hw 5

Dongyang Wang

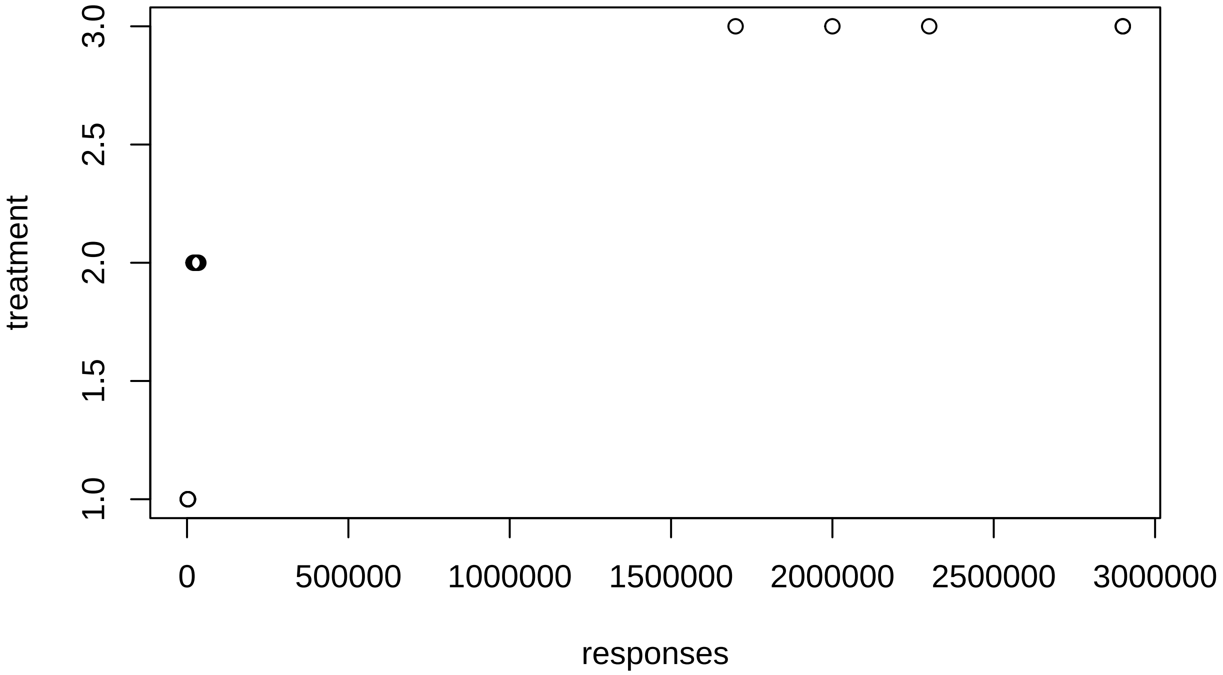
11/14/2021

1

(a)

```
dat <- data.frame(responses = c(2600, 2900, 2000, 2200,3200,
35000,23000,20000,30000,27000,
2900000,2300000,1700000,2900000,2000000),
treatment = as.factor(c(rep("1",5),rep("2",5),rep("3",5))))

plot(dat)
```



```
var(dat$responses[1:5])
```

```
## [1] 242000
```

```
var(dat$responses[6:10])
```

```
## [1] 34500000
```

```
var(dat$responses[11:15])
```

```
## [1] 2.88e+11
```

```
sd((dat$responses[11:15]))/sd(dat$responses[1:5])
```

```
## [1] 1090.909
```

First, the data seems not normally distributed. Also, across different treatments, the variances differ greatly (the highest/lowest standard deviation ratio is way larger than 7). This violates our linear model ANOVA assumption that the variance should be the same.

(b)

```
anova_model <- lm(dat$responses ~ dat$treatment)
anova(anova_model)
```

```
## Analysis of Variance Table
##
## Response: dat$responses
##               Df     Sum Sq    Mean Sq F value    Pr(>F)
## dat$treatment  2 1.8335e+13 9.1674e+12  95.483 4.271e-08 ***
## Residuals     12 1.1521e+12 9.6012e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

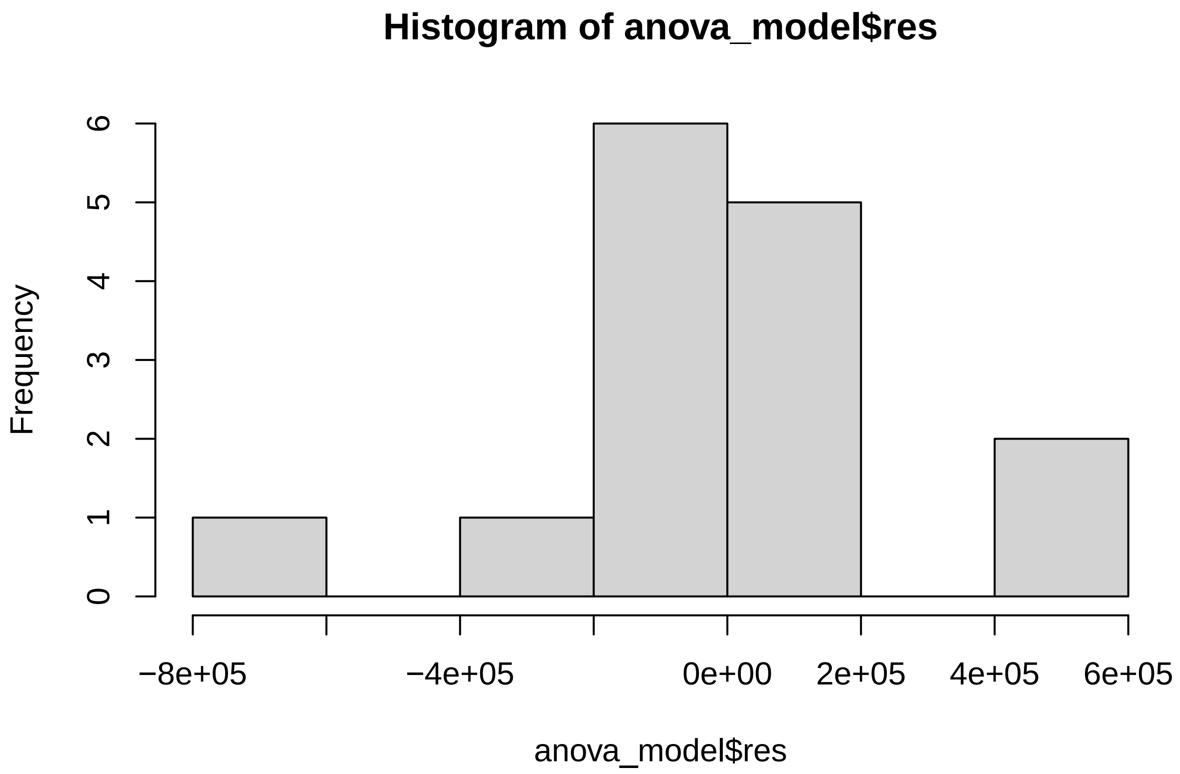
```
anova_model$res
```

