A Study on the Relationship between Class Type and Math Scores

STA 207 Project STAR II, Jan 31, 2020

Group ID:

Introduction

Tennessee Student/Teacher Achievement Ratio study (Project STAR) is a four-year longitudinal class-size study with randomized experimental design. Students from 79 schools were randomly assigned to one of three class types: small, regular, regular with aide; while the classroom teachers were also randomly assigned to classes of different types. The information of each student spans over demographic factors, school and class IDs, schools' and teachers' information, experimental conditions, test scores, motivation, etc. In this study, we will only work with variables specific to first-grader, in an attempt to answer the following questions:

- Is there a significant difference between teachers' performance across different class types, if we measure it as the average scaled math scores for the first-graders?
- Knowing the block design of the experiment, what would be an appropriate model for our purposes? Are the model assumptions satisfied?
- If the difference is significant, can we interpret it as a causal effect? And if so, under what conditions?

Statistical Analysis

Exploratory Data Analysis

Among all 11,601 students, 6,563 of them have complete data of scaled math scores, school IDs, and teacher information in the first grade. 337 classroom teachers from 76 schools were randomly assigned to teach classes of a particular type. Upon primary analysis, the pie chart in Figure 3 indicates that the teachers are roughly equally split across the three types.
By the experiment design, the number of first-graders for which each teacher is responsible can vary from class to class, resulting in different numbers of scaled math scores. To effectively evaluate the performance, the average values of the scores of all students by each teacher were considered as the performance measure. The teachers' performance measure is roughly bell-shaped (4). Teachers assigned with small classes appear to perform better on average per box plot (Figure 1), while displaying a wider range than the other two types. Notice teachers are not evenly distributed across the schools and there are also missing math scores for regular with aide type within some schools (Figure 5 and Table 3). More than half of the schools had 4 teachers participating in the project, around 30 schools had 5 to 7 teachers, and only 1 school had 12 teachers in the

project. Such heterogeneity suggests that further analysis will revolve around unbalanced and incomplete experimental design.

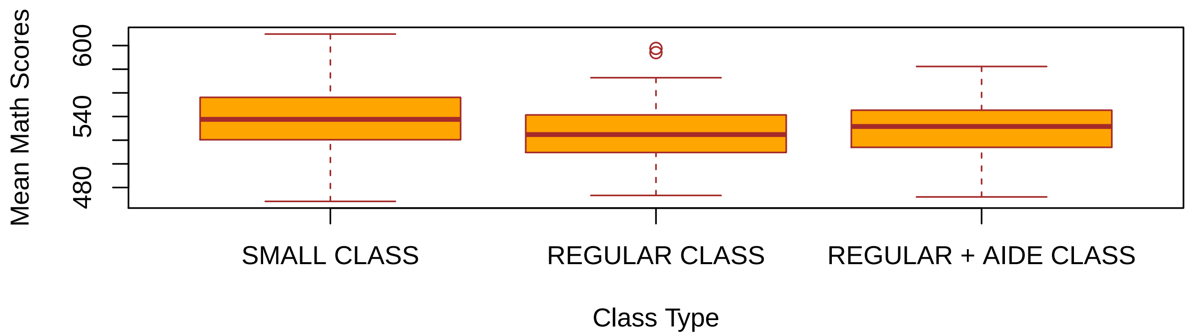

Figure 1: Boxplots for mean math scores by class types

Model Specification

For this incomplete, unbalanced experimental design, a multiple linear regression form of two-way ANOVA additive model is adopted. This allows for comparison of the mean scaled math scores of each teacher among three treatment groups, while taking into account the schools as the blocking factor.

$$Y_i = \mu + \sum_{k=2}^{3} \gamma_k X_{k,i} + \sum_{j=2}^{76} \rho_j Z_{j,i} + \epsilon_i, \; \forall i = 1, \cdots, 337,$$

