# HW6

Hanao Li hl3202

## Problem 1

```
if(!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
p_load(nlme)

data <- ChickWeight

day_16 <- merge(data[which(data$Time == 16), ], data[which(data$Time == 0),], by = "Chic
k")
anova_16 <- aov(day_16$weight.x~day_16$Diet.x + day_16$weight.y)
summary(anova_16)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## day_16$Diet.x    3  19983    6661   3.633 0.0203 *
## day_16$weight.y  1   4201    4201   2.291 0.1376
## Residuals       42  77015    1834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
day_20 <- merge(data[which(data$Time == 20), ], data[which(data$Time == 0),], by = "Chic
k")
anova_20 <- aov(day_20$weight.x ~ day_20$Diet.x + day_20$weight.y)
summary(anova_20)
```

```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## day_20$Diet.x    3  55881   18627   5.594 0.00261 **
## day_20$weight.y  1   6672    6672   2.004 0.16447
## Residuals       41 136519    3330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
day_21 <- merge(data[which(data$Time == 21), ], data[which(data$Time == 0),], by = "Chic
k")
anova_21 <- aov(day_21$weight.x ~ day_21$Diet.x + day_21$weight.y)
summary(anova_21)
```

```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## day_21$Diet.x    3  57164   19055   4.743 0.00636 **
## day_21$weight.y  1   7137    7137   1.776 0.19014
## Residuals       40 160703    4018
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# From the above results, the p-value is 0.02 for day16, 0.002 for day20 and 0.006 for d
ay21. They are all smaller than 0.05. Thus, we reject the null hypothesis. And we can co
nclude that there is a significant difference in mean weight of the four groups of chick
en on Day 16, Day 20 and Day 21 given adjusted for baseline.
```

## Problem 2

```
combined <- data[which(data$Time == 0 | data$Time == 16 | data$Time == 20 | data$Time ==
21), ]
anova_combined <- aov(combined$weight ~ combined$Diet*combined$Time + Error(combined$Chi
ck))
summary(anova_combined)
```

```
##
## Error: combined$Chick
##                             Df Sum Sq Mean Sq F value Pr(>F)
## combined$Diet                3 110563   36854   5.891 0.0018 **
## combined$Time                1  37146   37146   5.938 0.0189 *
## combined$Diet:combined$Time  1   2552    2552   0.408 0.5263
## Residuals                   44 275249    6256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: Within
##                              Df Sum Sq Mean Sq F value   Pr(>F)
## combined$Time                 1 919369  919369 1026.84  < 2e-16 ***
## combined$Diet:combined$Time   3  37421   12474   13.93 5.84e-08 ***
## Residuals                   134 119975     895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#From the results,we could see that the p-value is less than 0.05, we reject the null. We can conclude that there is a significant difference in mean wieght of the four groups of chicken given adjusted for the baseline for Day16, Day20 and Day21. Also, the result seems to be more significant than the previous problem, if we considered the interaction effets. The within error indicates that there might be a two-way interaction effect between Time and Diet.

# Compound symmatry
grouped <- groupedData(weight ~ as.numeric(Diet) * as.numeric(Time) | Chick, data = combined)
compound <- gls(weight ~ Diet * Time, data = grouped, corr = corCompSymm(, form = ~ 1 | Chick))
summary(compound)
```

```
## Generalized least squares fit by REML
##   Model: weight ~ Diet * Time
##   Data: grouped
##        AIC      BIC    logLik
##   1885.902 1917.831 -932.9508
##
## Correlation Structure: Compound symmetry
##  Formula: ~1 | Chick
##  Parameter estimate(s):
##      Rho
## 0.597605
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 41.41562 10.442017  3.966247  0.0001
## Diet2       -2.07708 18.071455 -0.114937  0.9086
## Diet3       -3.56320 18.071455 -0.197173  0.8439
## Diet4       -1.89466 18.075374 -0.104820  0.9166
## Time         6.44083  0.421688 15.273907  0.0000
## Diet2:Time   1.77577  0.699908  2.537145  0.0120
## Diet3:Time   4.36602  0.699908  6.237997  0.0000
## Diet4:Time   2.91453  0.711240  4.097821  0.0001
##
##  Correlation:
##            (Intr) Diet2  Diet3  Diet4  Time   Dt2:Tm Dt3:Tm
## Diet2      -0.578
## Diet3      -0.578  0.334
## Diet4      -0.578  0.334  0.334
## Time       -0.508  0.293  0.293  0.293
## Diet2:Time  0.306 -0.528 -0.177 -0.177 -0.602
## Diet3:Time  0.306 -0.177 -0.528 -0.177 -0.602  0.363
## Diet4:Time  0.301 -0.174 -0.174 -0.523 -0.593  0.357  0.357
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.93430887 -0.57478875  0.01325605  0.46592128  2.73086747
##
## Residual standard error: 46.99131
## Degrees of freedom: 188 total; 180 residual
```

```
# Unstructed
unstructed <- gls(weight ~ Diet * Time, data = grouped, corr = corSymm(form = ~ 1 | Chic
k), weights = varIdent(form = ~ 1 | Time))
summary(unstructed)
```

```
## Generalized least squares fit by REML
##   Model: weight ~ Diet * Time
##   Data: grouped
##       AIC      BIC    logLik
##   1386.58 1444.054 -675.2901
##
## Correlation Structure: General
##  Formula: ~1 | Chick
##  Parameter estimate(s):
##  Correlation:
##   1      2      3
## 2 -0.233
## 3 -0.241  0.915
## 4 -0.237  0.889  0.993
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | Time
##  Parameter estimates:
##        0       16       20       21
##  1.00000 38.59186 52.70615 57.17781
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 41.40049 0.2518996 164.35313  0.0000
## Diet2       -0.67434 0.4362344  -1.54582  0.1239
## Diet3       -0.54147 0.4362344  -1.24123  0.2161
## Diet4       -0.37349 0.4362801  -0.85608  0.3931
## Time         6.58575 0.6604751   9.97124  0.0000
## Diet2:Time   1.17323 1.0886925   1.07766  0.2826
## Diet3:Time   3.23980 1.0886925   2.97587  0.0033
## Diet4:Time   2.30284 1.0887252   2.11517  0.0358
##
##  Correlation:
##            (Intr) Diet2  Diet3  Diet4  Time   Dt2:Tm Dt3:Tm
## Diet2      -0.577
## Diet3      -0.577  0.333
## Diet4      -0.577  0.333  0.333
## Time       -0.239  0.138  0.138  0.138
## Diet2:Time  0.145 -0.252 -0.084 -0.084 -0.607
## Diet3:Time  0.145 -0.084 -0.252 -0.084 -0.607  0.368
## Diet4:Time  0.145 -0.084 -0.084 -0.252 -0.607  0.368  0.368
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.13332227 -0.51094380  0.07301875  0.72123544  2.52980366
##
## Residual standard error: 1.128037
## Degrees of freedom: 188 total; 180 residual
```

