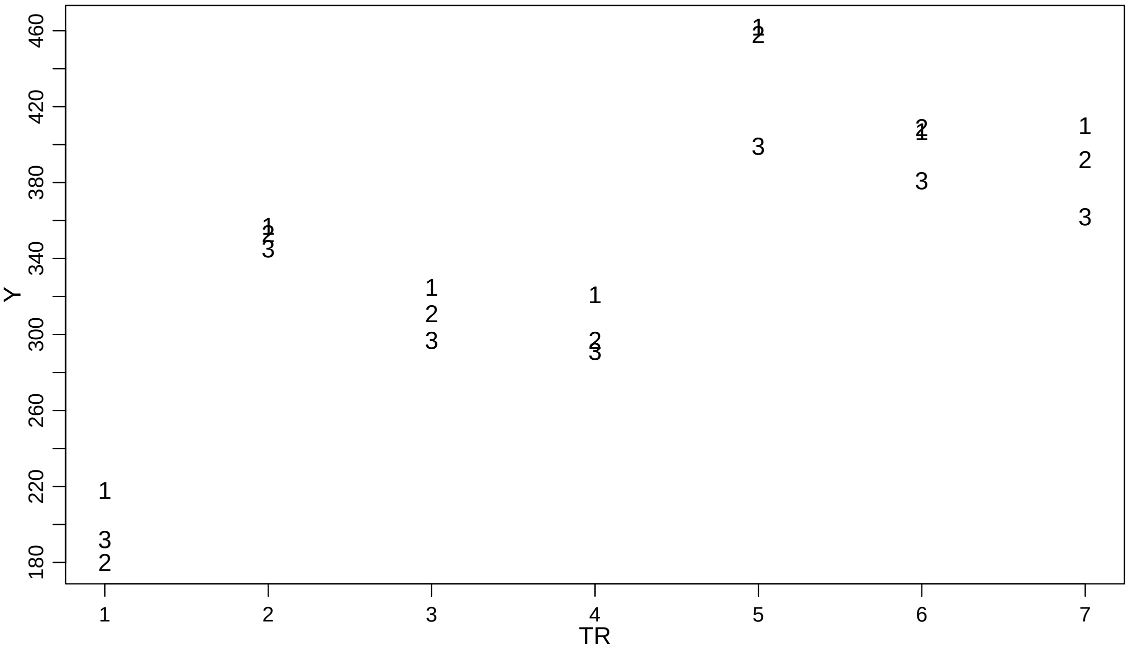
Solution Exercise 2

1. a) The plot shows clearly that there are big differences between the 7 treatments. The control issues the lowest values, while the values for treatments using no artificial manure (`TR` $= 2, 3, 4$) are clearly lower than treatments using artificial manure (`TR` $= 5, 6, 7$).