Y_i is the i-th teacher's average scaled math scores of first-graders in the class. μ is the true mean scaled math scores of first-graders in the baseline class type (i.e. small) of the baseline school (i.e. school ID: 112038). If class type k is assigned to the i-th teacher, then the indicator variable $X_{k,i} = 1$; otherwise $X_{k,i} = 0$. The corresponding coefficient γ_k is the increment of the mean of average scaled math scores for teacher of class type k from that of small class type, where $k = 2$ corresponds for regular type and $k = 3$ corresponds for regular with aide type. If the i-th teacher is in the j-th school, then $Z_{j,i} = 1$; otherwise, $Z_{j,i} = 0$. ρ_j represents the increment of the teachers' average scaled math scores overall class types in school j from the baseline school 112038. Lastly, ϵ_i's are the error terms assumed to follow $N(0, \sigma^2)$ i.i.d.
Since we are interested in the overall effect of class type on the teachers' performance across all schools (i.e. blocks) rather than teachers in any one particular school, it is safe for us to omit the interaction terms in that they only provide information on the increment of the effect of class type from the related schools. To substantiate our claim here, we will test on the interaction terms in the `Hypothesis Tests` section and rigorously examine whether they should be included in the model.

Model Estimates

The ANOVA table of our model is reported in Table 1. The estimates of the model coefficients are reported in the appendix for concision/brevity.

Table 1: ANOVA Table (Type II Sum of Squares)

	Sum of Squares	Degree of Freedom	F-value	P-value
Class Type	11779.56	2	21.20	0
Schools	134441.97	75	6.45	0
Residuals	71958.72	259		

Model Diagnostics

- **Normality**
  The histogram of the residuals (Fig. 2 left) is roughly bell-shaped. The Normal Q-Q plot (Fig. 2 middle) shows more probabilities on both tails, implying a heavy-tailed distribution. Since the Shapiro-Wilk test returns a p-value of 0.0001, we reject the normality assumption although the residuals do not depart severely from being normally distributed.

- **Homoscedasticity**
  Per Figure 2 (right), variance does not seem to differ across different fitted values. More formally, we examine this observation with Levene's test, where the null hypothesis states the variance is the same for all groups. Since we are not interested in the effect of the blocks (schools) , we limit the test to the class types. The result p-value of 0.45 implies that we fail to reject the equal variance assumption at the 0.05 level.

- **Independence**
  The design of the experiment implies that the students were randomly assigned to different class types and that each teacher was also randomly assigned a class type within each school. Therefore, we are assured that the error terms are independent of each other.

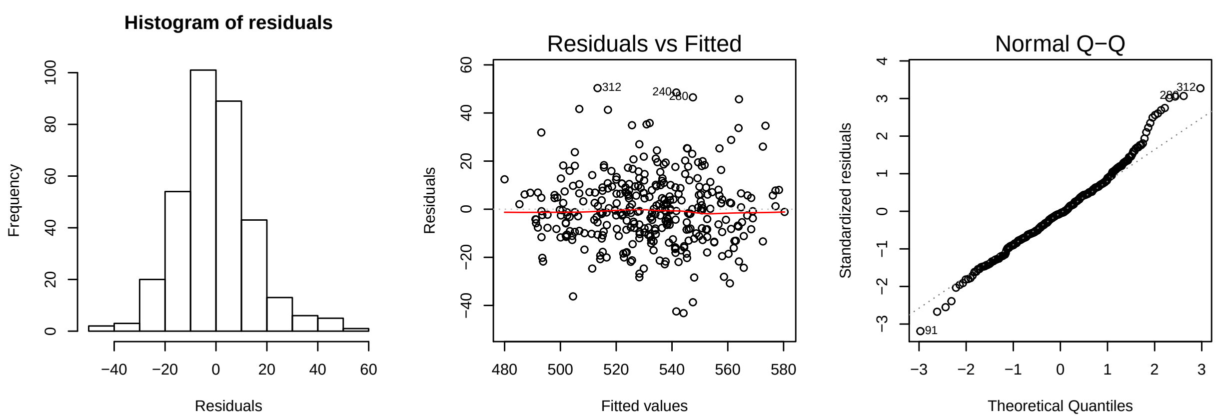

Figure 2: Model Diagnostics

Hypothesis Tests

The Existence of the Interaction Terms

To formally study the existence of the interaction between the schools and the class types, we construct a model that includes all the interaction terms, which we use as the full model to perform the following F-test. H_0 : The full model is not different from the reduced model VS. H_A : The two models are significantly different, where the reduced model is the two-way ANOVA model in our analysis.

