Topic 2

I. Abstract

A. Introduction

Exams are the primary measure of students' knowledge throughout their academic careers. They are used to enhance student learning (e.g. challenge students to apply their skills) or to measure student knowledge (e.g. to determine course grades or make instructional decisions). However, grades do not always give a true picture of a student's knowledge or abilities. Some students may have great difficulties demonstrating their knowledge in exams due to objective factors.

In this topic, we focus on math score, since math and numerical problem solving are a part of most cognitive ability tests.

Learning how various factors like economic, personal and social can influence the students' performance on math exams can help us evaluate the potentials and advantages of students objectively and understand the limit of math exam in measuring students' knowledge and skills.

B. Objective Statement

In this study, we analyzed the math score (response variable) as a linear combination of effects due to factors of gender, race/ethnicity, parental level of education, lunch type and test preparation course, with the following questions.

a. Does the independence of observations meet the assumptions of ANOVA ？

b. For each variable's group, are the math score data normally distributed? And with homogeneity of variance?

c. Would each variable affect the math score?

d. Is there any interactions between the 5 variables? How to interpret the different results showing by type I, type II and type III ANOVA tables?

C. Statistical Procedure

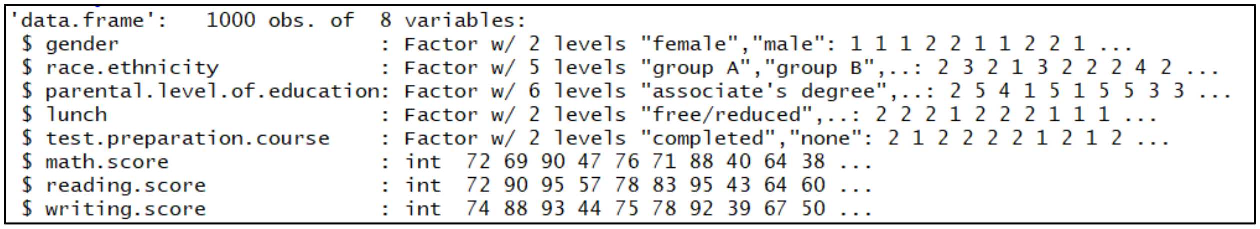
This study applied N-way analysis of variance on analyzing score of math in R. It was designed as multi-way fixed-effects model of ANOVA. Firstly, we checked the data to ensure it was formatted and designed properly. Then we fit several 2-way ANOVA models with all possible combinations of 2 factors from the 5 factors to figure out if there was an interaction. Since no interaction was found and each factor had main effect on response, we can fit 5 one-way ANOVA models to analyze factors one by one and check the validity of each ANOVA model. Then we used Tukey-Kramer test to show the difference in means when a main effect was significant, since the data for ANOVA is unbalanced. For those factors invalid for ANOVA tests, we conducted nonparametric tests to detect differences between levels.

II. Data Description and Preprocessing

The raw dataset consisted of 1000 observations (math, reading and writing 3 score types) with 5 variables. Each individual was represented by 1 measurement only. The 5 variables used in this study were gender, race/ethnicity, parental level of education, lunch type and test preparation course. There was no missing data (NA or NAN).

Data source: https://www.kaggle.com/spscientist/students-performance-in-exams

As we only focused on math score, the reading and writing columns were eliminated from the dataset. We assumed the 5 factors are independent of each other. We also assumed that gender, race, parental level of education, good lunch and completion of the preparation course would impact their math score. Furthermore, there could be interactions between these factors. Therefore, we took all 5 factors into account which were all categorical data and can be categorized at different levels respectively. The response variable (math score) was continuous numerical. We can get basic information of each factor by the frequency table of categorical variables.



```
'data.frame':   1000 obs. of  8 variables:
 $ gender                     : Factor w/ 2 levels "female","male": 1 1 1 2 2 1 1 2 2 1 ...
 $ race.ethnicity             : Factor w/ 5 levels "group A","group B",..: 2 3 2 1 3 2 2 2 4 2 ...
 $ parental.level.of.education: Factor w/ 6 levels "associate's degree",..: 2 5 4 1 5 1 5 5 3 3 ...
 $ lunch                      : Factor w/ 2 levels "free/reduced",..: 2 2 2 1 2 2 2 1 1 1 ...
 $ test.preparation.course    : Factor w/ 2 levels "completed","none": 2 1 2 2 2 2 1 2 1 2 ...
 $ math.score                 : int  72 69 90 47 76 71 88 40 64 38 ...
 $ reading.score              : int  72 90 95 57 78 83 95 43 64 60 ...
 $ writing.score              : int  74 88 93 44 75 78 92 39 67 50 ...
```

Figure 1: Structure of the raw Data Set

gender	female 518			male 481		
race	A 89	B 190	C 319	D 262	E 140	
PLOE	associate 222	bachelor 118	high school 196	master 59	some college 226	some high school 179
lunch	free/reduced 355			standard 645		
TPC	completed 358			none 642		

Figure 2: Frequency table of categorical variables

The natural model to consider was a 5-way ANOVA with interaction of all possible combinations of the 5 variables. However, if we constructed the model with interaction of gender * race * PLOE * lunch * TPC, we should split the data into 2*5*6*2*2=240 groups. Even though we had 1000 observations, there was at least 1 group without data and we cannot run the 5-way ANOVA consequently. For example, we found no data under the situation of race A, PLOE with master's degree and complete the preparation course. For the same reason, we cannot run the 3-way ANOVA for all possible combination of 3 factors. That is the reason why we run all 2-way ANOVA test for the analysis in the next section.

