

Homework 2

Igor Kuivjogi Fernandes

2023-02-02

0. An ANOVA output is shown below. Fill in the missing information.
One-way ANOVA

Source	DF	SS	MS	F	P
Factor	3	36.15	?	?	?
Error	?	?	?		
Total	19	196.04			

Completing the cells we got:

Source	DF	SS	MS	F	P
Factor	3	36.15	12.05	1.213418	0.3369274
Error	16	158.89	9.930625		
Total	19	196.04			

The p-value is the $P(F > 1.123418) = 0.3369274$, with $df_{\text{factor}} = 3$ and $df_{\text{error}} = 16$.

1. I belong to a golf club in my neighborhood. I divide the year into three golf seasons: summer (June–September), winter (November–March), and shoulder (October, April, and May). I believe that I play my best golf during the summer (because I have more time and the course isn't crowded) and shoulder (because the course isn't crowded) seasons and my worst golf is during the winter (because when all of the part-year residents show up, the course is crowded, play is slow, and I get frustrated). Data from the last year are shown in the following table.

We can write a hypothesis test as follows:

H_0 : the golf performance in the seasons are equal

H_1 : at least one golf performance differs

```
df <- data.frame(
  season = c(rep('summer', 10), rep('shoulder', 7), rep('winter', 8)),
  y = c(83, 85, 85, 87, 90, 88, 88, 84, 91, 90, 91, 87, 84,
        87, 85, 86, 83, 94, 91, 87, 85, 87, 91, 92, 86)
)

# frequency table
cat('summer:', df[df$season == 'summer', 'y'], '\n',
    'shoulder:', df[df$season == 'shoulder', 'y'], '\n',
    'winter:', df[df$season == 'winter', 'y'], '\n\n')
```

```
## summer: 83 85 85 87 90 88 88 84 91 90
## shoulder: 91 87 84 87 85 86 83
## winter: 94 91 87 85 87 91 92 86
```

```
# one-way ANOVA
```

```
df$season <- as.factor(df$season)
fit <- aov(y ~ season, data = df)
summary(fit)
```

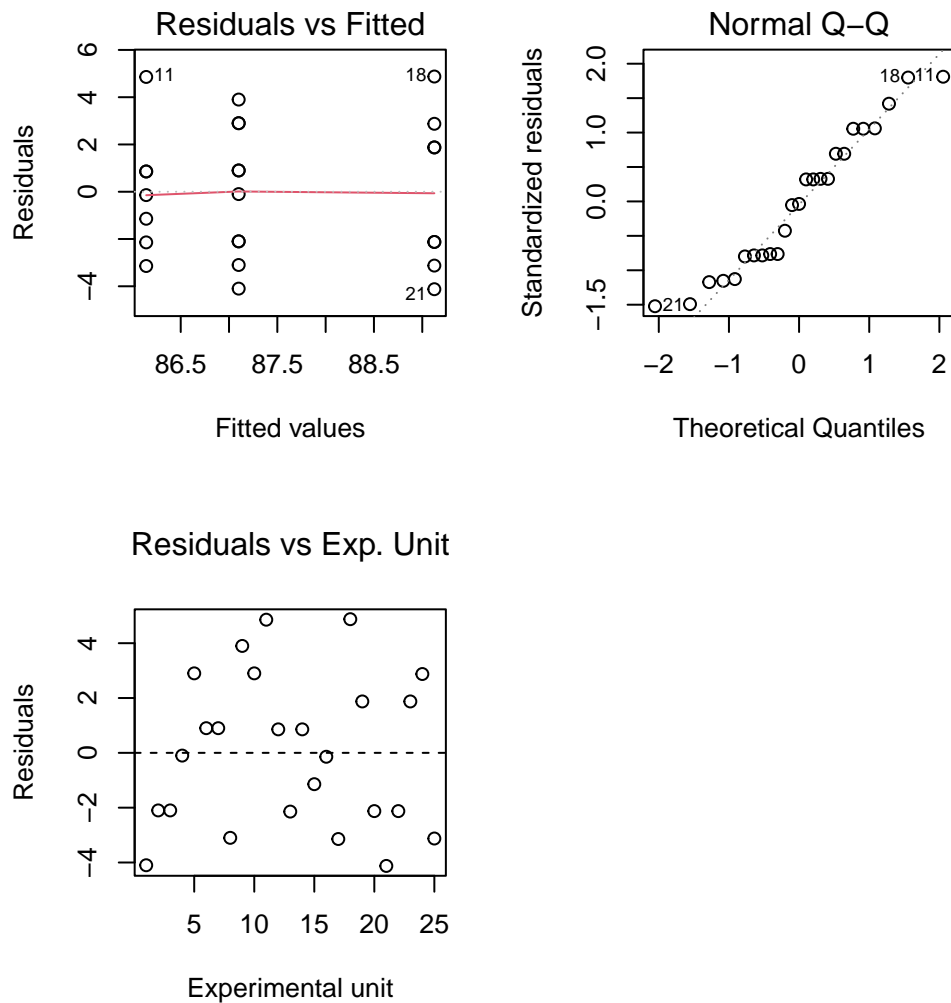
```
##           Df Sum Sq Mean Sq F value Pr(>F)
## season      2  35.61  17.804    2.121  0.144
## Residuals   22 184.63   8.392
```