Table 2: ANOVA Table (Existence of the Interaction Terms)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
259	71958.72				
113	33220.75	146	38737.96	0.9	0.72

We reject H_0 if the `anova()` function in R returns a p-value less than 0.05. Per Table 2 above, p-value is 0.72. We fail to reject the hypothesis that the two models are not different at the 0.05 level. No obvious interaction effects exist between the schools and the class types.

Nonparametric Tests on the Group Means

Let μ_1 denote the mean performance, in terms of the average scaled math score, of the teachers in the small classes, μ_2 the mean performance of teachers in regular classes, and μ_3 that of the teachers in regular classes with-aide. To test whether the average performance of the teachers is the same across different class type assignments, we would ideally perform the ANOVA F-test. However, since normality assumption is violated as mentioned above, we resort to nonparametric tests that do not rely on the normality in the error terms.

- **Rank Test**
  $H_0 : \mu_1 = \mu_2 = \mu_3$ VS. H_A : Not all μ_k's are equal, $\forall k = 1, 2, 3$. Test statistic: $F^* = \frac{MSTR_r}{MSE_r} \overset{H_0}{\sim} F(2, 334)$, where $MSTR_r, MSE_r$ are obtained from the model with the rank of the old response variable (i.e. teacher's performance) as the new response variable. Reject H_0 if $Pr(F^* > F(0.95; 2, 334)) < 0.05$. Since the p-value turns out to be 0.0003, we reject H_0 at the 0.05 level, and conclude that the teachers' average performance of at least one class type is different from others.

- **Kruskal-Wallis Test**
  $H_0 : \mu_1 = \mu_2 = \mu_3$ VS. H_A : Not all μ_k's are equal, $\forall k = 1, 2, 3$. Test statistic:

  $$H = (N-1)\frac{\sum_{i=1}^{3} n_i \cdot (\overline{r}_{i\cdot} - \overline{r})^2}{\sum_{i=1}^{3}\sum_{j=1}^{n_i}(r_{ij} - \overline{r})^2} \overset{H_0}{\sim} \chi^2_2,$$

  where n_i is the number of teachers in class type i; r_{ij} is the rank (among all observations) of the j-th teacher in class type i; N is the total number of teachers; $\overline{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all teachers in class type i; and $\overline{r} = \frac{1}{2}(N+1)$ is the average rank of all teachers. Reject H_0 if $Pr(\chi^{2\,*} > \chi^2_2(0.95)) < 0.05$. Since the p-value is computed to be 0.0003, we reject H_0 at the 0.05 level, and conclude that the teachers' average performance of at least one class type is different from others.

Note that we have obtained the same result using both the rank test and the Kruskal-Wallis test, we believe that the significant difference in the teachers' average performance across different class types is consistent, which leads to our following discussion of the post-hoc analysis, i.e. to identify the exact class type(s) different from others.

Post-hoc Analysis: Multiple Testing

Now that we have discovered that some class type displays different average teachers' performance from others, we will conduct further hypothesis tests to detect where the significant difference lies. Similar to above, we will perform multiple testing using two methods to see if the same result persists.
Note: In this section, we will use the related R functions that produce p-values directly, hence the statement

of the test statistics and their null distributions will be omitted; albeit, the following hypotheses and decision rule hold for both tests: $H_0 : \mu_i = \mu_j$, VS. $H_A : \mu_i \neq \mu_j$ $\forall i, j = 1, 2, 3$ and $i \neq j$ Reject H_0 if p-value< 0.05.

- **Bonferroni's Procedure**
  As is reported in Table 4, we reject $H_0 : \mu_1 = \mu_2$ and $H_0 : \mu_1 = \mu_3$ at the 0.05 level since the p-values are smaller than 0.05. Therefore, the average performance of teachers in small-sized classes is significantly different from those in regular-sized classes and regular-sized classes with the aide. Nonetheless, we do not have evidence against the hypothesis that the average teachers' performance of regular classes is not different from that of regular classes with-aide.

