Assignment 2 – STAT 453/558
Due date: February 9

Question 1. Show that for the one-way ANOVA we have $E(MSE) = \sigma^2$.

Question 2. Show that for the CRBD we have $E(MSE) = \sigma^2$.

Question 3. The effect of three different lubricating oils on fuel economy in diesel truck engines is being studied. Fuel economy is measured using brake-specific fuel consumption after the engine has been running for 15 minutes. Five different truck engines are available for the study, and the experimenters conduct the following randomized complete block design.

	Engines				
Lubricant	1	2	3	4	5
1	0.500	0.634	0.487	0.329	0.512
2	0.535	0.675	0.520	0.435	0.540
3	0.513	0.595	0.488	0.400	0.510

(a) Analyze the data from this experiment.
(b) Use the Tukey method to make comparisons among the three lubricating oils to determine specifically which oils differ in break-specific fuel consumption.
(c) Analyze the residuals from this experiment and comment on model adequacy.

Question 4. An article in the *Journal of the Electrochemical Society* (Vol. 139, No. 2, 1992, pp. 524-532) describes an experiment to investigate low-pressure vapor deposition of polysilicon. The experiment was carried out in a large capacity reactor at Sematech in Austin, Texas. The reactor has several wafer positions, and four of these positions are selected ***at random***. The response variable is film thickness uniformity. Three replicates of the experiment were run, and the data are as follows:

Wafer Positions	Uniformity		
1	2.76	5.67	4.49
2	1.43	1.70	2.19
3	2.34	1.97	1.47
4	0.94	1.36	1.65

(a) Is there a difference in the wafer positions? Use α=0.05.
(b) Estimate the variability due to wafer position ($\hat{\sigma}_\tau^2$).
(c) Estimate the random error component ($\hat{\sigma}^2$).
(d) Analyze the residuals from this experiment and comment on model adequacy.

Assignment 2 – STAT 453/558
Due date: February 9

Question 1. Show that for the one-way ANOVA we have $E(MSE) = \sigma^2$.

Question 2. Show that for the CRBD we have $E(MSE) = \sigma^2$.

Q1

$$E(MSE) = E\left(\frac{\sum_{j=1}^{k}\sum_{i=1}^{n}(y_{ij}-\bar{y}_{i\cdot})^2}{n-k}\right)$$

$$= \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}E(y_{ij}-\bar{y}_{i\cdot})^2}{E(n-k)}$$

$$= \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\sigma^2}{n-k}$$

$$= \frac{nk\sigma^2}{n-k}$$

Since k is quite small

we can conclude this as

$$= \sigma^2$$

Q2:

$$E(SSE) = E\left(\sum_i\sum_j e_{ij}^2\right) = E\sum_i\sum_j(y_{ij}-\hat{y}_{ij})^2$$

$$= \sum_i\sum_j E(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2$$

$$= \sum_i\sum_j E(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})(y_{ij})$$

$$= \sum_i\sum_j E(y_{ij}^2 - \sum_i E(y_{i\cdot}^2) - E(a\bar{y}_{\cdot j}^2) + E(ab\bar{y}_{\cdot\cdot}^2)$$

$$= (a-1)(y_{ij}-\bar{y}_{\cdot j})^2 + (b-1)(\bar{y}_{ij}-\bar{y}_{\cdot j})^2$$

$$E(MSE) = \frac{(a-1)(b-1)\sigma^2}{(a-1)(b-1)}$$

$$= \sigma^2$$

Assignment2_stat453

Koki Itagaki

2024-02-08

#Question3 #The effect of three different lubricating oils on fuel economy in diesel truck engines is being #studied. Fuel economy is measured using brake-specific fuel consumption after the engine has been #running for 15 minutes. Five different truck engines are available for the study, and the experimenters #conduct the following randomized complete block design.

```
#3(a)Analyze the data from this experiment.

# Create a dataframe with the given data
dataFrame <- data.frame(
  Fuel_Economy = c(0.500, 0.535, 0.513, 0.634, 0.675,
                   0.595, 0.487, 0.520, 0.488, 0.329,
                   0.435, 0.400, 0.512, 0.540, 0.510),
  Lubricant = rep(1:3, times = 5),
    Engine = rep(1:5, each = 3)

)
dataFrame
```

```
##    Fuel_Economy Lubricant Engine
## 1         0.500         1      1
## 2         0.535         2      1
## 3         0.513         3      1
## 4         0.634         1      2
## 5         0.675         2      2
## 6         0.595         3      2
## 7         0.487         1      3
## 8         0.520         2      3
## 9         0.488         3      3
## 10        0.329         1      4
## 11        0.435         2      4
## 12        0.400         3      4
## 13        0.512         1      5
```

```
## 14         0.540          2      5
## 15         0.510          3      5
```

```
# Perform the ANOVA
aov3 <- aov(Fuel_Economy ~ factor(Lubricant) + factor(Engine), data = dataFrame)
summary(aov3)
```

```
##                   Df  Sum Sq  Mean Sq F value   Pr(>F)
## factor(Lubricant)  2 0.00671 0.003353   6.353   0.0223 *
## factor(Engine)     4 0.09210 0.023025  43.626 1.78e-05 ***
## Residuals          8 0.00422 0.000528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# From the anova test above,since the p-value for Lubricant is 0.0233 which is smaller than a = 0.05,
#It is significant to explain the variability in a response variable. Also, since the p-value for truck is 1.78e-05 which is smaller than a = 0.05,
#It is significant to explain the variability in a response variable.



#3(b)Use the Tukey method to make comparisons among the three lubricating oils to determine specifically
#which oils differ in break-specific fuel consumption.

# Perform Tukey's HSD test
tukey <- TukeyHSD(aov3, "factor(Lubricant)", conf.level = 0.95)
tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Fuel_Economy ~ factor(Lubricant) + factor(Engine), data = dataFrame)
##
## $`factor(Lubricant)`
##        diff          lwr         upr     p adj
## 2-1  0.0486  0.007082078 0.090117922 0.0245809
## 3-1  0.0088 -0.032717922 0.050317922 0.8210970
```

```
## 14         0.540          2      5
## 15         0.510          3      5
```

```
# Perform the ANOVA
aov3 <- aov(Fuel_Economy ~ factor(Lubricant) + factor(Engine), data = dataFrame)
summary(aov3)
```

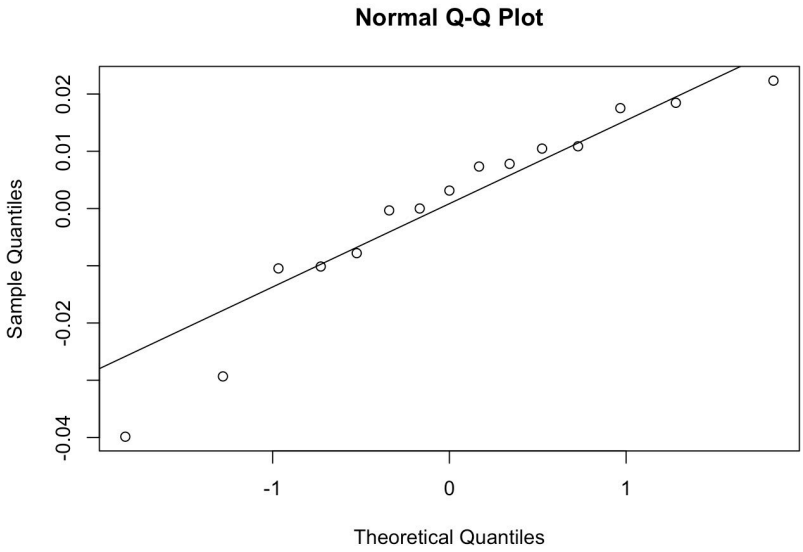
```
##                   Df  Sum Sq  Mean Sq F value   Pr(>F)
## factor(Lubricant)  2 0.00671 0.003353   6.353   0.0223 *
## factor(Engine)     4 0.09210 0.023025  43.626 1.78e-05 ***
## Residuals          8 0.00422 0.000528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# From the anova test above,since the p-value for Lubricant is 0.0233 which is smaller than a = 0.05,
#It is significant to explain the variability in a response variable. Also, since the p-value for truck is 1.78e-05 which is smaller than a = 0.05,
#It is significant to explain the variability in a response variable.



#3(b)Use the Tukey method to make comparisons among the three lubricating oils to determine specifically
#which oils differ in break-specific fuel consumption.

# Perform Tukey's HSD test
tukey <- TukeyHSD(aov3, "factor(Lubricant)", conf.level = 0.95)
tukey
```

