stats_101b_hw3_Charles_Liu

Charles Liu (304804942)

4/23/2020

Loading Necessary Packages:

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.6.3
## Warning: package 'lattice' was built under R version 3.6.3
```

```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 3.6.3
```

```
library(tidyverse)
library(crossdes)
```

```
## Warning: package 'crossdes' was built under R version 3.6.3
## Warning: package 'gtools' was built under R version 3.6.3
```

Problem 1 (Exercise 4.8)

```
values <- c(13, 22, 18, 39,
          16, 24, 17, 44,
          5, 4, 1, 22)
days <- rep(c("1 day", "2 day", "3 day", "4 day"), 3)
solutions <- c(rep("1", 4), rep("2", 4), rep("3", 4))
milk_data <- data.frame(values = values, solutions = solutions, days = days)
milk_data
```

```
##    values solutions  days
## 1      13         1 1 day
## 2      22         1 2 day
## 3      18         1 3 day
## 4      39         1 4 day
## 5      16         2 1 day
## 6      24         2 2 day
## 7      17         2 3 day
## 8      44         2 4 day
## 9       5         3 1 day
## 10      4         3 2 day
## 11      1         3 3 day
## 12     22         3 4 day
```

```r
m1 <- lm(values ~ days + solutions, data = milk_data)
summary(m1)
```

```
##
## Call:
## lm(formula = values ~ days + solutions, data = milk_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.583 -1.854 -0.250  1.250  4.417
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.5833     2.0783   7.498 0.000291 ***
## days2 day     5.3333     2.3998   2.222 0.067976 .
## days3 day     0.6667     2.3998   0.278 0.790493
## days4 day    23.6667     2.3998   9.862 6.27e-05 ***
## solutions2    2.2500     2.0783   1.083 0.320563
## solutions3  -15.0000     2.0783  -7.217 0.000359 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.939 on 6 degrees of freedom
## Multiple R-squared:  0.9722, Adjusted R-squared:  0.949
## F-statistic: 41.91 on 5 and 6 DF,  p-value: 0.0001371
```

```r
anova(m1)
```

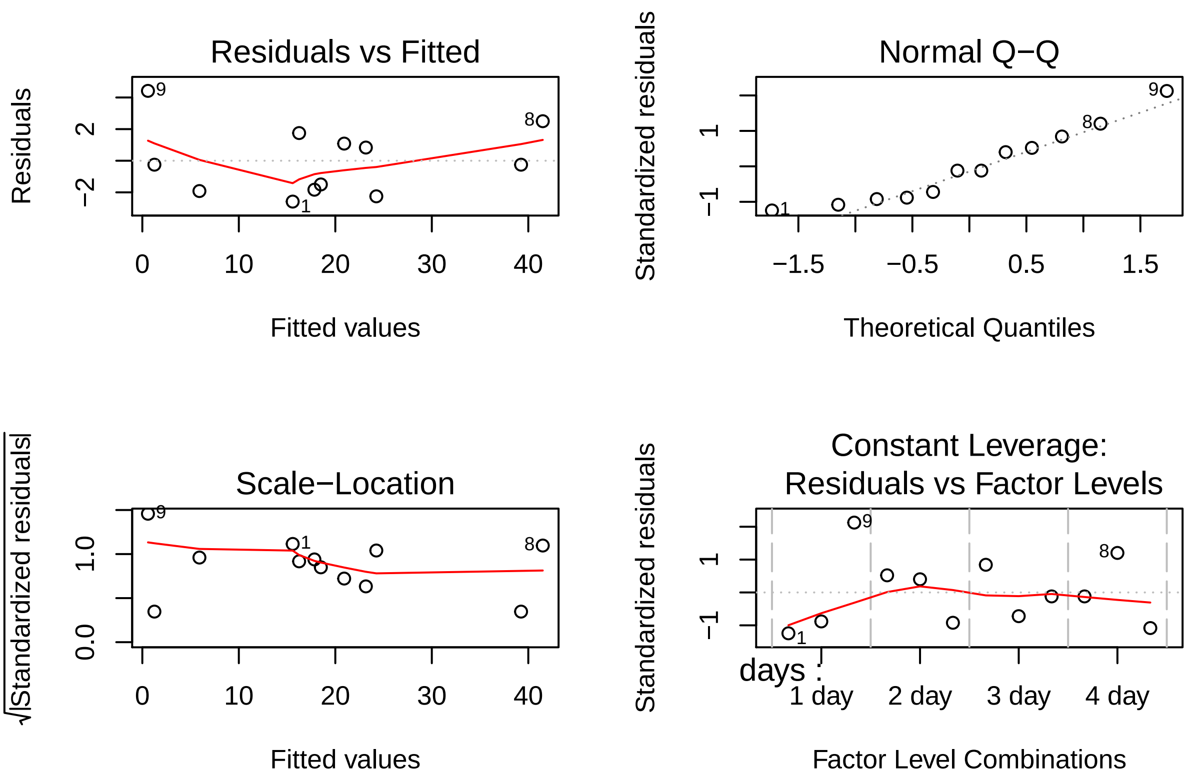
```
## Analysis of Variance Table
##
## Response: values
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## days       3 1106.92  368.97  42.711 0.0001925 ***
## solutions  2  703.50  351.75  40.717 0.0003232 ***
## Residuals  6   51.83    8.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(aov(m1))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = m1)
##
## $days
##                   diff        lwr       upr     p adj
## 2 day-1 day  5.3333333  -2.974240 13.640906 0.2193500
## 3 day-1 day  0.6666667  -7.640906  8.974240 0.9917442
## 4 day-1 day 23.6666667  15.359094 31.974240 0.0002622
## 3 day-2 day -4.6666667 -12.974240  3.640906 0.3037891
## 4 day-2 day 18.3333333  10.025760 26.640906 0.0010843
## 4 day-3 day 23.0000000  14.692427 31.307573 0.0003081
##
## $solutions
```