- **Tukey's Procedure**
  Alternatively, we could test the same hypotheses with a Tukey's procedure following the same decision rule as above. P-values are reported in the last column of the Table 5. We reject the null hypothesis at 0.05 same as above and arrive at identical conclusions to the Bonferroni's procedure.

Therefore, we conclude that the teachers' average performance differs across different class types. In particular, teachers of small classes tend to perform better than those of regular classes and regular classes with-aide, the latter two being statistically indistinguishable, in terms of the average scaled math scores. In the following section, we will investigate the experimental design and examine the assumptions of the causal inference to determine whether this significant difference can be interpreted as a causal effect.

Discussion

To determine if causal inference can be drawn on the effect of class type on the scaled math scores, we will examine the following assumptions under the potential outcome framework:

- Stable Unit Treatment Value Assumption (SUTVA)
  No spillover effect: We believe that the students' math performance solely depended on their effort and classroom learning. Provided students were randomly assigned to the classes, this assumption is likely to hold in that the learning outcome of one class hardly depended on that of another class.
  Same version of treatment: The definition of each class type was clear. Randomization implies that the teachers were homogeneous in all characteristics across class types. Additionally, the teachers taught the same materials.

- Ignorability: This unconfoundedness assumption holds by randomization and full compliance of the treatment assignment on teachers and students.

Conclusion

Through our analysis, in terms of the average scaled math scores, we have found that on average, teachers of small classes tend to perform better than those of regular classes and regular classes with-aide, while the latter two are not significantly different. This result is slightly different results from STAR I, where the pairwise test indicated that students' average scaled math scores were different for all class types. We suspect that this is due to the extent to which the subjects react to the treatment assignment. For instance, students may respond more actively to any change in the classroom setting, while teachers are less sensitive to the subtle difference. Further research is appropriate to uncover the nature of this discrepancy in the treatment effect. Following the investigation on the assumptions of causal inference based on the design of the experiment, we conclude that the effect of class type on the teachers' performance is indeed causal; that is, small class size causes the teacher to have better average scaled math scores among the students than regular class sizes, whether with or without the aide. As a final note, we have not been able to explore the time series dimension of this data set limited by the time constraint and our knowledge base. Nevertheless, we believe that if we research more on the panel data methods, e.g. fixed effect models, and implement them in future work, we would find more interesting patterns alongside more meaningful results.

