Transcript of Mick Crawley's R course 2010 Imperial College London, Silwood Park

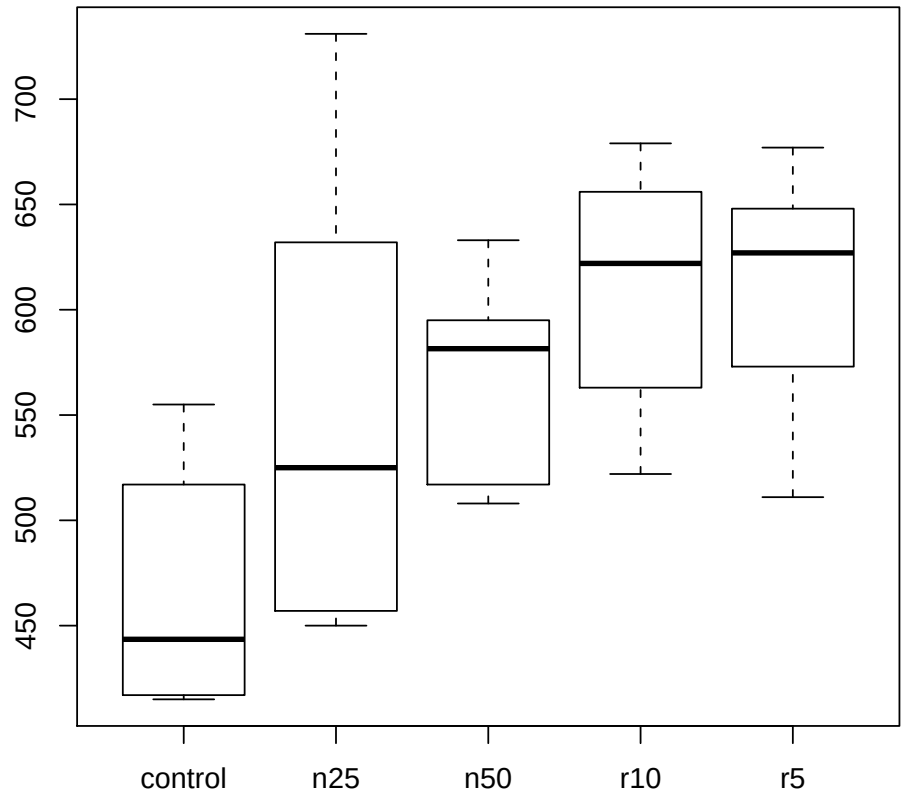
Emanuel G Heitlinger

Disclaimer: The following document is a private transcript of Mick Crawley's R-course. I am a participant in this course and my writeup has in no way been approved by Mick Crawley (from whom the ideas behind the code and teaching concepts are) or any of his staff.

Contrasts, single degree of freedom comparisons

We read in some new data. In these data we have 5 treatments: One control and 4 groups of manipulated plants. The response is the biomass.

```
> coex <- read.table("compexpt.txt", header = TRUE)
> attach(coex)
> plot(clipping, biomass)
```



Now in this experimental setup we have what we call *a priori* contrasts. We will want to compare the control treatment's mean with the means of the other 3 treatments first off all.

Such contrasts are non-problematic! Note that choosing the factor levels which have the highest and lowest means for *a posteriori* contrasts is only justifyalble after the Anova has already estblished significant differences.

Hand-made orthogonal contrasts

Implicit use

If there are k factor-levels, ther are $\sum_{i=2}^{k-1} i$ factor-level combinations. But only $k-1$ are orthogonal (independent). This means that if you have a factor with levels ABC and you compare AB and AC, you implicitly compared BC.

Back to the example:

```
> m1 <- aov(biomass ~ clipping)
> sm1 <- summary(m1)
> sm1

            Df Sum Sq Mean Sq F value   Pr(>F)
clipping     4  85356 21339.1  4.3015 0.008752 **
Residuals   25 124020  4960.8
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

> ssa <- sm1[[1]]$Sum[[1]]
```

Yes, there are highly significant differences. The treatment sum of squares (85356.47) is made up by $k - 1$=4 orthogonal contrast.