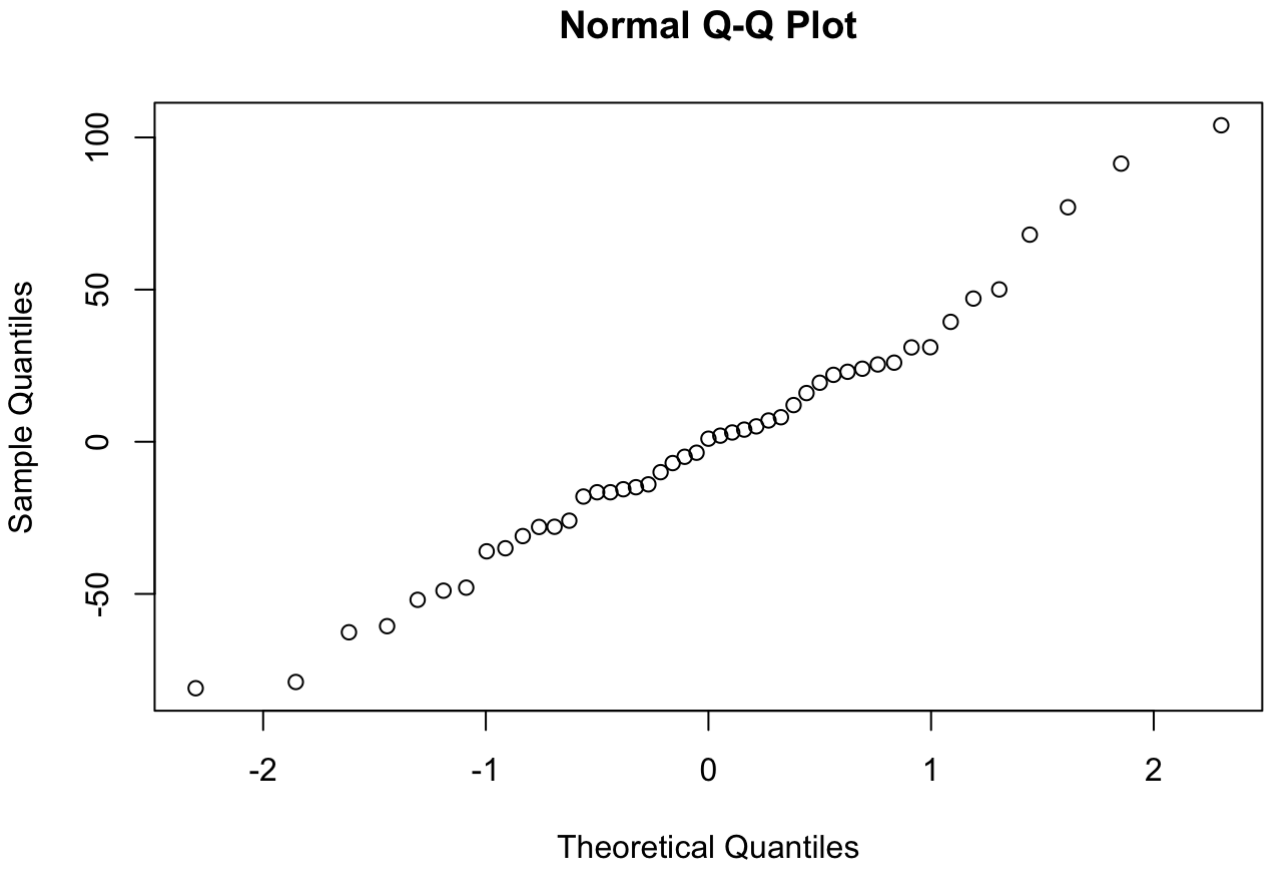
```
# Compare two models
anova(compound, unstructed)
```

