STATISTICS: AN INTRODUCTION USING R

By M.J. Crawley

Exercises

6. ANALYSIS OF COVARIANCE

Analysis of covariance combines elements from regression and analysis of variance. There is at least one continuous explanatory variable and at least one categorical explanatory variable. The procedure works like this:

- fit two or more linear regressions of *y* against *x* (one for each level of the factor)
- estimate different slopes and intercepts for each level
- use model simplification (deletion tests) to eliminate unnecessary parameters

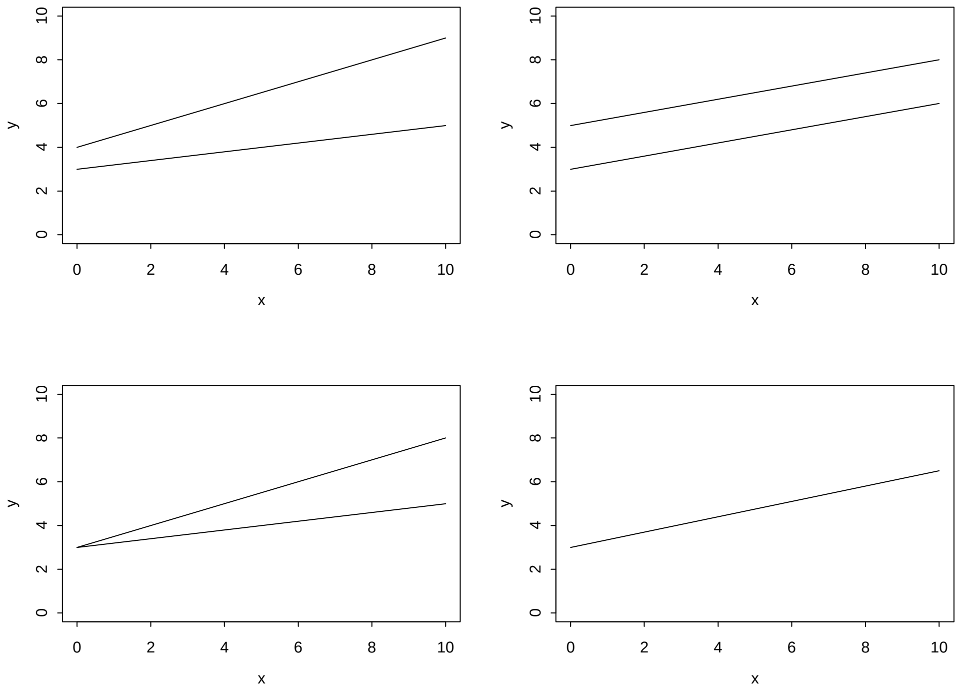
For example, we could use Ancova in a medical experiment where the response variable was 'days to recovery' and the explanatory variables were 'smoker or not' (categorical) and 'blood cell count' (continuous). In economics, local unemployment rate might be modelled as a function of country (categorical) and local population size (continuous).

Suppose we are modelling weight (the response variable) as a function of sex and age. Sex is a factor with 2 levels (male and female) and age is a continuous variable. The maximal model therefore has 4 parameters: two slopes (a slope for males and a slope for females) and two intercepts (one for males and one for females) like this:

$$weight_{male} = a_{male} + b_{male} \times age$$

$$weight_{female} = a_{female} + b_{female} \times age$$

This maximal model is shown in the top left panel:



Model simplification is an essential part of analysis of covariance, because the principle of parsimony requires that we keep as few parameters in the model as possible. In this case, the process of model simplification begins by asking whether we need all 4 parameters. Perhaps we could make do with 2 intercepts and a common slope (top right). Or a common intercept and two different slopes (bottom left). Alternatively, there may be no effect of sex at all, in which case we only need 2 parameters (one slope and one intercept) to describe the effect of age on weight (bottom right). There again, the continuous variable may have no significant effect on the response, so we only need 2 parameters to describe the main effects of sex on weight. This would show up as two separated, horizontal lines in the plot (one mean weight for each sex). In the limit, neither the continuous nor the categorical explanatory variables might have any significant effect on the response, in which case, model simplification will lead to the 1-parameter null model $\hat{y} = \bar{y}$ (a single, horizontal line).

