Project1

Yahui Li

2020/1/13

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.

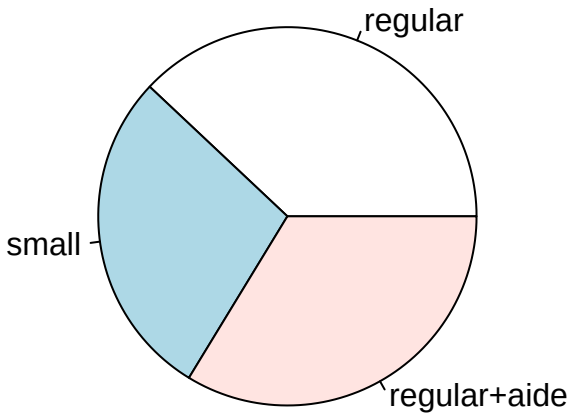
One of the way to deal with NA in math1 is to remove them

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

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

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

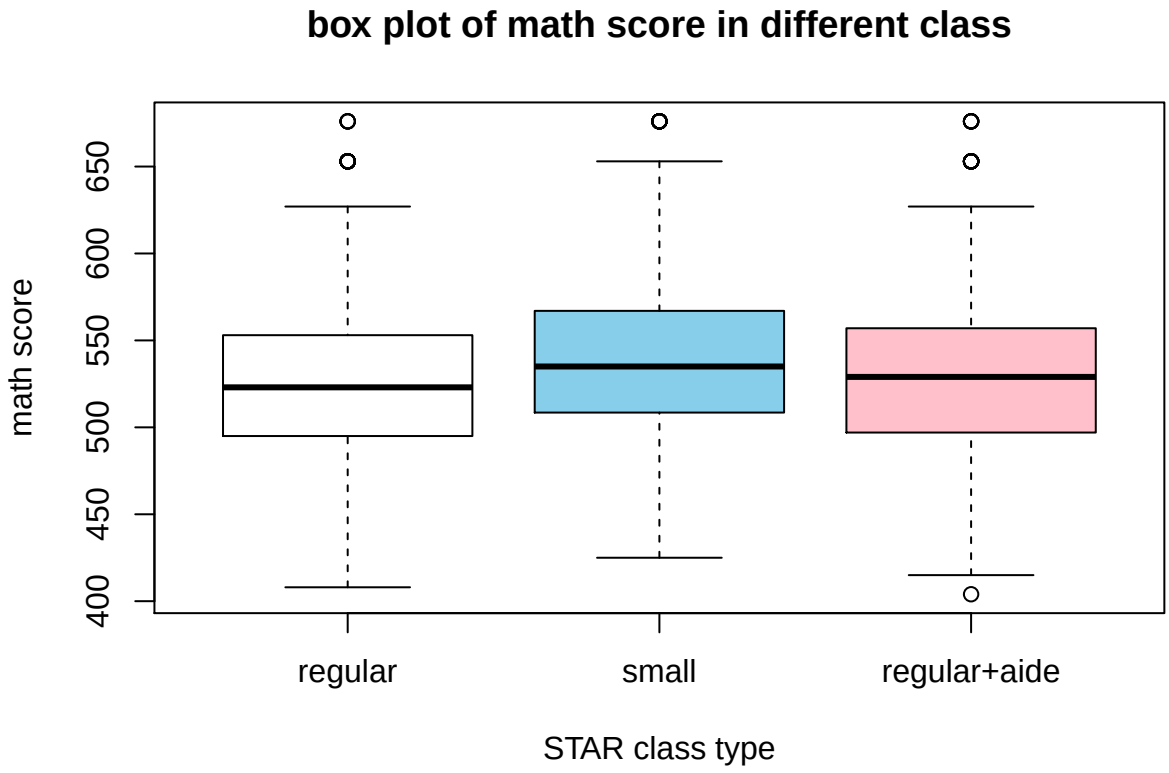
```r
pie(table(data_remove_na$star1),main = "pie chart of STAR class type")
```




```r
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
```

```r
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"))
```



From the result,

for mean, small > regular+aide > regular;

for all quantile information, small > regular+aide > regular;

for min, small > other two; For max, they are the same.

