Take Home Assessment

Data Analyst

UCLA CHPR

Shicheng Chen

04/13/2019

Assessment Description

Assume that you are providing statistical and programming support to the researcher on for a study on investigating the association between academic performance and individual characteristics as well as school characteristics. The researcher send you the following files to get started with:

imm10_a.csv	contains data for students from some schools
imm10_b.csv	contains data for students from other schools
imm10_lev2.csv	contains data for each school
label.txt	contains labels for each variable

Main Questions

The researcher is interested in finding out:
(1) Differences between students from public school and students from non-public school in terms of SES and gender.
(2) Do female students in public schools perform better than female students in non-public schools?

Data Checking and Cleaning

Data Checking:

- No duplicate variables.
- Only SES variable has NA's but the number of NA is really small, will not affect the overall decision.

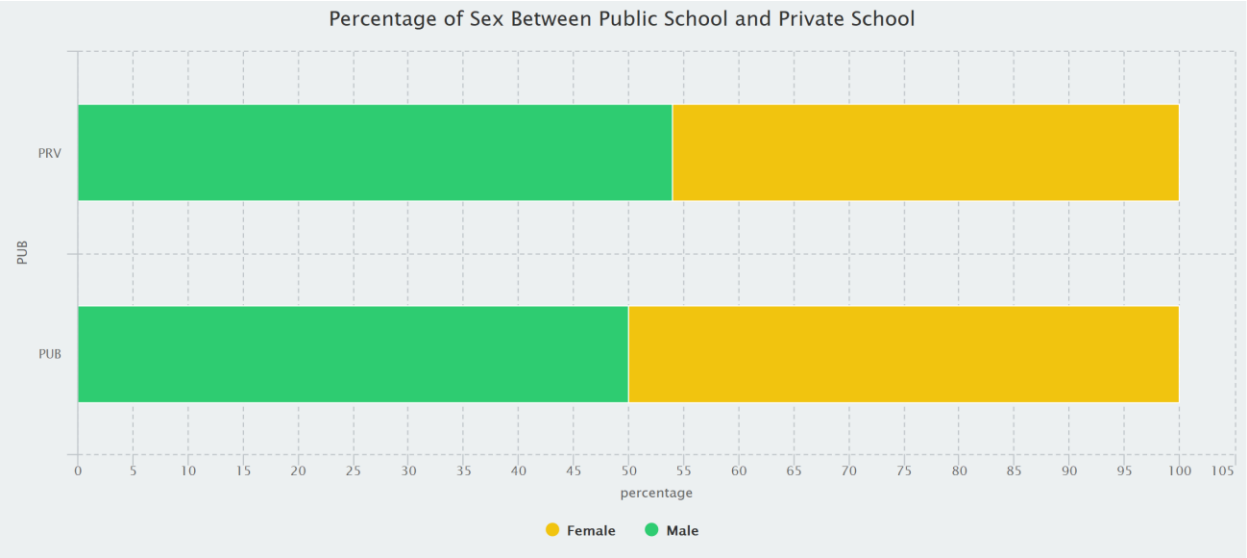
Data Cleaning:

- Merging imm10_a and imm10_b together.
- Separate public and private school data by imm10_lev2 variable MPUBLC.
- 193 observations in public schools file from 9 different schools, 67 observations in private school form 1 school.

