STAT 210
Applied Statistics and Data Analysis:
Problem list 6 - Solution
(Due on week 7)

Exercise 1

The data for this question is stored in the file `data_q1`. Read the data and store it in a file with the same name.

```r
data_q1 <- read.table('data_q1', header = TRUE)
str(data_q1)
```

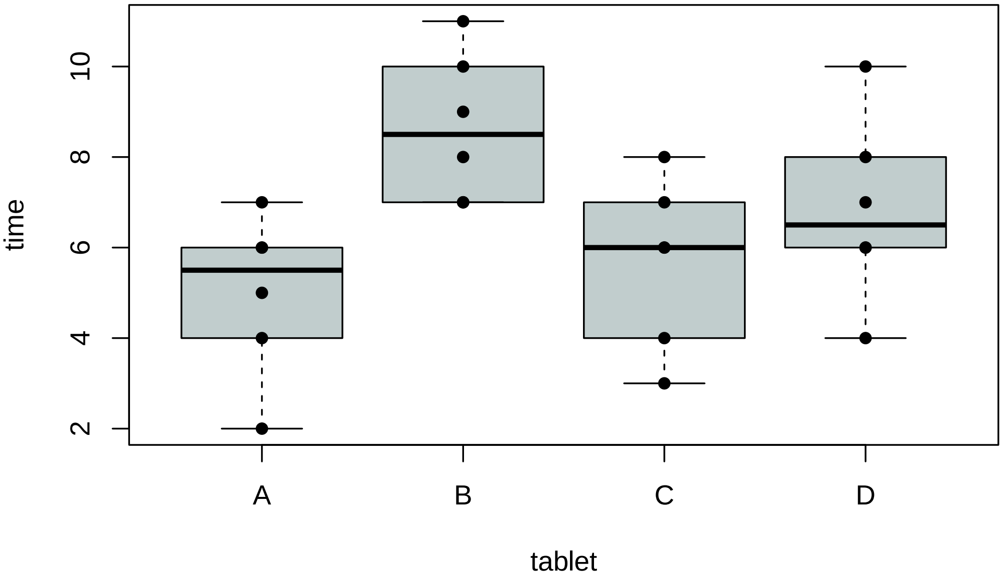
```
## 'data.frame':    24 obs. of  2 variables:
##  $ time  : int  6 2 5 4 6 7 10 8 11 7 ...
##  $ tablet: chr  "A" "A" "A" "A" ...
```

```r
data_q1$tablet <- factor(data_q1$tablet)
```

These data come from a study to compare the disintegration time (in seconds) of four types of coating for pharmaceutical tablets. The coatings are labeled `A`, `B`, `C`, and `D`.

(i) Do boxplots for the disintegration time as a function of treatment level. Comment on what you observe.

```r
boxplot(time ~ tablet, data = data_q1, col = 'azure3')
points(time ~ as.numeric(tablet), data = data_q1, pch = 16)
```



Coating `B` seems to last longer than the rest. Variances for the four levels are similar.

(ii) Do an analysis of variance to test whether the coatings have an effect on the disintegration time. Use $\alpha = 0.05$ for this test. What are your conclusions?

```
model1 <- aov(time ~ tablet, data = data_q1)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)  
## tablet       3  46.46  15.486   4.588 0.0133 *
## Residuals   20  67.50   3.375                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coatings have a significant effect on the disintegration time at the 5% level.

(iii) Find the estimated values for the average disintegration time for the four treatment levels. Find also the effects, and include the standard errors in each case.

We use the `model.tables` function with the option `se = TRUE` to get the standard errors.

```
model.tables(model1, se = TRUE)
```

```
## Tables of effects
## 
##  tablet 
## tablet
##       A       B       C       D 
## -1.5417  2.1250 -0.8750  0.2917 
## 
## Standard errors of effects
##         tablet
##           0.75
## replic.      6
```

