Validating the impact of Class Size on Test Performance

Greg DePaul

17 February 2023

Abstract

We explore the connection between class room setting versus math test score performance.

Introduction and Background

The Tennessee Student Achievement Ratio project was a four year study, collecting a broad range of statistics for 7000 students in 79 elementary schools. These statistics caputure a variety of information, ranging from the instructional background of these teachers, the test score performance of each student, as well as information about the classroom setting for each individual student. Several conclusions have since been released after the mass distribution of this dataset. One conclusion, in particular, argues that

"Nashville-Davidson County students who attended small classes (K-3) consistently made better grades than students in regular and regular/aide classes by the end of the 1994-1995 school year. In English, math, and science, the students in the small classes outscored their counterparts by over 10 points." Harvard dataverse

The goal of this project is to verify these results using some of the statistical tools taught in UC Davis's Stats 207 course. Specifically:

- Q1: Does class size have an impact on score performance?

But a follow up to this question is naturally, can we disregard the school to make this assertion? So we also formulate the question:

- Q2: Does school have an impact on score performance?

Note, by the design of this experiment, we make the assumption that these two are sample in such a way that an interaction between these would be spurious and therefore discount such considerations. So we will stay within the additive domain for this analysis.

Exploratory Data Analysis

In this section, we explore the feature variables as well as some of their multivariate relationships. The variables of interest are as follows:

- g1tmathss: total math scaled score in 1st grade.
- class_types: factor indicating the class type in 1st grade: small (1), regular (2), or regular-with-aide (3).
- g1schid: factor indicating school id
- gktchid: factor indicating teacher id

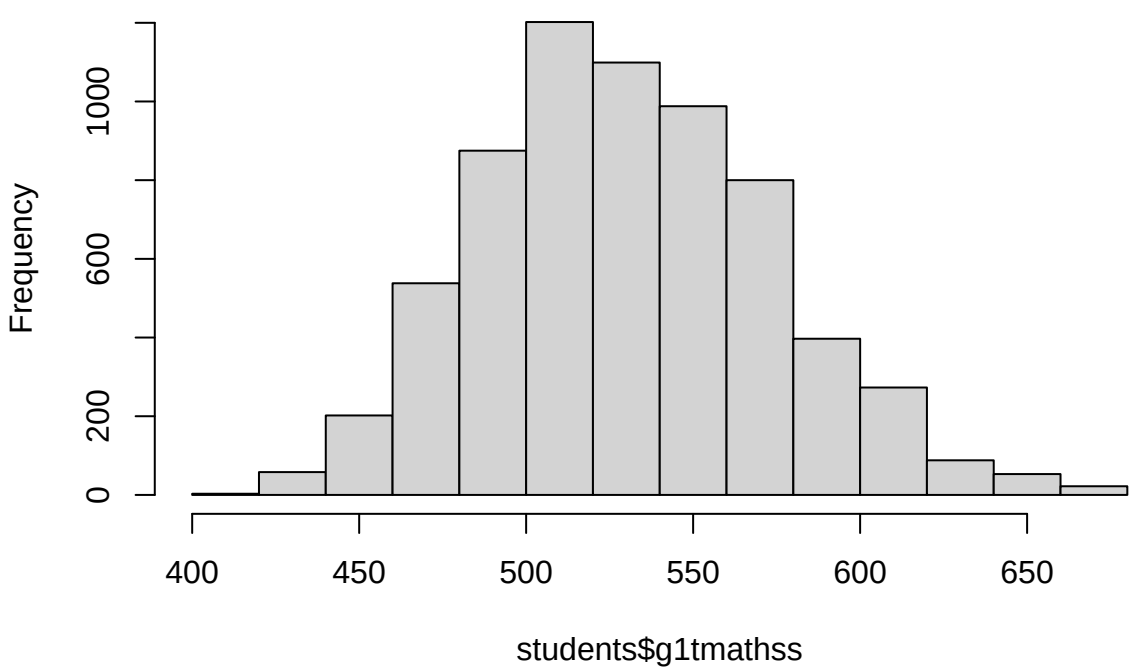
Summary Statistics on Feature Variables

We take a look at the summary statistics for the variables of interest:

	students (N = 11,601)
Math Scores	
min	404
25% quantiles	500
median	529
mean	530.527887238557
75% quantiles	557
max	676
# NA	5003
Class Type	
Small	1,925 (28)
Regular	2,584 (38)
Regular with Aide	2,320 (34)
# NA	4772
School Info	
Num Schools	77
Fewest Samples per School	47
Highest Samples per School	238
# NA	4772
Teacher Info	
Num Teachers	340
Fewest Samples per Teacher	12
Highest Samples per Teacher	30
# NA	4772

