Solution to Exercise 1

1. Read in the data:

```
blood <-c(62,60,63,59,63,67,71,64,65,66,68,66,71,67,68,68,56,62,60,61,63,64,63,59)
tr <- c(1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,4,4,4,4)
b.data <- data.frame(cbind(blood,tr))
b.data$tr <- as.factor(b.data$tr)
```

a) Plot the data with:
`plot(b.data$tr,b.data$blood)`
We see that the coagulation times vary a lot between different diets whereas the variation within a diet group is quite small.
In addition compute the overall mean and the group means. Do this by hand using a calculator.

overall mean = 64

treatment	group means
A	61
B	66
C	68
D	61

b) Compute the group sample variances s_i^2 and the pooled estimate of variance MS_{res}. Do this also by hand. For MS_{res} compute first SS_{res}.

$SS_{res} = 112$ $MS_{res} = 5.6$

treatment	s_i^2
A	3.333
B	8
C	2.8
D	6.85

c) Compute MS_{treat} and compare it to MS_{res}. Compute MS_{treat} by hand. First compute SS_{treat} and with it MS_{treat}.

$SS_{treat} = 228$ $MS_{treat} = 76$

We see that the estimated variance between groups is substantially bigger then the estimated variance within groups. This could indicate an effect of diet on blood coagulation time.

d) Use the R-function `aov(....)`.

```
fit.blood <- aov(b.data$blood ~ b.data$tr)
summary(fit.blood)
            Df Sum Sq Mean Sq F value    Pr(>F)
b.dat$tr     3    228    76.0  13.571 4.658e-05 ***
Residuals   20    112     5.6
```

Compare your by hand computed SS_{res}, SS_{treat}, MS_{res} and MS_{treat} with the output of `summary(fit.blood)`.

e) From the output above we see that the diet has an significant effect on blood coagulation time.

F-value = 13.571
P-value = $4.658 \cdot 10^{-5}$

2. a) The parameters in the one-way analysis of variance model $Y_{ij} = \mu + A_i + \epsilon_{ij}$ with $\sum A_i = 0$ are:
$\mu = 7.2,\ A_1 = -2.1, A_2 = -0.9, A_3 = 0.7, A_4 = 2.3$ and $\sigma^2 = 2.8^2$.

b) $E(MS_{res}) = \sigma^2 = 7.84$
$E(MS_{treat}) = \sigma^2 + 25 \cdot \frac{\sum_{i=1}^{4} A_i^2}{3} = 7.84 + 25 \cdot 3.666 = 99.5066$

Therefore we can conclude that the duration of employment has an effect on the job satisfaction. Because $E(MS_{treat})$ is way larger then $E(MS_{res})$.

3. Read in the data

```
N2 <- c(19.4,32.6,27,32.1,33,18.2,24.6,25.5,19.4,21.7,20.8,20.7,
        21,20.5,18.8,18.6,20.1,21.3)
strain <- c(1,1,1,1,1,5,5,5,5,5,5,7,7,7,7,7,7,7)
r.data <- data.frame(cbind(N2,strain))
r.data$strain <- as.factor(b.data$strain)
```

a) Plot the data:
`plot(r.data$strain,r.data$N2)`
The variance between strains looks larger then the variance within strains. This could be an indicator for a significant difference of nitrogen contents for different Rhizobium strains.

