S&DS 230 Final Project

Introduction

The motivation and background for this study is simple: I have a longstanding interest in eduation, and in locating educational data that I can analyze. The data for this project is unique in that it comes from a source with more information than is usually easy to access online. I am interested to get a basic feel for the data, sense of schools in CT, and some preliminary findings about what can be used to predict high school graduation. I would like to to develop this for college enrollment and persistence, but that is for another occasion.

Data

There are many more variables in the dataset than are highlighted in this report. The main ones used are listed here with the format: Name (unit of analysis, data type): {description}

- Enrollment (school, numeric): total number of students enrolled in a given year
- Chronic absenteeism rate (school, numeric): the percentage of students chronically absent in a given year
- Suspension rate (school, numeric): percentage of stuents suspended in a given year
- Normalized ISS/OSS counts (school, numeric): counts of in-school and out-of-school suspensions given divided by school enrollment
- Category (school, categorical): the type/sector of the school
- District ID (school, categorical): the district to which the school belongs
- Per-pupil spending (district, numeric): district-level spending per pupil, divided into several variables (ones used are spending on staff/instructional services and spending on support and administration)
- Course-taking percentages (school, numeric): percentages of students enrolled in particular subjects (this report considers math, ELA, science, language, history, and arts)
- Grade levels offered (school, categorical): a categorical variable which subsumes many different grade level combinations which a school may offer

Data cleaning

To my appreciation, EdSight has done a fair amount of work to arrange data into easily accessible, clearly structured .csv files. Most all of these files contain information on the school level, and therein each school is listed along with a numeric code, name, district code, and district name, facilitating cross-file analysis. However, the information of interest is spread out over a fair number of distinct data files. The largest challenge in getting the data was merging data from ~three dozen csv files into one large compendium dataset. The unit of analysis was schools, which are embedded within distinct districts. The basic motivation and challanges were as follows:

- It would be convenient to have a dataset where each school was allotted one row, with all relevant properties for that school defined in distinct columns; with this in mind,

- Schools are not consistently included in each file;
- There are plenty of missing values, which may be represented by more than one string;
- In some datasets, results are disaggregated by subject type, grade level, or another qualifier, leading to schools being listed in multiple rows;
- In such cases, the number of rows need not be consistent between different schools;
- Some schools in the datasets are identified as pertaining to multiple districts, while others belong only to one; some are not identified as belonging to any;
- Some datasets are at the district level, i.e. not disaggregated by school;
- Some schools appear to be listed more than once in some files where they should be listed only once;
- *And whatever other oddities and quirks emerged along the way…*

The full data-merging program is a few hundred lines long, and will not be shown here (rmd is available). The basic process underlying the code is such:

- Generate school-wise dataset for each data file:
    - If the data file is structured to list one row per school, remove any rows with (unexpected) duplicate school IDs;
    - Otherwise,
        - If the file is disaggregated by subject (e.g. Math, ELA) and/or grade level (e.g. Grade 4, Kindergarten, Grade 12), the data is rearranged: from the columns which vary by subject or grade (i.e., column names which do not give name/ID for schools/districts), a new series of columns is generated which allots one column to each grade, subject, or grade-subject combination that may occur (these columns are named according to whatever grade/subject/combination produced them). The standardized naming of the columns makes it easy to isolate particular grade levels, subjects, or years in the fully merged dataset;
        - If the file is disaggregated by some other type qualifier (e.g. a type of course offering, a type of disciplinary action, etc.), the same procedure is followed, where each category of the type disaggretator is used to generate a new column;
        - After disaggregation, there should be no duplicate school IDs–check to be sure.
    - Some datasets require merging or renaming of columns.
- Change all the column names which are unique to the data file (i.e., column names which do not give name/ID for schools/districts) by appending the name of the data file–this allows for easily extracting column names in the fully merged dataset.
- Set the rownames to the school IDs.
- Each generated dataset is merged with the previous one, by row, so that each school has its data appended to its unique row.
- In the end, some rows were generated which were fully or almost wholly empty; these were removed.
- The final merged dataframe was saved to a csv file so the merging program does not have to be run every time R is restarted.

The resultant data frame, by nature of construction, often has rows with many missing values. This is because schools have columns appended to them for which they lack data, or for which they are not meant to have data–for example, elementary schools lack data for any column pertaining to graduation and college attendance. While there may be some memory cost and redundancy involved with this, this is quite acceptable, given how easily accesible and manipulable the data become.

There are a large number of schools in the useable sample (~1000). There are more columns than schools; however, many of these columns will never be used simultaneously, so in reality the number of schools will always well outdo the number of variables used for analysis.

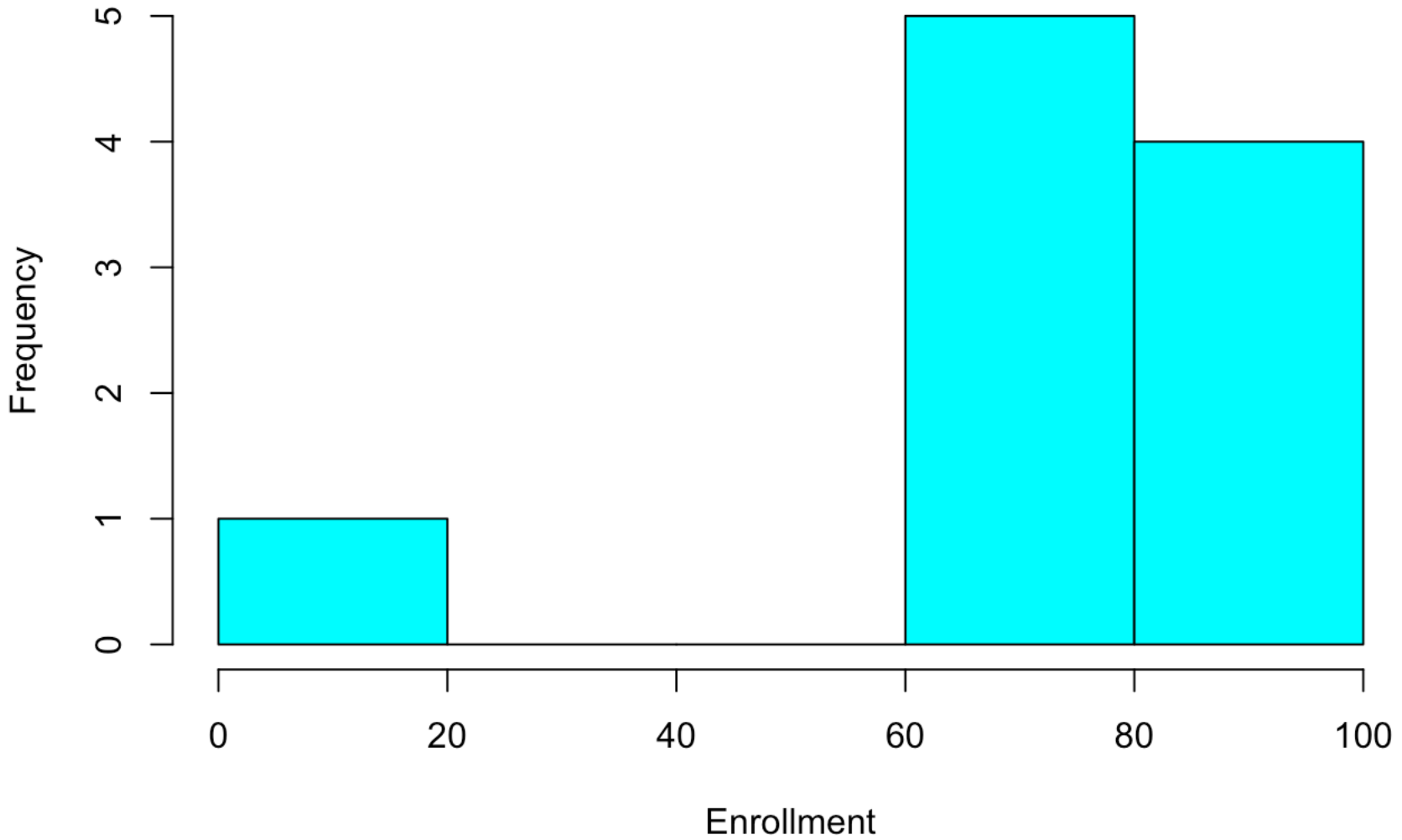
Preliminary: Enrollment & Attendance

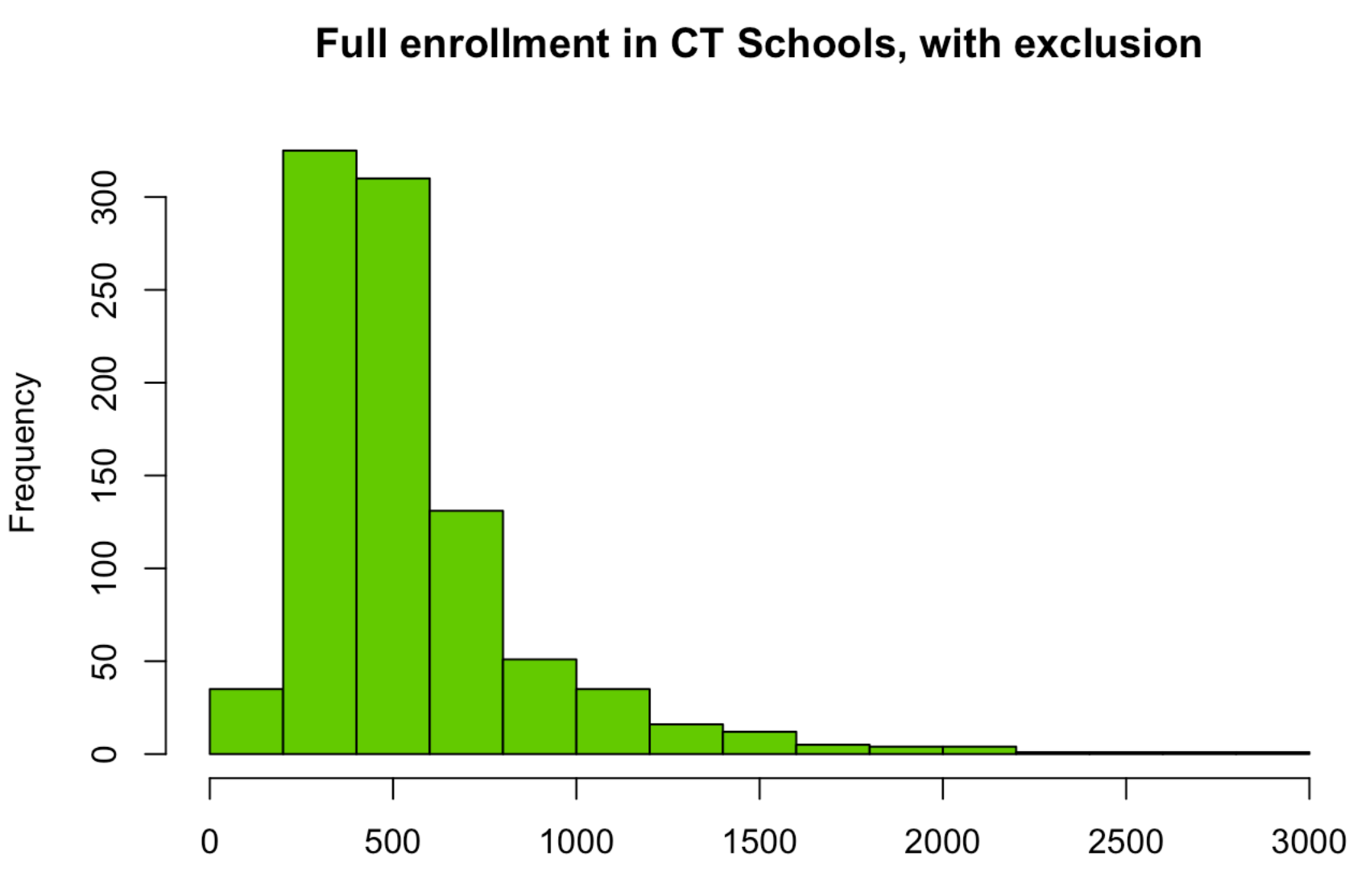
What does total per-school enrollment look like across the school population?

```
print(low_enroll_sum <- sum(full_data$enroll_2013_14 < 100))
```

```
## [1] 10
```

```
if(low_enroll_sum) {
    hist(full_data$enroll_2013_14[full_data$enroll_2013_14<100],col="cyan",main="Scho
ols with enrollment<100",xlab="Enrollment")
    sum(full_data$enroll_2013_14 < 100)/dim(full_data)[1]
    full_data <- full_data[full_data$enroll_2013_14 >= 100,]
    hist(full_data$enroll_2013_14,col="chartreuse3",main="Full enrollment in CT Schoo
ls, with exclusion",xlab="Enrollment")
} else {hist(full_data$enroll_2013_14,col="blue",main="Full enrollment in CT Schools"
,xlab="Enrollment")}
```


Schools with enrollment<100
Frequency
0
1
2
3
4
5
0
20
40
60
80
100
Enrollment


Full enrollment in CT Schools, with exclusion
Frequency
0
50
100
150
200
250
300
0
500
1000
1500
2000
2500
3000

In the raw datasets, there are schools with fewer than 100 students listed; by the time data were merged and rows with too many missing values are omitted, most of these disappear. The few that remain are removed, as it would be misleading to generate percentage statistics from such small schools (and anyway most of the data would be suppressed for privacy reasons).

There are many different grade-level combinations hosted by different schools:

```
unique(full_data$gradelevels)
```

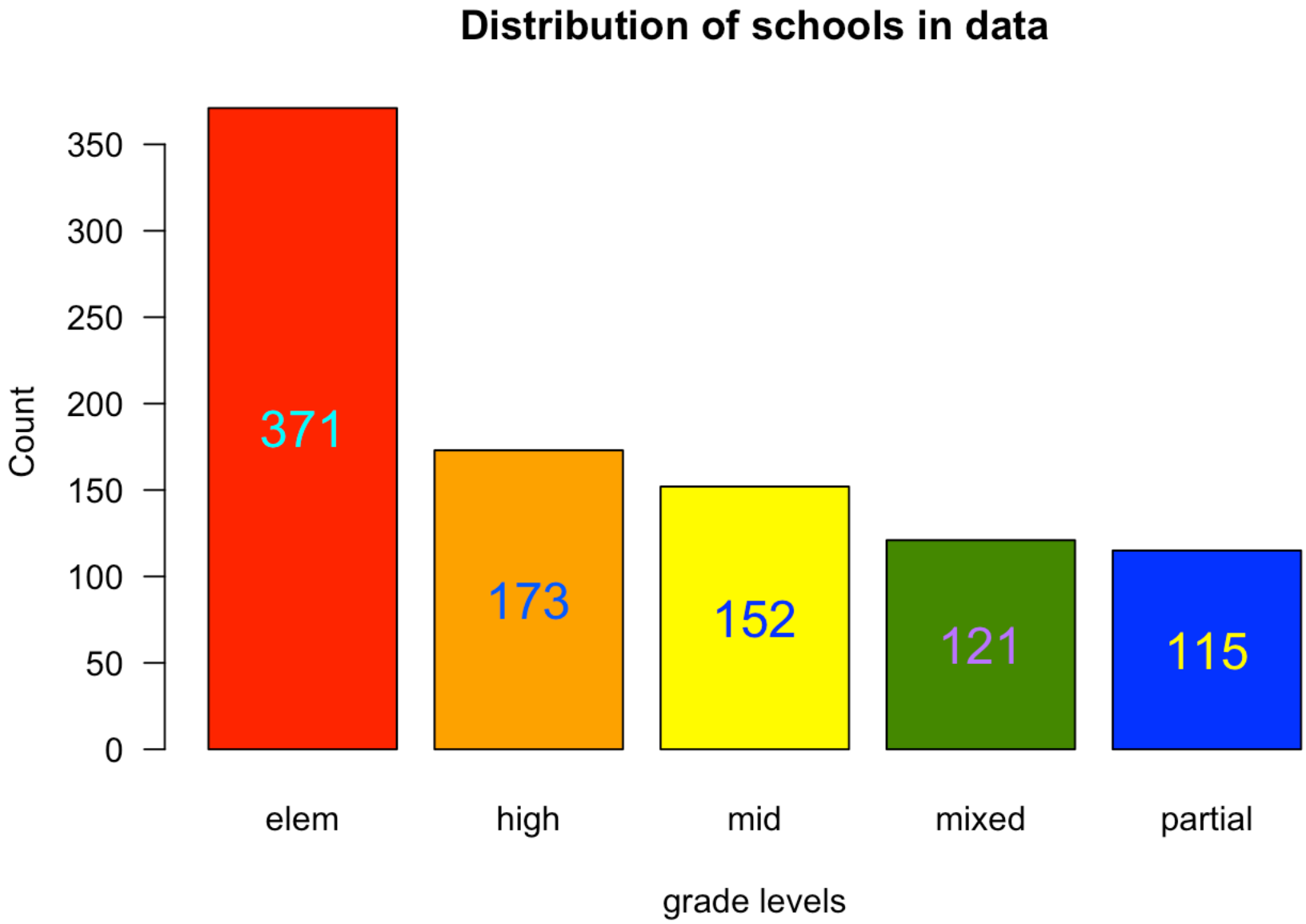
```
##  [1] "Pre K_6"     "K_6"         "Pre K_8"     "K_4"         "5_6"
##  [6] "7_8"         "9_12"        "K_5"         "Pre K_5"     "6_8"
## [11] "Pre K_3"     "K_3"         "4_5"         "3_4"         "K_2"
## [16] "Pre K_K"     "6_12"        "Pre K_4"     "5_8"         "K_8"
## [21] "Pre K_Pre K" "Pre K_1"     "2_4"         "4_6"         "1_6"
## [26] "3_5"         "Pre K_2"     "4_8"         "1_8"         "6_6"
## [31] "Pre K_12"    "K_12"        "3_6"         "1_5"         "2_5"
## [36] "7_12"        "8_12"
```

Let's bring this down to a workable size:

```
full_data$gradelevels <- gsub("Pre K_6|K_6|K_4|Pre K_5|Pre K_4|1_5|K_5|1_6","elem",fu
ll_data$gradelevels)
full_data$gradelevels <- gsub("7_8|6_8|5_8","mid",full_data$gradelevels)
full_data$gradelevels <- gsub("9_12","high",full_data$gradelevels)
full_data$gradelevels <- gsub("Pre K_12|K_12","mixed",full_data$gradelevels)
full_data$gradelevels <- gsub("Pre K_8|6_12|K_8|4_8|1_8|7_12|8_12","mixed",full_data$
gradelevels)
full_data$gradelevels[!grepl("elem|mid|high|full|mixed",full_data$gradelevels)] <- "p
artial"
unique(full_data$gradelevels)
```

```
## [1] "elem"    "mixed"   "partial" "mid"     "high"
```

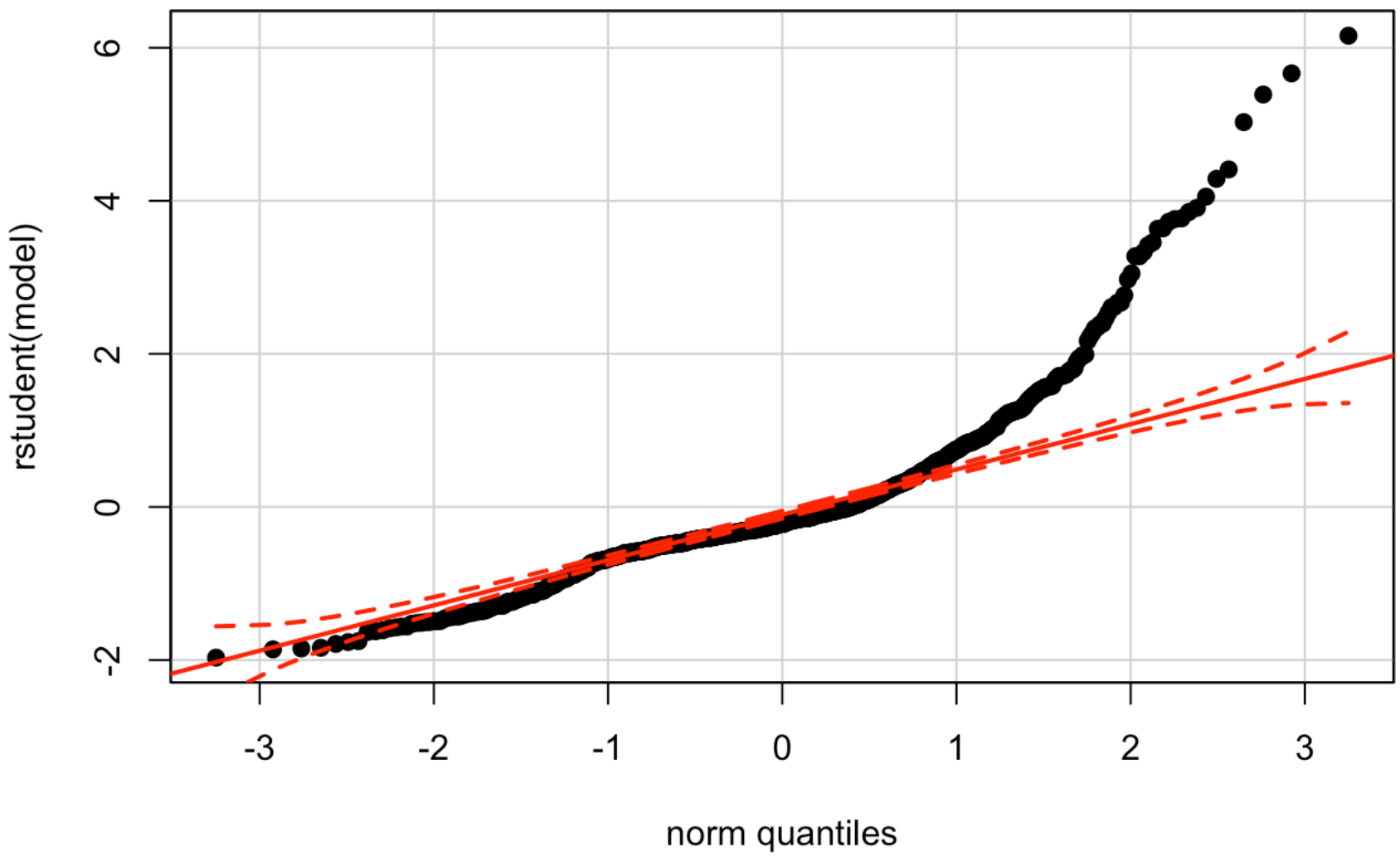
What are we left with?



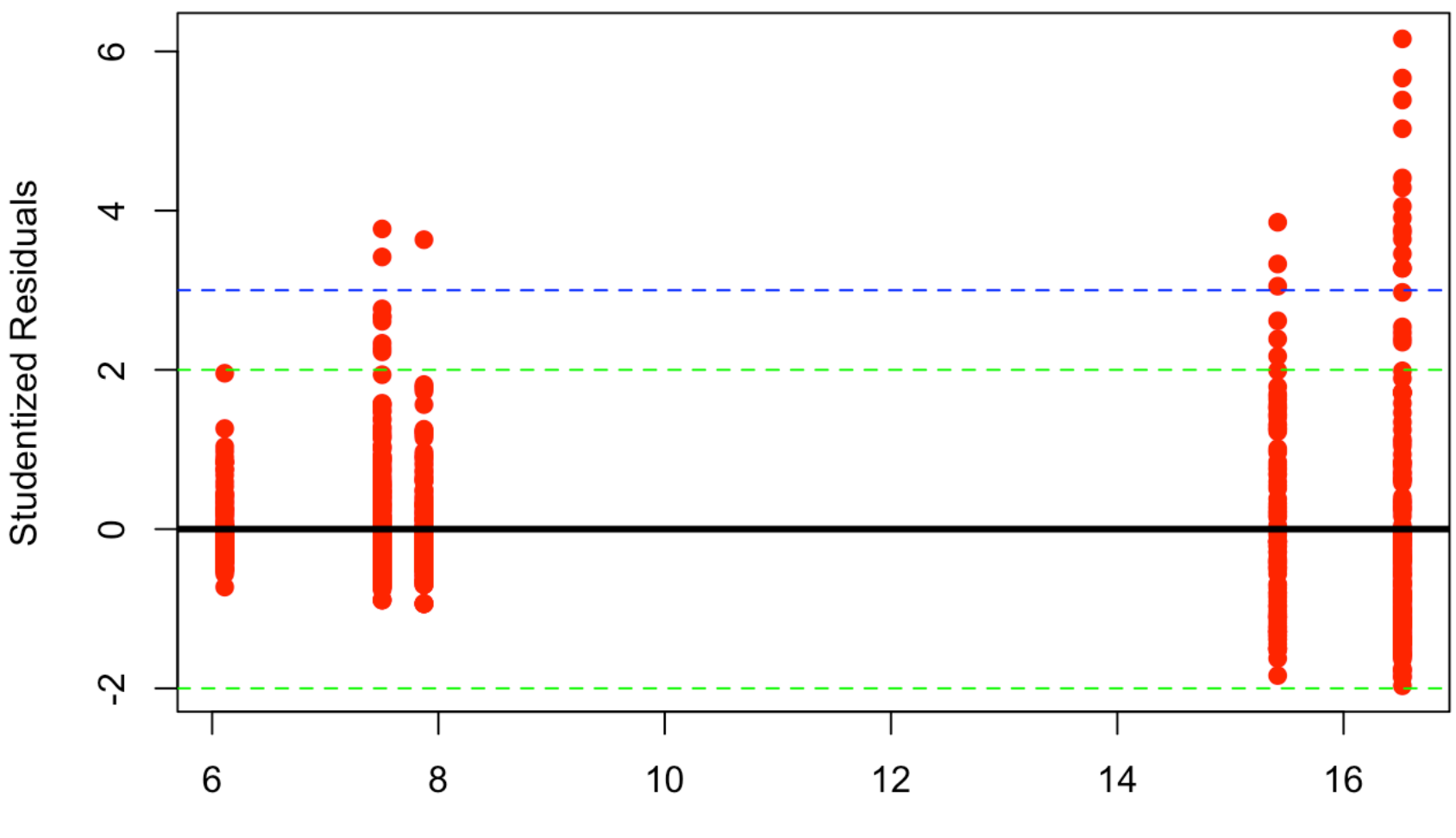
Absenteeism by school type: grade levels

Given enrollment, what about absenteeism? Absenteeism would preferably be broken down by grade so that we could clearly see changes from one year to the next, and so that mixed schools did not confound results; but since there are many elementary, middle, and high schools, I think that this measure is reliable enough to examine.