```
##        diff        lwr        upr     p adj
## 2-1    2.25  -4.126879   8.626879 0.5577862
## 3-1 -15.00 -21.376879  -8.623121 0.0008758
## 3-2 -17.25 -23.626879 -10.873121 0.0004067
```

```r
par(mfrow = c(2, 2))
plot(m1)
```



We can see that since the p-values for both Number of Days and Types of Solutions are small (or their F_0 values are very big), we would REJECT the Null Hypothesis. This means that both Days and Solutions have a significant effect on retarding bacteria growth.

We can see that the following pairs are significant from our Tukey Method for Days: 4 day-1 day, 4 day-2 day, & 4 day-3 day. The following pairs are significant from our Tukey Method for Solutions: 3-1 & 3-2. (all our Tukey Method uses a significance level of $\alpha = 0.05$)

From our plots, we can see it relatively satisfies the Normality asummption, but it does NOT satisfy the Linearity and Homoscedasticity assumptions (Residuals vs. Fitted & Scale-Location plots).

Problem 2 (Exercise 4.24)

Problem 2 (done by "hand")

```r
letter <- c("A", "B", "D", "C", "E",
            "C", "E", "A", "D", "B",
            "B", "A", "C", "E", "D",
            "D", "C", "E", "B", "A",
```

```r
                    "E", "D", "B", "A", "C")
value <- 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)
day <- rep(c("Day 1", "Day 2", "Day 3", "Day 4", "Day 5"), 5)
batch <- c(rep("1", 5), rep("2", 5), rep("3", 5), rep("4", 5), rep("5", 5))
chem_data <- data.frame(letter = letter, value = value, batch = batch, day = day)
chem_data
```

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

```r
# I'll be calculating everything by "hand" using Rstudio
total_batch <- chem_data %>% select(batch, value) %>% group_by(batch) %>% summarise(row_total = sum(valu
total_batch
```

```
## # A tibble: 5 x 2
##   batch row_total
##   <fct>     <dbl>
## 1 1            26
## 2 2            31
## 3 3            29
## 4 4            36
## 5 5            25
```

```r
total_day <- chem_data %>% select(day, value) %>% group_by(day) %>% summarise(column_total = sum(value)
total_day
```

```
## # A tibble: 5 x 2
```

```
##   day   column_total
##   <fct>        <dbl>
## 1 Day 1           33
## 2 Day 2           28
## 3 Day 3           27
## 4 Day 4           25
## 5 Day 5           34
```

```
total_letter <- chem_data %>% select(letter, value) %>% group_by(letter) %>% summarise(letter_sum = sum
total_letter
```

```
## # A tibble: 5 x 2
##   letter letter_sum
##   <fct>       <dbl>
## 1 A              42
## 2 B              28
## 3 C              44
## 4 D              17
## 5 E              16
```

```
total <- chem_data %>% select(letter, value) %>% summarise(total_values = sum(value)) # 147

# Setting up our paramters:
N <- 25 # (5x5) matrix
n <- 5 # can work for both Day and Batch since it's a (5x5) matrix
p <- 5
CF <- (147^2)/(N)

# Sum of Squares:
row_per_batch <- c(26, 31, 29, 36, 25)
SSR_batch <- ((1/n) * (sum(row_per_batch[1:5]^2))) - CF # SSR_batch

column_per_day <- c(33, 28, 27, 25, 34)
SSR_day <- ((1/n) * (sum(column_per_day[1:5]^2))) - CF # SSR_day

letter_per <- c(42, 28, 44, 17, 16)
SSR_letter <- ((1/n) * (sum(letter_per[1:5]^2))) - CF # SSR_letter

# Created a for() loop for Total Sum of Squares
for(i in 1:25) {
  new_value <- (value[1:25])^2
  SS_all <- sum(new_value)
  return(SS_all)
}

SST <- SS_all - CF # Total Sum of Squares

SSE <- SST - SSR_batch - SSR_day - SSR_letter # Sum of Square of Errors

# Mean Sum of Squares:
MS_batch <- SSR_batch/(p - 1)
MS_day <- SSR_day/(p - 1)
MS_letter <- SSR_letter/(p - 1)
MSE <- SSE/((p - 2) * (p - 1))
```

```r
# F-Test:
F_batch <- (MS_batch/MSE)
F_day <- (MS_day/MSE)
F_letter <- (MS_letter/MSE)



# All of our Values for the ANOVA table:
## Degrees of Freedom Values:
(p - 1) # Rows, Columns, Letters
```

```
## [1] 4
```

```r
(p - 2)*(p - 1) # SSE
```

```
## [1] 12
```

```r
(p^2 - 1) # SST
```

```
## [1] 24
```

```r
## Sum of Square Values:
SSR_batch
```

```
## [1] 15.44
```

```r
SSR_day
```

```
## [1] 12.24
```

```r
SSR_letter
```

```
## [1] 141.44
```

```r
SSE
```

```
## [1] 37.52
```

```r
SST
```

```
## [1] 206.64
```

```r
## Mean Sum of Square Values:
MS_batch
```

```
## [1] 3.86
```

```r
MS_day
```

```
## [1] 3.06
```

```r
MS_letter
```

```
## [1] 35.36
```

```r
MSE
```

```
## [1] 3.126667
```

```r
## F-test Values:
F_batch
```

```
## [1] 1.234542
```

```
F_day
```

```
## [1] 0.978678
```

```
F_letter
```

```
## [1] 11.30917
```

Problem 2 (done by RStudio)