Something interesting: there are only some certain scores like 601 612 627 653 676.

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 1 st 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 $X_{3,ij} = 0$.

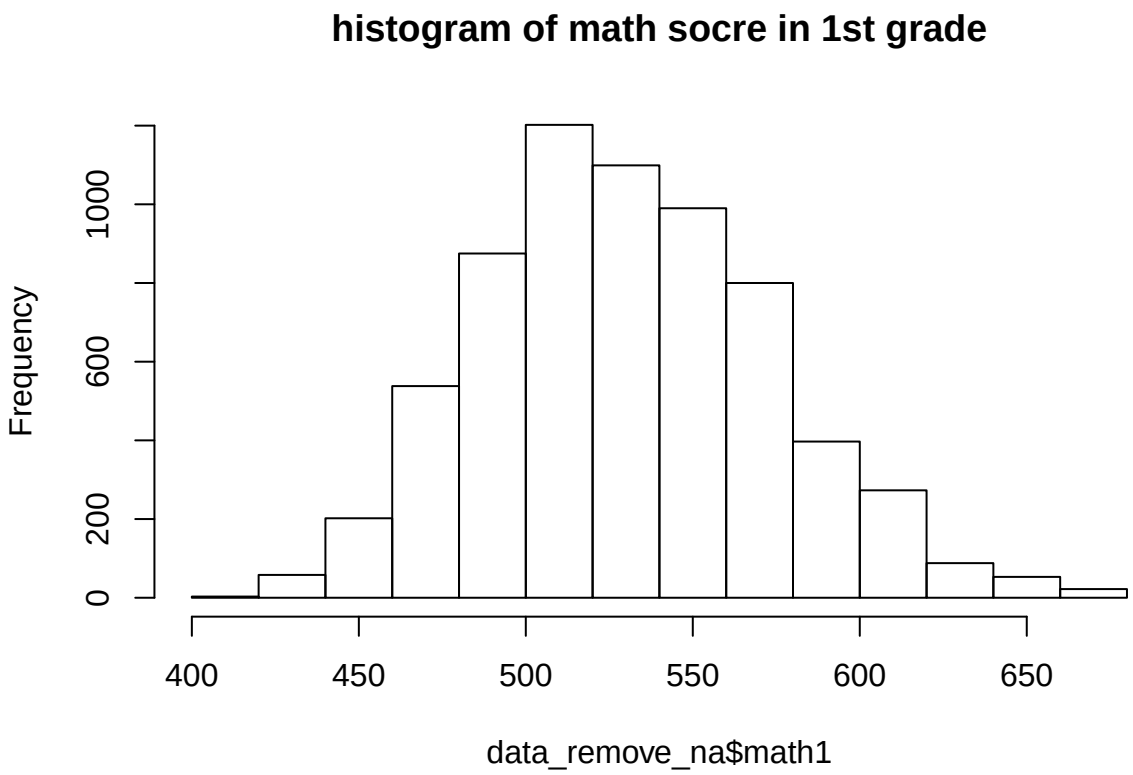
μ_i means the population mean of the i-th type class in 1st grade, $i = 1, 2, 3$.