NQ Plot of Studentized Residuals, Residual Plots



Fits vs. Studentized Residuals, Residual Plots

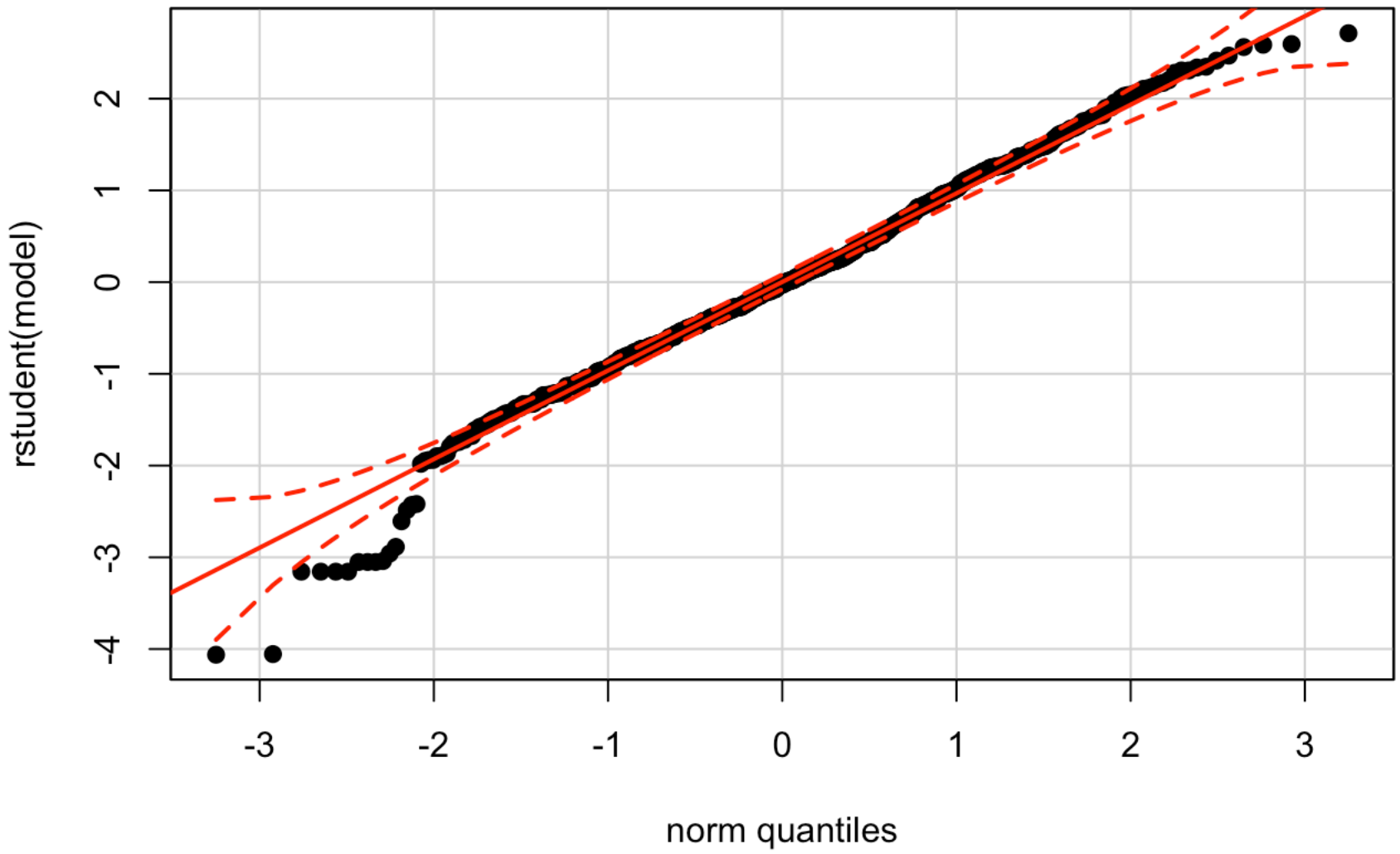


Fitted Values

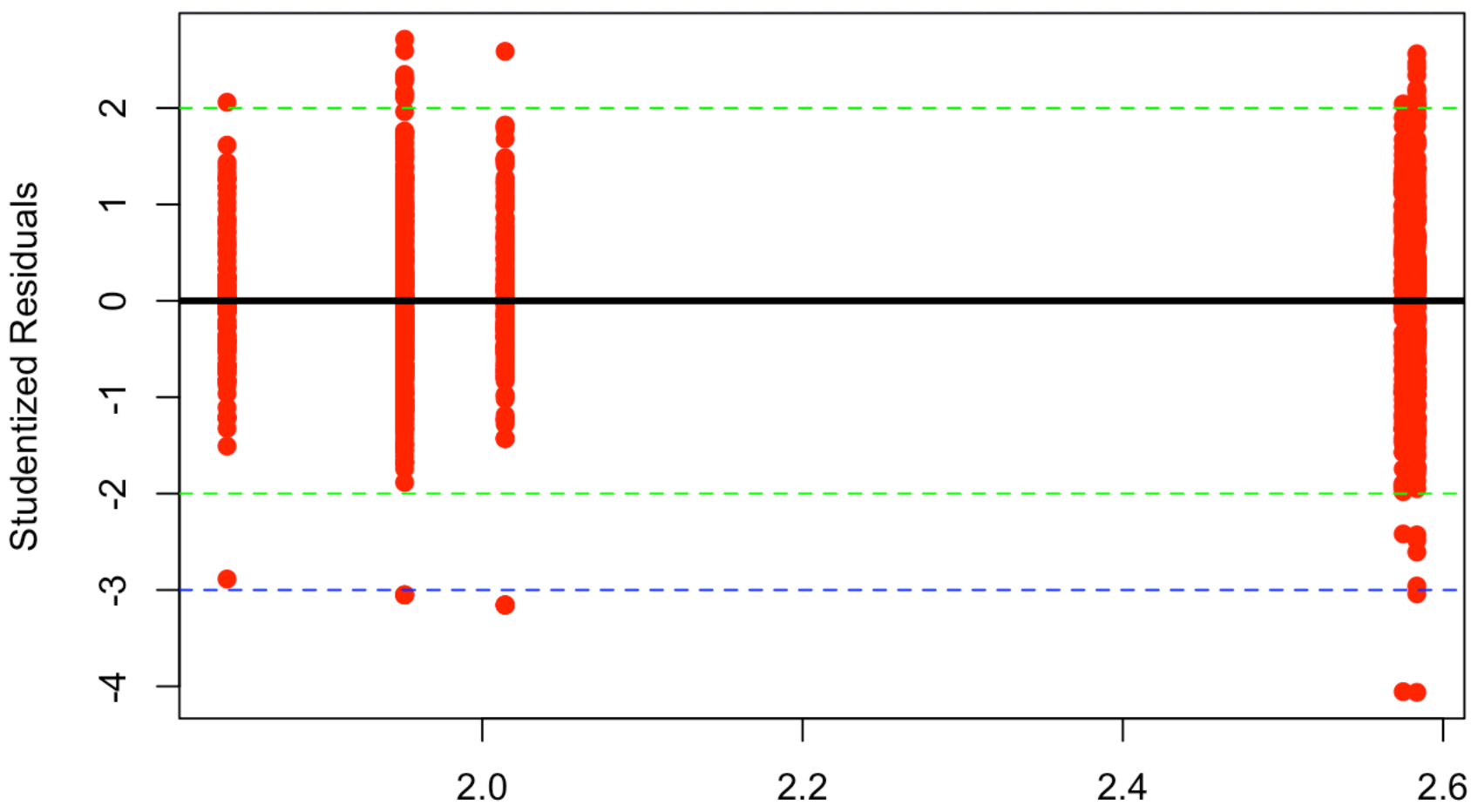
These residuals can’t pass as normally distributed. How about a log transformation?

```
y <- loga(y)
myResPlots2(lm(y~x))
```

NQ Plot of Studentized Residuals, Residual Plots



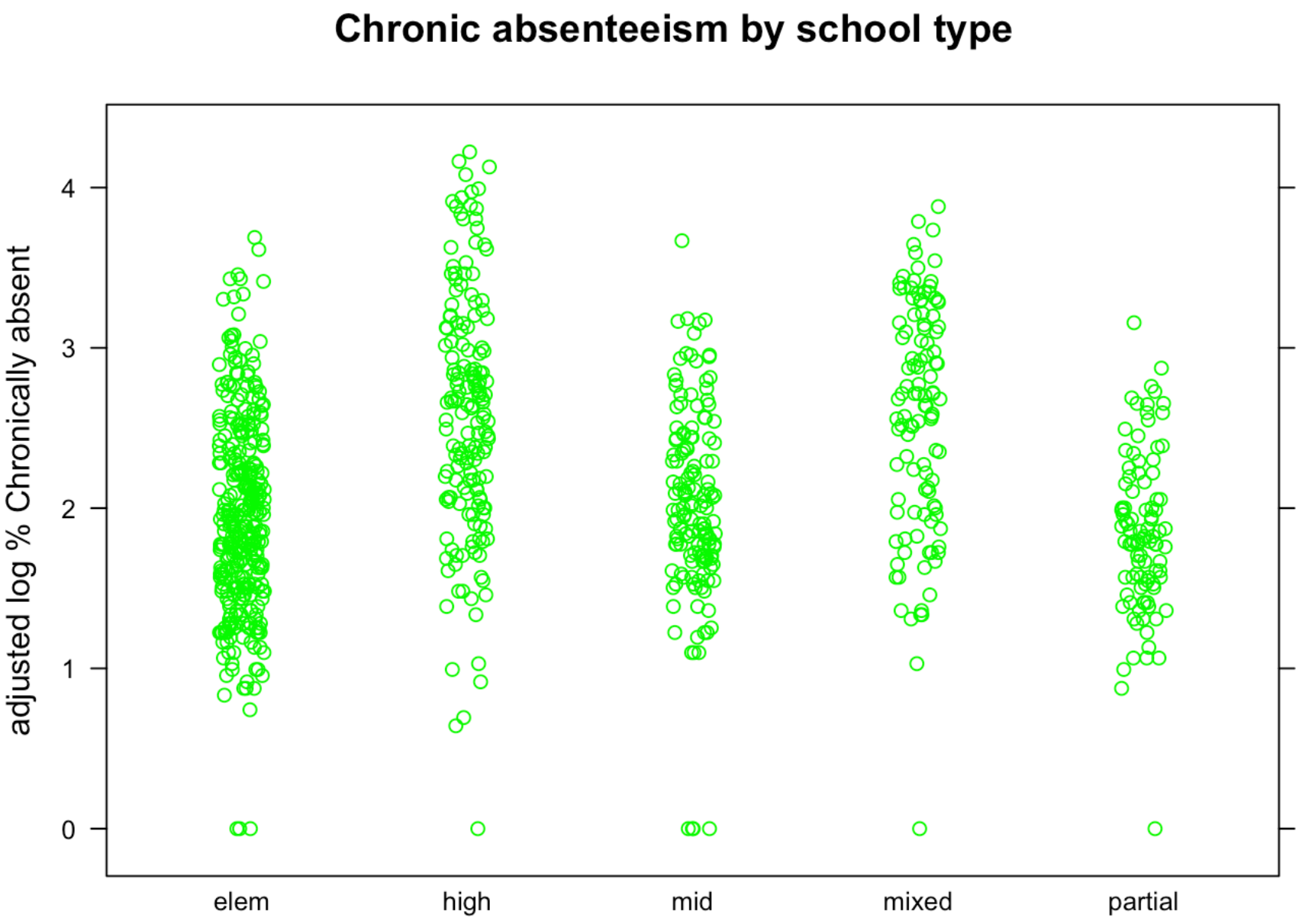
Fits vs. Studentized Residuals, Residual Plots



Fitted Values

Much better: residuals are mostly normally distributed, and there is no evidence of heteroskedasticity. Moving on with group comparisons:

```
inds <- remna(list(x,y))
means <- tapply(y[inds], x[inds], mean)
sds <- tapply(y[inds], x[inds], sd)

stripplot(y~x, jitter=.5, main = "Chronic absenteeism by school type",ylab="adjusted
log % Chronically absent",col="green")
```



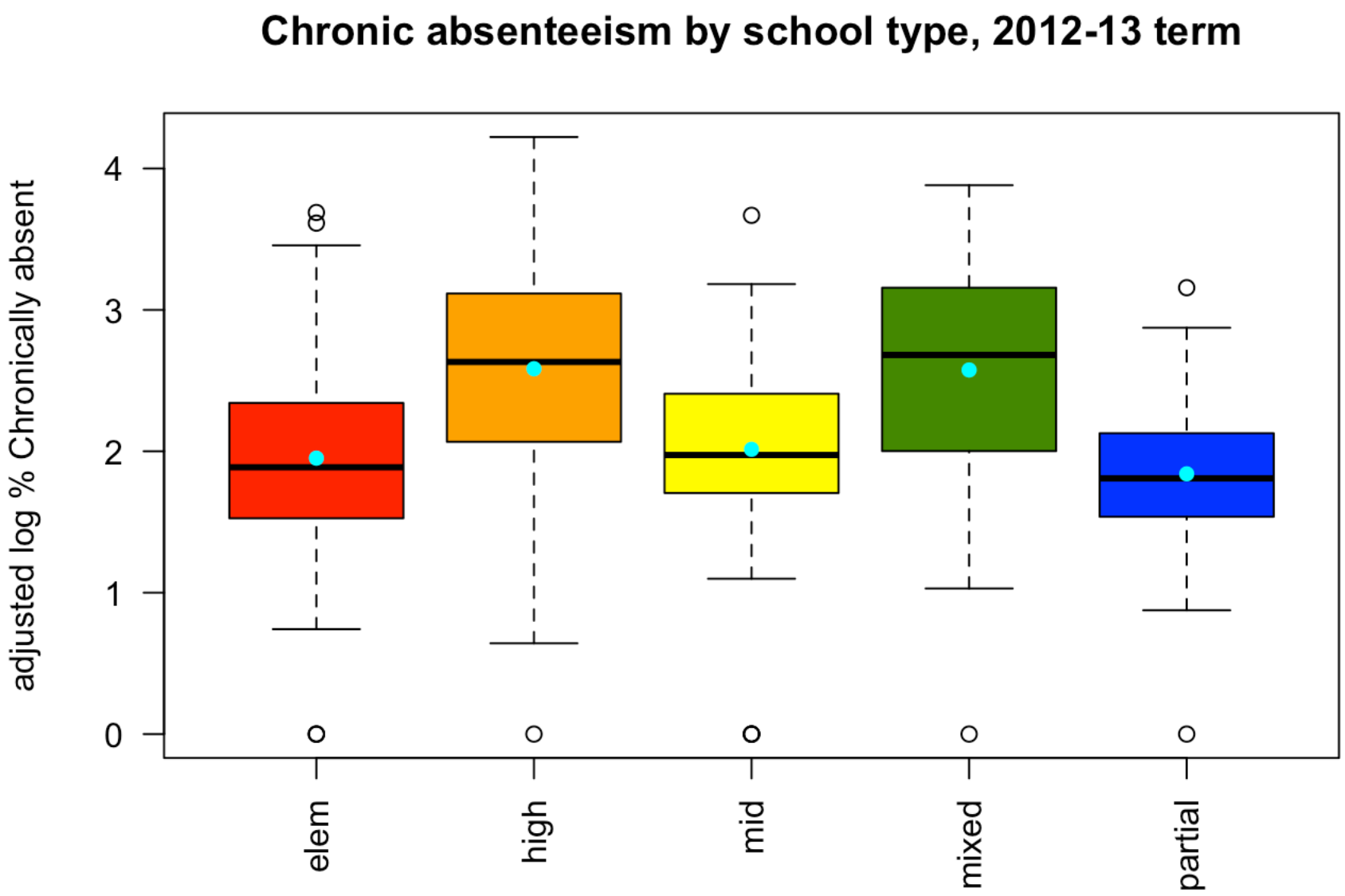

```
print("sds")
```

```
## [1] "sds"
```

```
print(sds)
```

```
##      elem      high       mid     mixed   partial
## 0.6047328 0.7683184 0.6016510 0.7195950 0.4960924
```

```
boxplot(y~x,las=2,ylab="adjusted log % Chronically absent",main="Chronic absenteeism
by school type, 2012-13 term",col=plotcols)
points(means,col="cyan",pch=16)
```



######(on any plot where it appears, ‘adjusted log’(v) = log(v+1)–this is useful for treating data which is lower-bounded at 0)

There are a large number of schools in each category (from stripplot and earlier barplot), which allows us greater confidence in the results we get. The standard deviations of each group are all under a factor of 2 of each other, so we can assume equal variances between groups. Given the checks of normality and equal variance, we may conduct ANOVA:

```
print("- - - summary - - -")
```

```
## [1] "- - - summary - - -"
```

```
aov1 <- aov(y~x)

print(Anova(aov1,type=3))
```

```
## Anova Table (Type III tests)
##
## Response: y
##              Sum Sq  Df  F value    Pr(>F)
## (Intercept) 1294.82   1 3125.989 < 2.2e-16 ***
## x             77.10   4   46.532 < 2.2e-16 ***
## Residuals    357.05 862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("-- -- -- -- -- -- -- -- --")
```

```
## [1] "-- -- -- -- -- -- -- -- --"
```

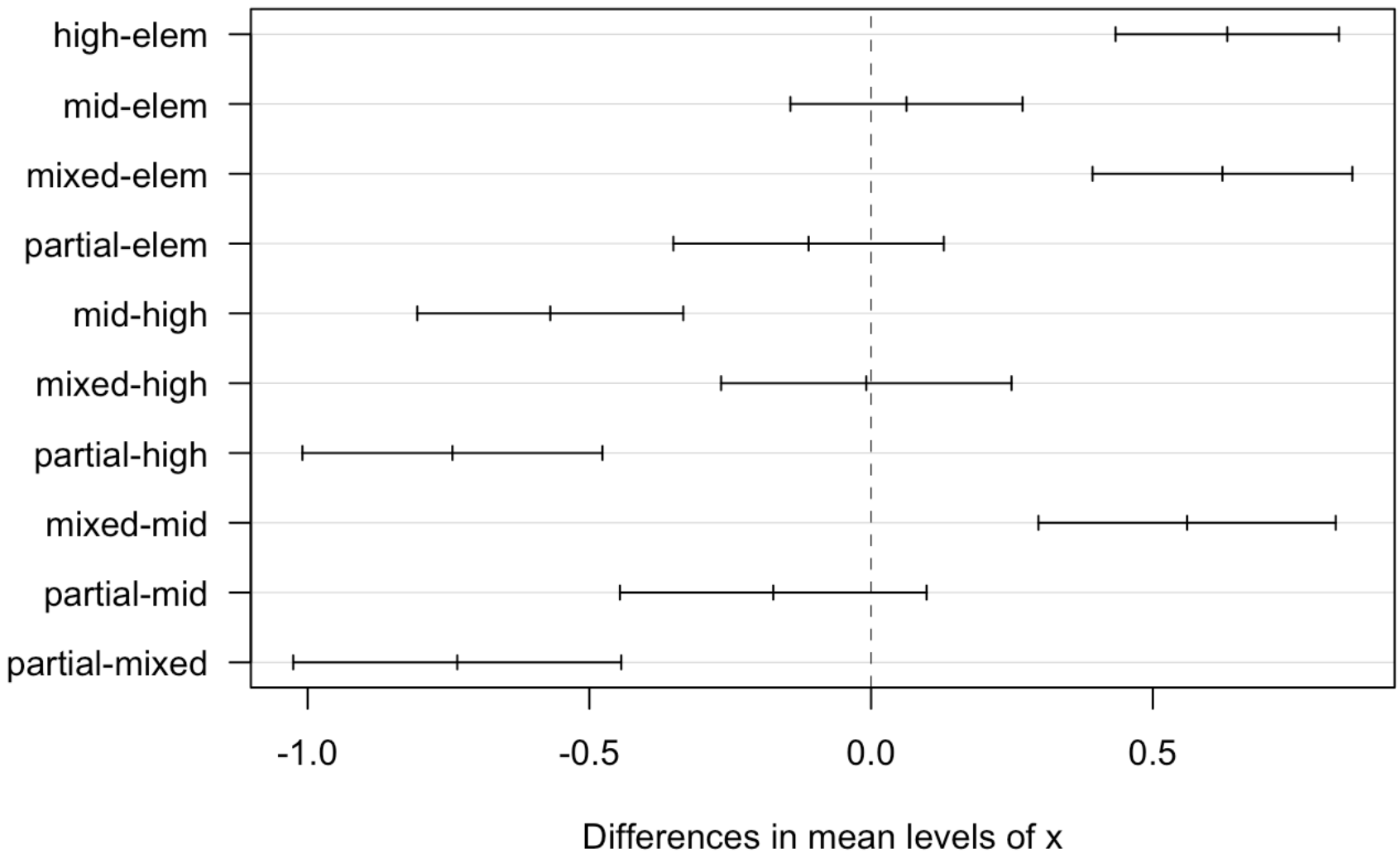
```
(welch_test <- oneway.test(y~x))
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  y and x
## F = 40.857, num df = 4.00, denom df = 327.47, p-value < 2.2e-16
```

```
print("- - - /summary - - -")
```

```
## [1] "- - - /summary - - -"
```

With such a low p-value, there is no question that a significant difference exists in graduation rates between at least two of the groups in question. As to which ones,



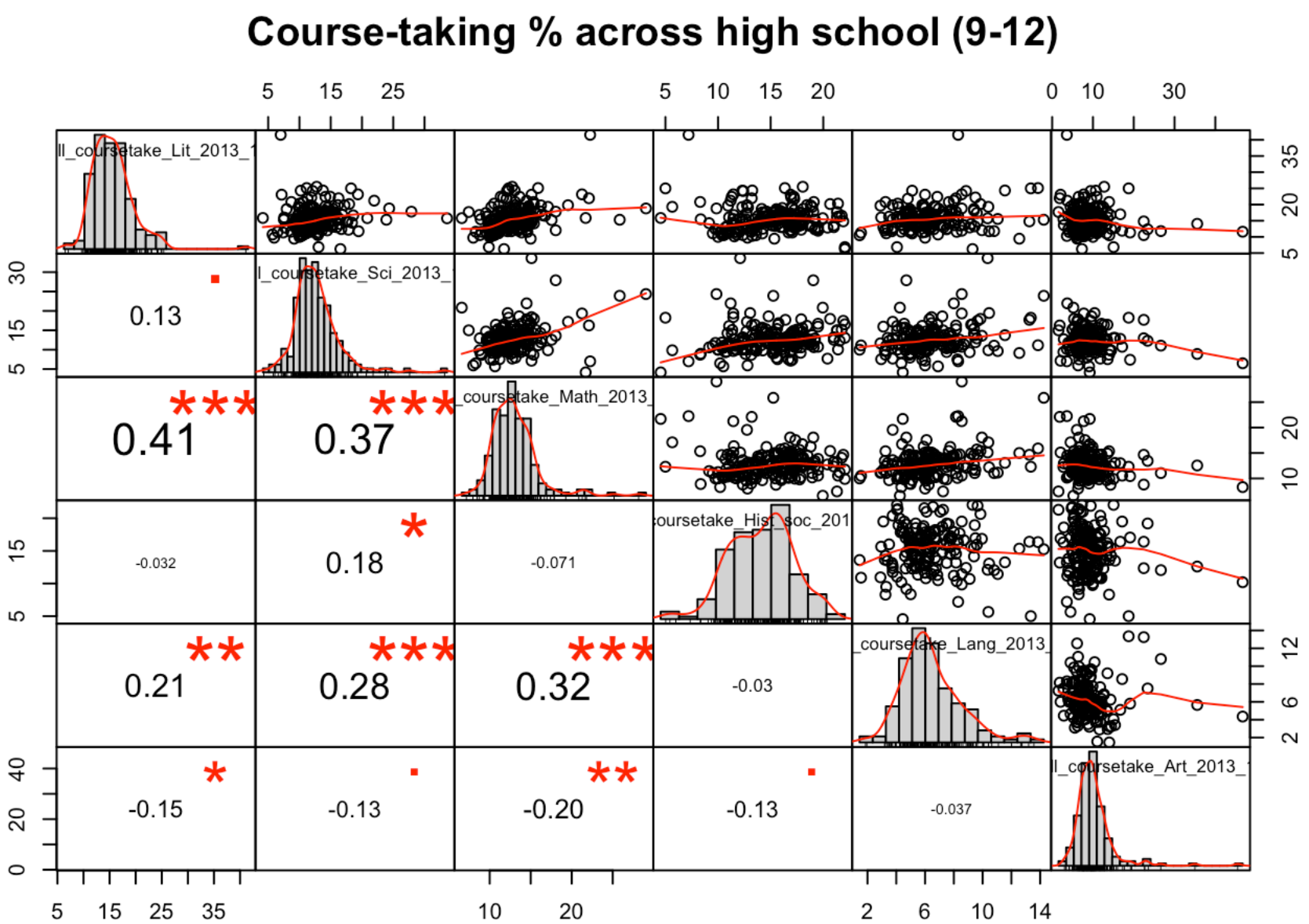
Since the ‘partial’ group is a random amalgamation of schools that didn’t fit into the other categories, there is not much I look to glean from comparisons involving them. But it is interesting to note that high schools and mixed schools (many of which include the high school grade levels) have consistently higher absenteeism rates than schools with lower grade levels. A lot of research, and a very cursory observation of any high school, tells us that students become more disengaged from school as they age, so this result is not surprising (it is in fact a good sanity check on the data).

Course-taking

There are a number of reasons why we might be interested to know what courses students have exposure to. The courses that students take in high school will be important for their graduation and college-related decisions, and for many it becomes immediately important for seeking employment. While the data available on course-taking here is available at the school level, it is not available at the student level, although this would provide much greater insight into how course-taking habits relate to graduation and future economic opportunity; this inadequate operationalization may lead to insignificant or odd results when fitting predictive models later.

