Project #2 BTRY 3020

Nick Gembs

5/7/2021

##Problem #1: A rental car company wants to determine if the color of car (red , blue , white , black) effects the length of a car rental (days). Over a period of time, they randomly selected a customer, and randomly assigned one of the four colors. The responses can be cound in Carrentals.csv. (Mont)

1.Is there evidence to support a claim that the color affects the length of a rental? α=0.05

```
library(MASS)
rentals <- read.csv("C:/Users/Nick/Downloads/carrentals.csv")
rentals <- stack(rentals)
names(rentals) <- c("length", "color")

rentals.aov <- aov(length~color, data = rentals)
rentals.sum <- summary(rentals.aov)
rentals.sum
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## color        3  16.67   5.558    1.11  0.358
## Residuals   36 180.30   5.008
```

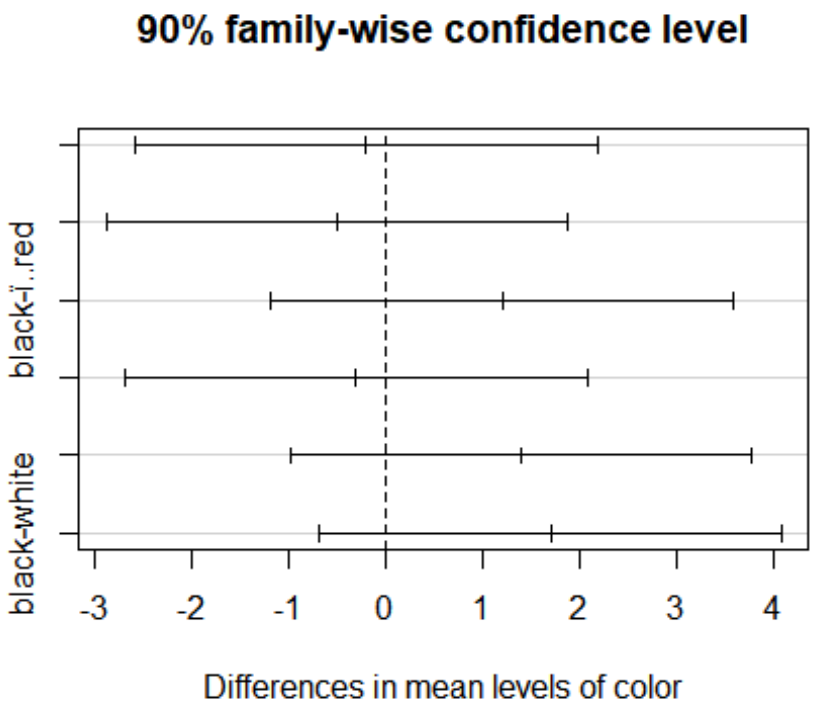
```
Based on these results, we do not have any evidence of an association between
car color and length of car rental, the p-value is .358>.05.
```

2.Identify which colors have a statistically significant difference, if any. Compute an estimate for those differences with 90% confidence, adjusting for the multiple comparisons.

```
#Tukey's method
rentals.tukey <- TukeyHSD(rentals.aov, conf.level = .9)
rentals.tukey
```

```
##   Tukey multiple comparisons of means
##     90% family-wise confidence level
##
## Fit: aov(formula = length ~ color, data = rentals)
##
## $color
##               diff        lwr      upr     p adj
## blue-ï..red  -0.2 -2.5784012 2.178401 0.9971215
## white-ï..red -0.5 -2.8784012 1.878401 0.9586389
## black-ï..red  1.2 -1.1784012 3.578401 0.6314346
## white-blue   -0.3 -2.6784012 2.078401 0.9904787
```

```
## black-blue    1.4 -0.9784012 3.778401 0.5082612
## black-white   1.7 -0.6784012 4.078401 0.3392640

plot(rentals.tukey)
```



3.Below you will find a complete set of orthogonal contrasts. Interpret each.

```
contrasts(rentals$color)

##        blue white black
## ï..red    0     0     0
## blue      1     0     0
## white     0     1     0
## black     0     0     1

levels(rentals$color)

## [1] "ï..red" "blue"   "white"  "black"

contrast.1<-c(.5,.5,-.5,-.5) #colorful vs. neutral
contrast.2<-c(1,-1,0,0) #red vs blue
contrast.3<-c(0,0,1,-1) #white vs black
contrast.all<-cbind(contrast.1,contrast.2,contrast.3)
contrast.all

##      contrast.1 contrast.2 contrast.3
## [1,]        0.5          1          0
## [2,]        0.5         -1          0
```

```
## [3,]       -0.5          0          1
## [4,]       -0.5          0         -1
```

4. Below, you will find the aggregated sample mean rental length for each color of car. Use give point estimates for the value of each contrast.

```
rm(mean)

## Warning in rm(mean): object 'mean' not found

mns <- aggregate(length~color, data=rentals, FUN = mean)
mns

##    color length
## 1 ï..red    4.1
## 2   blue    3.9
## 3  white    3.6
## 4  black    5.3

mns$length%*%contrast.all

##      contrast.1 contrast.2 contrast.3
## [1,]      -0.45        0.2       -1.7
```