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

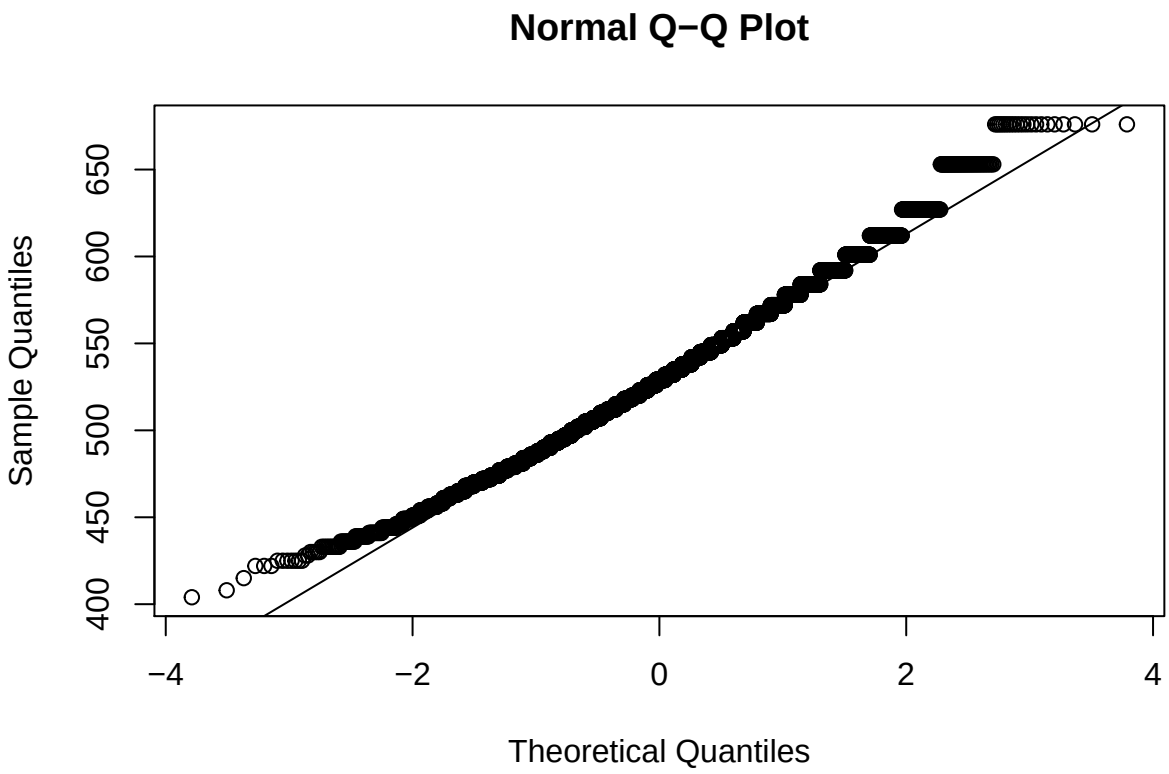
ϵ_{ij} is the random variable about error which is Normal with 0 mean and σ^2 variance under assumption.

Step4 Appropriate

```
hist(data_remove_na$math1, main = "histogram of math socre in 1st grade")
```