a) Do the data indicate that my opinion is correct? Use alpha 0.05.

No, we don't reject the null hypothesis that the golf performance is equal, because $p\text{-value} = 0.144 > \alpha$, which means that the golf performance was the same for all the seasons.

b) Analyze the residuals from this experiment and comment on model adequacy.

```
par(mfrow = c(2, 2))
plot(fit, which = 1)
plot(fit, which = 2)
plot(residuals(fit) ~ rownames(df), main = 'Residuals vs Exp. Unit',
     font.main = 1, data = df, xlab = 'Experimental unit', ylab = 'Residuals')
abline(h = 0, lty = 2)
```



The Residual vs Fitted plot shows homoscedasticity, i.e., the variance is constant along the predicted values, and are equally distributed around zero.

The Normal Q-Q plot shows that the residuals are normally distributed because the points are very aligned with the Q-Q line.

The Residual vs Exp. Unit plot presents an constant variance across the different experimental units.

In general, the residual plots represent a good model adequacy.

2. An article in Environment International (Vol. 18, No. 4, 1992) describes an experiment in which the amount of radon released in showers was investigated. Radon-enriched water was used in the experiment, and six different orifice diameters were tested in shower heads. The data from the experiment are shown in the following table:

```
df <- data.frame(
  diameter = c(0.37, 0.51, 0.71, 1.02, 1.40, 1.99),
  rep1 = c(80, 75, 74, 67, 62, 60),
  rep2 = c(83, 75, 73, 72, 62, 61),
  rep3 = c(83, 79, 76, 74, 67, 64),
  rep4 = c(85, 79, 77, 74, 69, 66)
)
```

```
df
```

```
##   diameter rep1 rep2 rep3 rep4
## 1    0.37   80   83   83   85
## 2    0.51   75   75   79   79
## 3    0.71   74   73   76   77
## 4    1.02   67   72   74   74
## 5    1.40   62   62   67   69
## 6    1.99   60   61   64   66
```