```
model.tables(model1, 'means', se = TRUE)
```

```
## Tables of means
## Grand mean
##          
## 6.541667 
## 
##  tablet 
## tablet
##     A     B     C     D 
## 5.000 8.667 5.667 6.833 
## 
## Standard errors for differences of means
##         tablet
##          1.061
## replic.      6
```

(iv) Find the estimated value for the variance and the standard deviation.

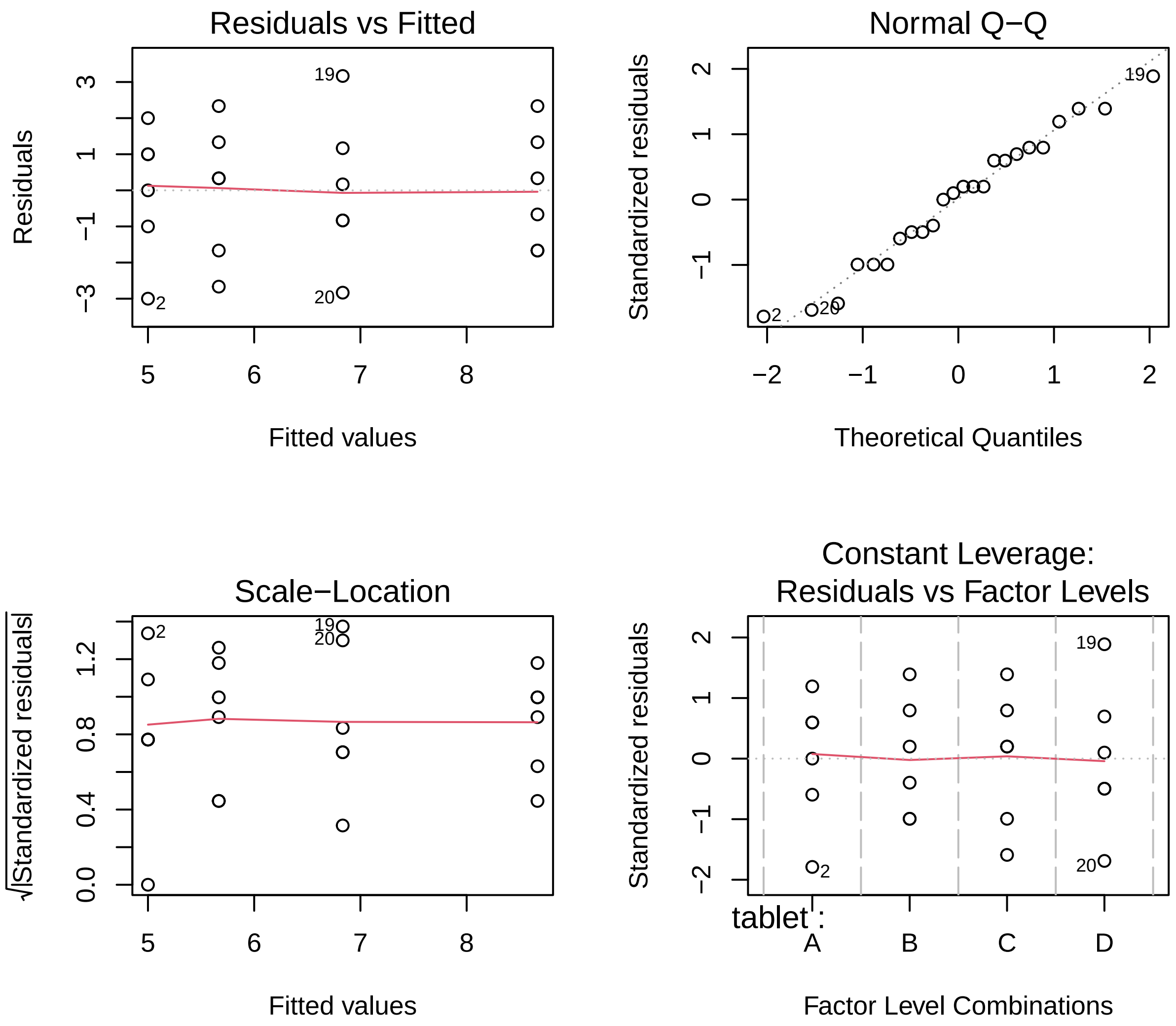
The estimated value for the variance is 3.375, obtained from the anova table. The standard deviation is

```
sqrt(3.37)
```

```
## [1] 1.835756
```

(v) Draw diagnostic plots for checking the assumptions and comment on the results.

```
par(mfrow = c(2,2))
plot(model1)
```



```
par(mfrow=c(1,1))
```

In the first plot, the red line indicates that the residuals are centered at 0 and are approximately symmetric. All the groups pf residuals seem to have the same variability. On the second plot, the points are well aligned and the graph supports the assumption of normality. In the third plot , all points are below 1.5 and the red line is horizontal, indicating that the variances are simmilar. Summing up, there are no signs of heteroscedasticity or asymmetry in the residual plots, and the normality assumption seems to be well satisfied.

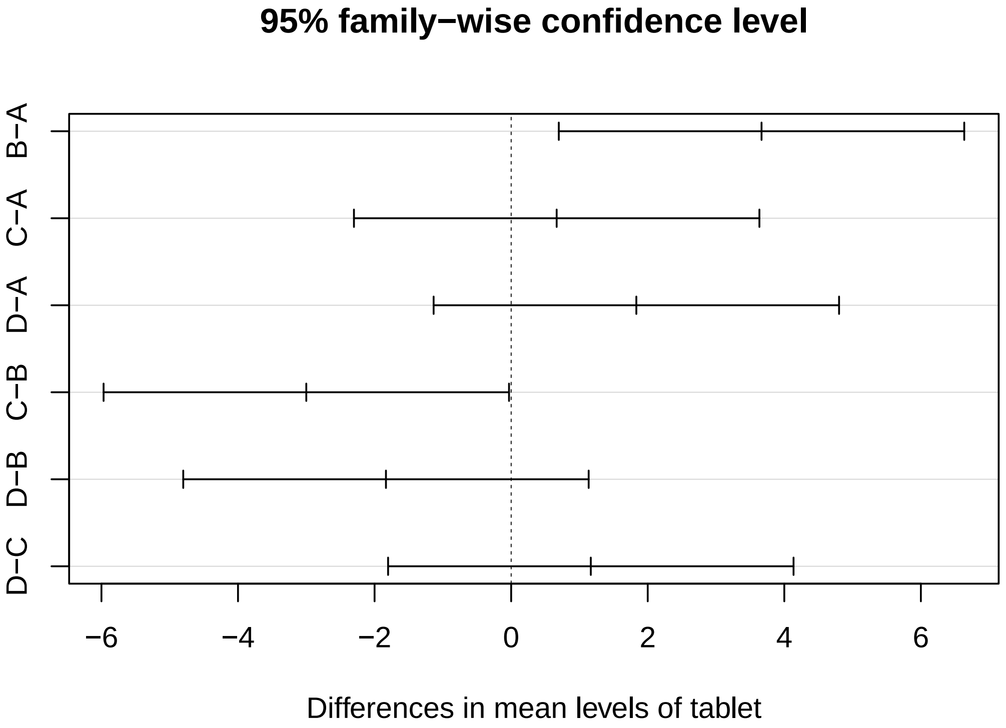
(vi) Do pairwise comparisons using Tukey's Honest Significant Difference method. Plot the confidence intervals. What differences are significant according to this method?

For this we use the function `TukeyHSD`:

```
TukeyHSD(model1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = time ~ tablet, data = data_q1)
##
## $tablet
##           diff        lwr        upr     p adj
## B-A  3.6666667  0.6979466  6.6353868 0.0122461
## C-A  0.6666667 -2.3020534  3.6353868 0.9216485
## D-A  1.8333333 -1.1353868  4.8020534 0.3358002
## C-B -3.0000000 -5.9687201 -0.0312799 0.0470689
## D-B -1.8333333 -4.8020534  1.1353868 0.3358002
## D-C  1.1666667 -1.8020534  4.1353868 0.6936360
```

```
plot(TukeyHSD(model1))
```



The only significant differences at the 5% level are B and A, and B and C. In the graph we see that the only confidence intervals that do not intersect the vertical line that passes through zero correspond to the comparisons A and B, and A and C.

(vii) What nonparametric procedure could be used in this situation? Carry out this test and compare it with your previous result.

We can use the Kruskal-Wallis test.