```
##             1             2             3             4             5
##  2.000000e+01  3.200000e+02 -5.800000e+02 -3.800000e+02  6.200000e+02
##             6             7             8             9            10
##  8.000000e+03 -4.000000e+03 -7.000000e+03  3.000000e+03 -7.807088e-13
##            11            12            13            14            15
##  5.400000e+05 -6.000000e+04 -6.600000e+05  5.400000e+05 -3.600000e+05
```

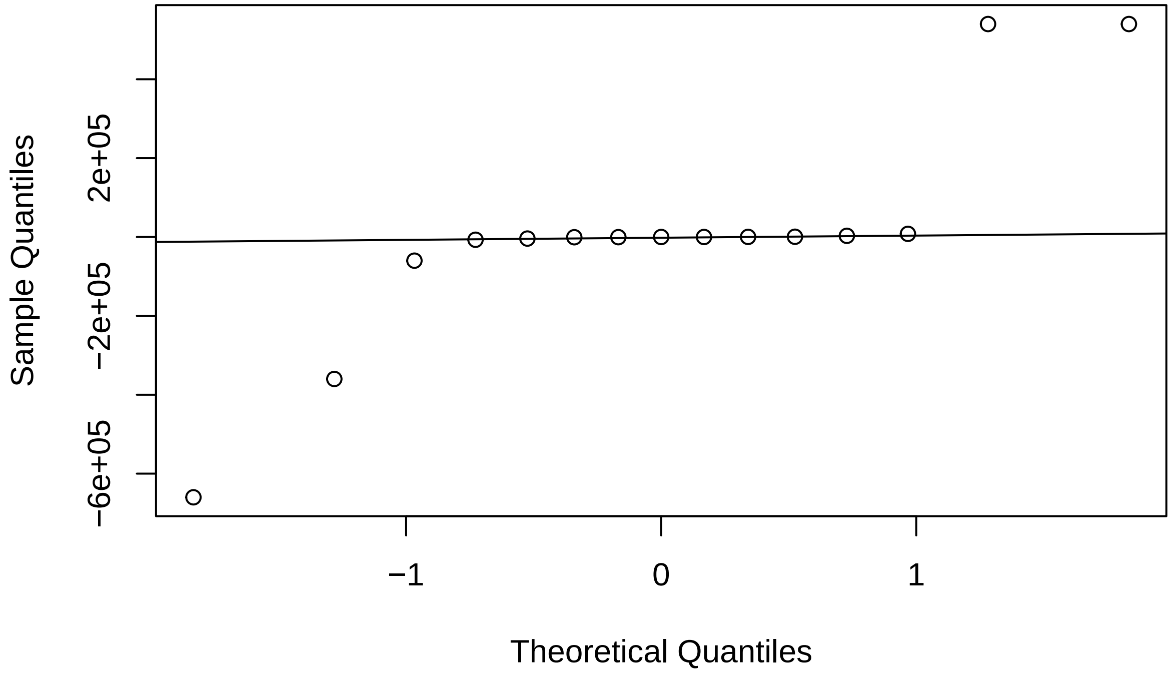
```
# expectation
sum(anova_model$res)
```

```
## [1] 1.548415e-10
```

```
# normality
hist(anova_model$res)
```



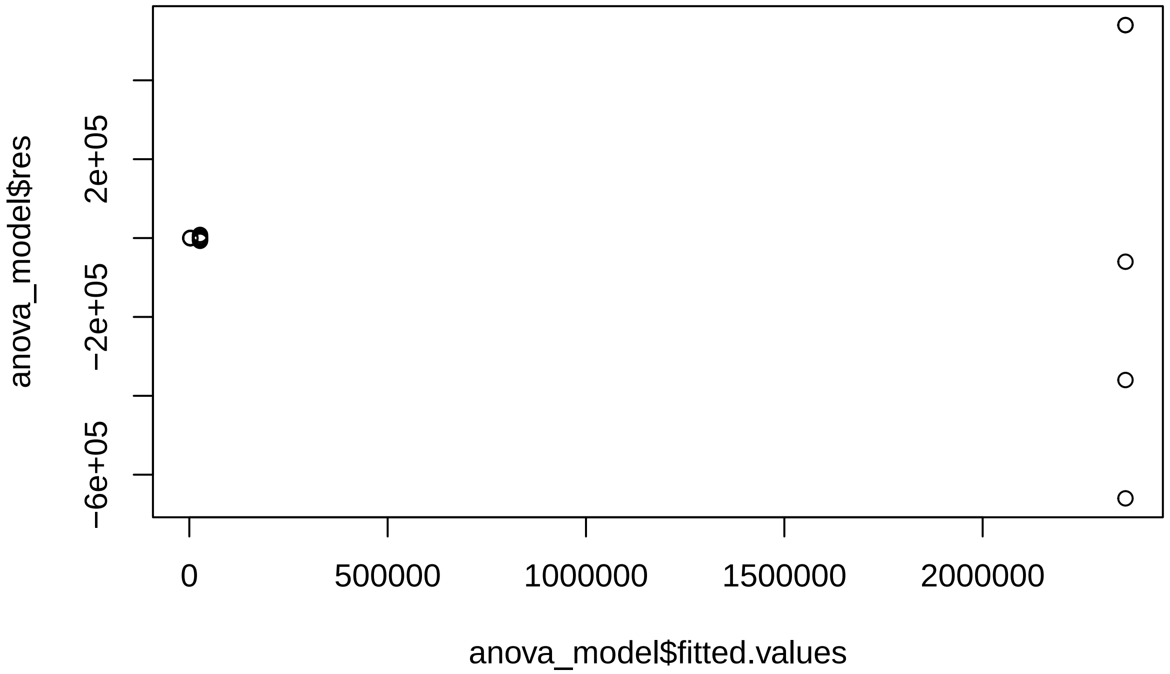
```
qqnorm(anova_model$res,main="") ; qqline(anova_model$res)
```



```
# homoskedasticity
anova_model$fitted.values
```

```
##       1       2       3       4       5       6       7       8       9      10
##    2580    2580    2580    2580    2580   27000   27000   27000   27000   27000
##      11      12      13      14      15
## 2360000 2360000 2360000 2360000 2360000
```