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## compound       1 10 1885.902 1917.831 -932.9508
## unstructed     2 18 1386.580 1444.053 -675.2901 1 vs 2 515.3214  <.0001
```

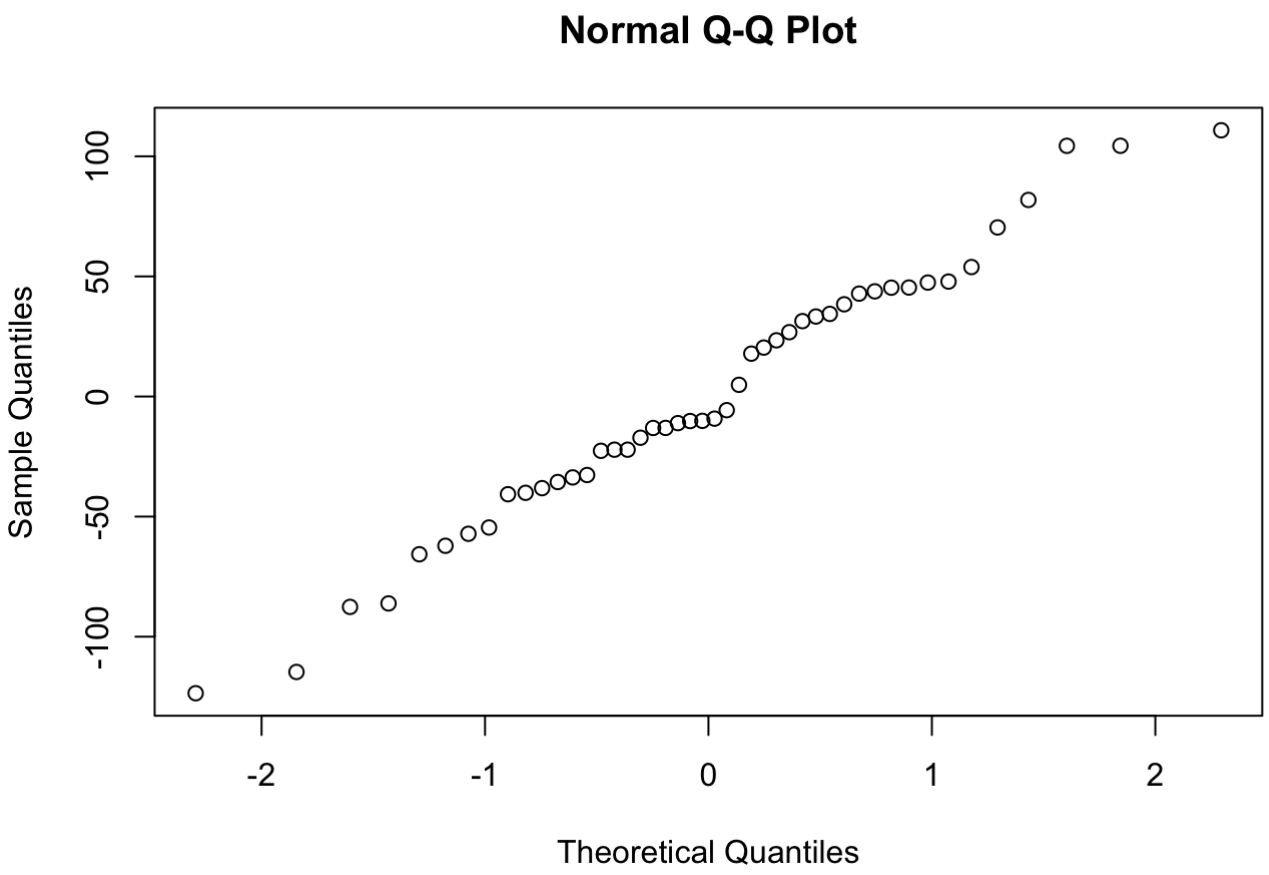
```
# From the above result, we can find that the unstructed model is a better model because it has a smaller AIC and a larger negative log likelihood, which makes it more precise. We can also see that the p-valued