Calculations involved in Ancova

We start with the simple example that we analysed as an Anova on p. 122, in which 6 different plant genotypes were exposed to 4 different light regimes. When we analysed the data as a 1-way Anova we found no significant effect of photoperiod, because the variation in growth rate between the different genotypes was so great. When we took differences between the genotype means into account by doing a 2-way Anova, we found a highly significant effect of photoperiod. Here we change the explanatory variable from a categorical variable (daylength with 4 levels) into a continuous variable (the number of hours of illumination). There are 2 benefits in doing this

- is saves us 2 degrees of freedom in describing the effect of photoperiod (1 slope and 1 intercept instead of 4 photoperiod means)
- it allows us to test for an interaction between genotype and photoperiod on growth response (recall that we couldn't do this using Anova because there was no replication, so the interaction term had to be used as the error term)
- other things being equal, a regression approach will generally be more powerful than an Anova-based approach, because Anova takes no account of the *ordering* of the factor levels, only of the differences between their means.

The calculations are as follows. We need to do 6 separate regressions of growth against photoperiod (one for each genotype), keeping account of the corrected sums of products, SSXY, for each one separately. Here are the data again, this time showing the number of hours of illumination as the continuous explanatory variable.

Genotype	8hr	12hr	16hr	24hr
A	2	3	3	4
B	3	4	5	6
C	1	2	1	2
D	1	1	2	2
E	2	2	2	2
F	1	1	2	3

Here is the Anova table that we calculated in Chapter 15:

Source	SS	d.f.	MS	F
Photoperiod	7.125	3	2.375	7.703
Genotype	27.875	5	5.575	18.08
Error	4.625	15	$s^2 = 0.308$	
Total	39.625	23		

The Genotype and Total sums of squares will remain as they were. What we plan to do is to reduce the Error sum of squares and increase the explained Photoperiod sum of squares by taking account of the fact that the different daylengths can be ordered.

We shall do an overall regression to begin with, just estimating a single overall slope for the relationship between growth and photoperiod. The Anova table will therefore look like this:

Source	SS	d.f.	MS	F
Photoperiod		1		
Genotype	27.875	5		
Error		17	s^2	
Total	39.625	23		

We have reduced the degrees of freedom for photoperiod from 3 to 1 (a useful model simplification in itself) and increased the error degrees of freedom accordingly (this is good for reducing the error variance). So we start by calculating an overall regression sum of squares using all 24 data points.

First we need the famous five $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$, $\sum xy$

```
photoperiod<-read.table("c:\\temp\\photoperiod.txt",header=T)
attach(photoperiod)
names(photoperiod)
```

```
[1] "Genotype"    "Growth"      "Photoperiod"
```

So x is Photoperiod and y is Growth. We get the famous 5 like this:

```
sum(Photoperiod);sum(Photoperiod^2)
```

```
[1] 360
[1] 6240
```

```
sum(Growth);sum(Growth^2)
```

```
[1] 57
[1] 175
```

```
sum(Growth*Photoperiod)
```

```
[1] 932
```

So the corrected sums of squares (see Chapter 14) are calculated like this

$$SST = 175 - \frac{57^2}{24} = 39.625$$

$$SSXY = 932 - \frac{57 \times 360}{24} = 77$$

$$SSX = 6240 - \frac{360^2}{24} = 840$$

Now we have everything we need to calculate the slope of the overall straight line relating growth to photoperiod (remember, $b = SSXY/SSX$)

$$b = \frac{77}{840} = 0.0917$$

The next step is to work out the sums of squares for the Anova table. First the explained sum of squares (the regression sum of squares, SSR):

$$SSR = b \times SSXY = 0.0917 \times 77 = 7.0583$$

We know SST = 39.625 already from our Anova calculations (and see above), so we obtain SSE by subtracting SSR from SST.

$$SSE = SST - SSR = 39.625 - 7.0583 = 32.567$$

The Anova table can now be completed:

Source	SS	d.f.	MS	F
Photoperiod	7.0583	1	7.0583	25.575
Genotype	27.875	5	5.575	20.199
Error	4.4917	17	s^2 = 0.276	
Total	39.625	23		

This is already a considerable improvement over the 2-way Anova: the F ratio for the effect of photoperiod has gone up from 7.703 to 25.575, and the error variance has gone down from 0.308 to 0.276 (see Chapter 15). But we can do better than this. We can ask whether there is any evidence that *different genotypes showed different responses* to photoperiod. This is an interaction term, which involves fitting 6 different slopes to the model instead of 1 common slope (as we just did here).