a) Does the size of the orifice affect the mean percentage of radon released? Use alpha 0.05.

```
# reshaping
df_long <- reshape(df, direction = 'long', idvar = 'diameter', varying = list(2:5),
                  timevar = 'rep', v.names = 'radon')
df_long$diameter <- as.factor(df_long$diameter) # use factor here to obtain 5 df!
rownames(df_long) <- NULL

# one-way ANOVA
fit <- aov(radon ~ diameter, data = df_long)
summary(fit)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diameter     5 1133.4   226.68    30.85 3.16e-08 ***
## Residuals    18  132.2     7.35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, the orifice diameter affects the mean percentage of radon because the p-value = $3.16 \times 10^{-8} < \alpha$, which means we reject the null hypothesis that the mean are equal.

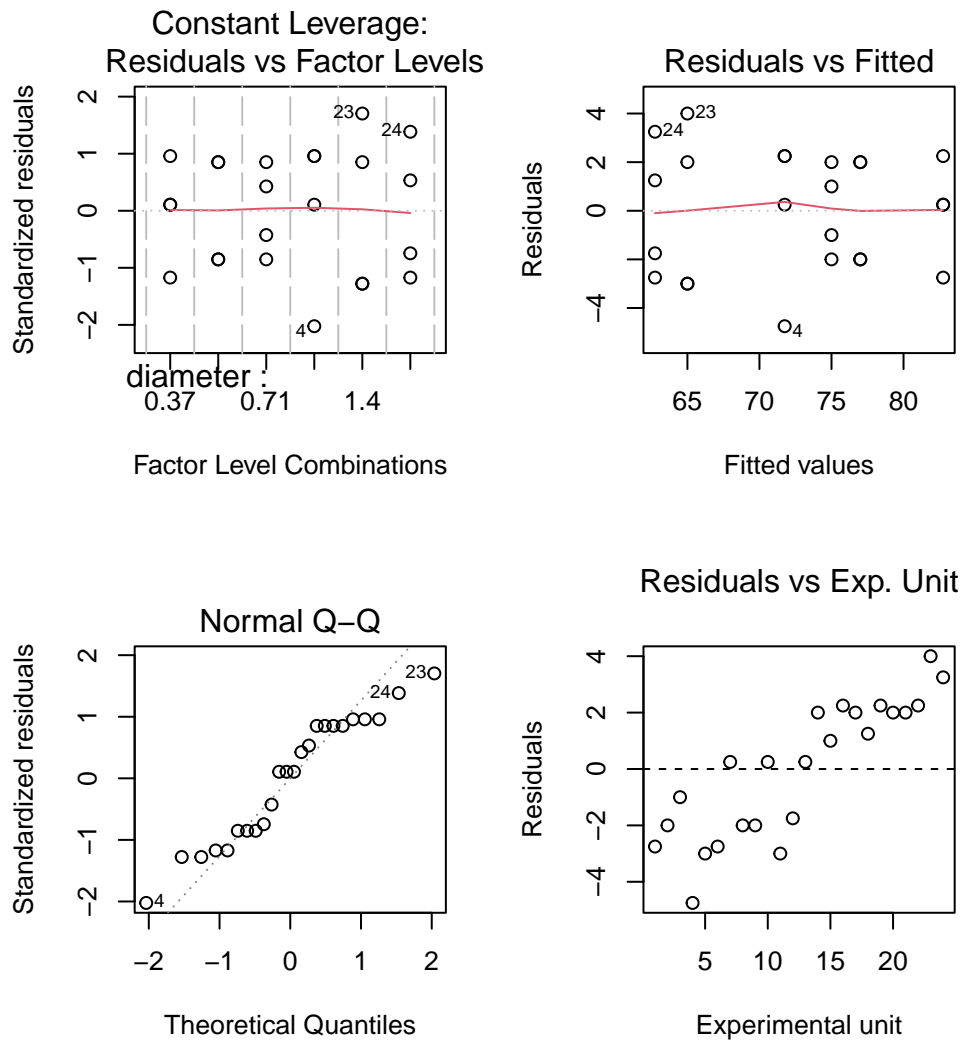
b) Find the P-value for the F statistic in part (a).

```
# it is the P(F > 30.85), where F is MS_factor / MS_error
# df1 is the DF of factor, and df2 is the DF of the errors
# lower.tail=F means we want the right side region of the quantile
pf(q = 30.85, df1 = 5, df2 = 18, lower.tail = F)
```