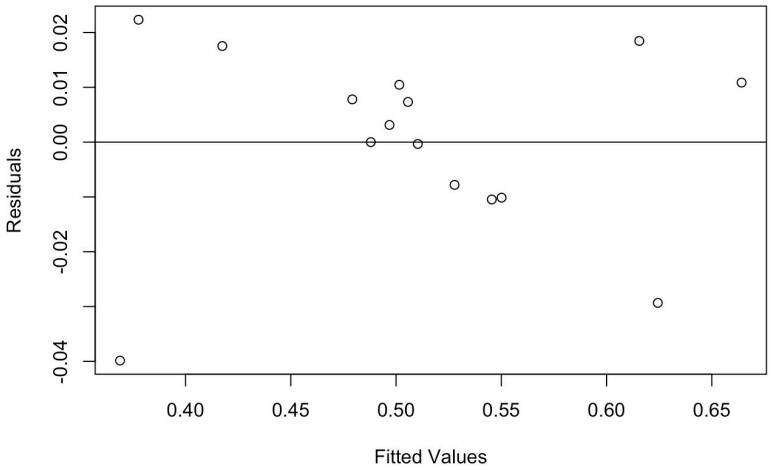
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Fuel_Economy ~ factor(Lubricant) + factor(Engine), data = dataFrame)
##
## $`factor(Lubricant)`
##        diff          lwr         upr     p adj
## 2-1  0.0486  0.007082078 0.090117922 0.0245809
## 3-1  0.0088 -0.032717922 0.050317922 0.8210970
```

```
## 3-2 -0.0398 -0.081317922 0.001717922 0.0594979
```

```
#From the output by tukey's test, It is clear that
#The difference between oil 1 and oil2 is significant since the p-value = 0.02458 < a = 0.05


#(c)
residuals3<-aov3$residuals
qqnorm(residuals3)
qqline(residuals3)
```



```
Fitted_values=aov3$fitted.values
plot(Fitted_values,residuals3,ylab="Residuals",xlab="Fitted Values")
abline(h=0)
```




```
#From the q-q plot for residuals, the almost of all data are near the streight line, it means the distribution is
normally distributed.


#Moreover,The   residual    vs  fitted  value   plot  shows
# that the pattern of scatter is the almost same.
#so the variance is adequate
#So this is an adequate model to test.
```

#Question 4 # An article in the Journal of the Electrochemical Society (Vol. 139, No. 2, 1992, pp. 524-532) #describes an experiment to investigate low-pressure vapor deposition of polysilicon. The experiment was #carried out in a large capacity reactor at Sematech in Austin, Texas. The reactor has several wafer #positions, and four of these positions are selected at random. The response variable is film thickness #uniformity. Three replicates of the experiment were run, and the data are as follows:

```
#(a) Is there a difference in the wafer positions? Use α=0.05.

##Ho: σ^2=0
#Ha: σ^2>0

# Create a data frame with the provided data
data4 <- data.frame(
  Wafer_Position = rep(1:4, each = 3),
  Uniformity = c(2.76, 5.67, 4.49, 1.43, 1.70, 2.19, 2.34, 1.97, 1.47, 0.94, 1.36, 1.65)
)
data4
```

```
##    Wafer_Position Uniformity
## 1               1       2.76
## 2               1       5.67
## 3               1       4.49
## 4               2       1.43
## 5               2       1.70
## 6               2       2.19
## 7               3       2.34
## 8               3       1.97
## 9               3       1.47
## 10              4       0.94
## 11              4       1.36
## 12              4       1.65
```

```
# Perform one-way ANOVA
result <- aov(Uniformity ~ factor(Wafer_Position), data = data4)

# Summary of the ANOVA
summary(result)
```

```
##                          Df Sum Sq Mean Sq F value  Pr(>F)
## factor(Wafer_Position)  3 16.220   5.407    8.29 0.00775 **
## Residuals               8  5.217   0.652
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#From the output of ANOVA table, the p-value is 0.00775 << 0.05, then we reject Ho:
#It is significant that there is differencew  between wafer positions.



#(b) Estimate the variability due to wafer position( )*
sigmat = (5.407-0.652)/3
sigmat
```

```
## [1] 1.585
```

```
#(c) Estimate the random error component ( )().
#From the ANOVA table i can get MSB
sol = (5.407-0.652)/3
sol
```

```
## [1] 1.585
```

```
#From the r output, the random error component is 0.652
```

#Question5 The effect of five different ingredients (A, B, C, D, E) on reaction time of a chemical process is #being studied. Each batch of new material is only large enough to permit five runs to be made. #Furthermore, each run requires approximately 1 1/2 hours, so only five runs can be made in one day. The #experimenter decides to run the experiment as a Latin square so that day and batch effects can be #systematically controlled. She obtains the data that follow. Analyze the data from this experiment (use a= 0.05) and draw conclusions.

```
# Create the data frame
data5 <- data.frame(
  Batch = rep(1:5, times = 5),
  Day = rep(1:5, each = 5),
```

```
  integer = c(8,7,1,7,3,11,2,7,3,8,4,9,10,1,5,
  6,8,6,6,10,4,2,3,8,8),

ingredients = c('A','B','D','C','E',                    'C','E','A','D','B',
                'B','A','C','E','D',
                'D','C','E','B','A',
                'E','D','B','A','C')

)

data5
```

```
##    Batch Day integer ingredients
## 1      1   1       8           A
## 2      2   1       7           B
## 3      3   1       1           D
## 4      4   1       7           C
## 5      5   1       3           E
## 6      1   2      11           C
## 7      2   2       2           E
## 8      3   2       7           A
## 9      4   2       3           D
## 10     5   2       8           B
## 11     1   3       4           B
## 12     2   3       9           A
## 13     3   3      10           C
## 14     4   3       1           E
## 15     5   3       5           D
## 16     1   4       6           D
## 17     2   4       8           C
## 18     3   4       6           E
## 19     4   4       6           B
## 20     5   4      10           A
## 21     1   5       4           E
## 22     2   5       2           D
## 23     3   5       3           B
## 24     4   5       8           A
## 25     5   5       8           C
```

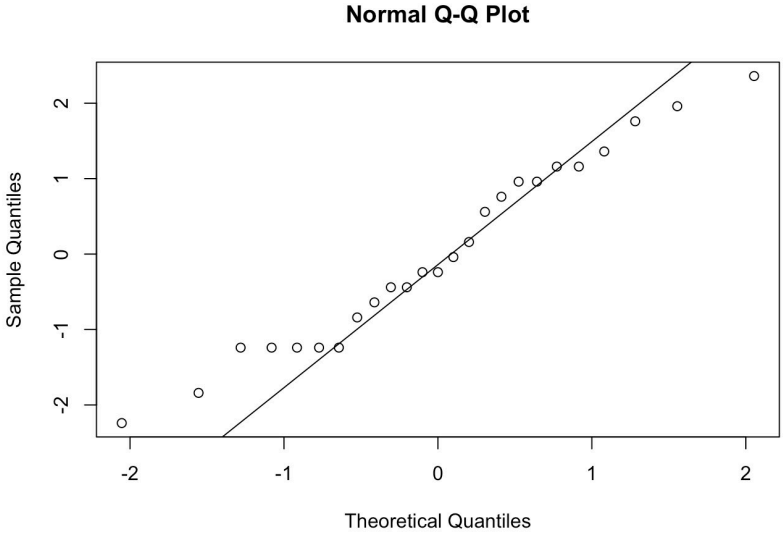
```r
#Ho: u1 = U2 = u3 = u4 = u4
#Ha: At least two u are not wqual to the others
aov5 <-aov(integer~factor(Batch)+factor(ingredients)+factor(Day),
           data=data5)
summary(aov5)
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## factor(Batch)        4  12.24    3.06   0.979 0.455014
## factor(ingredients)  4 141.44   35.36  11.309 0.000488 ***
## factor(Day)          4  15.44    3.86   1.235 0.347618
## Residuals           12  37.52    3.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