Then, we turned to the sample distribution to see if the samples of math score came from normal distribution. See Figure 3. From the boxplot, we found there were several outliers (8 points) on the lower side. The normal qqplot and histogram also showed the distribution of math score was negative skew. The majority of data were on the qqline but with light tail.

Removing the outliers would solve the negative skewness problem and improve fitness of normality of the model (result in Q-Q plot), see figure 4. Since the ANOVA test should meet the normality assumption, we would exclude the outliers from the dataset.

When we made a frequency histogram for factor variables, we found the math score was normally distributed for each factor in general. But for some factors, such as race and PLOE, there were some missing bars on the histogram, which means the sample size was not big enough. As for factors such as lunch and TPC, it clearly showed an offset between the spread of two levels. See figure 5.

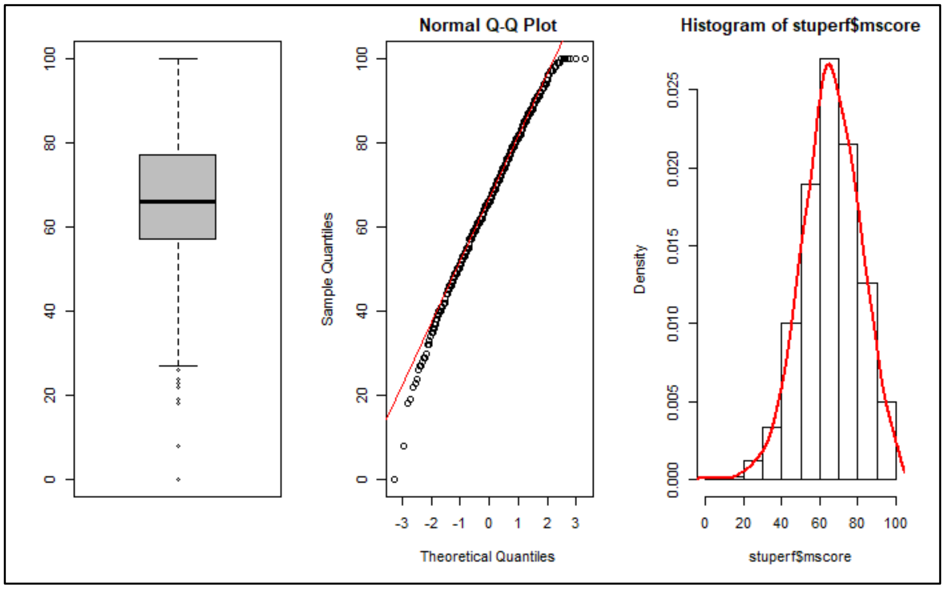


Figure 3: Distribution of math score with outliers

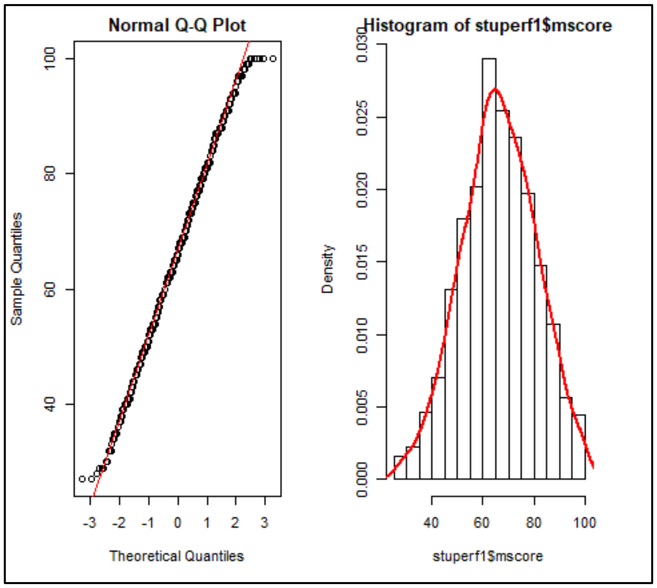


Figure 4: Distribution of math score without outliers

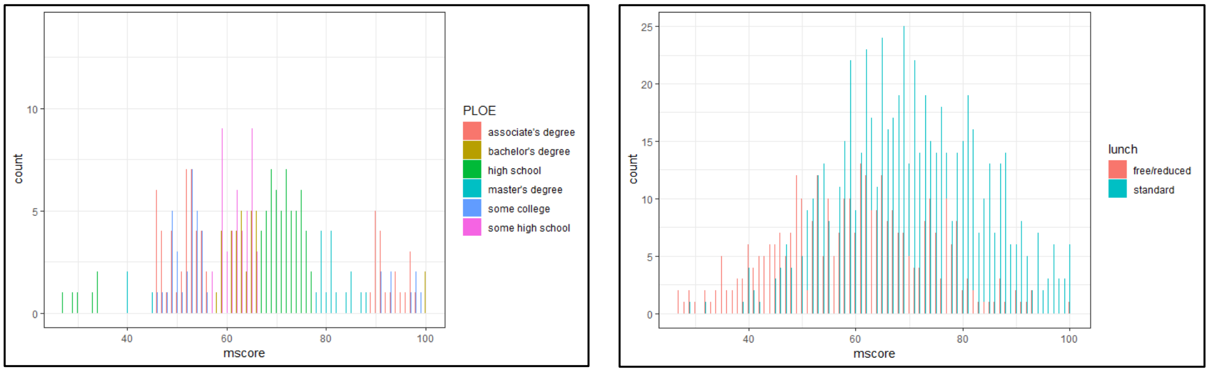


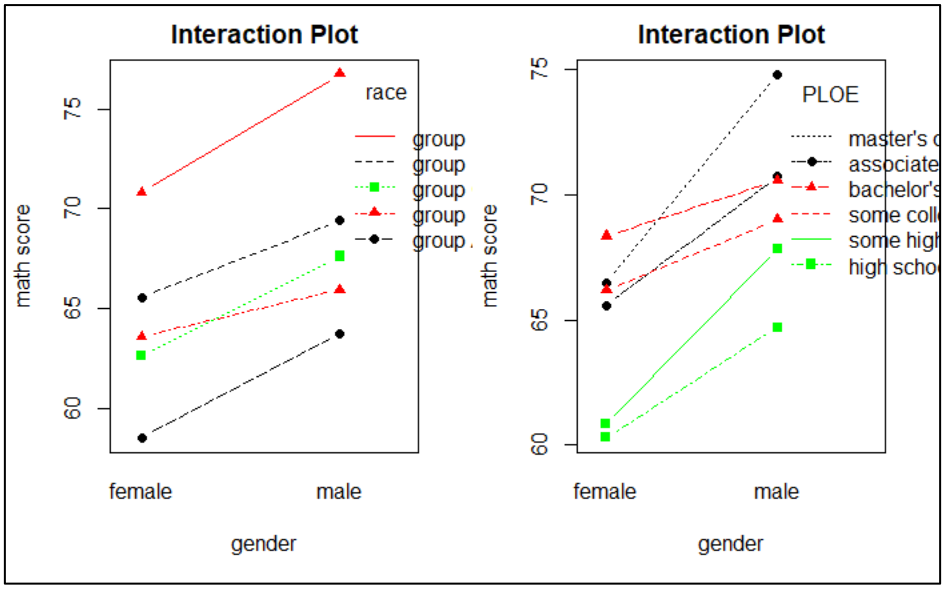
Figure 5: Frequency histogram for factors PLOE and lunch

III. Data Analysis

A. Interactions

Our primary statistical interest is whether each level of a factor is compensated equally, on average, after adjusting math score for other given effect of the rest factors. Therefore, we should first figure out the interaction among factors.

To start the analysis, we made plot for every two factors. The interactions plots showed that there were possible interactions between gender and race, gender and PLOE, as well as race and PLOE. See figure 6.



