Homework 2

Rudy Martinez, Jose Fernandez, Brenda Parnin

10/02/2020

Set Working Directory

```
setwd("/Users/rudymartinez/Desktop/MSDA/Fall 2020/STA 6443_Algorithms I/STAT-Algorithms-1/Week 4/HW2")
```

Read Files

```
heart = read.csv("heartbpchol.csv");
heart$BP_Status = as.factor(heart$BP_Status);
heart$Cholesterol = as.numeric(heart$Cholesterol)

bupa = read.csv("bupa.csv");
bupa$drinkgroup = as.factor(bupa$drinkgroup);
bupa$mcv = as.numeric(bupa$mcv);
bupa$alkphos = as.numeric(bupa$alkphos)

psych = read.csv("psych.csv");
psych$sex = as.factor(psych$sex);
psych$rank = as.factor(psych$rank);
psych$salary = as.numeric(psych$salary)

cars_new = read.csv("cars_new.csv");
cars_new$type = as.factor(cars_new$type);
cars_new$origin = as.factor(cars_new$origin);
cars_new$cylinders = as.factor(cars_new$cylinders);
cars_new$mpg_highway = as.numeric(cars_new$mpg_highway)
```

Libraries

```
library(DescTools)
library(MASS)
library(car)
```

Exercise 1: Analysis of Variance

The heartbpchol.csv data set contains continuous cholesterol (Cholesterol) and blood pressure status (BP_Status) (category: High/ Normal/ Optimal) for alive patients. For the heartbpchol.csv data set, consider a one-way ANOVA model to identify differences between group cholesterol means. The normality assumption is reasonable, so you can proceed without testing normality.

Exercise 1.A

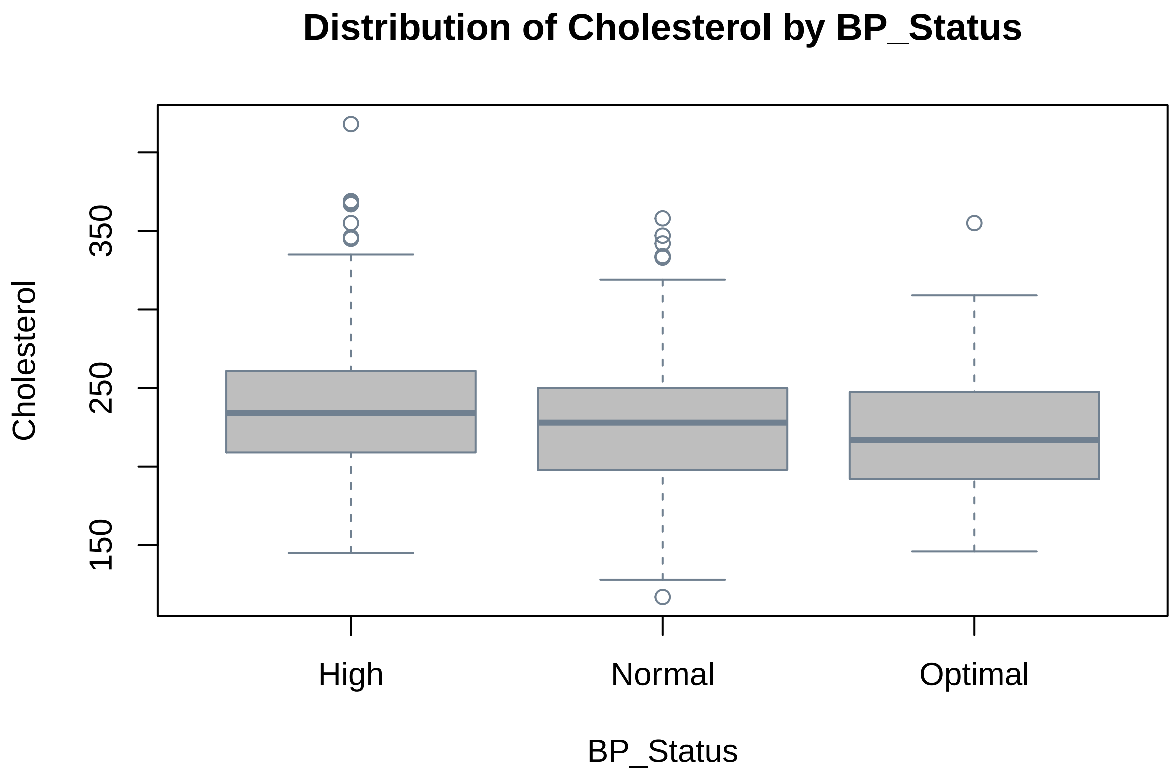
Perform a one-way ANOVA for `Cholesterol` with `BP_Status` as the categorical predictor. Comment on statistical significance of `BP_Status`, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

Data Exploration - Check Balance

```
table(heart$BP_Status)
```

```
##
##    High  Normal Optimal
##     229     245      67
```

```
boxplot(Cholesterol ~ BP_Status, data=heart,
        main="Distribution of Cholesterol by BP_Status",
        xlab = "BP_Status",
        ylab = "Cholesterol",
        col = "Grey",
        border = "slategray",
        horizontal = FALSE
        )
```



Observation: The distribution is **unbalanced**. Each `BP_Status` group has a different number of observations.

Run One-Way ANOVA

```
aov.res_heart= aov(Cholesterol~BP_Status, data=heart)

summary(aov.res_heart) #ANOVA result
```

```
##               Df  Sum Sq Mean Sq F value  Pr(>F)
## BP_Status      2   25211   12605   6.671 0.00137 **
## Residuals    538 1016631    1890
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: The p-value of 0.00137 is below the significance level of 0.05, meaning that we **reject** the null hypothesis. Therefore, `BP_Status` has a significant effect on `Cholesterol` levels (at least one group in `BP_Status` has a different mean of `Cholesterol`).

R-square (variation of response variable explained by `BP_Status`)