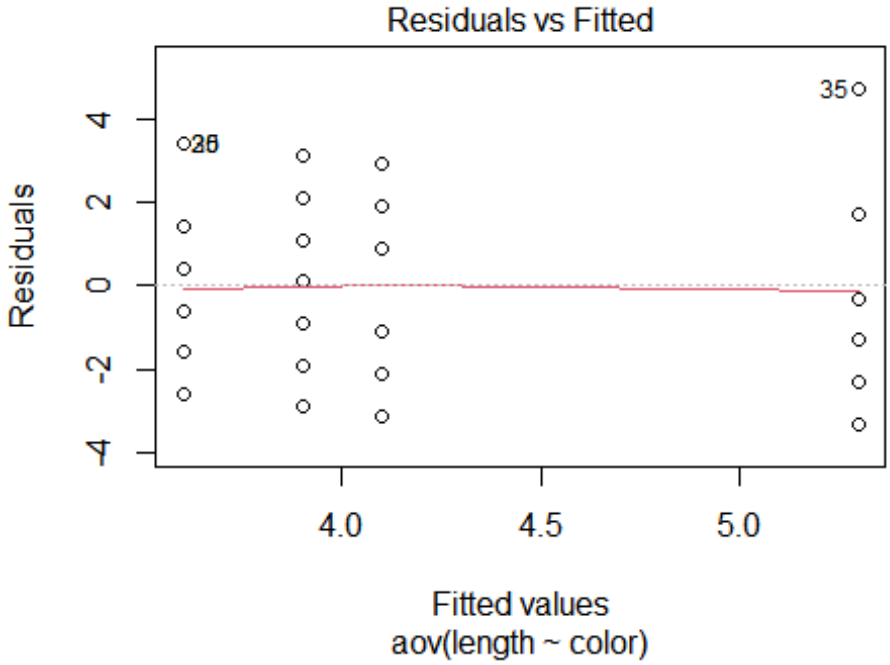
e. Are any of the contrast statistically different from zero?

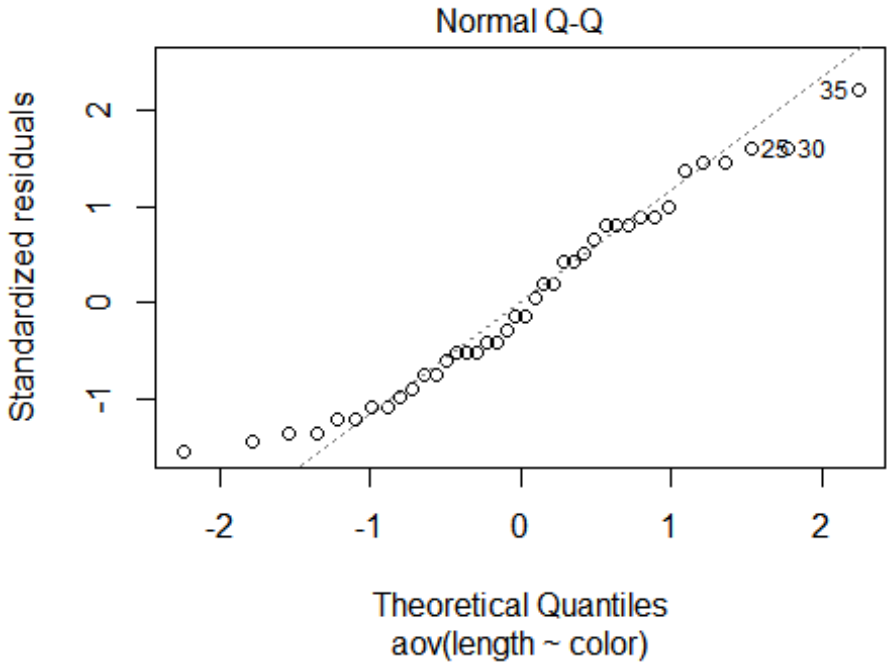
```
#Just do it component wise error rate.
contrasts(rentals$color)<-contrast.all
rentals.aov<-aov(length~color,data=rentals)
summary(rentals.aov,split=list(color=list("ColorVsBW"=1,"RVsB"=2,"BVsW"=3) )
)

##                    Df Sum Sq Mean Sq F value Pr(>F)
## color               3  16.68   5.558   1.110  0.358
##   color: ColorVsBW  1   2.03   2.025   0.404  0.529
##   color: RVsB       1   0.20   0.200   0.040  0.843
##   color: BVsW       1  14.45  14.450   2.885  0.098 .
## Residuals          36 180.30   5.008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

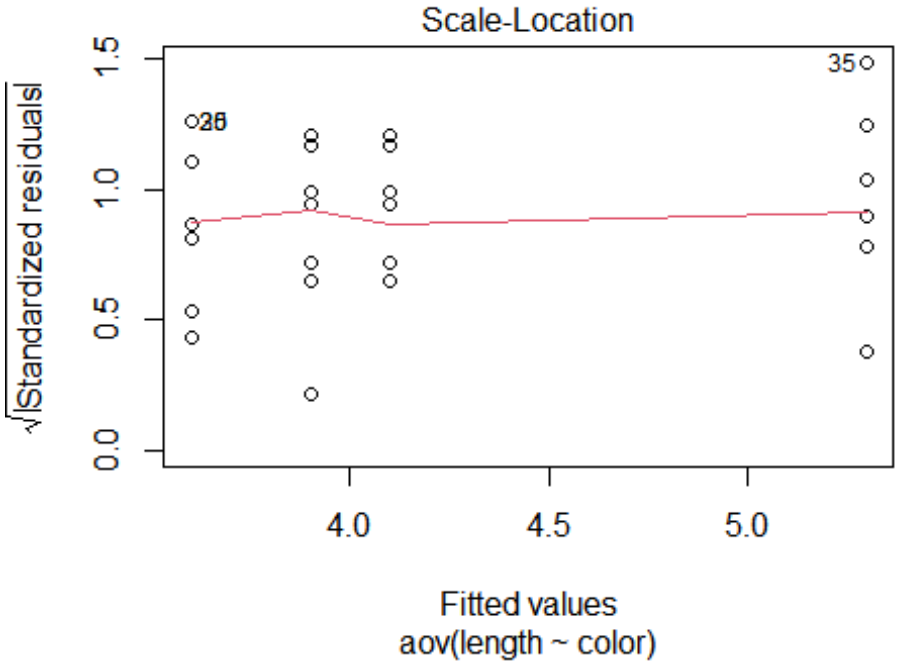
None of these appear to be statistically significant, the closest is black vs white.

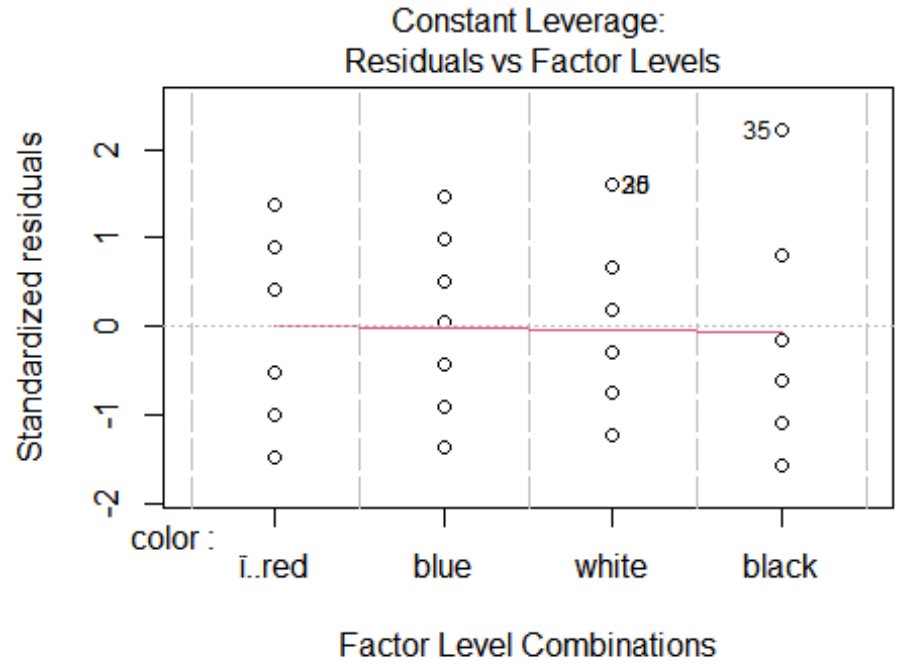
6. Does this appear to meet the assumptions? (Same assumptions as linear regression, since it is linear regression.)

