Project 1: Project STAR I

01/17/2020

Group Partners:

1. Bingdao Chen bdchen@ucdavis.edu contribution: explore data and test causal effect
2. Yahui Li yhuli@ucdavis.edu contribution: write down and hold one-way ANOVA model
3. Zihan Wang zihwang@ucdavis.edu contribution: test on difference
4. Jian Shi jnshi@ucdavis.edu contribution: model diagnose

Introduction

This Document is the first project of Group 7 in STA 207, Winter Quarter 2020.

The repository in Github is on https://github.com/yhli097/STA207-Project1-yhli097.git

There is also information in code like graph. Try PDF file if html file displays incompetely.

Background

Tennesses Student/Teacher Achievement Ratio study (Project STAR) was conducted in the late 1980s to evaluate the effect of class size on test scores. The study randomly assigned students to small classes, regular classes, and regular classes with a teacher's aide. In order to randomize properly, schools were enrolled only if they had enough studybody to have at least one class of each type. Once the schools were enrolled, students were randomly assigned to the three types of classes, and one teacher was randomly assigned to one class.

Data Description

The `STAR` dataset from the `AER` package which are from the very influential randomized experient STAR, assessing the effect of reducing class size on test scores in the early grades. The dataset contains scaled scores for math and reading from kindergarten to 3rd grade. We will only examine the math scores in 1st grade in this project[1].

The following is the description of `STAR`[2].

STAR is a data frame containing 11,598 observations on 47 variables.

Here are main variables for 1st grade.

Variable	Description
`gender`	factor indicating student's gender.
`ethnicity`	factor indicating student's ethnicity with levels `cauc` (Caucasian), `afam` (African-American), `asian` (Asian), `hispanic` (Hispanic), `amindian` (American-Indian) or `other`.
`birth`	student's birth quarter (of class yearqtr).

Variable	Description
`star1`	factor indicating the STAR class type in 1st grade: `regular`, `small`, or `regular-with-aide`. `NA` indicates that no STAR class was attended.
`read1`	total reading scaled score in 1st grade.
`math1`	total math scaled score in 1st grade.
`lunch1`	factor indicating whether the student qualified for free lunch in 1st grade.
`school1`	factor indicating school type in 1st grade: `inner-city`, `suburban`, `rural` or `urban`.

Question of Interest

Our questions have three parts:

a. How to build a one-way ANOVA model to study the effects of class types on the math scaled scores in 1st grade? Is this model appropriate? How does the model fit the data?

b. Is there any difference in the math scaled score in 1st grade across students in different class types?

c. Can we make any causal statements based on the analysis?

Methods and Results

Explore Data

We choose two columns `star1` and `math1` in `STAR` as our dataset.

Deal with NA

The data has several `NA` in both `star1` and `math1`. The `NA` in `star1` means that no STAR class was attended. We checked when `star1` is `NA`, `math1` is also `NA`, which is not informative. Therefore, we removed the cases where `star1` is `NA`.

After that, there are only 229 `NA` in 6829 math scores in `math1`, which have little influence. So we decided to remove them too.

Summary Statistics

From the pie chart on `star1`, there are almost the same number of cases in different type of classes.

From the box plot on `math1` divided by `star1`, we can see roughly that the quantile of math scores in small class is higher than in regular-with-aide class, which is higher than in regular class. So does the average math score from the summary table.

One-Way ANOVA

In this part our goal is to build a one-way ANOVA model on `math1` by `star1`. The model equation and notation as below:

$$Y_{ij} = \mu_1 + \tau_2 X_{2,ij} + \tau_3 X_{3,ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2), \quad i = 1, 2, 3, j = 1, \cdots, n_i.$$

where $i = 1$ means the class type in 1st grade is regular; $i = 2$ means the class type in 1st grade is small; $i = 3$ means the class type in 1st grade is regular-with-aide.

$n_1 = 2507, n_2 = 1868, n_3 = 2225, n = 6600$.

$X_{2,ij} = 1$ if $i = 2$, otherwise $X_{2,ij} = 0$. $X_{3,ij} = 1$ if $i = 3$, otherwise $X_{3,ij} = 0$.

Y_{ij} denotes the math score in 1st grade of the j-th experimental unit in the i-th class type, $i = 1, 2, 3, j = 1, \cdots, n_i$.

μ_i means the population mean of math score in i-th type class in 1st grade, $i = 1, 2, 3$.

$\tau_i = \mu_i - \mu_1$ means the difference in population mean of math score between i-th type and first type in 1st grade, $i = 2, 3$.

ϵ_{ij} is independent and identically distributed normal random variable with 0 mean and σ^2 variance.

Model Assumption:

(a) Response variable residuals are normally distributed.

(b) Variances of populations are equal.

(c) Responses for a given group are independent and identically distributed normal random variables.

All of the assumptions are necessary, because F-test and related procedures are pretty robust to the normality and equal variance assumptions, and pairwise comparisons could be substantially affected by unequal variances. Moreover, non-independence can have serious side effects and is hard to correct. So it is important to apply randomization whenever necessary.

Box-Cox Transformation

Before we fit the model, we need to ensure that model is appropriate on this dataset, that is, the response variable satisfies the assumptions of our model. In other words, we will check the normality and equal variance of the response varibales.

We first draw the density plot and Q-Q plot to check the normality of `math1` ,and we find the distribution of `math1` is right-skewed. So we use Box-Cox method on `math1` and from the result, we need to make a log-transformation on `math1`.

After the logarithmic transformation, the distribution shows more normal-like. Then we calculate the variance of log math grade in 1st grade of each class type. The result shows that they are very small and nearly equal to each other.

On the other hand, we can check $\mathrm{S}_i \propto \overline{Y}_i$. From commonly used transformations we choose $Y^* = \log Y$.

Therefore, it is appropiate to build our model on this dataset after log-transformation on `math1`.

Fitted Model

From the result in R, the fitted model we get is:

$$\log \hat{Y}_{ij} = 6.2608 + 0.0250 X_{2,ij} + 0.0081 X_{3,ij}$$

with means when the type is `regular`, the estimate math score is $e^{6.2608} = 523.6377$; when the type is `small`, the estimate math score is $e^{6.2608+0.0250} = 536.8936$; when the type is `regular-with-aide`, the estimate math score is $e^{6.2608+0.0081} = 527.8964$.