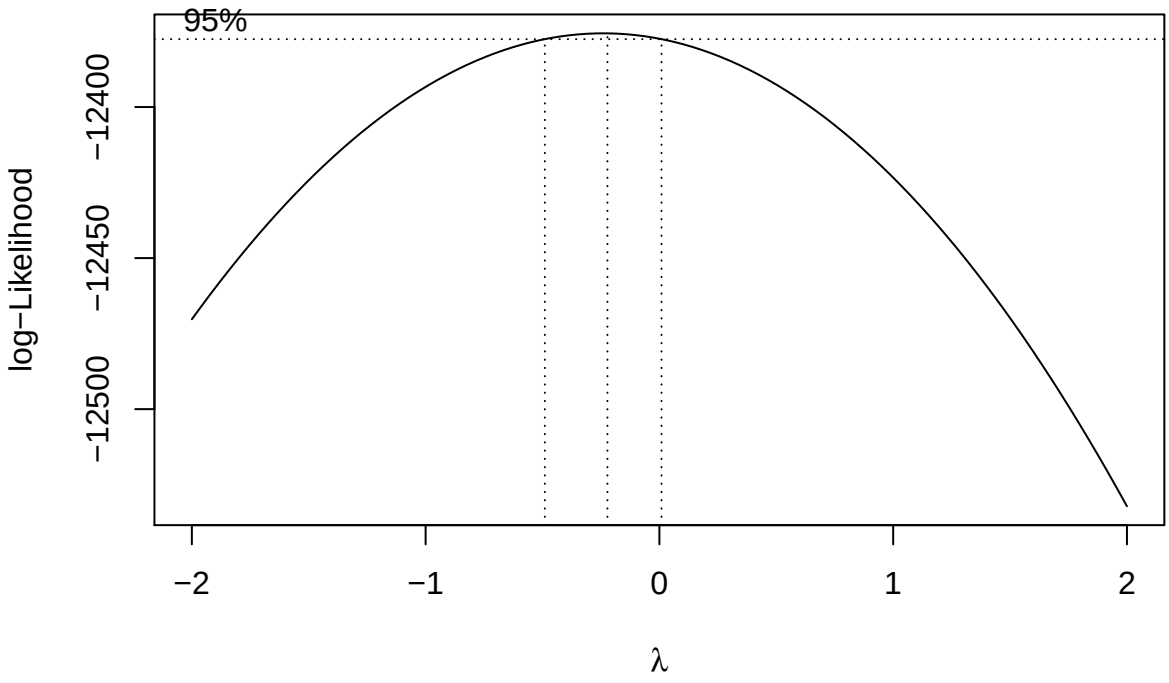

```
qqnorm(data_remove_na$math1)
qqline(data_remove_na$math1)
```



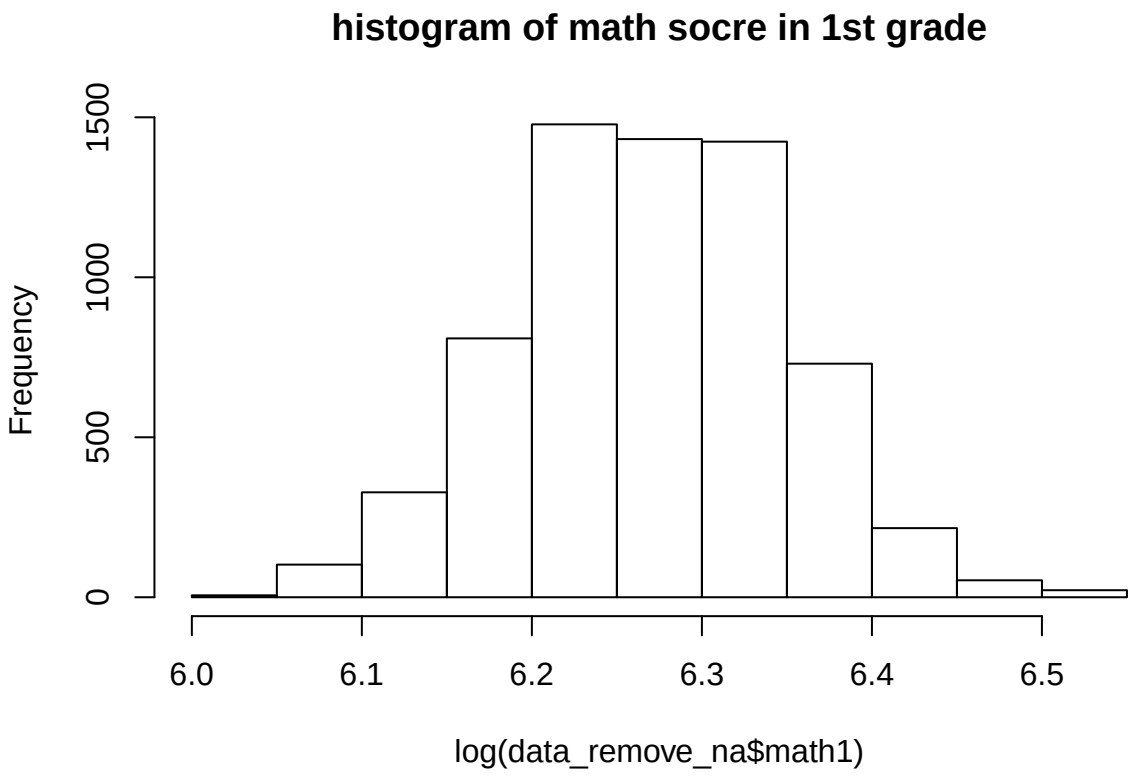
The graph shows the the distribution of math score is right-skewed.