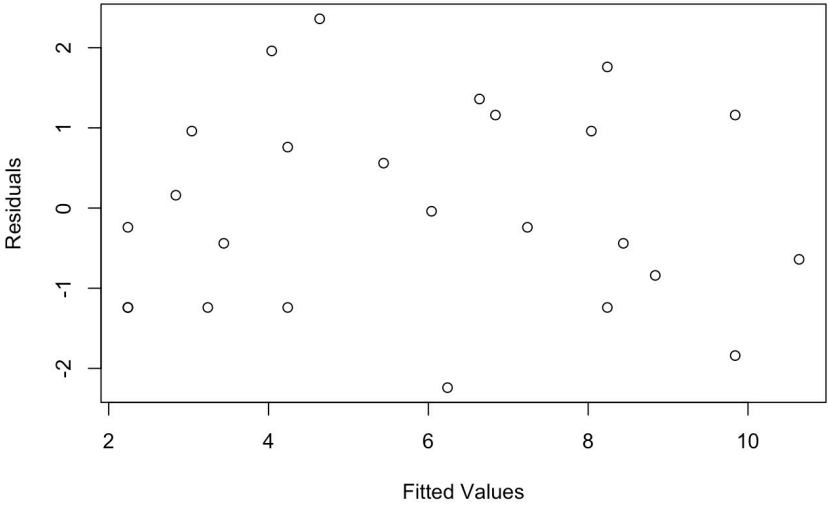
```r
TUKEY5 <- TukeyHSD(x=aov5, "factor(ingredients)", conf.level=0.95)
TUKEY5
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = integer ~ factor(Batch) + factor(ingredients) + factor(Day), data = data5)
##
## $`factor(ingredients)`
##     diff        lwr        upr     p adj
## B-A -2.8 -6.3646078  0.7646078 0.1539433
## C-A  0.4 -3.1646078  3.9646078 0.9960012
## D-A -5.0 -8.5646078 -1.4353922 0.0055862
## E-A -5.2 -8.7646078 -1.6353922 0.0041431
## C-B  3.2 -0.3646078  6.7646078 0.0864353
## D-B -2.2 -5.7646078  1.3646078 0.3365811
## E-B -2.4 -5.9646078  1.1646078 0.2631551
## D-C -5.4 -8.9646078 -1.8353922 0.0030822
## E-C -5.6 -9.1646078 -2.0353922 0.0023007
## E-D -0.2 -3.7646078  3.3646078 0.9997349
```

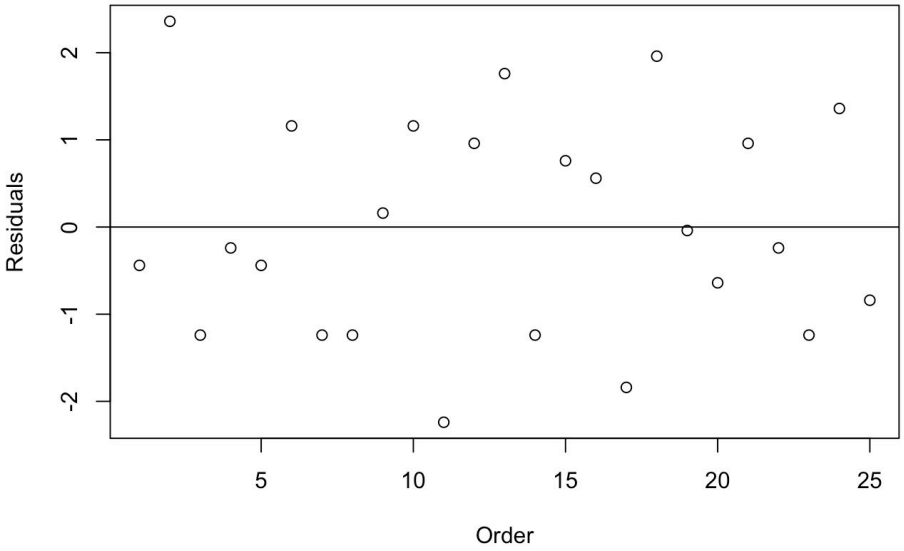
```
residuals5<-aov5$residuals
qqnorm(residuals5)
qqline(residuals5)
```




```
Fitted_values5=aov5$fitted.values
plot(Fitted_values5,residuals5,ylab="Residuals",xlab="Fitted Values")
```




```
plot(1:1:25,residuals5,ylab="Residuals",xlab="Order")
abline(h=0)
```




```
#From the output of ANOVA table, the p-value for Batch and Day are 0.455014 and
#0.034618 > a =  0.05, then we  fail toreject Ho.
#So these tw factors are not significant for the equation.
#It is significant that there is differencew  between wafer positions.
#On the other hand, the p-value of ingredients is 0.000488 << 0.05 = a.
#So we reject Ho. It is significant that the factor(ingredients) should be
#significant for the equation.

#From the turkey's test, it is clear that the differences of D and A,
#E and A, D and C, and E and C are significant.

#For checking normality, I used a q-q plot. According to the q-q plot,
#almost of all data are near the streight line, so it is normally distributed.
#Also residual  vs  fitted  value   plot shows that the scatter of the data are
#in the same pattern or similar pattern. So, the variance is also adequate.
```

```
#Moreover, the residual vs  order   plot shows that the smilar pattern of scatter
#based on the horizontal line at  h = 0. So, independent correlation is also adequate.
```