```
## [1] 3.160972e-08
```

c) Analyze the residuals from this experiment.

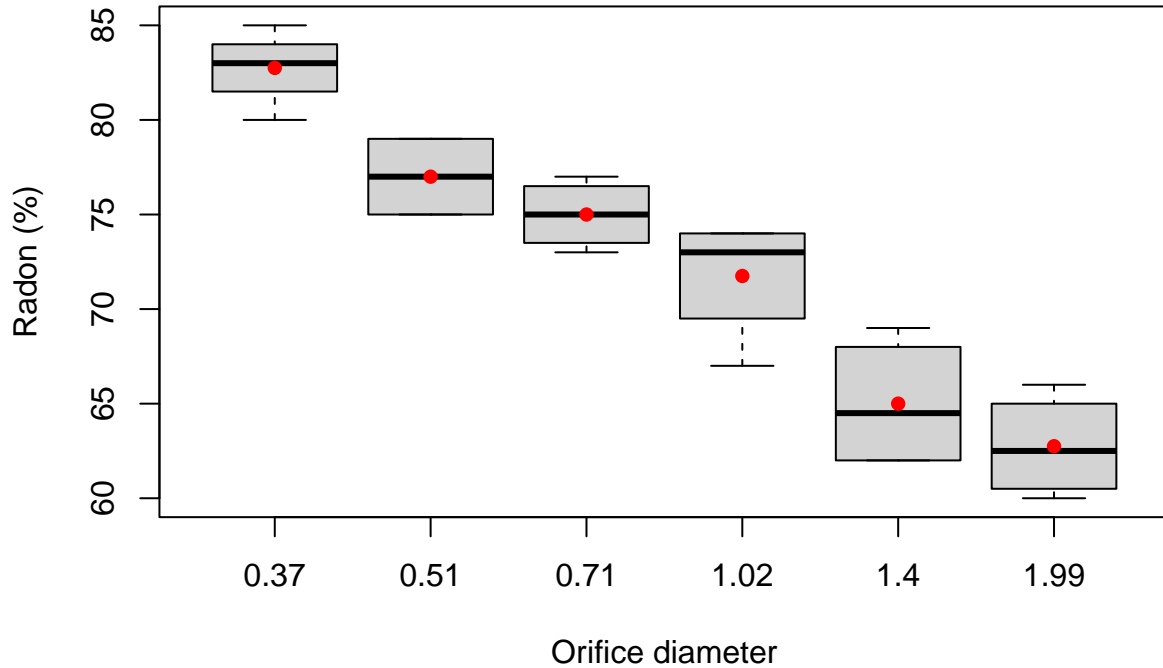
```
par(mfrow = c(2, 2))
plot(fit, which = 5)
plot(fit, which = 1)
plot(fit, which = 2)
plot(residuals(fit) ~ rownames(df_long), main = 'Residuals vs Exp. Unit',
     font.main = 1, data = df_long, xlab = 'Experimental unit', ylab = 'Residuals')
abline(h = 0, lty = 2)
```



The Residuals vs Factor Levels plot shows that some factor levels have more variability than others. The Residuals vs Fitted plot shows the residuals are somewhat homogeneous across the predicted values around around the zero. The Normal Q-Q shows a deviation from the Q-Q line, which means the normality of residuals is not met. Finally, the Residuals vs Exp. Unit plot does not show a horizontal pattern. In general, the residual plots don't show a good model adequacy.

d) Construct a graph to compare the treatment means. What conclusions can you draw?

```
boxplot(radon ~ diameter, data = df_long, xlab = 'Orifice diameter', ylab = 'Radon (%)')
means <- tapply(df_long$radon, df_long$diameter, mean)
points(means, col = 'red', pch = 16)
```



The mean percentage in radon (red points) differs from each different treatment (orifice diameter), although some treatments show more similar radon percentage distribution (e.g. 0.51 and 0.7; 1.4 and 1.99). Only looking through this data we could be suspicious that the treatment means were significant.

- Suppose that we have data on the weight loss of 100 people, each person assigned to one specific diet, each diet having assigned to it the same number of people. In performing an ANOVA, the analysts arrived at the table below, which is incomplete; Fill in the blanks.

Source of variability	SSQ	df	MSQ	Fcalc
Diet	?	3	?	15
Error	?	?	600	
Total	?	?		

We have 100 people, so the Total df is $n - 1 = 99$.

$$SSQ_E = df_E \times MSQ_E = 57600.$$

$$F_{calc} = \frac{MSQ_F}{MSQ_E}, \text{ so } MSQ_F = F_{calc} \times MSQ_E = 9000.$$

$$SSQ_F = df_F \times MSQ_F = 27000.$$

Source of variability	SSQ	df	MSQ	Fcalc
Diet	27000	3	9000	15
Error	57600	96	600	
Total	84600	99		

- Consider the data set HW2_Q5.csv, which represents the yield of soybean (in kg) grown using different potassium concentrations (in ppm). Are there significant differences due to the concentration of potassium used? Use $\alpha = 0.05$.