```
plot(rentals.aov)
```


Residuals vs Fitted
Residuals
Fitted values
aov(length ~ color)
35
4.0
4.5
5.0


Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(length ~ color)
35
25
30
-2
-1
0
1
2





Seems to meet all the assumptions

2. A drug company wants to investigate the bioactivity of a new drug. A completely randomize single factor experiment was conducted. Three different dosages of the drugs are easy to manufacture within the possible dosages of interest. The results are listed in the file Dosage.csv.(Mont)

    1. Does there appear to be a significant mean bioactivity difference among the three dosages? α=.01

```
Dosage <- read.csv("C:/Users/Nick/Downloads/Dosage.csv")
names(Dosage)[1] <- "g20"


dat1 <- data.frame(Dosage$g20 , Dosage$g30, Dosage$g40)
dat <- stack(dat1)
colnames(dat) <- c("bio", "dosage")
dat
```

```
##    bio     dosage
## 1   24 Dosage.g20
## 2   28 Dosage.g20
## 3   37 Dosage.g20
## 4   30 Dosage.g20
## 5   37 Dosage.g30
## 6   44 Dosage.g30
## 7   31 Dosage.g30
## 8   35 Dosage.g30
## 9   42 Dosage.g40
## 10  47 Dosage.g40
## 11  52 Dosage.g40
## 12  38 Dosage.g40
```

```
anova_results <- aov(bio ~ dosage, data = dat)
summary(anova_results)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## dosage       2  450.7  225.33   7.036 0.0145 *
## Residuals    9  288.2   32.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ft1v2 <- c(1, -1, 0)
ft1v3 <- c(-1, 0, 1)
ft2v3 <- c(0, 1, -1)
Contr <- cbind(ft1v2, ft1v3, ft2v3)

contrasts(dat$dosage) <- Contr
contrast_results <- aov(bio ~ dosage, data = dat)

summary(contrast_results, split = list(dosage = list(`1v2` = 1, `1v3` = 2,
`2v3` = 3)))
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## dosage         2  450.7   225.3   7.036 0.01446 *
##   dosage: 1v2  1   98.0    98.0   3.060 0.11418
##   dosage: 1v3  1  352.7   352.7  11.011 0.00896 **
##   dosage: 2v3  1
## Residuals      9  288.2    32.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Only using the first overall aov test, here does not appear to be a significant mean bioactivity difference among the three dosages because the p-value is .0145>.01. However, when split into contrasts, it appears that there is a significant difference between the means of treatment 1 and treatment 3 because the p-value is 0.00896<.01

2.For any pair of means that appear to be different, compute an estimate of the difference with 99% confidence, adjusting for the multiple comparisons?

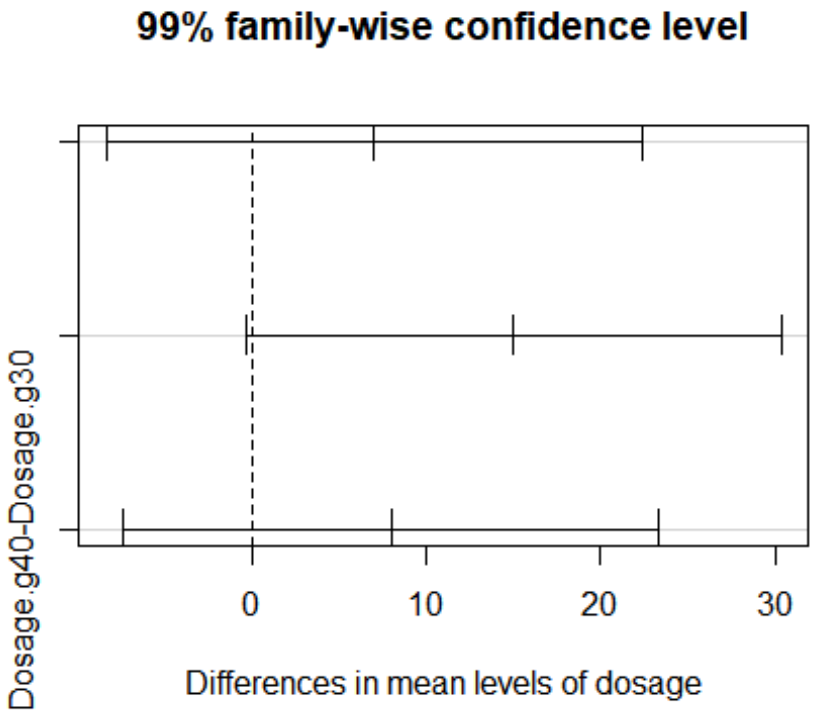
```
tukey.results <- TukeyHSD(contrast_results, ordered = TRUE, conf.level =
0.99)
tukey.results
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = bio ~ dosage, data = dat)
##
## $dosage
##                        diff        lwr      upr     p adj
## Dosage.g30-Dosage.g20    7 -8.3594856 22.35949 0.2402975
## Dosage.g40-Dosage.g20   15 -0.3594856 30.35949 0.0114434
## Dosage.g40-Dosage.g30    8 -7.3594856 23.35949 0.1680265
```