Let's the compare the control to the rest of it.

```
> c1 <- factor(1 + (clipping == "control"))
> m2 <- aov(biomass ~ c1)
> sm2 <- summary(m2)
> sm2

            Df Sum Sq Mean Sq F value   Pr(>F)
c1           1  70035   70035  14.073 0.000815 ***
Residuals   28 139342    4976
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

> ssc <- sm2[[1]]$Sum[[1]]
```

So we see that the control explaines 70035.01 out of the 85356.47. We can look if there are still significant contrasts left after this comparison. The formula is a bit more complicated as we have to "weight out" the contrast already analysed.

```
> c2 <- factor(1 + (clipping == "r10") + (clipping ==
+     "r5"))
> m3 <- aov(biomass ~ c2, weight = 1 * (clipping !=
+     "control"))
> sm3 <- summary(m3)
> sm3
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
c2           1  14553   14553  2.9592 0.09943 .
Residuals   22 108196    4918
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ
```

```
> ssrn <- sm3[[1]]$Sum[[1]]
```

This contrast is not significant it contains only 14553.37 of the total variation.

Explicit use

There are also sophisticated functions built into R to do these kinds of analyses. Let's make up own contrasts first accordign to our *a priori* of differences between control and others:

```
> contrasts(clipping) <- cbind(c(4, -1, -1, -1,
+     -1), c(0, 1, 1, -1, -1), c(0, 0, 0, 1, -1),
+     c(0, -1, 1, 0, 0))
> own.c <- contrasts(clipping)
> own.c
```

```
        [,1] [,2] [,3] [,4]
control    4    0    0    0
n25       -1    1    0   -1
n50       -1    1    0    1
r10       -1   -1    1    0
r5        -1   -1   -1    0
```

We built a contrast-matrix, for it to be orthogonal all products of coefficients have to be zero and the column-sums have to be zero. Let's test it like this:

```
> orth.test <- function(x) {
+     co <- combn(1:ncol(x), 2)
+     tmp <- vector()
+     for (r in 1:ncol(co)) {
+         for (i in 1:nrow(x)) {
+             tmp[i] <- x[[i, co[[1, r]]]] * x[[i,
+                 co[[2, r]]]]
+         }
+         if (sum(tmp) != 0) {
```

```
+             print("Non-zero coefficient product")
+             break
+         }
+     }
+     if (sum(tmp) == 0) {
+         if (sum(abs(colSums(x))) == 0) {
+             print("orthogonal contrasts")
+         }
+         else print("NON-zero clumns sums")
+     }
+ }
> orth.test(own.c)

[1] "orthogonal contrasts"
```

Hurray it is orthogonal! And this was a fun exercise in R-programming: combn is a neat function.

So remember we spicified the contrasts on clipping, therefore the next model and it's output is for the contrasts.

```
> model <- aov(biomass ~ clipping)
> mod1.smaov <- summary.aov(model)
> summary.lm(model)

Call:
aov(formula = biomass ~ clipping)

Residuals:
     Min       1Q   Median       3Q      Max
-103.333  -49.667    3.417   43.375  177.667

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 561.80000   12.85926  43.688  < 2e-16 ***
clipping1   -24.15833    6.42963  -3.757 0.000921 ***
clipping2   -24.62500   14.37708  -1.713 0.099128 .
clipping3     0.08333   20.33227   0.004 0.996762
clipping4     8.00000   20.33227   0.393 0.697313
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Residual standard error: 70.43 on 25 degrees of freedom
```

```
Multiple R-squared: 0.4077,        Adjusted R-squared: 0.3129
F-statistic: 4.302 on 4 and 25 DF,  p-value: 0.008752
```

Compare the probability for the comparison in contrast2 with the same comparison carried out in the model m3 above (with creation of a new level and weights) not perfectly the same, but close!