```
plot(anova_model$fitted.values, anova_model$res)
```



Zero expectation of residuals holds. No observable trend for the histogram, possibly because the data has only 15 entries. The qq plot shows, however, that the residuals are not normally distributed. Also from the mean-variance relationship, we verify the argument in (a) such that the variances are not constant. Therefore, we need to transform the data.

(c)

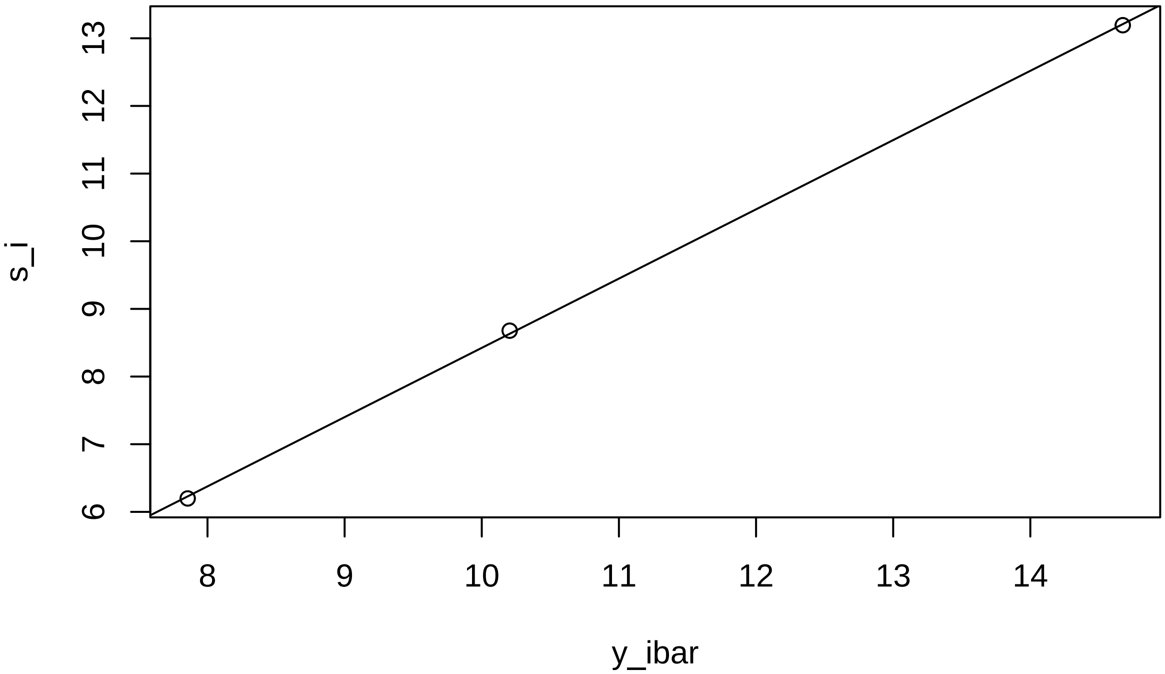
```
s_i <- log(c(sd(dat$responses[1:5]), sd(dat$responses[6:10]), sd(dat$responses[11:15])))
s_i
```

```
## [1]  6.198347  8.678235 13.193113
```

```
y_ibar <- log(c(mean(dat$responses[1:5]), mean(dat$responses[6:10]), mean(dat$responses[11:15])))
y_ibar
```

```
## [1]  7.855545 10.203592 14.674172
```

```
# show linear relationship
plot(y_ibar,s_i)
abline(lm(s_i ~ y_ibar))
```



```
# obtain alpha
lm(s_i ~ y_ibar)
```

```
##
## Call:
## lm(formula = s_i ~ y_ibar)
##
## Coefficients:
## (Intercept)       y_ibar
##      -1.813        1.024
```

Although there are only 3 points, we can see a clear linear relationship between the log values of the response and the log values of the sample standard deviation . Therefore, we can do a BoxCox transformation of the response with $\hat{\alpha} = 1.024$.

(d)

```
# transform
dat$y_star <- dat$responses^(1- 1.024)
new_model <- lm(dat$y_star ~ dat$treatment)
anova(new_model)
```

```
## Analysis of Variance Table
##
## Response: dat$y_star
##               Df   Sum Sq   Mean Sq F value   Pr(>F)
## dat$treatment  2 0.040013 0.0200066  1268.3 1.09e-14 ***
## Residuals     12 0.000189 0.0000158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

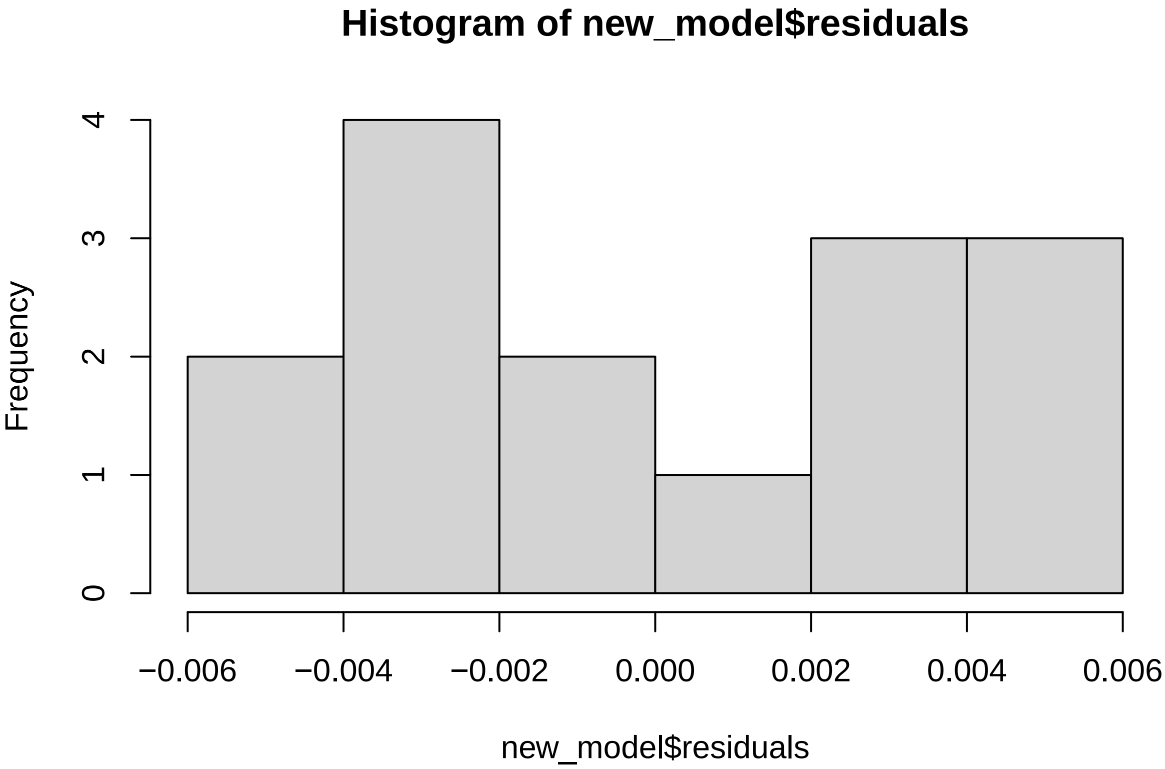
```
sum(new_model$residuals)
```

```
## [1] -5.014435e-19
```