#Question6 #The yield of a chemical process was measured using five batches of raw material, five acid #concentrations, five standing times, (A, B, C, D, E) and five catalyst concentrations (a, b, g, d, e). The #Graeco-Latin square that follows was used. Analyze the data from this experiment (use a = 0.05) and draw conclusions.

```
integer6 <-c(26,16,19,16,13,18,21,18,11,21,20,12,16,25,13,15,15,22,14,17,10,24,
             17,17,14)
times <- c('A', 'B','C', 'D','E','B','C','D','E','A','C','D','E','A','B','D',
'E','A','B','C','E','A','B','C','D')

Acid <- c(1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,1,2,3,4,5)
Batch6 = c(rep(1,5),rep(2,5),rep(3,5),rep(4,5),rep(5,5))
catalyst_concentrations =c("alpha","beta","gamma","delta","epsilon",                                    "gamma","delt
a","epsilon","alpha","beta",
                           "epsilon","alpha","beta","gamma","delta",                                    "beta","gamm
a","delta","epsilon","alpha",                                  "delta","epsilon","alpha","beta","gamma")
data6 = data.frame(integer6,times,catalyst_concentrations,Acid, Batch6)
data6
```

```
##    integer6 times catalyst_concentrations Acid Batch6
## 1        26     A                   alpha    1      1
## 2        16     B                    beta    2      1
## 3        19     C                   gamma    3      1
## 4        16     D                   delta    4      1
## 5        13     E                 epsilon    5      1
## 6        18     B                   gamma    1      2
## 7        21     C                   delta    2      2
## 8        18     D                 epsilon    3      2
## 9        11     E                   alpha    4      2
## 10       21     A                    beta    5      2
## 11       20     C                 epsilon    1      3
## 12       12     D                   alpha    2      3
## 13       16     E                    beta    3      3
## 14       25     A                   gamma    4      3
## 15       13     B                   delta    5      3
## 16       15     D                    beta    1      4
```

```
## 17       15     E                gamma    2      4
## 18       22     A                delta    3      4
## 19       14     B              epsilon    4      4
## 20       17     C                alpha    5      4
## 21       10     E                delta    1      5
## 22       24     A              epsilon    2      5
## 23       17     B                alpha    3      5
## 24       17     C                 beta    4      5
## 25       14     D                gamma    5      5
```

```
aov6 <-aov(integer6~factor(times)+factor(Acid)+factor(Batch6)+factor(catalyst_concentrations), data=data6)
summary(aov6)
```

```
##                               Df Sum Sq Mean Sq F value   Pr(>F)
## factor(times)                  4  342.8   85.70  14.650 0.000941 ***
## factor(Acid)                   4   24.4    6.10   1.043 0.442543
## factor(Batch6)                 4   10.0    2.50   0.427 0.785447
## factor(catalyst_concentrations)  4   12.0    3.00   0.513 0.728900
## Residuals                      8   46.8    5.85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

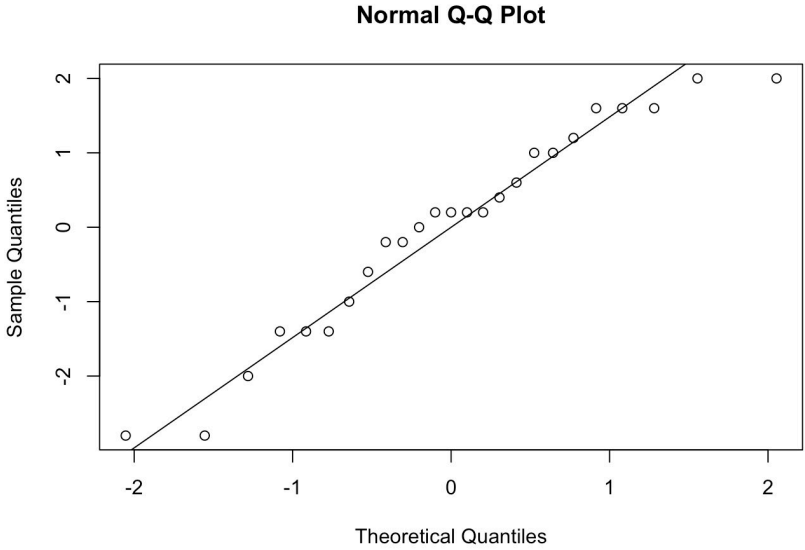
```
turkey6 <- TukeyHSD(x=aov6,conf.level=0.95)
turkey6
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = integer6 ~ factor(times) + factor(Acid) + factor(Batch6) + factor(catalyst_concentrations),
data = data6)
##
## $`factor(times)`
##      diff        lwr        upr     p adj
## B-A  -8.0 -13.284751 -2.7152488 0.0051639
## C-A  -4.8 -10.084751  0.4847512 0.0770797
## D-A  -8.6 -13.884751 -3.3152488 0.0032815
## E-A -10.6 -15.884751 -5.3152488 0.0008219
```

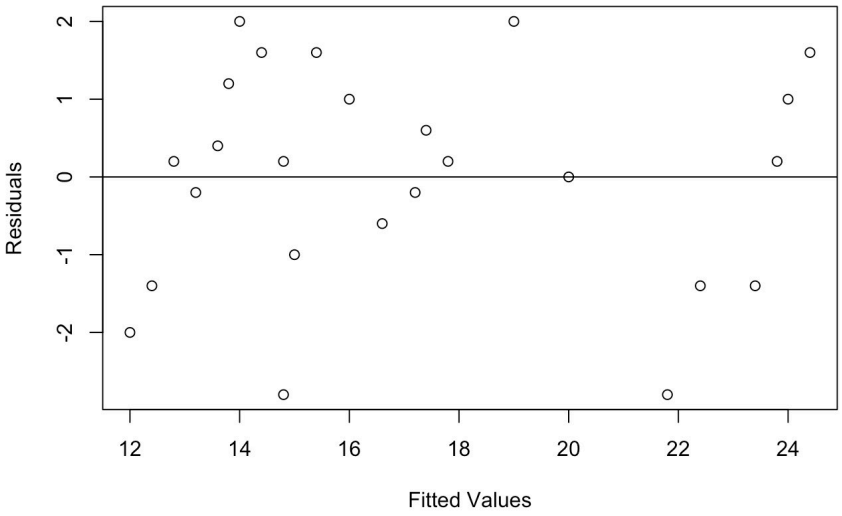
```
## C-B   3.2  -2.084751  8.4847512 0.3087034
## D-B  -0.6  -5.884751  4.6847512 0.9939694
## E-B  -2.6  -7.884751  2.6847512 0.4837165
## D-C  -3.8  -9.084751  1.4847512 0.1869031
## E-C  -5.8 -11.084751 -0.5152488 0.0317351
## E-D  -2.0  -7.284751  3.2847512 0.6948188
##
## $`factor(Acid)`
##     diff       lwr      upr     p adj
## 2-1 -0.2 -5.484751 5.084751 0.9999182
## 3-1  0.6 -4.684751 5.884751 0.9939694
## 4-1 -1.2 -6.484751 4.084751 0.9281909
## 5-1 -2.2 -7.484751 3.084751 0.6232282
## 3-2  0.8 -4.484751 6.084751 0.9823986
## 4-2 -1.0 -6.284751 4.284751 0.9610846
## 5-2 -2.0 -7.284751 3.284751 0.6948188
## 4-3 -1.8 -7.084751 3.484751 0.7640759
## 5-3 -2.8 -8.084751 2.484751 0.4197369
## 5-4 -1.0 -6.284751 4.284751 0.9610846
##
## $`factor(Batch6)`
##     diff       lwr      upr     p adj
## 2-1 -0.2 -5.484751 5.084751 0.9999182
## 3-1 -0.8 -6.084751 4.484751 0.9823986
## 4-1 -1.4 -6.684751 3.884751 0.8834072
## 5-1 -1.6 -6.884751 3.684751 0.8279246
## 3-2 -0.6 -5.884751 4.684751 0.9939694
## 4-2 -1.2 -6.484751 4.084751 0.9281909
## 5-2 -1.4 -6.684751 3.884751 0.8834072
## 4-3 -0.6 -5.884751 4.684751 0.9939694
## 5-3 -0.8 -6.084751 4.484751 0.9823986
## 5-4 -0.2 -5.484751 5.084751 0.9999182
##
## $`factor(catalyst_concentrations)`
##                diff       lwr      upr     p adj
## beta-alpha      0.4 -4.884751 5.684751 0.9987373
## delta-alpha    -0.2 -5.484751 5.084751 0.9999182
## epsilon-alpha   1.2 -4.084751 6.484751 0.9281909
## gamma-alpha     1.6 -3.684751 6.884751 0.8279246
## delta-beta     -0.6 -5.884751 4.684751 0.9939694
```

```
## epsilon-beta   0.8 -4.484751 6.084751 0.9823986
## gamma-beta     1.2 -4.084751 6.484751 0.9281909
## epsilon-delta  1.4 -3.884751 6.684751 0.8834072
## gamma-delta    1.8 -3.484751 7.084751 0.7640759
## gamma-epsilon  0.4 -4.884751 5.684751 0.9987373
```

```
residuals6<-aov6$residuals
qqnorm(residuals6)
qqline(residuals6)
```



```r
Fitted_values6=aov6$fitted.values
plot(Fitted_values6,residuals6,ylab="Residuals",xlab="Fitted Values")
abline(h = 0)
```




```r
##From the output of ANOVA table, the p-value for Acid with 0.442543,Batch6 with 0.785447
#and catalyst_concentrations with 0.728900 > a =  0.05, then we  fail to reject Ho.
#So these three factors are not significant for the equation.
#On the other hand, the p-value of times is 0.000941 << 0.05 = a.
#So we reject Ho. It is significant that the factor(times) should be
#significant for the equation.
```

```
#From the turkey's test, it is clear that the differences of B-A,
#D-A, E-A, and E-C are significant.