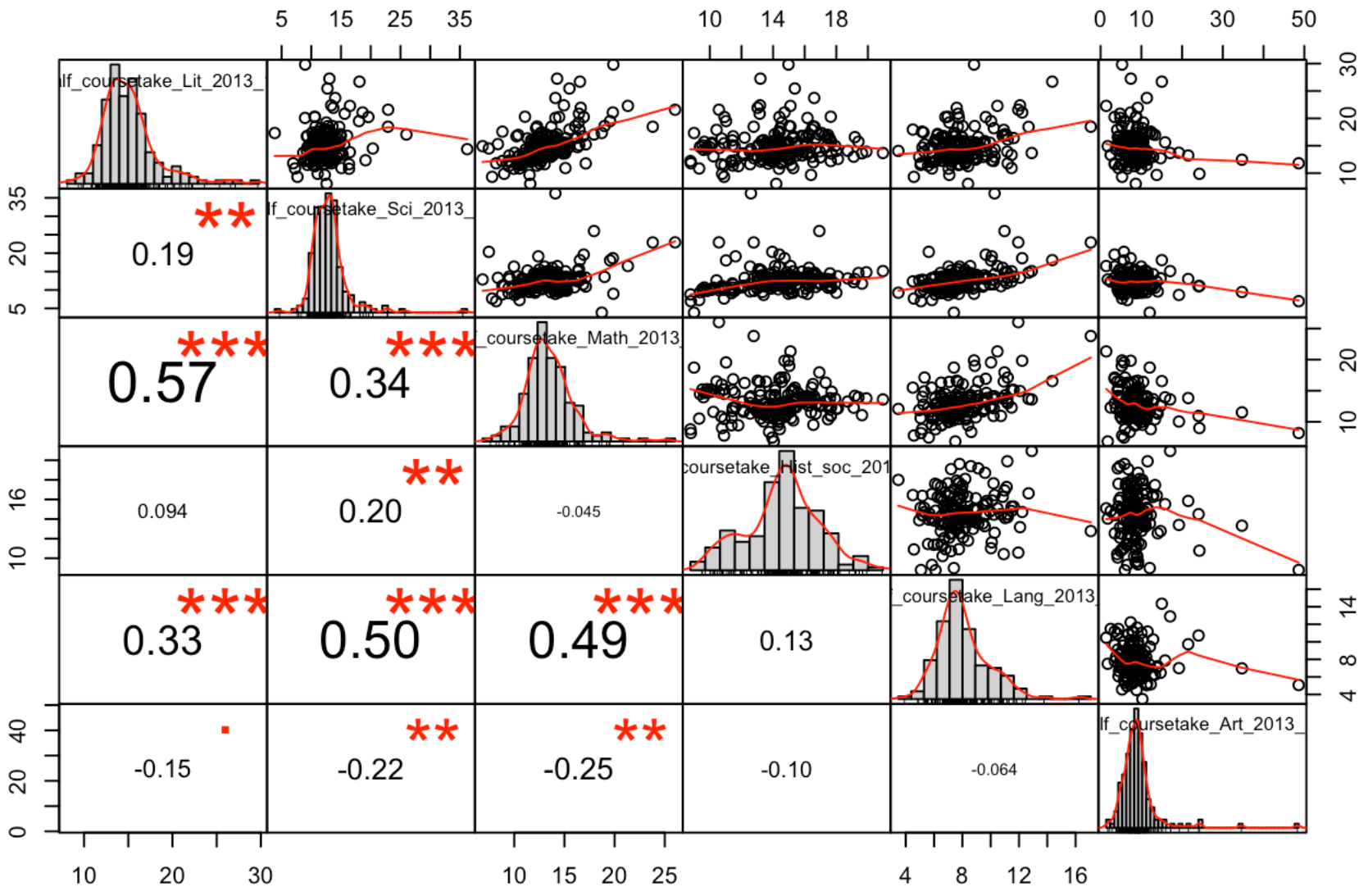
Course-taking was operationalized was as an average of the percentages of students in different grades taking courses in a select few subjects (English Language/Literature, Math, Science, History/Social Science, Language, Performance/Fine Arts). For each subject and year, there are two data points: the average of course-taking percentages in all the high-school grades (9-12), and the average of just grades 11 and 12.

```
df_ct <- full_data[,grepcol(full_data,paste0(paste0("_",ct_an,"_"),collapse="|"))]
df_ct <- df_ct[,grepcol(df_ct,c("2013_14","full"))]
pairsJDRS(df_ct,main="Course-taking % across high school (9-12)")
```




```
df_ct <- full_data[,grepcol(full_data,paste0(paste0("_",ct_an,"_"),collapse="|"))]
df_ct <- df_ct[,grepcol(df_ct,c("2013_14","half"))]
pairsJDRS(df_ct,main="Course-taking % across Grades 11 & 12")
```

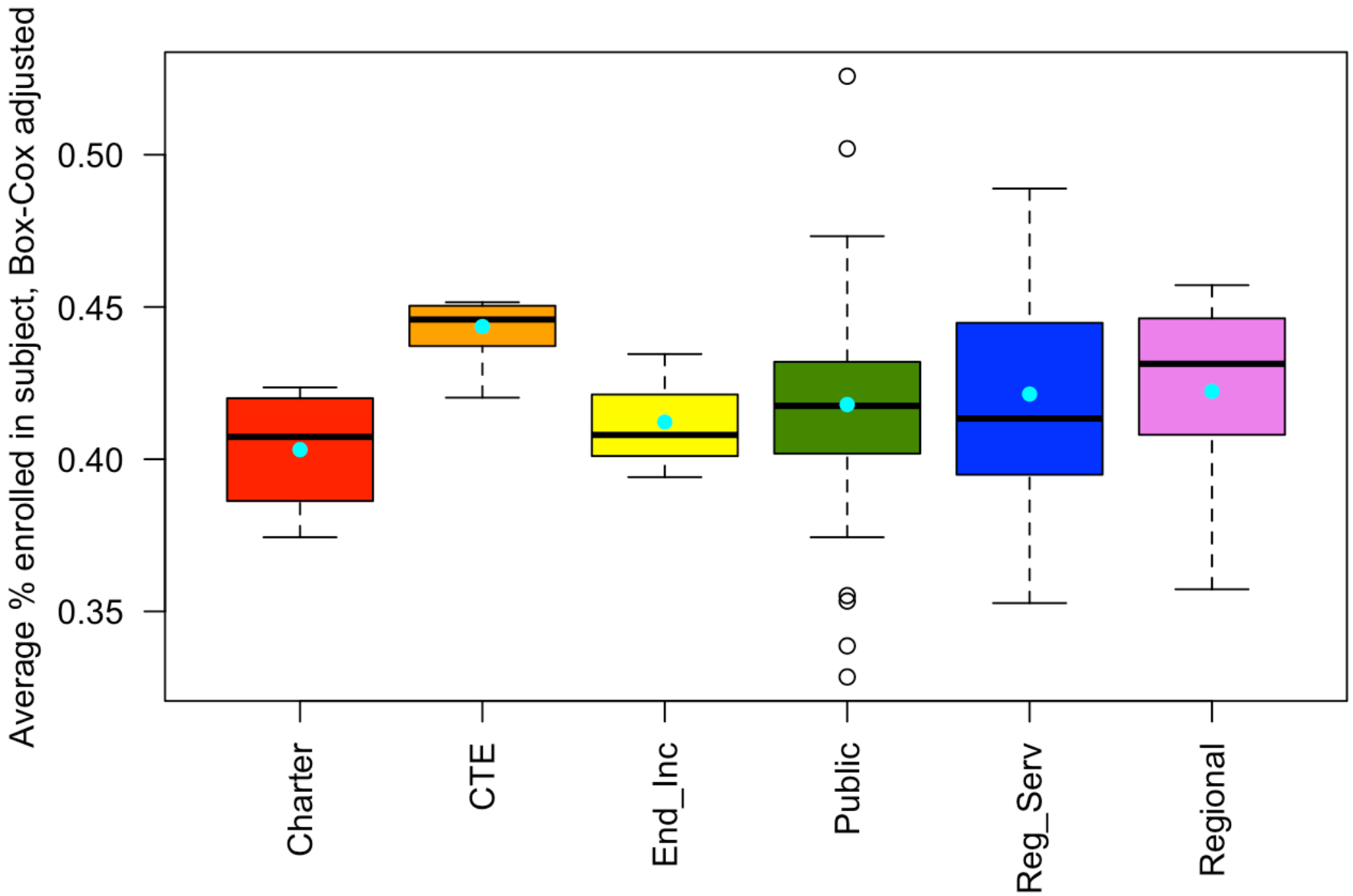
Course-taking % across Grades 11 & 12



The correlations in this plot likely represent the requirements of schools and the state of CT in what students are taking; e.g., math and ELA are consistently required under state testing programs, so it is not surprising that they have the highest correlation (this is pure speculation). There are some variable which may be imperfectly multicollinear (e.g. language and science coursetaking). There are also some outliers which should be ignored when applying regressions.

When considering graudation rates later I will stick to the measures of course-taking across only the upper grades of high school, as they focus on the students who are closest to graduating. But for now, as an exploratory approach we can look at the data aggregated across all grade levels.

Are there differences in course-taking patterns across school types?




```
## [1] "sds"
```

```
##     Charter         CTE     End_Inc      Public    Reg_Serv    Regional
## 0.021971589 0.008735643 0.020535812 0.026457374 0.044071426 0.031347126
```

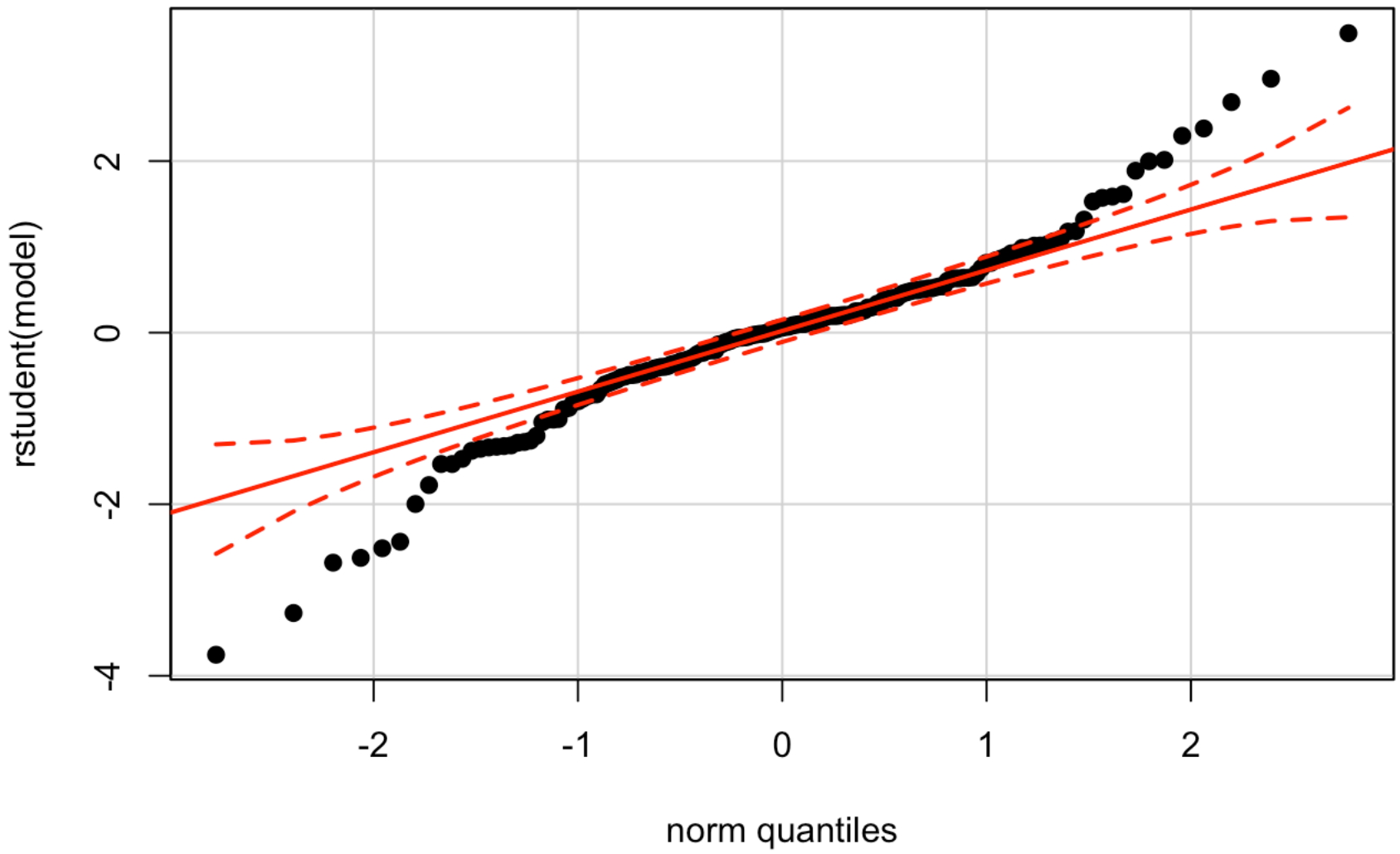
Keep in mind that charter schools, according to this plot, have higher rates of math class enrollment: the power suggested by the Box-Cox transformation was negative. There appears to be a trend, but be careful…

```
stripplot(y~x, jitter=.5,ylab="Average % enrolled in subject, Box-Cox adjusted",main=
"Math enrollment: mean(grades 11,12), 2013-14 term",col="green")
```

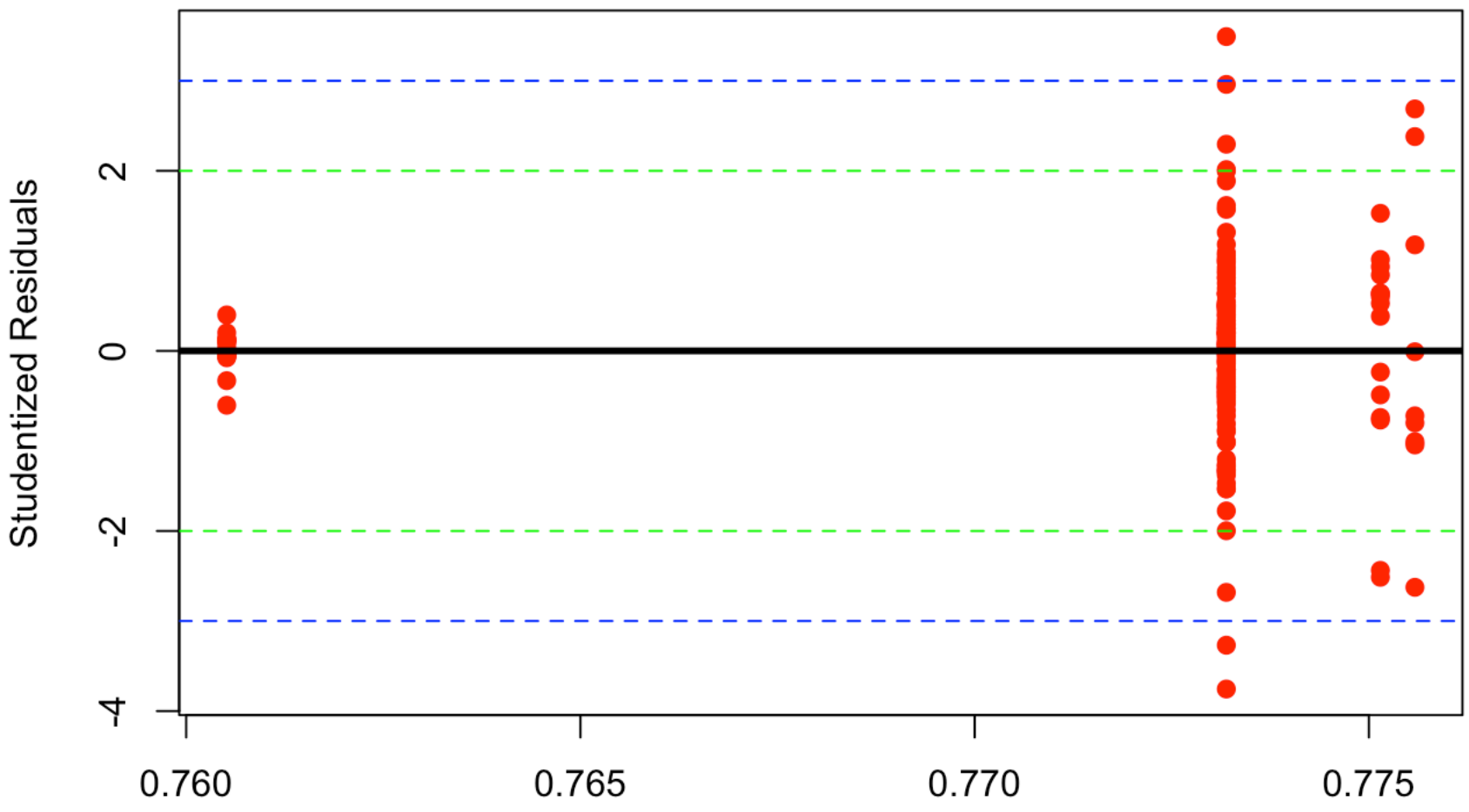
There are very few charter schools and endowed/incorporated schools with high-school grades in this dataset–I am not inclined to put much stock in their results without larger samples, so I will exclude them from any analysis limited to high school grade levels.

```
detach(full_data_noncharter)
myResPlots2(lm(y ~ x))
```

NQ Plot of Studentized Residuals, Residual Plots



Fits vs. Studentized Residuals, Residual Plots



```
inds <- remna(list(x,y))
means <- tapply(y[inds], x[inds], mean)
sds <- tapply(y[inds], x[inds], sd)

print("sds")
```

```
## [1] "sds"
```

```
print(sds)
```

```
##         CTE      Public    Reg_Serv    Regional
## 0.003249092 0.014691091 0.024197775 0.017133611
```

The variances between groups are uneven, and the maximal variance unevenness is more than 2. The different vertical spans may be attributable to restriction in the data range–schools not in the public category are very underrepresented. At the ends the residual distribution departs from normality. We can conduct traditional ANOVA, but to supplement it let us also have Welch and Kruska-Wallis tests:

```
print("- - - summary - - -")
```

```
## [1] "- - - summary - - -"
```

```
aov1 <- aov(y~x)

print(Anova(aov1,type=3))
```

```
## Anova Table (Type III tests)
##
## Response: y
##             Sum Sq  Df   F value  Pr(>F)
## (Intercept) 9.2541   1 41807.044 < 2e-16 ***
## x           0.0026   3     3.842 0.01072 *
## Residuals   0.0387 175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("-- -- -- -- -- -- -- -- --")
```

```
## [1] "-- -- -- -- -- -- -- -- --"
```

```
(welch_test <- oneway.test(y~x))
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  y and x
## F = 25.425, num df = 3.000, denom df = 24.763, p-value = 9.879e-08
```

```
print("- - - /summary - - -")
```

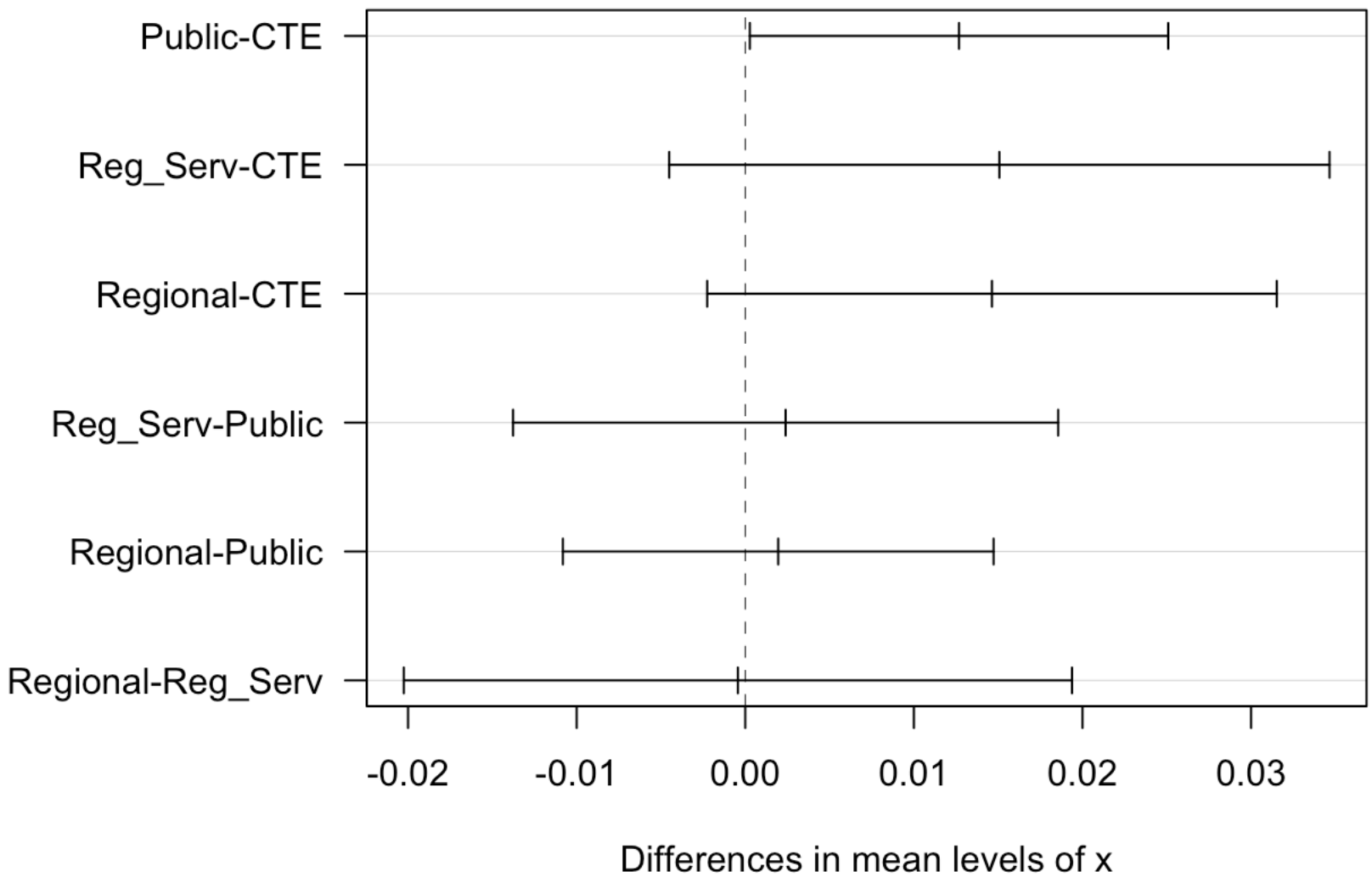
```
## [1] "- - - /summary - - -"
```

```
print(kruskal.test(y ~ as.factor(x)))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  y by as.factor(x)
## Kruskal-Wallis chi-squared = 21.616, df = 3, p-value = 7.84e-05
```

All three results agree, though the normal ANOVA test yields a much higher p-value than the other tests. In any case, we expect that a difference in average math course-taking among 11th & 12th graders exists somewhere among the groups we have examined. For the sake of completeness,

```
par(mar=c(5,10,4,2))
tuk <- TukeyHSD(aov1,conf.level = .99)
plot(tuk,las=1)
```

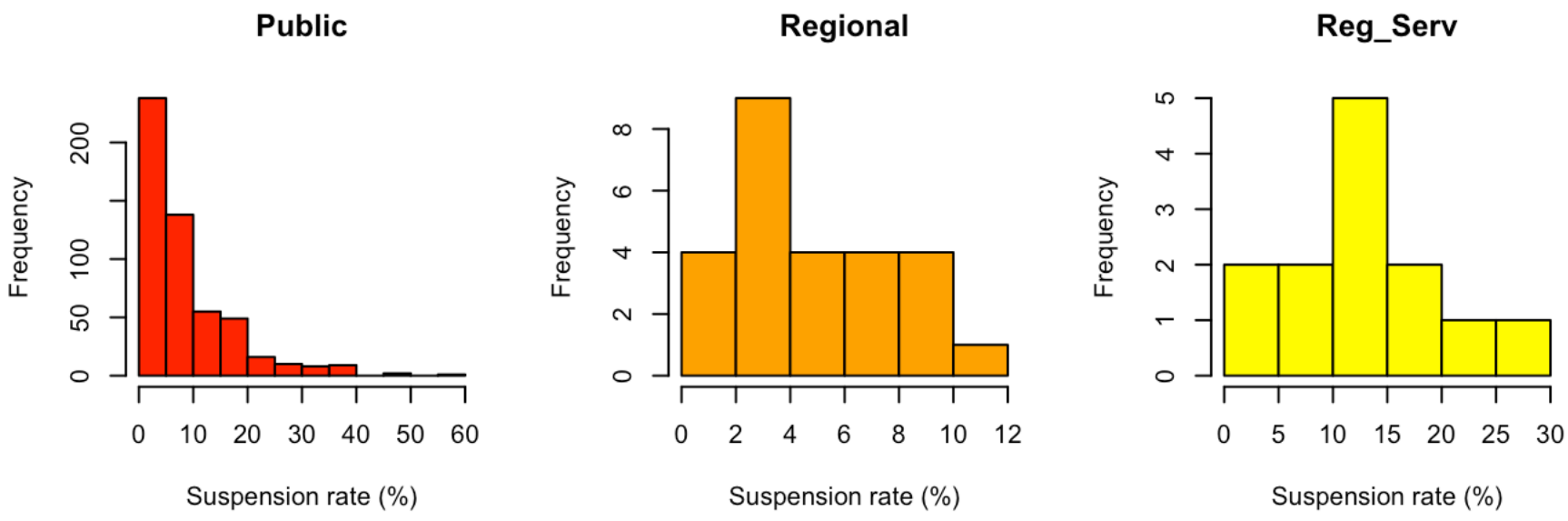


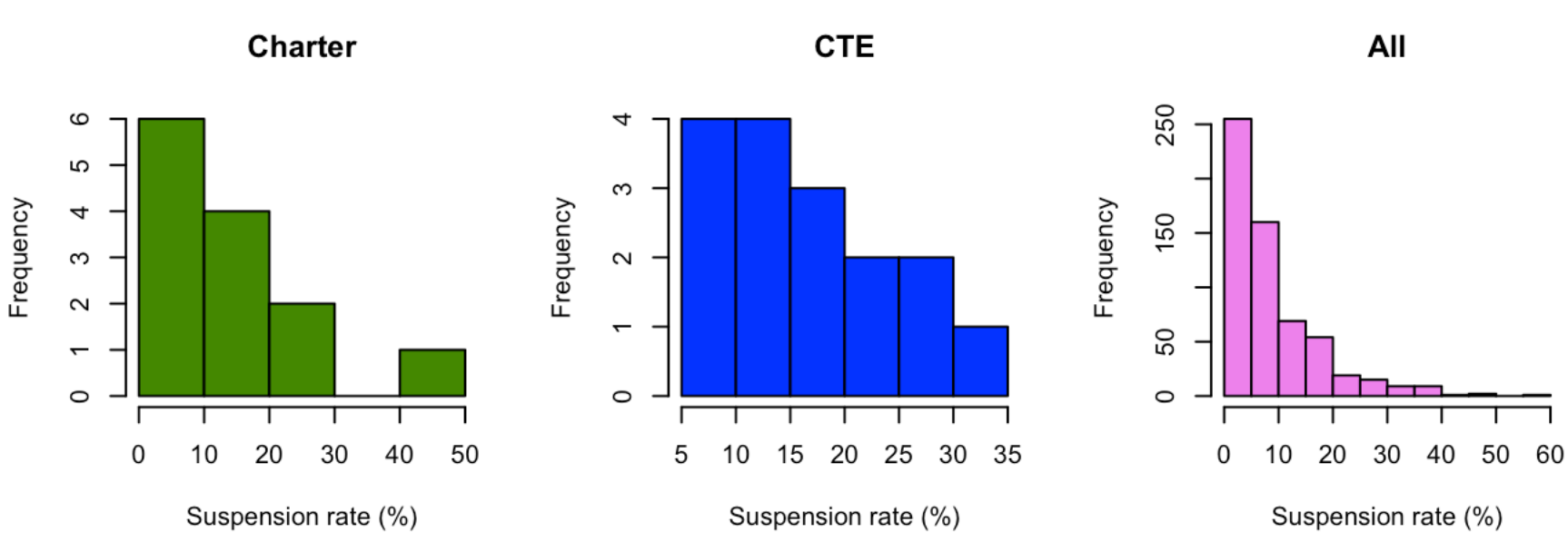
Interestingly, everything here is either clearly insignificantly different, not different at all, or barely measurably different. Public & CTE schools appear to be the one pair that barely attains statistical significance.