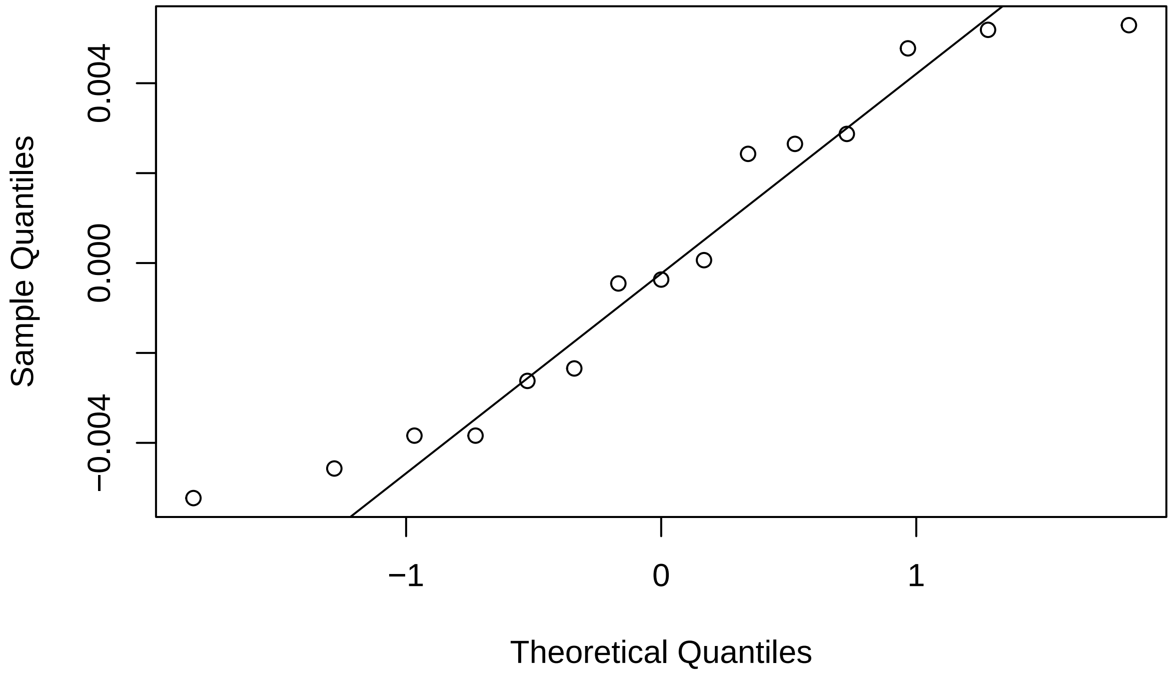
```r
# residuals
new_model$residuals
```

```
##             1             2             3             4             5 
## -4.546911e-04 -2.621910e-03  4.775585e-03  2.871750e-03 -4.570734e-03 
##             6             7             8             9            10 
## -5.228176e-03  2.650290e-03  5.290554e-03 -2.344790e-03 -3.678781e-04 
##            11            12            13            14            15 
## -3.839399e-03  6.396268e-05  5.186873e-03 -3.839399e-03  2.427962e-03
```

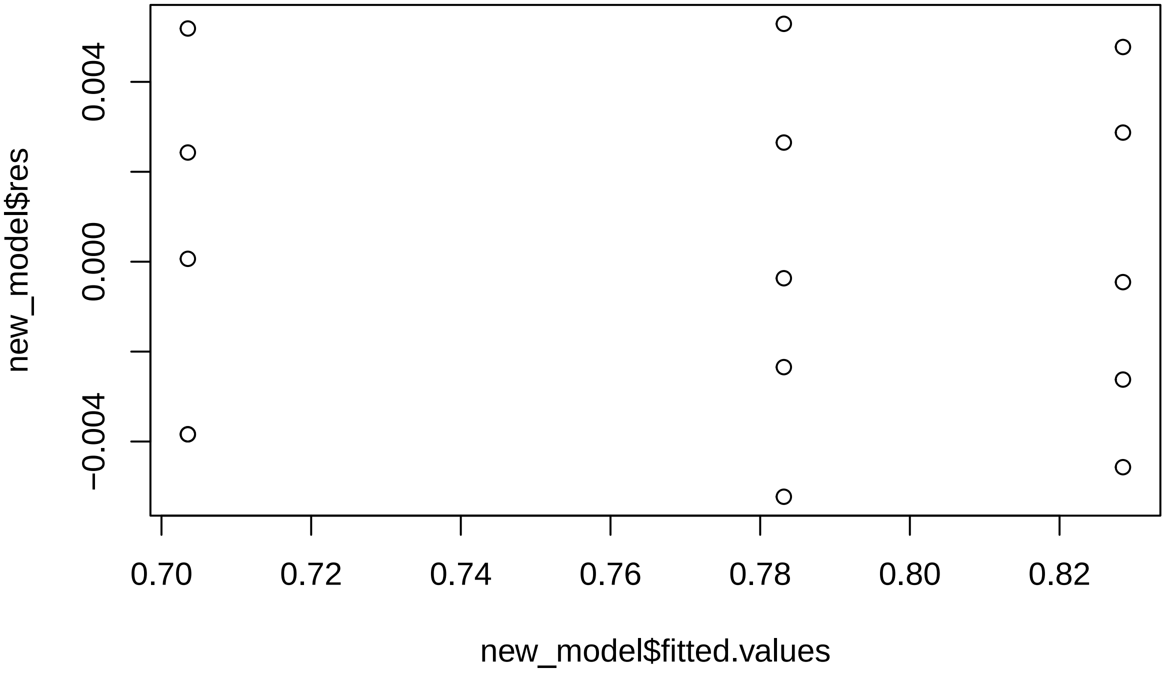
```r
hist(new_model$residuals)
```



```r
qqnorm(new_model$residuals,main="") ; qqline(new_model$residuals)
```



```
plot(new_model$fitted.values, new_model$res)
```



```
var((new_model$res[11:15]))/var(new_model$res[1:5])
```