So we use box cox method to make a transformation on math1.

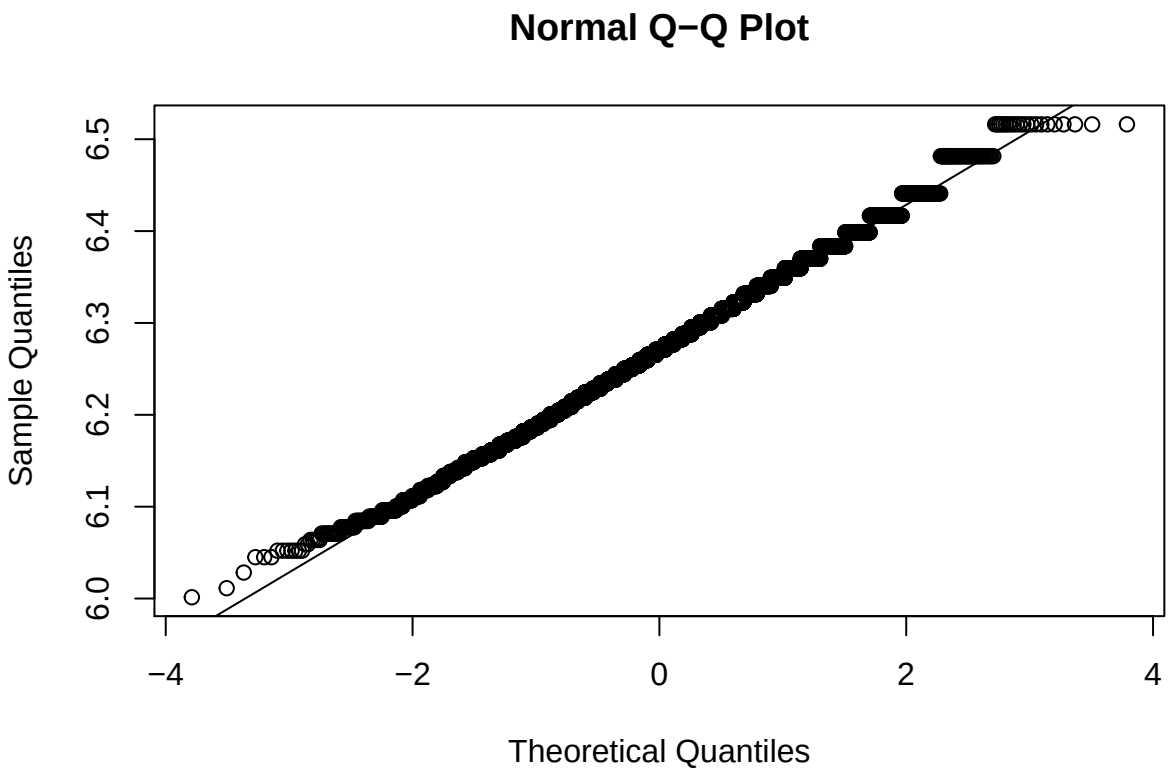
```
library(MASS)
boxcox(math1 ~ star1 , data = data_remove_na)
```




```
hist(log(data_remove_na$math1), main = "histogram of math socre in 1st grade")
```




```
qqnorm(log(data_remove_na$math1))
qqline(log(data_remove_na$math1))
```



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

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 $e^{6.2608+0.0081} = 527.8964$.