R code for the plot

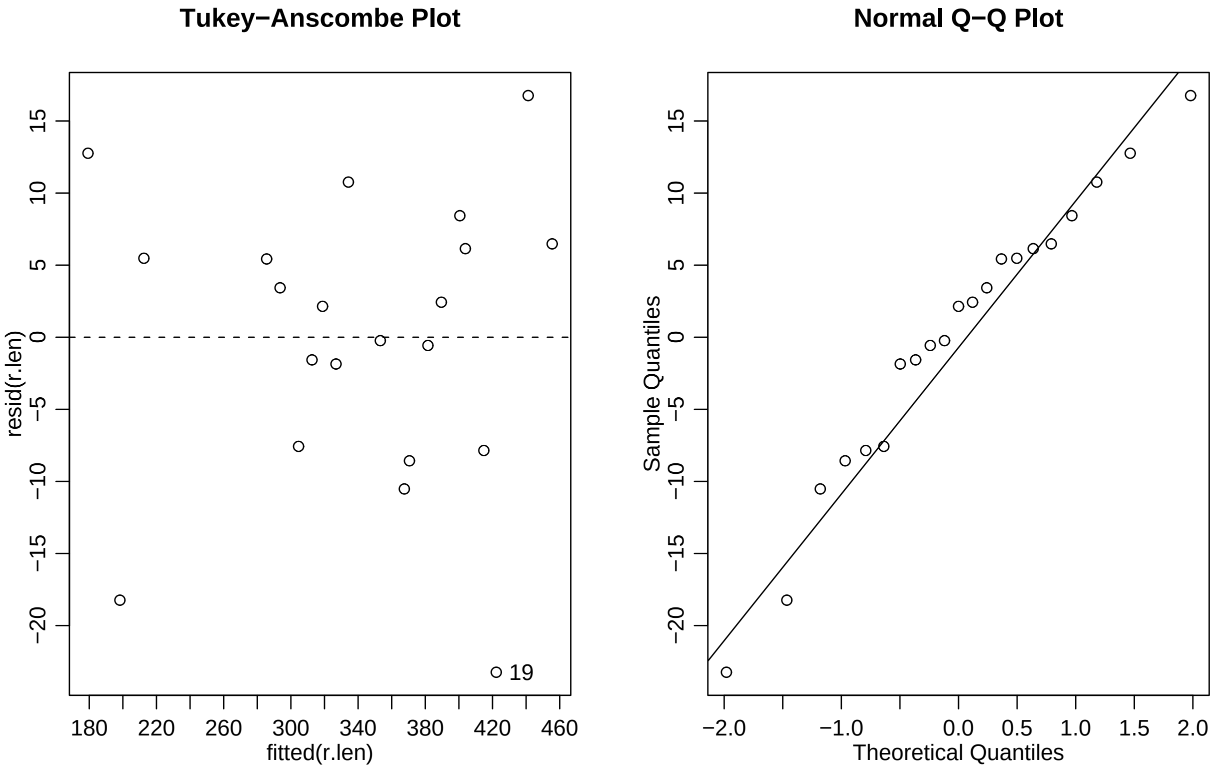
```
t.url <- "http://stat.ethz.ch/Teaching/Datasets/WBL/lentil.dat"
d.len <- read.table(t.url,header=T)
d.len$BLOCK <- factor(d.len$BLOCK)
d.len$TR <- factor(d.len$TR)
plot(as.numeric(d.len$TR),d.len$Y,type="n",xlab="TR",ylab="Y")
text(as.numeric(d.len$TR),d.len$Y,labels=d.len$BLOCK,cex=1.2)
```

2-way-ANOVA without interactions:

```
> r.len <- aov(Y ~ TR + BLOCK, d.len)
> summary(r.len)
            Df Sum Sq Mean Sq F value    Pr(>F)
TR           6 115792   19299 117.300 6.038e-10 ***
BLOCK        2   3904    1952  11.864  0.001436 **
Residuals   12   1974     165
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table confirms the above conclusion about the treatments (factor `TR`) being significantly different. The p-value is smaller than 0.001.

Checking the residual analysis plots we can see an extreme value (observation 19) in the Tukey-Anscombe plot. The normal plot shows no real deviation from the assumption of normality.



b) All contrasts are orthogonal since $\sum_{i=1}^{7} \lambda_{ji} \cdot \lambda_{ki} = 0$ for all $j \neq k$:.
Example for contrasts L1 and L2:

$$\sum_{i=1}^{7} \lambda_{1i} \cdot \lambda_{2i} = -6 \cdot 0 + 1 \cdot (-1) + 1 \cdot (-1) + 1 \cdot (-1) + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 = 0\,.$$

The contrasts describe the following comparisons:

contrast	comparison
L1	control vs rest
L2	artificial manure vs no artificial manure
L3	manual weeding vs herbicidal weeding
L4	spray herbicide before vs. spray herbicide afterwards
L5	interaction artificial manure * (manual weeding vs herbicidal weeding)
L6	interaction artificial manure * (spray herbicide before vs. spray herbicide afterwards)

The simplest way to detect orthogonality is by combining the contrasts to a matrix C (e.g. using `cbind`) and looking at $C^T C$. The matrix $C^T C$ is diagonal if and only if all contrasts are orthogonal.

c) The following procedure only works with orthogonal contrasts.
R code:

```
lent.contr <- cbind(c(-6,1,1,1,1,1,1), c(0,-1,-1,-1,1,1,1),
                    c(0,2,-1,-1,2,-1,-1), c(0,0,-1,1,0,-1,1),
                    c(0,-2,1,1,2,-1,-1), c(0,0,1,-1,0,-1,+1))

contrasts(d.len$TR) <- lent.contr
r.len <- aov(Y ~ TR + BLOCK, data = d.len)
summary(r.len,split=list(TR=list(L1=1,L2=2,L3=3,L4=4,L5=5,L6=6)))
summary.lm(r.len)
```

R output:

```
            Df Sum Sq Mean Sq F value  Pr(>F)
BLOCK        2   3904    1952   11.86  0.0014 **
TR           6 115792   19299  117.30 6.0e-10 ***
  TR: L1     1  73201   73201  444.93 7.5e-11 ***
  TR: L2     1  34061   34061  207.02 6.3e-09 ***
  TR: L3     1   8251    8251   50.15 1.3e-05 ***
  TR: L4     1    271     271    1.65  0.2238
  TR: L5     1      2       2    0.01  0.9088
  TR: L6     1      7       7    0.04  0.8429
Residuals   12   1974     165

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  357.143      4.848  73.668  < 2e-16 ***
BLOCK2       -14.286      6.856  -2.084 0.059240 .
BLOCK3       -33.286      6.856  -4.855 0.000395 ***
TR1           24.103      1.143  21.093 7.49e-11 ***
TR2           43.500      3.023  14.388 6.25e-09 ***
TR3           15.139      2.138   7.082 1.28e-05 ***
TR4           -4.750      3.703  -1.283 0.223775
TR5            0.250      2.138   0.117 0.908839
TR6           -0.750      3.703  -0.203 0.842878
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.83 on 12 degrees of freedom
Multiple R-Squared: 0.9838,     Adjusted R-squared: 0.973
F-statistic: 90.94 on 8 and 12 DF,  p-value: 1.47e-09
```

On a 5% level contrasts L1, L2 and L3 are significant. Contrasts L4, L5 and L6 are not significant.

Remark:

- In the case of nonorthogonal contrasts a separate model has to be computed for each contrast. More precisely: Contrasts which are orthogonal can be combined analysed using the above procedure. All other contrast have to be analysed separately.
- The matrix of contrasts for `TR` has the form (see also d)):

```
>      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    -6    0    0    0    0    0
[2,]     1   -1    2    0   -2    0
[3,]     1   -1   -1   -1    1    1
[4,]     1   -1   -1    1    1   -1
[5,]     1    1    2    0    2    0
[6,]     1    1   -1   -1   -1   -1
[7,]     1    1   -1    1   -1    1
```

d) R code:

```
model.matrix(r.len)
```

R output:

```
(Intercept) TR1 TR2 TR3 TR4 TR5 TR6 BLOCK2 BLOCK3
1             1  -6   0   0   0   0  0      0      0
2             1   1  -1   2   0  -2  0      0      0
3             1   1  -1  -1  -1   1  1      0      0
```

```
4           1  1 -1 -1  1  1 -1    0    0
5           1  1  1  2  0  2  0    0    0
6           1  1  1 -1 -1 -1 -1    0    0
7           1  1  1 -1  1 -1  1    0    0
8           1 -6  0  0  0  0  0    1    0
9           1  1 -1  2  0 -2  0    1    0
10          1  1 -1 -1 -1  1  1    1    0
11          1  1 -1 -1  1  1 -1    1    0
12          1  1  1  2  0  2  0    1    0
13          1  1  1 -1 -1 -1 -1    1    0
14          1  1  1 -1  1 -1  1    1    0
15          1 -6  0  0  0  0  0    0    1
16          1  1 -1  2  0 -2  0    0    1
17          1  1 -1 -1 -1  1  1    0    1
18          1  1 -1 -1  1  1 -1    0    1
19          1  1  1  2  0  2  0    0    1
20          1  1  1 -1 -1 -1 -1    0    1
21          1  1  1 -1  1 -1  1    0    1
```

2. a) The model is:

$$Y_{ij} = \mu + A_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

We calculate means and treatment effects:

Type						Mean	$\hat{A}_i$
T1	9	12	10	8	15	10.8	-3
T2	20	21	23	17	30	22.2	8.4
T3	6	5	8	16	7	8.4	-5.4
Mean						13.8	0

and the ANOVA-table[1]:

```
> v <- rep(1,5)
> y <- c(9,12,10,8,15,20,21,23,17,30,6,5,8,16,7)
> circ <- data.frame(Type=c(v,v*2,v*3),Y=y)
> circ$Type <- factor(circ$Type)
> circ.fit <- aov(formula =Y~Type , data=circ)
> summary(circ.fit)
            Df  Sum Sq Mean Sq F value    Pr(>F)
Type         2  543.6   271.8  16.083 0.0004023 ***
Residuals   12  202.8    16.9
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The P-value is smaller than 0.001 (and also smaller than 0.05), that means we reject the hypothesis: $A_1 = A_2 = A_3 = 0$.