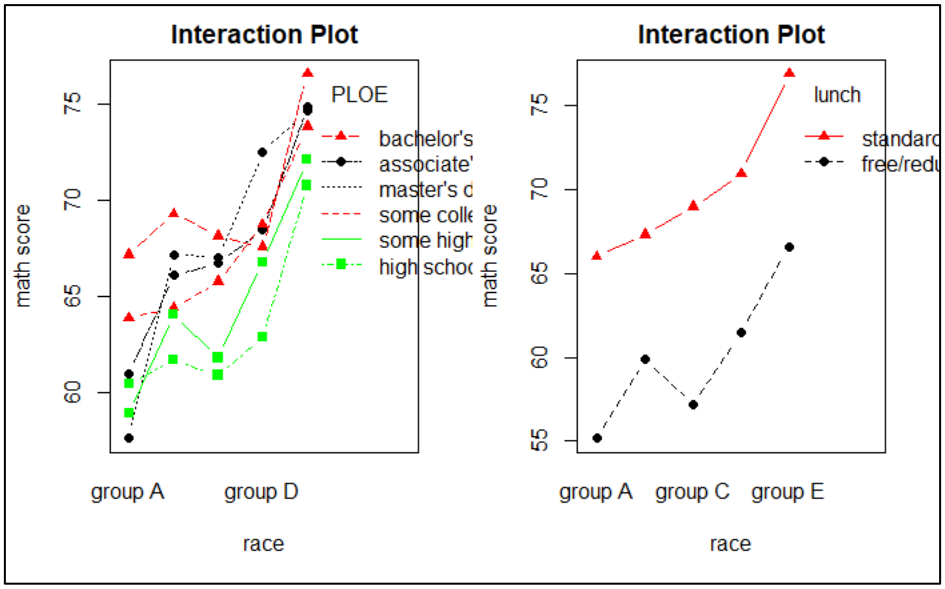


Figure 6: Partial interaction plot of two factors

Were the interactions shown in figure 6 significant? To answer this question, we ran the ANOVA tests. Since we had all levels for each factor and each factor had uneven levels, the ANOVA tests would be unbalanced fixed effect models. The ANOVA type III is appropriate for our analysis because this type tests for the presence of a main effect after the other main effect and interaction [1].

Besides, considering the data availability and feasibility of multi-ANOVA, we chose to run 10 2-way ANOVA tests covering all possible combinations of two factors from the 5 factors. Consequently, we found no significant interaction among these 5 factors. See figure 7.

Model no.	Factor 1	Factor 2	p-value of interaction
1	gender	race	0.7899
2	gender	POLE	0.5493
3	gender	lunch	0.4335
4	gender	TPC	0.6524
5	race	PLOE	0.998871
6	race	lunch	0.5688
7	race	TPC	0.3951
8	PLOE	lunch	0.4965
9	PLOE	TPC	0.8719
10	lunch	TPC	0.8056

Figure 7: P-values of all 2-way ANOVA models with interaction

B. Main effects

The next question was which model was better to test main effect of each factor: the model of union 5 factors or 5 one-way ANOVA models? In addition, we tested the assumptions of ANOVA [2] to ensure validity of each test. In our study, on residual of model [3], we used Shapiro-Wilk test to check Normality assumption and used Bartlett's test to check homogeneity of variance assumption [4]. Since ANOVA assumes that the data come from normally distributed populations and unequal variances can affect the Type I error rate and lead to false positives [5], we would take the ANOVA test valid If and only if both tests have nonsignificant results.

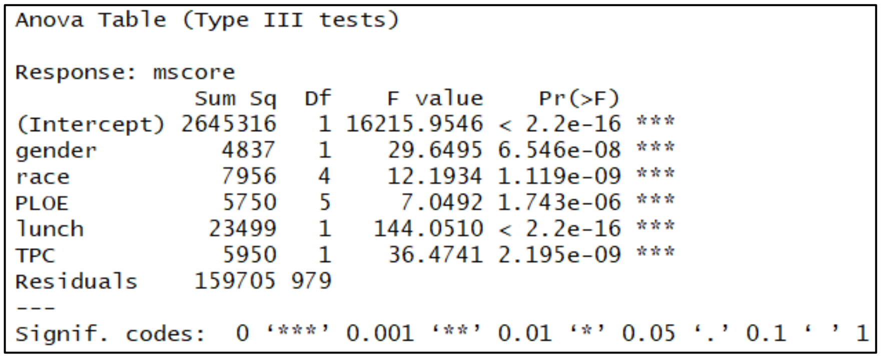
In general, the hypotheses test of main effect is shown as follows:

H_0: α1=α2=α3=…=αi=0 (i=total number of levels for a factor) #effects are zero

H_1: at least one α is not 0

If p-value < 0.05, we can reject H_0 and conclude the factor has a significant main effect on math score.

1. Model: mscore = Grand mean + gender effect + race effect + PLOE effect +lunch effect + TPC effect + Residual

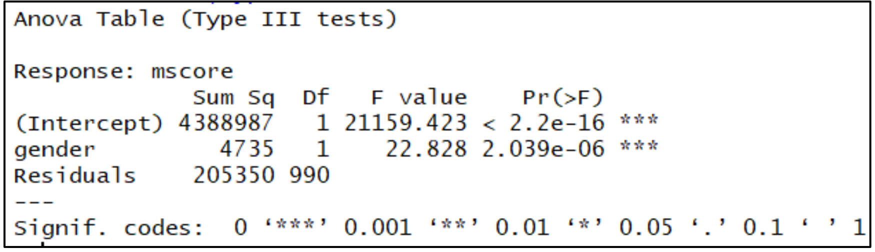


```
Anova Table (Type III tests)

Response: mscore
             Sum Sq  Df    F value    Pr(>F)    
(Intercept) 2645316   1 16215.9546 < 2.2e-16 ***
gender         4837   1    29.6495 6.546e-08 ***
race           7956   4    12.1934 1.119e-09 ***
PLOE           5750   5     7.0492 1.743e-06 ***
lunch         23499   1   144.0510 < 2.2e-16 ***
TPC            5950   1    36.4741 2.195e-09 ***
Residuals    159705 979                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8: ANOVA table for union factors model without interaction

We found that each factor had significant main effect on math score, but this model failed the test for normality assumption, the p-value for shapiro-wilk test was 0.005321. Therefore, we refused this model and its test results.

2. Model1: mscore = Grand mean + gender effect +Residual

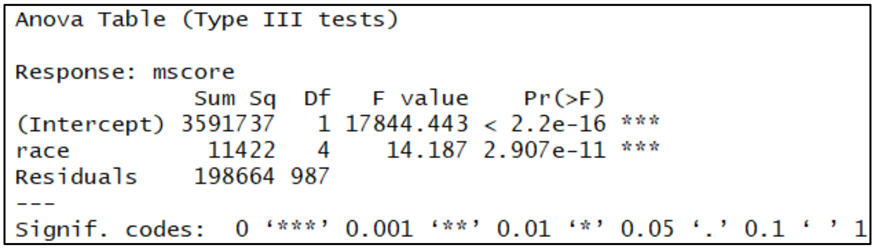