```r
lm.res_heart = lm(Cholesterol ~ BP_Status, data = heart)

summary(lm.res_heart)$r.squared
```

```
## [1] 0.02419833
```

Conclusion: 2.4% of the variation of `Cholesterol` can be explained by `BP_Status`.

Check Equal Variance Assumption

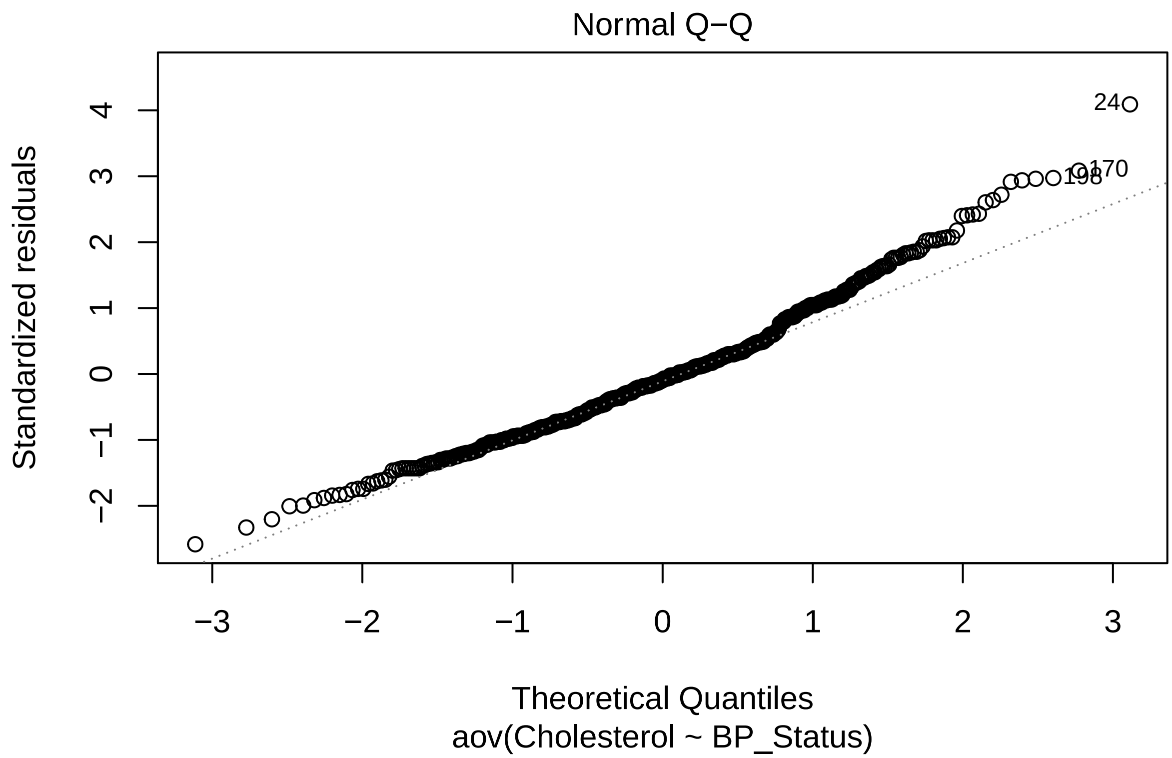
```r
LeveneTest(aov.res_heart)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.1825 0.8332
##       538
```

Conclusion: The p-value is above the significance level of 0.05, meaning that we **can't reject** the null. Therefore, all groups in `BP_Status` have the same variance.

Check Normality

```r
par(mfrow=c(1,1))# diagnostics plot - in one
plot(aov.res_heart, 2)
```



Conclusion: Through analysis of the Q-Q plot, we can see that a normal distribution is reasonable.

Exercise 1.B

Comment on any significantly different cholesterol means as determined by the post-hoc test comparing all pairwise differences. Specifically explain what that tells us about differences in cholesterol levels across blood pressure status groups, like which group has the highest or lowest mean values of `Cholesterol`.

```
ScheffeTest(aov.res_heart)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $BP_Status
##                      diff    lwr.ci    upr.ci   pval
## Normal-High    -11.543481 -21.35092 -1.736038 0.0159 *
## Optimal-High   -18.646679 -33.46702 -3.826341 0.0089 **
## Optimal-Normal  -7.103198 -21.81359  7.607194 0.4958
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments:

- `BP_Status` pairs *Normal-High* and *Optimal-High* have significantly different mean values of (effect on) `Cholesterol` (p-value below 0.05 means we reject the null).
- `BP_Status` pair *Optimal-Normal* **does not** have a significantly different mean value of `Cholesterol` (p-value above 0.05 means we do not reject the null). Simply put, Optimal and Normal `BP_Status` have equal means.
- Specifically, the following effects of `BP_Status` on `Cholesterol` can be seen:
  - Normal < High (The mean `Cholesterol` of High is greater than the mean `Cholesterol` of Normal)
  - Optimal < High (The mean `Cholesterol` of High is greater than the mean `Cholesterol` of Optimal)
  - Optimal = Normal (The mean `Cholesterol` of Normal is the same as the mean `Cholesterol` of Optimal)

Exercise 2: Analysis of Variance

For this problem use the `bupa.csv` data set. Check UCI Machine Learning Repository for more information (http://archive.ics.uci.edu/ml/datasets/Liver+Disorders). The mean corpuscular volume and alkaline phosphatase are blood tests thought to be sensitive to liver disorder related to excessive alcohol consumption. We assume that normality and independence assumptions are valid.