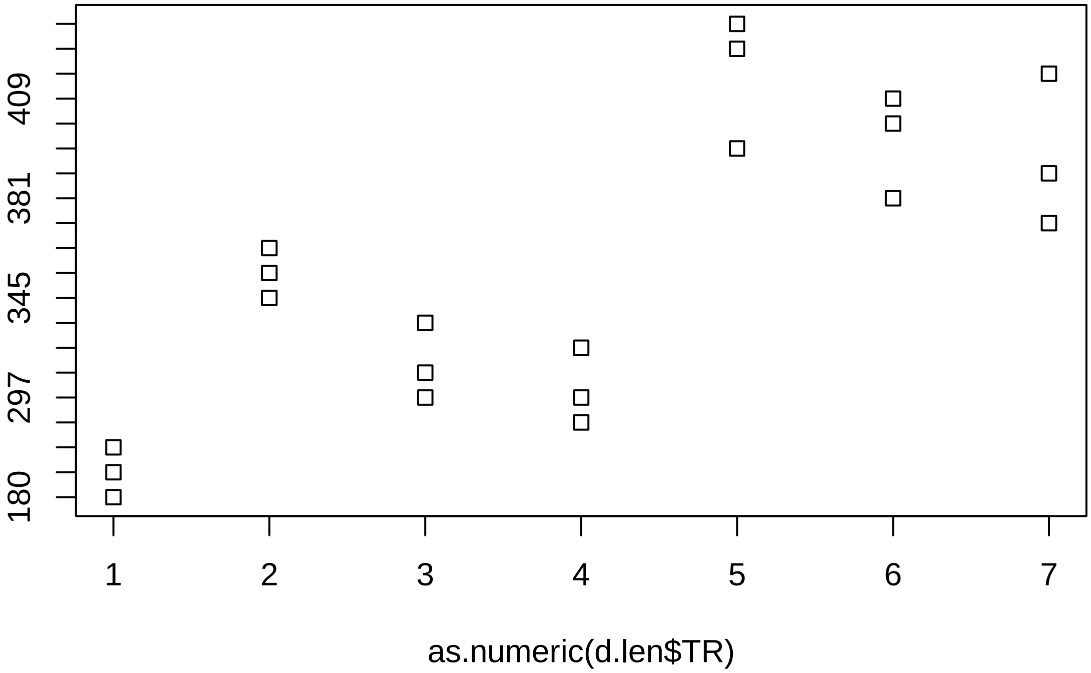
```
## [1] 1.055209
```

Due to limited data, the histogram is not very informative. But the qq plot shows clearly the normality of the residuals. The mean-variance plot shows the same variance, and the highest/lowest standard deviation ratio is smaller than 2. The assumptions for the linear model have been satisfied. Therefore, we can use the linear ANOVA model after applying the BoxCox transformation.

2

(a)

```
d.len <- read.table("lentil.dat", header = TRUE)
d.len$TR <- factor(d.len$TR)
stripchart(as.numeric(d.len$TR) ~ d.len$Y)
```



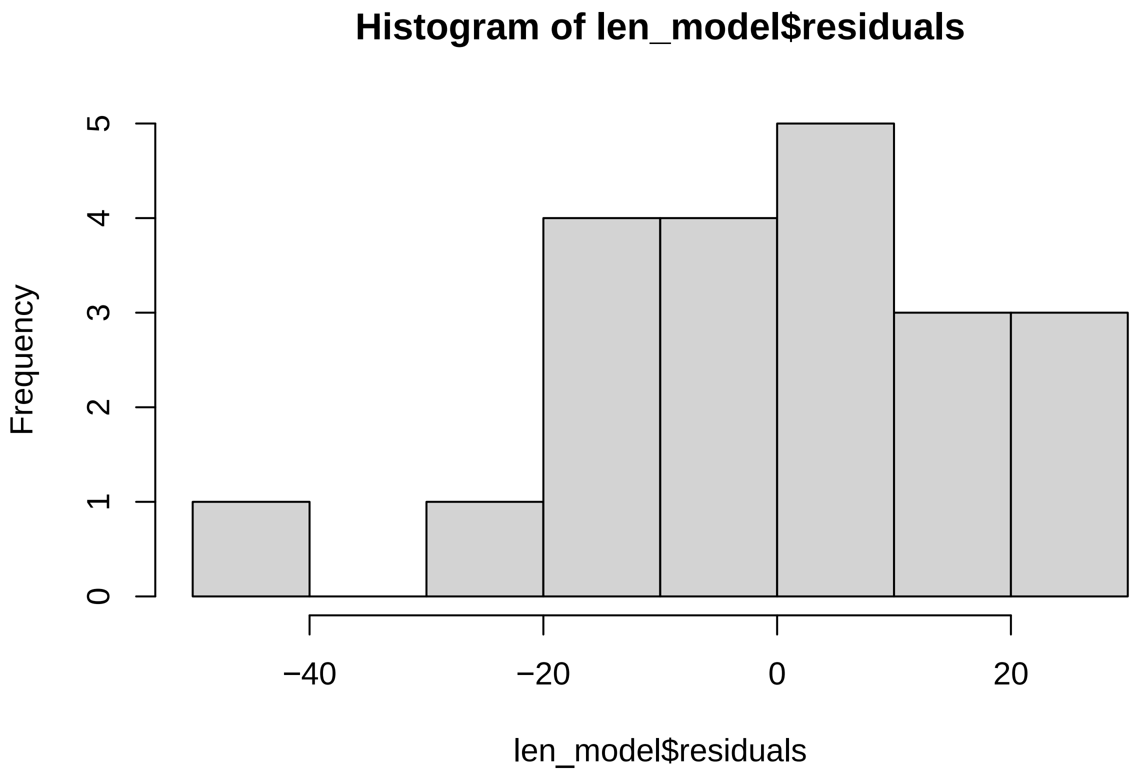
```
len_model <- lm(d.len$Y ~ d.len$TR)

sum((len_model$residuals))
```