Discipline

One more thing to overview is disciplinary atmosphere. There is a fair deal of research to support the notion that ‘school climate’, which is a mix of the disciplinary policy enforced for and support offered to students, has impacts on student engagement and performance. It would be particularly interesting to have charter and nonpublic schools included in this analysis, as many critics of schools outside the traditional public sector are quick to assert (especially of charters) that these schools bolster their numbers by selecting students in a way that public schools cannot.

```
full_data_nonendinc <- full_data[!grepl("End_Inc",full_data$Category),]
attach(full_data_nonendinc)
y <- susp_2013_14_
x <- Category
detach(full_data_nonendinc)
par(mfrow=c(2,3))
xu <- unique(x)
for (i in 1:length(xu)){
    hist(y[x==xu[i]],main=xu[i],col=plotcols[i],xlab='Suspension rate (%)')
}; hist(y,main="All",col='violet',xlab='Suspension rate (%)')
```






```
dev.off()
```

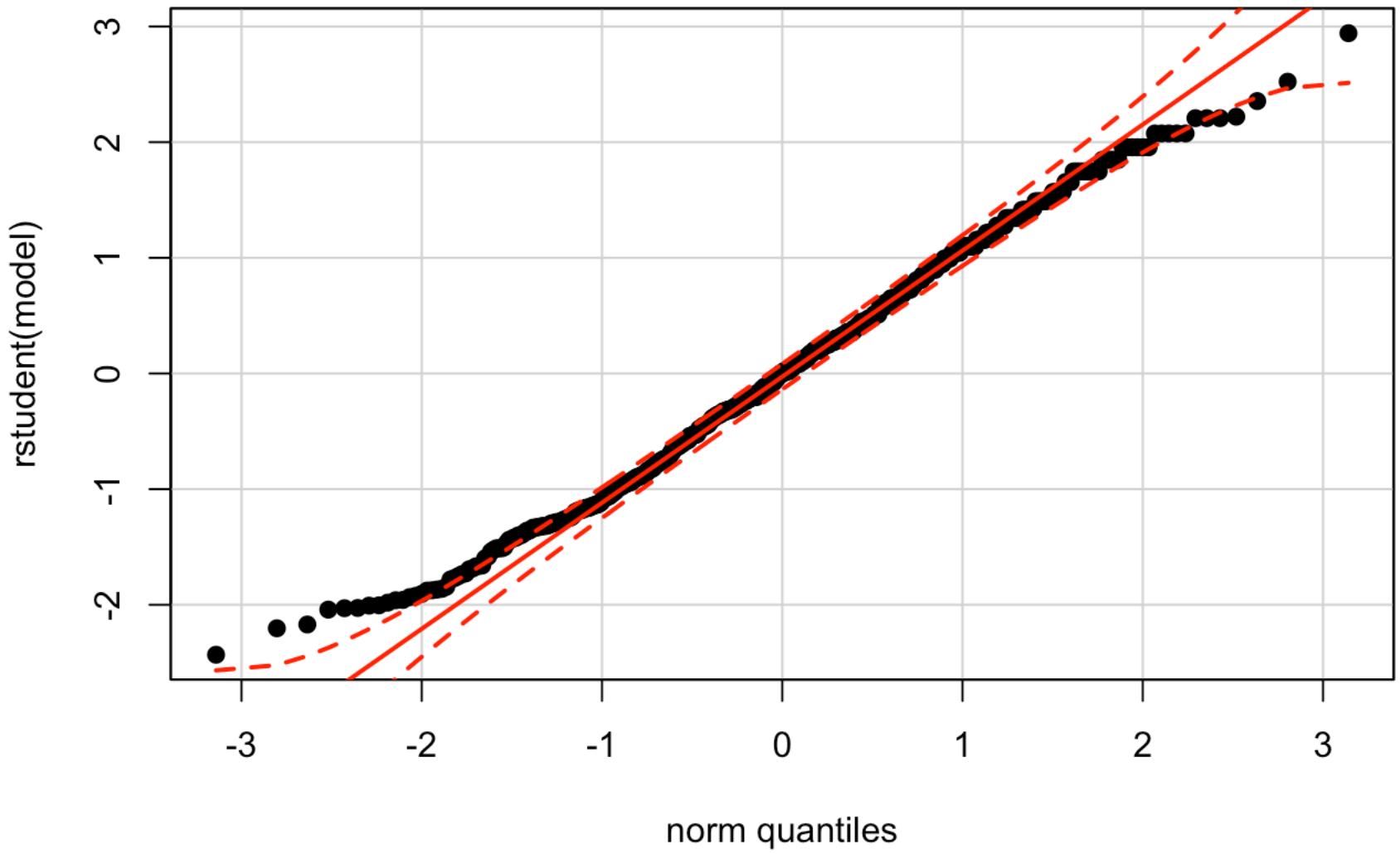
```
## null device
##           1
```

```
## [1] "lambda: -0.1010101010101"
```

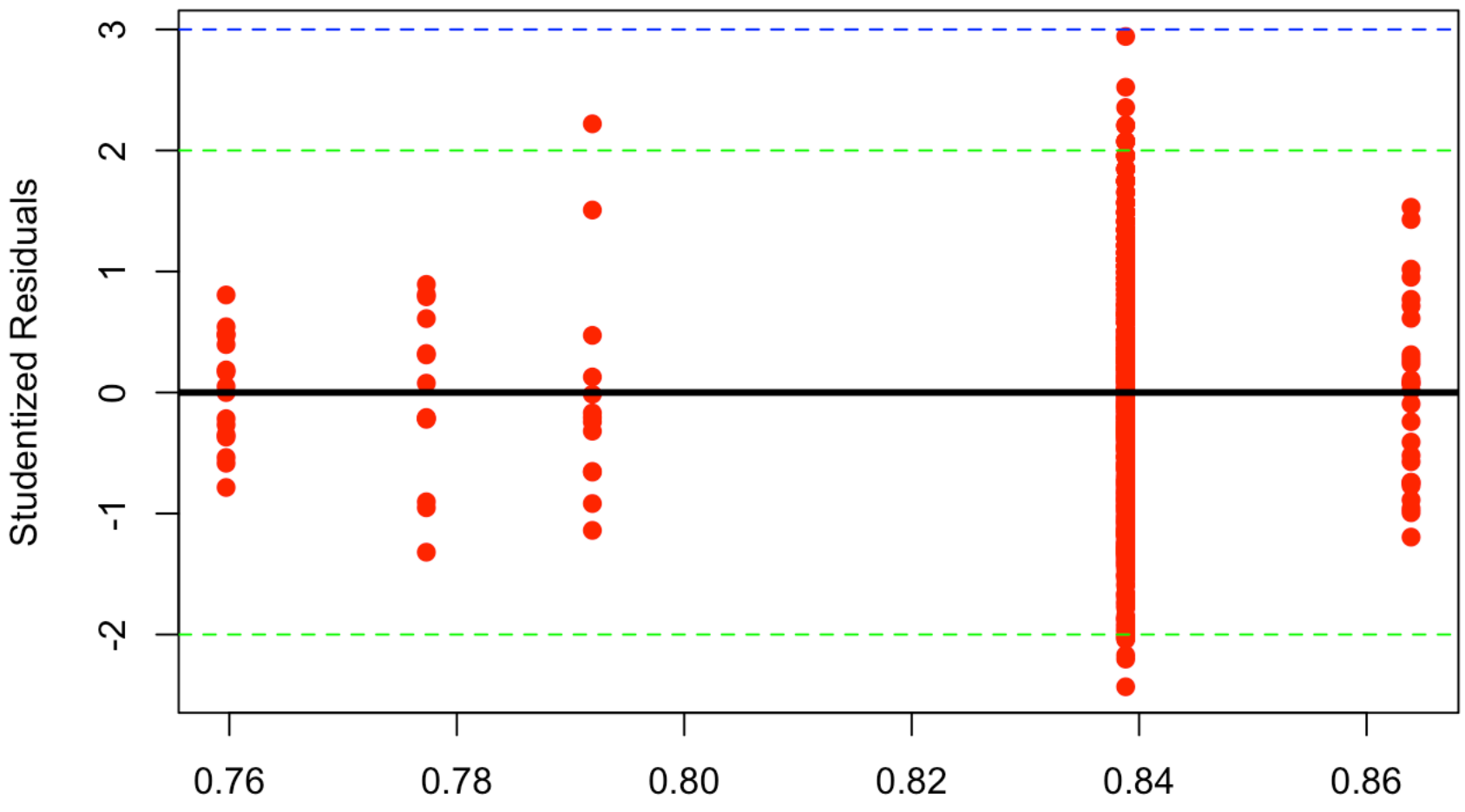
```
## [1] "mean comparison"
```

```
## $Charter
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    5.60    7.10   11.30   15.31   14.80   42.70       1
##
## $CTE
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.40   10.42   14.85   16.71   21.20   32.30
##
## $Public
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.600   3.100   5.650   8.791  11.850  59.900     299
##
## $Reg_Serv
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.70    9.00   12.10   13.01   18.20   29.00      11
##
## $Regional
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.300   2.775   4.100   5.173   7.725  12.000      24
```

NQ Plot of Studentized Residuals, Residual Plots



Fits vs. Studentized Residuals, Residual Plots

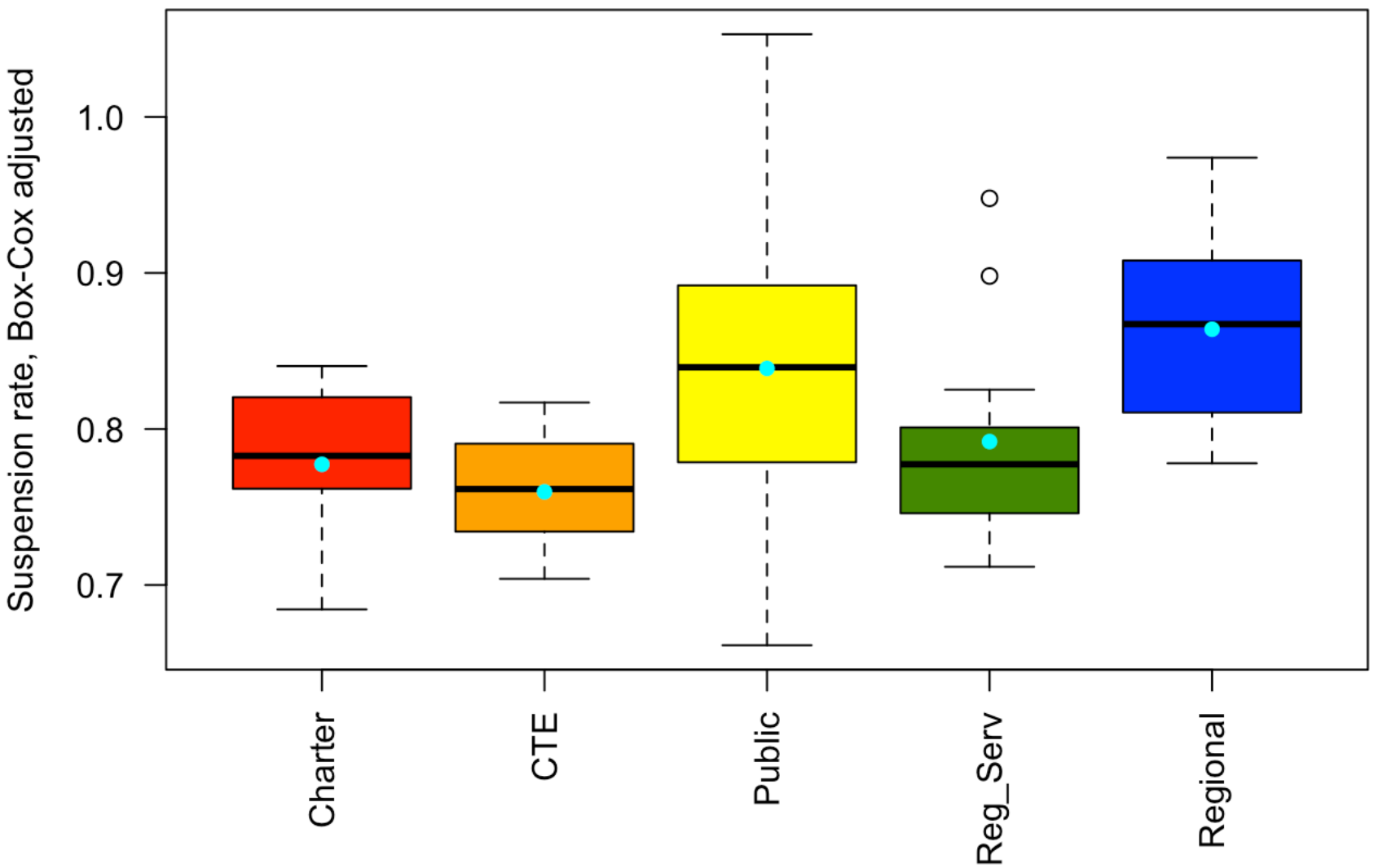


Fitted Values

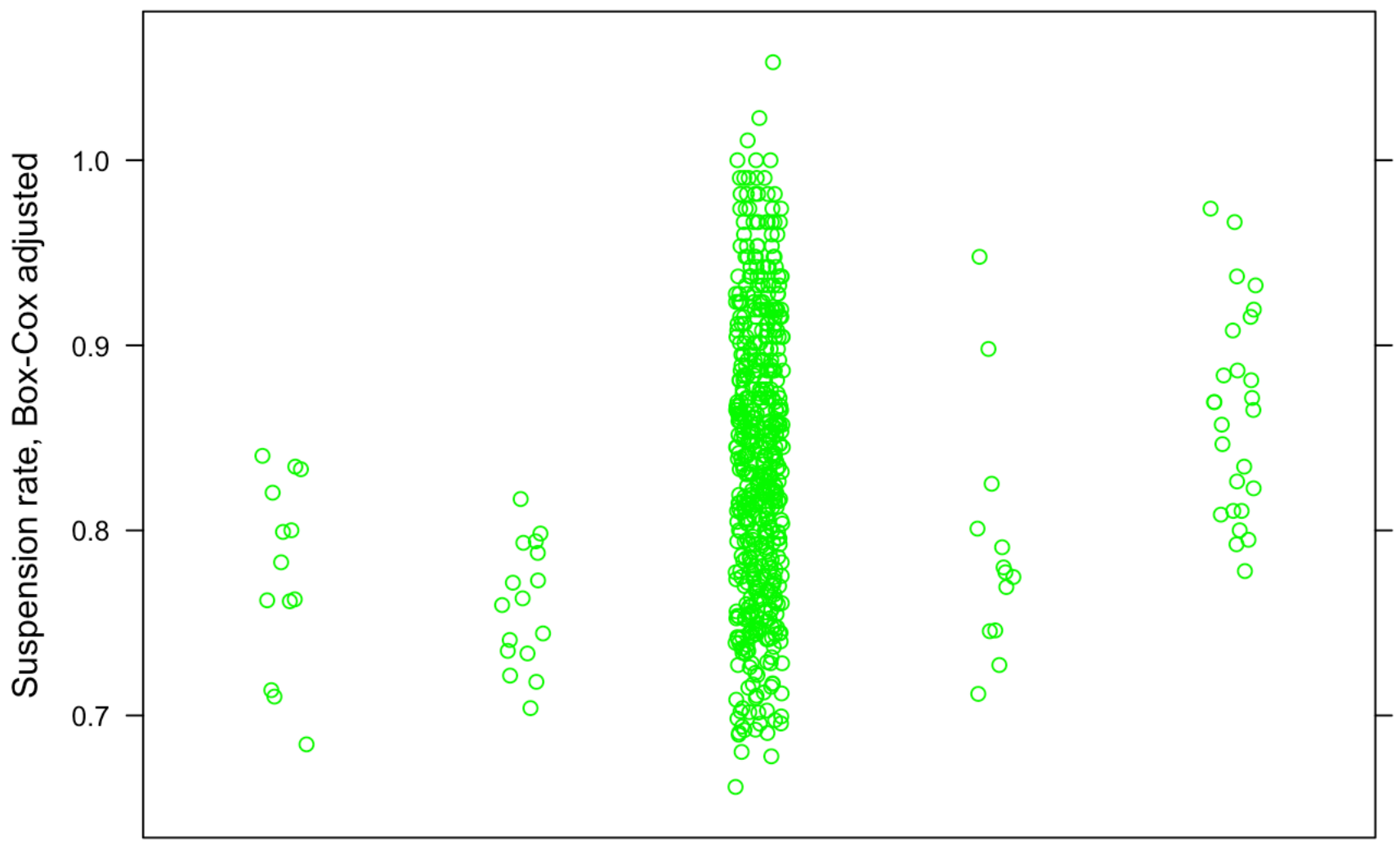
```
## [1] "sds"
```

```
##    Charter        CTE     Public   Reg_Serv   Regional
## 0.05089485 0.03304983 0.07547675 0.06632368 0.05515459
```

Suspension rates reciprocated by sector, 2013-14 term



Suspension rates reciprocated by sector, 2013-14 term



Charter CTE Public Reg_Serv Regional

The model assumptions appear to be met (again, the differing vertical spreads might be due to small subsample sizes, though it is uncertain–Welch ANOVA is conducted just in case). By eye, there appears to be a trend–public schools seem to have lower suspension rates than the categories to the left. Let's test it:

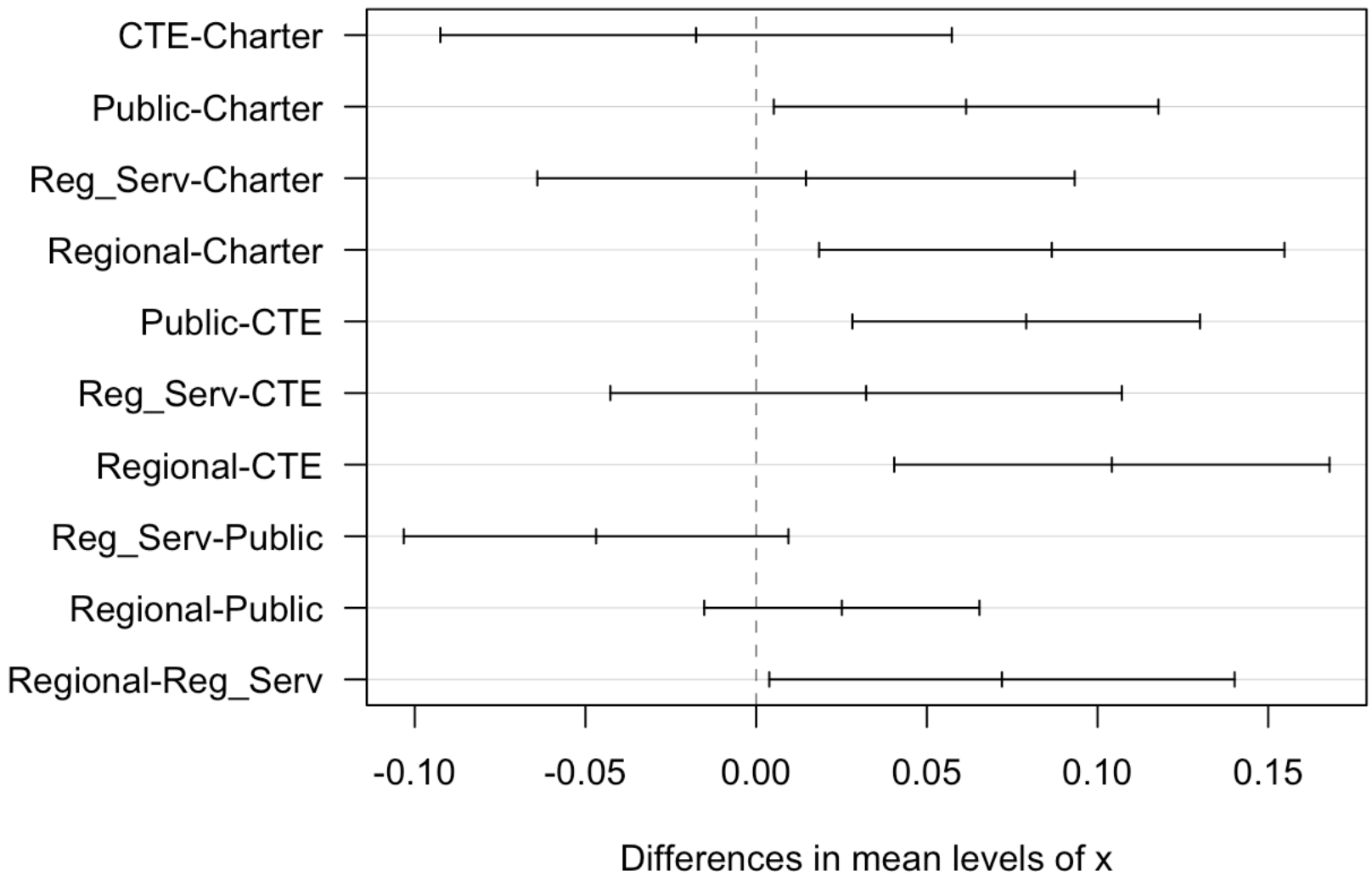
```
## [1] "- - - summary - - -"
```

```
## Anova Table (Type III tests)
##
## Response: y
##             Sum Sq  Df   F value    Pr(>F)
## (Intercept) 7.8550   1 1460.8301 < 2.2e-16 ***
## x           0.1874   4    8.7112 7.775e-07 ***
## Residuals   3.1671 589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "-- -- -- -- -- -- -- -- --"
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  y and x
## F = 25.183, num df = 4.000, denom df = 32.968, p-value = 1.273e-09
```

```
## [1] "- - - /summary - - -"
```



At the 95% confidence level, we can say that charter schools and CTE schools have higher rates of suspension than public schools.

Now there is a problem here–the anove analysis is secretly comparing apples to oranges. CTE schools really should not be compared with the entire sample of public schools, as they only operate at the high school level in this data sample, whereas public schools operate at all levels. Similarly, we have no elementary or mixed-level charters in the sample:

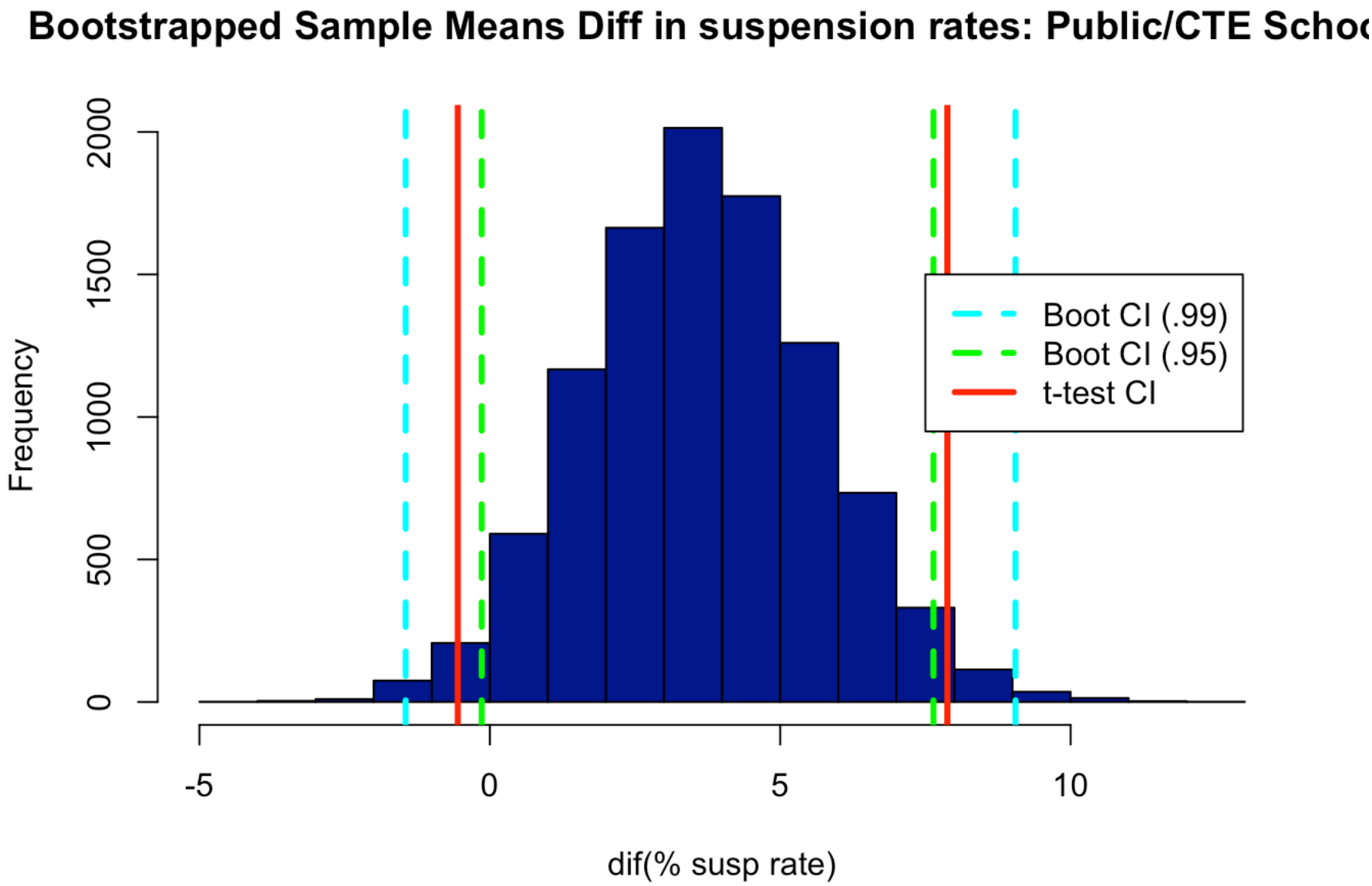
Let's test only the difference in suspension rates between CTE, public, and charter schools at the matching levels:

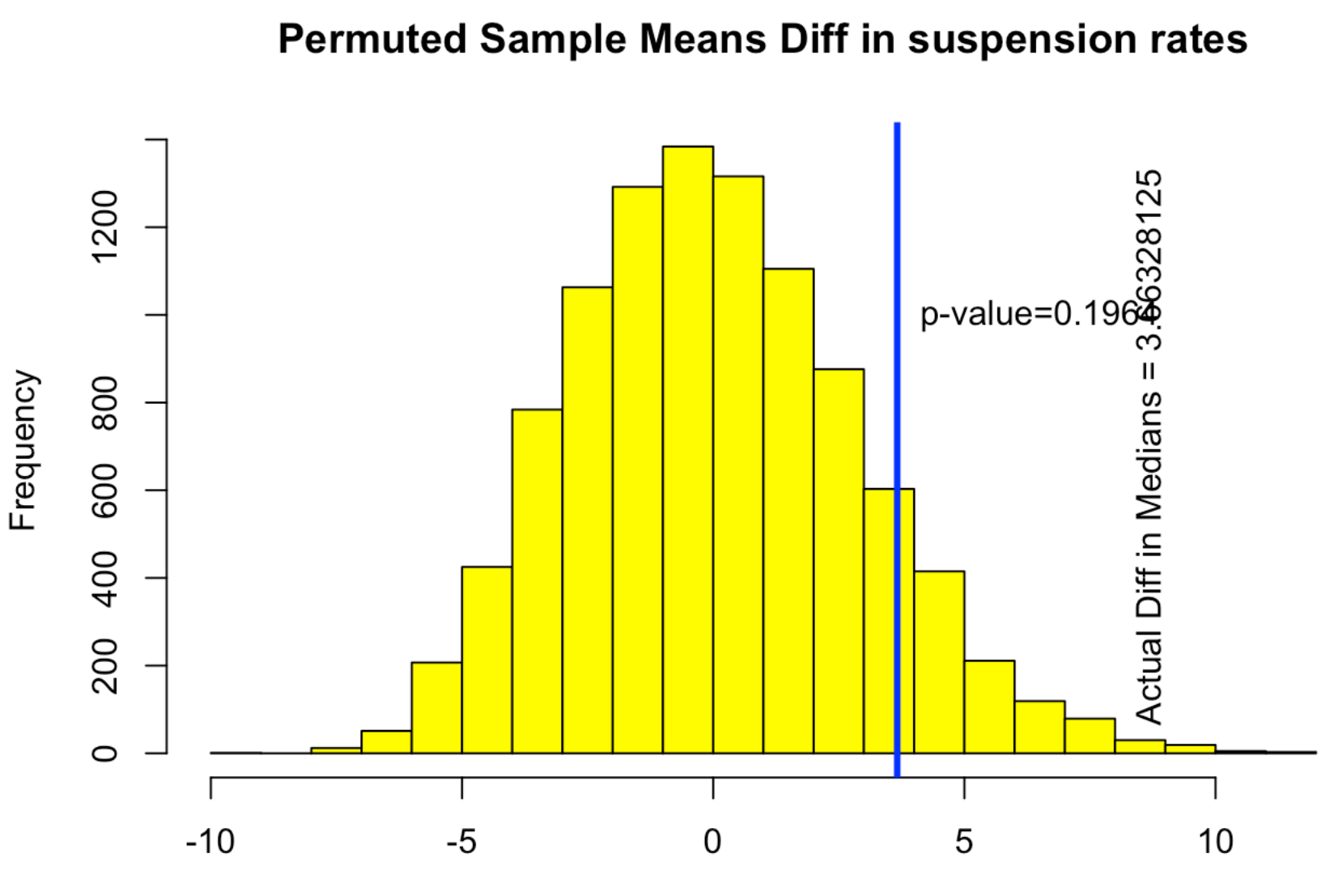
```
full_data_high <- full_data[full_data$gradelevels=="high",]
attach(full_data_high)
y <- susp_2013_14_
x <- Category
detach(full_data_high)

lp <- c(7.5,1500,10,1000)
tp <- c(5,700,6.5,1000,5,550,7,1000)

print(tapply(y, x, summary))
```

```
## $Charter
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.1    17.5    25.9    25.9    34.3    42.7
##
## $CTE
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.40   10.42   14.85   16.71   21.20   32.30
##
## $End_Inc
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   7.025   9.050   9.050  11.080  13.100
##
## $Public
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.900   4.875   9.500  13.040  16.620  59.900       6
##
## $Reg_Serv
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   12.10   12.75   13.40   16.30   18.40   23.40       2
##
## $Regional
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   3.675   5.600   5.686   7.725  10.000
```

```
bootperm(x,y,"Public","CTE",lp,tp,"suspension")
```

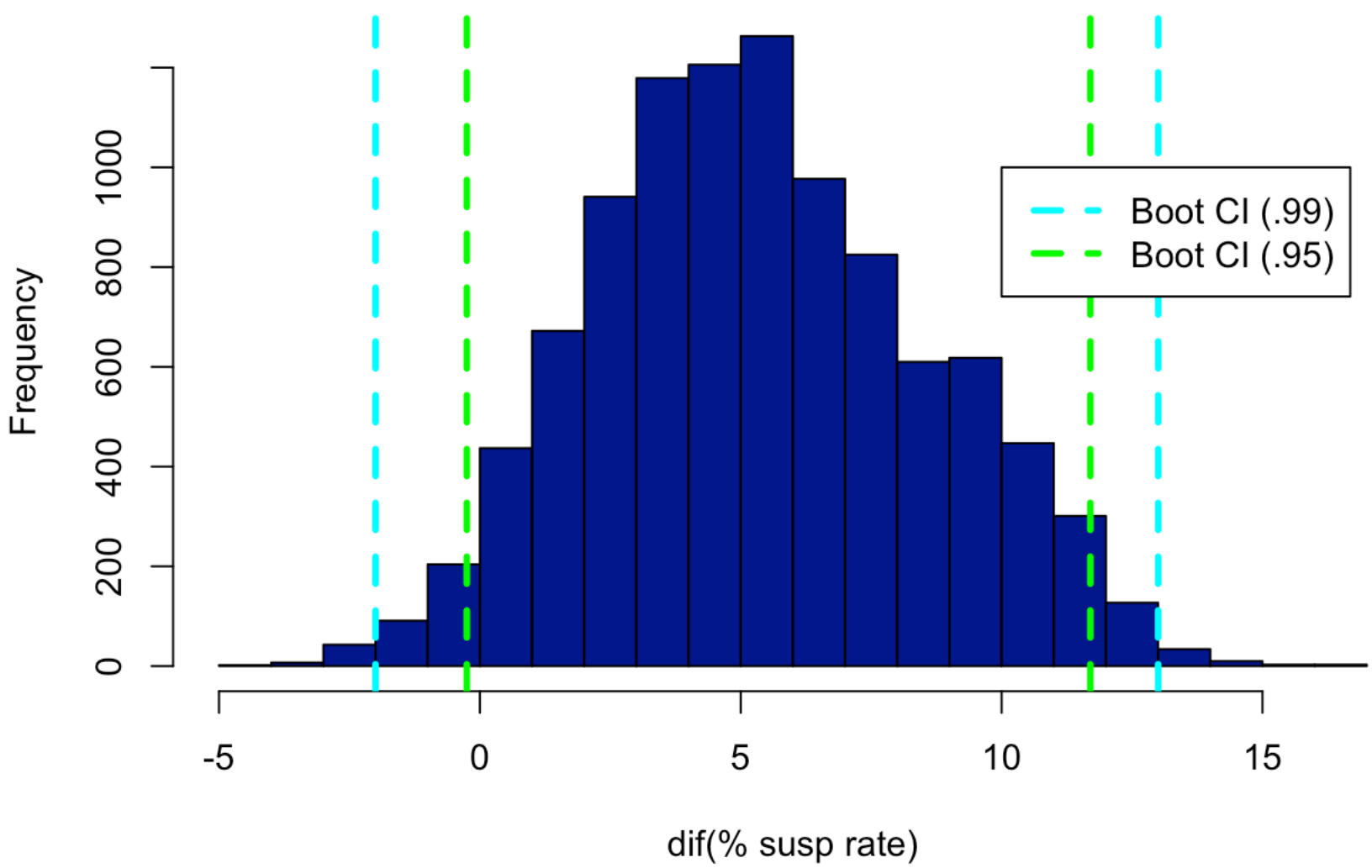
```
## [1] "observed mean: 3.66328125"
## [1] "bootstrapped confint on mean diff: -0.140136718749999  7.6384765625"
## [1] "bootstrapped confint on mean diff: -1.450046875  9.05165234374999"
## [1] "t-test"
##
##  Welch Two Sample t-test
##
## data:  g1314 by cattemp
## t = 1.79, df = 24.942, p-value = 0.0856
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5521161  7.8786786
## sample estimates:
##    mean in group CTE mean in group Public
##              16.70625              13.04297
##
## [1] "t-test conf.int"
## [1] -0.5521161  7.8786786
## attr(,"conf.level")
## [1] 0.95
```

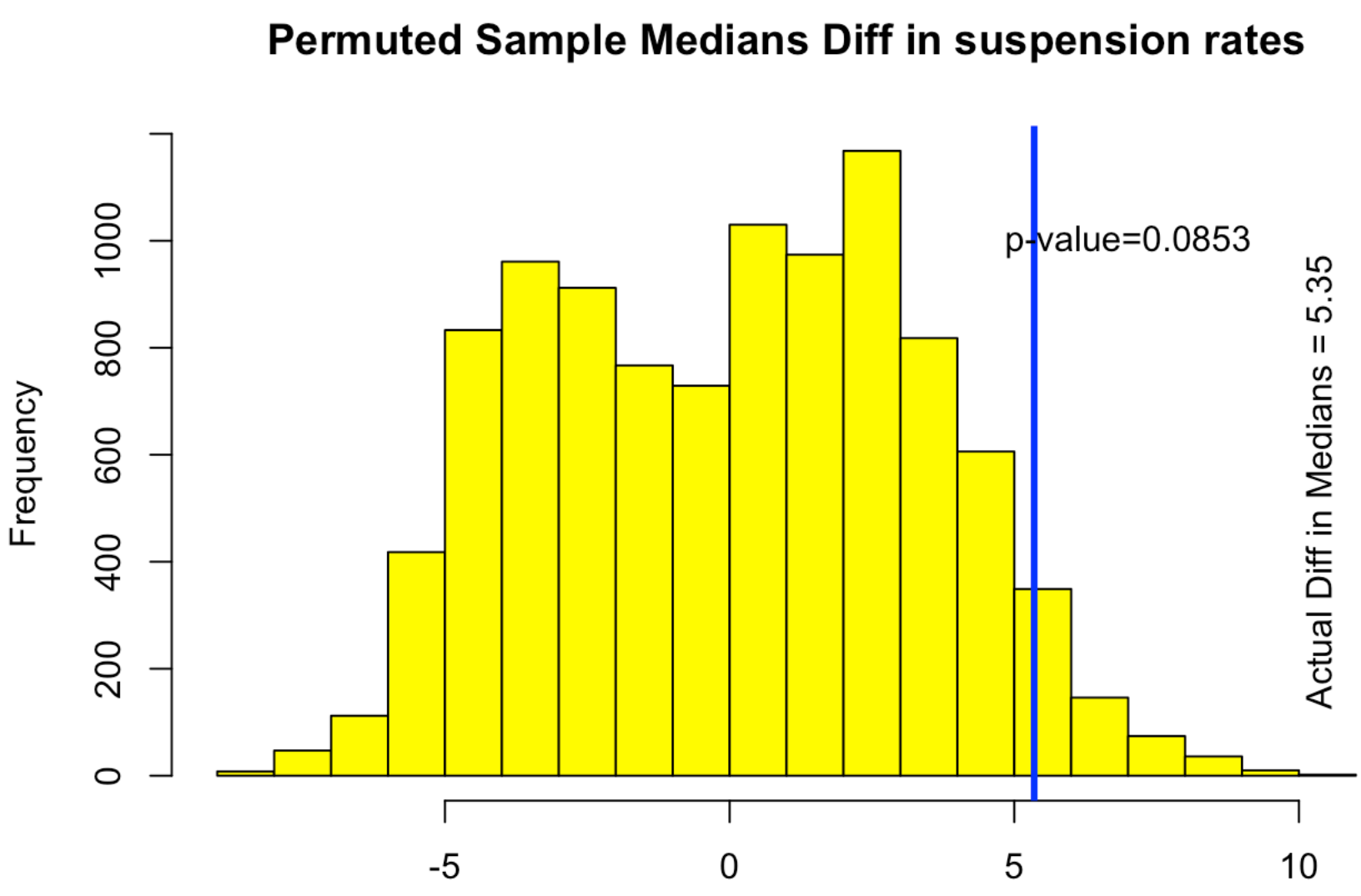

Bootstrapped Sample Means Diff in suspension rates: Public/CTE Schoo
Frequency
0
500
1000
1500
2000
-5
0
5
10
dif(% susp rate)
Boot CI (.99)
Boot CI (.95)
t-test CI


Permuted Sample Means Diff in suspension rates
Frequency
0
200
400
600
800
1200
-10
-5
0
5
10
p-value=0.196
Actual Diff in Medians = 3.46328125

dif(% susp rate)

```
## [1] "p-value=0.1964"
## [1] "observed median: 5.35"
## [1] "bootstrapped confint on median diff (.95): -0.250000000000002  11.7"
## [1] "bootstrapped confint on median diff (.99): -2  13.00025"
```


Bootstrapped Sample Medians Diff in suspension rates: Public/CTE Scho
Frequency
0
200
400
600
800
1000
-5
0
5
10
15
dif(% susp rate)
Boot CI (.99)
Boot CI (.95)


Permuted Sample Medians Diff in suspension rates
Frequency
0
200
400
600
800
1000
-5
0
5
10
p-value=0.0853
Actual Diff in Medians = 5.35

dif(% susp rate)

```
## [1] "p-value=0.0853"
```

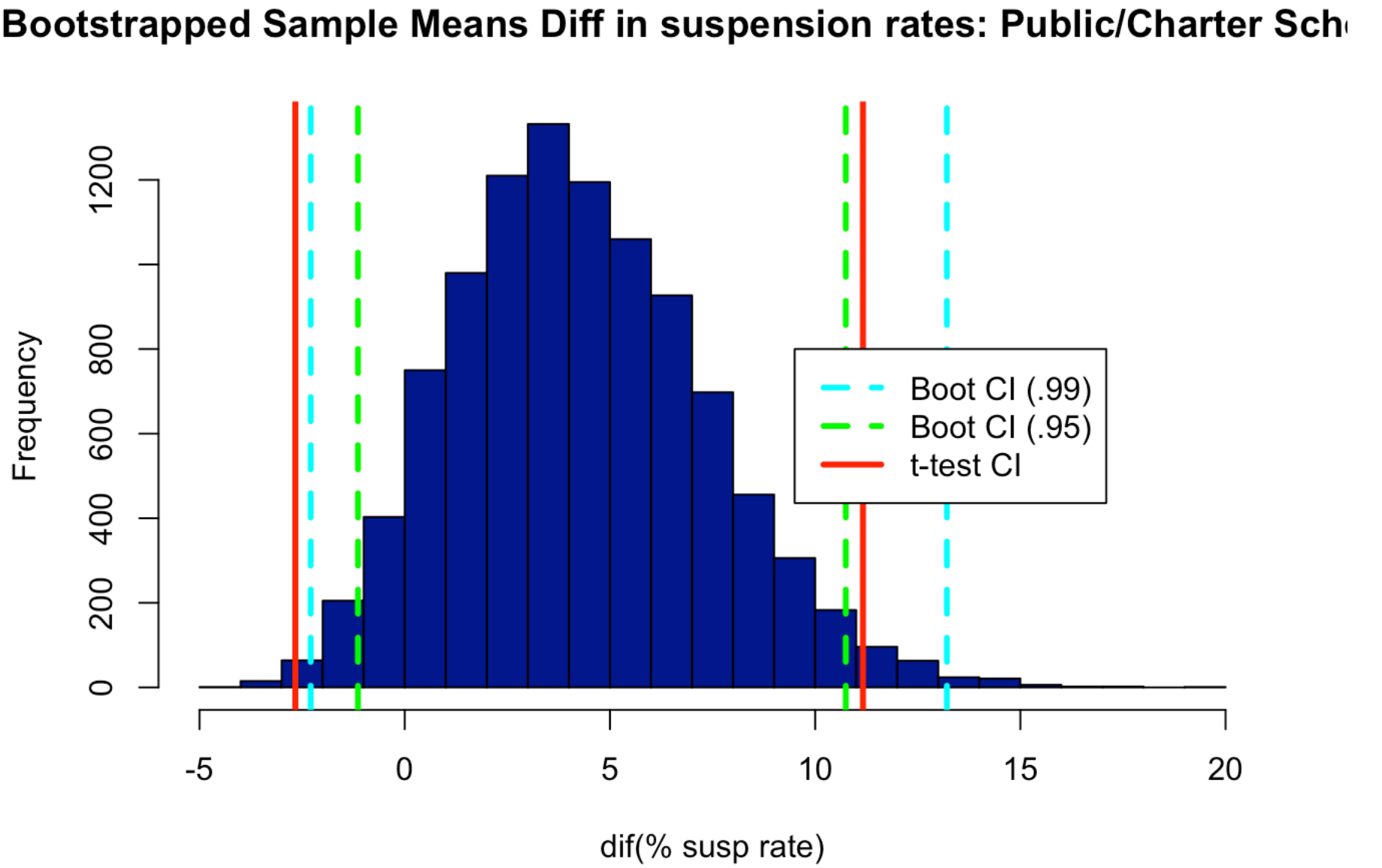
We fail to reject null difference between suspension rates of Public/CTE schools–so we cannot conclude that they are in fact statistically distinguishable. Were there more CTE schools in the sample, the difference may have been rendered significant.

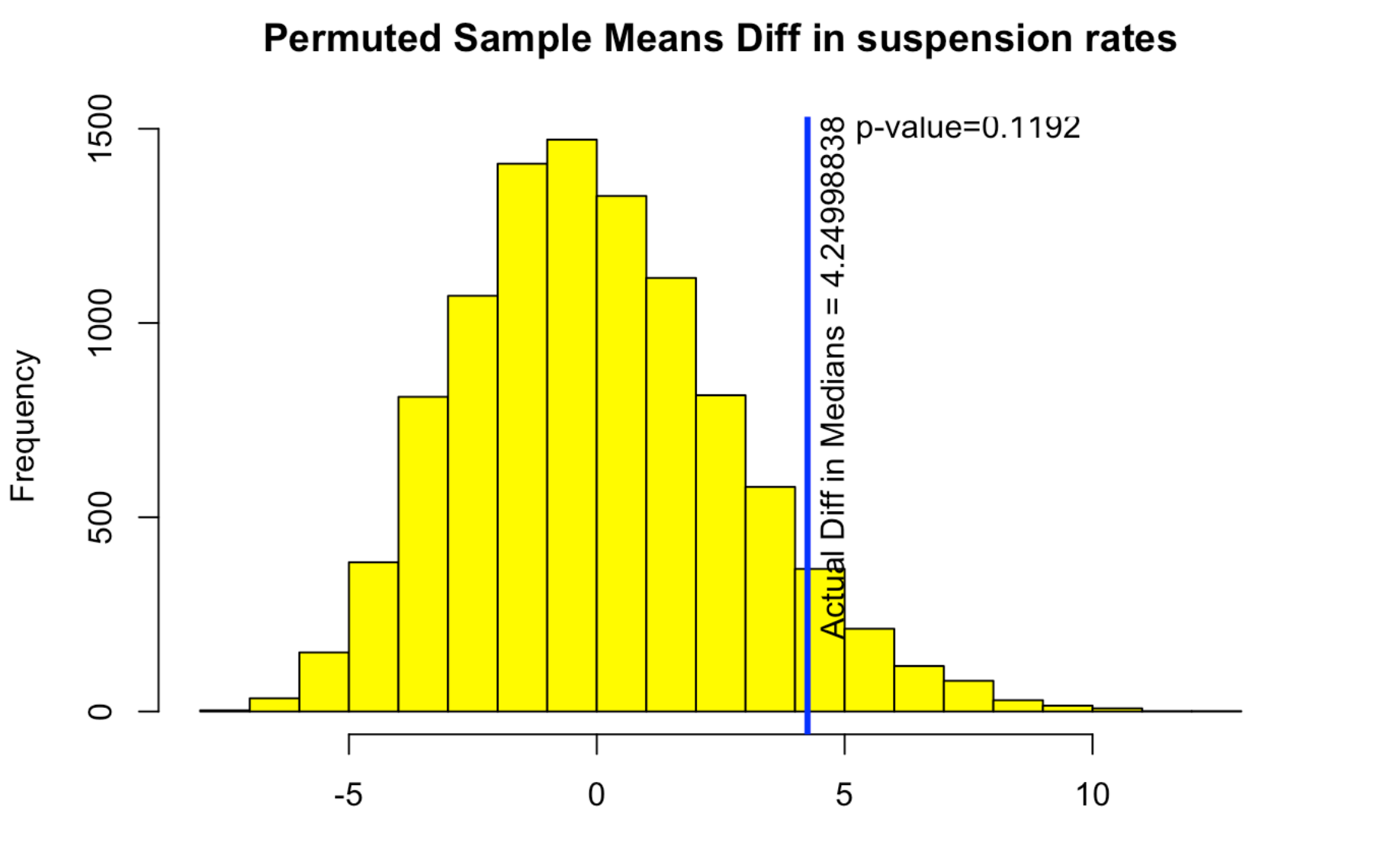
```
attach(full_data_hmm)
y <- susp_2013_14_
x <- Category
detach(full_data_hmm)

lp <- c(9.5,800,12,1500)
tp <- c(.5,1000,7.5,1500,.5,1000,7.5,1500)

print(tapply(y, x, summary))
```

```
## $Charter
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    5.60    7.10   11.30   15.31   14.80   42.70       1
##
## $CTE
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.40   10.42   14.85   16.71   21.20   32.30
##
## $End_Inc
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00    9.05   13.10   15.10   20.15   27.20
##
## $Public
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.60    4.25    7.60   11.06   15.55   59.90      38
##
## $Reg_Serv
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.70    9.90   12.30   13.85   18.22   29.00       3
##
## $Regional
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.300   2.775   4.100   5.173   7.725  12.000       3
```

```
bootperm(x,y,"Public","Charter",lp,tp,"suspension")
```

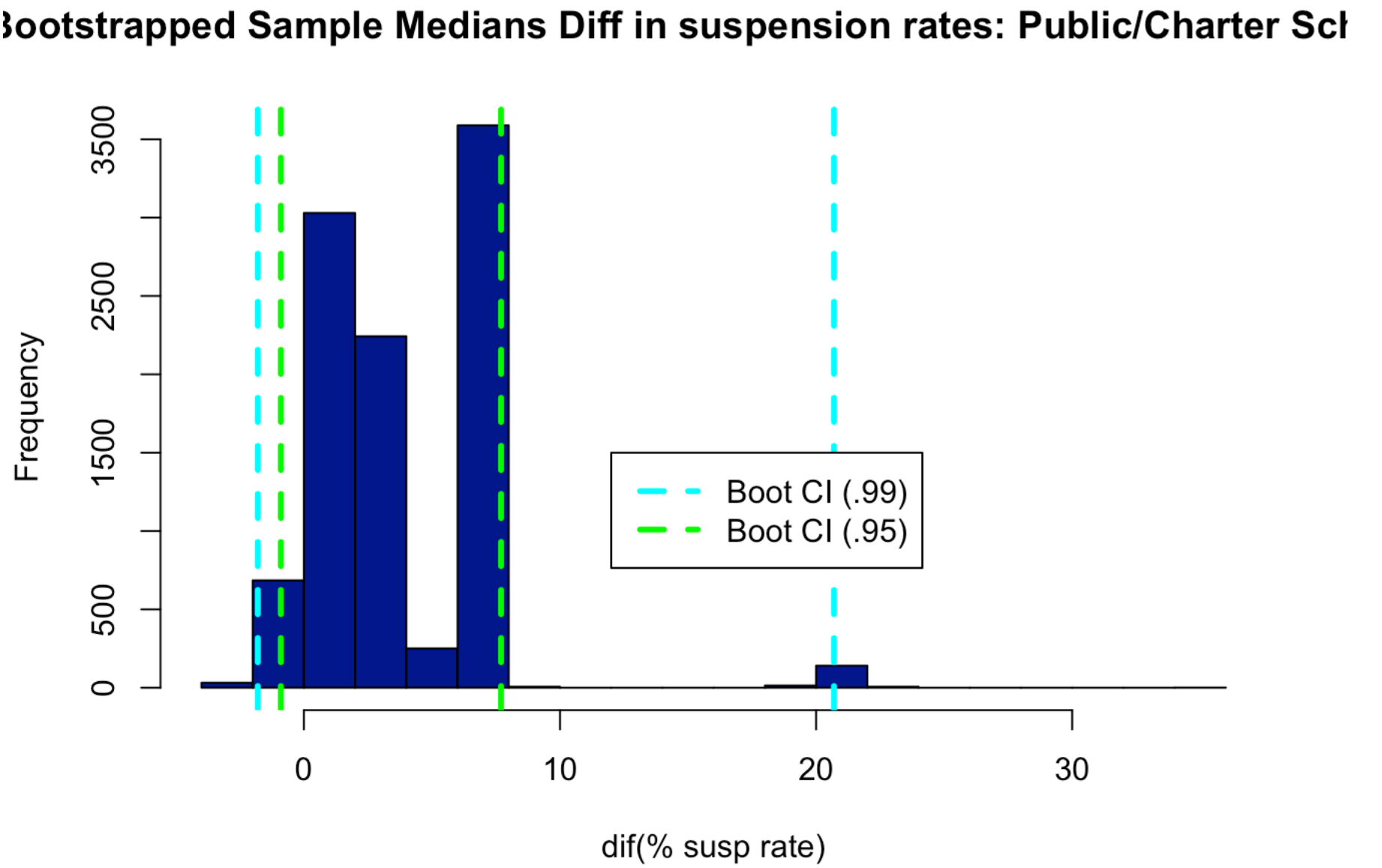
```
## [1] "observed mean: 4.24998838019986"
## [1] "bootstrapped confint on mean diff: -1.139249941901  10.7460771554729"
## [1] "bootstrapped confint on mean diff: -2.29159690913316  13.2101279339995"
## [1] "t-test"
##
##  Welch Two Sample t-test
##
## data:  g1314 by cattemp
## t = 1.3311, df = 12.681, p-value = 0.2066
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.665178 11.165155
## sample estimates:
## mean in group Charter  mean in group Public
##              15.30769              11.05770
##
## [1] "t-test conf.int"
## [1] -2.665178 11.165155
## attr(,"conf.level")
## [1] 0.95
```

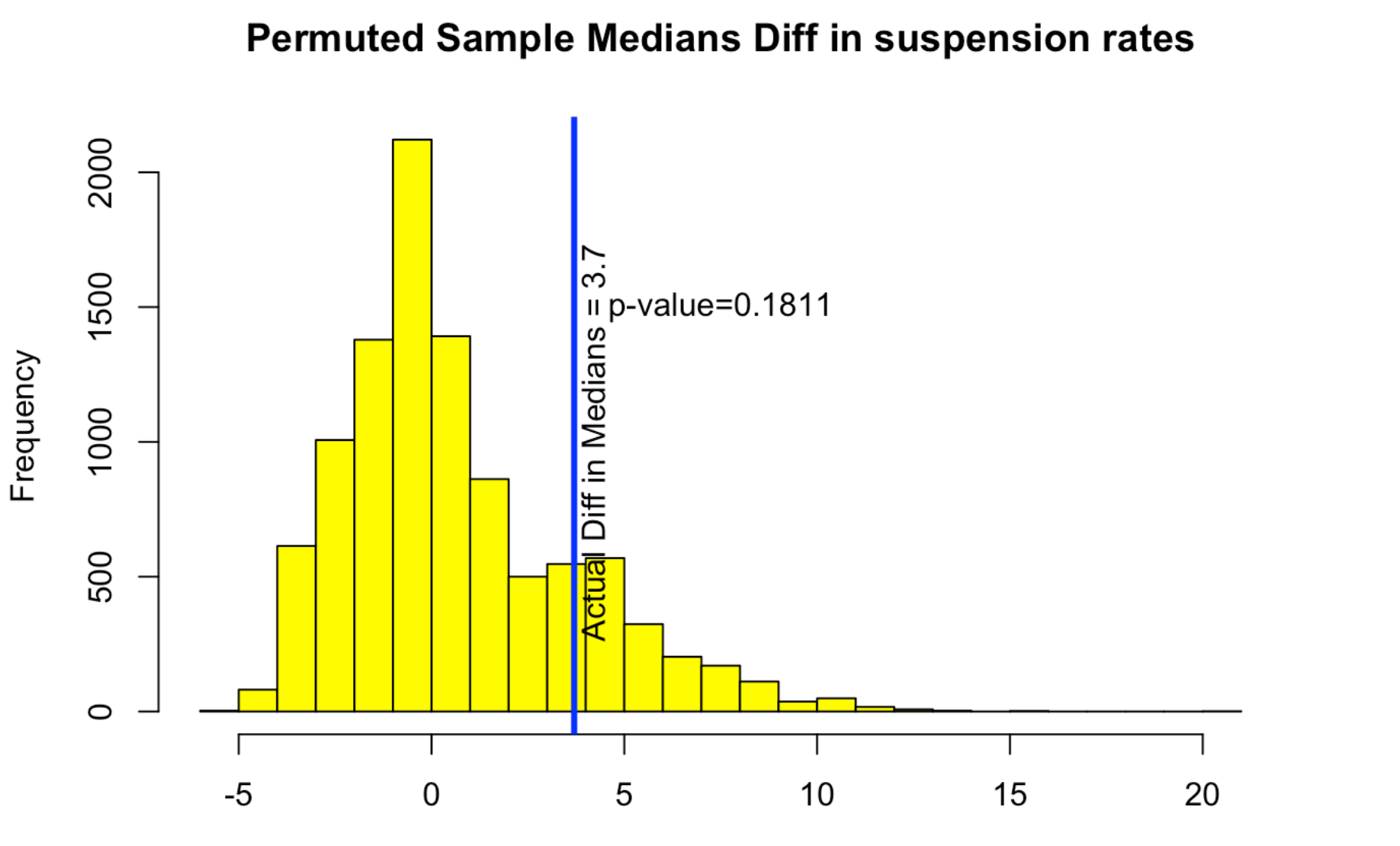

Bootstrapped Sample Means Diff in suspension rates: Public/Charter Sch
Frequency
0
200
400
600
800
1200
-5
0
5
10
15
20
dif(% susp rate)
Boot CI (.99)
Boot CI (.95)
t-test CI


Permuted Sample Means Diff in suspension rates
Frequency
0
500
1000
1500
-5
0
5
10
p-value=0.1192
Actual Diff in Medians = 4.24998838

dif(% susp rate)

```
## [1] "p-value=0.1192"
## [1] "observed median: 3.7"
## [1] "bootstrapped confint on median diff (.95): -0.9  7.7"
## [1] "bootstrapped confint on median diff (.99): -1.8  20.7"
```


Bootstrapped Sample Medians Diff in suspension rates: Public/Charter Scl
Frequency
0
500
1500
2500
3500
Boot CI (.99)
Boot CI (.95)
0
10
20
30
dif(% susp rate)


Permuted Sample Medians Diff in suspension rates
Frequency
0
500
1000
1500
2000
Actual Diff in Medians = 3.7
p-value=0.1811
-5
0
5
10
15
20

dif(% susp rate)

```
## [1] "p-value=0.1811"
```

Once again, sample size seems to be an issue and the source of distribution oddities (in particular the bootstrapped medians). The differences are not statistically significant under the bootstrap or permutation tests; given a larger sample, such differences would likely be significant.