Exercise 2.A

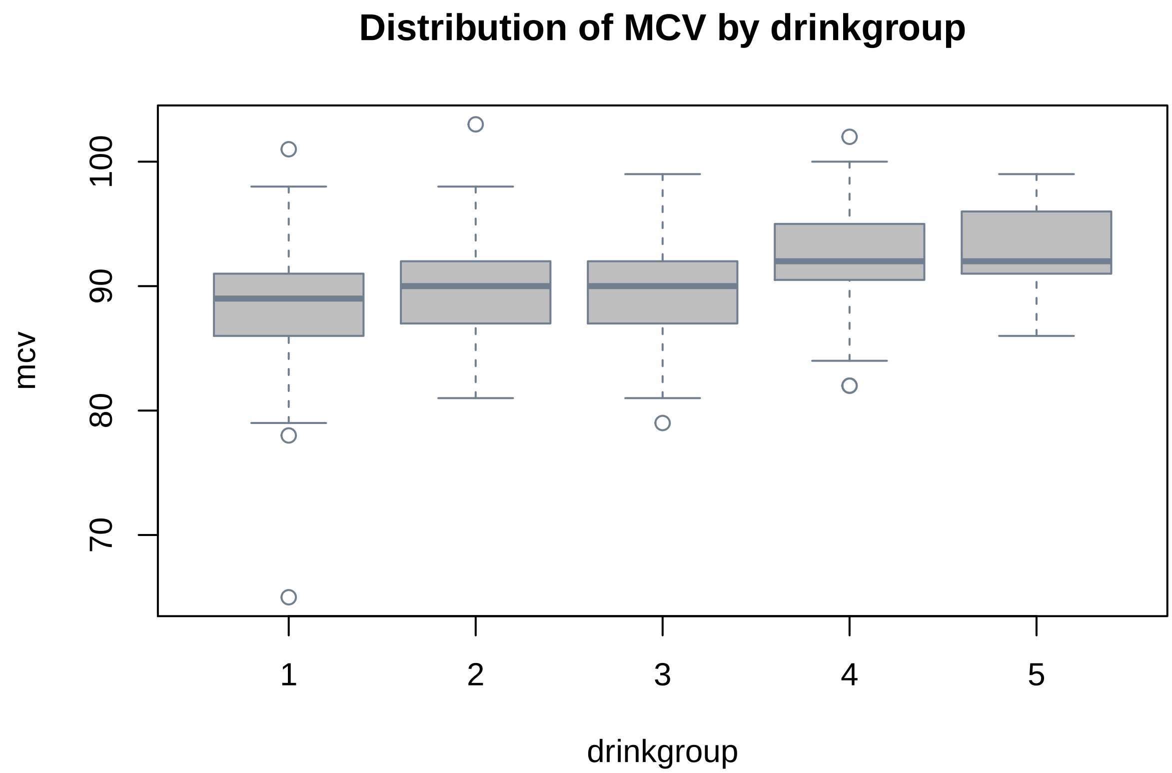
Perform a one-way ANOVA for Mean Corpuscular Volume or `mcv` as a function of `drinkgroup.` Comment on significance of the `drinkgroup`, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

Data Exploration - Check Balance

```
table(bupa$drinkgroup)
```

```
##
##   1   2   3   4   5
## 117  52  88  67  21
```

```
boxplot(mcv ~ drinkgroup, data=bupa,
        main="Distribution of MCV by drinkgroup",
        xlab = "drinkgroup",
        ylab = "mcv",
        col = "Grey",
        border = "slategray",
        horizontal = FALSE
        )
```



Observation: The distribution is **unbalanced**. Each `drinkgroup` has a different number of observations.

One-Way ANOVA

```
aov.res_bupa_mcv= aov(mcv~drinkgroup, data=bupa)

summary(aov.res_bupa_mcv) #ANOVA result
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## drinkgroup    4    733  183.29   10.26 7.43e-08 ***
## Residuals   340   6073   17.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: The p-value of 7.43e-08 is below the significance level of 0.05, meaning that we **reject** the null hypothesis. Therefore, `drinkgroup` has a significant effect on `mcv` (at least one group in `drinkgroup` has a different mean of `mcv`).

R-square (variation of response variable explained by `drinkgroup`)

```
lm.res_bupa_mcv = lm(mcv ~ drinkgroup, data = bupa)

summary(lm.res_bupa_mcv)$r.squared
```

```
## [1] 0.1077214
```

Conclusion: 10.8% of the variation of `mcv` can be explained by `drinkgroup`.

Check Equal Variance Assumption

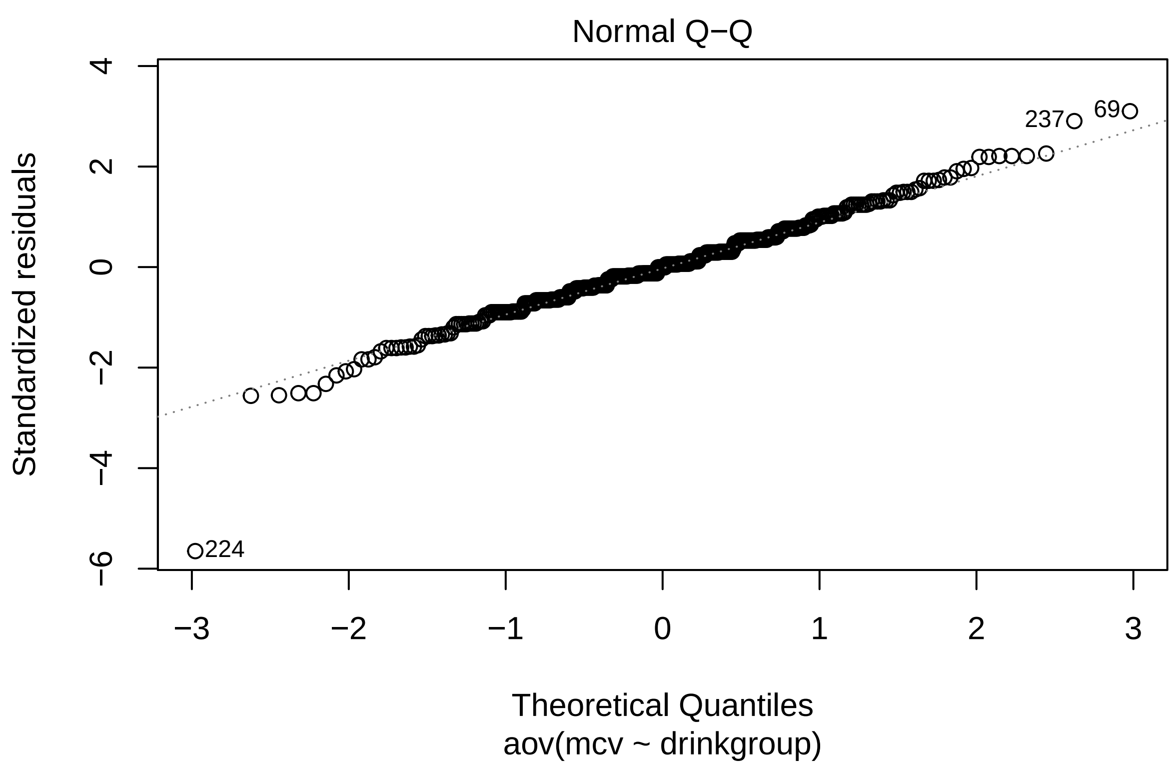
```r
LeveneTest(aov.res_bupa_mcv)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.3053 0.8744
##       340
```

Conclusion: The p-value is above the significance level of 0.05, meaning that we **can't reject** the null. Therefore, all groups in `drinkgroup` have the same variance.

Check Normality