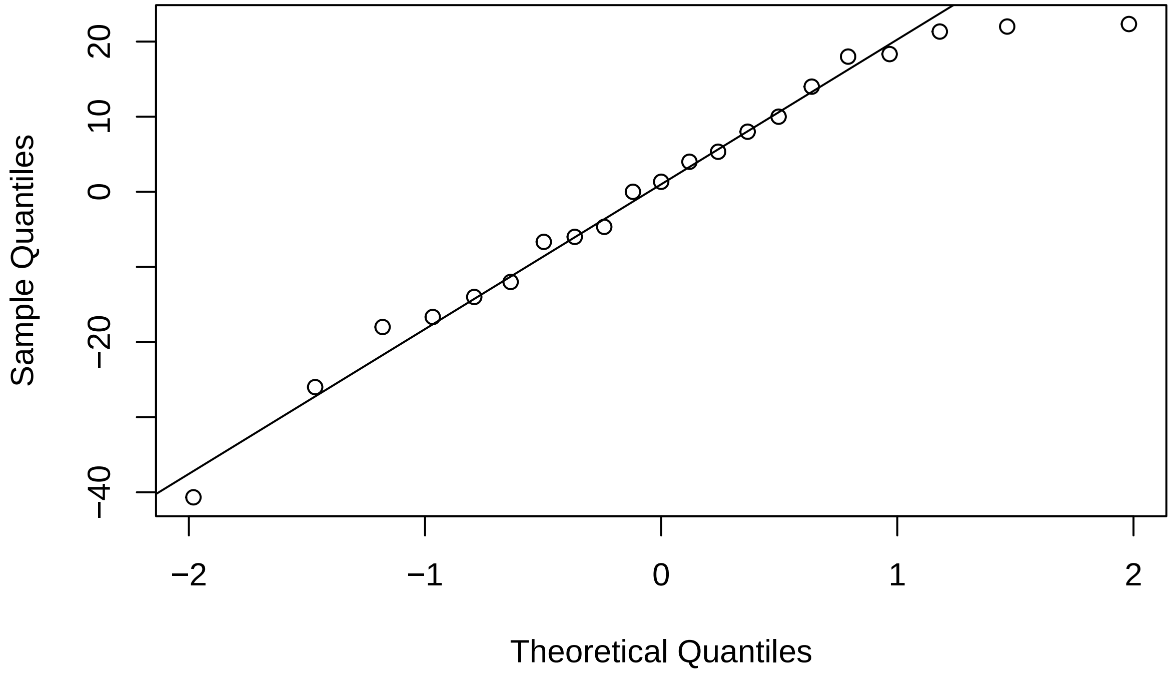
```
## [1] 2.04281e-14
```

```
# plots

hist(len_model$residuals)
```



```
qqnorm(len_model$residuals,main="") ; qqline(len_model$residuals)
```

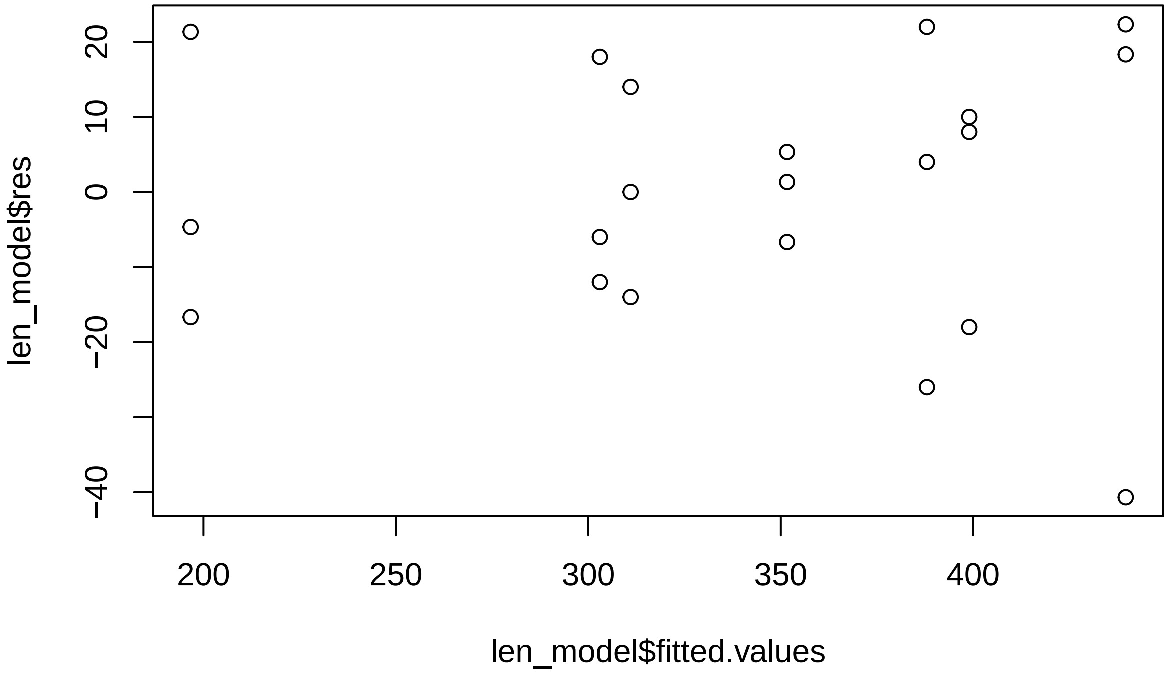


```
plot(len_model$fitted.values, len_model$res)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```



```
d.len %>% select(TR, Y) %>% group_by(TR) %>% summarize(varY = var(Y)) %>% mutate(max(varY)/min(varY))
```

```
## # A tibble: 7 x 3
##   TR       varY `max(varY)/min(varY)`
##   <fct>   <dbl>                 <dbl>
## 1 1      377.                    33.3
## 2 2       37.3                   33.3
## 3 3      196                     33.3
## 4 4      252                     33.3
## 5 5     1244.                    33.3
## 6 6      244                     33.3
## 7 7      588                     33.3
```

```
# or use
max(by(len_model$res, d.len$TR, var))/min(by(len_model$res, d.len$TR, var))
```

```
## [1] 33.33036
```

```
anova(len_model)
```

```
## Analysis of Variance Table
##
## Response: d.len$Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## d.len$TR   6 115792 19298.7  45.965 2.028e-08 ***
## Residuals 14   5878   419.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We should be careful in applying the linear model. Zero expectation of residuals holds. The histogram is right tailed but only one outlier is on the left. The qq plot seems fine. However, the variance varies a lot because the ratio is 33.33036 and above 7. This violates the constant variance assumption of ANOVA. If we use the ANOVA, we reject the null hypothesis at the 0.002 level,i.e., we can claim that there is difference between the treatments.