```
Anova Table (Type III tests)

Response: mscore
             Sum Sq  Df   F value    Pr(>F)    
(Intercept) 4388987   1 21159.423 < 2.2e-16 ***
gender         4735   1    22.828 2.039e-06 ***
Residuals    205350 990                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: One-way ANOVA table with gender factor

Main effect of gender was significant, and p-values for shapiro-wilk test and Bartlett test were 0.1037 and 0.8904 respectively.

3. Model2: mscore = Grand mean + race effect +Residual



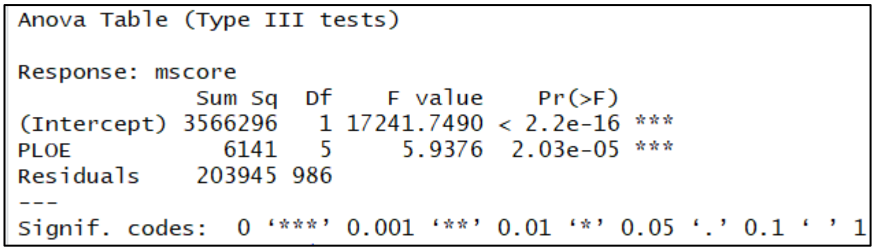
```
Anova Table (Type III tests)

Response: mscore
             Sum Sq  Df   F value    Pr(>F)    
(Intercept) 3591737   1 17844.443 < 2.2e-16 ***
race          11422   4    14.187 2.907e-11 ***
Residuals    198664 987                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: One-way ANOVA table with race factor

Main effect of race was significant, and p-values for shapiro-wilk test and Bartlett test were

0.09344 and 0.4079 respectively.

4. Model3: mscore = Grand mean + PLOE effect +Residual

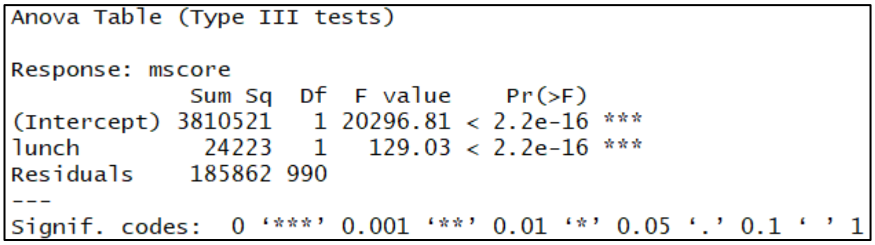


```
Anova Table (Type III tests)

Response: mscore
             Sum Sq  Df    F value    Pr(>F)    
(Intercept) 3566296   1 17241.7490 < 2.2e-16 ***
PLOE           6141   5     5.9376  2.03e-05 ***
Residuals    203945 986                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: One-way ANOVA table with PLOE factor

Main effect of PLOE was significant, but this model failed the shapiro-wilk normality test, the p-value was 0.01175.

5. Model4: mscore = Grand mean + lunch effect +Residual

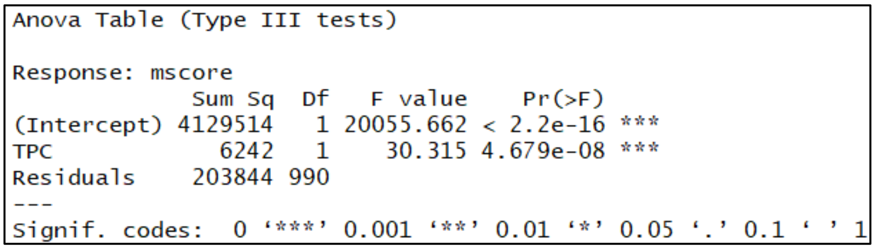


```
Anova Table (Type III tests)

Response: mscore
             Sum Sq  Df  F value    Pr(>F)    
(Intercept) 3810521   1 20296.81 < 2.2e-16 ***
lunch         24223   1   129.03 < 2.2e-16 ***
Residuals    185862 990                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12: One-way ANOVA table with lunch factor

Main effect of lunch was significant, and p-values for shapiro-wilk test and Bartlett test were 0.06336 and 0.4154 respectively.

6. Model5: mscore = Grand mean + TPC effect +Residual



