Working with orthogonal contrasts in R

Once you've done an **Analysis of Variance** (ANOVA), you may reach a point where you want to know:

What levels of the factor of interest were significantly different from one another?

Let us assume you've just analysed biomass data from a simple irrigation experiment. The dataset is constructed like this:

```
irrigation<-factor(c(rep("Control",10),rep("Irrigated 10
mm",10),rep("Irrigated 20 mm",10)))
biomass<-1:30
plot(x,y)
```

Now how would you set up the **overall ANOVA table** ? Well, in R (as you know) it's all very easy: We always use the following **model structure**:

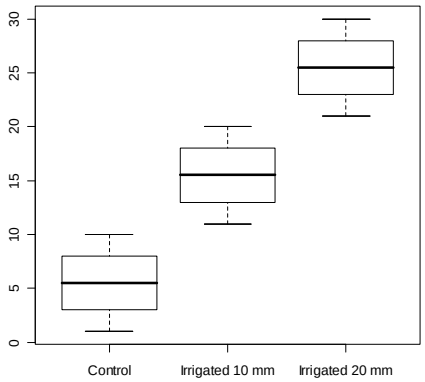
Model <- aov(Response variable ~Explanatory variable(s))

So in our case, the **response variable** is biomass, and the **explanatory variable** is irrigation (there is just one in this case). The ANOVA output is then:

```
summary(aov(biomass~irrigation))
            Df  Sum Sq Mean Sq F value    Pr(>F)
irrigation   2 2000.00 1000.00  109.09 1.162e-13 ***
Residuals   27  247.50    9.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we can see there is a huge effect of irrigation on biomass – but what does this mean in detail?

```
plot(irrigation,biomass)
```



We would expect all means to differ significantly from one another (in fact, in our case it´s true, because we have created the dataset like this on purpose). So, we would expect:

- Control differs significantly from 10 mm
- 10 mm differs significantly from 20 mm
- Control differs significantly from 20 mm
- 20 mm differs significantly from 10 mm
- 20 mm differs significantly from Control

All these **comparisons** are ***possible***, but:

There are only **k-1 orthogonal comparisons** (where k is the number of factor levels, which is 3 in our case).

So we know beforehand that **only two** of the comparisons listed above are orthogonal to each other (i.e., they´re **statistically independent**).

If we compare the Control with the 10 mm, and then the 10 mm with the 20 mm, we have implicitly also compared the Control with the 20 mm.

More formally, if you compare (A with B) and (B with C), this comparison already includes the comparison between (A and C).

Which of the possible comparisons should we conduct? Well, this very much depends on our hypothesis we have in mind. Let us assume that we think that Control differs from 10 mm and 20 mm. So:

- our **first comparison** would be between Control and 10 mm
- and the **second one** would be for Control and 20 mm

Now comes the tricky part: We need to specify a contrast matrix, showing which comparisons we want to make. A **contrast matrix** consists of so-called **contrast coefficients** that (in the end) all have to **sum to zero**. This means, those things we want to compare have to get the **opposite sign** (e.g. +1 and –1), while those things we don´t want to compare will receive a **value of zero**. In our case, the matrix of contrast coefficients would look like this:

Levels of Irrigation	**First comparison:** Control versus 10 mm	**Second comparison:** Control versus 20 mm
Control	**-1**	**-1**
10 mm	**1**	0
20 mm	0	**1**
Sum	0	0

So how do we set up this matrix in R? First, let´s extract the **default contrast matrix** for "Irrigation":

```
contrasts(irrigation)
               Irrigated 10 mm Irrigated 20 mm
Control                      0               0
Irrigated 10 mm              1               0
Irrigated 20 mm              0               1
```

This shows the default contrast matrix used in R, the so-called "**Treatment Contrasts**". It compares the **baseline level** ("Control") singly with the other levels

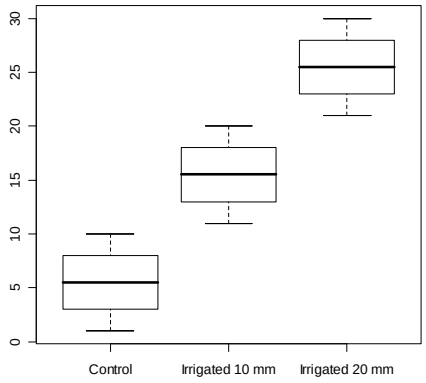
of the factor. So that´s already what we want! We don´t need to do anything more. *Note that the contrast matrix printed by R differs from what we´ve written above:* The contrast coefficients do *not* sum to zero – rather, the "1" indicates that the first comparison will be between Control and 10 mm, and the second comparison will be between Control and 20 mm.

Let´s directly see what this means. We use "**summary.lm**" instead of "summary" to split our ANOVA table from above according to the contrasts we defined. So, to repeat, here comes the "normal" table from above once more, and below comes the table with contrasts:

```
summary(aov(biomass~irrigation))
            Df  Sum Sq Mean Sq F value    Pr(>F)
irrigation   2 2000.00 1000.00  109.09 1.162e-13 ***
Residuals   27  247.50    9.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

################################################
summary.lm(aov(biomass~irrigation))

Residuals:
       Min         1Q     Median         3Q        Max
-4.500e+00 -2.500e+00 -4.163e-16  2.500e+00  4.500e+00

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 5.5000     0.9574   5.745 4.16e-06 ***
irrigationIrrigated 10 mm  10.0000     1.3540   7.385 6.05e-08 ***
irrigationIrrigated 20 mm  20.0000     1.3540  14.771 1.87e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.028 on 27 degrees of freedom
Multiple R-Squared: 0.8899,     Adjusted R-squared: 0.8817
F-statistic: 109.1 on 2 and 27 DF,  p-value: 1.162e-13
```

What does this all mean?

- The parameter labelled "**Intercept**" is the mean for the **Control** treatment (5.5)
- The parameter labelled "**irrigationIrrigated 10 mm**" is the *difference* between the mean of 10 mm and the Control mean (10.0)
- The parameter labelled "**irrigationIrrigated 20 mm**" is the difference between the 20 mm mean and the Control mean (20.0)
- Thus, our Control mean was 5.5; our 10 mm mean was (5.5 + 10.0) = 15.5; and our 20 mm mean was (5.5 + 20.0) = 25.5:



And, we see that all of the comparisons we made are **highly significant**. So everything is fine by now.

Let´s now assume we would rather like to have **a different kind of comparison**: We want to use the 20 mm mean as the "standard" against which the others should be tested. Thus, we construct a new contrast matrix like this:

Levels of Irrigation	**First comparison:** 10 mm versus 20 mm	**Second comparison:** 10 mm versus Control
Control	0	**-1**
10 mm	**1**	**1**
20 mm	**-1**	0
Sum	0	0

We first create **two column vectors**: One will be **c(0,1,-1)**, and the other one will be **c(-1,1,0)**. We bind these vectors together using **cbind()**, and inspect the result. Lets call our contrast matrix "contrastmatrix":

```
contrastmatrix<-cbind(c(0,1,-1),c(-1,1,0))
contrastmatrix
     [,1] [,2]
[1,]    0   -1
[2,]    1    1
[3,]   -1    0
```

Now, we use this contrast matrix for our factor "irrigation", like this:

```
contrasts(irrigation)<-contrastmatrix
summary.lm(aov(biomass~irrigation))

Residuals:
       Min         1Q     Median         3Q        Max
-4.500e+00 -2.500e+00  3.608e-16  2.500e+00  4.500e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.5000     0.5528   28.04  < 2e-16 ***
irrigation1 -10.0000     0.7817  -12.79 5.67e-13 ***
irrigation2  10.0000     0.7817   12.79 5.67e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.028 on 27 degrees of freedom
Multiple R-Squared: 0.8899,     Adjusted R-squared: 0.8817
F-statistic: 109.1 on 2 and 27 DF,  p-value: 1.162e-13
```

We see the change:

(1) The parameter called "**intercept**" is now the 10 mm treatment mean (15.5).

(2) The parameter called "**irrigation1**" is our first comparison (between 10mm and 20 mm):

i_{10}-i_{20}=-10 ➔ 15.5-i_{20}=-10 ➔ i_{20}=15.5+10=25.5

(3) Likewise, the parameter called "**irrigation2**" is our second comparison (between 10mm and Control):

i_{10}-C = +10 ➔ 15.5-C = +10 ➔ C = 15.5-10 = 5.5

Again, all comparisons we made are **highly significant**.

We conclude that Control, 10 mm and 20 mm all differed at $P < 0.05$ from one another. All irrigation treatments had highly significant effects on biomass

Note: R offers several built-in kinds of contrasts. They are specified using

```
contr.treatment(levels(irrigation))
```

```
                Irrigated 10 mm Irrigated 20 mm
Control                       0               0
Irrigated 10 mm               1               0
Irrigated 20 mm               0               1
```

Thus, you can use the following syntax to create **treatment contrasts** (the ones we used in our first contrast matrix above):

```
contrasts(irrigation)<- contr.treatment(levels(irrigation))
```

Likewise, you can use:

```
contr.helmert(…)
contr.poly(…)
contr.sum(…)
contr.SAS(…)
```

Details can be found using the help pages on

```
?contrasts
?contr.treatment
```