```r
par(mfrow=c(1,1))# diagnostics plot - in one
plot(aov.res_bupa_mcv, 2)
```



Conclusion: Through analysis of the Q-Q plot, we can see that a normal distribution is reasonable.

Exercise 2.B

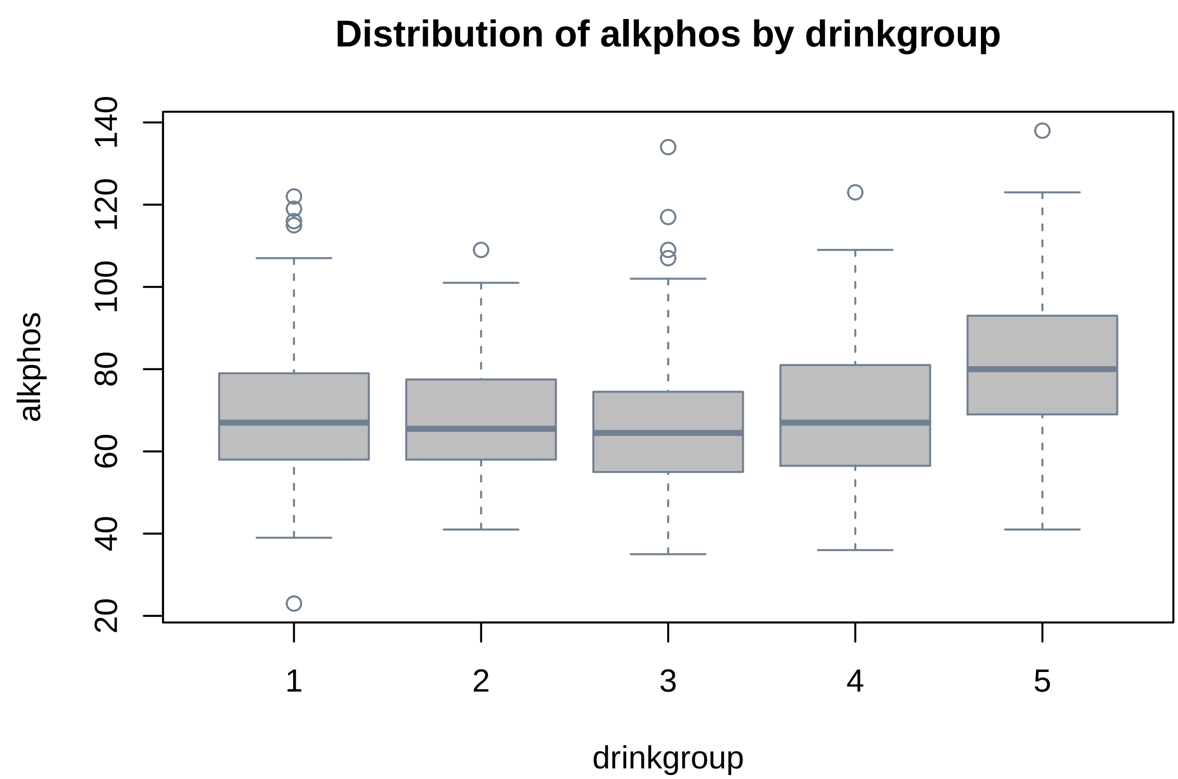
Perform a one-way ANOVA for `alkphos` as a function of `drinkgroup`. Comment on statistical significance of the `drinkgroup`, the amount of variation described by the model, and whether or not the equal variance assumption can be trusted.

Data Exploration - Check Balance

```
table(bupa$drinkgroup)
```

```
##
##   1   2   3   4   5
## 117  52  88  67  21
```

```
boxplot(alkphos ~ drinkgroup, data=bupa,
        main="Distribution of alkphos by drinkgroup",
        xlab = "drinkgroup",
        ylab = "alkphos",
        col = "Grey",
        border = "slategray",
        horizontal = FALSE
        )
```



Observation: The distribution is **unbalanced**. Each `drinkgroup` has a different number of observations.

One-Way ANOVA

```r
aov.res_bupa_alkphos= aov(alkphos~drinkgroup, data=bupa)

summary(aov.res_bupa_alkphos) #ANOVA result
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## drinkgroup    4   4946  1236.4   3.792 0.00495 **
## Residuals   340 110858   326.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: The p-value of 0.00495 is below the significance level of 0.05, meaning that we **reject** the null hypothesis. Therefore, `drinkgroup` has an effect on `alkphos` (at least one group in `drinkgroup` has a different mean of `alkphos`).

R-square (variation of response variable explained by `drinkgroup`)

```r
lm.res_bupa_alkphos = lm(alkphos ~ drinkgroup, data = bupa)

summary(lm.res_bupa_alkphos)$r.squared
```

```
## [1] 0.04270721
```

Conclusion: 4.3% of the variation of `alkphos` can be explained by `drinkgroup`.

Check Equal Variance Assumption

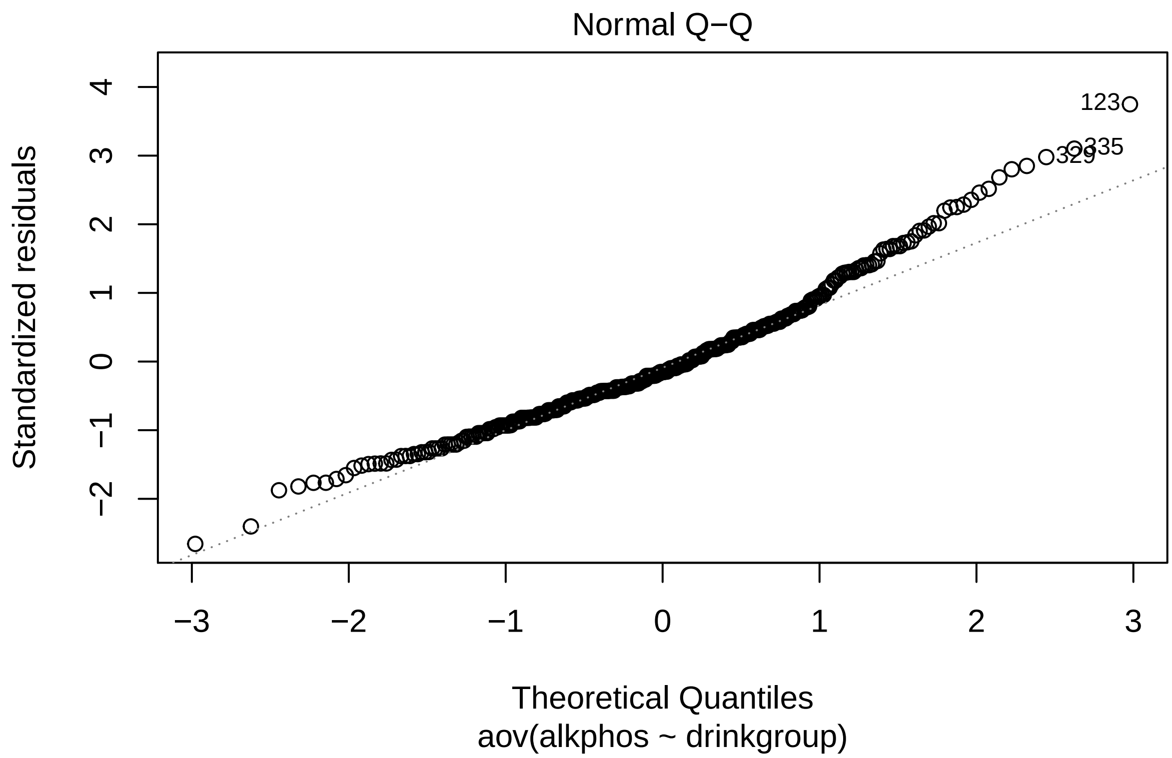
```r
LeveneTest(aov.res_bupa_alkphos)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.8089 0.5201
##       340
```