```
Anova Table (Type III tests)

Response: mscore
             Sum Sq  Df   F value    Pr(>F)    
(Intercept) 4129514   1 20055.662 < 2.2e-16 ***
TPC            6242   1    30.315 4.679e-08 ***
Residuals    203844 990                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 13: One-way ANOVA table with TPC factor

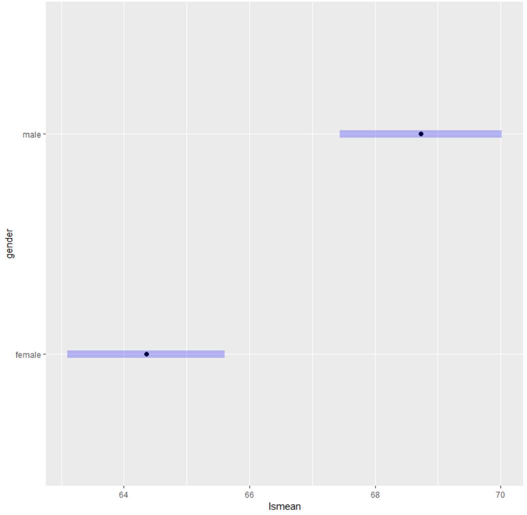
Main effect of TPC was significant, but this model failed the shapiro-wilk normality test, the p-value was 0.025.

From the above discussion, only gender, race and lunch had valid significant effects on math score means. But this did not tell us anything about the levels which means were significantly different for each factor. Post-hoc tests were needed.

C. Post-hoc tests

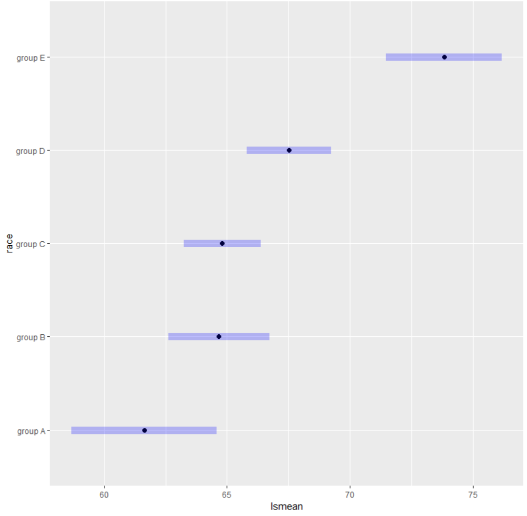
We ran a post-hoc test for each factor with valid significant main effect on math score to check which of the levels were different from the others. Since our data was unbalanced, the Tukey-Kramer test was more appropriate under this condition [6].

The previous tests told us whether a factor was important for effect on math score. The focus turned to be on how large the effect might be. A simple way to check is by constructing a confidence interval for each level of significant factors. Based on our tests results, we summarized the 3 significantly effective factors as follows:



1. Gender

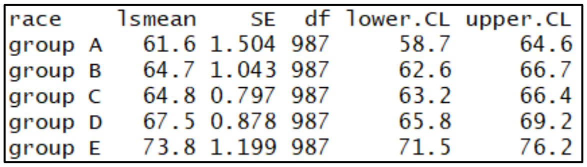
The mean math score for men is 4.37 higher than that for women. A rough 95% CI for the range of plausible values for the gender effect is [67.4,70.0] for men and [63.1,65.6] for women.



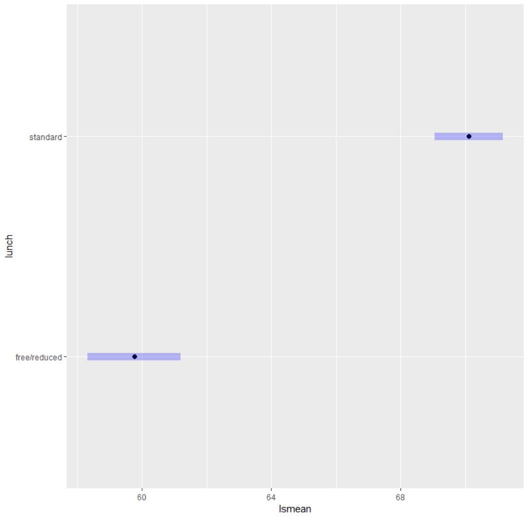
2. Race

The order from high to low for mean of math score is group E, group D, group C and group B, and group A. There is no significant different for group B, group C and group D. However, the different between group E and group A can reach 12.192!