```
# Simple Way by Using lm():
m2 <- lm(value ~ day + batch + letter, data = chem_data)
summary(m2)
```

```
##
## Call:
## lm(formula = value ~ day + batch + letter, data = chem_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2.24  -1.24  -0.24   0.96   2.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.440      1.275   6.619 2.47e-05 ***
## dayDay 2      -1.000      1.118  -0.894 0.388803
## dayDay 3      -1.200      1.118  -1.073 0.304364
## dayDay 4      -1.600      1.118  -1.431 0.178037
## dayDay 5       0.200      1.118   0.179 0.861049
## batch2         1.000      1.118   0.894 0.388803
## batch3         0.600      1.118   0.537 0.601409
## batch4         2.000      1.118   1.788 0.098971 .
## batch5        -0.200      1.118  -0.179 0.861049
## letterB       -2.800      1.118  -2.504 0.027725 *
## letterC        0.400      1.118   0.358 0.726798
## letterD       -5.000      1.118  -4.471 0.000764 ***
## letterE       -5.200      1.118  -4.650 0.000561 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.768 on 12 degrees of freedom
## Multiple R-squared:  0.8184, Adjusted R-squared:  0.6369
## F-statistic: 4.507 on 12 and 12 DF,  p-value: 0.007157
```

```
anova(m2) # We can see all our answers done by hand are the same!
```

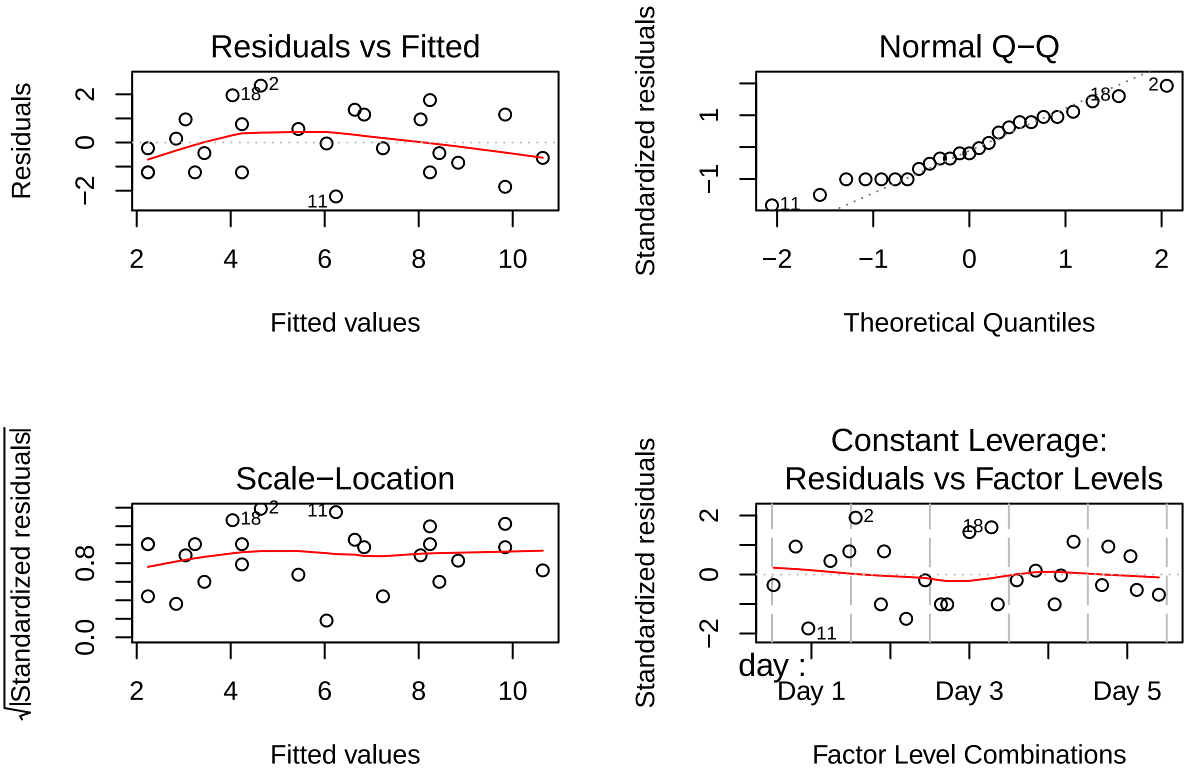
```
## Analysis of Variance Table
##
## Response: value
##           Df Sum Sq Mean Sq F value    Pr(>F)
## day        4  12.24   3.060  0.9787 0.4550143
## batch      4  15.44   3.860  1.2345 0.3476182
## letter     4 141.44  35.360 11.3092 0.0004877 ***
## Residuals 12  37.52   3.127
## ---
```

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