of the diet are all larger than 0.05. We do not reject the null and conclude that the there is not a significant difference in mean weight of the four groups of chicken on Day16, Day20 and Day21 given adjusted for the baseline and interaction between diet and time.
```

## Problem 3

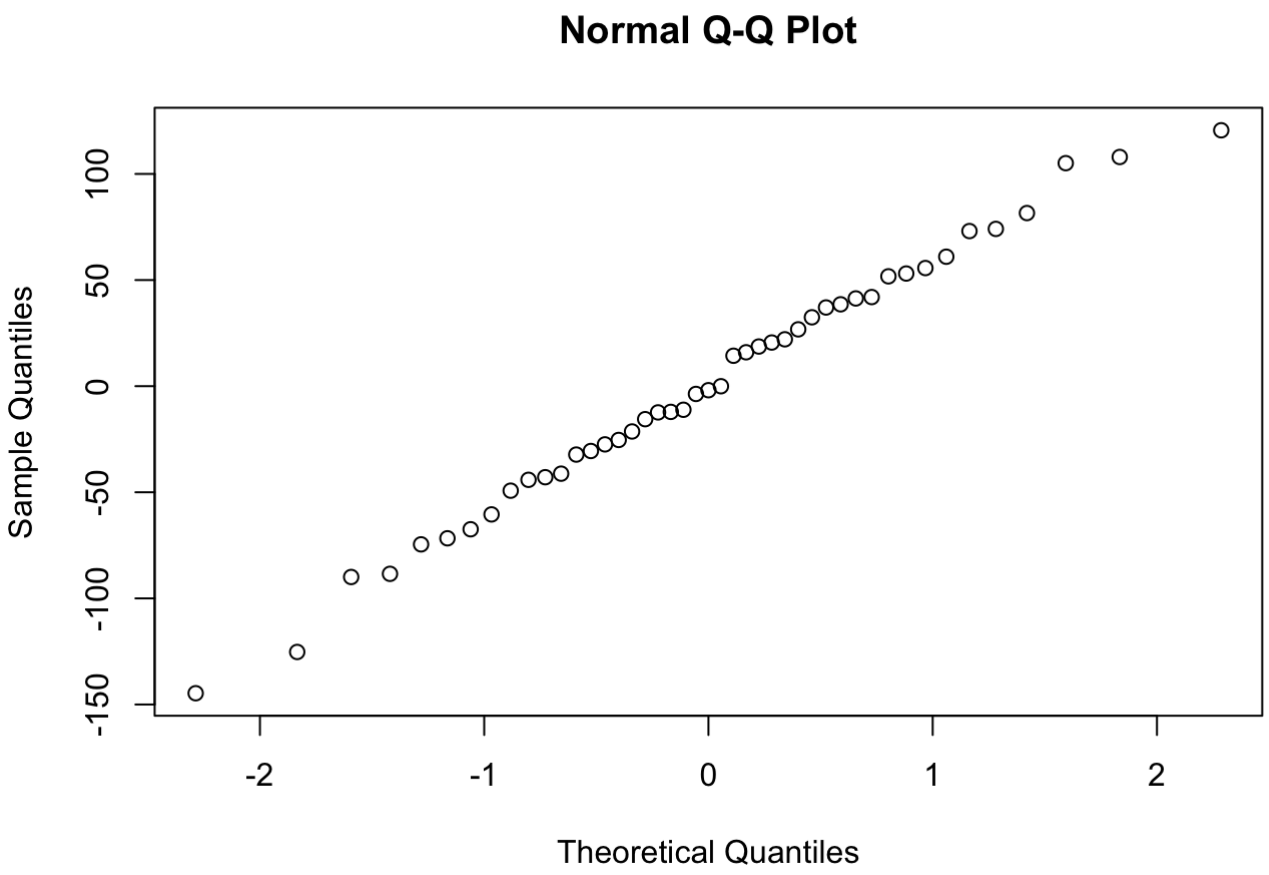
```
# Normality and Constant Variance
qqnorm(anova_16$residuals)
```



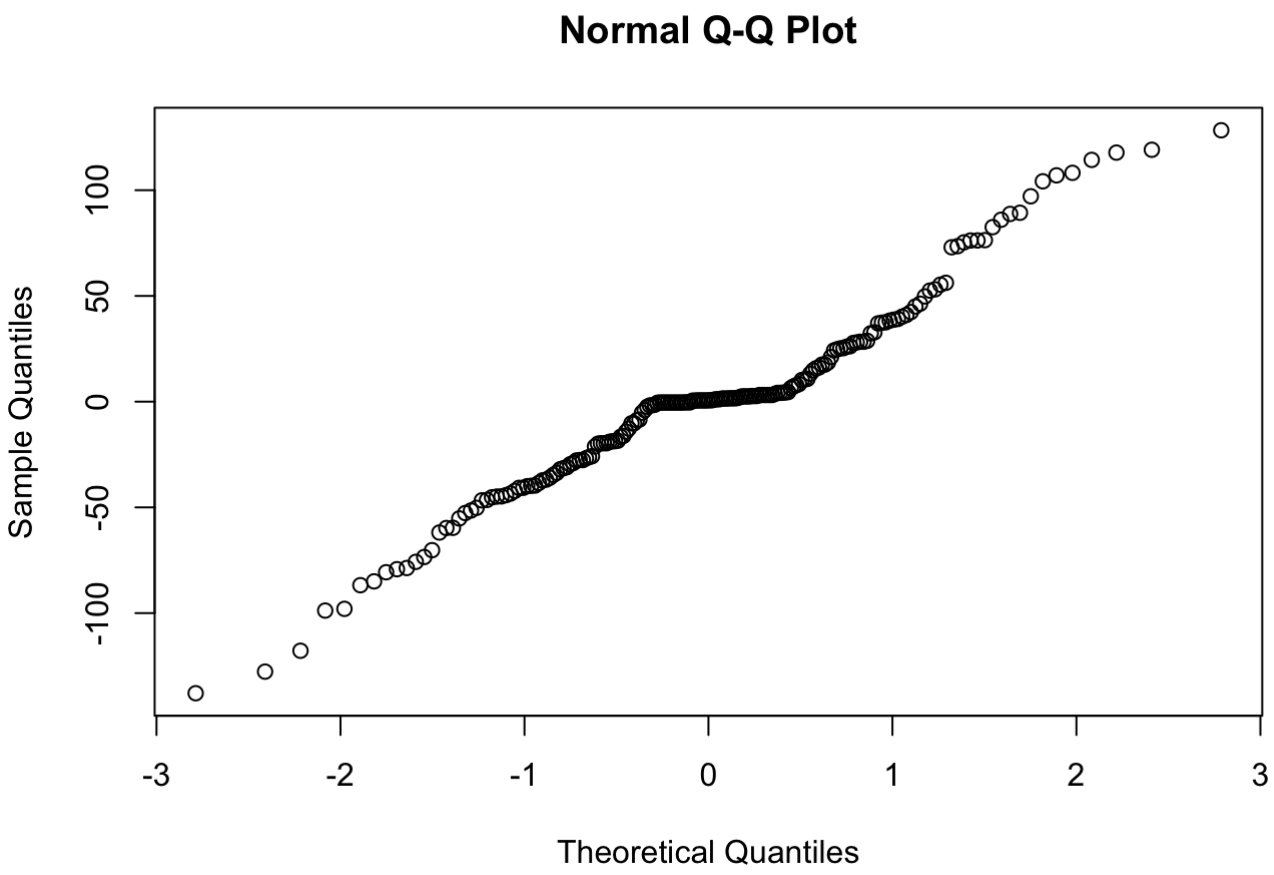