Appendix

Graphs and Tables

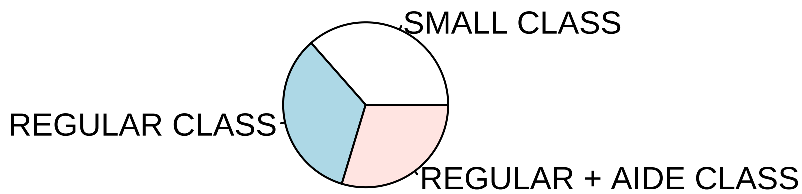

Figure 3: Proportion of Teachers in Each Class Type

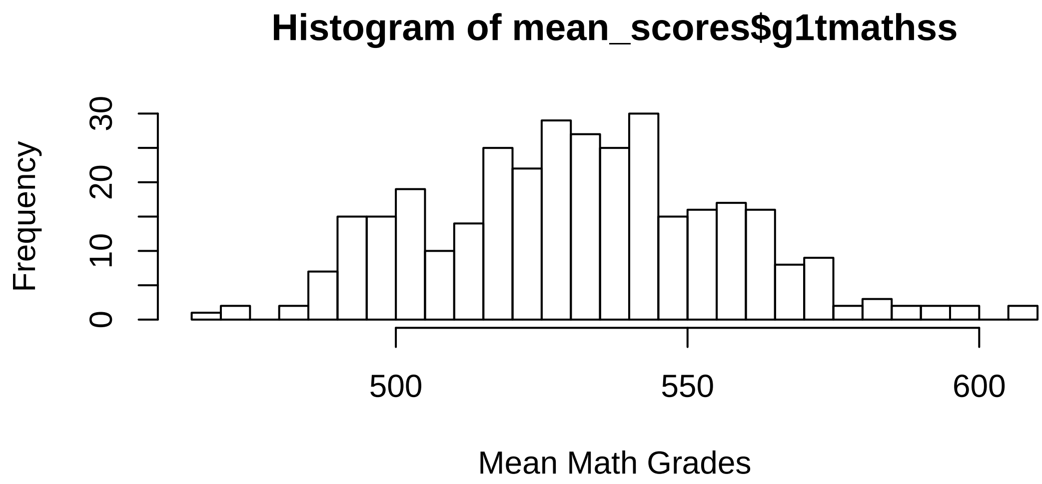

Figure 4: Histogram of 1st-Grader Mean Math Score

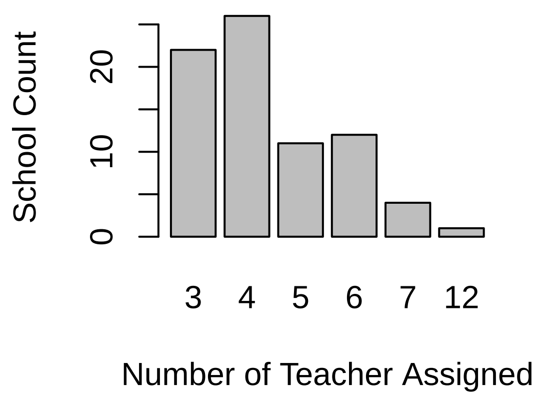

Figure 5: Frequency Plot: Frequency of Schools by Teacher Count

Table 3: Contigency Table of School and Class Size

	Small	Regular	Regular w/ Aide
112038	1	1	1
123056	1	1	1
128076	2	1	1
128079	2	1	1
...	...	...	...
244728	2	1	0
244736	1	2	0
...	...	...	...

Table 4: Bonferroni's Procedure

	SMALL CLASS	REGULAR CLASS
REGULAR CLASS	0.0001977	
REGULAR + AIDE CLASS	0.0157043	0.8523145

Table 5: Tukey's Procedure

	diff	lwr	upr	p adj
REGULAR CLASS-SMALL CLASS	-13.062167	-18.17039	-7.953943	0.0000000
REGULAR + AIDE CLASS-SMALL CLASS	-9.408044	-14.69859	-4.117496	0.0001123
REGULAR + AIDE CLASS-REGULAR CLASS	3.654123	-1.72926	9.037505	0.2474400

Outputs

```
##
##  Shapiro-Wilk normality test
##
## data:  mod1$residuals
## W = 0.9802, p-value = 0.0001357
```

- Rank Test

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## g1classtype   2  152390   76195    8.38 0.000281 ***
## Residuals   334 3036977    9093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Levene Test

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.8112 0.4452
##       334
```

- Kruskal-Wallis Test

```
##
##  Kruskal-Wallis rank sum test
##
## data:  rank by g1classtype
## Kruskal-Wallis chi-squared = 16.054, df = 2, p-value = 0.0003265
```

- Raw Output of Linear Regression