```
kruskal.test(time ~ tablet, data = data_q1)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  time by tablet
## Kruskal-Wallis chi-squared = 9.9382, df = 3, p-value = 0.0191
```

The *p*-value is similar and we would reach the same conclusion with this test.

(viii) Assuming that longer times to disintegration are desirable, what would you recommend to the tablet manufacturer?

B is better than A and C, but cannot be distinguished from D on the basis of this sample, so the choice is between B and D. Either repeat the experiment with these two, perhaps with a different number of replications, or if the difference between them has no practical significance, choose the cheapest.

Exercise 2

The data in the file sulfa.txt has the results of an experiment to study the effect of sulfamerazine (Sulfa) on the amount of hemoglobin (Hemo) in trouts. The trouts were placed at random in four different containers, and the fish food added contained, respectively, 0, 5, 10, and 15 grams of sulfamerazine per 100 pounds of fish (coded 1, 2, 3, and 4). The measurements were made on ten randomly selected fish from each container after 35 days.

a) Read the data file into a data frame named q3.df. Make sure the data are read correctly. Add a factor fSulfa with the information in the variable Sulfa.

(b) Do boxplots for the Hemo as a function of Sulfa (all the boxplots should appear on the same panel). Add the points to this graph. Comment on what you observe.

(c) Fit an analysis of variance model to this data. Use $\alpha = 0.02$ for your test. What do you conclude from this analysis?

(d) Find the estimate for the mean response for each treatment. Find also the effects, and include the standard errors in each case. What are the estimated values for the variance and standard deviation in this experiment?

(e) What are the assumption on which the analysis of variance model is based? Draw diagnostic plots for checking these assumptions and discuss the results.

Solution

(a) Read the data

```
q3.df <- read.table('sulfa.txt', header = T)
str(q3.df)
```

```
## 'data.frame':    40 obs. of  2 variables:
##  $ Sulfa: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Hemo : num  6.7 7.8 5.5 8.4 7 7.8 8.6 7.4 5.8 7 ...
```

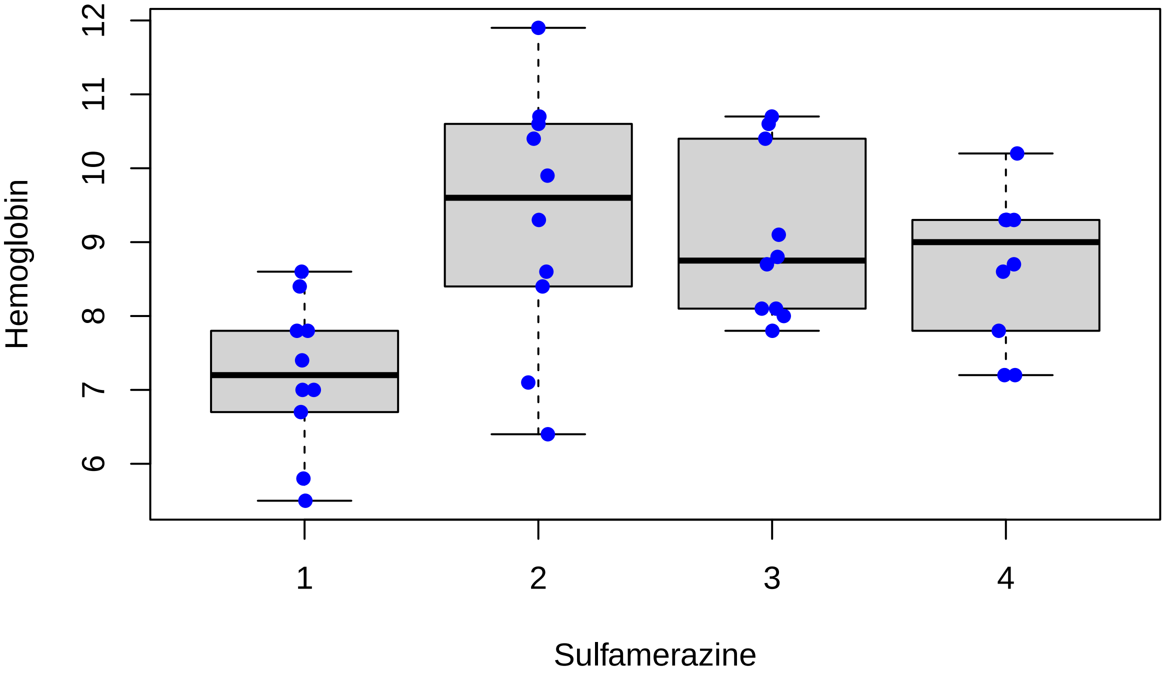
Add a factor

```
q3.df$fSulfa <- factor(q3.df$Sulfa)
str(q3.df)
```

```
## 'data.frame':    40 obs. of  3 variables:
##  $ Sulfa : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Hemo  : num  6.7 7.8 5.5 8.4 7 7.8 8.6 7.4 5.8 7 ...
##  $ fSulfa: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
```

(b) Boxplots:

```
plot(Hemo ~ fSulfa, data = q3.df, xlab='Sulfamerazine', ylab = 'Hemoglobin')
points(Hemo ~ jitter(Sulfa, amount = 0.05), data = q3.df, pch = 16, col = 'blue')
```



We observe that there is an initial increment in the amount of hemogoblin, when we add a small amount of `Sulfa`, but increasing the dosage has the effect of reducing the amount of hemogoblin. Since the amount of `Sulfa` increases uniformly in steps of 5 grams per 100 pounds of fish, we can also interpret the x-axis as a numerical variable and observe that there seems to be a quadratic relation between the variables. Also, the size of the boxes show some variability. This is something we will have to check with the diagnostic plots.

(c) Anova model:

```
model1 <- aov(Hemo ~ fSulfa, data = q3.df)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## fSulfa       3  26.80   8.934   5.696 0.00268 **
## Residuals   36  56.47   1.569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is small, so we conclude that there is a difference in the treatments.