#For checking normality, I used a q-q plot. According to the q-q plot,
#almost of all data are near the streight line, so it is normally distributed.
#Also residual  vs  fitted  value   plot shows that the scatter of the data are
#in the same pattern or similar pattern. So, the variance is also adequate.
#Moreover, the residual vs  order   plot shows that the smilar pattern of scatter
#based on the horizontal line at  h = 0. So, independent correlation is also adequate.
```

Question 5. The effect of five different ingredients (*A, B, C, D, E*) on reaction time of a chemical process is being studied. Each batch of new material is only large enough to permit five runs to be made. Furthermore, each run requires approximately 1 1/2 hours, so only five runs can be made in one day. The experimenter decides to run the experiment as a Latin square so that day and batch effects can be systematically controlled. She obtains the data that follow. Analyze the data from this experiment (use α = 0.05) and draw conclusions.

	Day				
Batch	1	2	3	4	5
1	*A*=8	*B*=7	*D*=1	*C*=7	*E*=3
2	*C*=11	*E*=2	*A*=7	*D*=3	*B*=8
3	*B*=4	*A*=9	*C*=10	*E*=1	*D*=5
4	*D*=6	*C*=8	*E*=6	*B*=6	*A*=10
5	*E*=4	*D*=2	*B*=3	*A*=8	*C*=8

Question 6. The yield of a chemical process was measured using five batches of raw material, five acid concentrations, five standing times, (*A, B, C, D, E*) and five catalyst concentrations (α, β, γ, δ, ε). The Graeco-Latin square that follows was used. Analyze the data from this experiment (use α = 0.05) and draw conclusions.

			Acid Concentration		
Batch	1	2	3	4	5
1	$A\alpha$=26	$B\beta$=16	$C\gamma$=19	$D\delta$=16	$E\varepsilon$=13
2	$B\gamma$=18	$C\delta$=21	$D\varepsilon$=18	$E\alpha$=11	$A\beta$=21
3	$C\varepsilon$=20	$D\alpha$=12	$E\beta$=16	$A\gamma$=25	$B\delta$=13
4	$D\beta$=15	$E\gamma$=15	$A\delta$=22	$B\varepsilon$=14	$C\alpha$=17
5	$E\delta$=10	$A\varepsilon$=24	$B\alpha$=17	$C\beta$=17	$D\gamma$=14

Assignment 2 – STAT 453/558
Due date: February 9

Question 1. Show that for the one-way ANOVA we have $E(MSE) = \sigma^2$.

Question 2. Show that for the CRBD we have $E(MSE) = \sigma^2$.

Question 3. The effect of three different lubricating oils on fuel economy in diesel truck engines is being studied. Fuel economy is measured using brake-specific fuel consumption after the engine has been running for 15 minutes. Five different truck engines are available for the study, and the experimenters conduct the following randomized complete block design.

	Engines				
Lubricant	1	2	3	4	5
1	0.500	0.634	0.487	0.329	0.512
2	0.535	0.675	0.520	0.435	0.540
3	0.513	0.595	0.488	0.400	0.510

(a) Analyze the data from this experiment.
(b) Use the Tukey method to make comparisons among the three lubricating oils to determine specifically which oils differ in break-specific fuel consumption.
(c) Analyze the residuals from this experiment and comment on model adequacy.

Question 4. An article in the *Journal of the Electrochemical Society* (Vol. 139, No. 2, 1992, pp. 524-532) describes an experiment to investigate low-pressure vapor deposition of polysilicon. The experiment was carried out in a large capacity reactor at Sematech in Austin, Texas. The reactor has several wafer positions, and four of these positions are selected ***at random***. The response variable is film thickness uniformity. Three replicates of the experiment were run, and the data are as follows:

Wafer Positions	Uniformity		
1	2.76	5.67	4.49
2	1.43	1.70	2.19
3	2.34	1.97	1.47
4	0.94	1.36	1.65

(a) Is there a difference in the wafer positions? Use α=0.05.
(b) Estimate the variability due to wafer position ($\hat{\sigma}_\tau^2$).
(c) Estimate the random error component ($\hat{\sigma}^2$).
(d) Analyze the residuals from this experiment and comment on model adequacy.

Question 5. The effect of five different ingredients (*A, B, C, D, E*) on reaction time of a chemical process is being studied. Each batch of new material is only large enough to permit five runs to be made. Furthermore, each run requires approximately 1 1/2 hours, so only five runs can be made in one day. The experimenter decides to run the experiment as a Latin square so that day and batch effects can be systematically controlled. She obtains the data that follow. Analyze the data from this experiment (use α = 0.05) and draw conclusions.