The first line (Intercept) shows the grand mean, the following lines show the differences in the contrast relative to the grand mean. For contrast 1: The overall versus the four manipulated treatment levels mean (this implicitly contrasts the control). For contrast 2: The n versus the r levels, note that the difference is twice the value shown. Two times the printed value is also the differende for the next two contrast. They are from beeing significant.

Standard errors differ, because sample sizes for contrasts differ. Look in the next code snippet how they are derived, from the error-mean square (extracted from the model), if you don't know where the sums of squares and mean squares come from go back to ancova-anova writeup.

```
> err.var <- mod1.smaov[[1]]$Mean[[2]]
> c(intercept = sqrt(err.var/length(biomass)), clipping1 = sqrt(err.var/120),
+     clipping2 = sqrt(err.var/24), clipping3 = sqrt(err.var/12),
+     clipping4 = sqrt(err.var/12))

intercept clipping1 clipping2 clipping3 clipping4
12.859255  6.429628 14.377084 20.332268 20.332268
```

The different numbers come from the different numbers involved in the comparisons: First row is boring, it is just the total SE. In the second row we compare a mean coming from four treatment levels (6x4=24 numbers) to the overall mean (30 numbers), it is a MYSTERY even to Mick Crawley where the 120 needed to get the right result comes from.

Third row compares two treatment levels with two other (in total 24 numbers). Then two times singele means are compared: in total 12 numbers. This makes sense again!

```
> contrasts(clipping) <- NULL
```

Helmert contrasts

Remind ourselfs of the treatment means and do the Helmert contrasts:

```
> tpl <- tapply(biomass, clipping, mean)
> tpl
```

```
 control      n25      n50      r10       r5
465.1667 553.3333 569.3333 610.6667 610.5000
```

```
> options(contrasts = c("contr.helmert", "contr.poly"))
> model2 <- lm(biomass ~ clipping)
> summary(model2)

Call:
lm(formula = biomass ~ clipping)

Residuals:
     Min       1Q   Median       3Q      Max
-103.333  -49.667    3.417   43.375  177.667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  561.800     12.859  43.688   <2e-16 ***
clipping1     44.083     20.332   2.168   0.0399 *
clipping2     20.028     11.739   1.706   0.1004
clipping3     20.347      8.301   2.451   0.0216 *
clipping4     12.175      6.430   1.894   0.0699 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Residual standard error: 70.43 on 25 degrees of freedom
Multiple R-squared: 0.4077,        Adjusted R-squared: 0.3129
F-statistic: 4.302 on 4 and 25 DF,  p-value: 0.008752
```

First is again the overall mean. Next row difference between the mean of treatment 1 and the average of the means of treatment 1 and 2.

```
> mean(c(tpl[1], tpl[2])) - tpl[1]

 control
44.08333
```

Third row is the difference between the average of the first two and the first three means.

```
> mean(c(tpl[1], tpl[2], tpl[3])) - mean(c(tpl[1],
+     tpl[2]))

[1] 20.02778
```

Fourth row the difference between the average of the first three and the first four.

```
> mean(c(tpl[1], tpl[2], tpl[3], tpl[4])) - mean(c(tpl[1],
+     tpl[2], tpl[3]))
```

```
[1] 20.34722
```

The final row is the difference between the overall mean and the first 4 treatments.

```
> mean(biomass) - mean(c(tpl[1], tpl[2], tpl[3],
+     tpl[4]))
```

```
[1] 12.175
```

Statisticians like Helmert contrasts because they are the only systematic contrasts, that are orthogonal. Let's see what they look like:

```
> helm.c <- contrasts(clipping)
> helm.c
```

```
        [,1] [,2] [,3] [,4]
control   -1   -1   -1   -1
n25        1   -1   -1   -1
n50        0    2   -1   -1
r10        0    0    3   -1
r5         0    0    0    4
```

```
> orth.test(helm.c)
```

```
[1] "orthogonal contrasts"
```