```
qqnorm(anova_20$residuals)
```



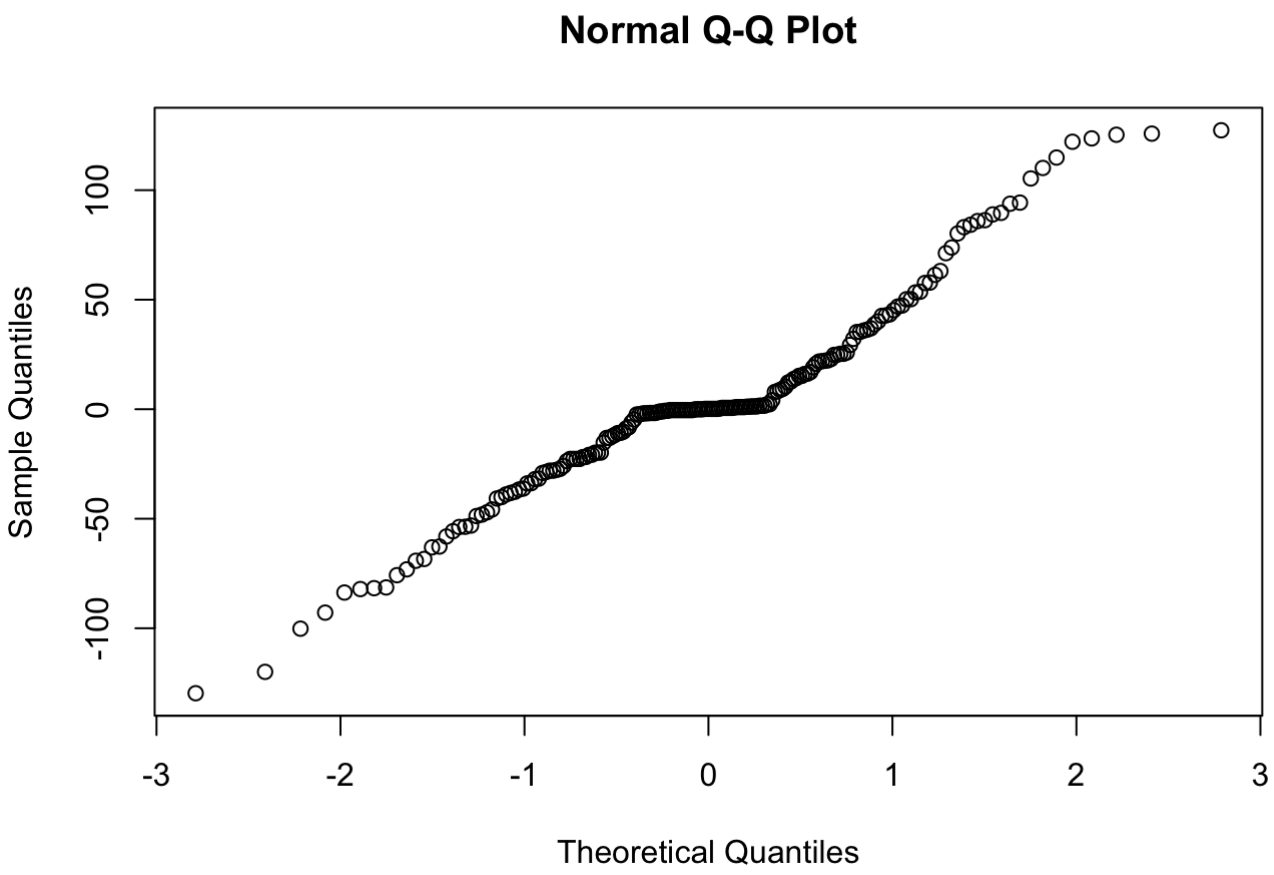

```
qqnorm(anova_21$residuals)
```



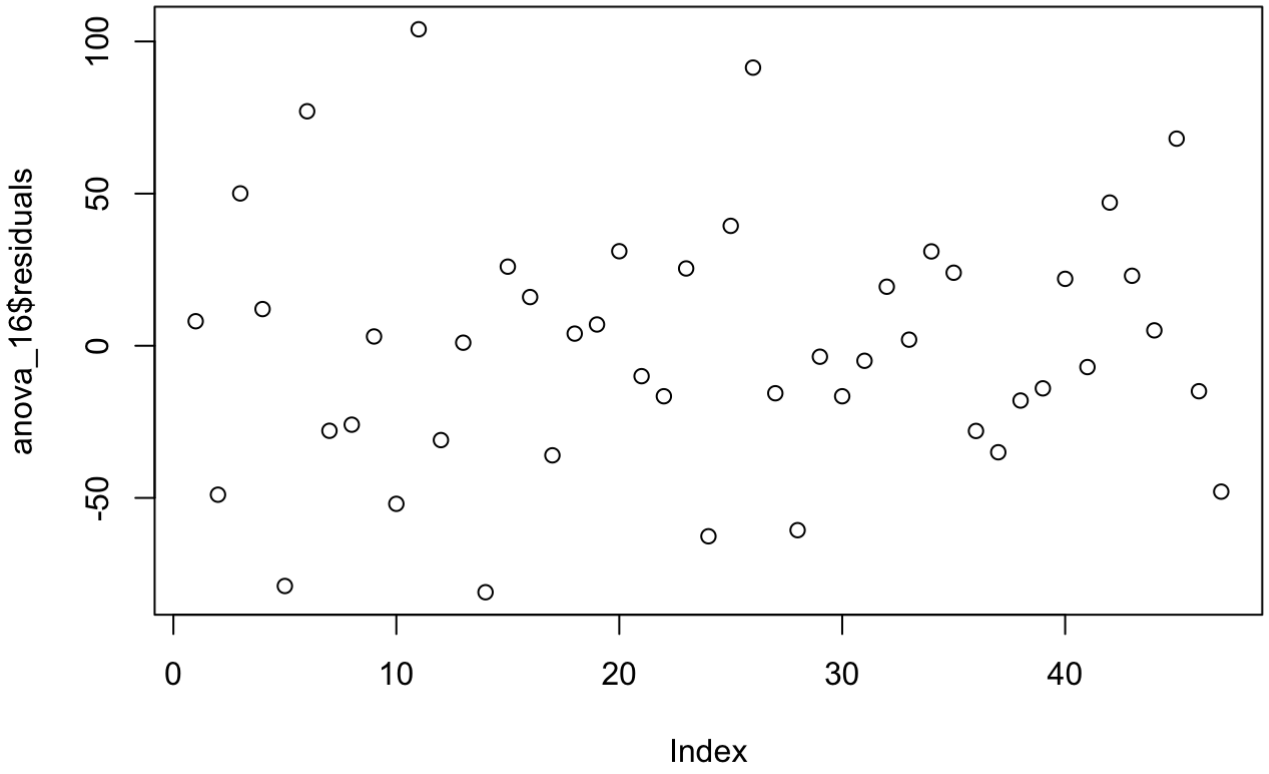

```
qqnorm(compound$residuals)
```



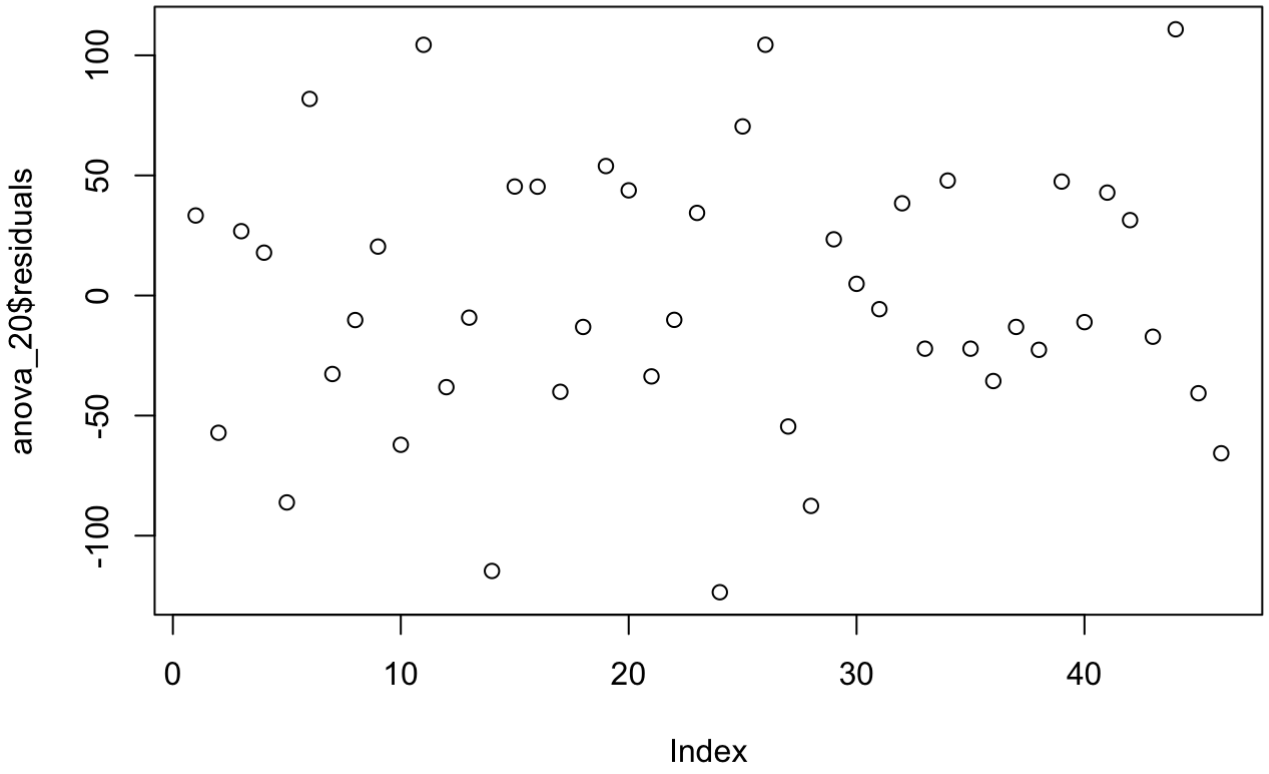

```
qqnorm(unstructed$residuals)
```



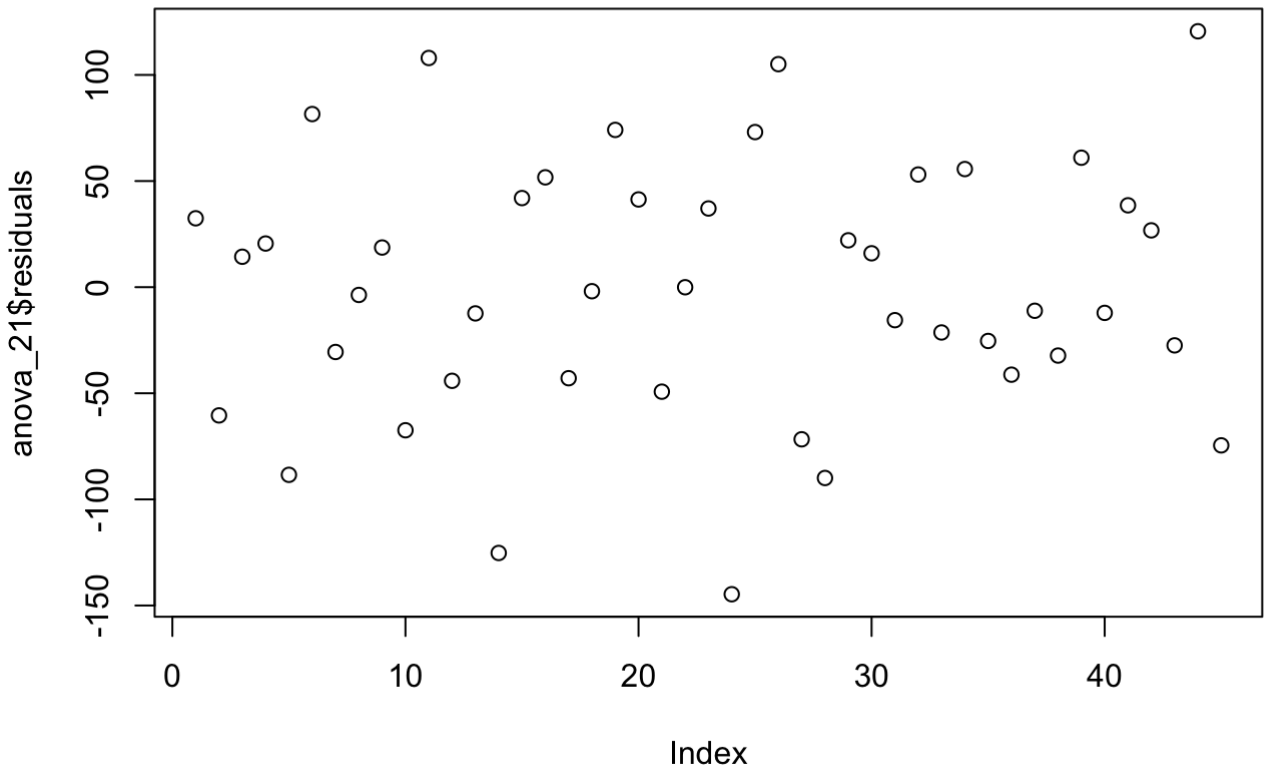

```
plot(anova_16$residuals)
```



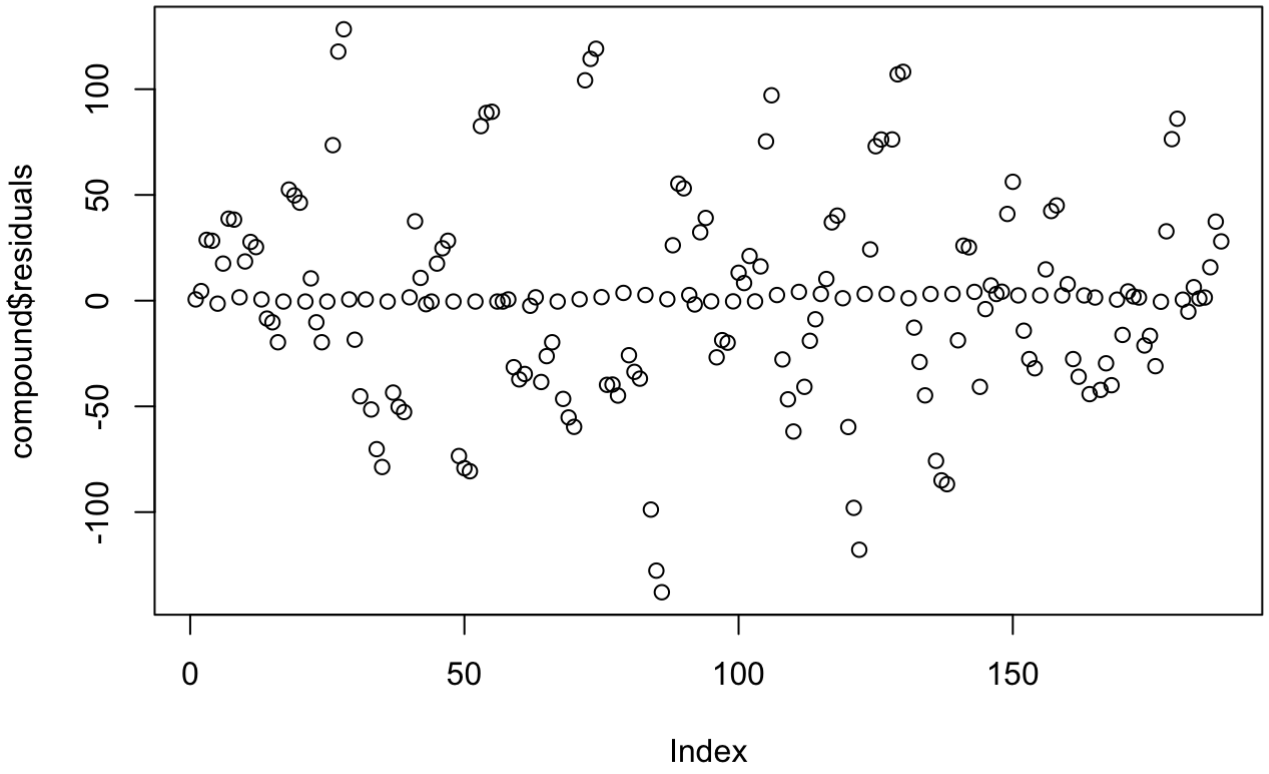

```
plot(anova_20$residuals)
```



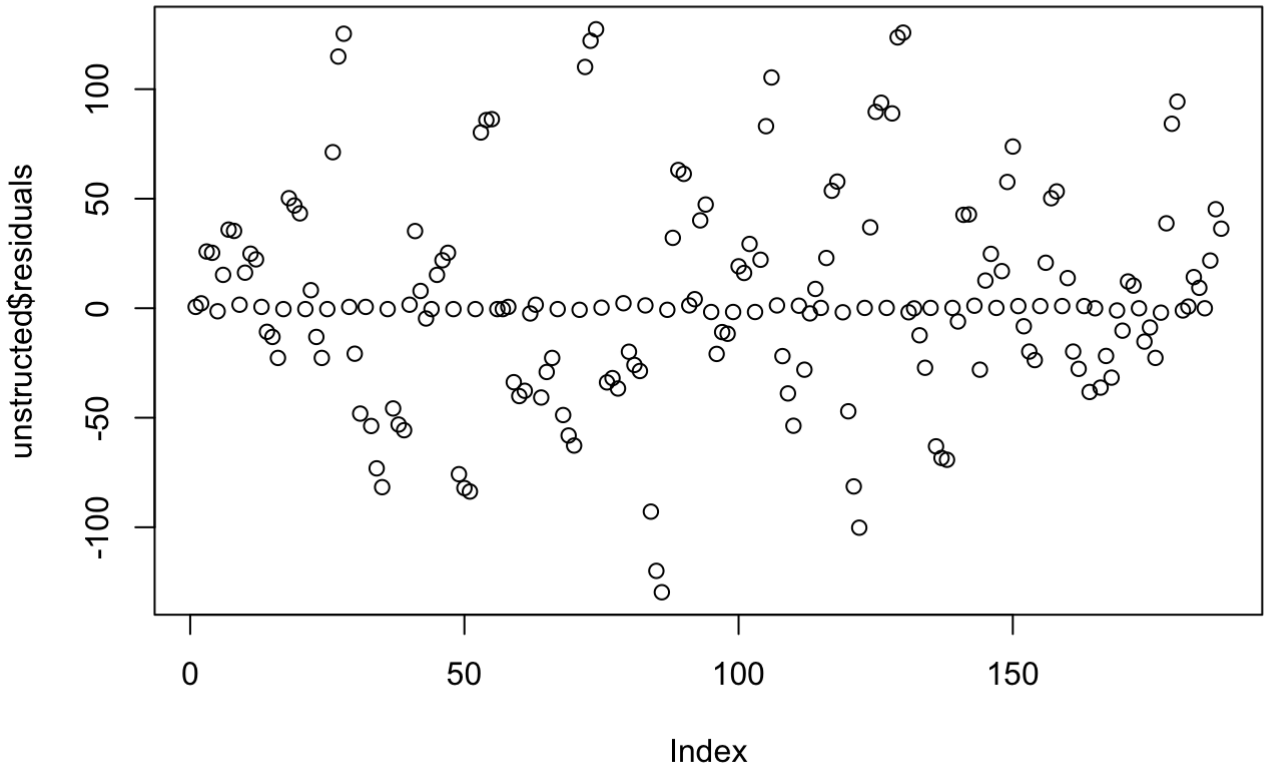

```
plot(anova_21$residuals)
```




```
plot(compound$residuals)
```




```
plot(unstructed$residuals)
```




```
shapiro.test(anova_16$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova_16$residuals
## W = 0.98261, p-value = 0.7019
```

```
shapiro.test(anova_20$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova_20$residuals
## W = 0.98247, p-value = 0.7082
```

```
shapiro.test(anova_21$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova_21$residuals
## W = 0.99117, p-value = 0.9792
```

```
shapiro.test(compound$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  compound$residuals
## W = 0.96342, p-value = 8.193e-05
```

```
shapiro.test(unstructed$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  unstructed$residuals