```r
TukeyHSD(aov(m2))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = m2)
##
## $day
##               diff       lwr      upr     p adj
## Day 2-Day 1 -1.0 -4.564608 2.564608 0.8936609
## Day 3-Day 1 -1.2 -4.764608 2.364608 0.8166339
## Day 4-Day 1 -1.6 -5.164608 1.964608 0.6212723
## Day 5-Day 1  0.2 -3.364608 3.764608 0.9997349
## Day 3-Day 2 -0.2 -3.764608 3.364608 0.9997349
## Day 4-Day 2 -0.6 -4.164608 2.964608 0.9816047
## Day 5-Day 2  1.2 -2.364608 4.764608 0.8166339
## Day 4-Day 3 -0.4 -3.964608 3.164608 0.9960012
## Day 5-Day 3  1.4 -2.164608 4.964608 0.7232162
## Day 5-Day 4  1.8 -1.764608 5.364608 0.5188508
##
## $batch
##     diff       lwr      upr     p adj
## 2-1  1.0 -2.564608 4.564608 0.8936609
## 3-1  0.6 -2.964608 4.164608 0.9816047
## 4-1  2.0 -1.564608 5.564608 0.4225127
## 5-1 -0.2 -3.764608 3.364608 0.9997349
## 3-2 -0.4 -3.964608 3.164608 0.9960012
## 4-2  1.0 -2.564608 4.564608 0.8936609
## 5-2 -1.2 -4.764608 2.364608 0.8166339
## 4-3  1.4 -2.164608 4.964608 0.7232162
## 5-3 -0.8 -4.364608 2.764608 0.9489243
## 5-4 -2.2 -5.764608 1.364608 0.3365811
##
## $letter
##     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
```

```r
par(mfrow = c(2, 2))
plot(m2)
```



We can see that only Ingredients (Letters) have a small p-value, so we would REJECT the Null Hypothesis for Ingredients. Therefore, we can say that Ingredients have a significant effect on the Chemical Process. As for Day and Batch, they are both NOT significant, so we would NOT REJECT the Null Hypothesis. For these two, we can say they do NOT play a significant role on the Chemical Process.

There is no significance for Day and Batch under Tukey Method (with a significance level $\alpha = 0.05$). We can see under Ingredients (Letters), there is significance between the following pairs: D-A, E-A, D-C, & E-C.

From our plots, we can see it relatively satisfies the Normality and Homoscedasticity asummptions, but it does NOT satisfy the Linearity assumptions (Residuals vs. Fitted).

Problem 3 (Exercise 4.37)

```
standing_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")
catalysts <- c("alpha", "beta", "gamma", "delta", "epsilon",
               "gamma", "delta", "epsilon", "alpha", "beta",
               "epsilon", "alpha", "beta", "gamma", "delta",
               "beta", "gamma", "delta", "epsilon", "alpha",
               "delta", "epsilon", "alpha", "beta", "gamma")
amount <- c(26, 16, 19, 16, 13,
            18, 21, 18, 11, 21,
            20, 12, 16, 25, 13,
```

```r
           15, 15, 22, 14, 17,
           10, 24, 17, 17, 14)
acid_concentration <- rep(c("Acid 1", "Acid 2", "Acid 3", "Acid 4", "Acid 5"), 5)
batches <- c(rep("1", 5), rep("2", 5), rep("3", 5), rep("4", 5), rep("5", 5))
chem_process_data <- data.frame(standing_times = standing_times, catalysts = catalysts, amount = amount
chem_process_data
```

```
##    standing_times catalysts amount acid_concentration batches
## 1               A     alpha     26             Acid 1       1
## 2               B      beta     16             Acid 2       1
## 3               C     gamma     19             Acid 3       1
## 4               D     delta     16             Acid 4       1
## 5               E   epsilon     13             Acid 5       1
## 6               B     gamma     18             Acid 1       2
## 7               C     delta     21             Acid 2       2
## 8               D   epsilon     18             Acid 3       2
## 9               E     alpha     11             Acid 4       2
## 10              A      beta     21             Acid 5       2
## 11              C   epsilon     20             Acid 1       3
## 12              D     alpha     12             Acid 2       3
## 13              E      beta     16             Acid 3       3
## 14              A     gamma     25             Acid 4       3
## 15              B     delta     13             Acid 5       3
## 16              D      beta     15             Acid 1       4
## 17              E     gamma     15             Acid 2       4
## 18              A     delta     22             Acid 3       4
## 19              B   epsilon     14             Acid 4       4
## 20              C     alpha     17             Acid 5       4
## 21              E     delta     10             Acid 1       5
## 22              A   epsilon     24             Acid 2       5
## 23              B     alpha     17             Acid 3       5
## 24              C      beta     17             Acid 4       5
## 25              D     gamma     14             Acid 5       5
```

```r
# Using RStudio to find conclusions (lm(...) method)
m3 <- lm(amount ~ standing_times + catalysts + acid_concentration + batches, data = chem_process_data)
summary(m3)
```

```
##
## Call:
## lm(formula = amount ~ standing_times + catalysts + acid_concentration +
##     batches, data = chem_process_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.8   -1.0    0.2    1.0    2.0
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                24.400      1.994  12.234 1.85e-06 ***
## standing_timesB            -8.000      1.530  -5.230 0.000793 ***
## standing_timesC            -4.800      1.530  -3.138 0.013850 *
## standing_timesD            -8.600      1.530  -5.622 0.000497 ***
## standing_timesE           -10.600      1.530  -6.929 0.000121 ***
```