(d) Mean responses with standard errors

```
(means <- model.tables(model1, 'means', se = TRUE))
```

```
## Tables of means
## Grand mean
##
## 8.5625
##
##  fSulfa
## fSulfa
##    1    2    3    4
## 7.20 9.33 9.03 8.69
##
## Standard errors for differences of means
##         fSulfa
##          0.5601
## replic.     10
```

Effects with standard errors:

```
model.tables(model1, se = TRUE)
```

```
## Tables of effects
##
##  fSulfa
## fSulfa
##       1       2       3       4
## -1.3625  0.7675  0.4675  0.1275
##
## Standard errors of effects
##          fSulfa
##          0.3961
## replic.     10
```

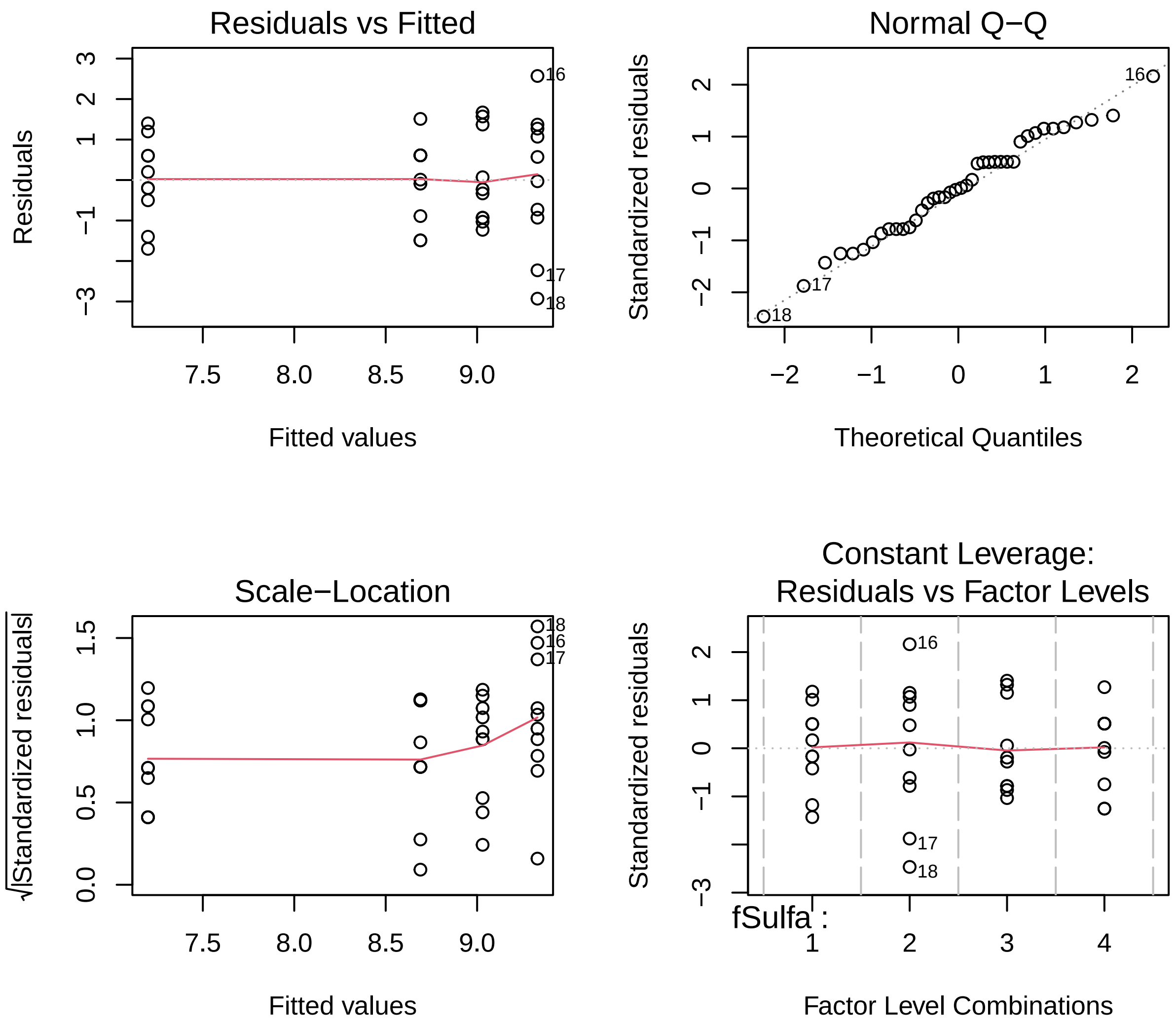
The estimated variance comes from the anova table and is 1.569. the standard deviation is

```
Va = 1.569
sqrt(Va)
```

```
## [1] 1.252597
```

(e) The model is based on the assumption that the experimental errors are independent, normally distributed random variables with mean zero and equal variance. We use the diagnostic plots to check the assumptions.

```
par(mfrow = c(2,2))
plot(model1)
```



```r
par(mfrow=c(1,1))
```