Conclusion: The p-value is above the significance level of 0.05, meaning that we **can't reject** the null. Therefore, all groups in `drinkgroup` have the same variance.

Check Normality

```r
par(mfrow=c(1,1))# diagnostics plot - in one
plot(aov.res_bupa_alkphos, 2)
```



Conclusion: Through analysis of the Q-Q plot, we can see that a normal distribution is reasonable.

Exercise 2.C

Perform post-hoc tests for models in a) and b). Comment on any similarities or differences you observe from their results.

```
ScheffeTest(aov.res_bupa_mcv)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $drinkgroup
##             diff      lwr.ci   upr.ci    pval
## 2-1  1.241452991 -0.94020481 3.423111  0.5410
## 3-1  0.938131313 -0.90892674 2.785189  0.6495
## 4-1  3.744610282  1.73913894 5.750082 1.9e-06 ***
## 5-1  3.746031746  0.64379565 6.848268  0.0081 **
## 3-2 -0.303321678 -2.59291786 1.986275  0.9966
## 4-2  2.503157290  0.08395442 4.922360  0.0380 *
## 5-2  2.504578755 -0.87987039 5.889028  0.2646
## 4-3  2.806478969  0.68408993 4.928868  0.0025 **
## 5-3  2.807900433 -0.37116998 5.986971  0.1151
## 5-4  0.001421464 -3.27222796 3.275071  1.0000
```

```
##
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
ScheffeTest(aov.res_bupa_alkphos)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $drinkgroup
##          diff      lwr.ci    upr.ci   pval
## 2-1 -2.645299 -11.9663647  6.675766 0.9419
## 3-1 -4.056138 -11.9476367  3.835360 0.6389
## 4-1 -1.148743  -9.7170578  7.419571 0.9965
## 5-1 12.572650  -0.6815582 25.826857 0.0734 .
## 3-2 -1.410839 -11.1930681  8.371390 0.9953
## 4-2  1.496556  -8.8394138 11.832525 0.9952
## 5-2 15.217949   0.7579944 29.677903 0.0329 *
## 4-3  2.907395  -6.1604467 11.975236 0.9117
## 5-3 16.628788   3.0463078 30.211268 0.0069 **
## 5-4 13.721393  -0.2651729 27.707959 0.0578 .
##
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Differences:

`mcv`

- `drinkgroup` Groups *4-1*, *5-1*, *4-2*, and *4-3* respectively have significantly different mean values of `mcv` (p-value below 0.05 means we reject the null).
- `drinkgroup` Groups *2-1*, *3-1*, *3-2*, *5-2*, *5-3*, and *5-4* **do not** have a significantly different mean value of `mcv` (p-value above 0.05 means we do not reject the null). Simply put, the preceding `drinkgroup` pairs have equal means.
- Specifically, the following effects of `drinkgroup` on `mcv` can be seen:
  - 4 > 1 (The mean `mcv` of 4 is greater than the mean `mcv` of 1)
  - 5 > 1 (The mean `mcv` of 5 is greater than the mean `mcv` of 1)
  - 4 > 2 (The mean `mcv` of 4 is greater than the mean `mcv` of 2)
  - 4 > 3 (The mean `mcv` of 4 is greater than the mean `mcv` of 3)
  - Equivalent means among the following `drinkgroup` pairs: *2-1*, *3-1*, *3-2*, *5-2*, *5-3*, and *5-4*

`alkphos`

- `drinkgroup` Groups *5-2* and *5-3* have significantly different mean values of `alkphos` (p-value below 0.05 means we reject the null).
- `drinkgroup` Groups *2-1*, *3-1*, *4-1*, *5-1*, *3-2*, *4-2*, *4-3*, and *5-4* **do not** have a significantly different mean value of `alkphos` (p-value above 0.05 means we do not reject the null). Simply put, the preceding `drinkgroup` pairs have equal means.

- Specifically, the following effects of `drinkgroup` on `alkphos` can be seen:
  - 5 > 2 (The mean `alkphos` of 5 is greater than the mean `alkphos` of 2)
  - 5 > 3 (The mean `alkphos` of 5 is greater than the mean `alkphos` of 3)
  - Equivalent means among the following `drinkgroup` pairs: *2-1*, *3-1*, *4-1*, *5-1*, *3-2*, *4-2*, *4-3*, and *5-4*

Similarities:

- Group pairs **2-1**, **3-1**, **3-2**, **5-4** all have equal mean values, and their high p-values above the significance level means that they do not have an effect on either `mcv` or `alkphos`.

Exercise 3:

The psychology department at a hypothetical university has been accused of underpaying female faculty members. The data represent salary (in thousands of dollars) for all 22 professors in the department. This problem is from Maxwell and Delaney (2004).

Exercise 3.A

Fit a two-way ANOVA model including `sex` (F, M) and `rank` (Assistant, Associate) the interaction term. What do the Type 1 and Type 3 sums of squares tell us about significance of effects? Is the interaction between `sex` and `rank` significant? Also comment on the variation explained by the model.

Two-Way ANOVA (Type 1)