```
## catalystsbeta               0.400      1.530   0.261 0.800322
## catalystsdelta             -0.200      1.530  -0.131 0.899206
## catalystsepsilon            1.200      1.530   0.784 0.455366
## catalystsgamma              1.600      1.530   1.046 0.326154
## acid_concentrationAcid 2   -0.200      1.530  -0.131 0.899206
## acid_concentrationAcid 3    0.600      1.530   0.392 0.705129
## acid_concentrationAcid 4   -1.200      1.530  -0.784 0.455366
## acid_concentrationAcid 5   -2.200      1.530  -1.438 0.188326
## batches2                   -0.200      1.530  -0.131 0.899206
## batches3                   -0.800      1.530  -0.523 0.615162
## batches4                   -1.400      1.530  -0.915 0.386837
## batches5                   -1.600      1.530  -1.046 0.326154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.419 on 8 degrees of freedom
## Multiple R-squared:  0.8927, Adjusted R-squared:  0.678
## F-statistic: 4.158 on 16 and 8 DF,  p-value: 0.02355
```

```
anova(m3)
```

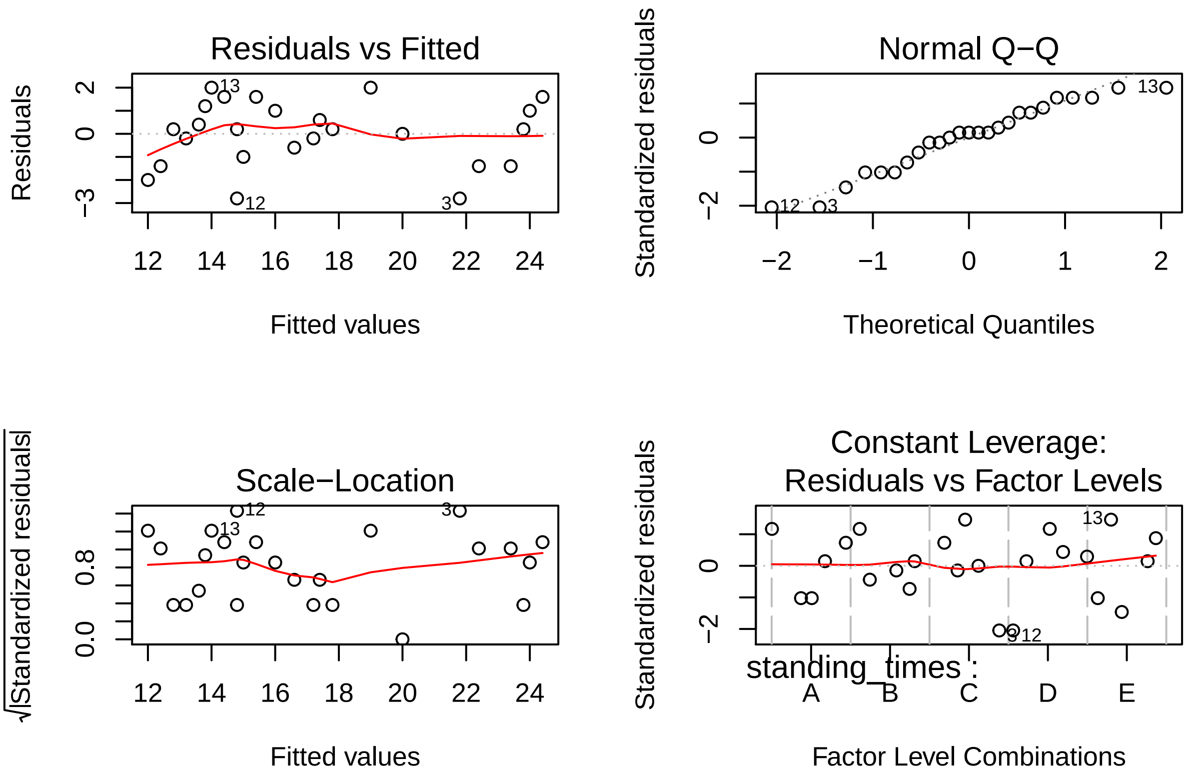
```
## Analysis of Variance Table
##
## Response: amount
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## standing_times      4  342.8   85.70 14.6496 0.000941 ***
## catalysts           4   12.0    3.00  0.5128 0.728900
## acid_concentration  4   24.4    6.10  1.0427 0.442543
## batches             4   10.0    2.50  0.4274 0.785447
## Residuals           8   46.8    5.85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(aov(m3))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = m3)
##
## $standing_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
##
## $catalysts
##               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
##
## $acid_concentration
##               diff       lwr      upr     p adj
## Acid 2-Acid 1 -0.2 -5.484751 5.084751 0.9999182
## Acid 3-Acid 1  0.6 -4.684751 5.884751 0.9939694
## Acid 4-Acid 1 -1.2 -6.484751 4.084751 0.9281909
## Acid 5-Acid 1 -2.2 -7.484751 3.084751 0.6232282
## Acid 3-Acid 2  0.8 -4.484751 6.084751 0.9823986
## Acid 4-Acid 2 -1.0 -6.284751 4.284751 0.9610846
## Acid 5-Acid 2 -2.0 -7.284751 3.284751 0.6948188
## Acid 4-Acid 3 -1.8 -7.084751 3.484751 0.7640759
## Acid 5-Acid 3 -2.8 -8.084751 2.484751 0.4197369
## Acid 5-Acid 4 -1.0 -6.284751 4.284751 0.9610846
##
## $batches
##     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
```