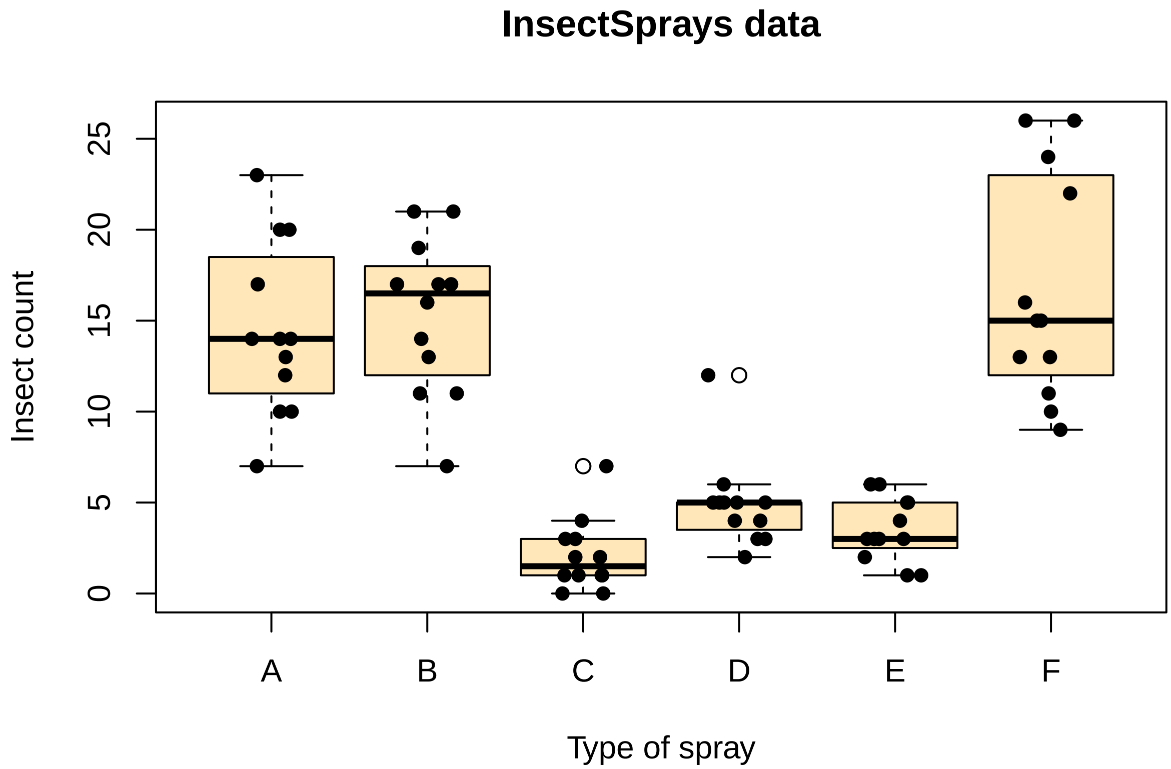
The normal quantile plot is very good, and shows that the residuals follow a normal distribution. The residuals vs. fitted values shows that the residuals for one of the treatment levels have a wider spread than the other levels, but the difference does not seem to be important. Three points, 16, 17, and 18, are singled out in the graphs as having largest residuals, and they all belong to the same treatment level. The scale-location plot shows an increase in the average value towards the largest fitted values, but the increase is moderate. Taking into account all the graphs, it seems reasonable to conclude that the assumptions on which the model is based are satisfied.

Exercise 3

For this exercise we use the dataset `InsectSprays`, which is available in `R`. In this experiment, 6 different insecticides were used and the number of dead insects in each plot were counted. There were 12 replications for each treatment level (insecticide), for a total of 72 observations.

(i) Draw a boxplot for the results and add axes labels and a title. Add the points for each treatment level. Observe that there is overplotting. Add some noise in the horizontal direction to avoid this problem. Comment on what you observe.

```r
attach(InsectSprays)
boxplot(count ~ spray, data = InsectSprays,
        xlab = "Type of spray", ylab = "Insect count",
        main = "InsectSprays data", col = 'wheat1')
points(count ~ jitter(as.numeric(spray)), data = InsectSprays, pch = 16)
```



(ii) Do an analysis of variance and test whether the different insecticides have an effect. Use level $\alpha = 0.01$ for this test. What are your conclusions?

```
fm1 <- aov(count ~ spray, data = InsectSprays)
summary(fm1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)    
## spray        5   2669   533.8    34.7 <2e-16 ***
## Residuals   66   1015    15.4                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the test of no treatment effect is practically zero, so we reject the null hypothesis.

(iii) What are the estimated values for the average values for the six treatments $\hat{\mu} + \hat{\tau}_i, i = 1, \ldots, 6$? What are the estimated values for the effects $\hat{\tau}_i, i = 1, \ldots, 6$?

We use `model.tables` for this. The average values are

```
model.tables(fm1,'means')
```

```
## Tables of means
## Grand mean
##     
## 9.5 
## 
##  spray 
## spray
##      A      B      C      D      E      F
## 14.500 15.333  2.083  4.917  3.500 16.667
```

and the effects are