```
df <- read.csv('HW2_Q5.csv')
df$dose <- as.factor(df$dose)
tibble::glimpse(df)

## Rows: 24
## Columns: 2
## $ dose <fct> 0, 0, 0, 0, 0, 0, 60, 60, 60, 60, 60, 60, 120, 120, 120, 120, 12~
## $ yield <dbl> 15.7, 13.1, 13.5, 14.9, 14.4, 13.9, 19.5, 17.8, 16.7, 17.7, 18.2~

fit <- aov(yield ~ dose, data = df)
summary(fit)

##           Df Sum Sq Mean Sq F value    Pr(>F)
## dose          3  139.50    46.50    64.87 1.75e-10 ***
## Residuals    20   14.34     0.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null hypothesis that the mean yield of soybeans are equal, i.e., at least one mean yield differs, because the p-value < 0.05.

5. Using information from question 4, write the statistical model (all 3: the cell means, the treatment effect, as well as the matrix form) explaining what each term means.

Cell means model

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where Y_{ij} is the yield for the j -th experimental unit subject to the i -th dose, $i = 1, \dots, 4$, $j = 1, \dots, r_i$, and r_i is the number of experimental units or replications in the i -th dose. In the Question 4, all the doses have the same number of replications ($r_1 = r_2 = r_3 = r_4 = 6$). The μ_i is the mean within the i -th dose. The errors ϵ_{ij} are i.i.d. normally distributed with mean 0 and constant variance σ^2 .

Treatment effects model

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

where τ_i is the effect of the i -th dose and μ is the global mean. The errors ϵ_{ij} are i.i.d. normally distributed with mean 0 and constant variance σ^2 .

Matrix form model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \\ y_{26} \\ \vdots \\ y_{41} \\ y_{42} \\ y_{43} \\ y_{44} \\ y_{45} \\ y_{46} \end{pmatrix}_{24 \times 1}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}_{24 \times 5}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix}_{5 \times 1}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{15} \\ \epsilon_{16} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{25} \\ \epsilon_{26} \\ \vdots \\ \epsilon_{41} \\ \epsilon_{42} \\ \epsilon_{43} \\ \epsilon_{44} \\ \epsilon_{45} \\ \epsilon_{46} \end{pmatrix}_{24 \times 1}$$

where $\boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2 \mathbf{I})$, where $\mathbf{0}$ is a vector of zeros, and \mathbf{I} is an identity matrix.

6. In an experiment that aims to compare plant growth under the application of 3 different fertilizers, the researcher has 6 experimental units available. Assuming a CRD will be used with an equal number of replications, provide below the Linear Model in matrix form and write all matrices with their elements (see slide 5 from Week3_part1).

The experiment has 3 treatments and 6 experimental units. As the experiment used an equal number of replications, then there are 2 replications per treatment:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}_{6 \times 1}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}_{6 \times 4}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}_{4 \times 1}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}_{6 \times 1}$$

7. Design a CRD, i.e., create a table with the randomized treatments for the experiment mentioned in question 6. Include your table below and indicate how it was generated.

```
set.seed(2023)
factors <- as.factor(rep(c('a', 'b', 'c'), 2)) # 3 treatments, 2 repetitions each
factors <- sample(factors) # randomize treatments
ids <- seq(1:length(factors))
df <- data.frame(id = ids, treat = factors)
print(df, row.names = FALSE)
```

```
## id treat
## 1      b
## 2      a
## 3      a
## 4      c
```



```
##      5      b
##      6      c
# check that the number of repetitions are equal
table(df$treat)
```

```
##
## a b c
## 2 2 2
```

8. The data set HW2_Q9.csv is from an experiment aiming to evaluate the influence of yacon flour consumption on the glycemic index. Run an ANOVA, check the assumptions, and if they are not met, re-run the model with the appropriate transformation.