```
aov.psych1 = aov(salary ~ sex * rank, data = psych)
aov.psych_3 = aov(salary ~ rank * sex, data = psych)

summary(aov.psych1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## sex          1 155.15  155.15  17.007 0.000637 ***
## rank         1 169.82  169.82  18.616 0.000417 ***
## sex:rank     1   0.63    0.63   0.069 0.795101    
## Residuals   18 164.21    9.12                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov.psych_3)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## rank         1 252.22  252.22  27.647 5.33e-05 ***
## sex          1  72.76   72.76   7.975   0.0112 *  
## rank:sex     1   0.63    0.63   0.069   0.7951    
## Residuals   18 164.21    9.12                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-Way ANOVA (Type 3)

```
Anova(aov.psych1, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: salary
##             Sum Sq Df  F value  Pr(>F)
## (Intercept) 8140.2  1 892.2994 < 2e-16 ***
## sex           28.0  1   3.0711 0.09671 .
## rank          70.4  1   7.7189 0.01240 *
## sex:rank       0.6  1   0.0695 0.79510
## Residuals    164.2 18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Type 1 ANOVA Test

We see that `sex` and `rank` have p-values below the significance level of 0.05. Therefore, we **reject** the null hypothesis for both `sex` and `rank` and conclude that that both have a significant effect on `salary`. Additionally, the interaction between `sex` and `rank` yields a p-value above the significance level of .05. This means that we **do not reject** the null, indicating that the interaction does not have a significant effect on `salary`.

Type 3 ANOVA Test

We see that `rank` has a p-value below the significance level of 0.05. Therefore, we **reject** the null hypothesis for `rank` and conclude that that it has a significant effect on `salary`. Additionally, the `sex` and the interaction between `sex` and `rank` both yield a p-value above the significance level of .05. This means that we **do not reject** the null, indicating that the `sex` and the interaction does not have a significant effect on `salary`.

Variation Explained by the Model

```
lm.psych1= lm(salary ~ sex * rank , data = psych)
summary(lm.psych1)$r.squared
```

```
## [1] 0.6647566
```

Observation: 66% of the variation of `salary` can be explained by the model (`rank` and `sex`).

Exercise 3.B

Refit the model without the interaction term. Comment on the significance of effects and variation explained. Report and interpret the Type 1 and Type 3 tests of the main effects. Are the main effects of `rank` and `sex` significant?

Two-Way ANOVA (Type 1)

```r
aov.psych2 = aov(salary ~ sex + rank, data = psych)
aov.psych4 = aov(salary ~ rank + sex, data = psych)

summary(aov.psych2)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## sex          1  155.2  155.15   17.88 0.000454 ***
## rank         1  169.8  169.82   19.57 0.000291 ***
## Residuals   19  164.8    8.68                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov.psych4)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## rank         1 252.22  252.22  29.071 3.34e-05 ***
## sex          1  72.76   72.76   8.386  0.00926 ** 
## Residuals   19 164.84    8.68                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-Way ANOVA (Type 3)

```r
Anova(aov.psych2, type = 3)
```

```
## Anova Table (Type III tests)
## 
## Response: salary
##              Sum Sq Df   F value    Pr(>F)    
## (Intercept) 10227.6  1 1178.8469 < 2.2e-16 ***
## sex            72.8  1    8.3862 0.0092618 ** 
## rank          169.8  1   19.5743 0.0002912 ***
## Residuals     164.8 19                        
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Type 1 ANOVA Test

We see that both `sex` and `rank` have p-values below the significance level of 0.05. Therefore, we **reject** the null hypothesis for both `sex` and `rank` and conclude that they both have a significant effect on `salary`, and at least one group in `rank` (Assoc or Assist) and one group in `sex` (Male or Female) have different mean values.

Type 3 ANOVA Test

We see that both `sex` and `rank` have p-values below the significance level of 0.05. Therefore, we **reject** the null hypothesis for both `sex` and `rank` and conclude that they both have a significant effect on `salary`, and at least one group in `rank` (Assoc or Assist) and one group in `sex` (Male or Female) have different mean values.

Variation Explained by the Model

```
lm.psych2= lm(salary ~ sex + rank , data = psych)
summary(lm.psych2)$r.squared
```

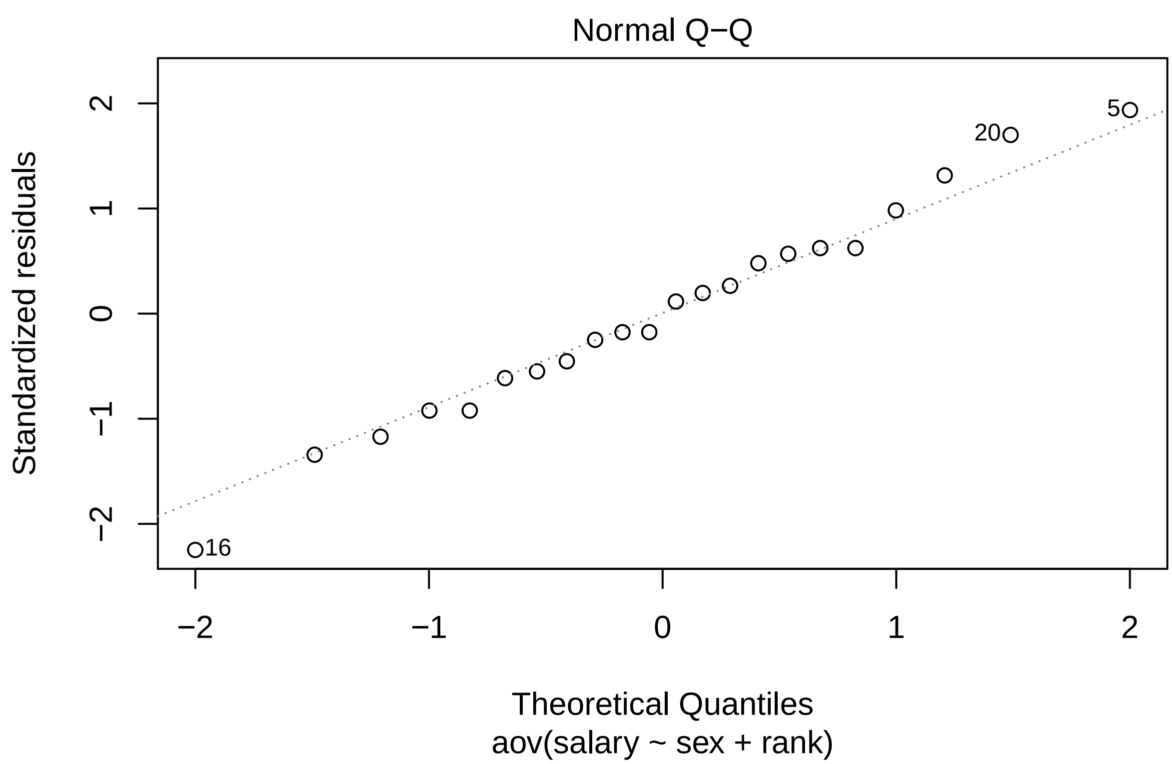
```
## [1] 0.6634627
```

Observation: 66% of the variation of `salary` can be explained by the model (`rank` and `sex`).

Exercise 3.C

Obtain model diagnostics to validate your Normality assumptions.