The following is a ANOVA table for this model.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean of Squares	F*
`Between treatments`	SSTR = 0.68	2	MSTR = 0.3391	F* = 52.56
`Within treatments`	SSE = 42.55	6597	MSE = 0.0065	
`Total`	SSTO = 43.23	6599		

Model Diagnostic

Recalling the above assumptions, we need to check the normality of the residuals to make sure that our model is reasonable. We start to check the assumption that $\mathrm{E}(\epsilon_{ij}) = 0$ and ϵ_{ij} are normally distributed. From the scatter plot of residuals vs fitted values, the residuals are divided into three groups and among each group, these residuals are around the zero, which means that the average residuals almost equals to zero. According to the histogram of residuals and studentized residuals, we can find that the distribution of the residuals of the fitted model approximates to the normal distribution. Besides, the same conclusion can be obtained by checking the Q-Q Plot of the residuals. Therefore, we can confirm that the residuals of the model are normally distributed.

Then, we turn to formal tests of the equality of variances. First, we calculate the variances for each type of class and find that the variances of three types of class are close to each other.

Variances of three types		
regular	small	regular+aide
1734.825	1945.082	1837.493

Then, because the sample sizes of the three types of class are not the same, we choose two formal tests, which are Bartlett test[3] and Levene test[4], to check the equality of model variances.

Bartlett test is used to test the null hypothesis H_0 that all k population variances are equal against the alternative H_a that at least two are different. If there are k samples with sizes n_i and sample variances S_i^2 then Bartlett test statistic is

$\chi^2 = \frac{(N-k)\log(S_p^2) - \sum_{i=1}^{k}(n_i-1)\log(S_i^2)}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(\frac{1}{n_i-1}) - \frac{1}{N-k}\right)}$ where $N = \sum_{i=1}^{k} n_i$ and $S_p^2 = \frac{1}{N-k}\sum_i (n_i-1)S_i^2$ are the pooled estimate for the variance.

The test statistic has approximately a χ^2_{k-1} distribution. Thus the null hypothesis is rejected if $\chi^2 > \chi^2_{k-1}(\alpha)$ (where $\chi^2_{k-1}(\alpha)$ is the upper tail critical value for the χ^2_{k-1} distribution).

Bartlett test of homogeneity of variances		
$\chi^2 = 2.3004$	df = 2	p-value = 0.3166

The null hypothesis H_0 of Levene test is that all k population variances are equal against the alternative H_a that at least two are different. It is equivalent to a one-way between-groups analysis of variance (ANOVA) with the dependent variable being the absolute value of the difference between a score and the mean of the group to which the score belongs (shown below as $Z_{ij} = |Y_{ij} - \bar{Y}_{i\cdot}|$). The test statistic, W, is equivalent to the F statistic that would be produced by such an ANOVA, and is defined as follows:

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - Z_{i\cdot})^2},$$

The test statistic W is approximately F-distributed with $k-1$ and $N-k$ degrees of freedom, so the null hypothesis is rejected if $W > F(k-1, N-k; 1-\alpha)$

Levene Test for Homogeneity of Variance (center = median)			
	Df	F value	Pr(>F)
group	2	1.1836	0.3062
	6597		

From the results of the two tests, both of the P-values are much larger than 0.05, which means that we can not reject the null hypothesis: the variances of the model are equal.

In conclusion, we confirm that our model satisfies the normality assumption that $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$.

Difference among factors

In order to investigate whether there is a difference among three factor level means, We choose to use F-test. Moreover, to investigate comparisons between every two factor level means simultaneously and control the family wise error rate, we use Tukey Procedure and Bonferroni Procedure.

F-test

To test null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ against alternative hypothesis H_a: not all μ_is are equal. We caculate F-statistic: $F^* = \frac{MSTR}{MSE} = 52.17$ and $F(0.95, 2, 6597) = 3.00$. Because $F^* > F(0.95, 2, 6597)$, We can reject the null hypothesis at the significance level 0.05. We can claim that there exists difference among factor level means.

Tukey Procedure

For Tukey Procedure, the largest difference is between small and regular. The null hypothesis is $H_0 : \mu_i = \mu_j$ against alternative hypothesis $H_a : \mu_i \neq \mu_j$. All the p values are less than 0.05, so we reject all null hypothesis, and every two factors are statistically significant.

Bonferroni Procedure

For Bonferroni Procedure, the null hypothesis is $H_0 : \mu_i = \mu_j$ against alternative hypothesis $H_a : \mu_i \neq \mu_j$. All p values of three pairwise comparisons are less than 0.05. We can make a conclusion that every two factors are statistically significant. We get the same result as Tukey Procedure.

Causal Statements

We will investigate whether there are any causal statements of math grade in three different class types. Instead of discussing all of them simultaneously, our strategy is to make pairwise tests among them. Therefore, we will take three tests including `small` class with `regular` class, `small` class with `regular-with-aid` class and `regular` class with `regular-with-aid` class.

Assumptions

In our causal inference, potential outcome is the `math1`, which represents the math grade in 1st grade. There are three assumptions:

(a) Causal order can't be reversed.

(b) No spillover effect.

(c) Same version of treatment.

(d) Potential outcomes follow a normal distribution.

All of them are satisfied, then, causal inference can start.

Causal Effect Test

Null hypothese $H_0 : Y_i(1) = Y_i(0)$ for all $i = 1, 2, 3, \cdots, N$, $\leftrightarrow$ alternative hypothese H_a : not for all i, $Y_i(1) = Y_i(0)$,
where $Y_i(1)$ represents the i-th potential outcome when $Z_i = 1$; $Y_i(0)$ represents the i-th potential outcome when $Z_i = 0$. $Z_i = 1$ and $Z_i = 0$ stand for different treatments and N represents the total number of potential outcomes.

Our estimand is $\tau = \overline{Y(1)} - \overline{Y(0)}$. The unbiased estimator of this estimand is $\hat{\tau} = \frac{\sum_i^N 1\{Z_i=1\}Y_i(1)}{N_1} - \frac{\sum_i^N 1\{Z_i=0\}Y_i(0)}{N_0}$, where N_1 represnets the number of observed outcomes when $Z_i = 1$ and N_0 represents the number of observed outcomes when $Z_i = 0$. It also has $N_1 + N_0 = N$.