Graduation rates

Overview

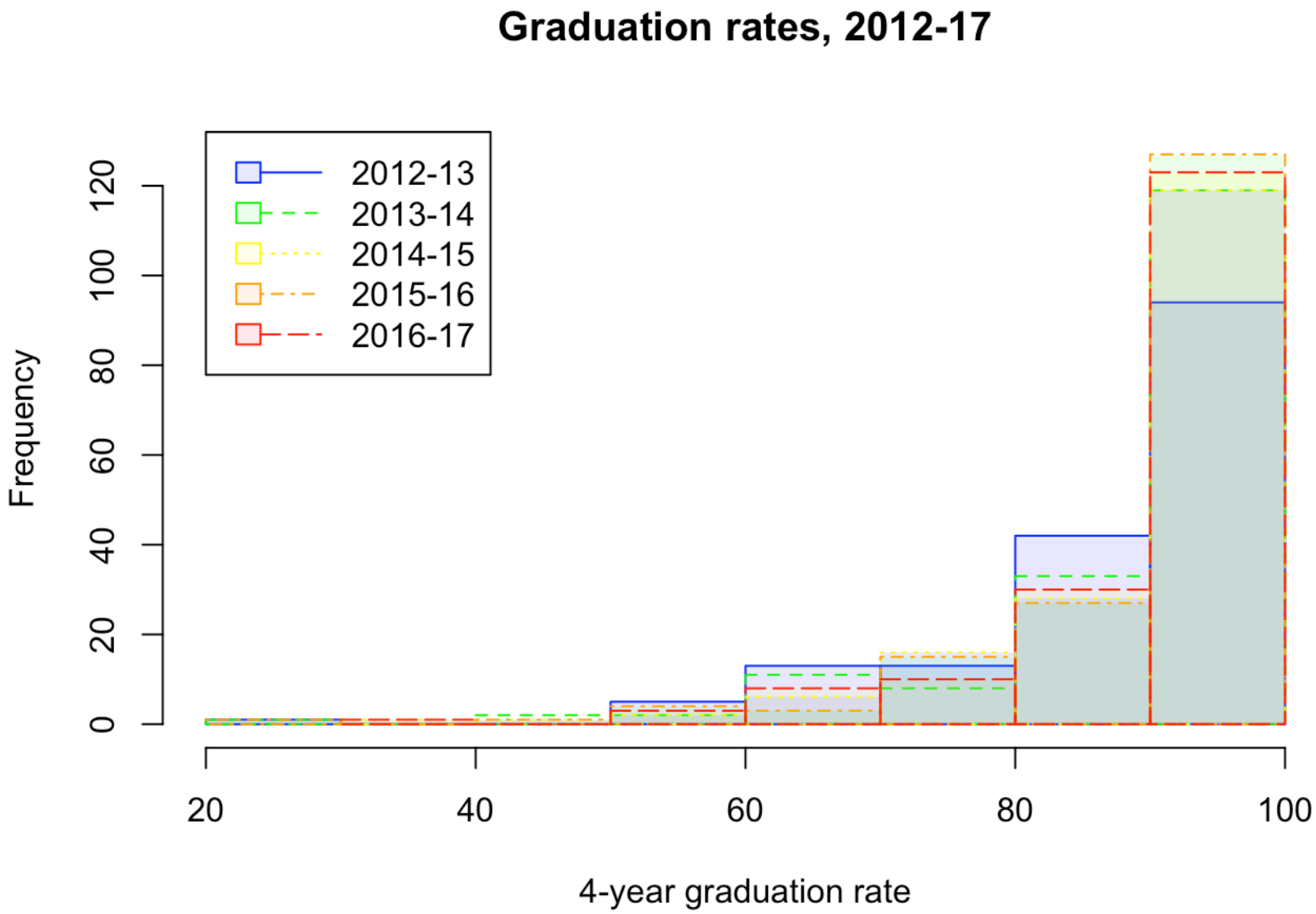
I was looking for a large number of high schools in the overall sample, so that I could analyze graduation rates. Although high schools are not as represented as elementary schools in the sample, there are still 170+ of them, which is plenty (mind that some schools in the “mixed” category also include the high school grade ranges).

Let’s start by examining graduation rates at the school level across all of CT for the past 5 years:

```
attach(full_data)
hist(grad_2012_13,border="blue",col=alpha("blue", aval),xlab="4-year graduation rate"
,main="Graduation rates, 2012-17",ylim=c(0,132))
for(i in 2:5){
    print(paste0("grad_",ys[i],ys2[i]))
    max_vals[i-1] <- max(hist(full_data[,paste0("grad_",ys[i],"_",ys2[i])],border=col
s[i],col=alpha(cols[i-2], aval),add=T,lty=i)$counts)
}
```

```
## [1] "grad_201314"
## [1] "grad_201415"
## [1] "grad_201516"
## [1] "grad_201617"
```

```
legend(20,132,yst,fill=alpha(cols, aval),border=cols,col=cols,lty=1:5)
```




```
detach(full_data)
```

The blue bin on the right (2012-13) seems to be substantially lower than those for later years. Have graduation rates across CT changed over time, at the school level?

```
## [1] "observed mean: 3.28101265822784"
```

```
## [1] "bootstrapped confint: 0.774556962025317  5.87025316455696"
```

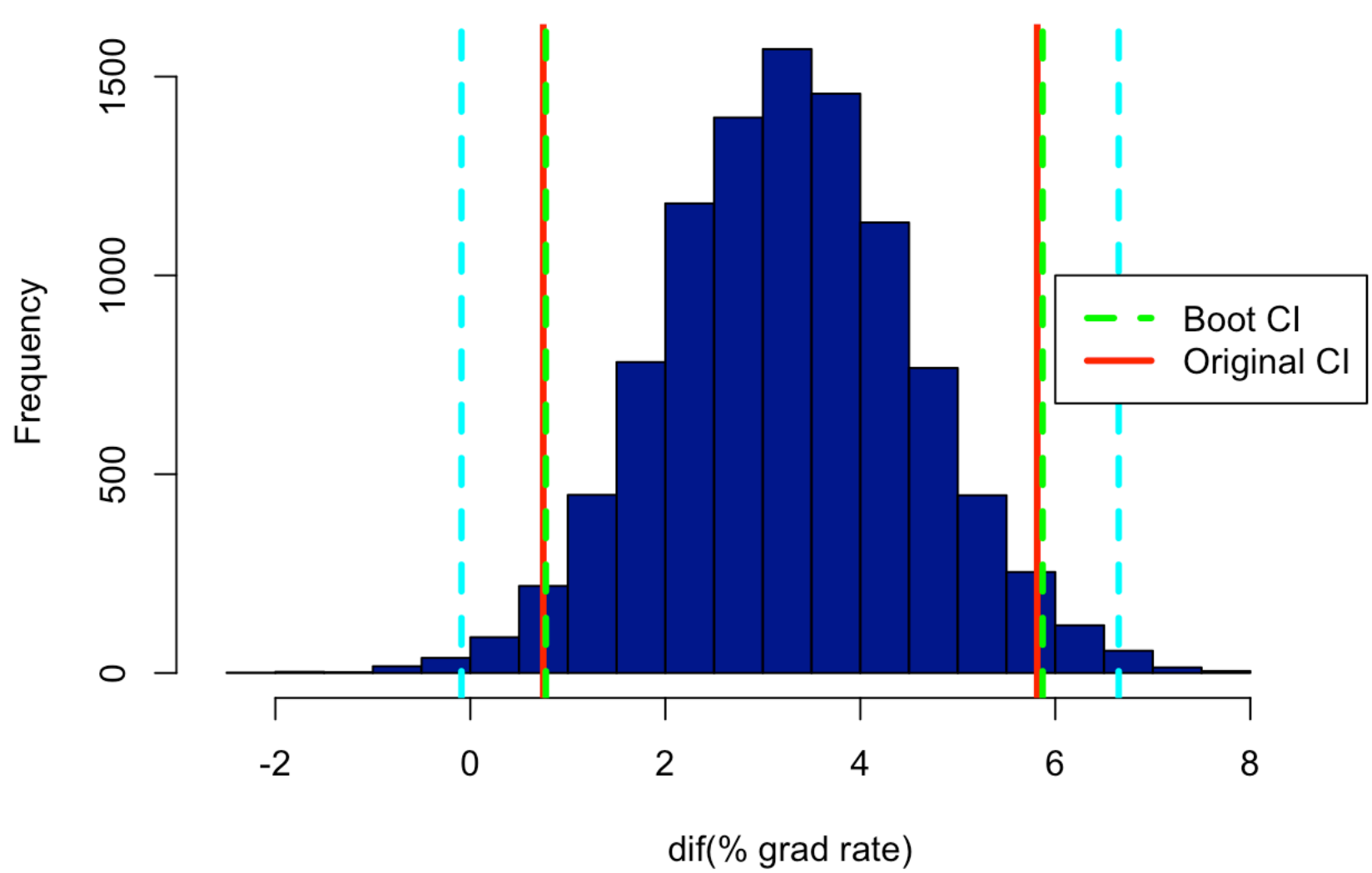
```
## [1] "bootstrapped confint: -0.091202531645563  6.65000316455695"
```

```
## [1] "t-test"
```

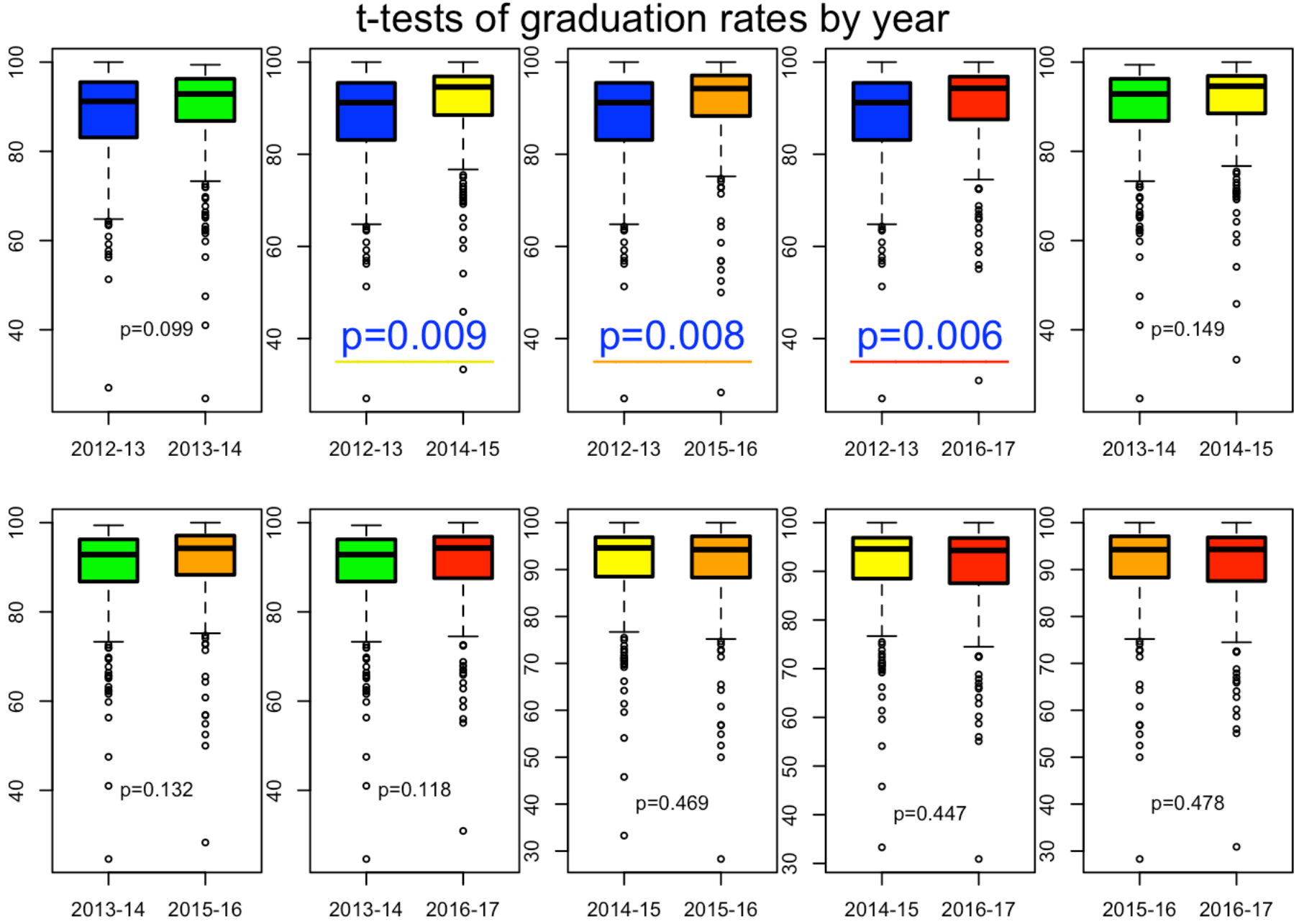
```
##
##  Welch Two Sample t-test
##
## data:  g1617 and g1213
## t = 2.5485, df = 311.27, p-value = 0.0113
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7478175 5.8142078
## sample estimates:
## mean of x mean of y
##  90.21266  86.93165
```

```
## [1] "t-test conf.int"
```

```
## [1] 0.7478175 5.8142078
## attr(,"conf.level")
## [1] 0.95
```



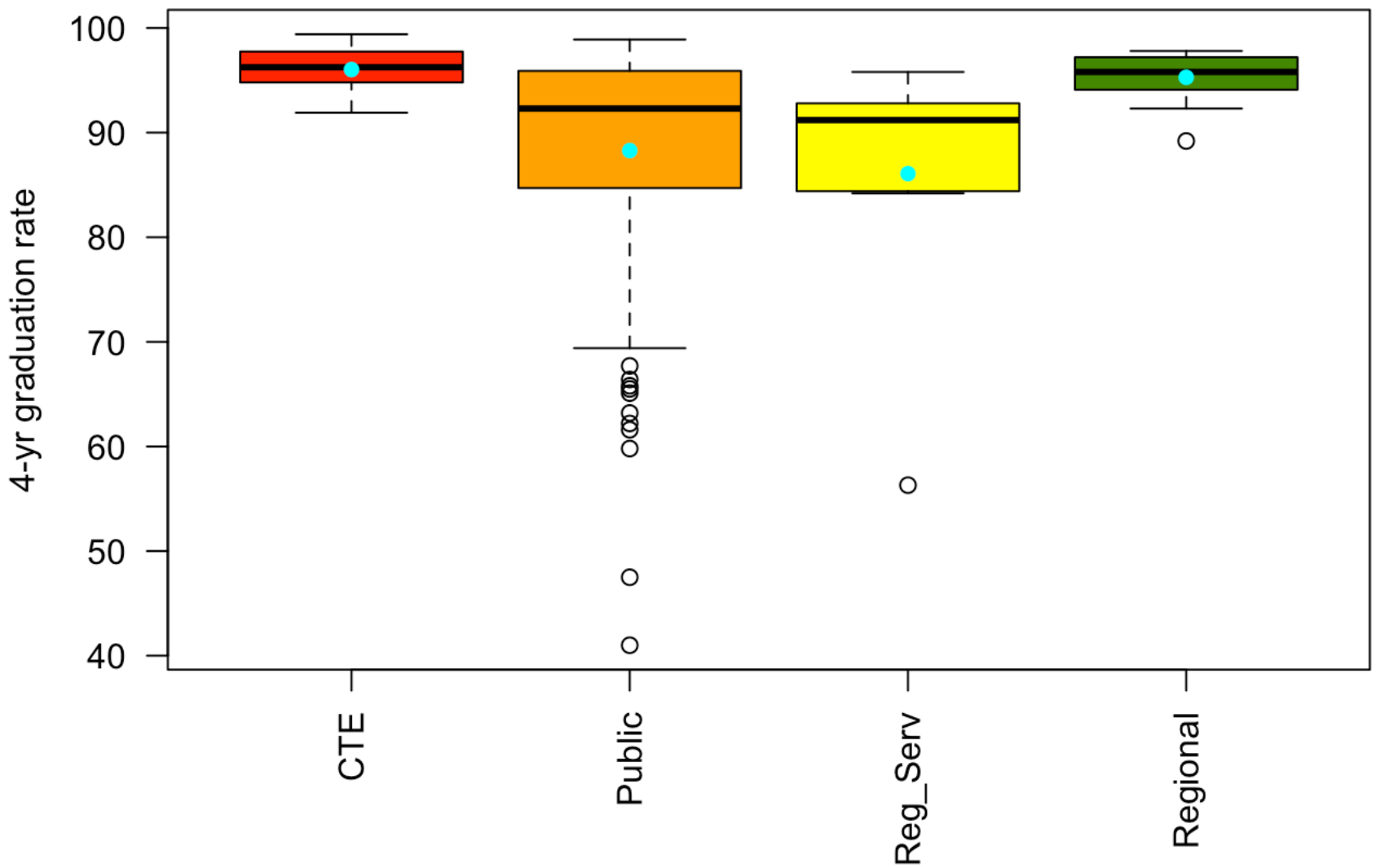
At a confidence level of 95% (and barely at), we can reject the null hypothesis that graduation rates across CT schools were equal last year to those from five years ago. But when did this change occur–gradually, or in a leap early on? Let's try another visualization:



There is evidence to assert a difference between the 2012-13 term and the years from 2014 forward, though not for the other comparisons. Thus it appears that most of the graduation rate increase between five years ago and last year happened between five and three years ago; there has been little change since then.

Let’s look at graduation rates between school catagories:

```
## [1] 17  0
```




```
## [1] "_____________________________"
```

```
## [1] "means"
```

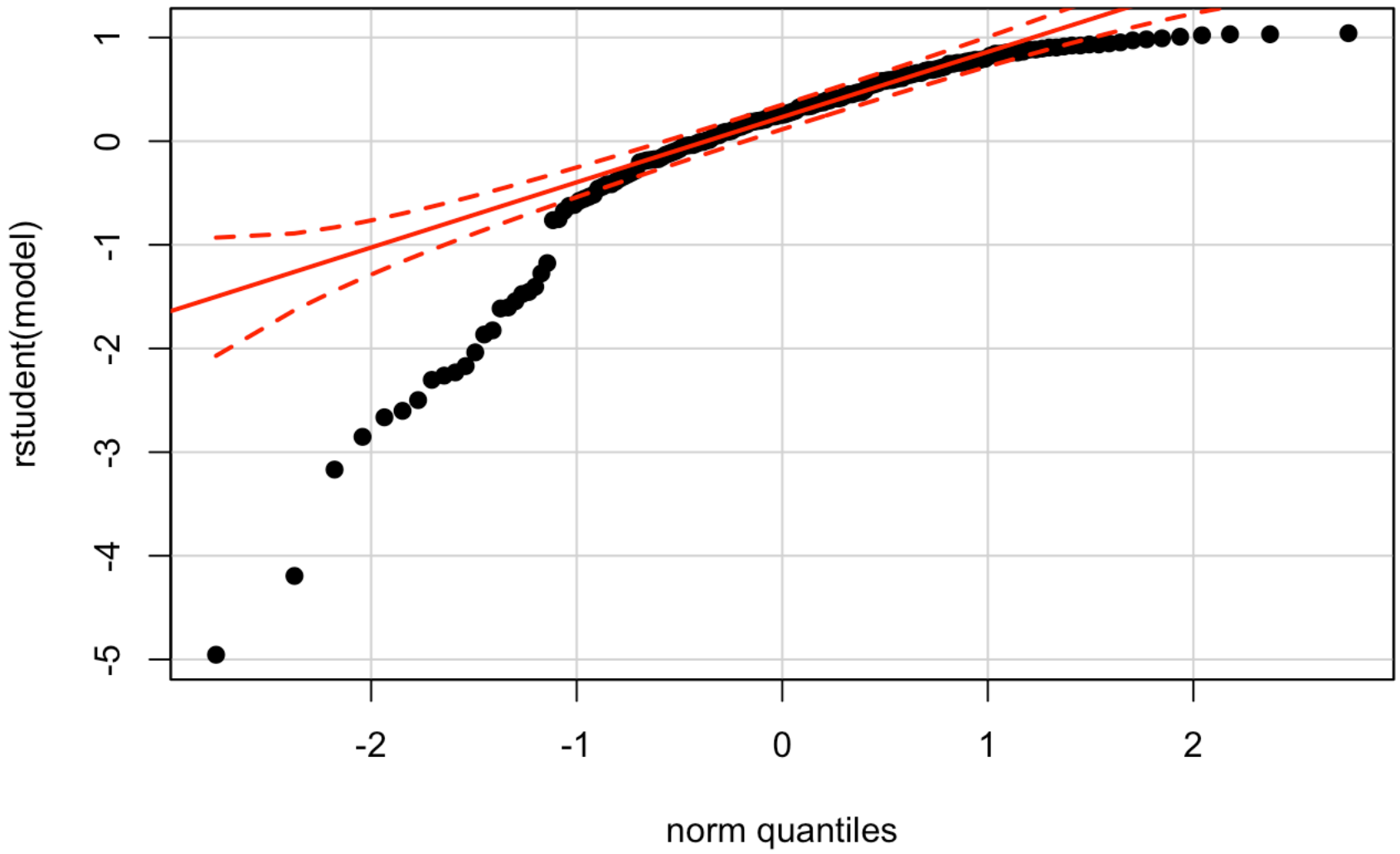
```
##      CTE   Public Reg_Serv Regional
## 96.03125 88.28321 86.07778 95.28571
```

```
## [1] "sds"
```