A rough 95% CI for the range of plausible values for the race effect is shown below:



```
race     lsmean    SE  df lower.CL upper.CL
group A    61.6 1.504 987     58.7     64.6
group B    64.7 1.043 987     62.6     66.7
group C    64.8 0.797 987     63.2     66.4
group D    67.5 0.878 987     65.8     69.2
group E    73.8 1.199 987     71.5     76.2
```



3. Lunch

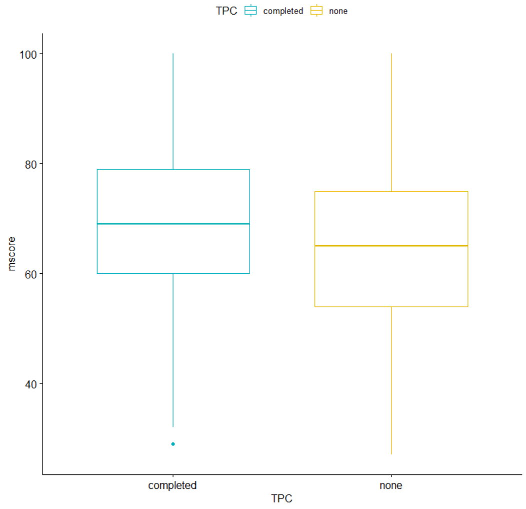
The mean math score for standard lunch is 10.4 higher than that for free/reduced lunch. A rough 95% CI for the range of plausible values for the lunch effect is [69.1,71.2] for standard lunch and [58.3,61.2] for free/reduced lunch.

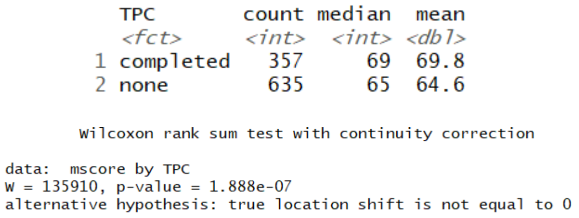
D. Nonparametric Tests

In the previous chapters, we found out there was no interactions among the 5 factors. For the factors of gender, race and lunch, we obtained the estimations of their effects on math score. However, due to violation of normality assumption, the ANOVA test for PLOE and TPC are invalid and the effects of PLOE and TPC are still in uncharted territory.

Here we considered two alternative test procedures, Wilcoxon rank sum test and Kruskal-Wallis test for analyzing TPC and PLOE respectively. The former can be used to compare two independent groups of samples (even they are unpaired) while the later extends the former in the situation where there are more than two groups.

1) Wilcoxon rank sum test for TPC



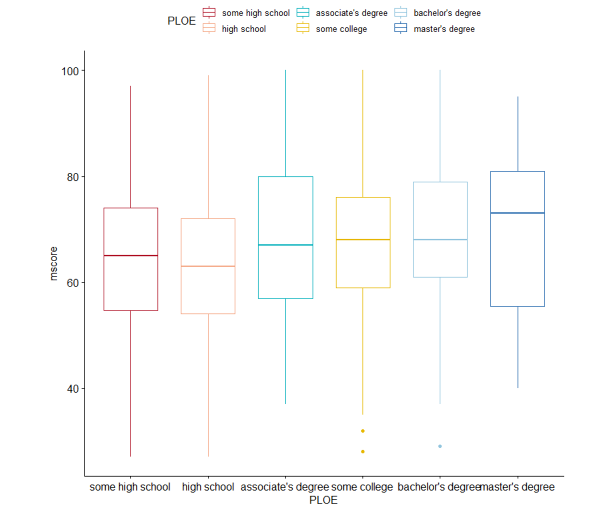


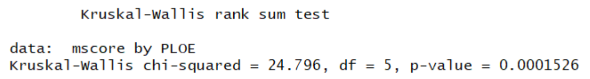
```
    TPC        count median  mean
    <fct>      <int>  <int> <dbl>
  1 completed    357     69  69.8
  2 none         635     65  64.6

        Wilcoxon rank sum test with continuity correction

data:  mscore by TPC
W = 135910, p-value = 1.888e-07
alternative hypothesis: true location shift is not equal to 0
```

The p-value of test is close to 0 and we reject H_0 that conclude there is significant difference between completed and none.

The students TPC-completed have a mean math score 5.2 higher than that of TPC-none.

2) Kruskal-Wallis test for PLOE

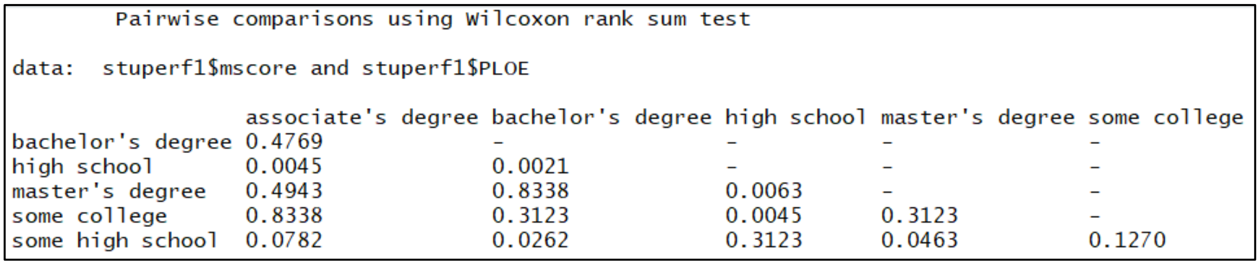




```
        Kruskal-Wallis rank sum test

data:  mscore by PLOE
Kruskal-Wallis chi-squared = 24.796, df = 5, p-value = 0.0001526
```