(b)(i)

```
# orthogonality
A = matrix(c(-6, +1, +1, +1, +1, +1,
             +1, 0, -1, -1 ,-1, +1,
             +1, +1, 0, +2, -1, -1,
             +2, -1, -1, 0, 0, -1,
             +1, 0 ,-1 ,+1, 0 ,-2 ,
             +1 ,+1 ,+2 ,-1 ,-1,
             0, 0, +1, -1, 0 ,-1 ,+1), nrow = 6, byrow = T)

for (i in 1:(nrow(A)-1)){
  for (j in (i+1):nrow(A)){
    print(sum(A[i,] * A[j,] ) )
  }
}
```

```
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
```

```
# An easier way is to claim that only diagonals are non-zero.
# That means each pair of different rows has a sum of 0.
A %*% t(A)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]   42    0    0    0    0    0
## [2,]    0    6    0    0    0    0
## [3,]    0    0   12    0    0    0
## [4,]    0    0    0    4    0    0
## [5,]    0    0    0    0   12    0
## [6,]    0    0    0    0    0    4
```

Yes, the contrasts are orthogonal.

(b)(ii) $C_1 = -6k_1 + k_2 + k_3 + k_4 + k_5 + k_6 + k_7$. This contrast tries to test if at least one of the treatments is useful, as we are comparing the sum of the 6 treatments with the 1 null/control treatment. $C_2 = -k_2 - k_3 - k_4 + k_5 + k_6 + k_7$. This contrast tries to test the effect of fertilizer. $C_3 = 2k_2 - k_3 - k_4 + 2k_5 - k_6 - k_7$.

This contrast tries to test the effect of hand vs the effect of spraying the herbicide. $C_4 = -k_3 + k_4 - k_6 + k_7$. This contrast tries to test the effect of spraying the herbicide before and after.

(c)

```
#install.packages('multcomp')
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```
#?glht
attach(d.len)

fit.len <- lm(Y ~ TR)
fit.mc <- glht(fit.len, linfct = mcp(TR = A))
summary(fit.mc, test = adjusted("none"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: lm(formula = Y ~ TR)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0  1012.33      76.67  13.204 2.72e-09 ***
## 2 == 0   261.00      28.98   9.007 3.36e-07 ***
## 3 == 0   181.67      40.98   4.433 0.000568 ***
## 4 == 0   -19.00      23.66  -0.803 0.435378
## 5 == 0     3.00      40.98   0.073 0.942679
## 6 == 0    -3.00      23.66  -0.127 0.900906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

```
detach(d.len)
```

The first three contrasts are significant. That means, the treatments are useful, in particular fertilizer and herbicide/hand weeding.

3

(a)

```
y <- c(9, 12, 10, 8, 15,
20, 21, 23, 17, 30,
6, 5, 8, 16, 7)
type <- c(rep("type 1",5),rep("type 2",5),rep("type 3",5))
circ <- data.frame(Type = type, Y = y)
circ$Type <- as.factor(circ$Type)
circ
```

```
##      Type  Y
## 1  type 1  9
## 2  type 1 12
## 3  type 1 10
## 4  type 1  8
## 5  type 1 15
## 6  type 2 20
## 7  type 2 21
## 8  type 2 23
## 9  type 2 17
## 10 type 2 30
## 11 type 3  6
## 12 type 3  5
## 13 type 3  8
## 14 type 3 16
## 15 type 3  7
```

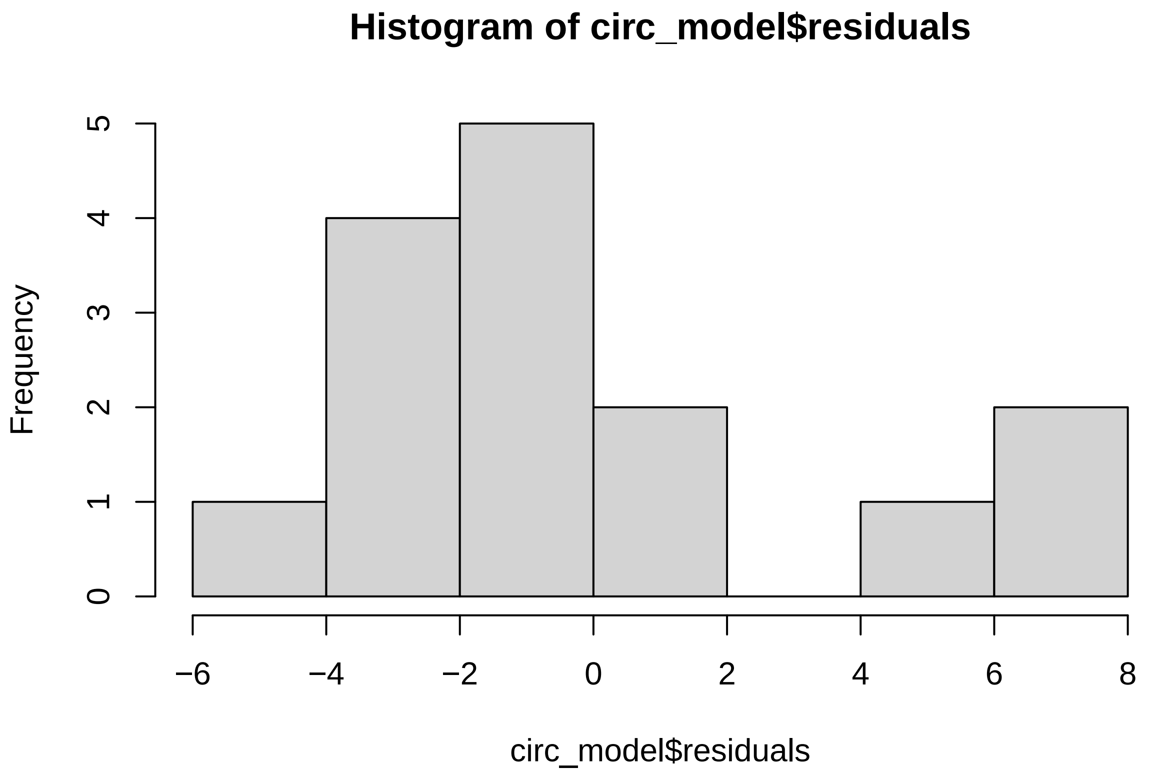
```
# anova
circ_model <- lm(Y ~ Type, data = circ)
anova(circ_model)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Type       2  543.6   271.8  16.083 0.0004023 ***
## Residuals 12  202.8    16.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