```
##
## Call:
## lm(formula = g1tmathss ~ g1classtype + as.factor(g1schid), data = mean_scores)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.208  -9.272  -0.814   7.807  50.343
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                     502.456      9.708  51.757  < 2e-16 ***
## g1classtypeREGULAR CLASS        -13.251      2.203  -6.014 6.14e-09 ***
## g1classtypeREGULAR + AIDE CLASS -11.380      2.288  -4.973 1.20e-06 ***
## as.factor(g1schid)123056          36.091     13.610   2.652 0.008498 **
## as.factor(g1schid)128076          31.729     12.735   2.492 0.013345 *
## as.factor(g1schid)128079          21.203     12.735   1.665 0.097119 .
## as.factor(g1schid)130085          61.138     12.735   4.801 2.68e-06 ***
## as.factor(g1schid)159171          50.141     11.786   4.254 2.93e-05 ***
## as.factor(g1schid)161176          31.026     12.735   2.436 0.015514 *
## as.factor(g1schid)161183          74.653     11.786   6.334 1.05e-09 ***
## as.factor(g1schid)162184          47.834     12.735   3.756 0.000213 ***
## as.factor(g1schid)164198          47.174     13.610   3.466 0.000618 ***
## as.factor(g1schid)165199          75.889     13.610   5.576 6.17e-08 ***
## as.factor(g1schid)166203          17.127     13.610   1.258 0.209376
## as.factor(g1schid)168211          46.235     12.735   3.630 0.000341 ***
## as.factor(g1schid)168214          71.058     13.610   5.221 3.65e-07 ***
## as.factor(g1schid)169219          58.750     12.176   4.825 2.39e-06 ***
## as.factor(g1schid)169229          39.090     10.761   3.633 0.000338 ***
## as.factor(g1schid)169231          30.937     12.176   2.541 0.011643 *
## as.factor(g1schid)169280          42.783     12.735   3.360 0.000898 ***
## as.factor(g1schid)170295          73.687     12.735   5.786 2.07e-08 ***
## as.factor(g1schid)173312          56.185     12.735   4.412 1.50e-05 ***
## as.factor(g1schid)176329          43.525     12.735   3.418 0.000733 ***
## as.factor(g1schid)180344          43.346     11.786   3.678 0.000286 ***
## as.factor(g1schid)189378          33.157     12.735   2.604 0.009757 **
## as.factor(g1schid)189382          47.250     12.735   3.710 0.000253 ***
## as.factor(g1schid)189396          23.750     12.735   1.865 0.063312 .
## as.factor(g1schid)191411           9.680     13.610   0.711 0.477556
## as.factor(g1schid)193422          34.149     13.610   2.509 0.012713 *
## as.factor(g1schid)193423          25.896     12.176   2.127 0.034378 *
## as.factor(g1schid)201449          55.155     11.504   4.795 2.75e-06 ***
## as.factor(g1schid)203452          48.217     12.176   3.960 9.69e-05 ***
## as.factor(g1schid)203457          49.248     13.610   3.619 0.000356 ***
## as.factor(g1schid)205488          27.431     12.735   2.154 0.032158 *
## as.factor(g1schid)205490          41.694     13.610   3.064 0.002418 **
## as.factor(g1schid)205491          50.287     13.610   3.695 0.000268 ***
## as.factor(g1schid)205492          26.318     13.610   1.934 0.054233 .
## as.factor(g1schid)208501          38.826     12.735   3.049 0.002535 **
## as.factor(g1schid)208503          34.144     13.610   2.509 0.012726 *
## as.factor(g1schid)209510          35.379     11.786   3.002 0.002947 **
## as.factor(g1schid)212522          23.213     12.176   1.907 0.057687 .
## as.factor(g1schid)215533          58.296     11.786   4.946 1.36e-06 ***
## as.factor(g1schid)216537          61.550     11.786   5.222 3.63e-07 ***
## as.factor(g1schid)218562          54.216     12.735   4.257 2.89e-05 ***
## as.factor(g1schid)221571          14.100     11.786   1.196 0.232673
## as.factor(g1schid)221574          29.077     12.735   2.283 0.023222 *
## as.factor(g1schid)225585          30.424     12.735   2.389 0.017606 *
## as.factor(g1schid)228606          77.890     12.735   6.116 3.52e-09 ***
## as.factor(g1schid)230612          59.957     13.610   4.405 1.55e-05 ***
## as.factor(g1schid)231616          40.821     13.610   2.999 0.002968 **
## as.factor(g1schid)234628          65.939     11.786   5.595 5.62e-08 ***
## as.factor(g1schid)244697          17.550     11.786   1.489 0.137700
## as.factor(g1schid)244708           2.079     11.786   0.176 0.860102
## as.factor(g1schid)244723          13.156     11.786   1.116 0.265375
```