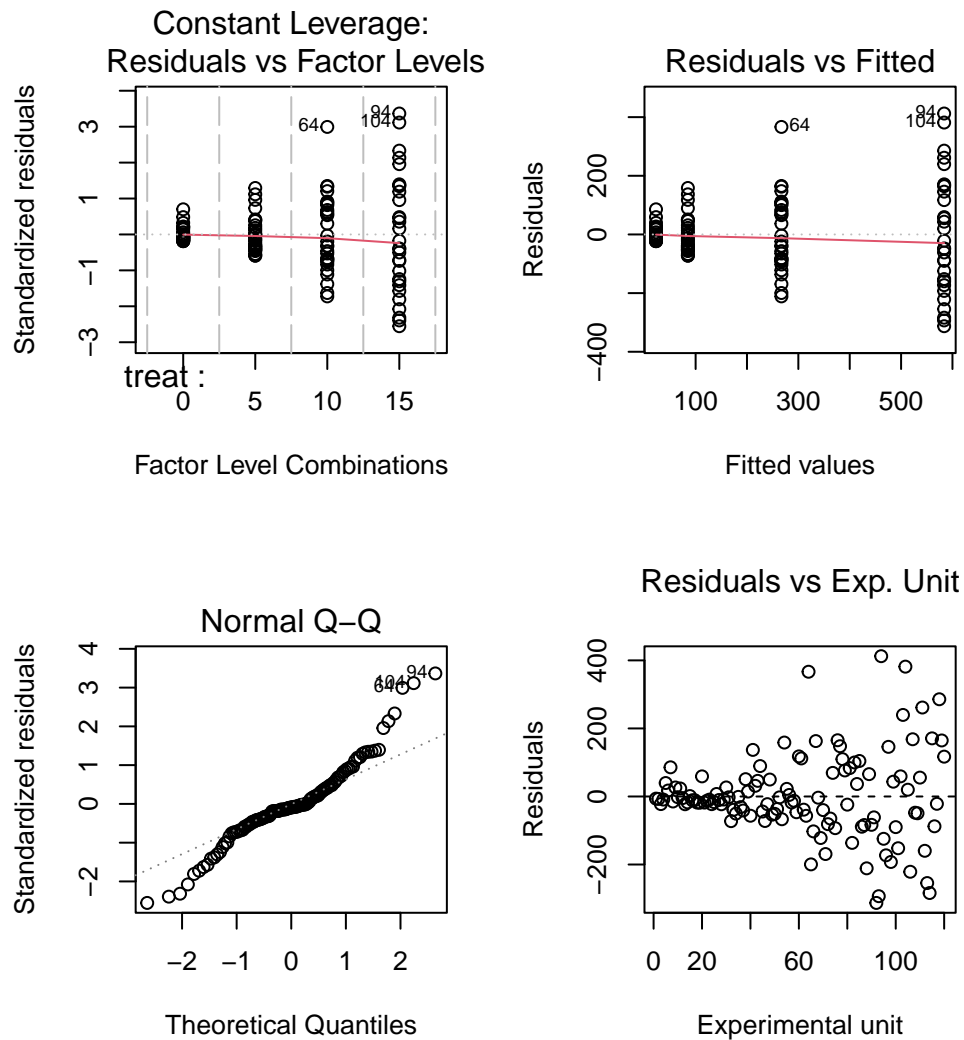
```
df <- read.csv('HW2_Q9.csv')
df$treat <- as.factor(df$treat)
str(df)
```

```
## 'data.frame': 120 obs. of 2 variables:
## $ treat: Factor w/ 4 levels "0","5","10","15": 1 1 1 1 1 1 1 1 1 1 ...
## $ y : num 16.603 15.727 0.514 13.851 62.446 ...
```

```
fit1 <- aov(y ~ treat, data = df)
summary(fit1)
```