The true variance of $\hat{\tau}$ is $\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S^2}{N}$, where $S_1^2 = \frac{1}{N-1}\sum_{i=1}^N (Y_i(1) - \bar{Y}(1))^2$, $S_0^2 = \frac{1}{N-1}\sum_{i=1}^N (Y_i(0) - \bar{Y}(0))^2$, $S^2 = \frac{1}{N-1}\sum_{i=1}^N ((Y_i(1) - Y_i(0)) - (\bar{Y}(1) - \bar{Y}(0)))^2$.

It is worthy to note that, S^2 equals to zero if the treatment effect is constant for all i. Therefore, our estimator of variance is $\widehat{var}(\hat{\tau}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$.

Then we get a statistic $t^* = \frac{\hat{\tau}}{\widehat{var}(\hat{\tau})}$. And it approximately follows the t-distribution. For the degree of freedom, we take the Welch's approximate t solution to estimate it [5]. The degree of freedom υ can be regarded as $\frac{(\gamma_1+\gamma_2)^2}{\gamma_1^2/(n_1-1)+\gamma_2^2/(n_2-1)}$, where $\gamma_i = \frac{\sigma_i^2}{n_i}$. For the significance level α, because we do pairwise tests, let $\alpha = 1 - 0.05/3$ as the significance level to control the family wise error rate. Compared with two values, we can make a conclusion whether there is a casual effect.

Test on Data

Test 1 Treatments: `small` class type, `regular` class type.

$t^* = \frac{\hat{\tau}}{\widehat{var}(\hat{\tau})} = 8.077159$, degree of freedom $\upsilon \approx 3892$. And $t^* > t(1 - 0.05/3, 3892)$. We reject the null hypothesis and make a claim that there is a casual effect, that is, compared with `regular` class, `small` class contributes to higher math score in 1st grade.

Test 2 Treatments: `small` class type, `regular-with-aide` class type.

$t^* = \frac{\hat{\tau}}{\widehat{var}(\hat{\tau})} = 4.820985$, degree of freedom $\upsilon \approx 3928$. And $t^* > t(1 - 0.05/3, 3928)$. We reject the null hypothesis and make a claim that there is a casual effect, that is, compared with `regular-with-aide` class, `small` class contributes to higher math score in 1st grade.

Test 3 Treatments: `regular` class type, `regular-with-aide` class type.
$t^* = \frac{\hat{\tau}}{\widehat{var}(\hat{\tau})} = 3.294502$, degree of freedom $\upsilon \approx 4628$. And $t^* > t(1 - 0.05/3, 4628)$. We reject the null hypothesis and make a claim that there is a casual effect, that is, compared with `regular-with-aide` class, `regular` class contributes to higher math score in 1st grade.

We can make a conclusion that there are causal effects among pairwise comparisons.

Conclusions

Select Model

We select `star1` and `math1` from `STAR` and remove all `NA` cases.

Then we hold Box-Cox method and decide the logarithmic transformation on 'math1'.

So we get the one-way ANOVA model:

$$\log Y_{ij} = \mu_1 + \tau_2 X_{2,ij} + \tau_3 X_{3,ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2), \quad i = 1, 2, 3, j = 1, \cdots, n_i.$$

Fit Model

The fitted model is

$$\log \hat{Y}_{ij} = 6.2608 + 0.0250 X_{2,ij} + 0.0081 X_{3,ij}$$

The estimation of the mean of total math score in `regualr`, `small`, `regular-with-aide` class in 1st grade is $523.6377, 536.8936, 527.8964$ respectively.

The model diagnose shows this model satisfies the normality and homoscedasticity assumption.

Hold Test

F test result rejects the null hypothesis and shows that the means in different class are not the same. So there is a treatment effect for class type on math score in 1st grade.

From Tukey and Bonferroni procedure, every pairwise means in different class are not the same. So the three types of class have different treatment effect on math score in 1st grade from each other.

Causal Statement

By causal effect test, we reject all three null hypotheses and get the conclusion that there are causal effects among pairwaise comparisons. What is more, `small` class contributes to higher math score in 1st grade than `regular-with-aide` class, which contributes to higher mathe score in 1st grade than `regual` class.

References

[1] https://chenshizhe.github.io/STA207W2020/ch-proj.html#project-1-project-star-i

[2] https://cran.r-project.org/web/packages/AER/AER.pdf

[3] https://en.wikipedia.org/wiki/Bartlett%27s_test

[4] https://en.wikipedia.org/wiki/Levene%27s_test

[5] https://en.m.wikipedia.org/wiki/Behrens%E2%80%93Fisher_problem

Code and Output

Step1 Read Data

```
#install.packages("AER")
library(AER)
data("STAR")
```

Step2 Explore Data

We will only examine the math scores in 1st grade in this project.

```
data <- data.frame(star1 = STAR$star1, math1 = STAR$math1)

sapply(data,class)
```

```
##     star1     math1
##  "factor" "integer"
```

```
sapply(data,summary)
```

```
## $star1
##      regular        small regular+aide         NA's
##         2584         1925         2320         4769
##
## $math1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   404.0   500.0   529.0   530.5   557.0   676.0    4998
```

```
data.star1.na <- data[is.na(data$star1),]
all(is.na(data.star1.na$math1))
```

```
## [1] TRUE
```

Which shows that the math score has not been recorded if class type is not recorded. So we can remove the data where star1 is NA.

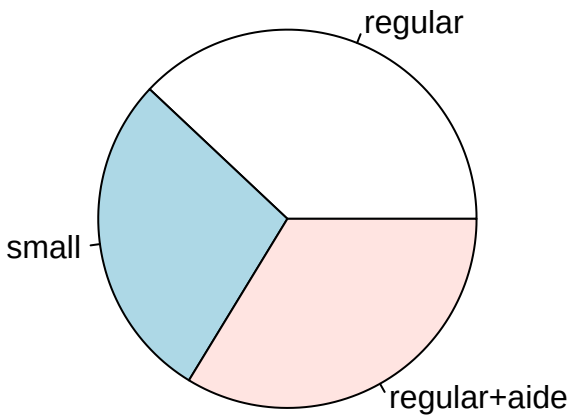
One of the way to deal with NA in math1 is to remove them

```
data_remove_na <- na.omit(data[-is.na(data$star1),])
```

```
table(data_remove_na$star1)
```

```
##
##      regular        small regular+aide
##         2507         1868         2225
```