	Day				
Batch	1	2	3	4	5
1	*A*=8	*B*=7	*D*=1	*C*=7	*E*=3
2	*C*=11	*E*=2	*A*=7	*D*=3	*B*=8
3	*B*=4	*A*=9	*C*=10	*E*=1	*D*=5
4	*D*=6	*C*=8	*E*=6	*B*=6	*A*=10
5	*E*=4	*D*=2	*B*=3	*A*=8	*C*=8

Question 6. The yield of a chemical process was measured using five batches of raw material, five acid concentrations, five standing times, (*A, B, C, D, E*) and five catalyst concentrations ($\alpha, \beta, \gamma, \delta, \varepsilon$). The Graeco-Latin square that follows was used. Analyze the data from this experiment (use α = 0.05) and draw conclusions.

			Acid Concentration		
Batch	1	2	3	4	5
1	$A\alpha$=26	$B\beta$=16	$C\gamma$=19	$D\delta$=16	$E\varepsilon$=13
2	$B\gamma$=18	$C\delta$=21	$D\varepsilon$=18	$E\alpha$=11	$A\beta$=21
3	$C\varepsilon$=20	$D\alpha$=12	$E\beta$=16	$A\gamma$=25	$B\delta$=13
4	$D\beta$=15	$E\gamma$=15	$A\delta$=22	$B\varepsilon$=14	$C\alpha$=17
5	$E\delta$=10	$A\varepsilon$=24	$B\alpha$=17	$C\beta$=17	$D\gamma$=14

Solutions

Question 1

$$E(MSE) = \frac{1}{a(n-1)} E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i\cdot})^2\right)$$

$$= \frac{1}{a(n-1)} E\left(\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2\right) - E\left(\sum_{i=1}^{a}\sum_{j=1}^{n} 2\cdot y_{ij}\cdot\bar{y}_{i\cdot}\right)$$

$$+ E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}\bar{y}_{i\cdot}^2\right)$$

$$= \frac{1}{a(n-1)}\sum_{i=1}^{a}\sum_{j=1}^{n} E(y_{ij}^2) - 2n\sum_{i=1}^{a} E(\bar{y}_{i\cdot})^2 + n\cdot\sum_{i=1}^{a} E(\bar{y}_{i\cdot})^2$$

Note $E(y_{ij})^2 = Var(y_{ij}) + [E(y_{ij})]^2$

$= \sigma^2 + (\mu+\tau_i)^2$

Also $E(\bar{y}_{i\cdot})^2 = Var(\bar{y}_{i\cdot}) + [E(\bar{y}_{i\cdot})]^2$

$= \frac{\sigma^2}{n} + (\mu+\tau_i)^2$

$$= \frac{1}{a(n-1)}\left[an\sigma^2 + n\sum_{i=1}^{a}(\mu+\tau_i)^2 - n\cdot\sum_{i=1}^{a}\frac{\sigma^2}{n} + (\mu+\tau_i)^2\right]$$

$$= \frac{1}{a(n-1)}\left[a\cdot\sigma^2(n-1) + n\sum_{i=1}^{a}(\mu+\tau_i)^2 - n\sum_{i=1}^{a}(\mu+\tau_i)^2\right]$$

$$= \sigma^2$$

Question 2

$$y_{ij} = \mu + \tau_i + B_j + \varepsilon_{ij} \qquad \text{CRBD}$$

Note: $E(y_{ij}) = \mu + \tau_i + B_j$ & $V(y_{ij}) = \sigma^2$

$E(y_{i\cdot}) = E\left(\sum_{j=1}^{b} y_{ij}\right) = \sum_{j=1}^{b} \mu + \tau_i + B_j = b(\mu + \tau_i) \rightarrow E(\bar{y}_{i\cdot}) = \mu + \tau_i$

$V(\bar{y}_{i\cdot}) = \frac{1}{b^2} \sum_{j=1}^{b} V(y_{ij}) = \sigma^2/b$

$E(\bar{y}_{\cdot\cdot}) = \frac{\sum_i \sum_j E(y_{ij})}{a \cdot b} = \mu \qquad V(\bar{y}_{\cdot\cdot}) = \frac{\sigma^2}{ab}$

From Slide 18 - Chapter 4

$$SSE = \sum_i \sum_j y_{ij}^2 - \sum_i b\bar{y}_{i\cdot}^2 - \sum_j a\bar{y}_{\cdot j}^2 + ab\bar{y}_{\cdot\cdot}^2$$

(I) (II) (III) (IV)

We will use $E(X^2) = V(X) + [E(X)]^2$

(I) $E\left(\sum_i \sum_j y_{ij}^2\right) = \sum_i \sum_j \left(\sigma^2 + (\mu + \tau_i + B_j)^2\right)$

(II) $E\left(\sum_i b\bar{y}_{i\cdot}^2\right) = b \cdot \sum_i \left(\frac{\sigma^2}{b} + (\mu + \tau_i)^2\right)$

(III) $E\left(\sum_j a\bar{y}_{\cdot j}^2\right) = a \cdot \sum_j \left[\frac{\sigma^2}{a} + (\mu + B_j)^2\right]$

(IV) $E(ab\bar{y}_{\cdot\cdot}^2) = ab\left[\frac{\sigma^2}{ab} + \mu^2\right]$

$$E(SSE) = a \cdot b \cdot \sigma^2 + \sum_i \sum_j \left[\mu^2 + 2\mu\tau_i + 2\mu\beta_j + \tau_i^2 + 2\tau_i\beta_j + \beta_j^2\right]$$

$$- a\sigma^2 - ba\mu^2 - b\sum_i \tau_i^2 - b\sigma^2 - ab\mu^2 - a\sum_j \beta_j^2$$

$$+ \sigma^2 + ab\mu^2$$

(Now we use $\sum_i \tau_i = 0$ and $\sum_j \beta_j = 0$)

$$E(SSE) = \sigma^2(ab - a - b + 1) = \sigma^2(a-1)(b-1)$$

$$E(MSE) = \frac{1}{(a-1)(b-1)} \cdot \sigma^2(a-1)(b-1) = \sigma^2$$

Note:

$$(\mu + \tau_i + \beta_j)(\mu + \tau_i + \beta_j) = \mu^2 + \mu\tau_i + \mu\beta_j + \mu\tau_i + \tau_i^2 + \tau_i\beta_j + \mu\beta_j + \beta_j\tau_i + \beta_j^2$$

Question 3

(a) We can use ANOVA or linear models. With ANOVA:

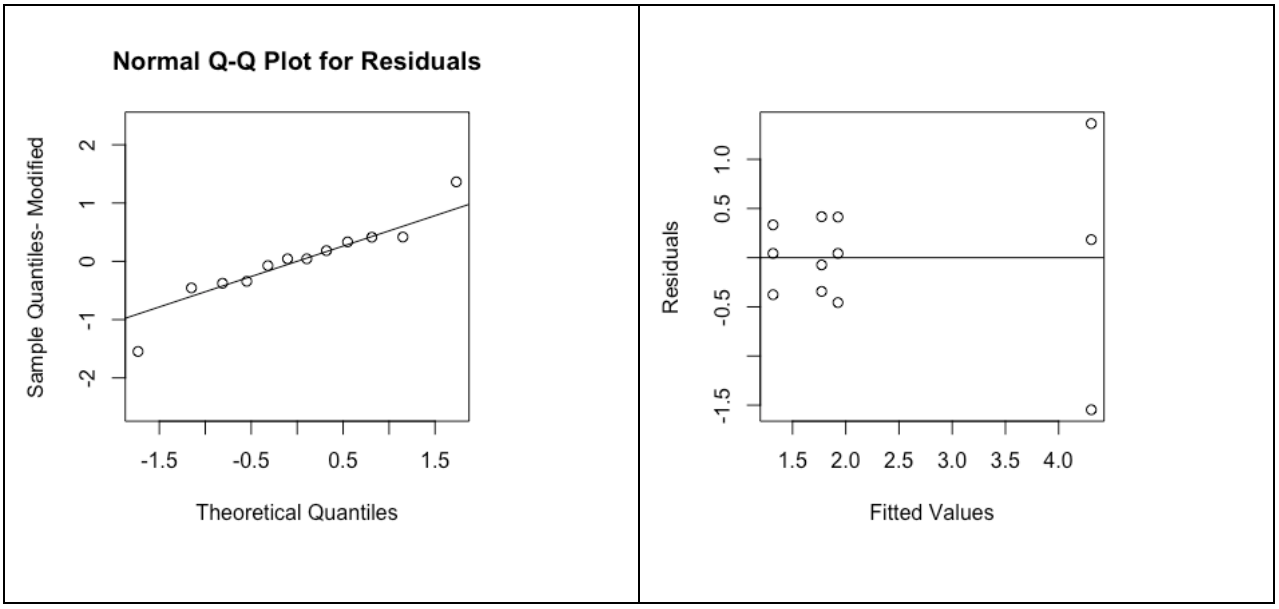
```
> res.aov <-aov(Uniformity~factor(Position),data=Uniformity_Data)
> summary(res.aov)
                Df Sum Sq Mean Sq F value  Pr(>F)
factor(Position)  3 16.220   5.407    8.29 0.00775 **
Residuals         8  5.217   0.652
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Yes, based on the ANOVA output from R, there is an effect of wafer position on the film thickness uniformity. The p-value is 0.00775 and therefore we reject the null hypothesis of H0: σ_τ^2 =0 based on α=0.05.