Explanatory Data Analyst Between Public and Private School

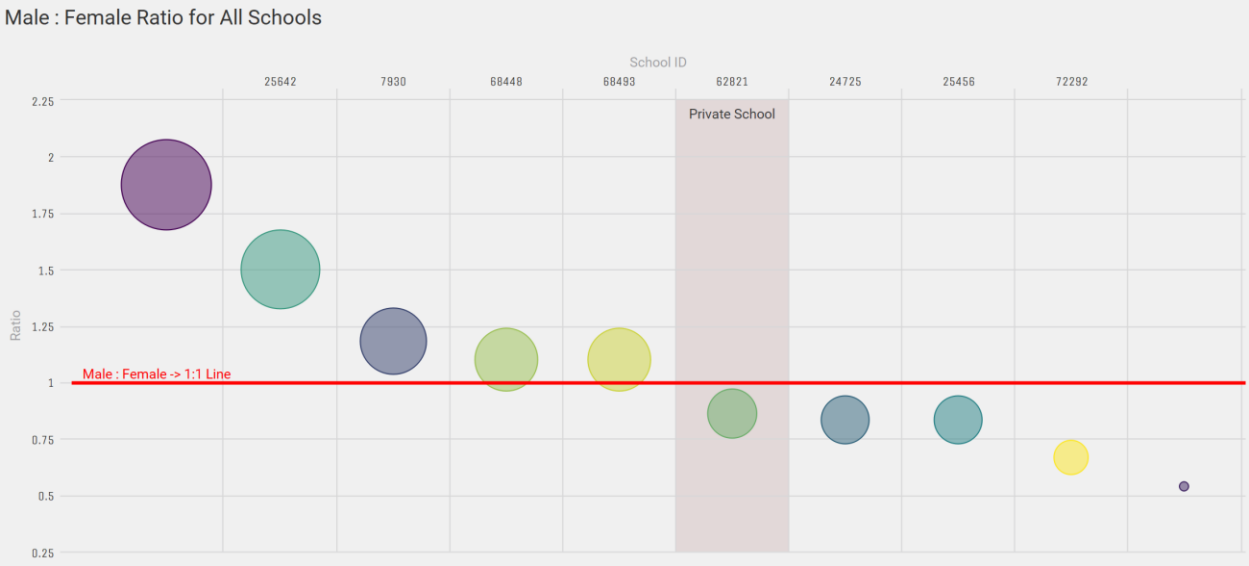
Question 1 Analysis:

- Sex Percentage



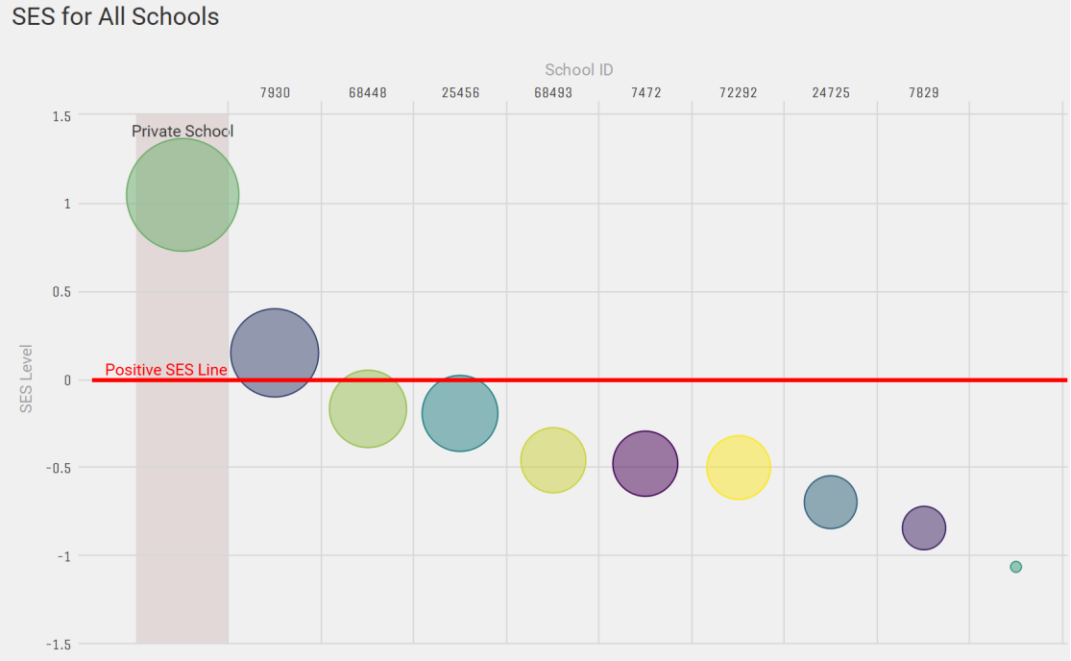
From the plot we can see private school has a little bit male students more than female students, and for public school, overall it is actually 50 and 50 percentage on both sex.