```r
par(mfrow = c(2, 2))
plot(m3)
```



We can see that only Standing Times (Latin Letters) have a small p-value, so we would REJECT the Null Hypothesis for Standing Times Therefore, we can say that Standing Times have a significant effect on the Chemical Process. As for Catalysts (Greek Letters), Acid Concentration, and Batches, they are all NOT significant, so we would NOT REJECT the Null Hypothesis. For these three, we can say they do NOT play a significant role on the Chemical Process.

There is no significance for Catalysts (Greek Letters), Acid Concentration, and Batches under Tukey Method (with a significance level $\alpha = 0.05$). We can see under Standing Times (Latin Letters), there is significance between the following pairs: B-A, D-A, E-A, & E-C.

From our plots, we can see it relatively satisfies the Linearity, Normality, and Homoscedasticity asummptions. Even if lines are not perfectly horizontal, they are mostly horizontal. Therefore, all our assumptions are satisfied.

Problem 4 (Exercise 4.44)

```
amounts <- c(114, 0, 0, 0, 120, 0, 117,
             126, 120, 0, 0, 0, 119, 0,
             0, 137, 117, 0, 0, 0, 134,
             141, 0, 129, 149, 0, 0, 0,
             0, 145, 0, 150, 143, 0, 0,
             0, 0, 120, 0, 118, 123, 0,
             0, 0, 0, 136, 0, 130, 127)
time <- rep(c("Day 1", "Day 2", "Day 3", "Day 4", "Day 5", "Day 6", "Day 7"), 7)
hardwood_concentration <- c(rep("2%", 7), rep("4%", 7), rep("6%", 7), rep("8%", 7), rep("10%", 7), rep(
paper_data <- data.frame(time = time, hardwood_concentration = hardwood_concentration, amounts = amount
```

```r
paper_data
```

```
##      time hardwood_concentration amounts
## 1  Day 1                      2%     114
## 2  Day 2                      2%       0
## 3  Day 3                      2%       0
## 4  Day 4                      2%       0
## 5  Day 5                      2%     120
## 6  Day 6                      2%       0
## 7  Day 7                      2%     117
## 8  Day 1                      4%     126
## 9  Day 2                      4%     120
## 10 Day 3                      4%       0
## 11 Day 4                      4%       0
## 12 Day 5                      4%       0
## 13 Day 6                      4%     119
## 14 Day 7                      4%       0
## 15 Day 1                      6%       0
## 16 Day 2                      6%     137
## 17 Day 3                      6%     117
## 18 Day 4                      6%       0
## 19 Day 5                      6%       0
## 20 Day 6                      6%       0
## 21 Day 7                      6%     134
## 22 Day 1                      8%     141
## 23 Day 2                      8%       0
## 24 Day 3                      8%     129
## 25 Day 4                      8%     149
## 26 Day 5                      8%       0
## 27 Day 6                      8%       0
## 28 Day 7                      8%       0
## 29 Day 1                     10%       0
## 30 Day 2                     10%     145
## 31 Day 3                     10%       0
## 32 Day 4                     10%     150
## 33 Day 5                     10%     143
## 34 Day 6                     10%       0
## 35 Day 7                     10%       0
## 36 Day 1                     12%       0
## 37 Day 2                     12%       0
## 38 Day 3                     12%     120
## 39 Day 4                     12%       0
## 40 Day 5                     12%     118
## 41 Day 6                     12%     123
## 42 Day 7                     12%       0
## 43 Day 1                     14%       0
## 44 Day 2                     14%       0
## 45 Day 3                     14%       0
## 46 Day 4                     14%     136
## 47 Day 5                     14%       0
## 48 Day 6                     14%     130
## 49 Day 7                     14%     127
```

```r
# Using RStudio to find conclusions (lm(...) method)
m4 <- lm(amounts ~ time + hardwood_concentration, data = paper_data)
summary(m4)
```

```
##
## Call:
## lm(formula = amounts ~ time + hardwood_concentration, data = paper_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62.16 -56.88 -50.16  72.27  82.41
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.159e+01  3.857e+01   1.597    0.119
## timeDay 2                 3.000e+00  4.002e+01   0.075    0.941
## timeDay 3                -2.143e+00  4.002e+01  -0.054    0.958
## timeDay 4                 7.714e+00  4.002e+01   0.193    0.848
## timeDay 5                 1.773e-14  4.002e+01   0.000    1.000
## timeDay 6                -1.286e+00  4.002e+01  -0.032    0.975
## timeDay 7                -4.286e-01  4.002e+01  -0.011    0.992
## hardwood_concentration12% -1.100e+01  4.002e+01  -0.275    0.785
## hardwood_concentration14% -6.429e+00  4.002e+01  -0.161    0.873
## hardwood_concentration2%  -1.243e+01  4.002e+01  -0.311    0.758
## hardwood_concentration4%  -1.043e+01  4.002e+01  -0.261    0.796
## hardwood_concentration6%  -7.143e+00  4.002e+01  -0.178    0.859
## hardwood_concentration8%  -2.714e+00  4.002e+01  -0.068    0.946
##
## Residual standard error: 74.88 on 36 degrees of freedom
## Multiple R-squared:  0.006648,   Adjusted R-squared:  -0.3245
## F-statistic: 0.02008 on 12 and 36 DF,  p-value: 1
```