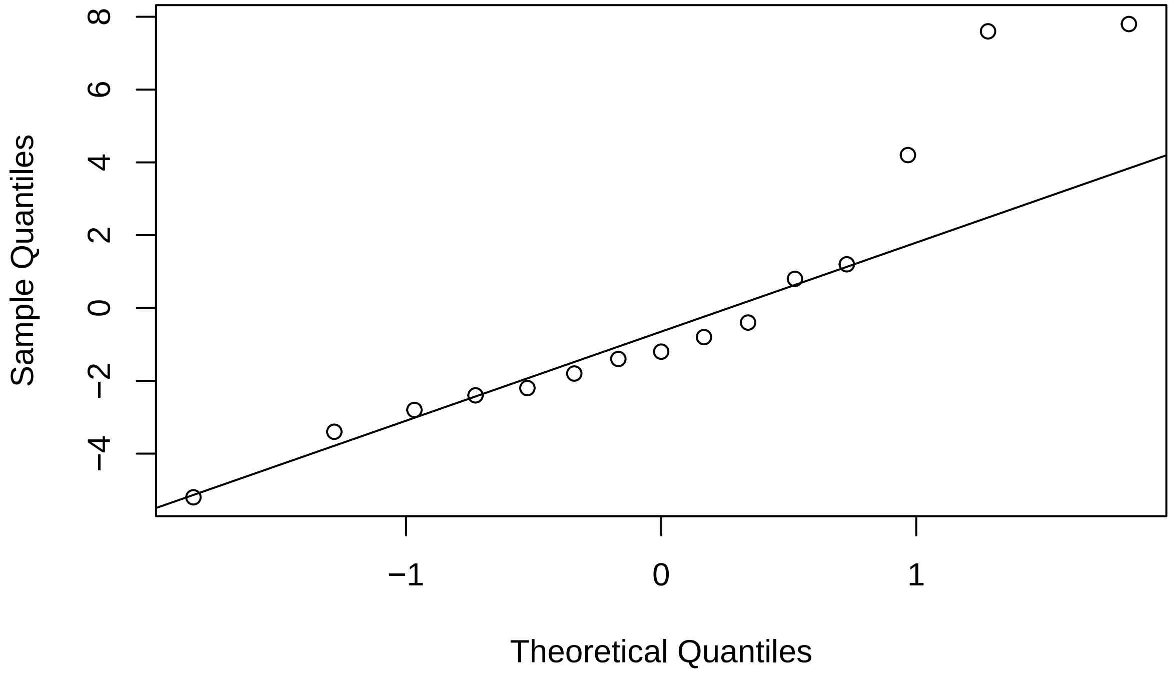
```
# assumption: expectation zero
sum(circ_model$residuals)
```

```
## [1] 1.221245e-15
```

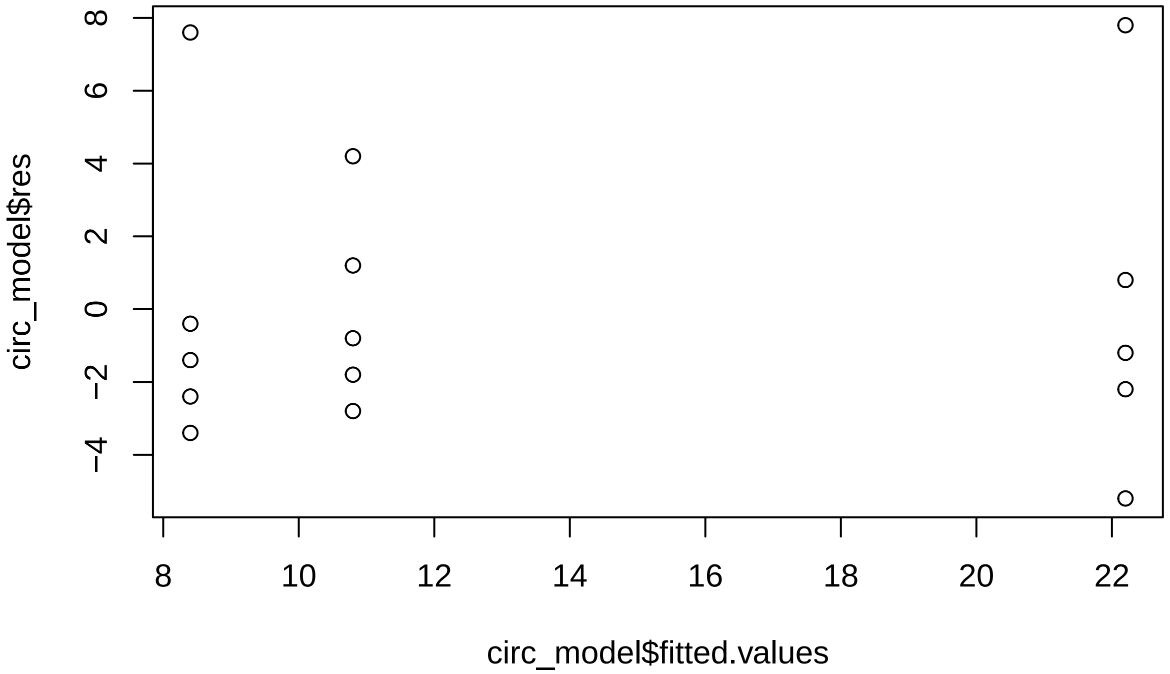
```
# assumption: normality
hist(circ_model$residuals)
```



```
qqnorm(circ_model$residuals,main="") ; qqline(circ_model$residuals)
```



```
# assumption: variance
plot(circ_model$fitted.values, circ_model$res)
```



```
max(by(circ_model$res, circ$Type, var))/min(by(circ_model$res, circ$Type, var))
```

```
## [1] 3.077922
```

The model tells us that we can reject the null at 0.01 level. The assumption of zero expectation and homoskedasticity hold. Normality, however, does not seem to hold.

(b)

```
contrasts <- matrix(c(-1,2,-1, -1,0,1), nrow = 2, byrow = T)

# orthogonal
contrasts %*% t(contrasts)
```

```
##      [,1] [,2]
## [1,]    6    0
## [2,]    0    2
```

```
contrasts
```

```
##      [,1] [,2] [,3]
## [1,]   -1    2   -1
## [2,]   -1    0    1
```

(c)

```
library(multcomp)
circ.mc <- glht(circ_model, linfct = mcp(Type = contrasts))
summary(circ.mc, test = adjusted("holm"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: lm(formula = Y ~ Type, data = circ)
```

```
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0   25.200      4.503   5.596 0.000234 ***
## 2 == 0   -2.400      2.600  -0.923 0.374155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- holm method)
```

```r
critical <- c(0.01/2, 0.01)
critical
```

```
## [1] 0.005 0.010
```

With a family wise level at 0.01, we can only rejects the null hypothesis for the first contrast. We do not reject the second contrast.