```
par(mfrow=c(1,1))
plot(aov.psych2, 2)
```



Conclusion: Through analysis of the Q-Q plot, we can see that a normal distribution is reasonable.

Exercise 3.D

Choose a final model based on your results from parts (a) and (b). Comment on any significant group differences through the post-hoc test. State the differences in `salary` across different main effect groups and interaction (if included) between them.

Decision

Based on the results from (a) and (b), we see that there **does not exist** an interaction effect. Therefore, the final model that is selected is the Two-Way ANOVA **without interaction**, specifically the Type 3 ANOVA test to ensure that we see unique contribution of each categorical variable. Because every effect is adjusted for all other effects, we believe this model is best suited for our dataset.

Post-Hoc Test

```
TukeyHSD(aov.psych2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = salary ~ sex + rank, data = psych)
##
## $sex
##         diff      lwr      upr     p adj
## M-F 5.333333 2.693648 7.973019 0.0004544
##
## $rank
##                  diff      lwr      upr     p adj
## Assoc-Assist 5.377778 2.738092 8.017463 0.0004193
```

Conclusion:

`sex`

- Due to a p-value below the significance level of 0.05, the M-F pair has a significant effect on `salary`.
- Specifically, the following effect of `sex` on `salary` can be seen:
  - M > F (The mean `salary` of Male is greater than the mean `salary` of Female)

`rank`

- Due to a p-value below the significance level of 0.05, the Assoc-Assist pair has a significant effect on `salary`.
- Specifically, the following effect of `rank` on `salary` can be seen:
  - Assoc > Assist (The mean `salary` of Associate is greater than the mean `salary` of Assistant)

Exercise 4:

Use the `cars_new.csv`. See HW1 for detailed information of variables.

Exercise 4.A

Start with a three-way main effects ANOVA and choose the best main effects ANOVA model for `mpg_highway` as a function of `cylinders`, `origin`, and `type` for the cars in this set. Comment on which terms should be kept in a model for `mpg_highway` and why based on Type 3 SS. For the model with just predictors you decide to keep, comment on the significant effects in the model and comment on how much variation in highway fuel efficiency the model describes.

We will utilize the Backwards Elimination model selection process to determine the main effects that will be included in the model.

Three-Way ANOVA (Type 3) Full Model

```
aov.cars_new1 = aov(mpg_highway ~ cylinders + origin + type, data = cars_new)
Anova(aov.cars_new1, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: mpg_highway
##             Sum Sq  Df   F value  Pr(>F)
## (Intercept)  69548   1 6501.6715 < 2e-16 ***
## cylinders     1453   1  135.8499 < 2e-16 ***
## origin           1   1    0.0786 0.77948
## type           108   1   10.1018 0.00175 **
## Residuals     1883 176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Type 3 ANOVA Test (Full Model)

We see that both `cylinders` and `type` have p-values below the significance level of 0.05. Therefore, we **reject** the null hypothesis for both `cylinders` and `type` and conclude that they both have a **significant effect** on `mpg_highway`, and at least one group in `cylinders` (4 or 6) and one group in `type` (Sedan or Sports) have different mean values. Because `origin` had a p-value below the significance level (does not meet our cutoff criteria), we **do not reject** the null; therefore, we will remove `origin` from the model because it has an **insignificant effect**. This is in line with the **Backward Elimination** model selection process.

Three-Way ANOVA (Type 3) Model (`cylinders` and `type`)

```
aov.cars_new1_1 = aov(mpg_highway ~ cylinders + type, data = cars_new)
Anova(aov.cars_new1_1, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: mpg_highway
##             Sum Sq  Df F value    Pr(>F)
## (Intercept)  88449   1 8311.96 < 2.2e-16 ***
## cylinders     1482   1  139.27 < 2.2e-16 ***
## type           116   1   10.88  0.001175 **
## Residuals     1883 177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Type 3 ANOVA Test (Full Model)

Based on the results of this Type 3 ANOVA test, `cylinders` and `type` have ap-value below the significance level of 0.05, meaning we reject the null. Both of these predictors have a significant effect on `mpg_highway`.

Variation Explained by the Model (Predictors = `cylinders` and `type`)

```
lm.cars_new1_1= lm(mpg_highway ~ cylinders + type , data = cars_new)
summary(lm.cars_new1_1)$r.squared
```

```
## [1] 0.4572163
```

Observation: 46% of the variation of `mpg_highway` can be explained by the model (`cylinders` and `type`).

Exercise 4.B

Starting with main effects chosen in part (a), find your best ANOVA model by adding in any additional interaction terms that will significantly improve the model. For your final model,comment on the significant effects and variation explained by the model.

Two-Way ANOVA with Interaction (Type 3)

```
aov.cars_new2 = aov(mpg_highway ~ cylinders * type, data = cars_new)

Anova(aov.cars_new2, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: mpg_highway
##                Sum Sq  Df  F value    Pr(>F)
## (Intercept)     85471   1 8358.838 < 2.2e-16 ***
## cylinders        1558   1  152.397 < 2.2e-16 ***
## type              198   1   19.392 1.844e-05 ***
## cylinders:type     84   1    8.201  0.004696 **
## Residuals        1800 176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

The best model to use is the Two-Way ANOVA model (Type 3) with the implementation of categorical predictors `cylinders`, `type`, and their **interaction**. We see that both `cylinders`, `type`, and their interaction have p-values below the significance level of 0.05. Therefore, we **reject** the null hypothesis and conclude that the individual predictors and their interaction have a significant effect on `mpg_highway`.