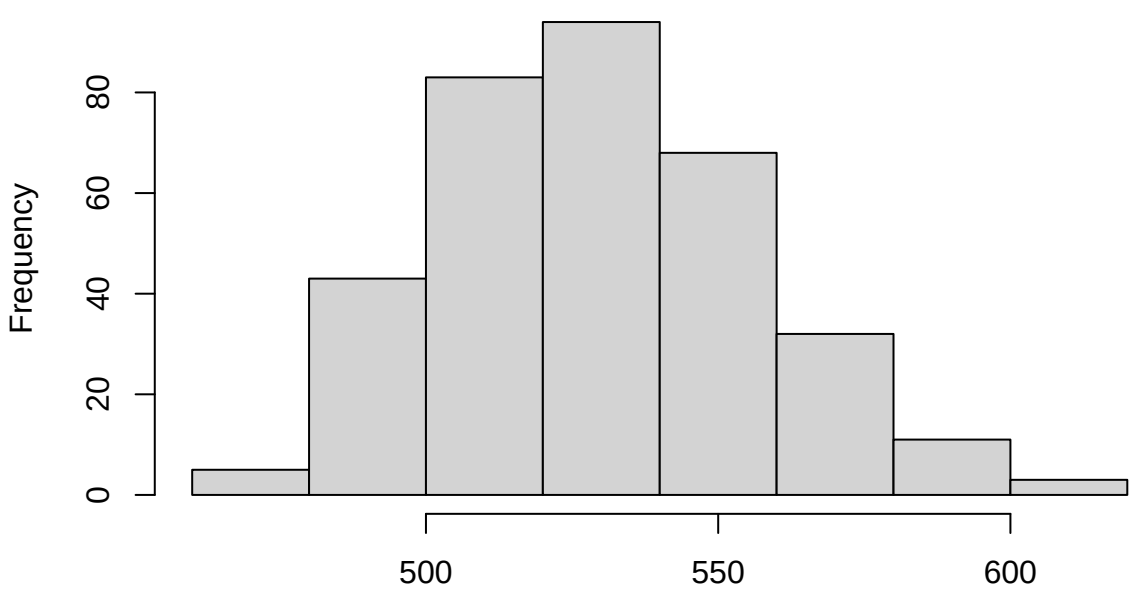
Aggregating per School

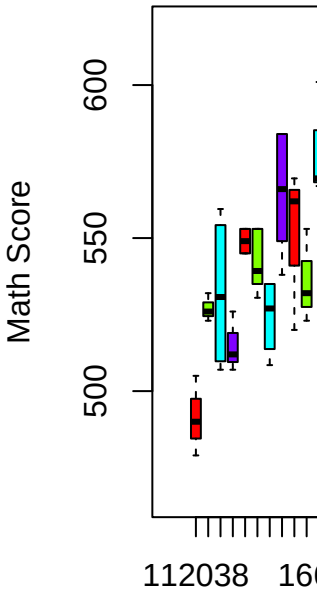
We can see that if we don't aggregate at the teacher level, and simply consider all students simultaneously, these are the distribution curves that we face. Recall that some of these teachers will only have 12 students to be able to draw meaningful conclusions on. This could be an issue depending on the summary statistic. For example, if we were to use the mean to evaluate a teacher's performance, this can be highly susceptible to outliers because the mean is a nonrobust statistic.