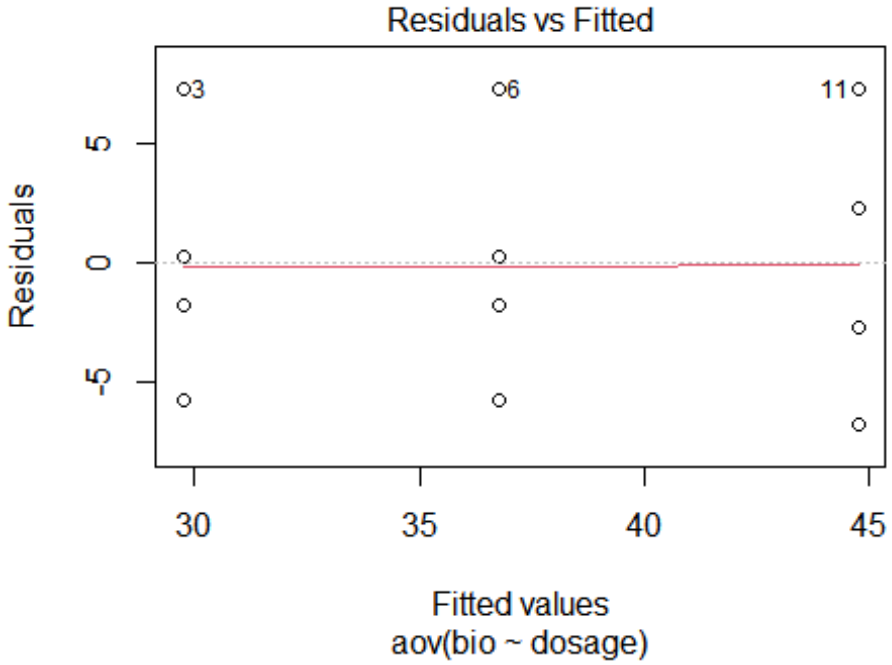
The difference between g40 and g20 dosage with 99% confidence has an upper bound of 30.35949 and a lower bound of -.3594856.

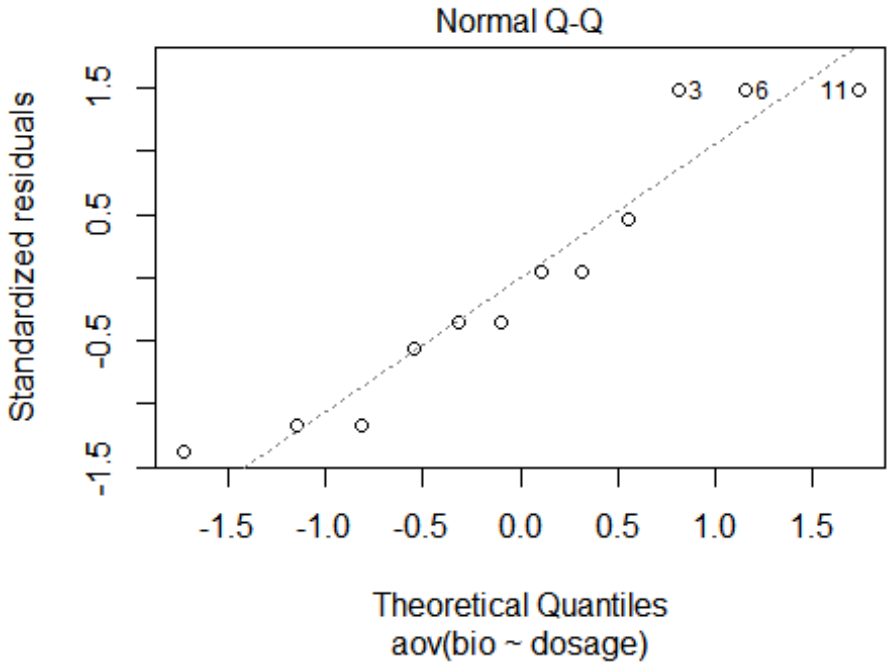
3.Does this appear to meet the assumptions? (Same assumptions as linear regression, since it is linear regression.)