The same conclusion can be obtained by using the critical F-value instead of the P-value:

$$F^{crit}_{2;12}(95\%) = 3.89 < 16.083$$

Consequently the hypothesis that all $A_i = 0$ is rejected.

[1] For calculations by hand see below.

Calculations by hand:

$$\begin{aligned}
543.6 &= \sum_i J_i A_i^2 = 5\cdot 3^2 + 5\cdot 8.4^2 + 5\cdot 5.4^2 \\
202.8 &= \sum_{ij} (y_{ij} - \hat{\mu} - A_i)^2 \qquad\qquad \text{(where } \hat{\mu} = 13.8\text{)} \\
MS &= \frac{SS}{Df} \\
F &= \frac{MS_{type}}{MS_{res}} = \frac{271.8}{16.9} = 16.08
\end{aligned}$$

Remark: If our calculations are correct, the total square error is equal to the sum of the SS, i.e.

$$\sum SS = SS_{type} + SS_{res} = 543.6 + 202.8 = 746.4$$

$$SS_{tot} = \sum (y_{ij} - \hat{\mu})^2 = \sum (y_{ij} - 13.8)^2 = 746.4$$

b) With the function "TukeyHSD" we can compare pairs of treatment means.

```
> TukeyHSD(circ.fit,"Type", conf.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Y ~ Type, data = circ)

$Type
     diff        lwr        upr     p adj
2-1  11.4   4.463555 18.336445 0.0023656
3-1  -2.4  -9.336445  4.536445 0.6367043
3-2 -13.8 -20.736445 -6.863555 0.0005042
```

The result can be interpreted as follows:
Type 2 is different from the other two types. The difference between type 1 and 3 is not significantly different from 0.

c)

i	1	2	3
$y_{i.}$	10.8	22.2	8.4

Test	Contrast	λ_1	λ_2	λ_3	$\hat{L}$	$\omega := \sum_i(\lambda_i^2/J)$	$SS_L = \hat{L}^2/\omega$
T2 vs. other	L1	1	-2	1	-25.2	1.2	529.2
T1 vs. T3	L2	1	0	-1	2.4	0.4	14.4

with R we can define the contrasts as follows:

```
> circ.contr <- cbind(c(1,-2,1),c(1,0,-1))
> contrasts(circ$Type) <- circ.contr
```

d) Using the SS_L we calculate the MS_L for the contrasts. By dividing MS_L by MS_{res}, we obtain the F-value. The results are listed in the next R output.

$$\begin{aligned}
\text{Mean Sq} &= \frac{\text{Sum Sq}}{\text{Df}} \\
\text{F Value} &= \frac{\text{Mean Sq}}{\text{MS of the Residuals}}
\end{aligned}$$

With R we obtain:

```
> circ.ctr.fit <- aov(formula =Y~Type , data=circ)
> summary(circ.ctr.fit,split=list(Type=list(L1=1,L2=2)))
            Df Sum Sq Mean Sq F value    Pr(>F)
Type         2  543.6   271.8 16.0828 0.0004023 ***
  Type: L1   1  529.2   529.2 31.3136 0.0001169 ***
  Type: L2   1   14.4    14.4  0.8521 0.3741550
Residuals   12  202.8    16.9
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

3. **a)** Model: $Y_{ij} = \mu + Treat_i + block_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), block_j \sim \mathcal{N}(0, \sigma_b^2)$.
Overall mean:

$\hat{\mu} = 20$

Table of residuals and means:

Residuals:	A	B	C	Block means
Technician 1	2	1	-3	15
Technician 2	-2	-1	3	25
Treatment means	10	40	10	

b) We can construct the following ANOVA table[2].