```
model.tables(fm1)
```

```
## Tables of effects
##
##  spray
## spray
##      A      B      C      D      E      F
##  5.000  5.833 -7.417 -4.583 -6.000  7.167
```

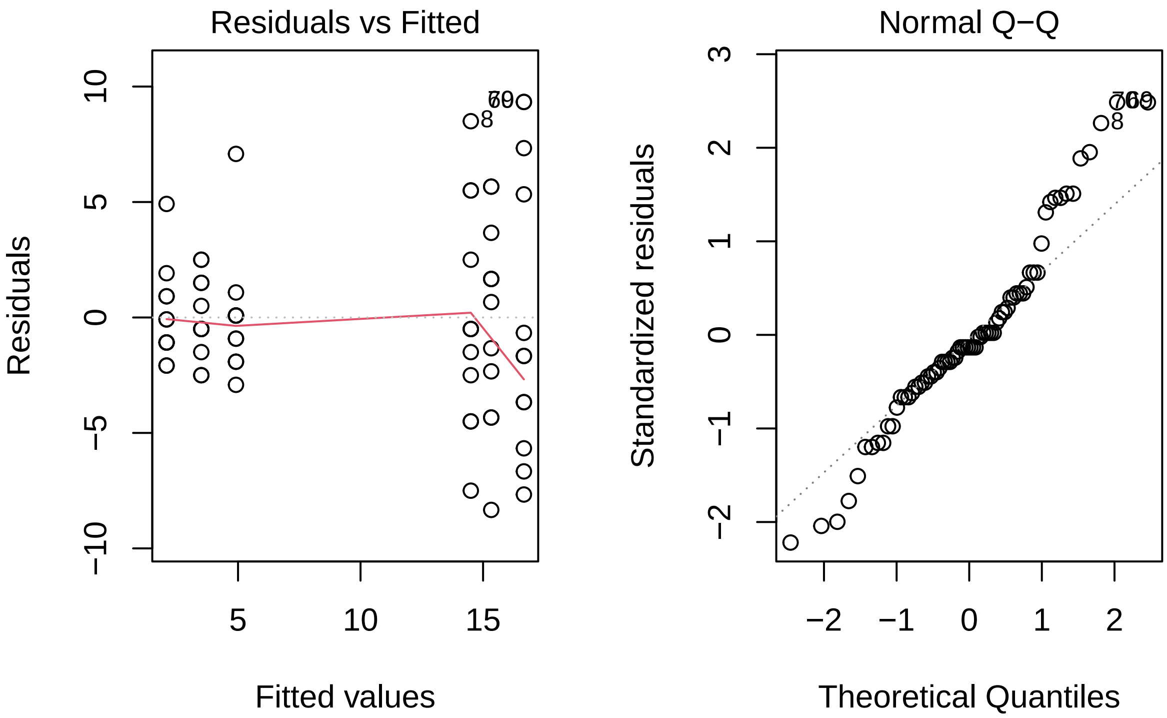
(iv) What is the estimated value for the variance of the experimental error σ^2? The estimated variance for the experiments error is found in the anova table as the MSE for residuals, which in this case is 15.4. The standard deviation is

```
sqrt(15.4)
```

```
## [1] 3.924283
```

(v) Make two residual plots for checking assumptions: residuals vs. fitted values and normal quantile plots. Do you think the usual assumptions for the model are reasonable in this experiment?

```
par(mfrow = c(1,2))
plot(fm1, which = c(1,2))
```



```
par(mfrow = c(1,1))
```

In this case the plots do not look good. The first one shows that variance increases with fitted value: The points on the right of the graph, which correspond to larger fitted values, have a wider spread than the points on the left of the graph. On the other hand, the qq normal plot shows a good fit at the center of the sample, but both tails are from the straight line.

(vi) Consider an alternative model using the square root of the number of counts. Obtain the analysis of variance table and compare with the previous model

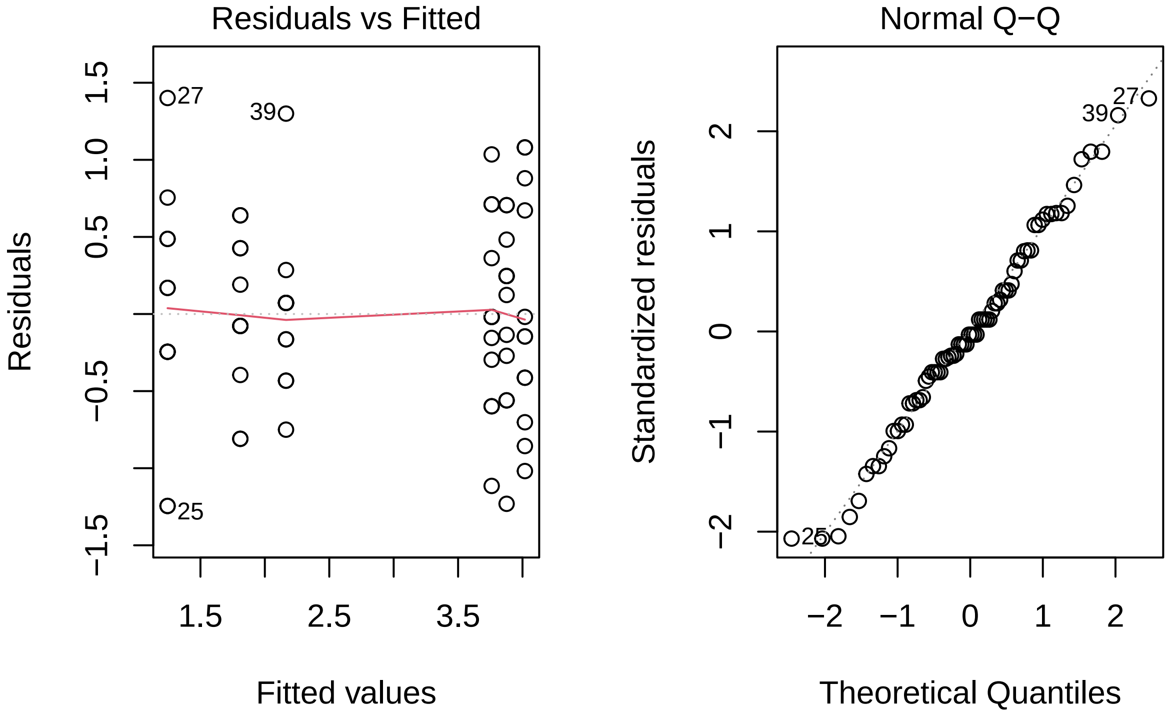
```
fm2 <- aov(sqrt(count) ~ spray, data = InsectSprays)
summary(fm2)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)    
## spray        5  88.44  17.688    44.8 <2e-16 ***
## Residuals   66  26.06   0.395                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, the p-value for the overall test is practically zero. Observe the reduction in MSE, which is the estimated error variance.

(vii) Finally, draw the same diagnostic plots for this model and comment.

```
par(mfrow = c(1,2))
plot(fm2, which = c(1,2))
```



```
par(mfrow = c(1,1))
```

Both plots look much better now. The first shows a more homogeneous spread of points for all fitted values, and in the second the fit is very good, so the hypothesis of normality seems to be valid now.