Variation Explained by the Model (Predictors = `cylinders`, `type`, and interaction)

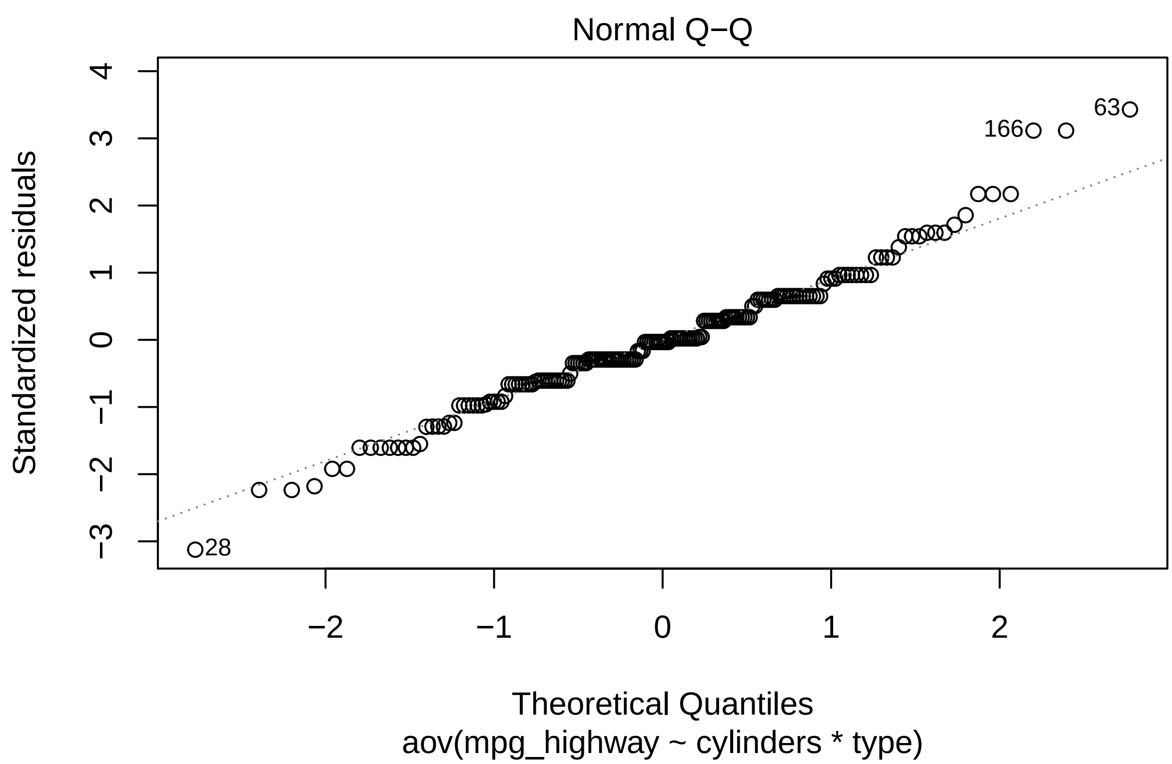
```
lm.cars_new2= lm(mpg_highway ~ cylinders * type , data = cars_new)
summary(lm.cars_new2)$r.squared
```

```
## [1] 0.4813821
```

Observation: 48% of the variation of `mpg_highway` can be explained by the model (`cylinders`, `type`, and interaction).

Normality Check

```
par(mfrow=c(1,1))
plot(aov.cars_new2, 2)
```



Conclusion: Through analysis of the Q-Q plot, we can see that a normal distribution is reasonable.

Exercise 4.C

Comment on any significant group differences through the post-hoc test. What does this tell usabout fuel efficiency differences across `cylinders`, `origin`, or `type` groups? See Hint in Exercise 3.

Post-Hoc Test

```r
TukeyHSD(aov.cars_new2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mpg_highway ~ cylinders * type, data = cars_new)
##
## $cylinders
##          diff       lwr       upr p adj
## 6-4 -5.722662 -6.664343 -4.780981     0
##
## $type
##                   diff       lwr       upr     p adj
## Sports-Sedan -2.817931 -4.470787 -1.165075 0.0009407
##
## $`cylinders:type`
##                         diff       lwr       upr     p adj
## 6:Sedan-4:Sedan   -6.1723315 -7.469178 -4.875485 0.0000000
## 4:Sports-4:Sedan  -5.2275641 -8.306639 -2.148489 0.0001079
## 6:Sports-4:Sedan  -6.6025641 -9.681639 -3.523489 0.0000006
## 4:Sports-6:Sedan   0.9447674 -2.120956  4.010491 0.8546517
## 6:Sports-6:Sedan  -0.4302326 -3.495956  2.635491 0.9834567
## 6:Sports-4:Sports -1.3750000 -5.521993  2.771993 0.8253946
```

`cylinders`

- Due to a p-value below the significance level of 0.05, `cylinders` group **6-4** has a significant effect on `mpg_highway`.
- Specifically, the following effects of `cylinders` on `mpg_highway` can be seen:
  - $6 < 4$ (The mean `mpg_highway` of 4 is greater than the mean `mpg_highway` of 6)

`type`

- Due to a p-value below the significance level of 0.05, `type` group **Sports-Sedan** has a significant effect on `mpg_highway`.
- Specifically, the following effects of `type` on `mpg_highway` can be seen:
  - Sports < Sedan (The mean `mpg_highway` of Sedan is greater than the mean `mpg_highway` of Sports)

`cylinders` and `type` Interaction

- Due to a p-value below the significance level of 0.05, `type` groups **6:Sedan-4:Sedan**, **4:Sports-4:Sedan**, and **6:Sports-4:Sedan** have a significant effect on `mpg_highway`.
- Specifically, the following effects of interaction effects on `mpg_highway` can be seen:

  - 6:Sedan < 4:Sedan (The mean `mpg_highway` of 4:Sedan is greater than the mean `mpg_highway` of 6:Sedan)
  - 4:Sports < 4:Sedan (The mean `mpg_highway` of 4:Sedan is greater than the mean `mpg_highway` of 4:Sports)
  - 6:Sports < 4:Sedan (The mean `mpg_highway` of 4:Sedan is greater than the mean `mpg_highway` of 6:Sports)

Conclusion:

In summary, the analysis above indicates the following:

- 4 Cylinder cars have a higher `mpg_highway` than 6 Cylinder cars, meaning they have better highway fuel efficiency
- Sedans have a a higher `mpg_highway` than Sports car types, meaning they have better highway fuel efficiency
- 4 Cylinder Sedans have a higher `mpg_highway` than 6 Cylinder Sedans, meaning they have better highway fuel efficiency
- 4 Cylinder Sedans have a higher `mpg_highway` than 4 Cylinder Sports car types, meaning they have better fuel efficiency
- 4 Cylinder Sedans have a higher `mpg_highway` than 6 Cylinder Sports car types, meaning they have better fuel efficiency