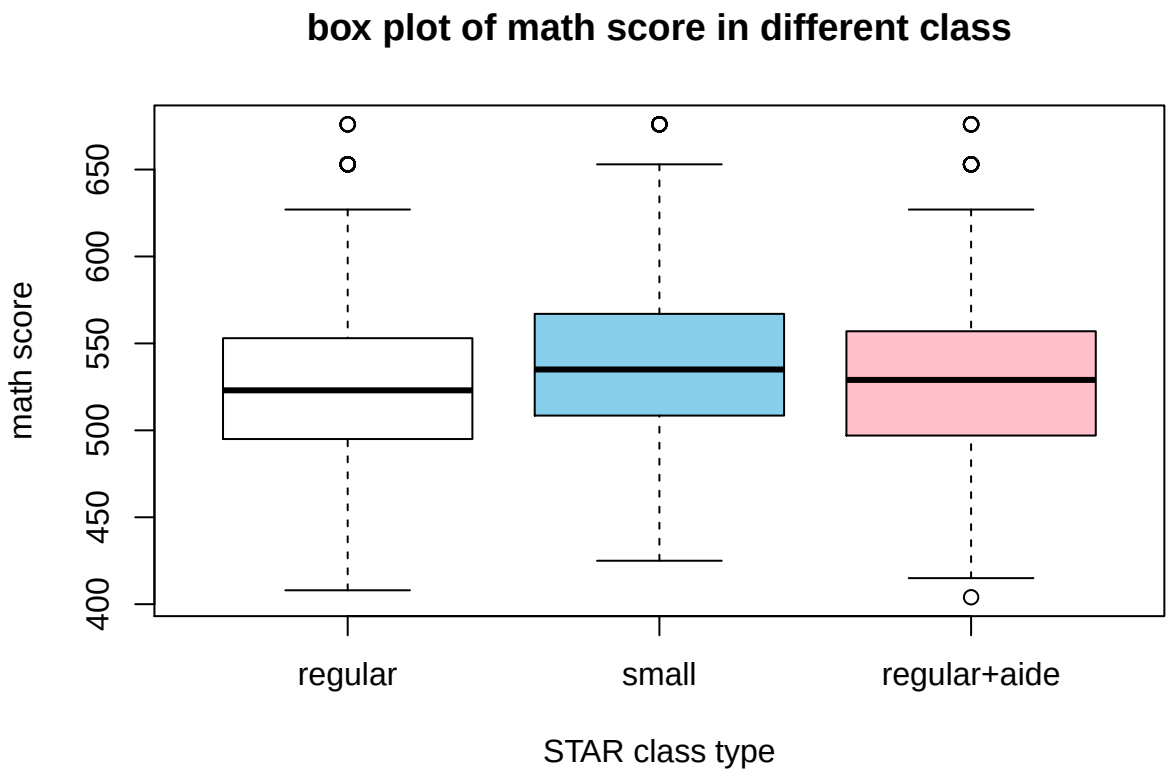
```
pie(table(data_remove_na$star1),main = "pie chart of STAR class type")
```




```
tapply(data_remove_na$math1, data_remove_na$star1,summary)
```

```
## $regular
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   408.0   495.0   523.0   525.3   553.0   676.0
##
## $small
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   425.0   509.2   535.0   538.7   567.0   676.0
##
## $`regular+aide`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   404.0   497.0   529.0   529.6   557.0   676.0
```

```
boxplot(data$math1~data$star1,main = "box plot of math score in different class",
        xlab = "STAR class type", ylab = "math score", col = c("white", "skyblue", "pink"))
```



Step3 One Way ANOVA Model

$$Y_{ij} = \mu_1 + \tau_2 X_{2,ij} + \tau_3 X_{3,ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2), \quad i = 1, 2, 3, j = 1, \cdots, n_i.$$

where $i = 1$ means the class type in 1st grade is regular; $i = 2$ means the class type in 1st grade is small; $i = 3$ means the class type in 1st grade is regular-with-aide.

From the table in step2, $n_1 = 2507, n_2 = 1868, n_3 = 2225, n = 6600$.

$X_{2,ij} = 1$ if $i = 2$, otherwise $X_{2,ij} = 0$. $X_{3,ij} = 1$ if $i = 3$, otherwise $X_{3,ij} = 0$.

Y_{ij} denotes the math grade in 1st grade of the j-th experimental unit in the i-th class type.

μ_i means the population mean of the i-th type class in 1st grade, $i = 1, 2, 3$.

$\tau_i = \mu_i - \mu_1$ means the difference in population mean between i-th type and first type in 1st grade, $i = 2, 3$.

ϵ_{ij} is independent and identically distributed normal random variable with 0 mean and σ^2 variance under normal assumption.

Model Assumption
(a) Response variable residuals are normally distributed.
(b) Variances of populations are equal.
(c) Responses for a given group are independent and identically distributed normal random variables.
All of the assumptions are necessary, because F-test and related procedures are pretty robust to the normality and equal variance assumptions, and pairwise comparisons could be substantially affected by unequal variances. Moreover, non-independence can have serious side effects and is hard to correct. So it is important to apply randomization whenever necessary.