## W = 0.95848, p-value = 2.459e-05
```

```
bartlett.test(weight.x ~ Diet.x, data = day_16)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight.x by Diet.x
## Bartlett's K-squared = 4.4411, df = 3, p-value = 0.2176
```

```
bartlett.test(weight.x ~ Diet.x, data = day_20)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight.x by Diet.x
## Bartlett's K-squared = 3.2498, df = 3, p-value = 0.3547
```

```
bartlett.test(weight.x ~ Diet.x, data = day_21)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight.x by Diet.x
## Bartlett's K-squared = 3.0524, df = 3, p-value = 0.3836
```

```
# For Problem 1 From shapiro's test and qqplots, we could see that p-values are larger t
han 0.05, and we do not reject the null and conclud ethat the normality assumption is sa
tisfied. From the scatterplot of the residuals, it seems that the residuals do not have
 any patterns and thus they seem to have a constant variance.

# For Problem 2 From shapiro's test and qqplots, we could see that p-values are less tha
n 0.05, and we reject the null and conclude that the normality assumption is not satisfi
ed. From the scatterplot of the residuals, it seems that the residuals are from the rang
e of -100 to 100. The residuals seem to have a constant variance.

# Also from the bartlett test, we could see that p-values are larger than 0.05 so we do
 not reject the null and we can conclude equal variance assumption is not violated.


# Correlation
matrix <- with(combined, matrix(c(weight[Time == 16], weight[Time == 20], weight[Time ==
21]), ncol = 3))
var(matrix)
```

```
##           [,1]      [,2]      [,3]
## [1,] 2217.9406  317.1633 2539.2942
## [2,]  317.1633 4357.1401  735.6338
## [3,] 2539.2942  735.6338 5044.7135
```

```
cor(matrix)
```

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.1020251 0.7591364
## [2,] 0.1020251 1.0000000 0.1569071
## [3,] 0.7591364 0.1569071 1.0000000
```

```
# From the results above, we can see that the correlation between Day16 and Day20 is onl
y 0.1 and the correlation between Day20 and Day21 is only 0.16 while the correlation bet
ween Day16 and Day21 is 0.76. We could conclude that the compound symmatry structure is
 not appropriate. Therefore a unstructured correlation structure is better to fit the ou
r model. And we could see that from the previous problem, we had the same result that an
unstructured model will have a better performance.

# Parallelism
summary(aov(weight.x ~ weight.y * Diet.x, data = day_16))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## weight.y        1   9606    9606   5.047 0.0304 *
## Diet.x          3  14578    4859   2.553 0.0694 .
## weight.y:Diet.x 3   2787     929   0.488 0.6925
## Residuals      39  74228    1903
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(weight.x ~ weight.y * Diet.x, data = day_20))
```

```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## weight.y         1  20415   20415   6.493 0.01500 *
## Diet.x           3  42138   14046   4.467 0.00878 **
## weight.y:Diet.x  3  17043    5681   1.807 0.16235
## Residuals       38 119476    3144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(weight.x ~ weight.y * Diet.x, data = day_21))
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## weight.y         1  20538   20538   5.734 0.0218 *
## Diet.x           3  43763   14588   4.073 0.0135 *
## weight.y:Diet.x  3  28185    9395   2.623 0.0649 .
## Residuals       37 132517    3582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# From the results we could see that p-vlaues are larger than 0.05 for the interaction t
erms in all three summaries. So we do not reject the null and conclude that the parallel
ism assumption is not violated.
```