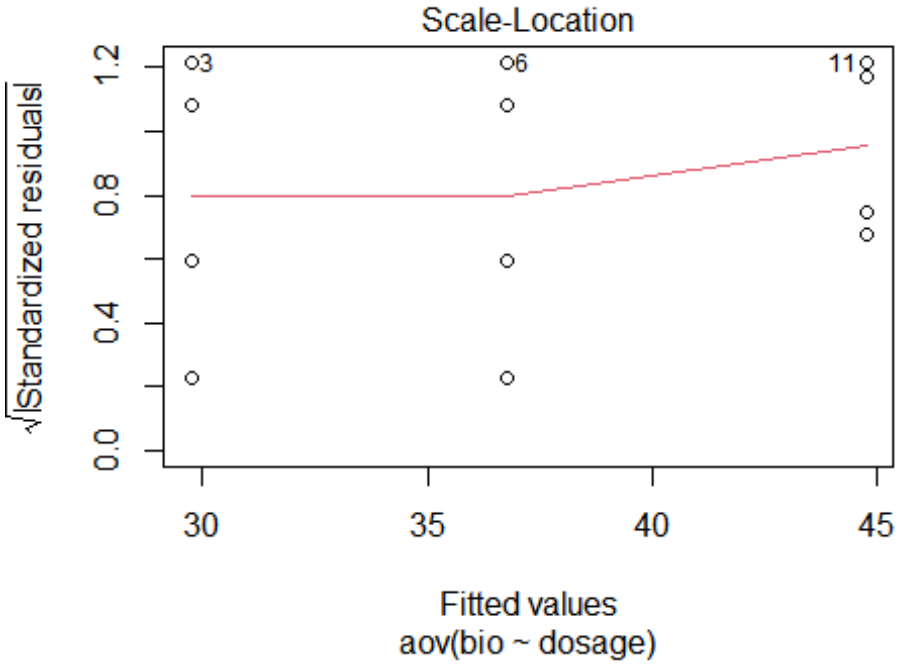
```
plot(tukey.results)
```

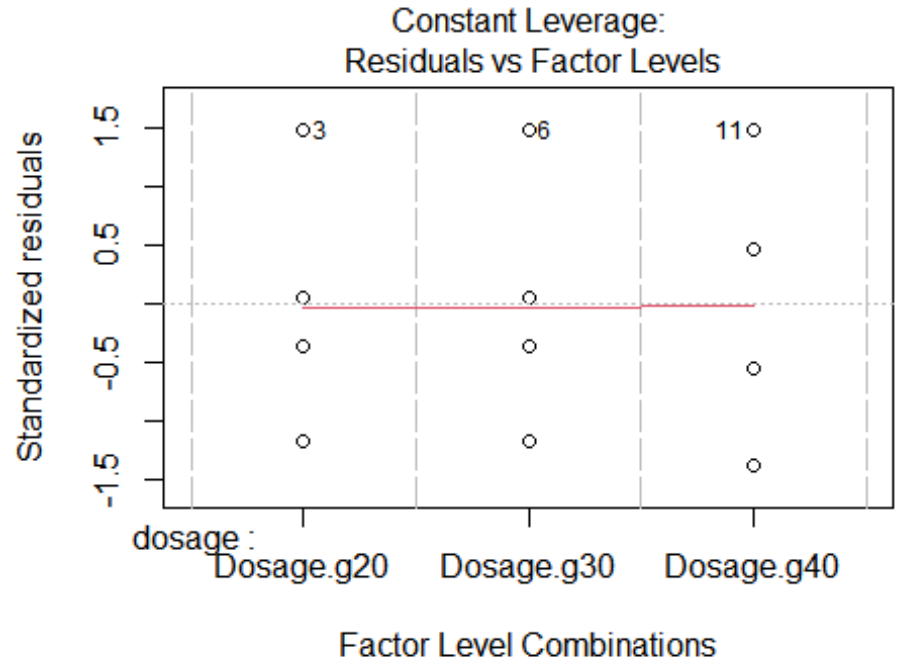



```
plot(contrast_results)
```


Residuals vs Fitted
Residuals
Fitted values
aov(bio ~ dosage)
3
6
11
5
0
-5
30
35
40
45


Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(bio ~ dosage)
3
6
11
1.5
0.5
-0.5
-1.5
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5





This does appear to meet the assupmtions as the residuals appear to be evenly spread, the variance appears to be constant, the residual-fitted and scale location plots are generally flat, and the qqplot appears to be linear.

3.Three brands of AA batteries were being studied to determine if they have the different lifetimes (weeks) . The relevent data is collected in the Batteries.csv file. (Mont) Brands 1 and Brand 2 are supposedly made by the Acme Co., and simply have different brand names. Brand 3 is made by a different company, Gump Co.

1.Is there evidence of a difference in the average battery lifetime of the different types? Inspect the dataset once you import it. You may need to chage its layout, or change the type of a column. (α=0.05)

```
Batteries <- read.csv("C:/Users/Nick/Downloads/Batteries.csv")
names(Batteries)[1] <- "hundred"
names(Batteries)[2] <- "brands"
anova_results <- aov(hundred ~ brands, data = Batteries)
summary(anova_results)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## brands       2 1150.5   575.3   39.95 9.03e-06 ***
## Residuals   11  158.4    14.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is evidence that there is a difference in the average battery lifetime of the different times because the p-value is 9.03e-06<.05.

2.Construct a meaningful complete set of orthogonal contrasts. Indicate their meaning. Show that they are orthogonal. (Don't forget to check the ordering of the factors before you construct any contrasts. You need your coefficients to be in the correct order for a contrast to be useful.)

```
Batteries$brands <- as.factor(Batteries$brands)
ft1v23 <- c(1, -.5, -.5) # Brand 1 vs brand 2 and brand 3
ft1v3 <- c(0, 1, -1) # brand 2 vs brand 3
Contr <- cbind(ft1v23, ft1v3)
t(Contr) %*% Contr
```

```
##        ft1v23 ft1v3
## ft1v23    1.5     0
## ft1v3     0.0     2
```

```
contrasts(Batteries$brands) <- Contr
```

3.Are any of these contrasts significant? α_e=0.05

```
tukey.results <- TukeyHSD(contrast_results, ordered = TRUE, conf.level =
0.95)
tukey.results
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
```