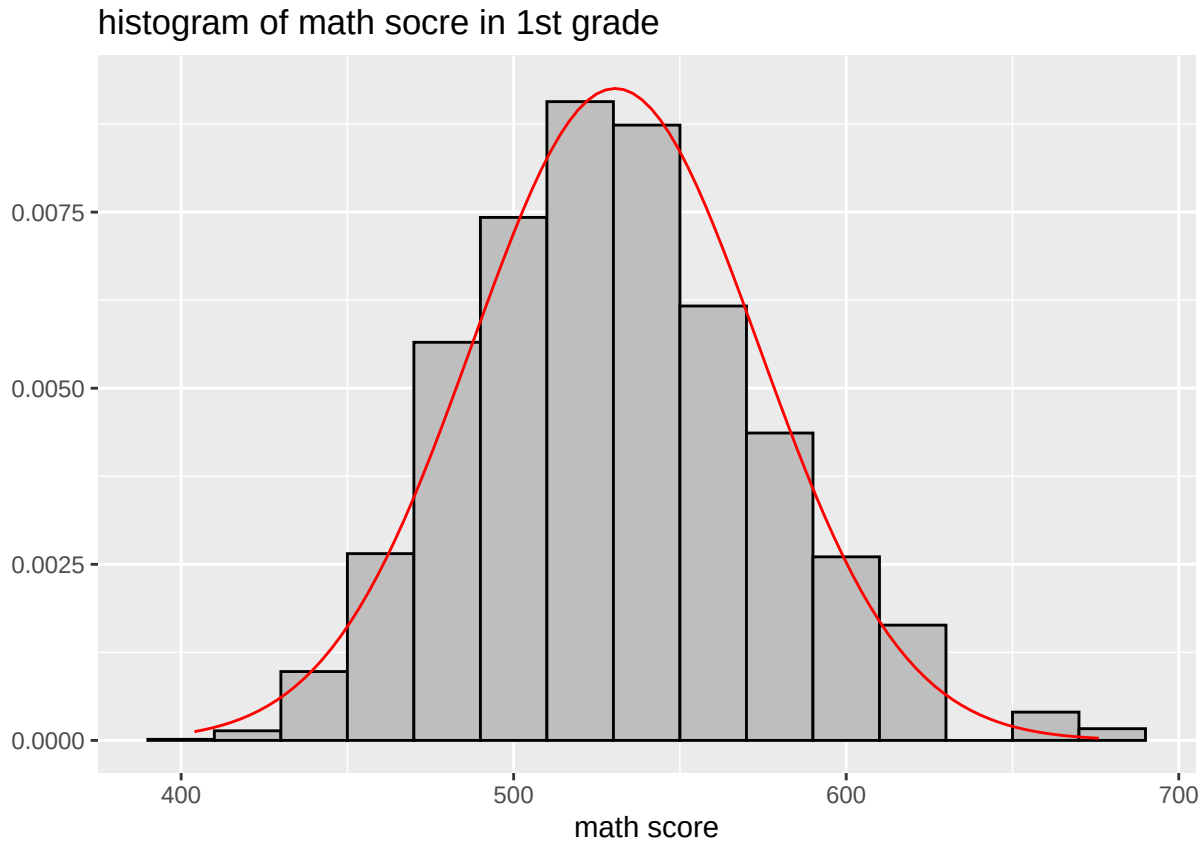
Step4 Appropriate

Before we fit the model, we need to ensure that model is appropriate on this dataset, that is, the response variable satisfies the assumptions of our model. In other words, we will check the normality and equal variance of the response varibales.

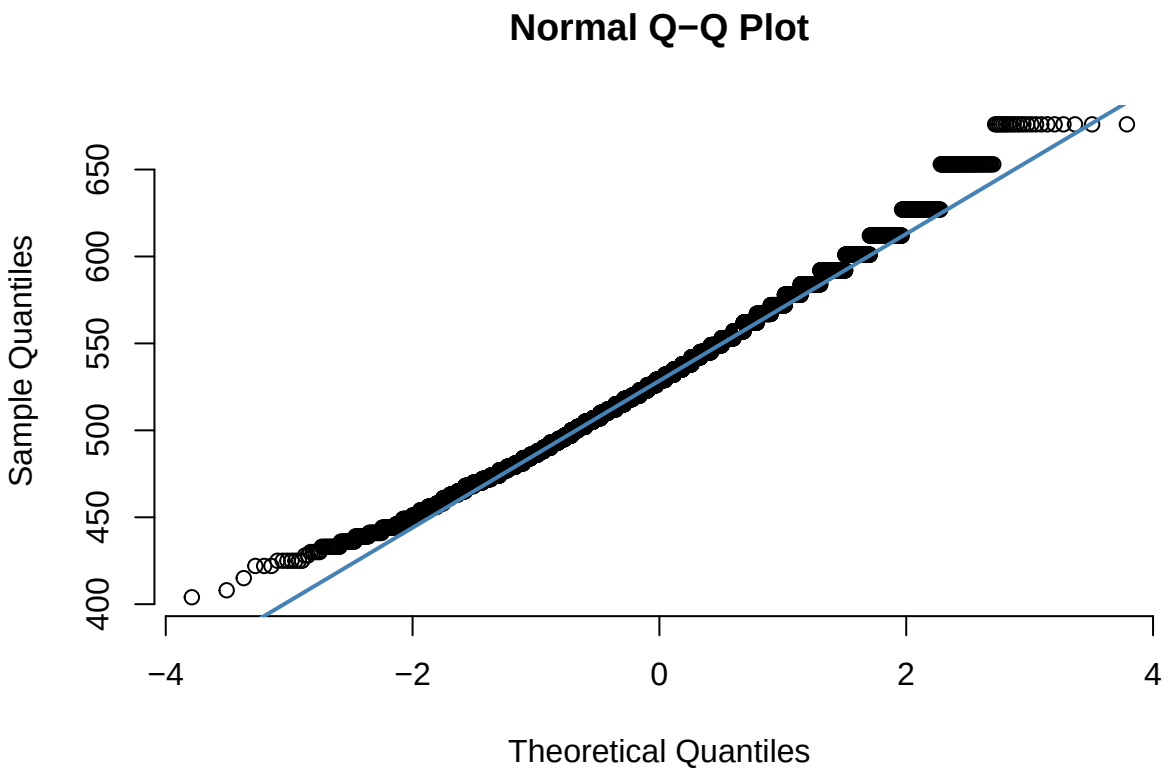
We first make a density plot and a Q-Q plot to check the normality of `math1`.

```
library(ggplot2)
x <- seq(404, 676, length.out=100)
df <- with(data_remove_na, data.frame(x = x, y = dnorm(x, mean(math1), sd(math1))))

ggplot(data_remove_na, aes(x=math1, y = ..density..)) +
  geom_histogram(binwidth = 20, fill = "grey", color = "black") +
  geom_line(data = df, aes(x = x, y = y), color = "red") +
  labs(x="math score",y="",title = "histogram of math socre in 1st grade")
```




```
qqnorm(data_remove_na$math1, pch = 1, frame = FALSE)
qqline(data_remove_na$math1, col = "steelblue", lwd = 2)
```