```
> v <- rep(1,3)
> Po <- data.frame(TE=c(v,v*2),TR=1:3,Y=c(7,36,2,13,44,18))
> Po$TE <- as.factor(Po$TE)
> Po$TR <- as.factor(Po$TR)
> Po.aov <- aov(formula =Y~TR+TE , data=Po)
> summary(Po.aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
TR           2   1200     600  42.857 0.02280 *
TE           1    150     150  10.714 0.08201 .
Residuals    2     28      14
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We calculate $\hat{\sigma}$ as follows:

$\hat{\sigma} = \sqrt{14} = 3.74$

4. **a)** With the functions

```
> st <- read.table("strawb.dat",header=TRUE)
> st$plot <- as.factor(st$plot)
> plot((st$gtype),st$yield,xlab="gtype",ylab="yield")
```

and

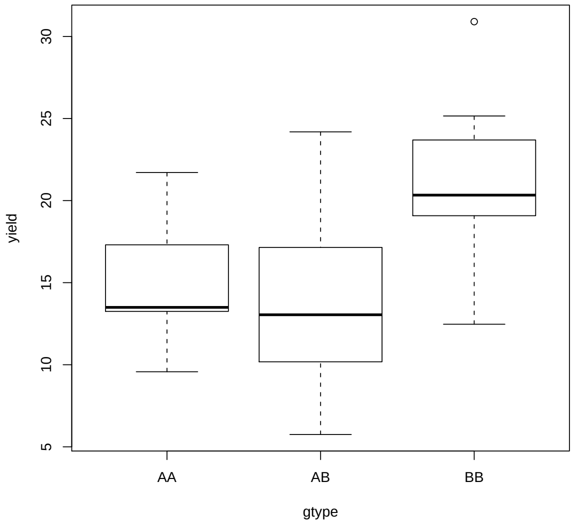
```
> plot(st$plot,st$yield,xlab="plot of land",ylab="yield")
```

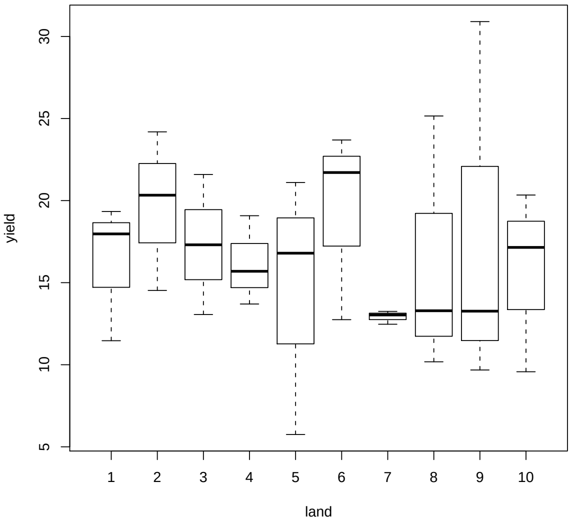
we plot the data.

The first figure shows a plot of gene type (x-axis) against yield (y-axis).
We notice that the gene type ”BB” seems to influence the yield. (Median and box[3] of the gene ”BB” are quite different from the ones of the genes ”AA” and ”AB”).
There is also some variability between different plots of land as can bee seen in the second graphic.

[2] Look at the Exercise 2 for an explanation of how the values are calculated.

[3] The box delimits the 50% of the data nearer to the median





b)

```
> st.a <- aov(formula=yield˜gtype+plot,data=st)
> summary(st.a)
            Df Sum Sq Mean Sq F value  Pr(>F)
gtype        2 289.65 144.824  5.4056 0.01450 *
plot         9 115.97  12.886  0.4810 0.86870
Residuals   18 482.25  26.792
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The factor ”genotype” is significant on a 5% level, but not on a 1% level. The block factor ”plot” does not have much influence on the yield.

c) We analyse the data without the block factor.

```
> st.n <- aov(formula=yield˜gtype,data=st)
> summary(st.n)
            Df Sum Sq Mean Sq F value   Pr(>F)
gtype        2 289.65 144.824  6.5364 0.004841 **
Residuals   27 598.22  22.156
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The factor ”genotype” is now significant on a 1% level.

d) The degree of freedom of the residuals are now $27 = 18 + 9$ because we are not considering block effects any more. With other words ”the effect of the plot is now considered as part of the error”.

Model c) appears to be favorable, but we would like to find out why blocking was not useful.