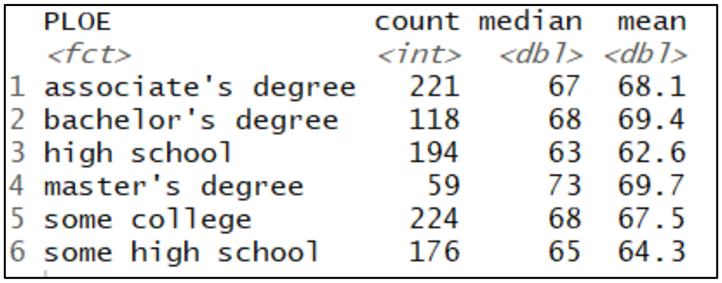
The p-value of test is close to 0 and we reject H_0 that conclude there are significant differences among the PLOE groups.

To find out which pairs of PLOE groups are different, we calculated pairwise comparisons between group levels with corrections for multiple testing.



```
        Pairwise comparisons using Wilcoxon rank sum test

data:  stuperf1$mscore and stuperf1$PLOE

                 associate's degree bachelor's degree high school master's degree some college
bachelor's degree 0.4769             -                 -           -              -
high school       0.0045             0.0021            -           -              -
master's degree   0.4943             0.8338            0.0063      -              -
some college      0.8338             0.3123            0.0045      0.3123         -
some high school  0.0782             0.0262            0.3123      0.0463         0.1270
```

The pairwise comparison showed that, some high school and high school, some high school and bachelor's degree, some high school and master's degree, high school and associate's degree, high school and bachelor's degree, and high school and master's degree were significantly different (p-value<0.05).



```
  PLOE               count median  mean
  <fct>              <int>  <dbl> <dbl>
1 associate's degree   221     67  68.1
2 bachelor's degree    118     68  69.4
3 high school          194     63  62.6
4 master's degree       59     73  69.7
5 some college         224     68  67.5
6 some high school     176     65  64.3
```

If we calculate mean by groups, we found that students whose parents have high school degree had lowest math score while students whose parents have master's degree had highest math score. Students whose parents have associate's degree and above performed much better than students whose parents have high school's degree and below.

IV. Discussion

According to the test results of the five factors, we found gender, race and lunch type significantly affected the math score through ANOVA tests, as well as TPC and PLOE through nonparametric tests. Males have better performance than females in math tests. Race E is really good at math while race A is struggling with math. Having a good lunch can greatly improve performance in math tests. Completion of test preparation course can significantly make math score higher. Parents who has associate's and above degree yield positive effects on children's math performance. Therefore, if you are eager to improve your math score, maybe it is impossible to change your gender, race, parental level of education but you can choose to have healthy and nutrient lunch and to complete the test preparation course seriously. Good lunch not only helps you to catch up in math, but also pays you off both mentally and physically. A full and thorough preparation can prevent from failure somehow during the exams.

V. Summary

This study mainly explored factors that had significant effects on math score. The raw dataset included scores from three exams at a public school and a variety of personal, social and economic factors that have interaction effects. Here we chose math score as the only response variable. All factors were collected on completed levels, although the observations were not evenly arranged on each level.

Therefore, we conducted fixed-effect, unbalanced, multi-way models of ANOVA and Tukey-Kramer test was performed in the multiple comparison procedure. This study aimed to examine possible interactions among factors and the main effect of each factor. We also checked the validity of ANOVA tests based on test of assumptions about data, since violating any of these assumptions can result in false positives or false negatives. For those factors who failed the normality assumption test, we conducted nonparametric test to detect if the differences are significant.

The result of the analysis showed that males perform better than females in math test, race group E is the best at math while group A has least mean of math score. Students who haves standard lunch have higher math score compared to those having free/reduced lunch. In addition, completion of test preparation course can significantly make math score higher. Parents who has associate's and above degree yield positive effects on children's math performance.

VI. References

[1] https://mcfromnz.wordpress.com/2011/03/02/anova-type-iiiiii-ss-explained/

[2] https://yieldingresults.org/wp-content/uploads/2015/03/Checking_ANOVA_assumptions.html

[3] https://www.theanalysisfactor.com/checking-normality-anova-model/

[4] https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm

[5] http://blog.minitab.com/blog/quality-business/common-assumptions-about-data-part-2-normality-and-equal-variance

[6] https://rcompanion.org/rcompanion/d_05.html

Appendix II: R code