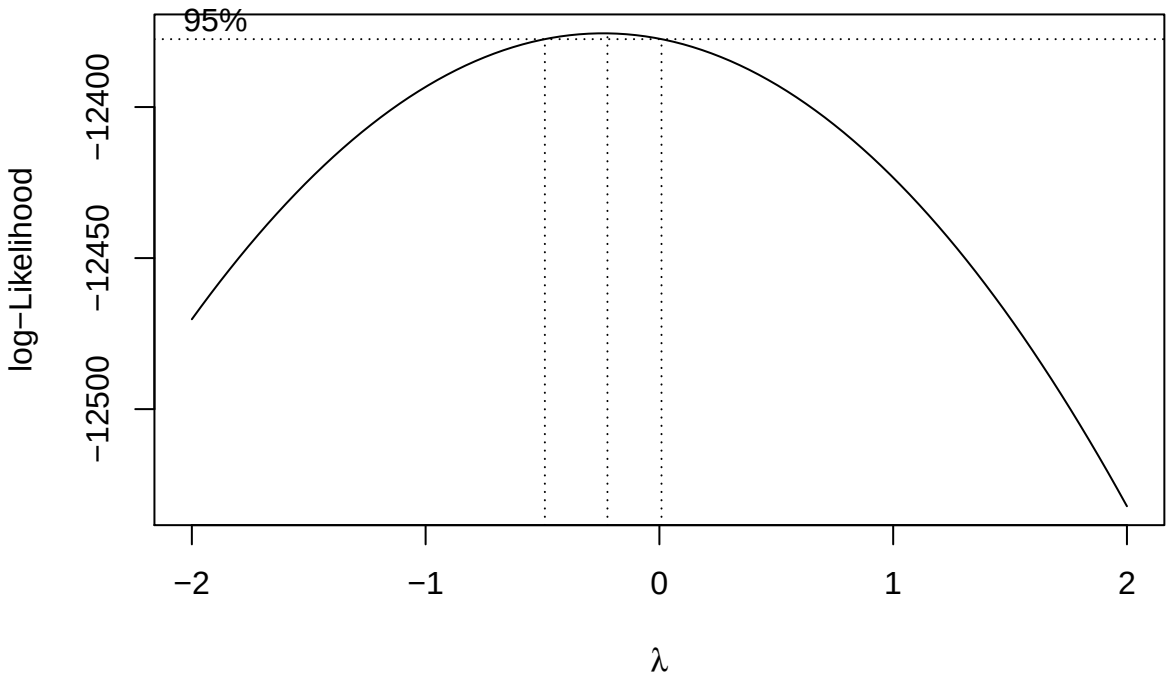


The histogram shows that it seems normal distribution.

The Q-Q plot shows the the distribution of math score is right-skewed.

So we use Box-Cox method to make a transformation on `math1`.

```
library(MASS)
boxcox(math1 ~ star1 , data = data_remove_na)
```




```
data_mean <- tapply(data_remove_na$math1,data_remove_na$star1,mean)
data_sd <- tapply(data_remove_na$math1,data_remove_na$star1,sd)
data_sd/data_mean
```

```
##      regular        small regular+aide
##   0.07929423   0.08187285   0.08093645
```

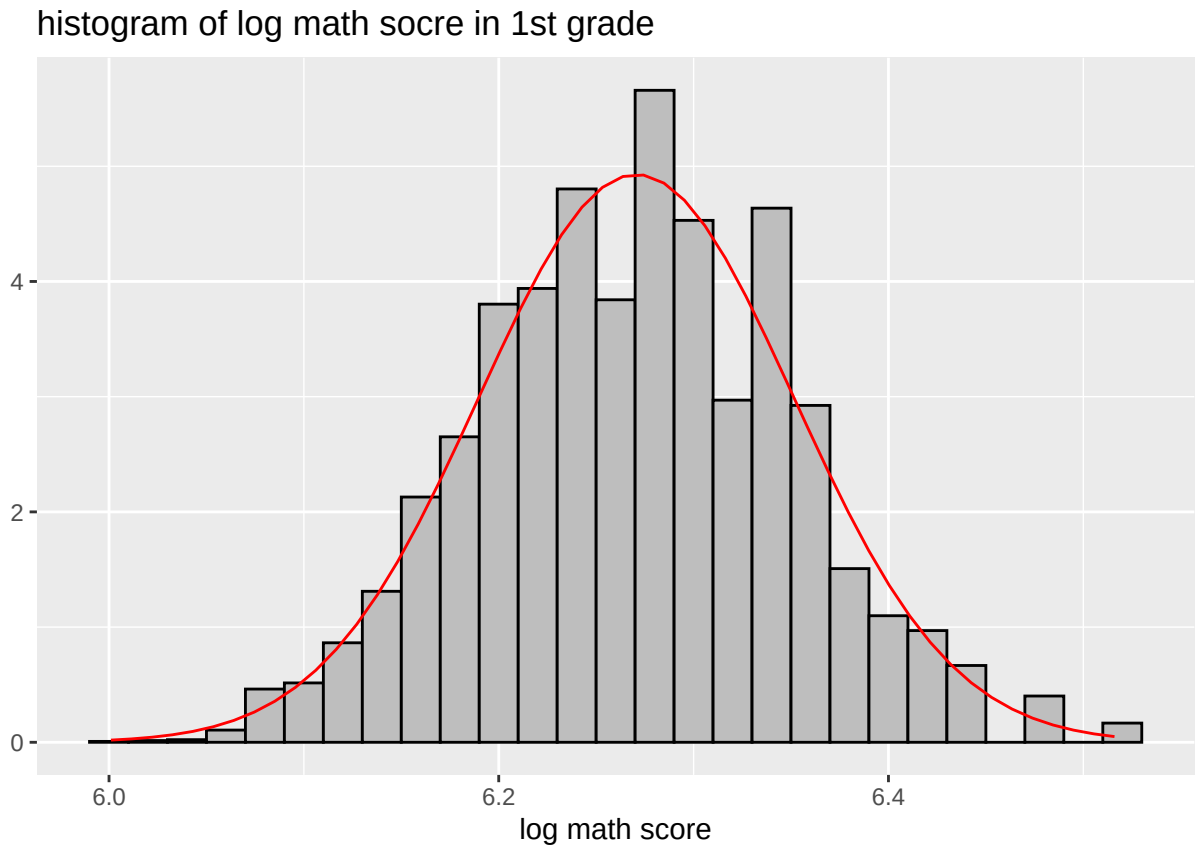
It indicates that we need make a log-transformation for `math1`.

```
summary(log(data_remove_na$math1))
```