```r
anova(m4)
```

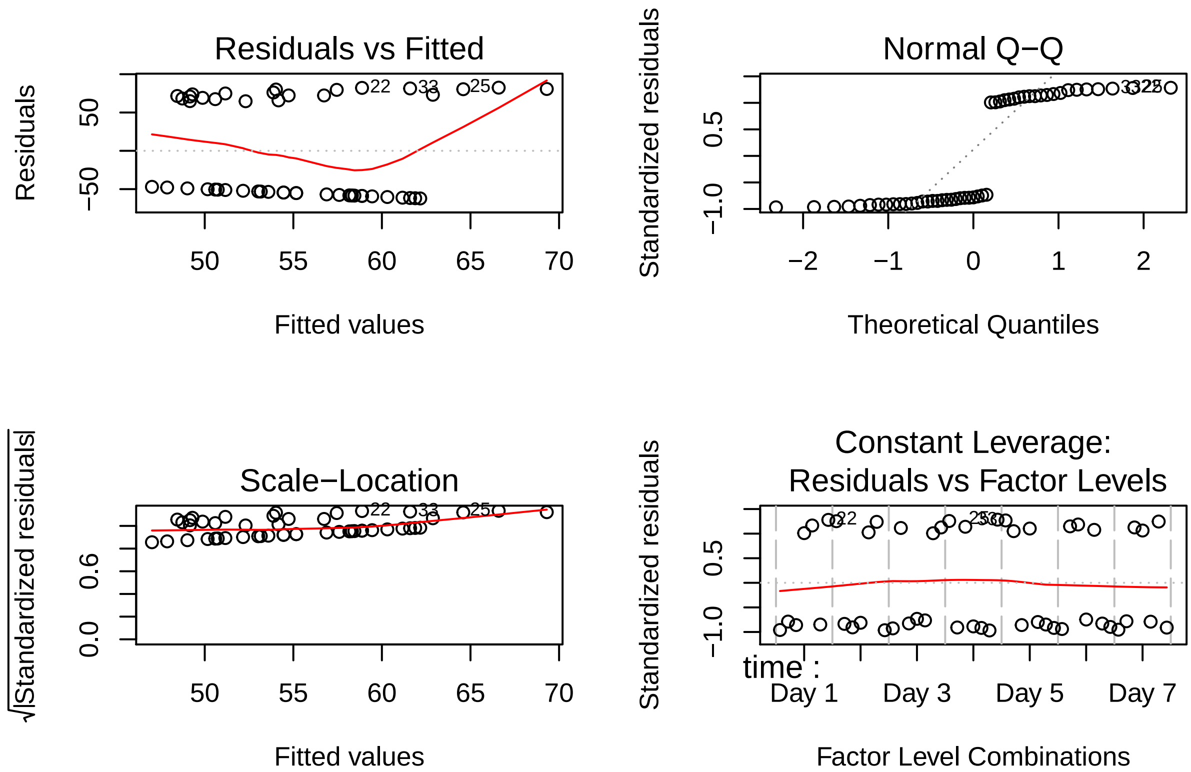
```
## Analysis of Variance Table
##
## Response: amounts
##                        Df Sum Sq Mean Sq F value Pr(>F)
## time                    6    478    79.6  0.0142 1.0000
## hardwood_concentration  6    873   145.5  0.0260 0.9999
## Residuals              36 201827  5606.3
```

```r
TukeyHSD(aov(m4))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = m4)
##
## $time
##                         diff       lwr      upr     p adj
## Day 2-Day 1  3.000000e+00 -121.9042 127.9042 1.0000000
## Day 3-Day 1 -2.142857e+00 -127.0470 122.7613 1.0000000
## Day 4-Day 1  7.714286e+00 -117.1899 132.6185 0.9999950
## Day 5-Day 1  2.131628e-14 -124.9042 124.9042 1.0000000
```

```
## Day 6-Day 1 -1.285714e+00 -126.1899 123.6185 1.0000000
## Day 7-Day 1 -4.285714e-01 -125.3328 124.4756 1.0000000
## Day 3-Day 2 -5.142857e+00 -130.0470 119.7613 0.9999996
## Day 4-Day 2  4.714286e+00 -120.1899 129.6185 0.9999997
## Day 5-Day 2 -3.000000e+00 -127.9042 121.9042 1.0000000
## Day 6-Day 2 -4.285714e+00 -129.1899 120.6185 0.9999999
## Day 7-Day 2 -3.428571e+00 -128.3328 121.4756 1.0000000
## Day 4-Day 3  9.857143e+00 -115.0470 134.7613 0.9999787
## Day 5-Day 3  2.142857e+00 -122.7613 127.0470 1.0000000
## Day 6-Day 3  8.571429e-01 -124.0470 125.7613 1.0000000
## Day 7-Day 3  1.714286e+00 -123.1899 126.6185 1.0000000
## Day 5-Day 4 -7.714286e+00 -132.6185 117.1899 0.9999950
## Day 6-Day 4 -9.000000e+00 -133.9042 115.9042 0.9999875
## Day 7-Day 4 -8.142857e+00 -133.0470 116.7613 0.9999931
## Day 6-Day 5 -1.285714e+00 -126.1899 123.6185 1.0000000
## Day 7-Day 5 -4.285714e-01 -125.3328 124.4756 1.0000000
## Day 7-Day 6  8.571429e-01 -124.0470 125.7613 1.0000000
##
## $hardwood_concentration
##                diff       lwr      upr     p adj
## 12%-10% -11.0000000 -135.9042 113.9042 0.9999592
## 14%-10%  -6.4285714 -131.3328 118.4756 0.9999983
## 2%-10%  -12.4285714 -137.3328 112.4756 0.9999165
## 4%-10%  -10.4285714 -135.3328 114.4756 0.9999702
## 6%-10%   -7.1428571 -132.0470 117.7613 0.9999968
## 8%-10%   -2.7142857 -127.6185 122.1899 1.0000000
## 14%-12%   4.5714286 -120.3328 129.4756 0.9999998
## 2%-12%   -1.4285714 -126.3328 123.4756 1.0000000
## 4%-12%    0.5714286 -124.3328 125.4756 1.0000000
## 6%-12%    3.8571429 -121.0470 128.7613 0.9999999
## 8%-12%    8.2857143 -116.6185 133.1899 0.9999924
## 2%-14%   -6.0000000 -130.9042 118.9042 0.9999989
## 4%-14%   -4.0000000 -128.9042 120.9042 0.9999999
## 6%-14%   -0.7142857 -125.6185 124.1899 1.0000000
## 8%-14%    3.7142857 -121.1899 128.6185 0.9999999
## 4%-2%     2.0000000 -122.9042 126.9042 1.0000000
## 6%-2%     5.2857143 -119.6185 130.1899 0.9999995
## 8%-2%     9.7142857 -115.1899 134.6185 0.9999804
## 6%-4%     3.2857143 -121.6185 128.1899 1.0000000
## 8%-4%     7.7142857 -117.1899 132.6185 0.9999950
## 8%-6%     4.4285714 -120.4756 129.3328 0.9999998
```