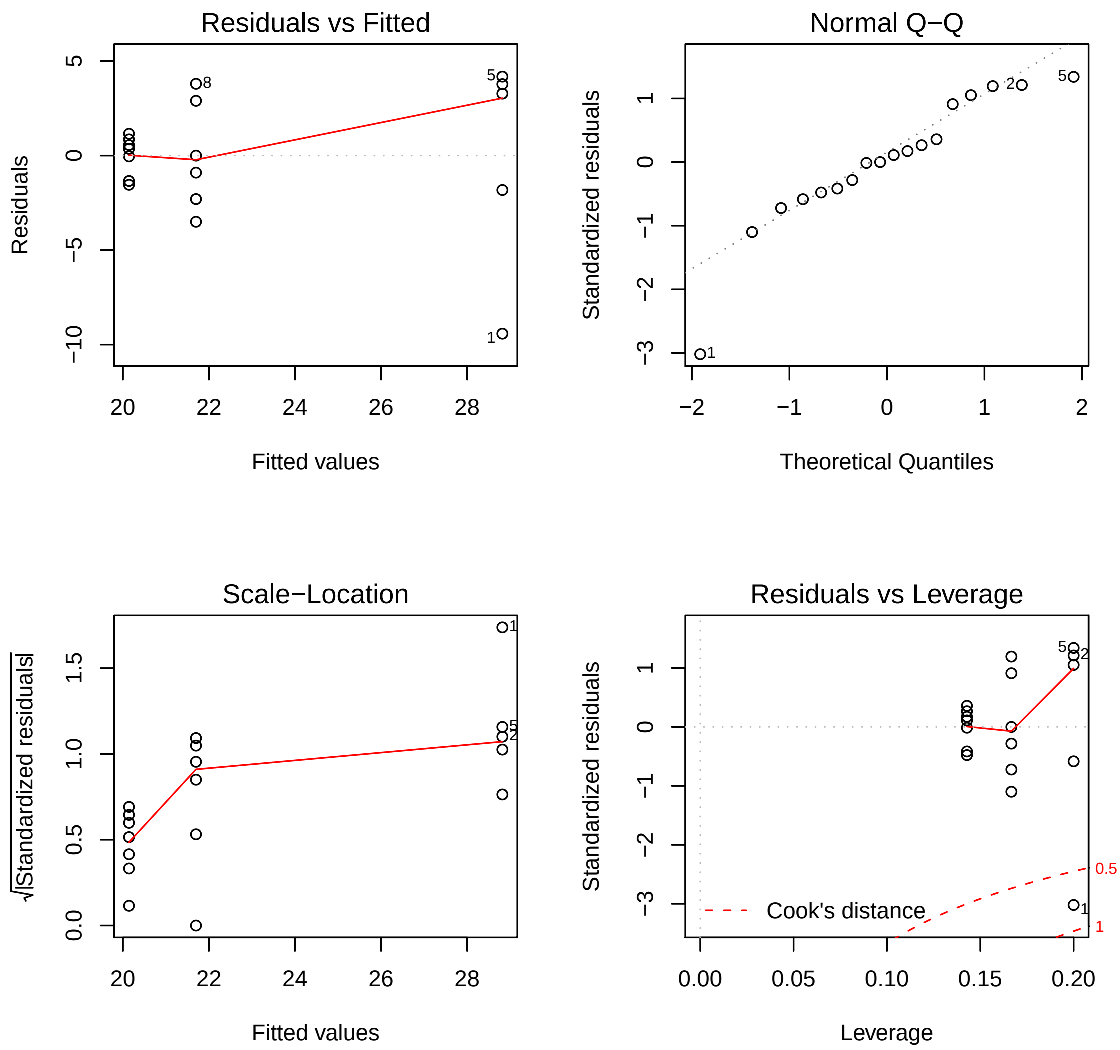
b) Carry out an analysis of variance:

```
fit.n2 <- aov(r.data$N2 ~ r.data$strain)
summary(fit.n2)
              Df Sum Sq Mean Sq F value   Pr(>F)
r.data$strain  2 236.55 118.275  9.7231 0.001959 **
Residuals     15 182.47  12.164
```

The F-value equals 9.7231. By looking at the P-value (= 0.00195) we see that there are significant differences in nitrogen contents for different strains of Rhizobium.

c) Check the model assumptions:

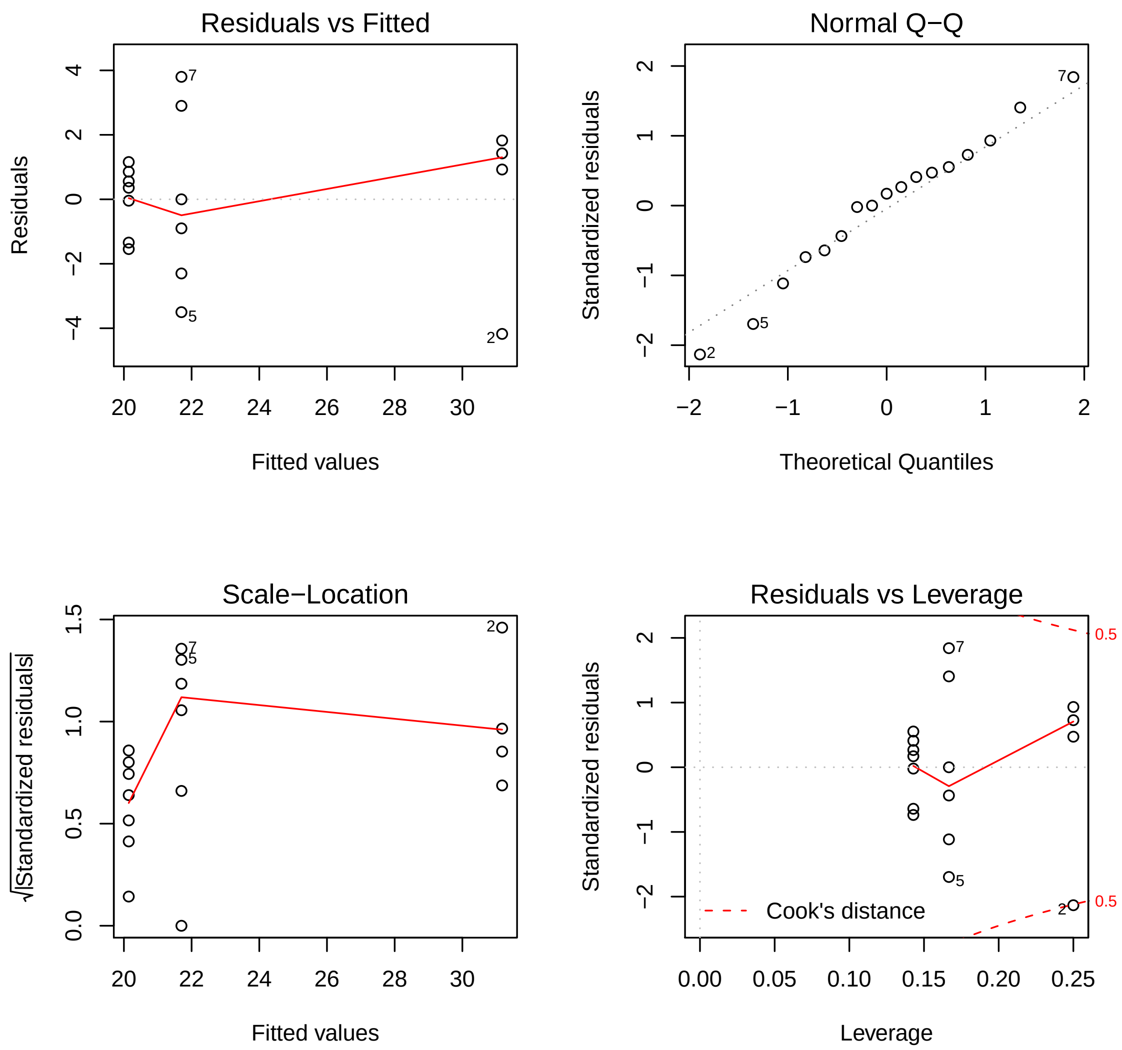
```
par(mfrow=c(2,2))
plot(fit.n2)
```



From the diagnostic plots we see that there exists an outlier. On the basis of the plots, observation number 1 can be clearly identified as an outlier. After removing the outlier we repeat the analysis.

```
rr.data <- r.data[-1,]
fit.n2mod <- aov(rr.data$N2~rr.data$strain)
summary(fit.n2mod)
               Df Sum Sq Mean Sq F value    Pr(>F)
rr.data$strain  2 333.19  166.60    32.6 5.393e-06 ***
Residuals      14  71.54    5.11

par(mfrow=c(2,2))
plot(fit.n2mod)
```



We see that now the model assumptions are fulfilled.