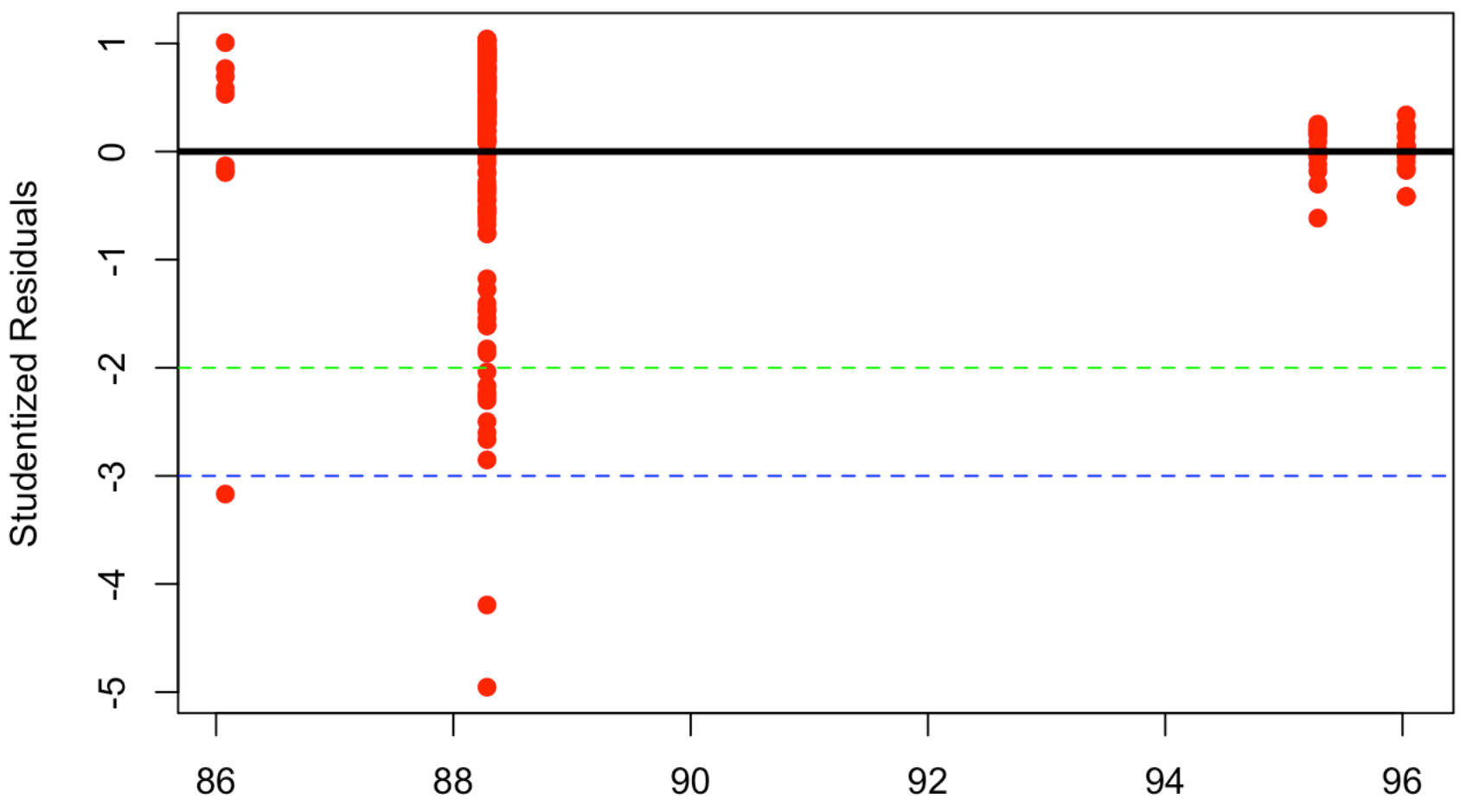
```
##       CTE    Public  Reg_Serv  Regional
##  2.161703 11.129069 11.975681  2.419109
```

```
myResPlots2(aov(y~x))
```

NQ Plot of Studentized Residuals, Residual Plots



Fits vs. Studentized Residuals, Residual Plots



Fitted Values

That is not going to work–there is too much deviation from normality, and many extreme residuals. Let's try a Box-Cox transformation:

```
bc <- boxCox(lm(y~x -1), lambda = bc_lambda, interp=T, plotit=F)
print(paste0("lambda: ",(lambda <- bc$x[which.max(bc$y)])))
```

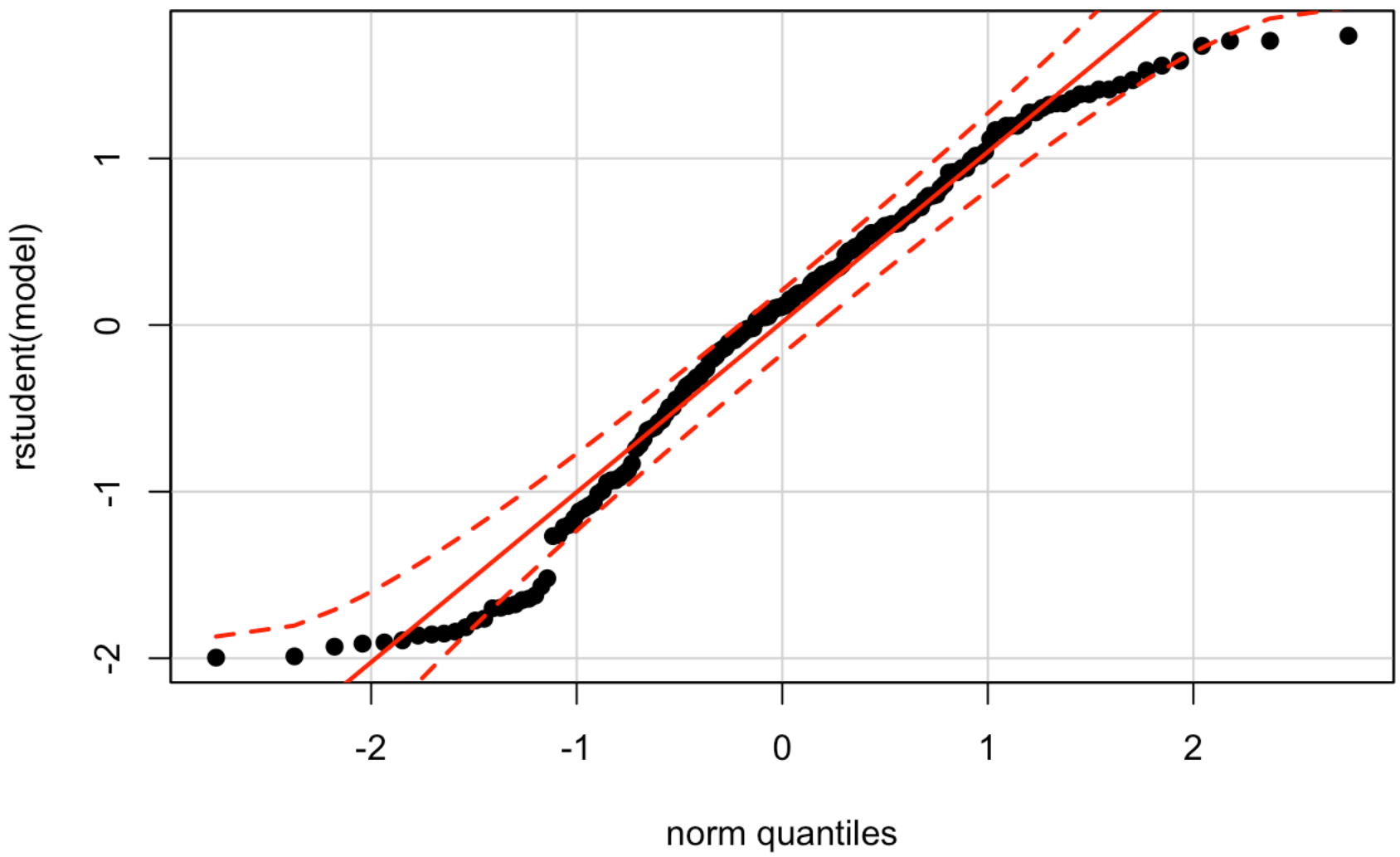
```
## [1] "lambda: 7.97979797979798"
```

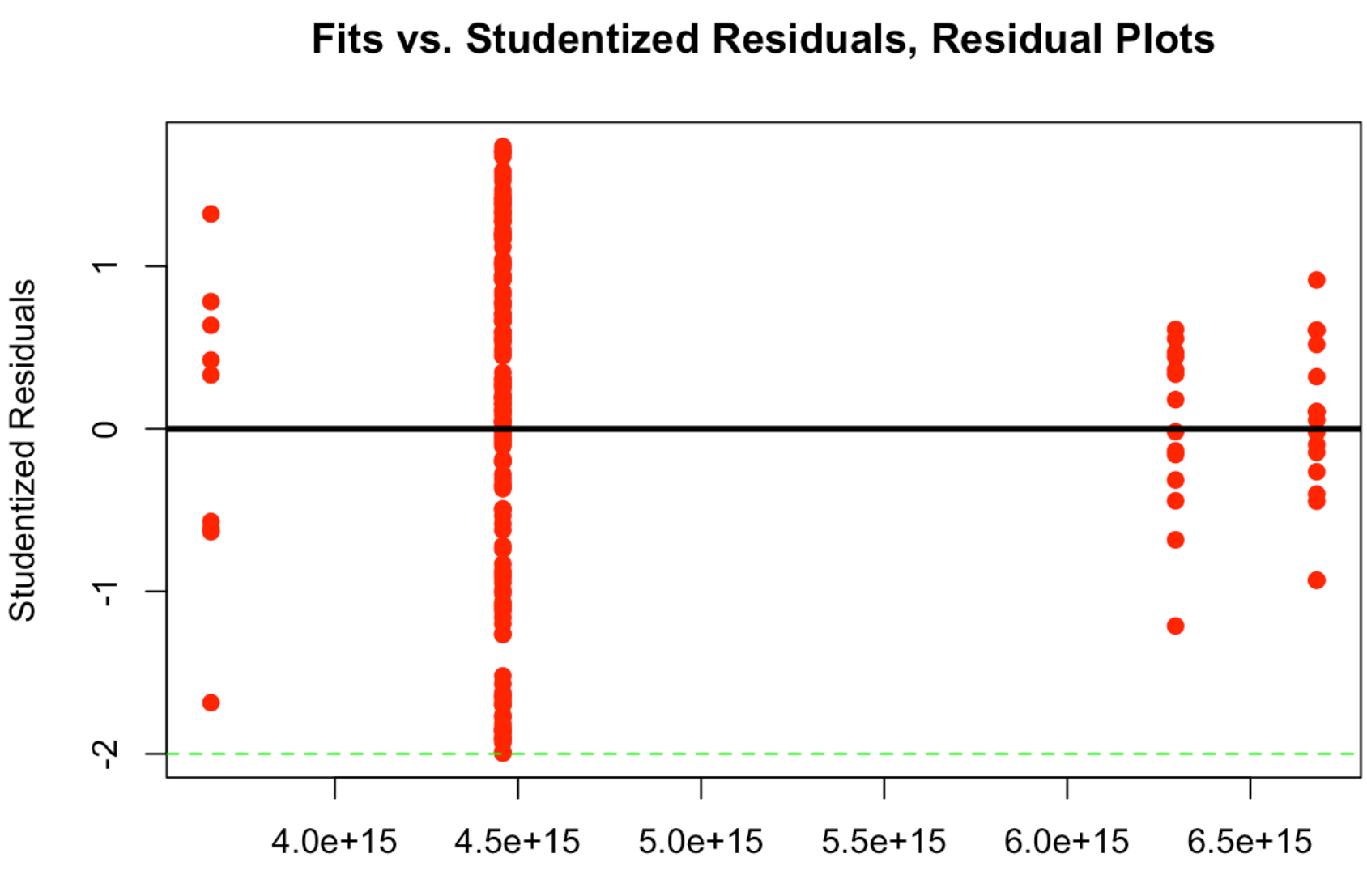
Wow.

```
print(paste0("- - - summary: ^",round(lambda,1)," - - -"))
```

```
## [1] "- - - summary: ^8 - - -"
```

```
y <- full_data_noncharter$grad_2013_14^lambda
aov1 <- aov(y~x)
myResPlots2(aov1)
```


NQ Plot of Studentized Residuals, Residual Plots
rstudent(model)
1
0
-1
-2
-2
-1
0
1
2
norm quantiles


Fits vs. Studentized Residuals, Residual Plots
Studentized Residuals
1
0
-1
-2
4.0e+15
4.5e+15
5.0e+15
5.5e+15
6.0e+15
6.5e+15

```
print(Anova(aov1,type=3))
```

```
## Anova Table (Type III tests)
##
## Response: y
##                 Sum Sq  Df  F value    Pr(>F)
## (Intercept) 7.1421e+32   1 140.0332 < 2.2e-16 ***
## x           1.1481e+32   3   7.5032 9.698e-05 ***
## Residuals   8.4665e+32 166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("-- -- -- -- -- -- -- -- --")
```

```
## [1] "-- -- -- -- -- -- -- -- --"
```

```
(welch_test <- oneway.test(y~x))
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  y and x
## F = 17.239, num df = 3.000, denom df = 25.221, p-value = 2.681e-06
```

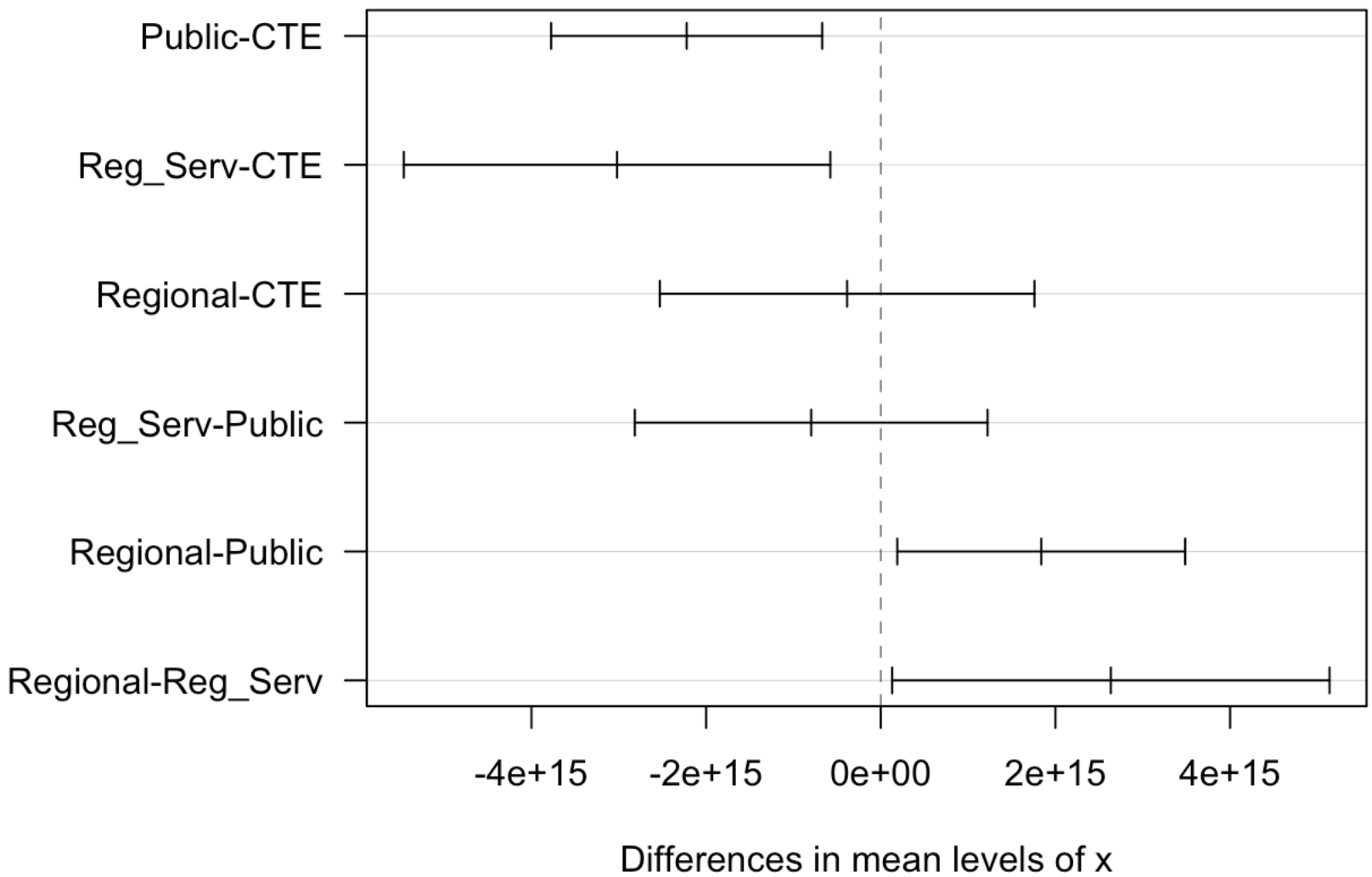
```
print("- - - /summary - - -")
```

```
## [1] "- - - /summary - - -"
```

Minding the enormous units, this model feedback looks much better–good normal conformity, and the studentized residuals are not as extreme as before.

Which groups have significant differences in graduation rates?

```
tuk <- TukeyHSD(aov1)
par(mar=c(5,10,4,2))
plot(tuk,las=1)
```



Public schools are reported to be distinct from career and technical schools and regional schools (they have statistically distinguishable graduation rates); regional schools are also distinct from CTE schools. The regional-public relation approaches somewhat close to the 0 line–let's do a more thorough analysis for this case by bootstrapping the mean difference and permuting the samples:

Contrary to my expectations, the results from before are confirmed (though marginally at alpha=.05) from the second permutation test): for the means, the bootstrapped confidence interval and t-test interval align rather closely, and both are well-contained within the negative range; likewise for medians, we have evidence to declare different graduation rates for regional and public schools.

Predicting graduation rates

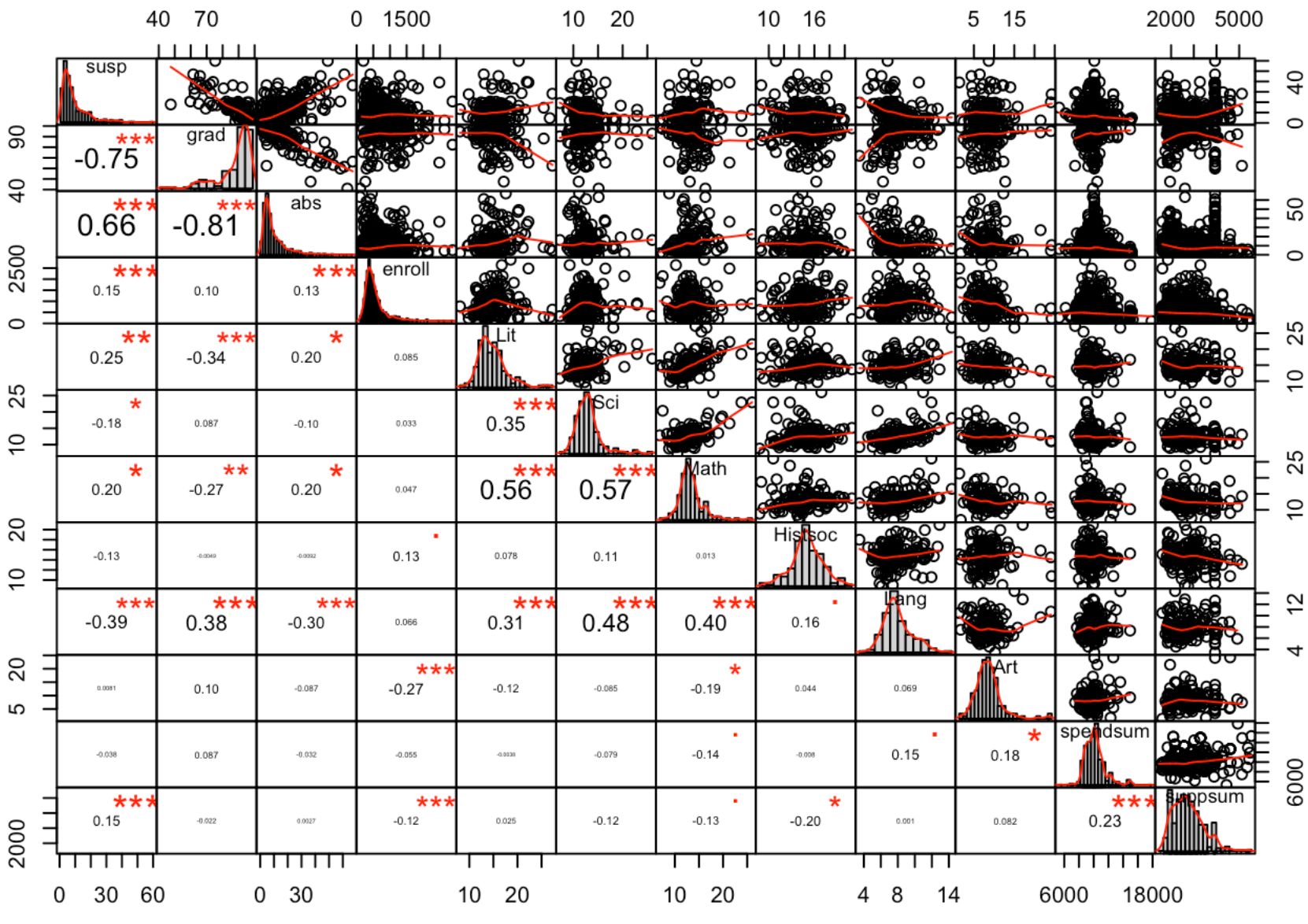
We could do more comparisons between school categories on individual characteristics and then compare to see how they related to school graduation rates, but this would be a long and tedious procedure, and not the most informative. Plus, there is the issue of low subsample sizes for certain school types, so the comparisons will be incomplete. Let's try to fit models that predict graduation rates. We will work with a few key variables–enrollment, coursetaking in grades 11/12, suspension rate, chronic absenteeism, and spending on instructional and support services. Recall the caveats associated with these variables: * Spending is an average at the district level, so it is not expected to be a very accurate representation of individual school conditions; * Course taking is averaged over grade levels (not too much of a problem) and taken at the school level (the bigger problem). A better way to assess the effect of course taking on graduation would be to have

data at the student level and then conduct some kind of logistic regression to determine the increase in probability of graduation given an increase in the number of courses taken in particular subejcts or the number of years of particular subjects taken. This information is not presently availalble, so we settle with what we have–though it might not be very revealing. * Absenteeism is provided across the whole school, not across individual grades, and (most importantly) not at the student level. This, again, would be preferable, but is unavailable.

Let us proceed:

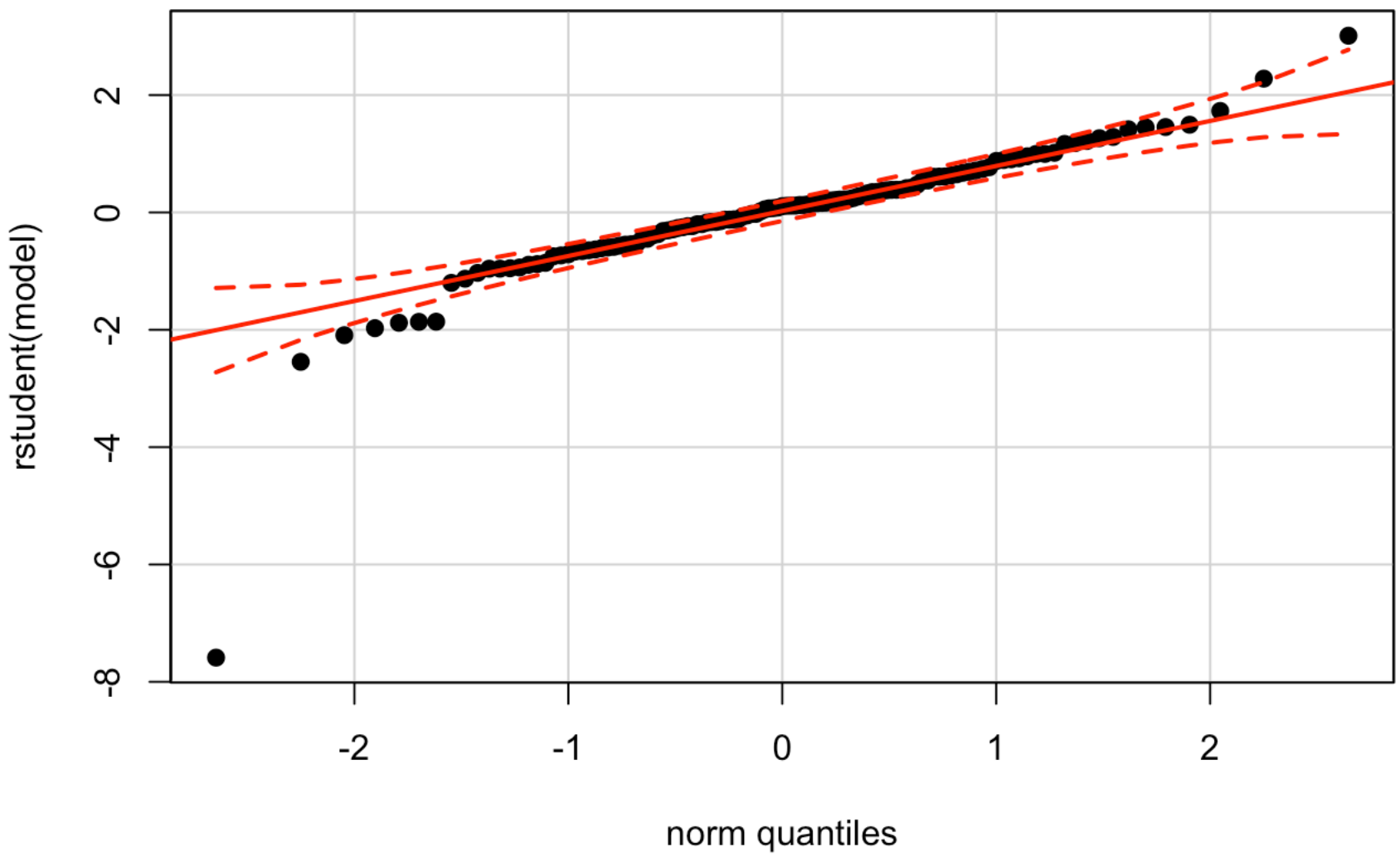
```
grad_predict(full_data,"2013_14","r",opt="half")
```

```
## [1] "_____________________________________________________________________"
## [1] "_____________________________________________________________________"
## [1] "_____________________Summary, mode=r_____________________"
## [1] "_____________________________________________________________________"
## [1] "_____________________________________________________________________"
```