Exercise 4

A product developer is investigating the tensile strength of a new synthetic fiber that will be used to make cloth for men's shirts. Strength is usually affected by the percentage of cotton used in the blend of materials for the fiber. The engineer conducts an experiment with five levels of cotton content and replicates the experiment five times. The data are given below.

```
strength <- c(7, 7, 15, 11, 9, 12, 17, 12, 18, 18, 14, 19, 19, 18, 18, 19,
              25, 22, 19, 23, 7, 10, 11, 15, 11)
```

Table 1: Exercise 3: Tensile strength of fibers

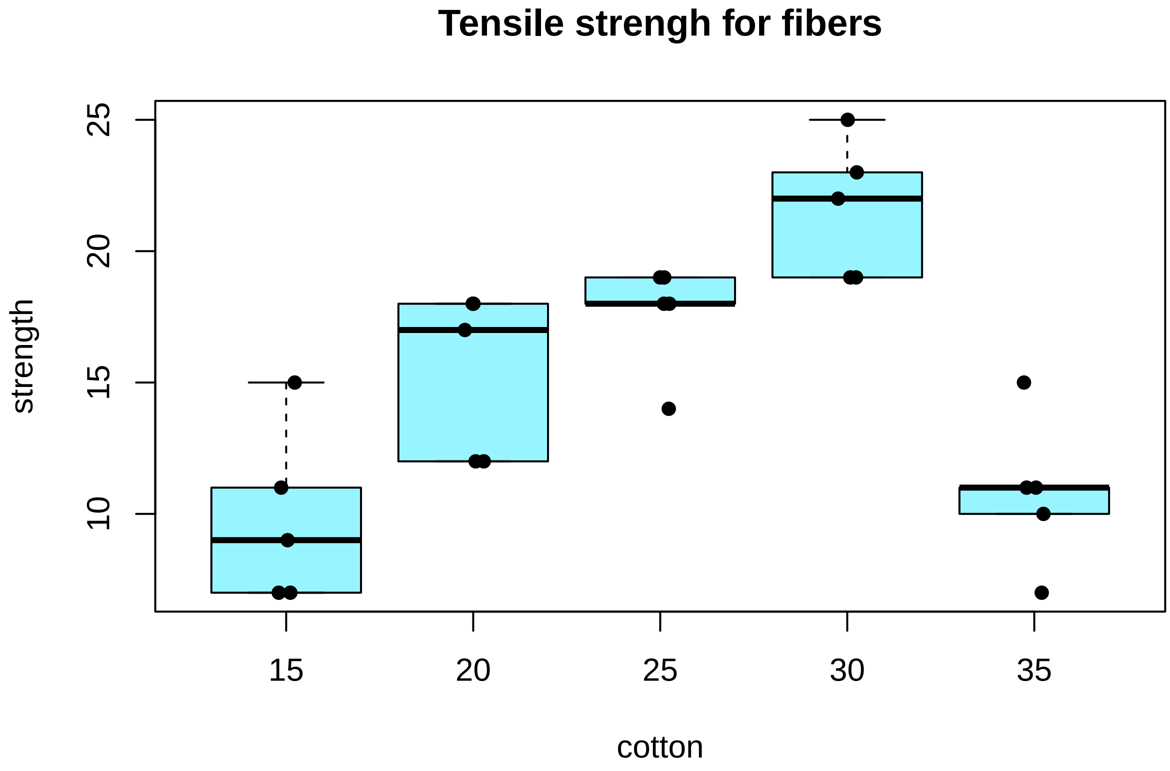
Cotton %	Strength				
15	7	7	15	11	9
20	12	17	12	18	18
25	14	19	19	18	18
30	19	25	22	19	23
35	7	10	11	15	11

```
cotton <- factor(rep(seq(15,35,5), each = 5))
q4.df <- data.frame(cotton, strength)
```

Observe that even though cotton is numerical (the values represent the percentage of cotton in the mixture) it is included in the data frame as a factor. This is necessary to fit an anova model.

(i) Draw a boxplot for the results and add axes labels and a title. Add the points for each treatment level.Observe that there is overplotting. Add some noise in the horizontal direction to avoid this problem. Comment on what you observe.

```
boxplot(strength ~ cotton, data = q3.df, col = 'cadetblue1', main = 'Tensile strengh for fibers', outli
points(strength ~ jitter(as.numeric(cotton), factor = 0.3), data = q3.df, pch = 16)
```



The strength increases with cottoncontent until the percentage is 30% and then it decreses sharply when we move to 35%. This suggests that there may be a quadratic relation between these two variables. The boxes in the plot are of different sizes, suggesting that variances may not be similar among the different groups, but adding the points to the plot sshows that this is probably an effect of the small data sizes, since the spread of the clouds of points are similar.

(ii) Is there evidence to support the claim that cotton content affects the mean tensile strength? Use $\alpha = 0.01$.

To answer this we need to fit an anova model.

```
model3 <- aov(strength ~ cotton, data = q3.df)
summary(model3)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## cotton       4  475.8  118.94   14.76 9.13e-06 ***
## Residuals   20  161.2    8.06                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is very small, so we reject the null hypothesis of no treatment effect. At least one of the differences between the groups is significant.

(iii) What are the estimated values for the average values for the treatments. What are the estimated values for the effects? Include the estimated standard errors in your tables.

The estimated values for the averages are given by

```
model.tables(model3,'means',se = TRUE)
```

```
## Tables of means
## Grand mean
##       
## 15.04 
## 
##  cotton 
## cotton
##   15   20   25   30   35 
##  9.8 15.4 17.6 21.6 10.8 
## 
## Standard errors for differences of means
##         cotton
##          1.796
## replic.      5
```

while for the effects we have