Yes really!

```
> contrasts(clipping) <- NULL
```

Sum contrasts

```
> options(contrasts = c("contr.sum", "contr.poly"))
> model3 <- lm(biomass ~ clipping)
> summary(model3)

Call:
lm(formula = biomass ~ clipping)

Residuals:
     Min       1Q   Median       3Q      Max
-103.333  -49.667    3.417   43.375  177.667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  561.800     12.859  43.688  < 2e-16 ***
clipping1    -96.633     25.719  -3.757 0.000921 ***
clipping2     -8.467     25.719  -0.329 0.744743
clipping3      7.533     25.719   0.293 0.772005
clipping4     48.867     25.719   1.900 0.069019 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Residual standard error: 70.43 on 25 degrees of freedom
Multiple R-squared: 0.4077,        Adjusted R-squared: 0.3129
F-statistic: 4.302 on 4 and 25 DF,  p-value: 0.008752
```

Intercept is the overall mean again. Each row is the difference between one treatment and the grand mean.

```
> sapply(1:4, function(i) tpl[i] - mean(biomass))

   control        n25        n50        r10
-96.633333  -8.466667   7.533333  48.866667
```

Note that the comparison of the remaining treatment level would lead to overparameterisation (we have chosen the grand mean for the intercept). So the following is for its last element not correct in our model.

```
> sapply(1:5, function(i) tpl[i] - mean(biomass))

   control        n25        n50        r10         r5
-96.633333  -8.466667   7.533333  48.866667  48.700000
```

The standard error of the overall mean is the same than with any other contrasts, the other standard errors involve :

```
> c(intercept = sqrt(err.var/length(biomass)), all.others = sqrt(err.var/12 +
+     err.var/20))

 intercept all.others
  12.85926   25.71851
```

Again the 12 and the 20 are a MYSTERY.

Let's see what sum contrasts look like:

```
> sum.c <- contrasts(clipping)
> sum.c

        [,1] [,2] [,3] [,4]
control    1    0    0    0
n25        0    1    0    0
n50        0    0    1    0
r10        0    0    0    1
r5        -1   -1   -1   -1

> orth.test(sum.c)

[1] "Non-zero coefficient product"
```

Their column sums are zero like for the Helmert contrasts, but they are not orthogonal as the product of their coefficients are not all zero.

```
> contrasts(clipping) <- NULL
```

Treatment contrasts

These are the **default contrasts in R**.

```
> options(contrasts = c("contr.treatment", "contr.poly"))
> model <- lm(biomass ~ clipping)
> summary(model)

Call:
lm(formula = biomass ~ clipping)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-103.333  -49.667    3.417   43.375  177.667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   465.17      28.75  16.177  9.4e-15 ***
clippingn25    88.17      40.66   2.168  0.03987 *
clippingn50   104.17      40.66   2.562  0.01683 *
clippingr10   145.50      40.66   3.578  0.00145 **
clippingr5    145.33      40.66   3.574  0.00147 **
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Residual standard error: 70.43 on 25 degrees of freedom
Multiple R-squared: 0.4077,        Adjusted R-squared: 0.3129
F-statistic: 4.302 on 4 and 25 DF,  p-value: 0.008752
```

The first coefficient is the mean of the first factor level (alphabetical order). The other coefficients are the difference between the (named) factor level and the first factor level. Like this:

```
> sapply(2:5, function(i) tpl[i] - tpl[1])

      n25       n50       r10        r5
 88.16667 104.16667 145.50000 145.33333
```

The standard error of the intercept is the standard error of the first treatment mean. The remaining are the standard errors for the difference of two means compared.

```
> sqrt(err.var/length(biomass[clipping == "control"]))

[1] 28.75417

> sapply(2:5, function(i) sqrt(err.var/length(biomass[clipping ==
+     "control"]) + err.var/length(biomass[clipping ==
+     levels(clipping)[i]])))

[1] 40.66454 40.66454 40.66454 40.66454
```

And finally a look at these treatment contrasts:

```
> treat.c <- contrasts(clipping)
> treat.c

        n25 n50 r10 r5
control   0   0   0  0
n25       1   0   0  0
n50       0   1   0  0
r10       0   0   1  0
r5        0   0   0  1

> orth.test(treat.c)

[1] "NON-zero clumns sums"
```

They are not orthogonal!