```
## Fit: aov(formula = bio ~ dosage, data = dat)
##
## $dosage
##                         diff       lwr      upr     p adj
## Dosage.g30-Dosage.g20    7 -4.172869 18.17287 0.2402975
## Dosage.g40-Dosage.g20   15  3.827131 26.17287 0.0114434
## Dosage.g40-Dosage.g30    8 -3.172869 19.17287 0.1680265
```

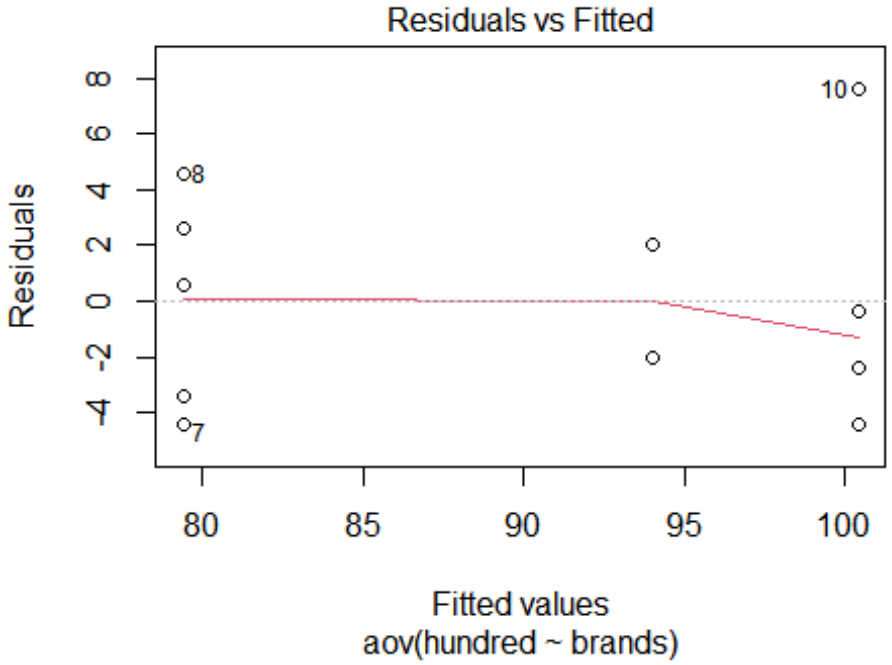
```
contrast_results <- aov(hundred ~ brands, data = Batteries)

summary(contrast_results, split = list(brands = list(`1v23` = 1, `1v3` = 2)))
```

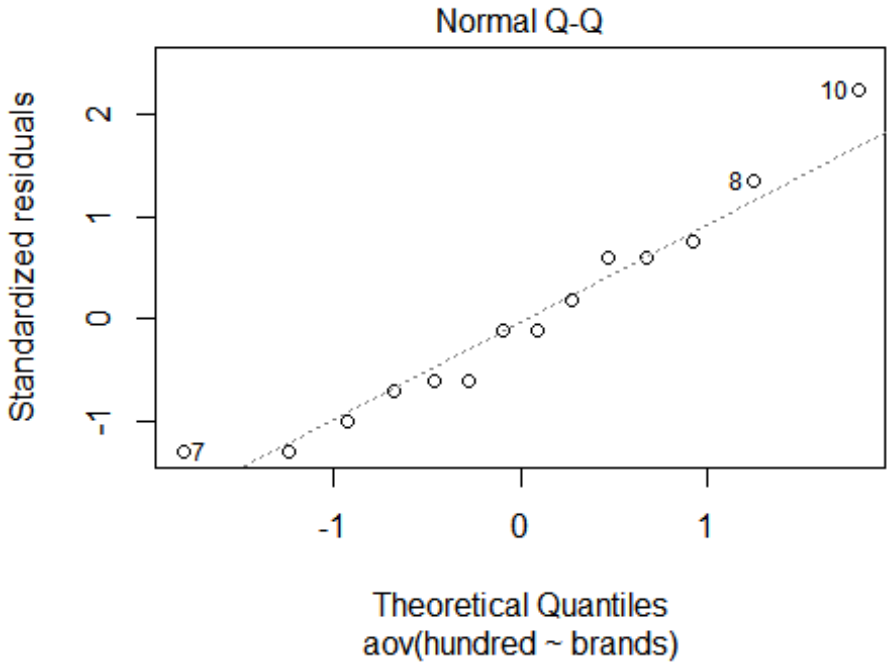
```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## brands          2 1150.5   575.3  39.949 9.03e-06 ***
##   brands: 1v23  1   48.0    48.0   3.335    0.095 .
##   brands: 1v3   1 1102.5  1102.5  76.563 2.76e-06 ***
## Residuals      11  158.4    14.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

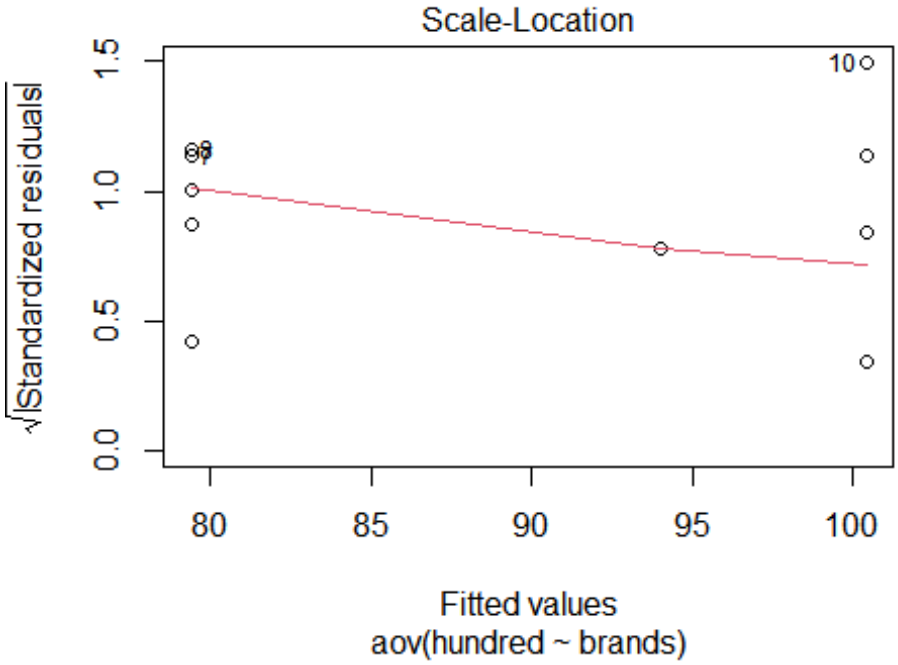
Contrasts between brand 1 and brand 3 is significant.