The calculations are not hard, just tedious. We need to find SSR separately for each of the 6 genotypes. For this, we need 6 values of *b* and 6 values of SSXY.

First, we need the sums and sums of squares

```
tapply(Growth,Genotype,sum)
```

```
 A  B  C  D  E  F
12 18  6  6  8  7
```

```
tapply(Growth^2,Genotype,sum)
```

```
 A  B  C  D  E  F
38 86 10 10 16 15
```

```
tapply(Photoperiod,Genotype,sum)
```

```
 A  B  C  D  E  F
60 60 60 60 60 60
```

```
tapply(Photoperiod^2,Genotype,sum)
```

```
  A    B    C    D    E    F
1040 1040 1040 1040 1040 1040
```

Finally, we calculate the sums of products

```
tapply(Photoperiod*Growth,Genotype,sum)
```

```
  A   B   C   D   E   F
196 296  96 100 120 124
```

All the SSX's are the same:

$$SSX = 1040 - \frac{60^2}{4} = 140$$

Now we calculate the 6 separate SSXY's

$$SSXY_1 = 196 - \frac{12 \times 60}{4} = 16.0$$

$$SSXY_2 = 296 - \frac{18 \times 60}{4} = 26.0$$

$$SSXY_3 = 96 - \frac{6 \times 60}{4} = 6.0$$

$$SSXY_4 = 100 - \frac{6 \times 60}{4} = 10.0$$

$$SSXY_5 = 120 - \frac{8 \times 60}{4} = 0.0$$

$$SSXY_6 = 124 - \frac{7 \times 60}{4} = 19.0$$

So the 6 slopes are 16/140 = 0.11429, 26/140 = 0.1857, 6/140 = 0.04286, 0 and 19/140 = 0.1357.

Finally we can work out the 6 SSR's

Clone	SSXY	SSX	b	SSR
A	$196-\frac{60\times 12}{4}=16$	$1040-\frac{60^2}{4}=140$	0.114	1.829
B	$296-\frac{60\times 18}{4}=26$	$1040-\frac{60^2}{4}=140$	0.186	4.829
C	$96-\frac{60\times 6}{4}=6$	$1040-\frac{60^2}{4}=140$	0.043	0.257
D	$100-\frac{60\times 6}{4}=10$	$1040-\frac{60^2}{4}=140$	0.071	0.714
E	$120-\frac{60\times 8}{4}=0$	$1040-\frac{60^2}{4}=140$	0	0
F	$124-\frac{60\times 7}{4}=19$	$1040-\frac{60^2}{4}=140$	0.136	2.579
			Total	10.208

The 6 component regression sums of squares add up to 10.208. Note that there is substantial variation between the slopes for the different genotypes (*b* ranges from 0 to 0.186).

The next step is to work out the sum of squares attributable to differences between the slopes. The logic is simple. A single regression slope explained an SSR of 7.058 (above). Different slopes, as we have just seen, explained an SSR of 10.208. So the difference, attributable to having 6 slopes instead of 1, is 10.208 – 7.058 = 3.149. Because we have estimated 6 slopes now, where we had 1 before, this sum of squares is based on 6 – 1 = 5 d.f..

Now we can draw all the information together into a single Anova table:

Source	SS	d.f.	MS	F
Genotype	27.87	5	5.574	43.21
Regression	7.059	1	7.059	54.72
Differences in slope	3.149	5	0.63	4.88
Error	1.55	12	s^2 = 0.129	
Total	39.63	23		

The Ancova has produced a much better model than either the simple regression or the 1-way Anova, and a substantially better model than the 2-way Anova. Generally, if the nature of your data allows it, it is a good idea to use analysis of covariance whenever you can.