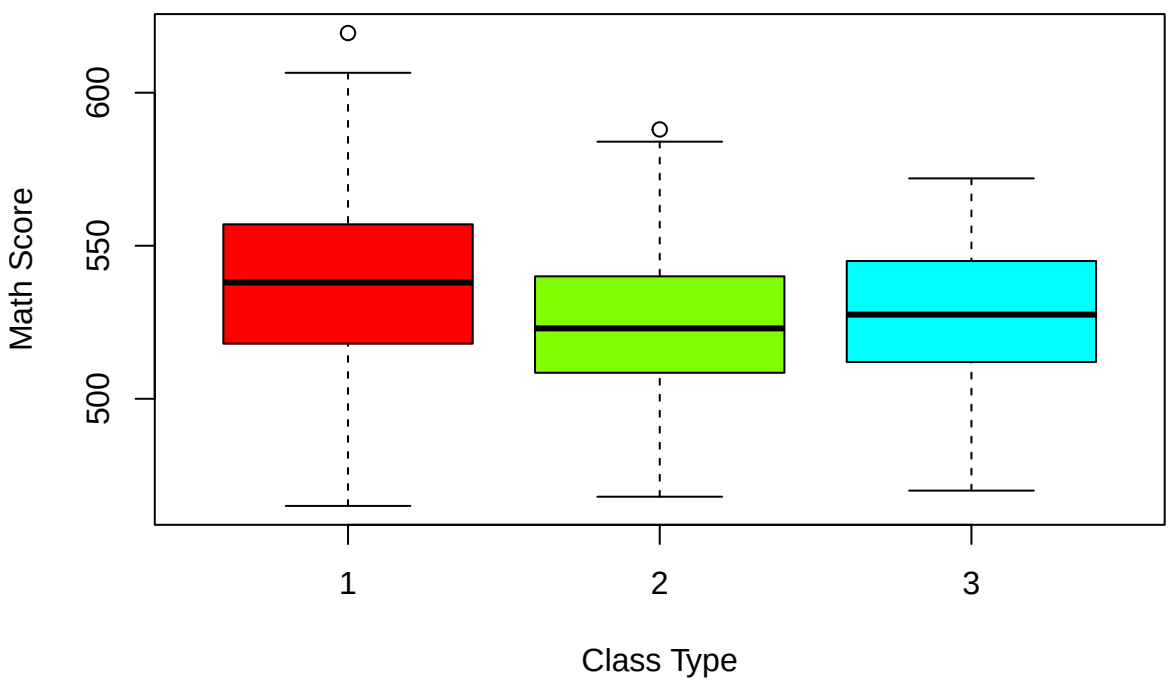


So instead we turn to aggregate our data. As a sanity check, we see that after aggregating over teachers and taking the mean score of their students, we don't see too great a departure from the histogram above of all students versus the histogram over median scores. This suggests that it is perfectly reasonable to continue this analysis using these aggregated statistics.

Now, looking at the relationships between math score versus school and math score, versus class type, we see that there are interesting relationships to explore.





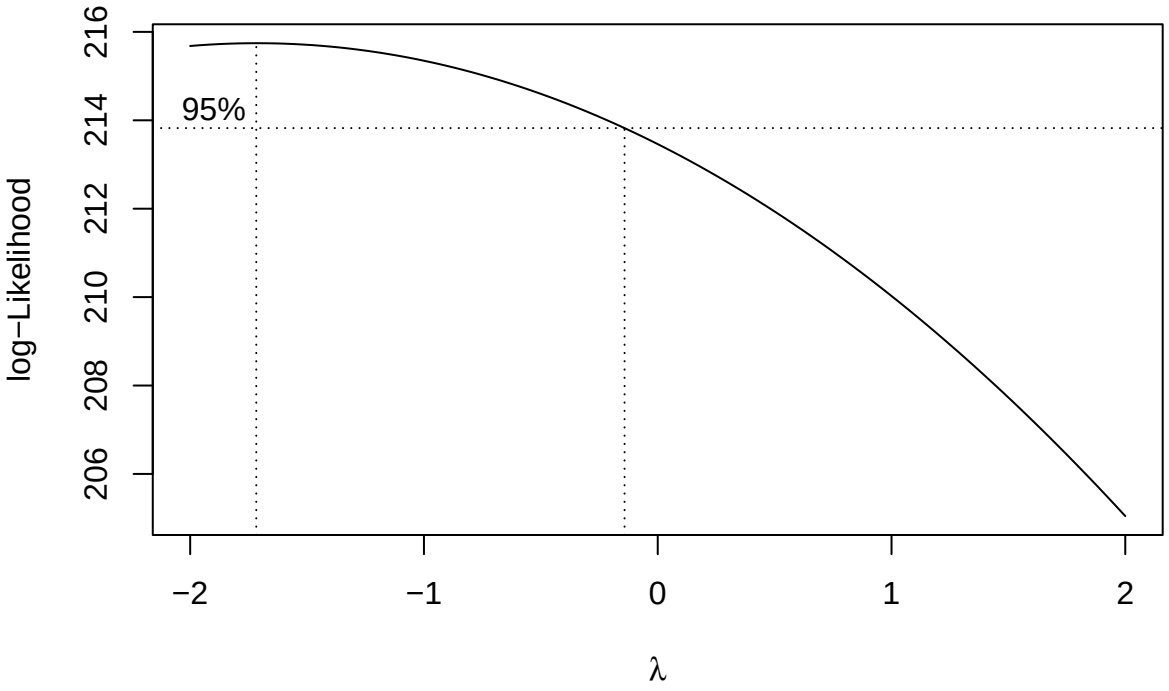


Firstly, we see that the scores vary wildly by school. This suggests that schools themselves play a significant part in determining student performance.

We also see that students in smaller classrooms tend to do much better than those in larger classrooms, regardless of the presence of an aide. This agrees with the report included with the Harvard dataset.

Checking for Nonlinearity in Score Target Variable

We see that upon applying the Box-Cox procedure, we recognize a serious nonlinearity with our data. Specifically, the value of 1 deviates heavily from the optimal 95% confidence interval obtain from the procedure. Observe, the "optimal" value of λ is printed below:




```
## [1] -1.717172
```

Recall, the box cox procedure tells us an appropriate transformation for the response variable Y. Specifically,

$$Y_{\text{transformed}} = \begin{cases} \frac{1}{\lambda}(Y^{\lambda} - 1) & \lambda \neq 0 \\ \log(Y) & \text{otherwise} \end{cases}$$

Notice, we simply choose $\lambda = 0$ since that is reasonably close to the 95% confidence interval within the box-cox plot.