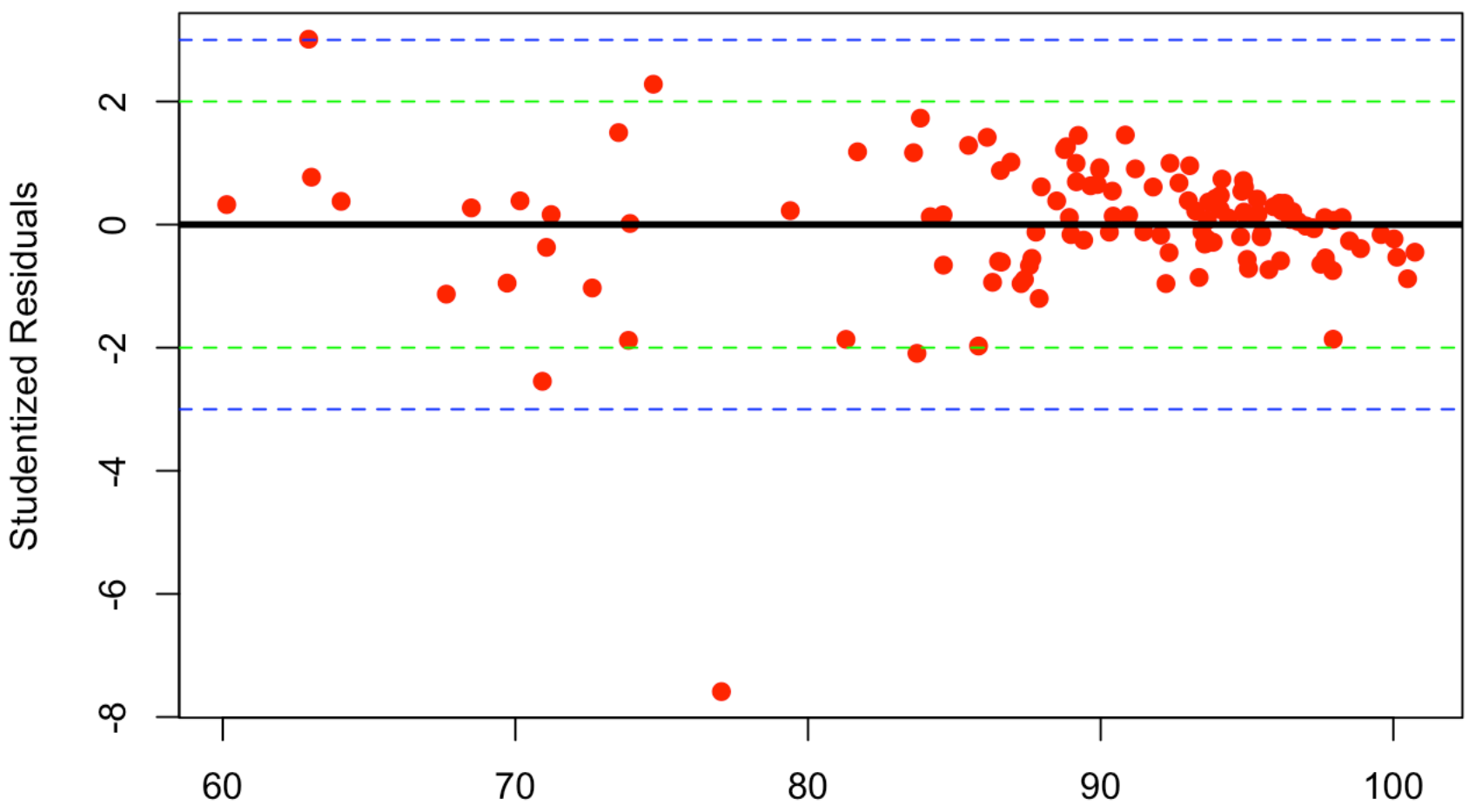



```
## [1] "- - - - Summary for ^1 - - - -"
```

NQ Plot of Studentized Residuals, Residual Plots



Fits vs. Studentized Residuals, Residual Plots

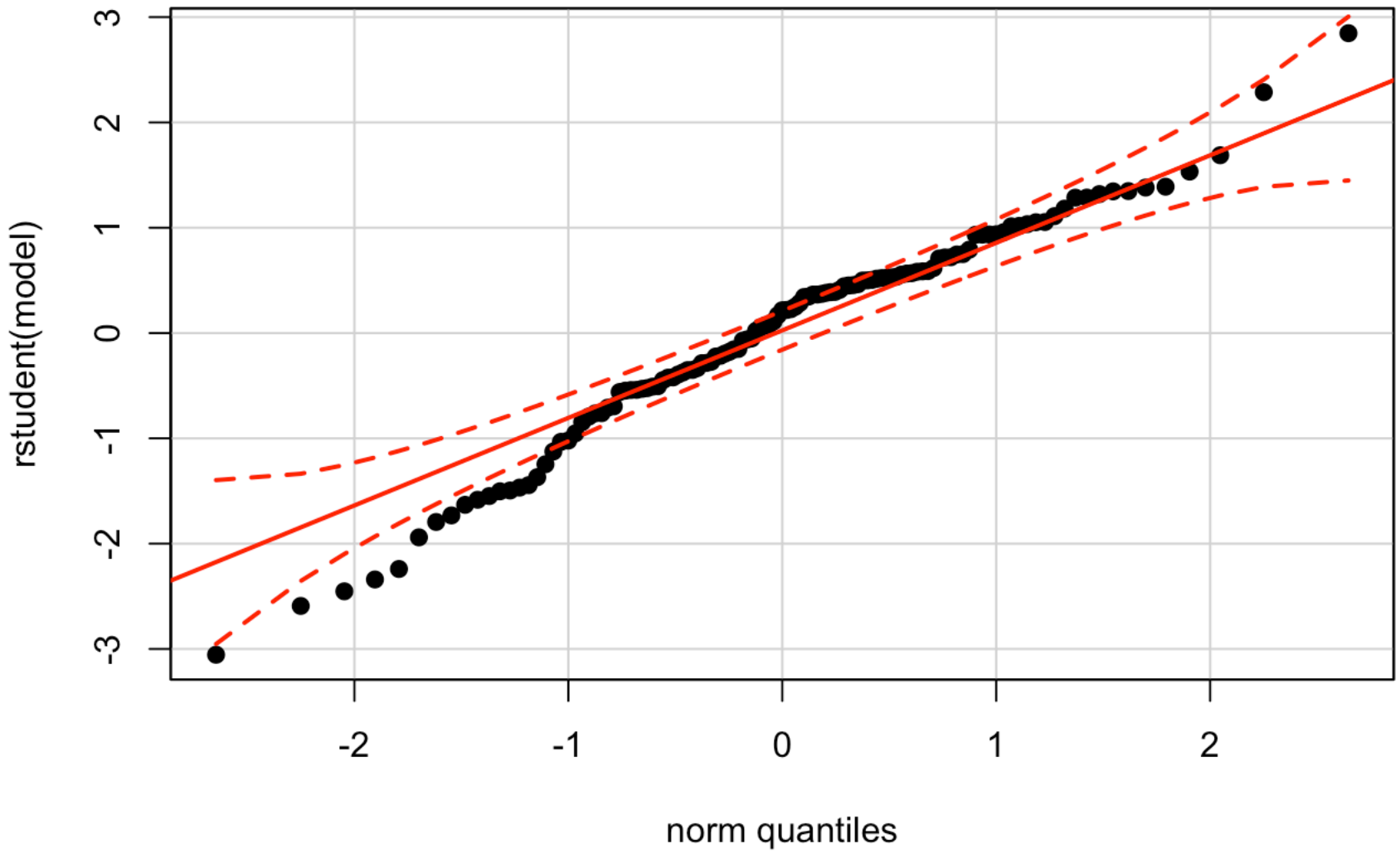


Fitted Values

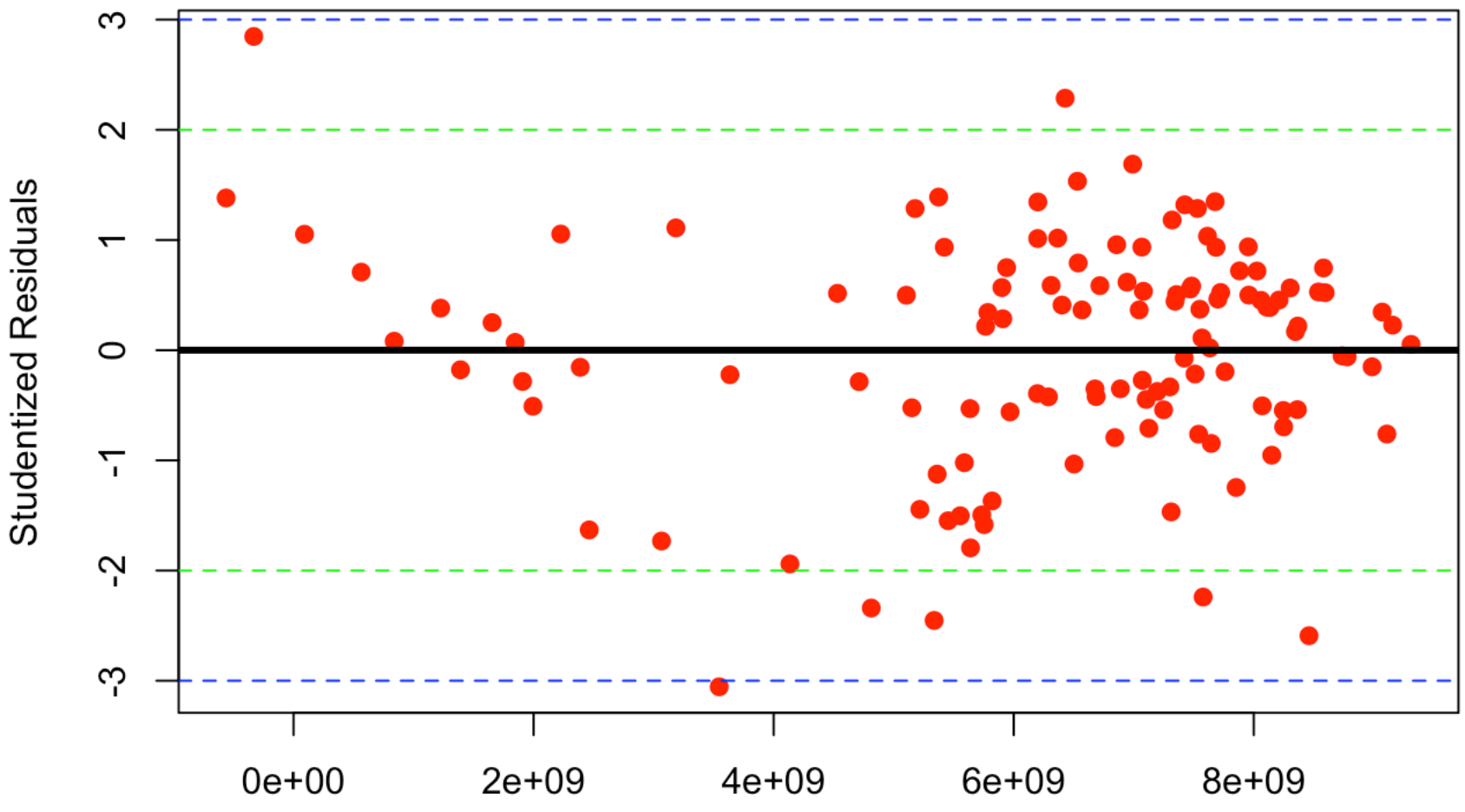
```
## Anova Table (Type III tests)
##
## Response: grad
##              Sum Sq  Df  F value    Pr(>F)
## (Intercept) 4781.6   1 187.8445 < 2.2e-16 ***
## susp         392.3   1  15.4123 0.0001503 ***
## abs         1510.0   1  59.3188 6.036e-12 ***
## enroll         3.8   1   0.1495 0.6997611
## Lit            2.4   1   0.0948 0.7587277
## Sci            2.2   1   0.0856 0.7704285
## Math          73.2   1   2.8764 0.0926899 .
## Histsoc       21.9   1   0.8591 0.3559939
## Lang         144.1   1   5.6594 0.0190695 *
## Art            4.3   1   0.1677 0.6829837
## spendsum       0.6   1   0.0249 0.8749904
## suppsum       54.3   1   2.1313 0.1471391
## Residuals   2825.5 111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = grad ~ ., data = grad_susp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.5406  -2.3785   0.5361   2.6812  12.3645
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.685e+01  7.067e+00  13.706  < 2e-16 ***
## susp        -2.904e-01  7.397e-02  -3.926  0.00015 ***
## abs         -4.080e-01  5.297e-02  -7.702 6.04e-12 ***
## enroll      -3.826e-04  9.895e-04  -0.387  0.69976
## Lit          7.619e-02  2.474e-01   0.308  0.75873
## Sci          7.891e-02  2.697e-01   0.293  0.77043
## Math        -5.129e-01  3.024e-01  -1.696  0.09269 .
## Histsoc     -2.163e-01  2.334e-01  -0.927  0.35599
## Lang         9.313e-01  3.915e-01   2.379  0.01907 *
## Art          7.743e-02  1.891e-01   0.409  0.68298
## spendsum    -6.392e-05  4.054e-04  -0.158  0.87499
## suppsum      1.177e-03  8.061e-04   1.460  0.14714
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.045 on 111 degrees of freedom
##   (761 observations deleted due to missingness)
## Multiple R-squared:  0.7827, Adjusted R-squared:  0.7611
```

```
## F-statistic: 36.34 on 11 and 111 DF,  p-value: < 2.2e-16
##
## [1] "- - - - Summary for ^5 - - - -"
```

NQ Plot of Studentized Residuals, Residual Plots



Fits vs. Studentized Residuals, Residual Plots



Fitted Values

```
## Anova Table (Type III tests)
##
## Response: (grad^lambda)
##                 Sum Sq  Df F value    Pr(>F)
## (Intercept) 2.5395e+19   1 18.1460 4.304e-05 ***
## susp        3.1375e+19   1 22.4197 6.516e-06 ***
## abs         6.0004e+19   1 42.8768 1.878e-09 ***
## enroll      3.5557e+17   1  0.2541  0.615219
## Lit         1.6162e+17   1  0.1155  0.734621
## Sci         8.7737e+17   1  0.6269  0.430170
## Math        9.2892e+18   1  6.6377  0.011297 *
## Histsoc     6.3380e+17   1  0.4529  0.502365
## Lang        1.5548e+19   1 11.1098  0.001167 **
## Art         3.9760e+17   1  0.2841  0.595085
## spendsum    1.2966e+17   1  0.0926  0.761406
## suppsum     2.9771e+18   1  2.1273  0.147516
## Residuals   1.5534e+20 111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = (grad^lambda) ~ ., data = grad_susp)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -3.305e+09 -6.014e+08  2.205e+08  6.772e+08  2.752e+09
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7058362622 1656967362   4.260 4.30e-05 ***
## susp         -82126066   17344672  -4.735 6.52e-06 ***
## abs          -81323505   12419526  -6.548 1.88e-09 ***
## enroll         -116944     232004  -0.504  0.61522
## Lit           19716885   58018544   0.340  0.73462
## Sci           50079060   63247608   0.792  0.43017
## Math        -182685130   70907753  -2.576  0.01130 *
## Histsoc      -36823966   54718614  -0.673  0.50237
## Lang         305952634   91791138   3.333  0.00117 **
## Art           23633461   44338782   0.533  0.59509
## spendsum         28929      95043   0.304  0.76141
## suppsum         275664     189000   1.459  0.14752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.183e+09 on 111 degrees of freedom
##   (761 observations deleted due to missingness)
## Multiple R-squared:  0.7947, Adjusted R-squared:  0.7743
## F-statistic: 39.06 on 11 and 111 DF,  p-value: < 2.2e-16
```

Model assumptions appear to be met, though there does appear to be evidence of heteroskedasticity (mostly removed by the Box-Cox trasnformation). Let's take the results one at a time.

Suspension rate, rate of chronic absenteeism, and language course taking emerged as significant predictors of the graduation rate; mathematics course taking is marginally significant ($p<0.1$). The coefficients on suspension and absenteeism are negative, which makes sense: increases in absenteeism and suspension are expected to correspond with decreases in the graduation rate (associatively, if not causally). A 1% increase in suspension rate (absenteeism rate) is predicted to correspond to a 0.29% (.41%) decrease in graduation rate.

The coefficients on course taking are harder to interpret—not mathematically, but in terms of what practical implications the may have, if any. Does taking more language courses lead to a higher chance of graduation, but not taking more courses in ELA or Math? We may conclude nothing of the sort from this analysis: we can only say that in this sample, language course taking was a significant predictor of graduation rate. This may well be because language course taking is related to something else which has causal effect on graduation, or to something of which language course taking and graduation are concomitant effects.

Of course, it is also possible that the analysis here captures some part of a causal effect of course taking on graduation rate—but this effect is so laden with potential confounders that a solid conclusion is not attainable. And it is also strange that only language course taking arises as a significant predictor, and none of the (other) four core subjects (if language is considered a core subject in these schools). This may be an issue with multicollinearity in the data.

The results for the model with graduation rate adjusted by the Box-Cox procedure (lambda ~ 5) is much the same, save that the p-values diminish, and so math course taking becomes a significant ($p=.011$).

Using the regsubsets function, every possible combination of variables is now fit as a model.

```
modfits(grad_susp)
```

```
## [1] "r2 results"
##  [1] "grad"     "abs"      "enroll"   "Lit"      "Sci"      "Math"
##  [7] "Histsoc"  "Lang"     "Art"      "spendsum" "suppsum"
## [1] "11 predictors in model"
##
## Call:
## lm(formula = grad ~ ., data = rss_r2_fit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.5406  -2.3785   0.5361   2.6812  12.3645
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.685e+01  7.067e+00  13.706  < 2e-16 ***
## susp        -2.904e-01  7.397e-02  -3.926  0.00015 ***
## abs         -4.080e-01  5.297e-02  -7.702 6.04e-12 ***
## enroll      -3.826e-04  9.895e-04  -0.387  0.69976
## Lit          7.619e-02  2.474e-01   0.308  0.75873
## Sci          7.891e-02  2.697e-01   0.293  0.77043
```

```
## Math        -5.129e-01  3.024e-01  -1.696  0.09269 .
## Histsoc     -2.163e-01  2.334e-01  -0.927  0.35599
## Lang         9.313e-01  3.915e-01   2.379  0.01907 *
## Art          7.743e-02  1.891e-01   0.409  0.68298
## spendsum    -6.392e-05  4.054e-04  -0.158  0.87499
## suppsum      1.177e-03  8.061e-04   1.460  0.14714
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.045 on 111 degrees of freedom
##   (761 observations deleted due to missingness)
## Multiple R-squared:  0.7827, Adjusted R-squared:  0.7611
## F-statistic: 36.34 on 11 and 111 DF,  p-value: < 2.2e-16
##
## [1] "+ + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + +"
## [1] "adjr2 results"
## [1] "grad"    "abs"     "Math"    "Lang"    "suppsum"
## [1] "5 predictors in model"
##
## Call:
## lm(formula = grad ~ ., data = rss_adjr2_fit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.3111  -2.2611   0.1216   2.6799  12.3922
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 95.9175703  3.5730977  26.844  < 2e-16 ***
## susp        -0.2496086  0.0611673  -4.081 7.91e-05 ***
## abs         -0.4233714  0.0448093  -9.448 2.41e-16 ***
## Math        -0.5566486  0.2197914  -2.533  0.01255 *
## Lang         0.9630352  0.3342879   2.881  0.00466 **
## suppsum      0.0010773  0.0006839   1.575  0.11768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.987 on 126 degrees of freedom
##   (752 observations deleted due to missingness)
## Multiple R-squared:  0.7642, Adjusted R-squared:  0.7548
## F-statistic: 81.66 on 5 and 126 DF,  p-value: < 2.2e-16
##
## [1] "+ + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + +"
## [1] "bic results"
## [1] "optimal bic -160.03598650755"
## [1] "grad"    "abs"     "suppsum"
## [1] "3 predictors in model"
##
## Call:
## lm(formula = grad ~ ., data = rss_bic_fit)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.9292  -1.9546   0.4768   2.7536  17.2251
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 98.9098140  2.0548971  48.134  < 2e-16 ***
## susp        -0.3899358  0.0608340  -6.410 2.29e-09 ***
## abs         -0.3978467  0.0479755  -8.293 1.03e-13 ***
## suppsum      0.0005227  0.0007022   0.744    0.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.594 on 134 degrees of freedom
##   (746 observations deleted due to missingness)
## Multiple R-squared:  0.7149, Adjusted R-squared:  0.7085
## F-statistic:   112 on 3 and 134 DF,  p-value: < 2.2e-16
##
## [1] "+ + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + + +"
## [1] "cp results"
## [1] "grad"    "abs"     "suppsum"
## [1] "3 predictors in model"
##
## Call:
## lm(formula = grad ~ ., data = rss_cp_fit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.9292  -1.9546   0.4768   2.7536  17.2251
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 98.9098140  2.0548971  48.134  < 2e-16 ***
## susp        -0.3899358  0.0608340  -6.410 2.29e-09 ***
## abs         -0.3978467  0.0479755  -8.293 1.03e-13 ***
## suppsum      0.0005227  0.0007022   0.744    0.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.594 on 134 degrees of freedom
##   (746 observations deleted due to missingness)
## Multiple R-squared:  0.7149, Adjusted R-squared:  0.7085
## F-statistic:   112 on 3 and 134 DF,  p-value: < 2.2e-16
```

The r^2 criterion is not useful for evaluating model goodness of fit, as it merely includes every variable. The adjusted r^2 value isolates five predictors: the four which were discussed above (which are all significant in this model at alpha=.05), and district-level per-pupil spending on administration/support services ('suppsum', ns, beta ~ .001). The fact that the four predictive variables are now all significant under no power transformation suggests that there was some multicollinearity in the full dataset.

The BIC and CP statistics both isolate the same tripartite model: suspension rate and absenteeism rate (p<.0005), and suppsum (which is now highly insignificant at p=0.46). This is probably the most believable result here. The respective coefficients on suspension and absenteeism are now -.39 and -.40.

It is possible that some of the variables in this dataset have interaction effects with each other. Out of curiosity, let's see if the variables other than course taking have interaction effects with school size.

```
grad_predict(full_data,"2013_14","r",opt="half",A=T,pl=F)
```

```
## [1] "_____________________________________________________________________________"
## [1] "_____________________________________________________________________________"
## [1] "_____________________________Summary, mode=r_____________________________"
## [1] "_____________________________________________________________________________"
## [1] "_____________________________________________________________________________"
## [1] "- - - - Summary for ^1 - - - -"
## Analysis of Deviance Table (Type III tests)
##
## Response: grad
##                LR Chisq Df Pr(>Chisq)
## abs              34.031  1  5.423e-09 ***
## susp              0.654  1    0.41877
## enroll            0.000  1    0.99162
## Lit               0.068  1    0.79393
## Sci               0.078  1    0.78020
## Math              1.709  1    0.19115
## Histsoc           0.532  1    0.46565
## Lang              4.853  1    0.02760 *
## Art               0.146  1    0.70203
## spendsum          0.053  1    0.81815
## suppsum           0.274  1    0.60038
## susp:enroll       3.726  1    0.05359 .
## enroll:spendsum   0.010  1    0.91875
## enroll:suppsum    0.000  1    0.99966
## abs:enroll        1.880  1    0.17029
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## glm(formula = grad ~ abs + susp + enroll + Lit + Sci + Math +
##     Histsoc + Lang + Art + spendsum + suppsum + enroll * susp +
##     enroll * spendsum + enroll * suppsum + enroll * abs, data = grad_susp)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -29.0210   -2.0986    0.0771    2.6150   11.7829
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.701e+01  9.764e+00   9.935  < 2e-16 ***
```

```
## abs              -5.140e-01  8.812e-02  -5.834 5.85e-08 ***
## susp             -9.624e-02  1.190e-01  -0.809   0.4206
## enroll           -6.989e-05  6.652e-03  -0.011   0.9916
## Lit               6.605e-02  2.529e-01   0.261   0.7944
## Sci               7.557e-02  2.708e-01   0.279   0.7807
## Math             -4.139e-01  3.166e-01  -1.307   0.1940
## Histsoc          -1.707e-01  2.340e-01  -0.730   0.4672
## Lang              8.736e-01  3.966e-01   2.203   0.0297 *
## Art               7.260e-02  1.898e-01   0.383   0.7028
## spendsum         -1.922e-04  8.360e-04  -0.230   0.8186
## suppsum           7.866e-04  1.502e-03   0.524   0.6015
## susp:enroll      -2.220e-04  1.150e-04  -1.930   0.0562 .
## enroll:spendsum   6.379e-08  6.253e-07   0.102   0.9189
## enroll:suppsum   -6.995e-10  1.649e-06   0.000   0.9997
## abs:enroll        1.289e-04  9.397e-05   1.371   0.1732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 25.31663)
##
##     Null deviance: 13000.8  on 122  degrees of freedom
## Residual deviance:  2708.9  on 107  degrees of freedom
##   (761 observations deleted due to missingness)
## AIC: 763.39
##
## Number of Fisher Scoring iterations: 2
##
## [1] "- - - - Summary for ^4.9 - - - -"
## Analysis of Deviance Table (Type III tests)
##
## Response: grad^lambda
##                 LR Chisq Df Pr(>Chisq)
## abs              19.4965  1  1.008e-05 ***
## susp              2.5071  1   0.113334
## enroll            0.0000  1   0.995636
## Lit               0.0220  1   0.882050
## Sci               0.7022  1   0.402040
## Math              3.7964  1   0.051363 .
## Histsoc           0.2672  1   0.605232
## Lang              9.6192  1   0.001925 **
## Art               0.2290  1   0.632290
## spendsum          0.0080  1   0.928884
## suppsum           0.3028  1   0.582126
## susp:enroll       2.3342  1   0.126562
## enroll:spendsum   0.0592  1   0.807836
## enroll:suppsum    0.0098  1   0.921118
## abs:enroll        0.1839  1   0.668016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Call:
## glm(formula = grad^lambda ~ abs + susp + enroll + Lit + Sci +
##     Math + Histsoc + Lang + Art + spendsum + suppsum + enroll *
##     susp + enroll * spendsum + enroll * suppsum + enroll * abs,
##     data = grad_susp)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.138e+09  -3.716e+08   1.423e+08   4.510e+08   1.549e+09
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.499e+09  1.438e+09   3.129  0.00226 **
## abs             -5.730e+07  1.298e+07  -4.415 2.42e-05 ***
## susp            -2.776e+07  1.753e+07  -1.583  0.11629
## enroll          -5.358e+03  9.798e+05  -0.005  0.99565
## Lit              5.526e+06  3.724e+07   0.148  0.88233
## Sci              3.342e+07  3.988e+07   0.838  0.40391
## Math            -9.086e+07  4.663e+07  -1.948  0.05398 .
## Histsoc         -1.781e+07  3.446e+07  -0.517  0.60630
## Lang             1.812e+08  5.841e+07   3.101  0.00246 **
## Art              1.337e+07  2.795e+07   0.479  0.63327
## spendsum        -1.099e+04  1.231e+05  -0.089  0.92905
## suppsum          1.217e+05  2.211e+05   0.550  0.58327
## susp:enroll     -2.588e+04  1.694e+04  -1.528  0.12951
## enroll:spendsum  2.240e+01  9.210e+01   0.243  0.80830
## enroll:suppsum  -2.405e+01  2.429e+02  -0.099  0.92130
## abs:enroll       5.936e+03  1.384e+04   0.429  0.66888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 5.491326e+17)
##
##     Null deviance: 2.9811e+20  on 122  degrees of freedom
## Residual deviance: 5.8757e+19  on 107  degrees of freedom
##   (761 observations deleted due to missingness)
## AIC: 5390.1
##
## Number of Fisher Scoring iterations: 2
```

Interestingly, while one interaction term ('susp:enroll') is marginally significant in the model without transformation, this term loses its significance under the Box-Cox suggestion. In any case, this would suggest that as enrollment increases, the slope on grad~susp decreases.

Wrap-up

We have seen so far:

- High schools have the highest rates of absenteeism, statistically distinguishably from elementary and

middle schools
- Suspension rates within schools cannot be distinguished currently, as certain cateogry groups are highly underrepresented
- Graduation rates have increased over the past 5 years, but this change happened between 3-5 years ago
- There appears to be systematic differences in graduation rates, with CTE schools having higher rates than public schools
- Limited subsamples and inadequate operationalizations cause problems with trying to fit a predictive model for graduation rates. The only model that I could make sense of is the model that made sense already–graduation rates appear to be predicted negatively by suspension and absenteeism rates.

There are more variables on the student level that would be preferable to have, such as course-taking habits and disciplinary data. For starters, it would also be nice if per-pupil expenditure data were available at the school level. Perhaps these can be acquired at some point.

Fin.