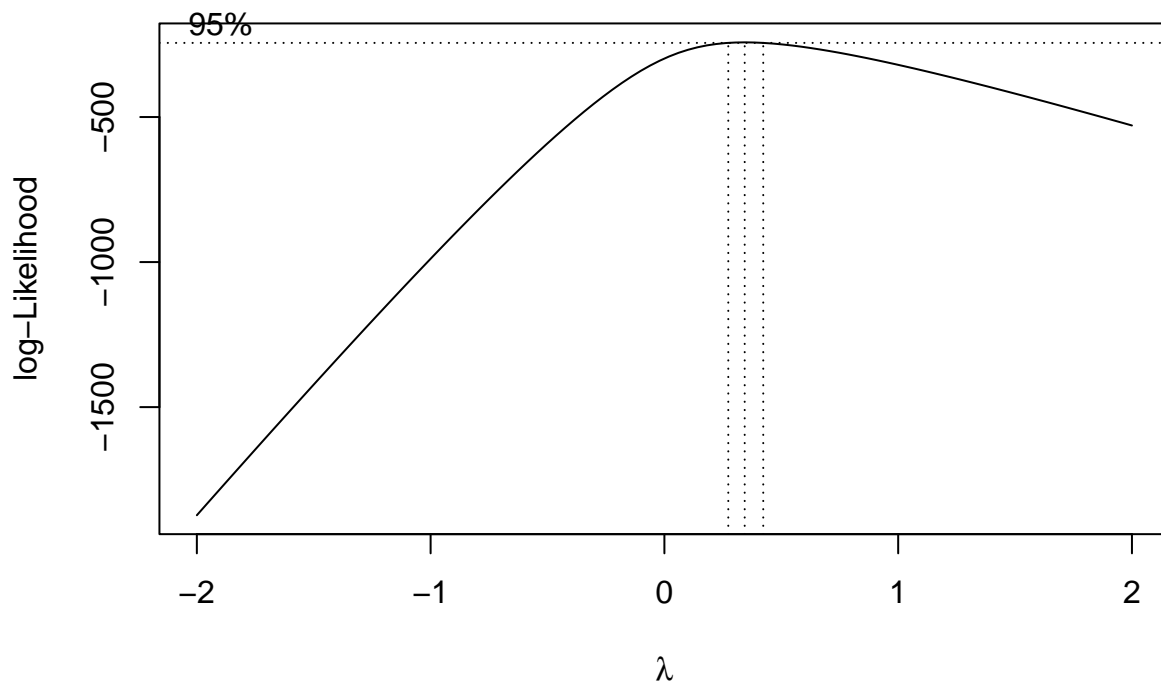
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## treat          3 5703799 1901266   122.5 <2e-16 ***
## Residuals    116 1800383   15521
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(2, 2))
plot(fit1, which = 5)
plot(fit1, which = 1)
plot(fit1, which = 2)
plot(residuals(fit1) ~ rownames(df), main = 'Residuals vs Exp. Unit',
     font.main = 1, data = df, xlab = 'Experimental unit', ylab = 'Residuals')
abline(h = 0, lty = 2)
```



The residuals variance is not homogeneous across the fitted values. Actually, the variance seems to increase as the fitted values increase. This also happens when we look at the residuals within each factor level. The normality assumption is not met. Moreover, the residuals across the experimental units does not show a horizontal pattern.

```
bc <- MASS::boxcox(fit1, data = df) # do box-cox power transformation search
```



```
lambda <- bc$x[which.max(bc$y)] # pick lambda with the highest log-likelihood
df$y_trans <- df$y ^ lambda # transform the response variable
```

```
# fit again but using the transformed response variable
fit2 <- lm(y_trans ~ treat, data = df)
anova(fit2)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: y_trans
```

```
##      Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## treat      3  672.39  224.130   171.84 < 2.2e-16 ***
```