Inferential analysis

We want to model this multivariate relationship between first grade math scores with the particular school and class room type in which each student was instructed. Specifically, we wish to employ the model

$$Y_{ij,\text{transformed}} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

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. For this model, we make the following assumptions:

- Each cell (i, j) is normally distributed with the same variance σ^2, and we use the error term $\epsilon_{i,j} \sim N(0, \sigma^2)$
- The natural constraints on α_i and β_j are

$$\sum_{i=1}^{3} n_i \alpha_i = 0 \quad \text{and} \quad \sum_{j=1}^{77} n_j \beta_j = 0$$

- Assuming no interaction term implies for all i, j,

$$(\alpha\beta)_{i,j} = 0$$

Upon fitting the model, we get the following ANOVA table. Specifically, we notice the F-statistic on the additive terms are clearly signficant, with p-values on the order of 10^{-30}.

```
##              Df Sum Sq  Mean Sq F value  Pr(>F)
## g1classtype   2 0.0381 0.019062  17.528 7.2e-08 ***
## g1schid      75 0.5341 0.007122   6.549 < 2e-16 ***
## Residuals   261 0.2838 0.001087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This suggests that schools will have an impact on the mean score, not just the class room type. We can see certain schools have a greater impact than others, indicated by the magnitude of their coefficients:

```
##   (Intercept)  g1classtype2  g1classtype3 g1schid123056 g1schid128076
##  6.2126336359 -0.0229053253 -0.0243311648  0.0702872495  0.0749139075
## g1schid128079 g1schid130085 g1schid159171 g1schid161176 g1schid161183
##  0.0417908532  0.1133301758  0.0976283876  0.0672339259  0.1384859496
## g1schid162184 g1schid164198 g1schid165199 g1schid166203 g1schid168211
##  0.1145733101  0.0869728457  0.1643495345  0.0371138656  0.0907881709
## g1schid168214 g1schid169219 g1schid169229 g1schid169231 g1schid169280
##  0.1472042421  0.1153546048  0.0837869326  0.0592249181  0.0864708711
## g1schid170295 g1schid173312 g1schid176329 g1schid180344 g1schid189378
##  0.1423812752  0.1140126153  0.0895098050  0.0862168298  0.0788329370
## g1schid189382 g1schid189396 g1schid191411 g1schid193422 g1schid193423
##  0.0903013461  0.0488911313  0.0109693897  0.0718044615  0.0650952091
## g1schid201449 g1schid203452 g1schid203457 g1schid205488 g1schid205490
##  0.1168209684  0.0919608055  0.0953281768  0.0608972990  0.0889977954
## g1schid205491 g1schid205492 g1schid208501 g1schid208503 g1schid209510
##  0.1139466279  0.0595795905  0.0784510486  0.0734869781  0.0792298810
## g1schid212522 g1schid215533 g1schid216537 g1schid218562 g1schid221571
##  0.0477437279  0.1123213859  0.1184912829  0.1025626179  0.0250996669
## g1schid221574 g1schid225585 g1schid228606 g1schid230612 g1schid231616
##  0.0529369245  0.0519378483  0.1536047625  0.1144420107  0.1001621884
## g1schid234628 g1schid244697 g1schid244708 g1schid244723 g1schid244727
##  0.1309162269  0.0400083602  0.0011516563  0.0374463013  0.0788270298
## g1schid244728 g1schid244736 g1schid244745 g1schid244746 g1schid244755
## -0.0297625378  0.0428593451  0.0206377034  0.0955827625  0.0233996925
## g1schid244764 g1schid244774 g1schid244776 g1schid244780 g1schid244796
##  0.0659065471  0.0033831053  0.0136624717  0.0005706956  0.0222994915
## g1schid244799 g1schid244801 g1schid244806 g1schid244831 g1schid244839
##  0.0882862426  0.0182265933  0.0519312604  0.0267839461  0.0912541441
## g1schid252885 g1schid253888 g1schid257899 g1schid257905 g1schid259915
##  0.0947634662  0.0915756302  0.0698309541  0.1195866226  0.0649879577
## g1schid261927 g1schid262937 g1schid264945
##  0.0603995950  0.1333724560  0.1036924343
```

Suppose that class size has no effect on wage. Then it would follow that the effect terms would be zero. Therefore, we formulate the null and alternative hypotheses to be:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$

$$H_A : \text{ not all } \alpha_i \text{ are zero}$$

the ANOVA table tells us the F-statistic in this case is:

$$F := 17.528$$

which has a p-value of $1.4 \cdot 10^{-30}$, so we can assume that these coefficients are indeed significant.

Similarly, suppose that class size has no effect on wage. Then it would follow that the effect terms would be zero. Therefore, we formulate the null and alternative hypotheses to be:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{76} = 0$$

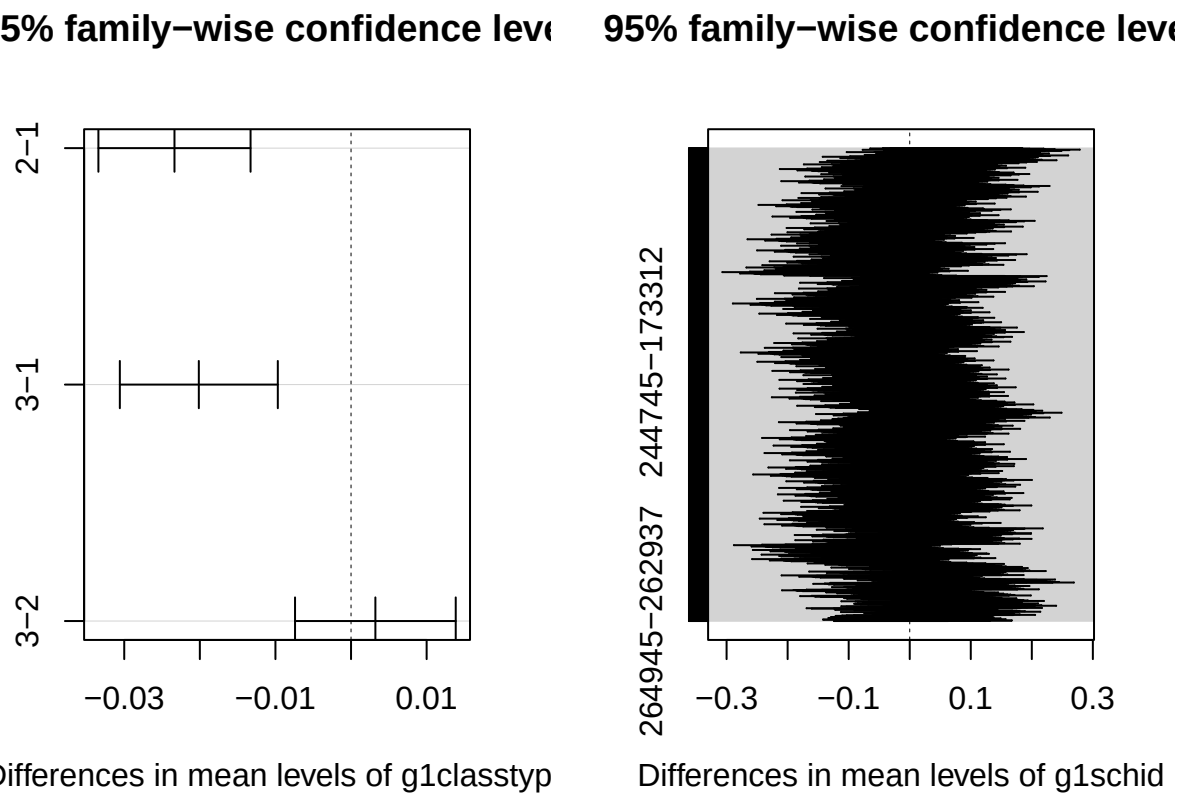
$$H_A : \text{ not all } \beta_j \text{ are zero}$$

the ANOVA table tells us the F-statistic in this case is:

$$F := 6.549$$

which has a p-value of 0, so we can assume that these coefficients are indeed significant as well.

Now, we can further press on by performing simultaneous inferencing to test whether schools themselves have a significant impact on the math score, beyond the control of the classroom setting. Using Tukey's Range Test, which compares pairwise over all 77 schools, we see that there are some schools for which their p-values are significant. (There are too many combinations to enumerate here)

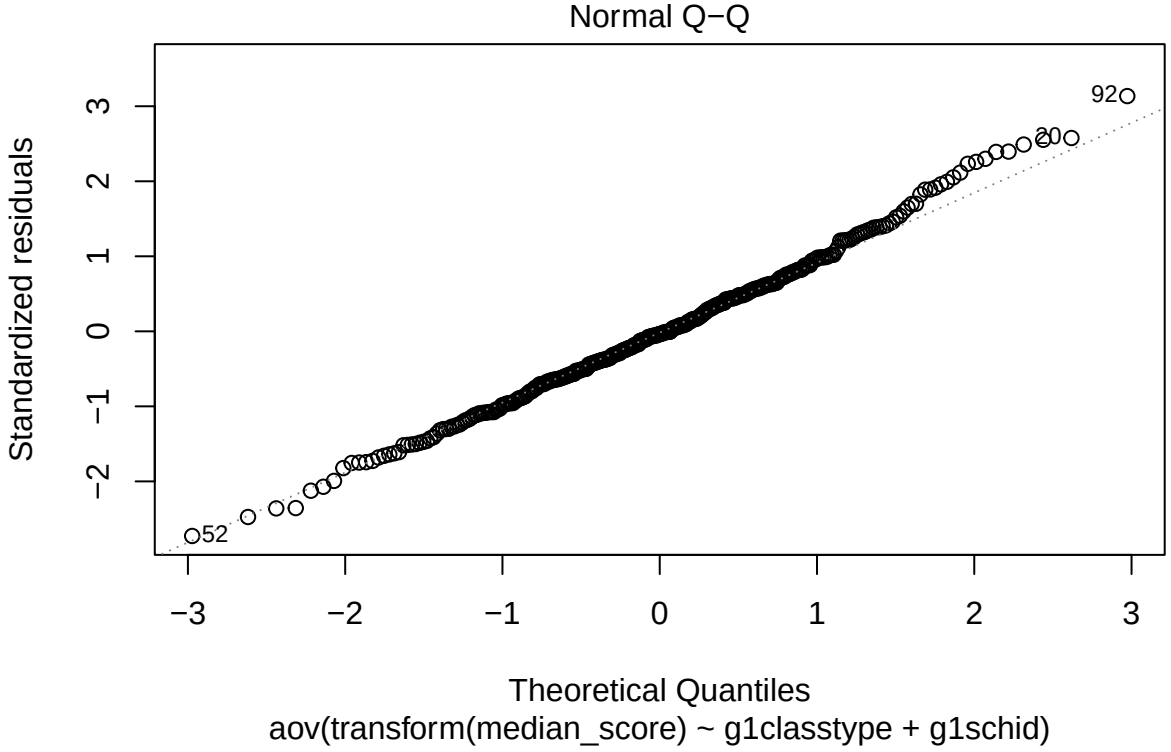


This indicates that we cannot assume that schools don't play a significant impact on a students performance.

Sensitivity analysis

Testing for Normality

We examine the residual plot of the fitted model. Observe the Q-Q plot appears to be incredibly normal.



We consult both the Anderson-Darling and Shapiro-Wilk normality tests. Observe, we see that we have no scepticism about any departure from normality:

```
##
##  Anderson-Darling normality test
##
## data:  aov_residuals
## A = 0.4114, p-value = 0.3392
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.99487, p-value = 0.3211
```

Avoiding Use of Mean Statistic

An interesting question is "Why did we choose to use the median?" We recall that the mean is a non robust statistic against outliers. Observe, we can use the Anderson-Darling normality test to see that we should reject the normality assumption while using the mean, as seen below:

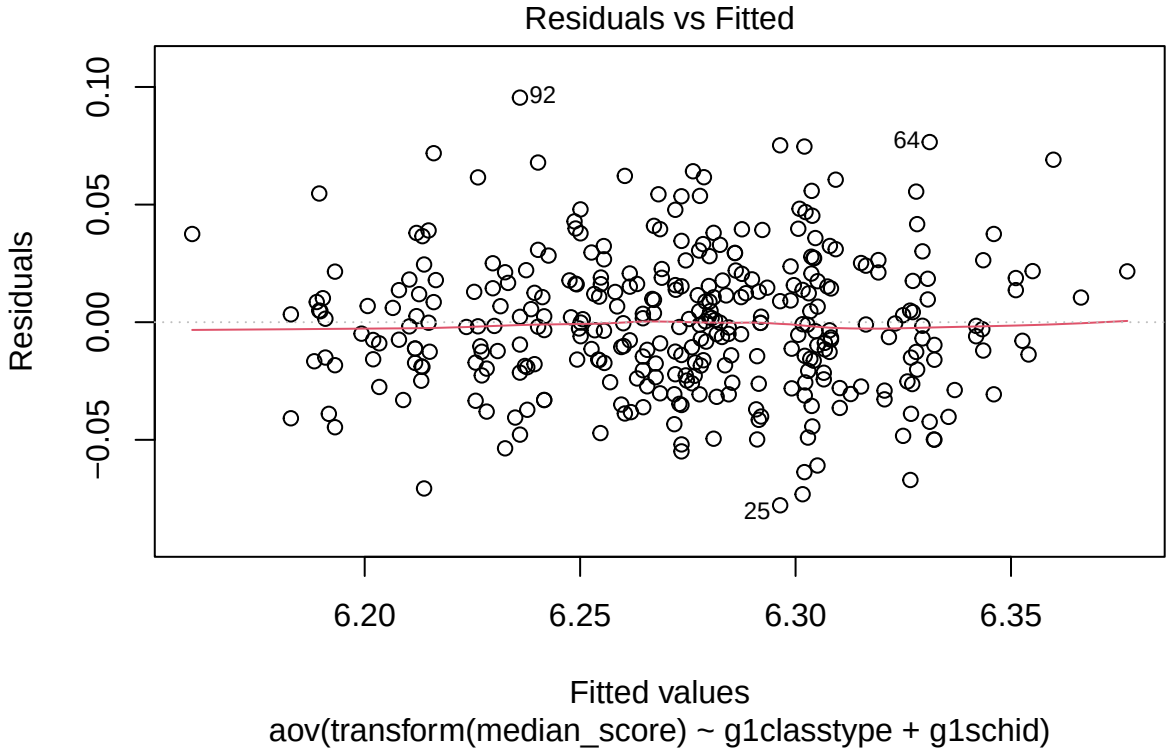
```
##
##  Anderson-Darling normality test
##
## data:  aov_residuals
## A = 1.4848, p-value = 0.0007765
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.98016, p-value = 0.0001262
```

We see that with a critical value of $\alpha = 10^{-3}$, the Anderson-Darling and Shapiro-Wilks tests both suggest to reject the Normality Assumption. This is why I have chosen to model using the median statistic.

Testing for Equal Variances across Cells

Looking at the Residuals verus fitted plot, we don't see any terrific deviations in our fit line:




```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group 223  1.8755 0.0001077 ***
##       115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  median_score by interaction(g1classtype, g1schid)
## Fligner-Killeen:med chi-squared = 286.9, df = 223, p-value = 0.002493
```

We see that the p-value for Levene's tests is less than the significance level of 0.001. However, its not the case for the Fligner-Killeen test. This inspires some doubt as to the treatment of data collection for these different groups of teachers. It could be some teachers may be more forthcoming with score information than others. However, we know these test tend to be conservative, and looking the our residuals fit line appears to be mostly straight. So we can assume our data is trustworthy and we're not experiencing too much unequal variances.

Discussion

After using analysis of variation to identify whether there is a tangible connection between classroom size and test performance, we come to the conclusion that such a relationship is significant. Further, we can confidently state as a result of Kuskal's test that the mean of small classroom settings is significantly higher than those of larger settings, regardless of the presence of an aide. However, we also see that this trend is not independent of the specific school of attendance. So classroom setting will not guarantee performance unfortunately. But there are a variety of reasons for this. Certain schools may attract better trained professionals than others, or have the resources to provide smaller classrooms settings. This inability to have smaller classrooms would explain why there may lack equal variance across schools.

For future studies, it may be useful to consider, as opposed to singular testing events, the series of tests students take over their academic careers. It might also be usefult to use more of the available features to study test performance.

Acknowledgement

The instructors and TAs have been very helpful, making themselves available and approachable to ask questions. I'm very thankful at how attentive they are to student needs.

Reference

Achilles, C.M. et al. (2008) Tennessee's student teacher achievement ratio (STAR) project, Harvard Dataverse. Harvard Dataverse. Available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl%3A1902.1%2F10766 (Accessed: February 16, 2023).

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

Session info

Report information of your `R` session for reproducibility.

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Ventura 13.2
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] DescTools_0.99.47 MASS_7.3-58.1     nortest_1.0-4     qwraps2_0.5.2
##  [5] gtsummary_1.7.0   forcats_1.0.0     stringr_1.5.0     purrr_1.0.1
##  [9] readr_2.1.4       tidyr_1.3.0       tibble_3.1.8      ggplot2_3.4.1
```

```
## [13] tidyverse_1.3.2   dplyr_1.1.0       haven_2.5.1       AER_1.2-10
## [17] survival_3.4-0    sandwich_3.0-2    lmtest_0.9-40     zoo_1.8-11
## [21] car_3.1-1         carData_3.0-5
##
## loaded via a namespace (and not attached):
##  [1] fs_1.6.0              lubridate_1.9.2      httr_1.4.4
##  [4] tools_4.2.2           backports_1.4.1      utf8_1.2.3
##  [7] R6_2.5.1              DBI_1.1.3            colorspace_2.1-0
## [10] withr_2.5.0           tidyselect_1.2.0     Exact_3.2
## [13] compiler_4.2.2        cli_3.6.0            rvest_1.0.3
## [16] gt_0.8.0              expm_0.999-7         xml2_1.3.3
## [19] scales_1.2.1          mvtnorm_1.1-3        proxy_0.4-27
## [22] digest_0.6.31         rmarkdown_2.20       pkgconfig_2.0.3
## [25] htmltools_0.5.4       highr_0.10           dbplyr_2.3.0
## [28] fastmap_1.1.0         rlang_1.0.6          readxl_1.4.1
## [31] rstudioapi_0.14       generics_0.1.3       jsonlite_1.8.4
## [34] googlesheets4_1.0.1   magrittr_2.0.3       Formula_1.2-4
## [37] Matrix_1.5-1          Rcpp_1.0.10          munsell_0.5.0
## [40] fansi_1.0.4           abind_1.4-5          lifecycle_1.0.3
## [43] stringi_1.7.12        yaml_2.3.7           rootSolve_1.8.2.3
## [46] grid_4.2.2            crayon_1.5.2         lmom_2.9
## [49] lattice_0.20-45       splines_4.2.2        hms_1.1.2
## [52] knitr_1.42            pillar_1.8.1         boot_1.3-28
## [55] gld_2.6.6             reprex_2.0.2         glue_1.6.2
## [58] evaluate_0.20         data.table_1.14.6    broom.helpers_1.12.0
## [61] modelr_0.1.10         vctrs_0.5.2          tzdb_0.3.0
## [64] cellranger_1.1.0      gtable_0.3.1         assertthat_0.2.1
## [67] xfun_0.37             broom_1.0.3          e1071_1.7-13
## [70] class_7.3-20          googledrive_2.0.0    gargle_1.3.0
## [73] timechange_0.2.0      ellipsis_0.3.2
```