But they are very helpfull, because you can compare any means with any oter directly. But a **warning**: The probability is for the comparison with the first level, you can **not** conclude from this to the probability to retain a particular factor-level in the model.

Contrasts in Ancova

Helmert contrasts

```
> ipo <- read.table("ipomopsis.txt", header = T)
> options(contrasts = c("contr.helmert", "contr.poly"))
> modelH <- lm(ipo$Fruit ~ ipo$Root + ipo$Grazing)
> summary(modelH)

Call:
lm(formula = ipo$Fruit ~ ipo$Root + ipo$Grazing)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1920  -2.8224   0.3223   3.9144  17.3290

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -109.778      8.318  -13.20 1.45e-15 ***
ipo$Root       23.560      1.149   20.51  < 2e-16 ***
ipo$Grazing1   18.052      1.679   10.75 6.11e-13 ***
---
```

```
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Residual standard error: 6.747 on 37 degrees of freedom
Multiple R-squared: 0.9291,        Adjusted R-squared: 0.9252
F-statistic: 242.3 on 2 and 37 DF,  p-value: < 2.2e-16
```

The intercept is the average of the two intercepts. The effect of root is the slope of the graph of Fruit against Root (same slope for both graphs). The effect of Grazing is the difference between Grazed an intercept and average intercept (i.e. half the difference between the two intercepts).

Sum contrasts

```
> options(contrasts = c("contr.sum", "contr.poly"))
> modelS <- lm(ipo$Fruit ~ ipo$Root + ipo$Grazing)
> summary(modelS)

Call:
lm(formula = ipo$Fruit ~ ipo$Root + ipo$Grazing)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1920  -2.8224   0.3223   3.9144  17.3290

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -109.778      8.318  -13.20 1.45e-15 ***
ipo$Root       23.560      1.149   20.51  < 2e-16 ***
ipo$Grazing1  -18.052      1.679  -10.75 6.11e-13 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Residual standard error: 6.747 on 37 degrees of freedom
Multiple R-squared: 0.9291,        Adjusted R-squared: 0.9252
F-statistic: 242.3 on 2 and 37 DF,  p-value: < 2.2e-16
```

The same as for Helmert, except the sign for Grazing is reversed.

Treatment contrasts

```
> options(contrasts = c("contr.treatment", "contr.poly"))
> modelT <- lm(ipo$Fruit ~ ipo$Root + ipo$Grazing)
```

```
> smt <- summary(modelT)
> smt

Call:
lm(formula = ipo$Fruit ~ ipo$Root + ipo$Grazing)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1920  -2.8224   0.3223   3.9144  17.3290

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -127.829      9.664  -13.23 1.35e-15
ipo$Root              23.560      1.149   20.51  < 2e-16
ipo$GrazingUngrazed   36.103      3.357   10.75 6.11e-13

(Intercept)         ***
ipo$Root            ***
ipo$GrazingUngrazed ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ

Residual standard error: 6.747 on 37 degrees of freedom
Multiple R-squared: 0.9291,        Adjusted R-squared: 0.9252
F-statistic: 242.3 on 2 and 37 DF,  p-value: < 2.2e-16

> rv <- c(round(smt[[4]][[1, 1]], 2), round(smt[[4]][[2,
+     1]], 2), round(smt[[4]][[3, 1]], 2))
> rn <- c(rownames(smt[[4]])[[1]], rownames(smt[[4]])[[2]],
+     rownames(smt[[4]])[[3]])
> rn <- gsub("ipo\\$", "", rn)
```

The **default** contrasts show the intercept for the factor level that comes first in the alphabet first (Grazed is (Intercept) -127.83). The second parameter (Root 23.56) is the slope of the Graph of Fruit against Root. The third parameter (GrazingUngrazed 36.1) is the difference between two intercepts it tells you that the ungrazed plants have about 36.1 more Fruit.