```
## Residuals 116  151.30    1.304
```

```
## ---
```

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

```
par(mfrow = c(2, 2))
```

```
plot(fit2, which = 5)
```

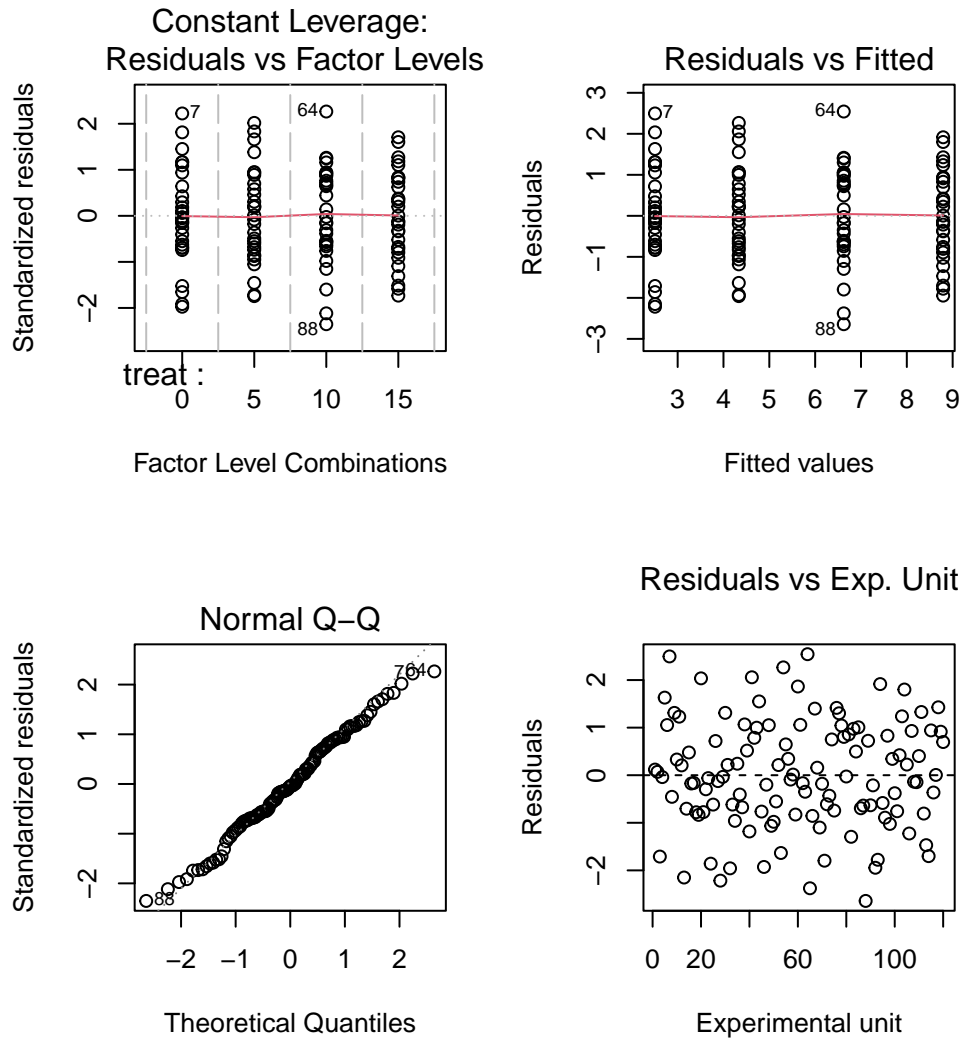
```
plot(fit2, which = 1)
```

```
plot(fit2, which = 2)
```

```
plot(residuals(fit2) ~ rownames(df), main = 'Residuals vs Exp. Unit',
```

```
      font.main = 1, data = df, xlab = 'Experimental unit', ylab = 'Residuals')
```

```
abline(h = 0, lty = 2)
```



The residual plots are way better. The homoscedasticity is present, with residuals varying constantly around 0, and the normality seems to be met (with slightly deviations). The residuals also are constant between the different factor levels, and the residuals follow a horizontal pattern when looking across the experimental units.

9. There is an “Exercises” section at the end of chapter 3. For this homework, answer question 5 of the Exercises list of chapter 3.