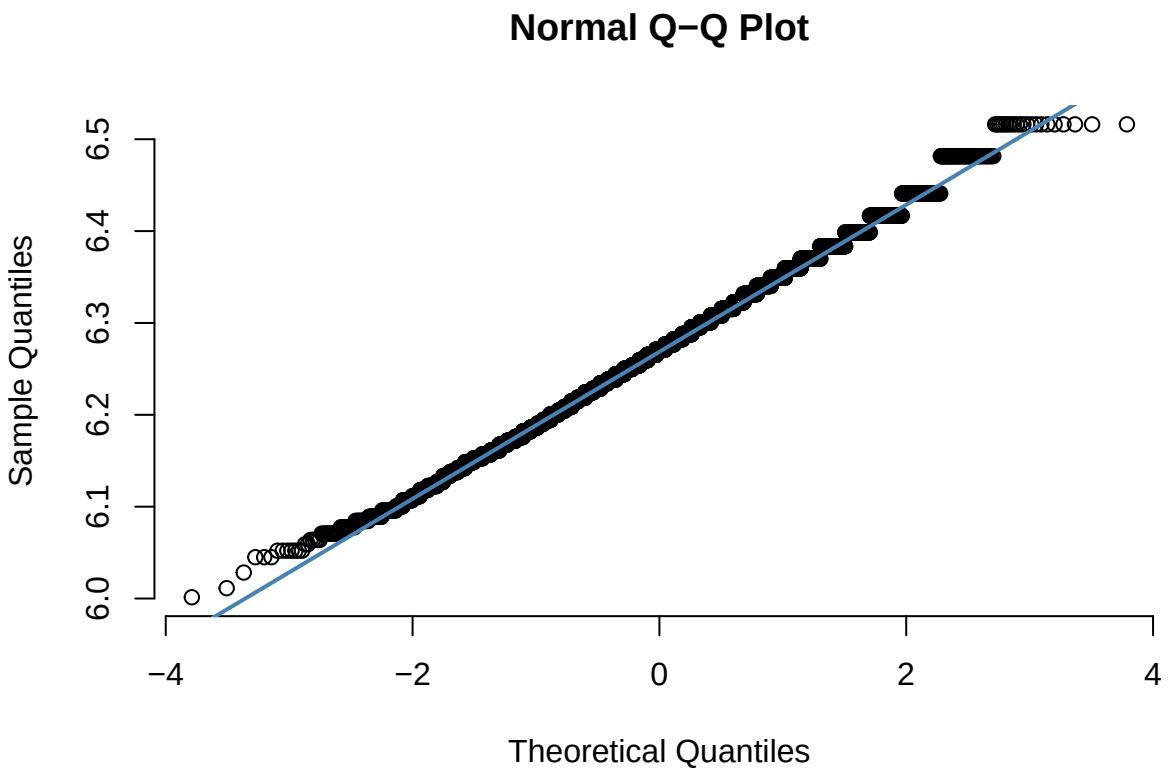
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.001   6.215   6.271   6.271   6.323   6.516
```

```
x <- seq(6.001, 6.516, length.out=50)
df <- with(data_remove_na, data.frame(x = x, y = dnorm(x, mean(log(math1)), sd(log(math1)))))

ggplot(data_remove_na, aes(x=log(math1), y = ..density..)) +
  geom_histogram(binwidth = 0.02, fill = "grey", color = "black") +
  geom_line(data = df, aes(x = x, y = y), color = "red") +
  labs(x="log math score",y="",title = "histogram of log math socre in 1st grade")
```




```
qqnorm(log(data_remove_na$math1), pch = 1, frame = FALSE)
qqline(log(data_remove_na$math1), col = "steelblue", lwd = 2)
```



The graph shows the distribution of log math score in 1st grade is Normal-like.

Then we calculate the variance of math grade in 1st grade of each class type.

```
library(tidyverse)
```

```
data_remove_na %>%
  group_by(star1) %>%
  summarize(var_math1 = var(log(math1), na.rm = T))
```

```
## # A tibble: 3 x 2
##   star1        var_math1
##   <fct>            <dbl>
## 1 regular        0.00625
## 2 small          0.00666
## 3 regular+aide   0.00650
```

The result shows that they are very small and nearly equal to each other. Therefore, it is appropiate to build our model on this dataset.

Step5 Fit Model

```
anova.fit<- aov(log(math1)~star1,data=data_remove_na)
summary(anova.fit)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## star1          2   0.68  0.3391   52.56 <2e-16 ***
```

```
## Residuals   6597  42.55  0.0065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova.fit$coefficients
```

```
##       (Intercept)        star1small star1regular+aide
##        6.26079457        0.02499081        0.00812069
```

From the result, the fitted model we get is:

$$\log \hat{Y}_{ij} = 6.2608 + 0.0250 X_{2,ij} + 0.0081 X_{3,ij}$$

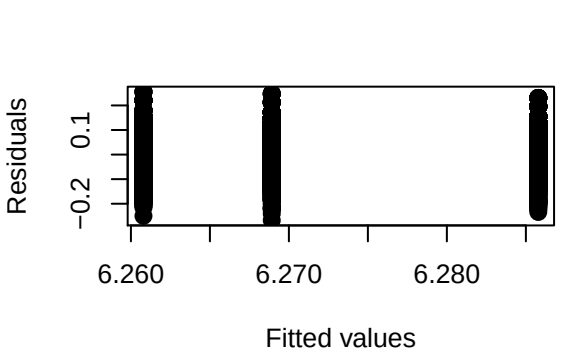
with means when the type is regular, the estimate math score is $e^{6.2608} = 523.6377$; when the type is small, the estimate math score is $e^{6.2608+0.0250} = 536.8936$; when the type is regular-with-aide, the estimate math score is $e^{6.2608+0.0081} = 527.8964$.

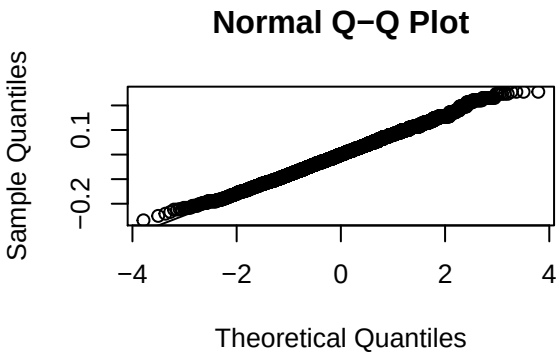
Step6 Model Diagnostic Analysis and Sensitivity Analysis

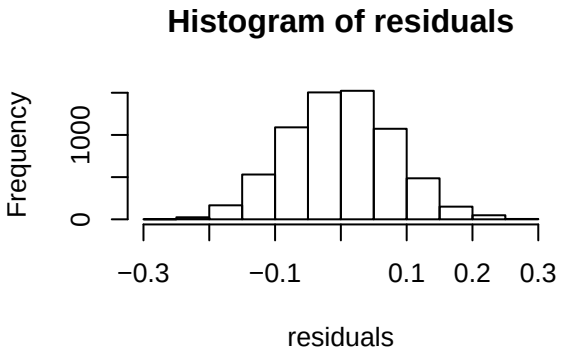
Recalling above assumptions, there are three things we need to check: normality, equal variance and independence.

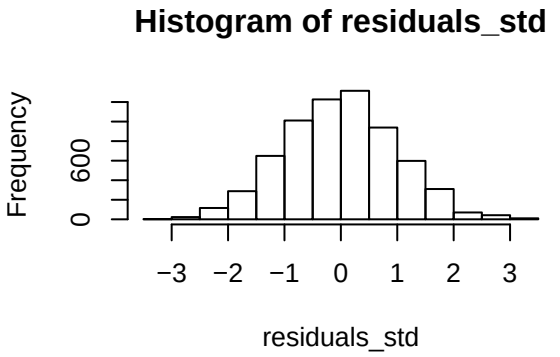
By Q-Q plot, we can check normality. And by residuals vs fitted value plot, we can check equal variance.