```r
par(mfrow = c(2, 2))
plot(m4)
```



We can see that none of the factors has a p-value less than 0.05, so they are all NOT significant. We would NOT REJECT the Null Hypothesis, and we can say they do NOT play a significant role on the Strength of Paper.

There is no significance for Time and Hardwood Concentration under Tukey Method (with a significance level $\alpha = 0.05$).

From our plots, we can see that NONE of our assumptions are satisfied (Linearity, Normality, and Homoscedasticity asummptions).

Problem 5

```
design_1 <- matrix(c(1, 2, 3, 4, 5,
                     2, 3, 4, 5, 1,
                     3, 4, 5, 1, 3,
                     4, 5, 1, 2, 4), nrow = 5, ncol = 4)

design_2 <- matrix(c(1, 5, 2, 3, 1,
                     2, 1, 3, 4, 2,
                     3, 2, 4, 5, 4,
                     4, 3, 5, 1, 5), nrow = 5, ncol = 4)
```

5a)

```r
# (a > k); a*gamma = b * k; (a - 1) * lambda = (k - 1) * gamma

design_1
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    2    3    4    5
## [3,]    3    4    5    1
## [4,]    4    5    1    2
## [5,]    5    1    3    4
```

```r
# a = 5, b = 5, k = 4, gamma = 3, lambda = 4
# (i) a * gamma = b * k --> (5) * (3) = (5) * (4) --> NOT TRUE (=/=)
# (ii) (a - 1) * (lambda) = (k - 1) * (gamma) --> (5 - 1) * (4) = (4 - 1) * (3) --> NOT TRUE (=/=)


design_2
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    5    1    2    3
## [3,]    2    3    4    5
## [4,]    3    4    5    1
## [5,]    1    2    4    5
```

```r
# a = 5, b = 5, k = 4, gamma = 4, lambda = 3
# (i) a * gamma = b * k --> (5) * (4) = (5) * (4) --> TRUE (=)
# (ii) (a - 1) * (lambda) = (k - 1) * (gamma) --> (5 - 1) * 3 = (4 - 1) * (4) --> TRUE (=)







# Creating our Incidence Matrices:
incidence_1 <- matrix(c(1, 1, 1, 1, 0,
                        0, 1, 1, 1, 1,
                        1, 0, 1, 1, 1,
                        1, 1, 0, 1, 1,
                        1, 0, 1, 1, 1), nrow = 5, ncol = 5)
incidence_1
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    1    1    1
## [2,]    1    1    0    1    0
## [3,]    1    1    1    0    1
## [4,]    1    1    1    1    1
## [5,]    0    1    1    1    1
```

```r
incidence_2 <- matrix(c(1, 1, 1, 1, 0,
                        1, 1, 1, 0, 1,
                        0, 1, 1, 1, 1,
                        1, 0, 1, 1, 1,
                        1, 1, 0, 1, 1), nrow = 5, ncol = 5)
incidence_2
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    1    0    1    1
## [2,]    1    1    1    0    1
## [3,]    1    1    1    1    0
## [4,]    1    0    1    1    1
## [5,]    0    1    1    1    1
```

5b)

```r
# Testing our Incidence Matrices:
incidence_1 %*% t(incidence_1) # This one is NOT BIBD.
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    4    2    3    4    3
## [2,]    2    3    2    3    2
## [3,]    3    2    4    4    3
## [4,]    4    3    4    5    4
## [5,]    3    2    3    4    4
```

```r
incidence_2 %*% t(incidence_2) # This one is BIBD! We would choose Design 2 for BIBD.
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    4    3    3    3    3
## [2,]    3    4    3    3    3
## [3,]    3    3    4    3    3
## [4,]    3    3    3    4    3
## [5,]    3    3    3    3    4
```

```r
# Double Checking our Work:
isGYD(design_1)
```

```
##
## [1] The design is neither balanced w.r.t. rows nor w.r.t. columns.
```

```r
# NOT BIBD

isGYD(design_2)
```

```
##
## [1] The design is a balanced incomplete block design w.r.t. rows.
```

```r
# Is BIBD
```