b) $\hat{\sigma}_\tau^2 = \frac{5.407-0.652}{3} = 1.585$ (equation 3.51).

c) $\hat{\sigma}^2 = MSE = 0.652$

d) The qqnorm plot for the residuals indicates that the normality assumption is satisfied, which can be confirmed by the Shapiro-Wilk normality test. However, there are some concerns with the variance. It might be a good idea to transform the variable *film thickness uniformity*.



Question 4

(a) Using ANOVA.

```
> res.aov <-
aov(Fuel_Economy~factor(Oil)+factor(Truck),data=FuelEconomy_Data)
> summary(res.aov)
              Df  Sum Sq  Mean Sq F value   Pr(>F)
factor(Oil)    2 0.00671 0.003353   6.353   0.0223 *
factor(Truck)  4 0.09210 0.023025  43.626 1.78e-05 ***
Residuals      8 0.00422 0.000528
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assume α=0.05. From the ANOVA table above, there is a significant difference between lubricating oils with regards to fuel economy (p-value=0.0223).

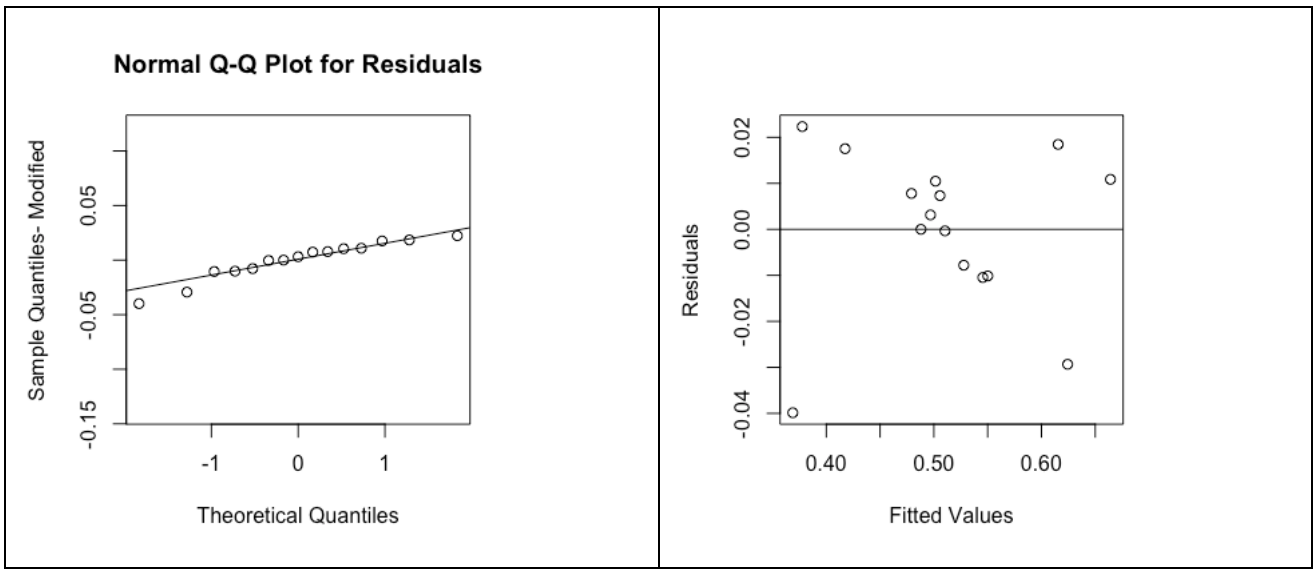
(b) Tukey method for comparisons.

```
> TUKEY <- TukeyHSD(x=res.aov,'factor(Oil)', conf.level=0.95)
> TUKEY
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Fuel_Economy ~ factor(Oil) + factor(Truck), data =
FuelEconomy_Data)

$`factor(Oil)`
       diff          lwr         upr     p adj
2-1  0.0486  0.007082078 0.090117922 0.0245809
3-1  0.0088 -0.032717922 0.050317922 0.8210970
3-2 -0.0398 -0.081317922 0.001717922 0.0594979
```

For the pairwise comparisons we assume α=0.05 for a family-wise confidence level of 95%. There is a significant difference between lubricating oils 1 and 2.

(c) Residuals.

The residual plots below do not seem to identify any violations to the normality and constant variance assumptions.



Question 5

```
> summary(res.aov)
                    Df Sum Sq Mean Sq F value   Pr(>F)
factor(Ingredients)  4 141.44   35.36  11.309 0.000488 ***
factor(Day)          4  12.24    3.06   0.979 0.455014
factor(Batch)        4  15.44    3.86   1.235 0.347618
Residuals           12  37.52    3.13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

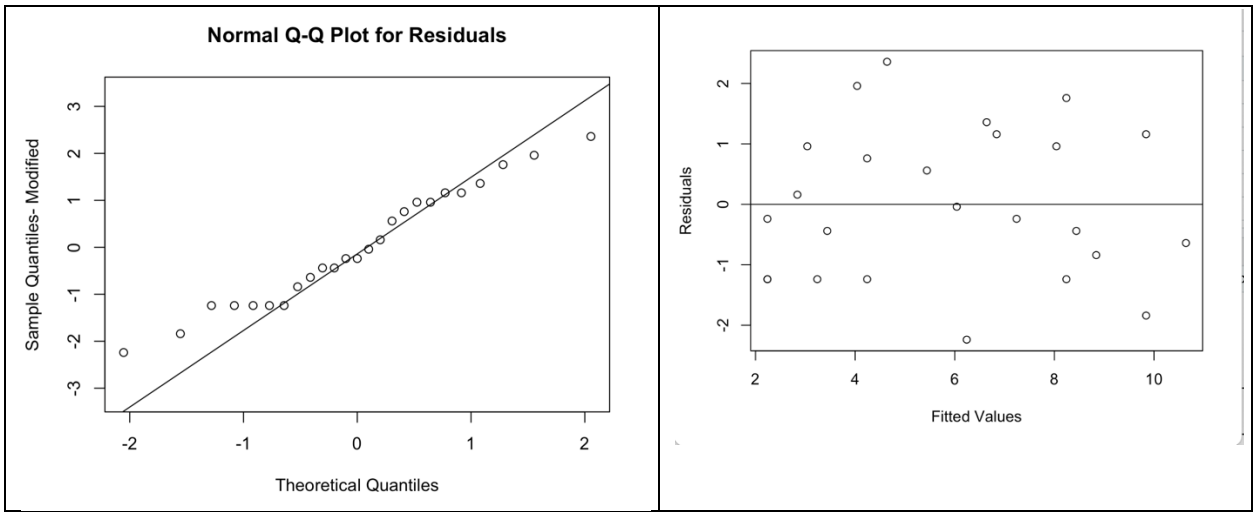
The ANOVA presented in the *R* output identifies the ingredients as having a significant effect on reaction time since p-value is approximately 0.0005, much smaller than $\alpha = 0.05$. We can identify which ingredients have different effects on the reaction time. Based on $\alpha = 0.05$, we identify differences between the following ingredient pairs: D-A, E-A, D-C, and E-C.

$$y_{ijk} = \mu + \alpha_i + \tau_j + \beta_k + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

```
> TUKEY <- TukeyHSD(x=res.aov, conf.level=0.95)
> TUKEY
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Reaction_time ~ factor(Ingredients) + factor(Day) +
factor(Batch), data = ReactionTime_data)

$`factor(Ingredients)`
    diff        lwr        upr     p adj
B-A -2.8 -6.3646078  0.7646078 0.1539433
C-A  0.4 -3.1646078  3.9646078 0.9960012
D-A -5.0 -8.5646078 -1.4353922 0.0055862
E-A -5.2 -8.7646078 -1.6353922 0.0041431
C-B  3.2 -0.3646078  6.7646078 0.0864353
D-B -2.2 -5.7646078  1.3646078 0.3365811
E-B -2.4 -5.9646078  1.1646078 0.2631551
D-C -5.4 -8.9646078 -1.8353922 0.0030822
E-C -5.6 -9.1646078 -2.0353922 0.0023007
E-D -0.2 -3.7646078  3.3646078 0.9997349
```

We check if the assumptions of normality and constant variance are satisfied based using the plots below.



There is a slight variation of the residuals from the theoretical quantiles but the normality assumption is not violated as confirmed by the Shapiro-Wilk test below. The constant variance assumption does not seem to be violated.

```
> shapiro.test(RT_residuals)

      Shapiro-Wilk normality test

data:  RT_residuals
W = 0.96606, p-value = 0.5476
```

Question 5. The effect of five different ingredients (*A, B, C, D, E*) on reaction time of a chemical process is being studied. Each batch of new material is only large enough to permit five runs to be made. Furthermore, each run requires approximately 1 1/2 hours, so only five runs can be made in one day. The experimenter decides to run the experiment as a Latin square so that day and batch effects can be systematically controlled. She obtains the data that follow. Analyze the data from this experiment (use $\alpha = 0.05$) and draw conclusions.

	Day				
Batch	1	2	3	4	5
1	*A*=8	*B*=7	*D*=1	*C*=7	*E*=3
2	*C*=11	*E*=2	*A*=7	*D*=3	*B*=8
3	*B*=4	*A*=9	*C*=10	*E*=1	*D*=5
4	*D*=6	*C*=8	*E*=6	*B*=6	*A*=10
5	*E*=4	*D*=2	*B*=3	*A*=8	*C*=8

Question 6. The yield of a chemical process was measured using five batches of raw material, five acid concentrations, five standing times, (*A, B, C, D, E*) and five catalyst concentrations ($\alpha, \beta, \gamma, \kappa, \vartheta$). The Graeco-Latin square that follows was used. Analyze the data from this experiment (use $\alpha = 0.05$) and draw conclusions.