```
par(mfrow=c(2,2))
residuals <- anova.fit$residuals
##Plot the residuals (or the other two versions) against fitted values
plot(anova.fit$fitted.values, anova.fit$residuals,
     type = "p",pch=16,cex=1.5,xlab="Fitted values",ylab="Residuals")
#QQplot
qqnorm(residuals);qqline(residuals)
#residuals
hist(residuals)
#studentized residuals
residuals_std <- rstudent(anova.fit)
hist(residuals_std)
```









From the scatterplot of residuals vs fitted valuues, the residuals are divided into three groups and among each group, these residuals are around the zero, which means that the average residuals are almost equal to zero.

According to the histogram of residuals and studentized residuals, we can find that the distribution of the residuals of the fitted model approximates to the normal distribution. Besides, the same conclusion can be obtained by checking the Q-Q Plot of the residuals. Therefore, we can confirm that the residuals of the model are normally distributed.

We now turn to formal tests of the equality of variances. First, we calculate the varianves for each type of class and find that the variances of three types of class are close to each other.

```
# Calculate the variances for each group:
(vars = tapply(data_remove_na$math1,data_remove_na$star1,var))
```

```
##      regular        small regular+aide
##     1734.825     1945.082     1837.493
```

Then, because the sample sizes of the three types of class are not same, we choose two formal tests, which are Bartlett test and Levene test, to check the equality of model variances.

```
data_remove_na$residuals <- anova.fit$residuals
#bartlett test
bartlett.test(residuals ~ star1, data = data_remove_na)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residuals by star1
## Bartlett's K-squared = 2.3004, df = 2, p-value = 0.3166
```

```r
#levene test
leveneTest(residuals ~ star1, data = data_remove_na)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    2  1.1836 0.3062
##       6597
```

From the two tests, both of the P-values are much larger than 0.05, which means that we can not reject the null hyperthesis: the variances of the model are equal.

In conclusion, we confirm that our model satisifies the normality assumption.

In order to test the sensitivity of our model, we decide to relax the assumption of our model. To be specific, we want to figure out that whether the influence of class size still exists even if the data is not normally distributed. Thus, we conduct the nonparameteric tests as follows, which are the rank test and Krusal-Wallis test.

```r
#rank test
data_remove_na$rank <- rank(data_remove_na$math1)
summary(aov(rank ~ star1, data = data_remove_na))
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## star1          2 3.556e+08 177779655   49.72 <2e-16 ***
## Residuals   6597 2.359e+10   3575611
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#kruskal test
kruskal.test(math1 ~ star1, data = data_remove_na)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  math1 by star1
## Kruskal-Wallis chi-squared = 97.993, df = 2, p-value < 2.2e-16
```

The results both show that the math scores of the different types of class are different at 99% confident level. So, even if the data was not normally distributed, there would still be influence of class size. In a word, our one-way anova model is reasonable in this case.

Step7 Hypothesis Test

In this part, in order to investigate whether there is a difference among the different factor level means, we use F-test. Moreover, to investigate comparisons between two factor level means simultaneously and control the family-wise type-I error, we use Tukey's Procedure, Bonferroni's Procedure.

F-test
To test null hypotheis $H_0 : \mu_1 = \mu_2 = \mu_3$ against alternative hypothesis: not all μ_i's are equal. We caculate F-statistic: $F^* = \frac{MSTR}{MSE} = 52.16923$ and $F(0.95, 2, 6597) = 2.997093$. Because $F^* > F(0.95, 2, 6597)$, We can thus reject the null hypothesis at the significance level 0.05. We can claim that there exists difference among factor level means.

Tukey's Procedure

```r
TukeyHSD(anova.fit)
```

```
##   Tukey multiple comparisons of means
```

```
##     95% family-wise confidence level
##
## Fit: aov(formula = log(math1) ~ star1, data = data_remove_na)
##
## $star1
##                              diff          lwr         upr     p adj
## small-regular          0.02499081  0.019236297  0.03074533 0.0000000
## regular+aide-regular   0.00812069  0.002637091  0.01360429 0.0015104
## regular+aide-small    -0.01687012 -0.022778294 -0.01096196 0.0000000
```

There are three pairwise comparisons of factor levels means. And from the result, we can see that all of them should be declared as being different.

Bonferroni's Procedure

```
pairwise.t.test(log(data_remove_na$math1),data_remove_na$star1,p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  log(data_remove_na$math1) and data_remove_na$star1
##
##              regular small
## small        < 2e-16 -
## regular+aide 0.0016  7.1e-11
##
## P value adjustment method: bonferroni
```

According to the result, all of p-value of three pairwise comparison is lesser than 0.05. We can make a conclusion that all of factor level means comparisions are different.