- Sex Ratio



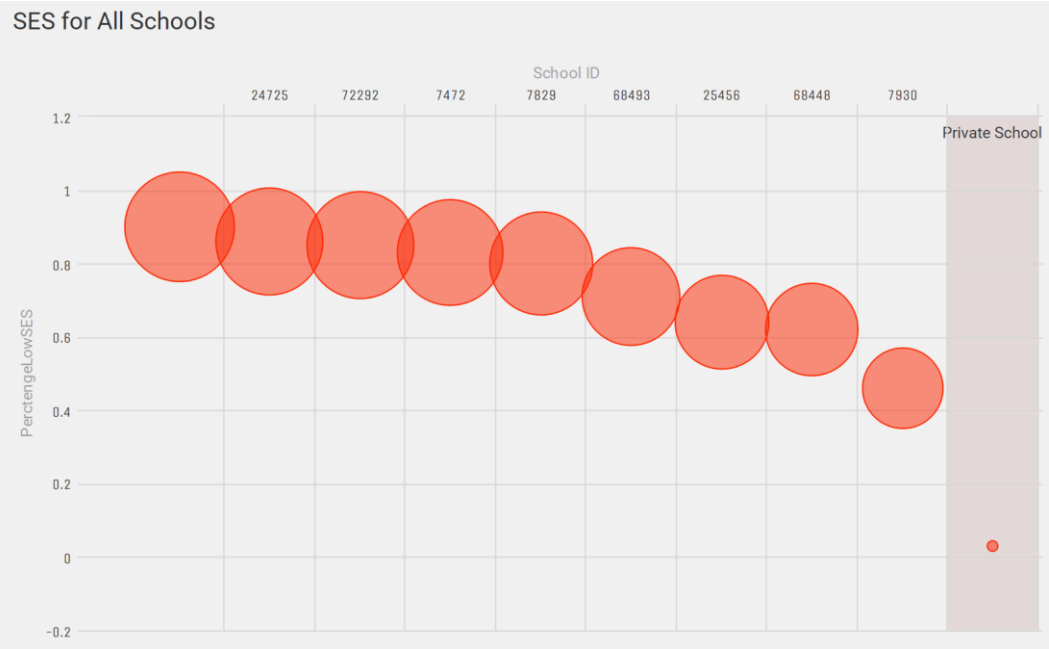
Most of public schools (5 our of 9) have more females, and the highest has ratio 1:1875. Our private school has ratio 1:0.8611, it is close to 1 : 1, but it still has more males than females.

- SES School Average



Private scholl has really high SES levels than all of other public schools. Only one public school has positive average SES levels, all of others are having negative SES levels.

- Low SES Percentage



All of the public schools have percentage over 0.45 of low SES level (SES level less than 0). Only the private school has a low percentage SES level close to 0.

Question 1 Result:

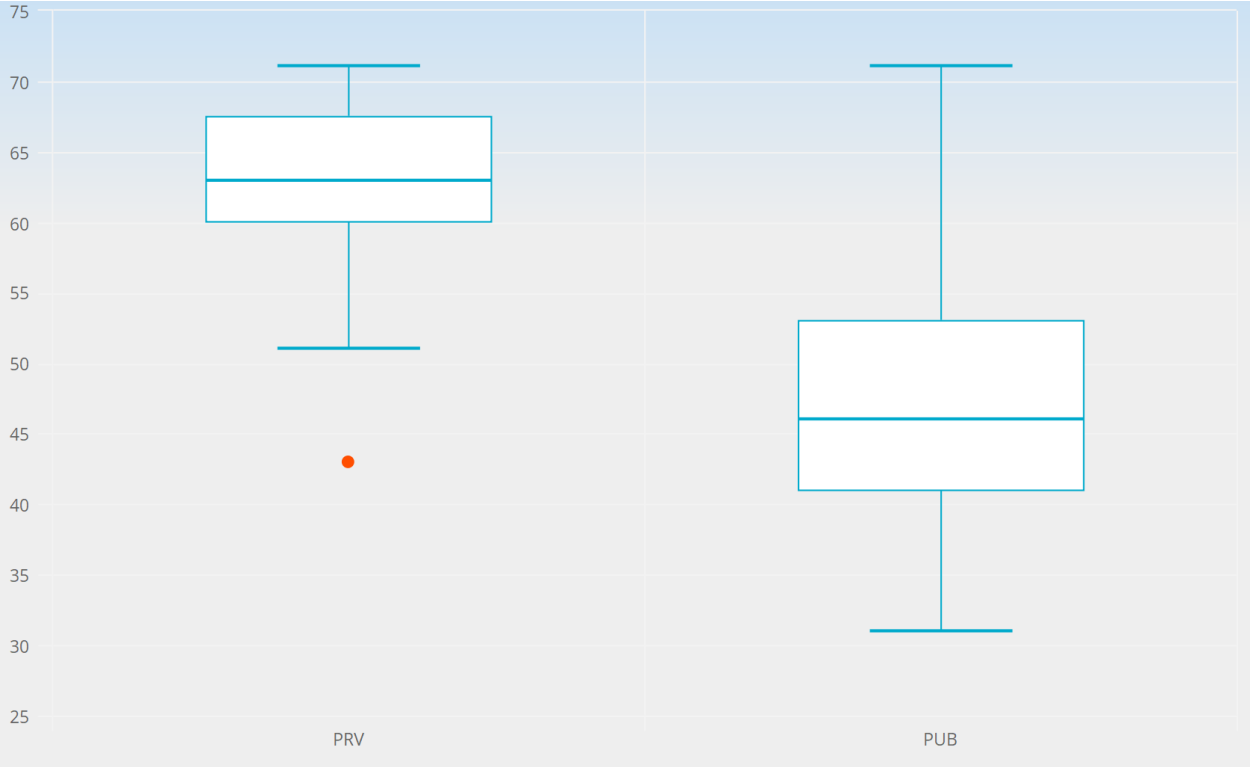
Answer:

In terms of gender and SES level, we can find that there is no big differences on genders. The public school contains the perfect half and half, but private school has male female ratio 1:0.8611, which is no obvious differences between two kinds of schools.

On the other hand, the SES levels has huge differences. On average, the private school has mean of 1.045 on SES levels, the highest in public school has mean of 0.147 on SES level. Depends on one of the reasons on SES level is income, it explains why private school has higher average SES score. Because of private school is not free, it banned people whom disable to pay additional money on the kids' education automatically.

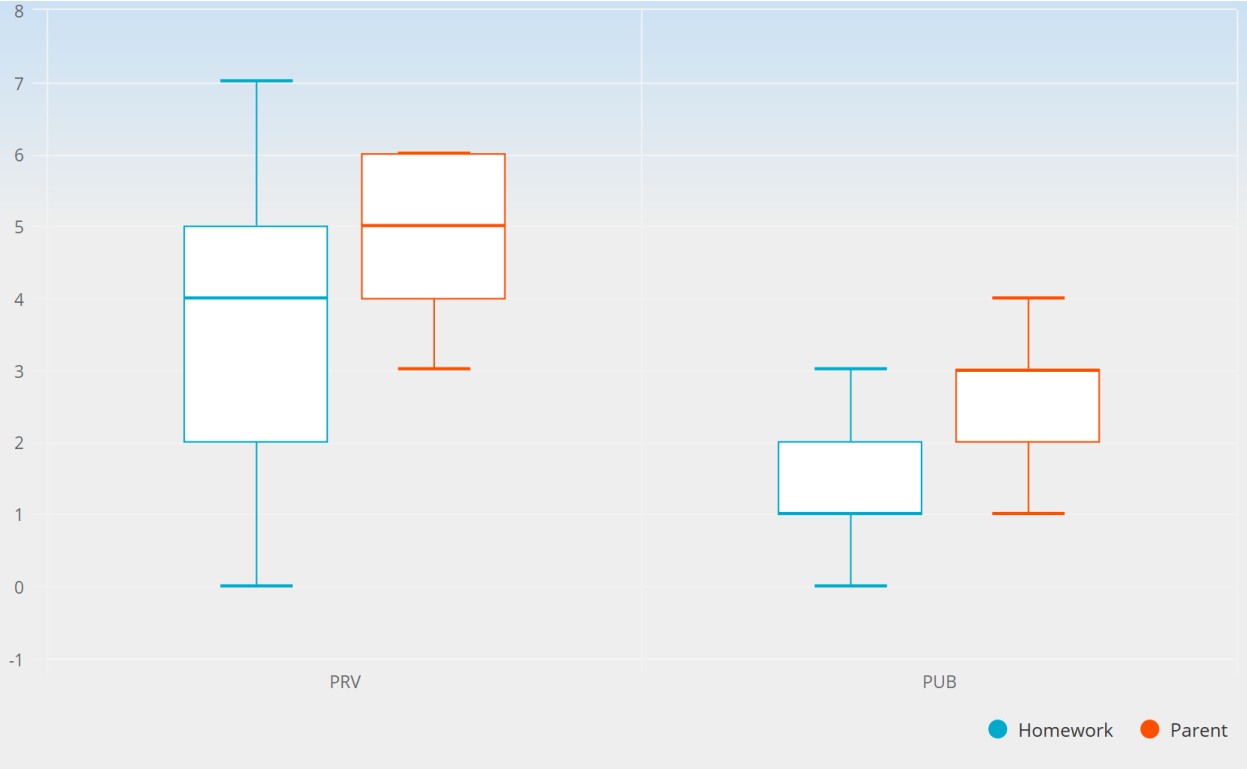
Question 2 Analysis:

- Math Score



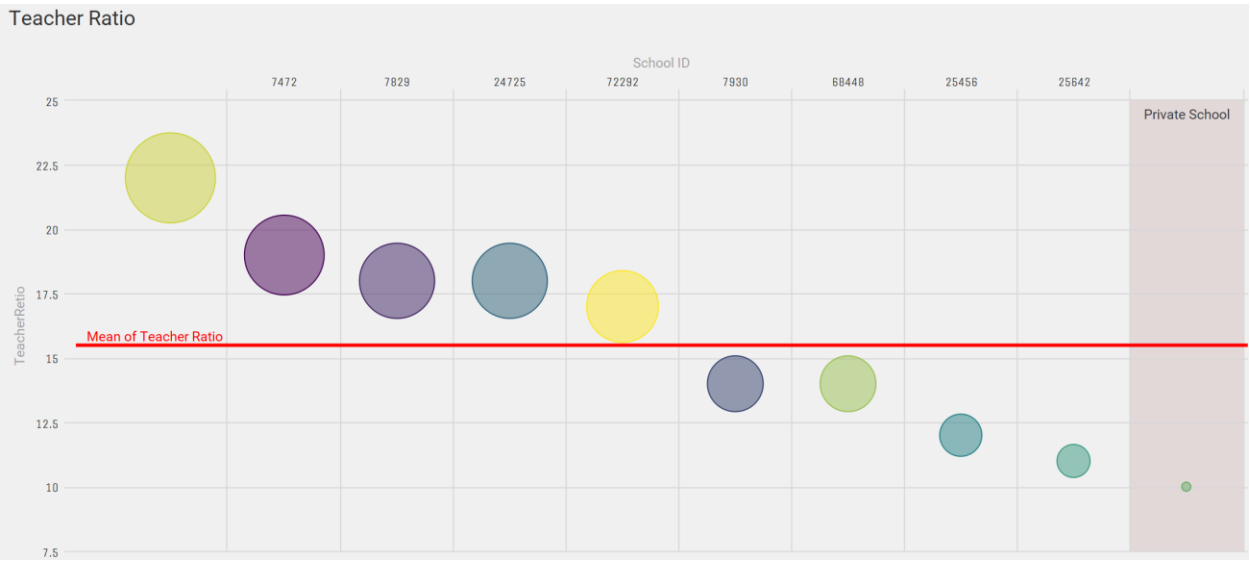
Comparing both box-plot, female from private school has higher on everthing than public school. Also noticed the twenty-five percentile of private school is higher than the median of public school average scores.

- Study time & Parent Education



On average, the parent education level and time spent each week on homework, the private is higher than public schools.

- Teacher Ratio



Teacher ratio is one teacher face to how many students. Commenly, less is better, which means each teacher have more attention on each stduent. Obviously, the prive has the least number.

Model: What cause the math score on female students in both public and private schools?