```
## as.factor(g1schid)244727          36.688     12.176   3.013 0.002841 **
## as.factor(g1schid)244728          -9.225     13.631  -0.677 0.499167
## as.factor(g1schid)244736          22.210     13.632   1.629 0.104472
## as.factor(g1schid)244745          10.985     12.735   0.863 0.389161
## as.factor(g1schid)244746          51.936     13.610   3.816 0.000170 ***
## as.factor(g1schid)244755          10.868     11.508   0.944 0.345866
## as.factor(g1schid)244764          32.327     13.610   2.375 0.018264 *
## as.factor(g1schid)244774           2.723     12.176   0.224 0.823228
## as.factor(g1schid)244776           4.315     11.786   0.366 0.714555
## as.factor(g1schid)244780          -3.883     13.610  -0.285 0.775627
## as.factor(g1schid)244796          11.811     13.632   0.866 0.387078
## as.factor(g1schid)244799          43.112     12.735   3.385 0.000821 ***
## as.factor(g1schid)244801           8.715     12.735   0.684 0.494358
## as.factor(g1schid)244806          25.961     11.504   2.257 0.024856 *
## as.factor(g1schid)244831          12.325     12.735   0.968 0.334038
## as.factor(g1schid)244839          48.526     12.194   3.980 8.97e-05 ***
## as.factor(g1schid)252885          44.835     12.735   3.521 0.000509 ***
## as.factor(g1schid)253888          38.680     13.610   2.842 0.004839 **
## as.factor(g1schid)257899          34.572     12.176   2.839 0.004879 **
## as.factor(g1schid)257905          55.525     11.504   4.827 2.37e-06 ***
## as.factor(g1schid)259915          33.355     12.735   2.619 0.009332 **
## as.factor(g1schid)261927          35.734     12.176   2.935 0.003636 **
## as.factor(g1schid)262937          70.128     12.735   5.507 8.79e-08 ***
## as.factor(g1schid)264945          51.282     12.176   4.212 3.50e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.67 on 259 degrees of freedom
## Multiple R-squared:  0.6687, Adjusted R-squared:  0.5702
## F-statistic: 6.788 on 77 and 259 DF,  p-value: < 2.2e-16
```

Reference

Imai, K. Tingley, D. and Yamamoto, T. (2013) Experimental designs for identifying causal mechanisms. J. R. Statist. Soc., A, 176 Part 1, pp.5-51.

Kruskal; Wallis (1952). "Use of ranks in one-criterion variance analysis". Journal of the American Statistical Association. 47 (260): 583–621. doi:10.1080/01621459.1952.10483441

https://www.r-bloggers.com/r-tutorial-series-two-way-anova-with-unequal-sample-sizes/

https://rtutorialseries.blogspot.com/2011/01/r-tutorial-series-two-way-anova-with.html

http://www.real-statistics.com/two-way-anova/two-factor-anova-with-replication/brown-forsythe-f-test-two-way-anova/

https://stattrek.com/statistics/dictionary.aspx?definition=randomized%20block%20design

https://www.classsizematters.org/wp-content/uploads/2016/09/STAR-Technical-Report-Part-I.pdf

Session Information

```
## R version 3.6.0 (2019-04-26)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.6 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/atlas-base/atlas/libblas.so.3.0
## LAPACK: /usr/lib/atlas-base/atlas/liblapack.so.3.0
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] kableExtra_1.1.0 knitr_1.27       foreign_0.8-71   dplyr_0.8.3
## [5] car_3.0-6        carData_3.0-3
##
## loaded via a namespace (and not attached):
##  [1] zip_2.0.4         Rcpp_1.0.3        pillar_1.4.3      compiler_3.6.0
##  [5] cellranger_1.1.0  highr_0.8         forcats_0.4.0     tools_3.6.0
##  [9] digest_0.6.23     viridisLite_0.3.0 lifecycle_0.1.0   evaluate_0.14
## [13] tibble_2.1.3      pkgconfig_2.0.3   rlang_0.4.4       openxlsx_4.1.4
## [17] rstudioapi_0.10   curl_4.3          yaml_2.2.0        haven_2.2.0
## [21] xfun_0.12         rio_0.5.16        xml2_1.2.2        httr_1.4.1
## [25] stringr_1.4.0     vctrs_0.2.2       hms_0.5.3         webshot_0.5.2
## [29] tidyselect_1.0.0  glue_1.3.1        data.table_1.12.8 R6_2.4.1
## [33] readxl_1.3.1      rmarkdown_2.1     readr_1.3.1       purrr_0.3.3
## [37] magrittr_1.5      scales_1.1.0      htmltools_0.4.0   rvest_0.3.5
## [41] abind_1.4-5       assertthat_0.2.1  colorspace_1.4-1  stringi_1.4.5
## [45] munsell_0.5.0     crayon_1.3.4
```