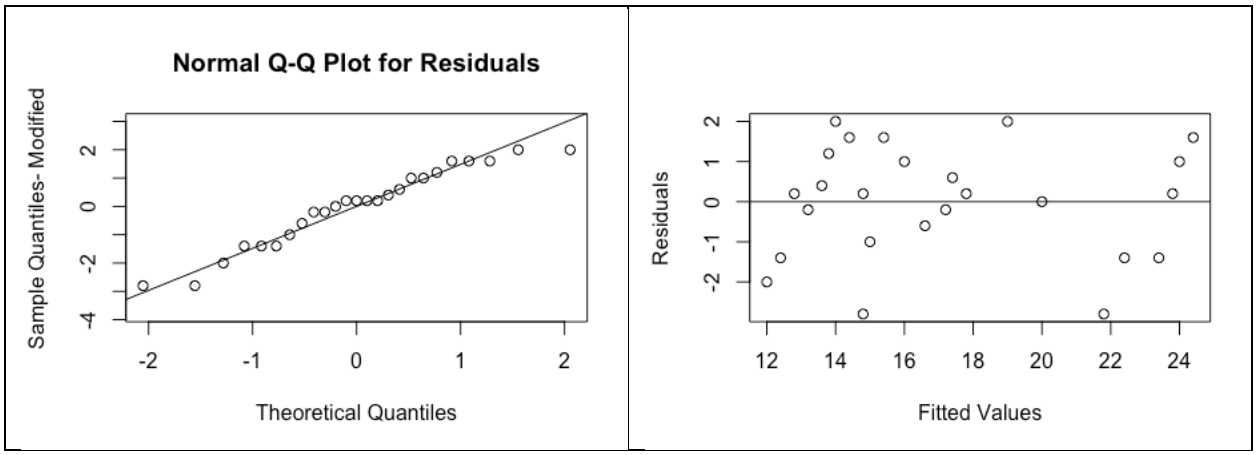
	Acid Concentration				
Batch	1	2	3	4	5
1	*Aα*=26	*Bβ*=16	*Cγ*=19	*Dκ*=16	*Eϑ*=13
2	*Bγ*=18	*Cϑ*=21	*Dϑ*=18	*Eα*=11	*Aβ*=21
3	*Cϑ*=20	*Dα*=12	*Eβ*=16	*Aγ*=25	*Bκ*=13
4	*Dβ*=15	*Eγ*=15	*Aϑ*=22	*Bϑ*=14	*Cα*=17
5	*Eκ*=10	*Aϑ*=24	*Bα*=17	*Cβ*=17	*Dγ*=14

Question 6

```
> summary(res.aov)
                     Df Sum Sq Mean Sq F value   Pr(>F)
factor(StandingTimes)  4  342.8   85.70  14.650 0.000941 ***
factor(Batch)          4   10.0    2.50   0.427 0.785447
factor(AcidConc)       4   24.4    6.10   1.043 0.442543
factor(CatalystConc)   4   12.0    3.00   0.513 0.728900
Residuals              8   46.8    5.85
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA presented in the *R* output identifies the standing times as having a significant effect on yield since p-value is approximately 0.0009, much smaller than $\alpha = 0.05$. We can identify which standing times have different effects on the reaction time. Based on $\alpha = 0.05$, we identify differences between the following standing times pairs: B-A,D-A, E-A, and E-C.

```
$`factor(StandingTimes)`
     diff        lwr        upr     p adj
B-A  -8.0 -13.284751 -2.7152488 0.0051639
C-A  -4.8 -10.084751  0.4847512 0.0770797
D-A  -8.6 -13.884751 -3.3152488 0.0032815
E-A -10.6 -15.884751 -5.3152488 0.0008219
C-B   3.2  -2.084751  8.4847512 0.3087034
D-B  -0.6  -5.884751  4.6847512 0.9939694
E-B  -2.6  -7.884751  2.6847512 0.4837165
D-C  -3.8  -9.084751  1.4847512 0.1869031
E-C  -5.8 -11.084751 -0.5152488 0.0317351
E-D  -2.0  -7.284751  3.2847512 0.6948188
```



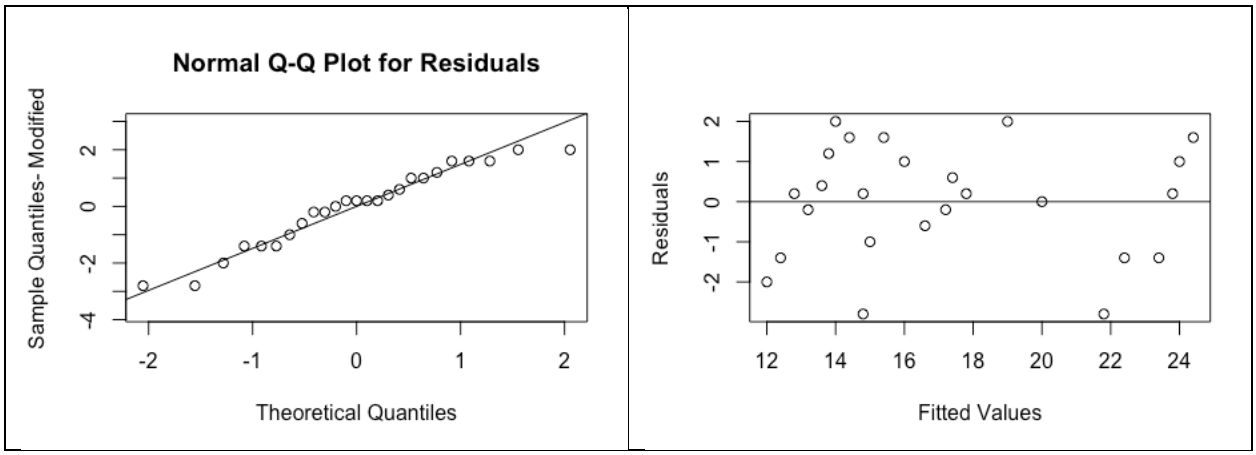
The residual plots above do not seem to identify any violations to the normality and constant variance assumptions.

Question 6

```
> summary(res.aov)
                     Df Sum Sq Mean Sq F value   Pr(>F)
factor(StandingTimes)  4  342.8   85.70  14.650 0.000941 ***
factor(Batch)          4   10.0    2.50   0.427 0.785447
factor(AcidConc)       4   24.4    6.10   1.043 0.442543
factor(CatalystConc)   4   12.0    3.00   0.513 0.728900
Residuals              8   46.8    5.85
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA presented in the *R* output identifies the standing times as having a significant effect on yield since p-value is approximately 0.0009, much smaller than $\alpha = 0.05$. We can identify which standing times have different effects on the reaction time. Based on $\alpha = 0.05$, we identify differences between the following standing times pairs: B-A,D-A, E-A, and E-C.

```
$`factor(StandingTimes)`
     diff        lwr        upr     p adj
B-A  -8.0 -13.284751 -2.7152488 0.0051639
C-A  -4.8 -10.084751  0.4847512 0.0770797
D-A  -8.6 -13.884751 -3.3152488 0.0032815
E-A -10.6 -15.884751 -5.3152488 0.0008219
C-B   3.2  -2.084751  8.4847512 0.3087034
D-B  -0.6  -5.884751  4.6847512 0.9939694
E-B  -2.6  -7.884751  2.6847512 0.4837165
D-C  -3.8  -9.084751  1.4847512 0.1869031
E-C  -5.8 -11.084751 -0.5152488 0.0317351
E-D  -2.0  -7.284751  3.2847512 0.6948188
```



The residual plots above do not seem to identify any violations to the normality and constant variance assumptions.