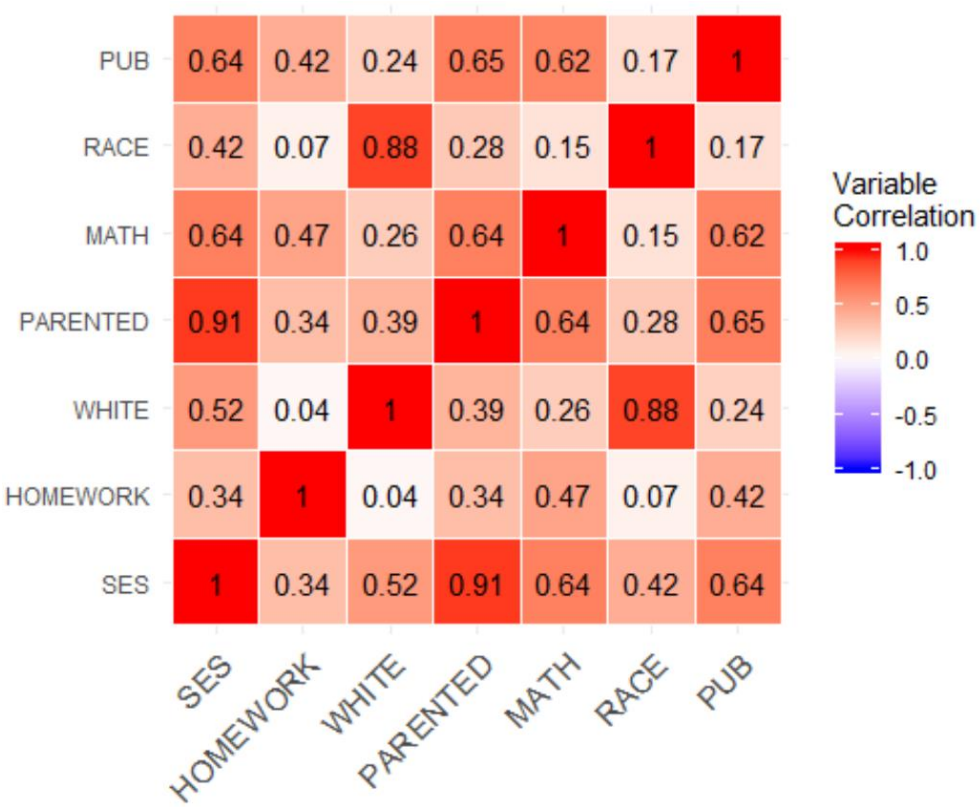
- Clean Data

  Use a combination data of all observations, setting additional variable labeling public school as 0, private school as 1.

  Delete the rows with NA on SES.

  Delete the School ID, Student ID, and Sex (only contain females)

- Correlation Heatmap



On the row/column of variable MATH, all other variables gives a positive correlation on score of MATH. The lowest correlation on influencing MATH is variable RACE and WHITE, which can

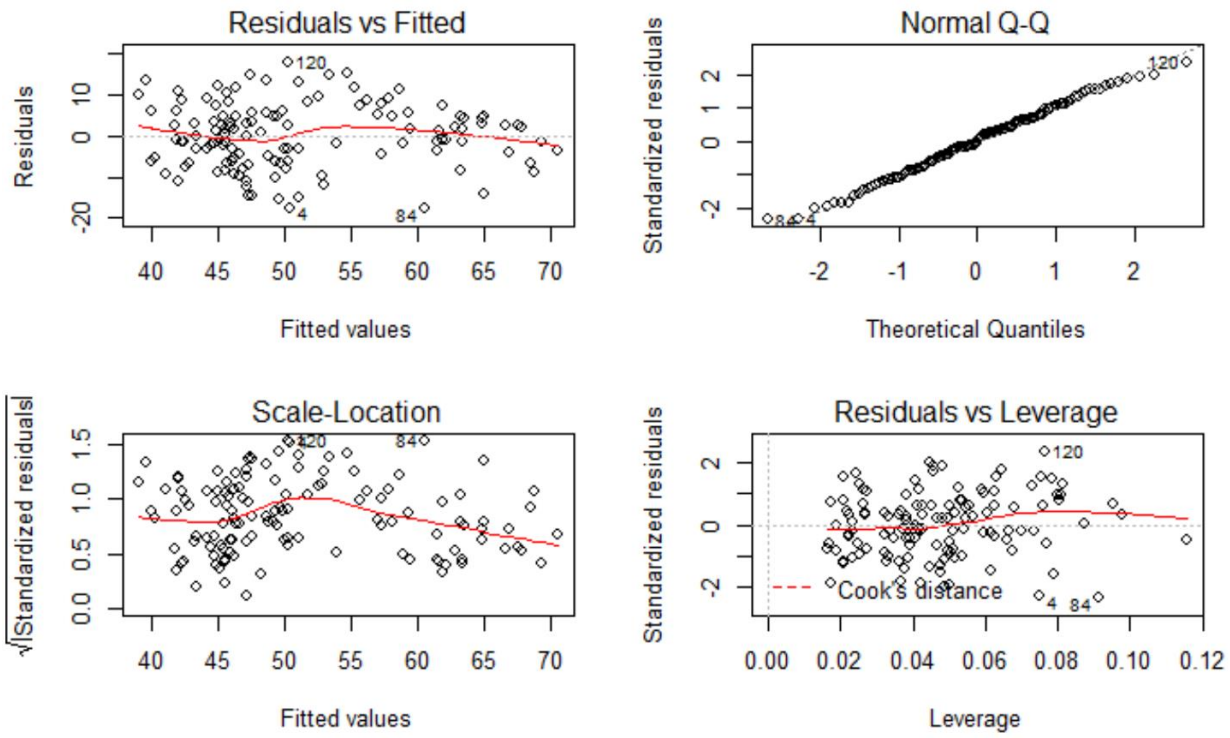
be defined as same variable. I choose the one with higher correlation which is WHITE and delete the RACE variable from the model.