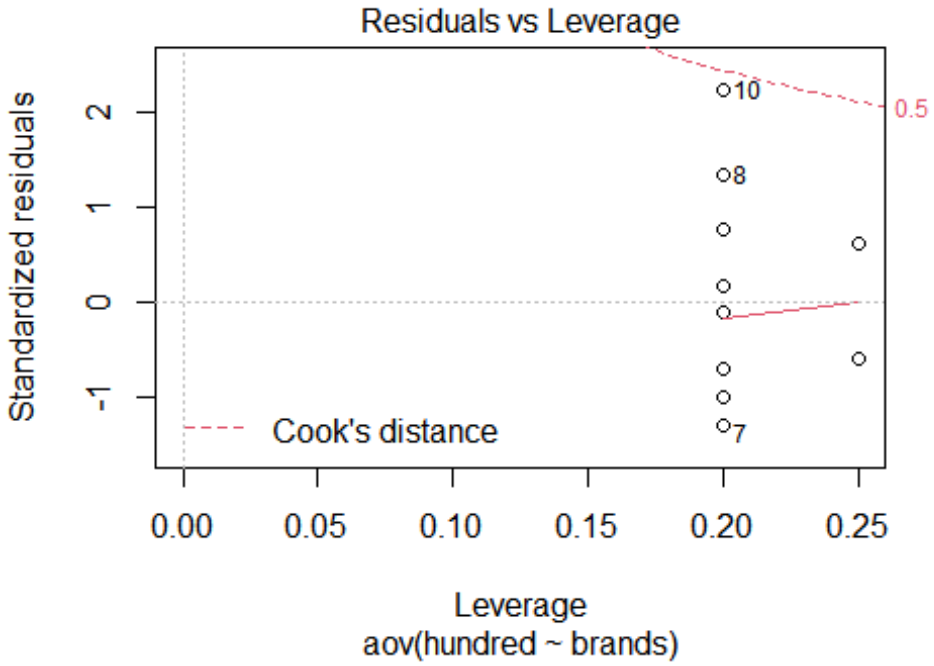
4.Does this appear to meet the assumptions? (Same assumptions as linear regression, since it is linear regression.)

```
plot(contrast_results)
```


Residuals vs Fitted
Residuals
8
6
4
2
0
-2
-4
10
8
7
80
85
90
95
100
Fitted values
aov(hundred ~ brands)


Normal Q-Q
Standardized residuals
2
1
0
-1
10
8
7
-1
0
1
Theoretical Quantiles
aov(hundred ~ brands)





This does not appear to perfectly meet the assumptions as both the residualvsfitted and scale-location plots show residuals the are lower towards the end of the graph. Variance does not appear to be entirely consistant.

Problem #4: An soil microbiology experiment was conducted to determine the effect of differing amount of fertilizer on the nitrogen fixation by a type of bacteria. The experiment used four types of crops, and three concentrations of fertilizer. Four replications of each treatment pair were used. At a certain point in the crops lifecycle, a measure of nitrogen fixation was taken. The results are in the Crops.csv data set. (Kuehl)

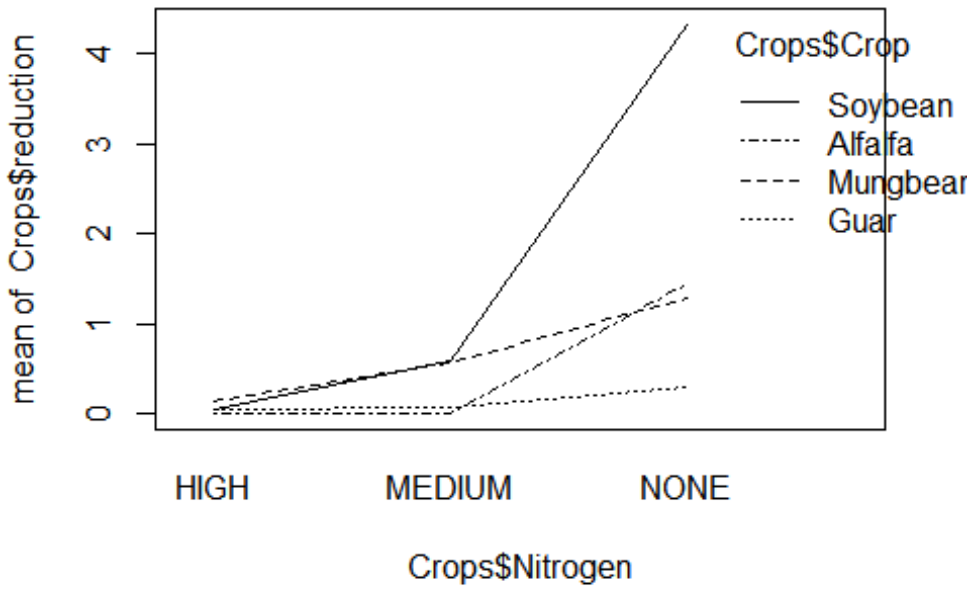
1. What type of treatment design is being used?

```
Crops <- read.csv("C:/Users/Nick/Downloads/Crops.csv")
```

A Factorial treatment design is being used.

2. Create an interaction plot. Does there appear to be an interaction between the two factors.

```
interaction.plot(Crops$Nitrogen, Crops$Crop, Crops$reduction)
```



There does appear to be an interaction between the two factors as the interaction plot lines do not appear to be parallel, indicating an interaction.