```
model.tables(model3, se = TRUE)
```

```
## Tables of effects
## 
##  cotton 
## cotton
##    15    20    25    30    35 
## -5.24  0.36  2.56  6.56 -4.24 
## 
## Standard errors of effects
##         cotton
##           1.27
## replic.      5
```

(iv) What are the estimated variance and standard deviation of the experimental error?

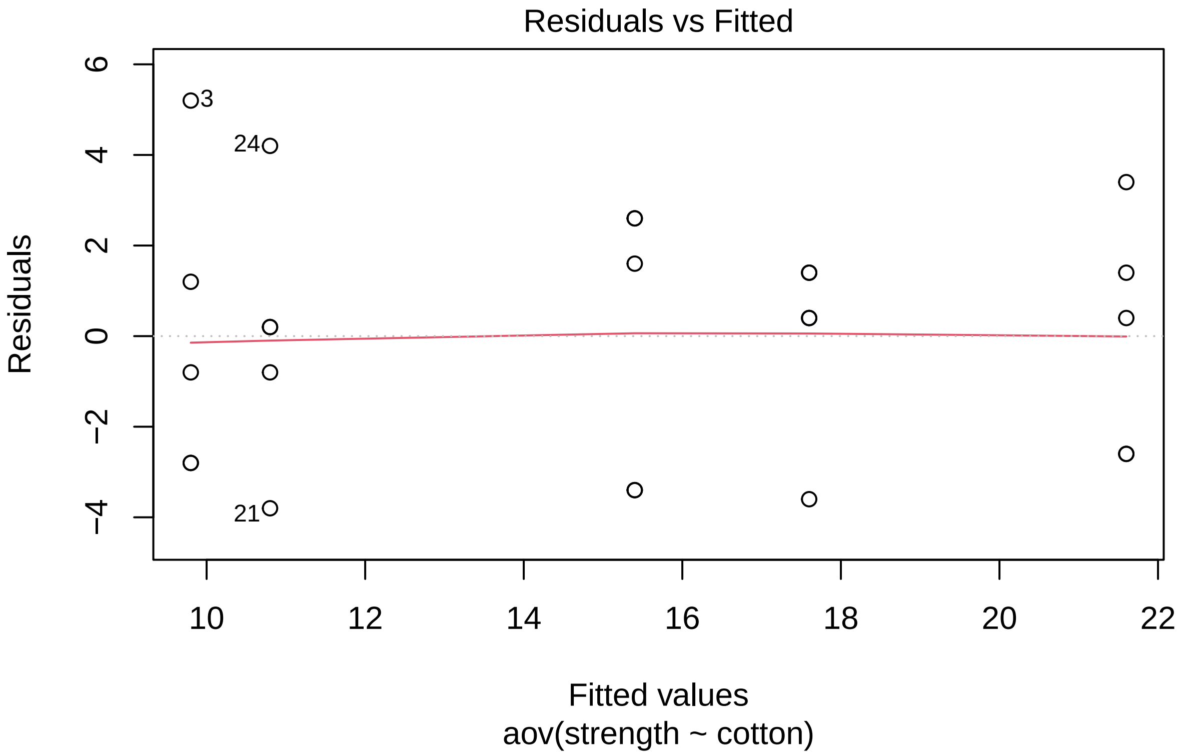
From the anova table we have that estimated variance is 8.06, and therefore, the estimated standard deviation is

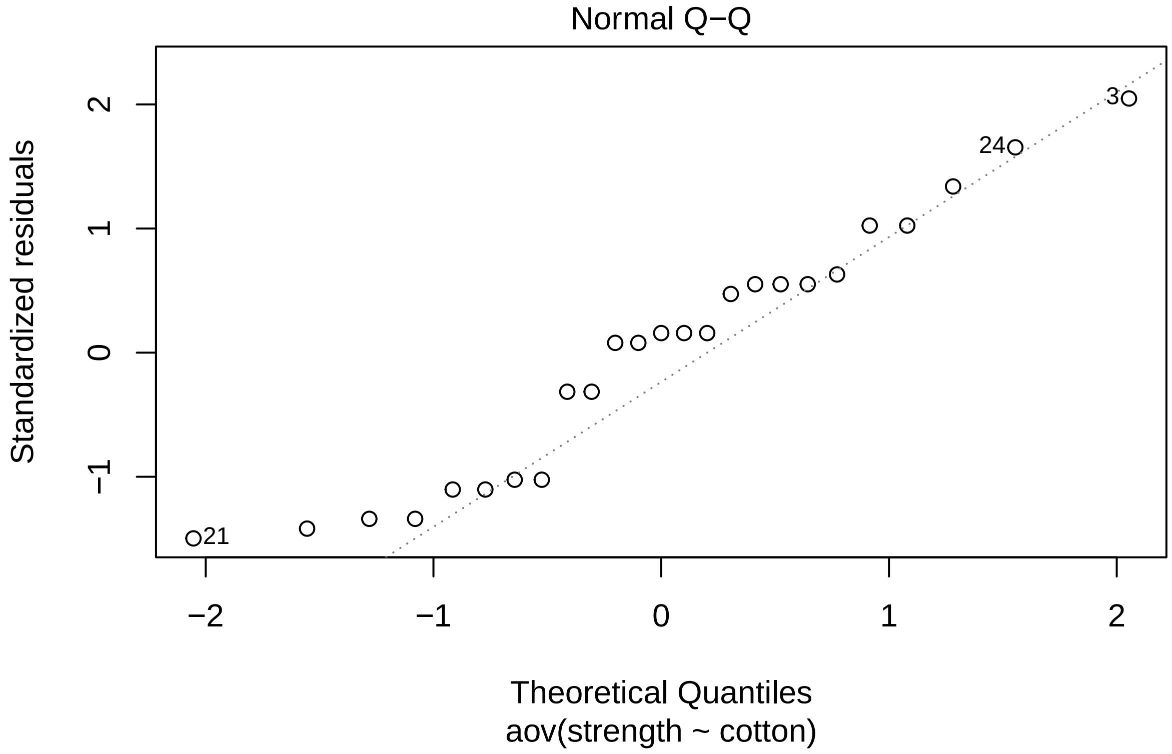
```
sqrt(8.06)
```

```
## [1] 2.839014
```

(v) Make two residual plots for checking assumptions: residuals vs. fitted values and normal quantile plots. Do you think the usual assumptions for the model are reasonable in this experiment?

```r
plot(model3, which = c(1,2))
```


Residuals vs Fitted
Residuals
Fitted values
aov(strength ~ cotton)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
aov(strength ~ cotton)