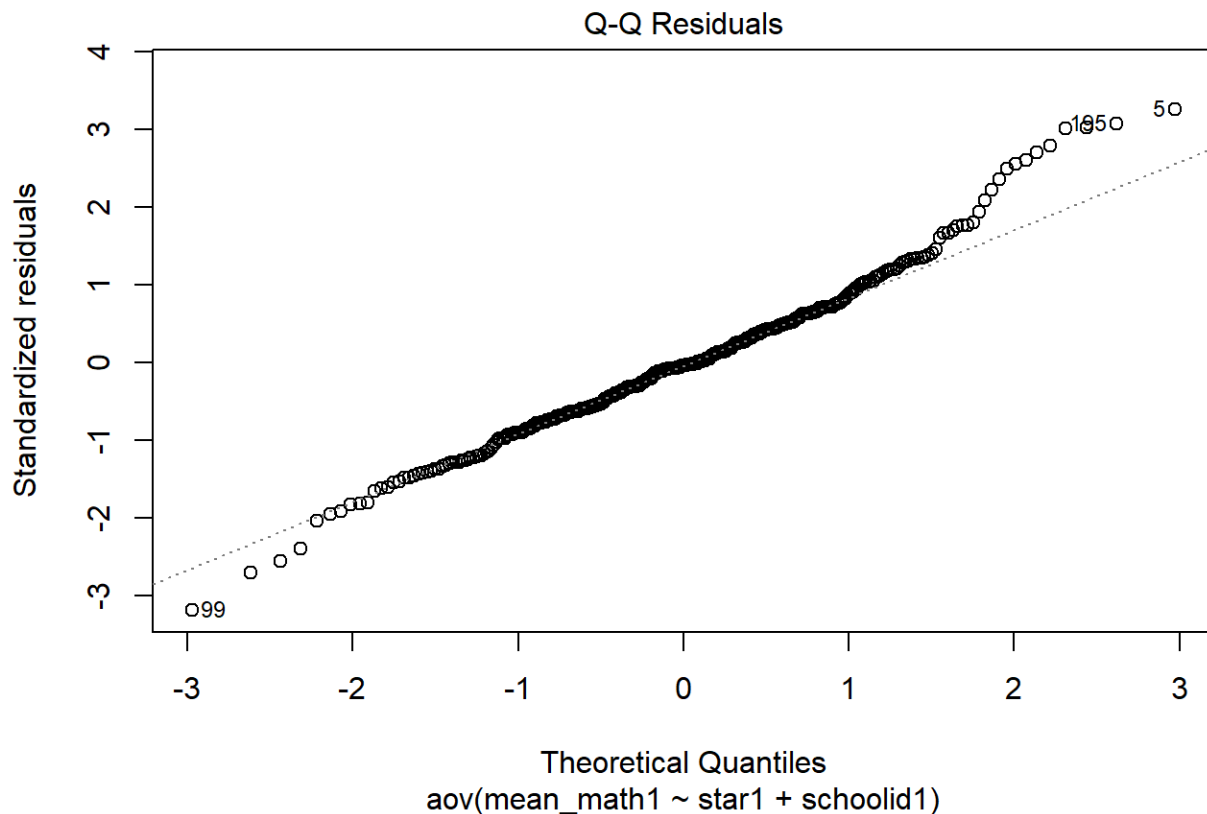
Use Mean Scores in ANOVA

First, I replace the median scores of each class with the mean scores. Then, I do two-way ANOVA as before.

In short, the new model is still significant (significance at 0.01) on both factors, and the Tukey's range test still shows that the "small" class outperforms the other two types.

However, the new model built using mean scores exhibits a weaker adherence to the normality assumption of the error terms.

- The QQ plot on the standardized residuals shows that the new model's error terms may not meet the normality assumption.



- The p-value for the Shapiro-Wilk Test is 0.0001, and that for the Anderson-Darling Test is 0.0007, much lower than 0.01, indicating the violation of the error terms on the normality assumption.

In conclusion, the results derived by the ANOVA built on the median scores are robust, and that the STAR data fits better in the ANOVA built on the median scores rather than the mean scores.

Acknowledgement

The author thanks Prof. Shizhe Chen, and the teaching assistant Yanhao Jin for their thought-provoking instruction. Also thanks her friend Ziyue Yang for educational discussion.

The author also uses ChatGPT for refining the reports but claims that the statistical judgement and ideas throughout all the report are original from herself.

Reference

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)

Session info

```
sessionInfo()
```

```
## R version 4.4.1 (2024-06-14 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 11 x64 (build 26100)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=Chinese (Simplified)_China.utf8
## [2] LC_CTYPE=Chinese (Simplified)_China.utf8
## [3] LC_MONETARY=Chinese (Simplified)_China.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=Chinese (Simplified)_China.utf8
##
## time zone: America/Los_Angeles
## tzcode source: internal
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] nortest_1.0-4      gplots_3.2.0      kableExtra_1.4.0  ggplot2_3.5.1
## [5] knitr_1.49         skimr_2.1.5        dplyr_1.1.4       AER_1.2-14
## [9] survival_3.6-4     sandwich_3.1-1     lmtest_0.9-40     zoo_1.8-12
## [13] car_3.1-3          carData_3.0-5
##
## loaded via a namespace (and not attached):
## [1] sass_0.4.9         utf8_1.2.4         generics_0.1.3     tidyr_1.3.1
## [5] bitops_1.0-9       xml2_1.3.6         KernSmooth_2.23-24 gtools_3.9.5
## [9] stringi_1.8.4      lattice_0.22-6     caTools_1.18.3     digest_0.6.37
## [13] magrittr_2.0.3     evaluate_1.0.0     grid_4.4.1         fastmap_1.2.0
## [17] jsonlite_1.8.9     Matrix_1.7-0       Formula_1.2-5      purrr_1.0.2
## [21] fansi_1.0.6        viridisLite_0.4.2  scales_1.3.0       jquerylib_0.1.4
## [25] abind_1.4-8        cli_3.6.3          rlang_1.1.4        munsell_0.5.1
## [29] splines_4.4.1      base64enc_0.1-3    withr_3.0.1        repr_1.1.7
## [33] cachem_1.1.0       yaml_2.3.10        tools_4.4.1        colorspace_2.1-1
## [37] vctrs_0.6.5        R6_2.5.1           lifecycle_1.0.4    stringr_1.5.1
## [41] pkgconfig_2.0.3    gtable_0.3.5       pillar_1.9.0       bslib_0.8.0
## [45] glue_1.8.0         systemfonts_1.1.0  xfun_0.50          tibble_3.2.1
## [49] tidyselect_1.2.1   rstudioapi_0.16.0  farver_2.1.2       htmltools_0.5.8.1
## [53] svglite_2.1.3      labeling_0.4.3     rmarkdown_2.28     compiler_4.4.1
```