Appendix

```
library('AER')
library(haven)
library(dplyr)
library(tidyverse)
library(gtsummary)
library(ggplot2)
library(qwraps2)
library(nortest)
options(qwraps2_markup = "markdown")
library(car)
library(MASS)
library(DescTools)
knitr::opts_chunk$set(fig.pos = 'H')
students <- read_sav("STAR_Students.sav")

our_summary1 <-
  list("Math Scores" =
       list("min"       = ~ min(g1tmathss, na.rm = TRUE),
            "25% quantiles" = ~ quantile(g1tmathss, na.rm = TRUE)[2],
            "median" = ~ quantile(g1tmathss, na.rm = TRUE)[3],
```

```
            "mean" = ~ mean(g1tmathss, na.rm = TRUE),
            "75% quantiles" = ~ quantile(g1tmathss, na.rm = TRUE)[4],
            "max"       = ~ max(g1tmathss, na.rm = TRUE),
            "# NA" = ~ sum(is.na(g1tmathss))
      ),
      "Class Type" =
      list("Small"       = ~ qwraps2::n_perc0(g1classtype == 1, na_rm = TRUE),
            "Regular"    = ~ qwraps2::n_perc0(g1classtype == 2, na_rm = TRUE),
            "Regular with Aide"       = ~ qwraps2::n_perc0(g1classtype == 3, na_rm = TRUE),
            "# NA" = ~ sum(is.na(g1classtype))
      ),
      "School Info" =
      list("Num Schools" = ~ length(unique(g1schid)),
            "Fewest Samples per School" = ~ min(table(g1schid)),
            "Highest Samples per School" = ~ max(table(g1schid)),
            "# NA" = ~ sum(is.na(g1schid))
      ),
      "Teacher Info" =
      list("Num Teachers" = ~ length(unique(g1tchid)),
            "Fewest Samples per Teacher" = ~ min(table(g1tchid)),
            "Highest Samples per Teacher" = ~ max(table(g1tchid)),
            "# NA" = ~ sum(is.na(g1tchid))
      )
  )

summary_table(students, our_summary1)


hist(students$g1tmathss)

boxplot(students$g1tmathss ~ students$gktchid,main='Teacher versus Math Score',
xlab='School',ylab='Math Score',col=rainbow(4))

students_cleaned <- students%>%dplyr::select(g1tmathss, g1classtype, g1schid, g1tchid)%>%mutate(na.coun
students_cleaned <- students_cleaned%>%na.omit()%>%dplyr::select(-na.count) # drop rows with missing da

students_cleaned <- students_cleaned %>% group_by(g1classtype, g1schid, g1tchid) %>%
  summarise(median_score=median(g1tmathss),
            .groups = 'drop')

hist(students_cleaned$median_score)

students_cleaned$g1classtype=as.factor(students_cleaned$g1classtype)
students_cleaned$g1schid=as.factor(students_cleaned$g1schid)

boxplot(students_cleaned$median_score ~ students_cleaned$g1schid,main='School versus Math Score',
xlab='School',ylab='Math Score',col=rainbow(4))

boxplot(students_cleaned$median_score ~ students_cleaned$g1classtype,main='Class Type versus Math Score
xlab='Class Type',ylab='Math Score',col=rainbow(4))

X_1 <- students_cleaned$g1schid
X_2 <- students_cleaned$g1classtype
```

```r
Y <- students_cleaned$median_score
bc <- boxcox(Y ~ X_1 + X_2)
lambda <- bc$x[which.max(bc$y)]
print(lambda)

lambda <- 0

#Y_transform <- 1/lambda * (Y^lambda - 1)

transform <- function(z) {
  #return(1/lambda * (z^lambda - 1))
  return(log(z))
}

inverse_transform <- function(z) {
  #return((lambda*z / 1 + 1)^(1/lambda))
  return(exp(z))
}
model1 = aov(transform(median_score) ~ g1classtype + g1schid, data = students_cleaned)
summary(model1)
#Anova(model1, type = "III")
print(model1$coef)

my_tukey <- TukeyHSD(model1)
par(mfrow = c(1,2))
plot(my_tukey)
model2 = aov(median_score ~ g1classtype + g1schid, data = students_cleaned)
plot(model1, 2)
aov_residuals <- residuals(object = model1)
ad.test(aov_residuals)
shapiro.test(aov_residuals)
students_cleaned_2 <- students%>%dplyr::select(g1tmathss, g1classtype, g1schid, g1tchid)%>%mutate(na.co
students_cleaned_2 <- students_cleaned_2%>%na.omit()%>%dplyr::select(-na.count) # drop rows with missin

students_cleaned_2 <- students_cleaned_2 %>% group_by(g1classtype, g1schid, g1tchid) %>%
  summarise(mean_score=mean(g1tmathss),
            .groups = 'drop')

students_cleaned_2$g1classtype=as.factor(students_cleaned_2$g1classtype)
students_cleaned_2$g1schid=as.factor(students_cleaned_2$g1schid)

model2 = aov(mean_score ~ g1classtype + g1schid, data = students_cleaned_2)


aov_residuals <- residuals(object = model2)
ad.test(aov_residuals)
shapiro.test(aov_residuals)

plot(model1, 1)
print(leveneTest(median_score ~ g1classtype*g1schid, data=students_cleaned))

fligner.test(median_score ~ interaction(g1classtype, g1schid), data = students_cleaned)
sessionInfo()
```