3. Does there appear to an interaction between the two factors? (Use a test.)

```
anova_results <- aov(reduction ~ Nitrogen + Crop + Nitrogen:Crop,
 data = Crops)
summary(anova_results)
```

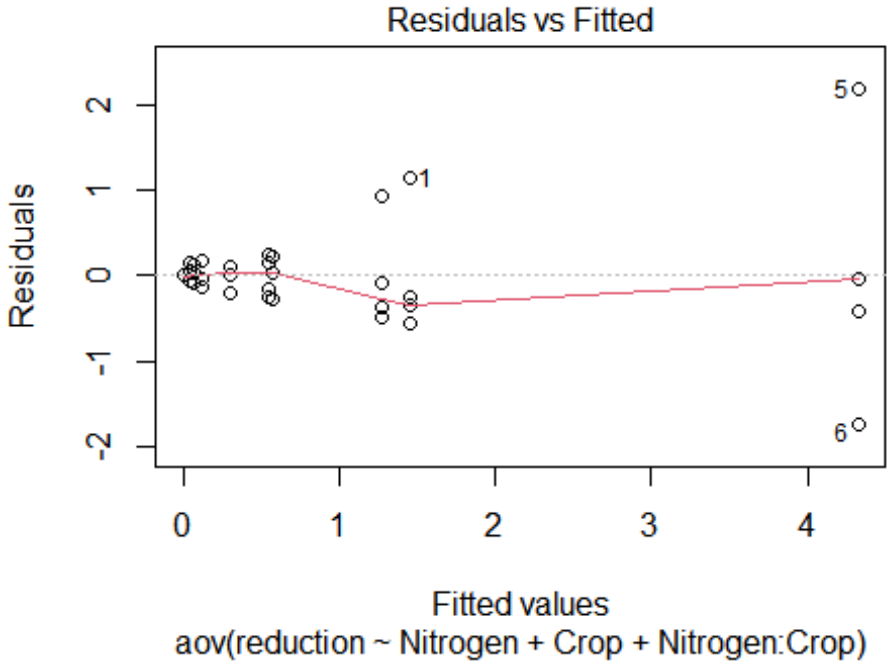
```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## Nitrogen       2  29.85  14.923   47.14 8.83e-11 ***
## Crop           3  15.12   5.039   15.92 9.44e-07 ***
## Nitrogen:Crop  6  22.10   3.684   11.63 3.26e-07 ***
```

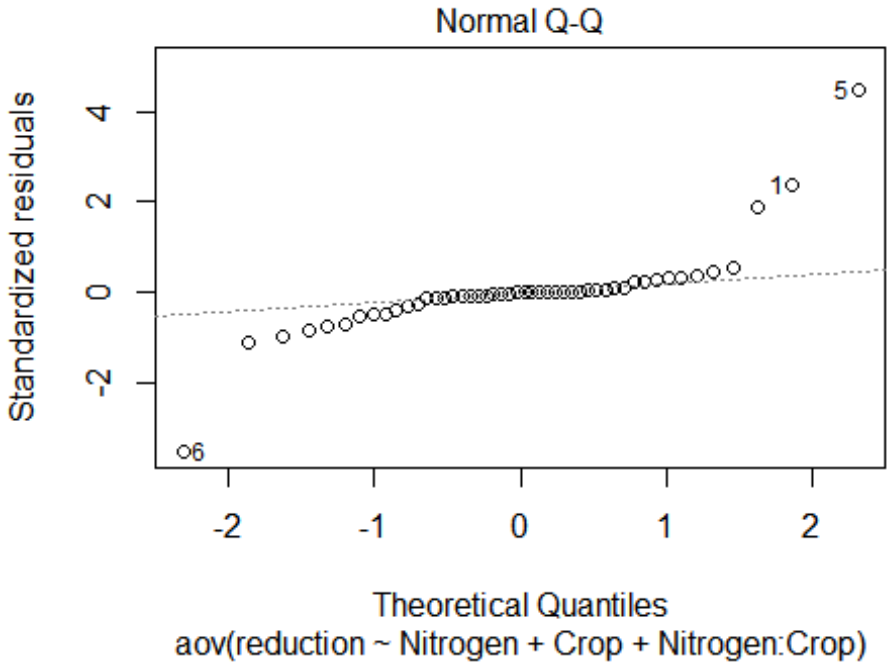
```
## Residuals     36  11.40   0.317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There does appear to be an interaction between the two factors as the p-value for the interaction is 3.26e-07<.05. Reject the null of no interaction and accept the hypothesis of an interaction between the factors.

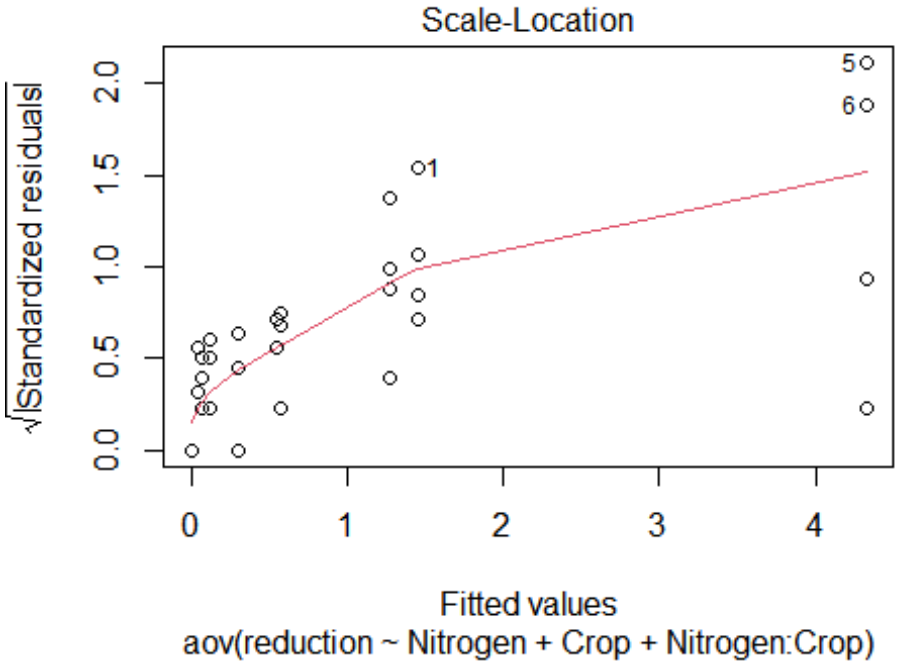
4. What assumptions are not met? (Same assumptions as linear regression, since it is linear regression.)

```
plot(anova_results)
```






```
## hat values (leverages) are all = 0.25
##  and there are no factor predictors; no plot no. 5
```



First, since the scale-location plot has a positive slope, it appears that the variance is not constant across the fitted values. Also, the qq-plot is not entirely linear, as it has strong tail-offs at either end.