- Linear Model

```
Call:
lm(formula = MATH ~ WHITE + SES + HOMEWORK + PARENTED + PUB,
    data = tmp)

Residuals:
     Min       1Q   Median       3Q      Max
-17.6066  -5.6427  -0.1319   4.9220  17.6951

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.6707     4.3113   9.897  < 2e-16 ***
WHITE        -0.2488     1.8787  -0.132  0.89487
SES           2.6707     1.8968   1.408  0.16171
HOMEWORK      1.5177     0.5013   3.028  0.00301 **
PARENTED      1.2956     1.1081   1.169  0.24462
PUB           6.7837     2.2531   3.011  0.00317 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.86 on 121 degrees of freedom
Multiple R-squared:  0.5288,    Adjusted R-squared:  0.5093
F-statistic: 27.15 on 5 and 121 DF,  p-value: < 2.2e-16
```

From the linear model, we can see the variable WHIT, SES and PARENTED are not statistically significant. Overall, the p-value is less than alpha = 0.05, which we can reject our null hypothesis and say our list variable is influenced the MATH score. Also our R-squared is larger than Adjusted R-squared, which label I did not overfitting my model. Finally, the R-squared shows there are 52.88% of all the data fit the model.

- Linear Model Plot



From the regular model plot we can see that there is not a strong curve pattern on Residuals vs Fitted plot, and a not strong curve pattern on Scale-Location graph, In the Residuals vs Leverage plot, we can see that there are leverages but not bad. On the Normal-Q-Q Plot, the plot are fitting the line very well with a perfect normal distribution

- ANOVA

```
             Df Sum Sq Mean Sq F value   Pr(>F)
WHITE         1   1068    1068  17.293 6.02e-05 ***
SES           1   5536    5536  89.608 3.01e-16 ***
HOMEWORK      1   1022    1022  16.542 8.51e-05 ***
PARENTED      1    202     202   3.262  0.07340 .
PUB           1    560     560   9.065  0.00317 **
Residuals   121   7475      62
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                  2.5 %    97.5 %
(Intercept) 34.1353409 51.205988
WHITE       -3.9681889  3.470607
SES         -1.0845937  6.425895
HOMEWORK     0.5253460  2.510094
PARENTED    -0.8981452  3.489283
PUB          2.3231173 11.244377
```

In our ANOVA model, all of the variables are statistically significant on 90% confidence. In the 95% confidence interval table, the HOMEWORK and PUB (In public school or not, 0 means public, 1 means private) will absolutely increase the MATH score.

Question 2 Result:

Answer:

For female students in public schools, whatever on data plot shows or the model, have high probability will not perform better than female in private school with these reasons:

1. On parent education level, the private school students’ parent’s education level is higher than public school students’ parents.

2. On homework time spend time, female in private school spend longer time than student in public school.

3. On teacher ratio, on average, one teacher in private school only need to face 10 students, but in public school, the least score of the teacher ratio is 11.

4. From the model, the variable PUB (In public school or not, 0 means public, 1 means private) have 95% condifence to increase female students' MATH score from 2.32 to 11.24.

My Additional Question & Limitations:

- Limitations:

    1. Do not have enough samples to analyze race differences.

    2. Only have one private school, which cannot label or delegate the average private school levels.

    3. Only have MATH score to determine the proffer better on question 2).

- Questions:

    1. What do you think that I should deal with 6 NA values in SES variable better?

    2. This answer may have bias because of the limitation of private school, if we only have this data and does have chance to have more, how can we make the result less bias?

    3. How des the SES level calculated, I estimated one reason is income, what are the other explanations?

    4. We have both WHITE and RACE, and RACE contains 'WHITE', is it necessary?

    5. There is no unit on HOMEWORK, basically I posit its in hours, but I am not sure.

    6. I did not use t he MPERCMIN variable, both schools has over 78% of majority races, and minority does not have enough samples to analyze.