Analysis of Covariance in R

We could use either **lm** or **aov**; the choice affects only the format of the summary table. We shall use both and compare their output. The model with 6 different slopes is specified using the asterisk operator

```
Growth ~ Genotype * Photoperiod
```

Note than in S-Plus the output of `summary.lm` would be different because it uses Helmert contrasts, rather than the Treatment contrasts used by R. To get the same output in S-Plus you would need to type the following (**but this is unnecessary in R**):

```
options(contrasts=c("contr.treatment","contr.poly"))
```

We call the output *model* and work like this, starting with **lm**

```
model<-lm(Growth~Genotype*Photoperiod)
```

```
summary(model)
```

```
Coefficients:
                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)            1.28571    0.48865   2.631  0.02193 *
GenotypeB              0.42857    0.69105   0.620  0.54674
GenotypeC             -0.42857    0.69105  -0.620  0.54674
GenotypeD             -0.85714    0.69105  -1.240  0.23855
GenotypeE              0.71429    0.69105   1.034  0.32169
GenotypeF             -1.57143    0.69105  -2.274  0.04213 *
Photoperiod            0.11429    0.03030   3.771  0.00267**
GenotypeB.Photoperiod  0.07143    0.04286   1.667  0.12145
GenotypeC.Photoperiod -0.07143    0.04286  -1.667  0.12145
GenotypeD.Photoperiod -0.04286    0.04286  -1.000  0.33705
GenotypeE.Photoperiod -0.11429    0.04286  -2.667  0.02054 *
GenotypeF.Photoperiod  0.02143    0.04286   0.500  0.62612
```

The output from Ancova takes a lot of getting used to. The most important thing to remember is that there are 12 rows in the summary table because the model has estimated 12 parameters from the data: 6 intercepts and 6 slopes.

With Treatment Contrasts (as here), the rows of the table are interpreted as follows:

- the Intercept in row 1 is the *intercept* for Genotype A
- the next 5 rows are *differences between intercepts*; for instance, row 2 (labelled Genotype B) is the difference in intercept between Genotypes B and A
- the *slope* is in row 6 (labelled Photoperiod); this is the slope of the graph of growth against photoperiod for Genotype A
- the last 5 rows are *differences between slopes*; for instance, row 7 (labelled GenotypeB : Photoperiod) is the difference between the slopes of the graphs for Genotypes B and A

So, for example, the parameterised equation for genotype D is as follows. The intercept is 1.28571 – 0.85714 = 0.42857, and the slope is 0.11429 – 0.04286 = 0.07143. The rightmost column indicates the parameter values that are significantly different from zero (when compared with Genotype A). The table shows that there is 1 significant interaction term (Genotype E has a significantly different response to Photoperiod compared with the other genotypes: in fact, it didn't respond at all to increasing light exposure. It also shows that Genotype F has a significantly lower intercept than has Genotype A.

This output generated by **lm** is good when you want to look at the individual parameter values. Let's see what happens when we fit the same model using **aov**.

```
model<-aov(Growth~Genotype*Photoperiod)
```

```
summary(model)
```

```
                     Df  Sum Sq Mean Sq F value     Pr(>F)
Genotype              5 27.8750  5.5750 43.3611 2.848e-007 ***
Photoperiod           1  7.0583  7.0583 54.8981 8.171e-006 ***
Genotype:Photoperiod  5  3.1488  0.6298  4.8981     0.0113 *
Residuals            12  1.5429  0.1286
```

Here we get a tidy Anova table to compare with the one we calculated earlier by hand, but there is no information on *which* parameters contribute to the significant differences. In practice, there is no need to fit two separate models as we did here using **lm** and **aov**. You only need to use one of them, then use **summary.aov** to produce an Anova table or **summary.lm** to produce a list of parameter estimates and standard errors. Try this:

```
summary.aov(model)
summary.lm(model)
```

Ancova with different values of the covariates

The calculations are more complicated if the different factor levels are associated with different values of the covariate. In the last example, all the genotypes were exposed to exactly the same photoperiods, which made the calculations much more straightforward. The next worked example concerns an experiment on the impact of grazing on the seed production of a biennial plant. Forty plants were allocated to two treatments, grazed and ungrazed, and the grazed plants were exposed to rabbits during the first two weeks of stem elongation. They were then protected from subsequent grazing by the erection of a fence and allowed to regrow. Because initial plant size was thought likely to influence fruit production, the diameter of the top of the rootstock was measured before each plant was potted up. At the end of the growing season, the fruit production (dry wt, mg) was recorded on each of the 40 plants, and this forms the response variable in the following analysis.

Ungrazed plants:

Fruit	59.77	60.98	14.73	19.28	34.25	35.53	87.73	63.21	24.25
Roots	6.225	6.487	4.919	5.130	5.417	5.359	7.614	6.352	4.975

Fruit	64.34	52.92	32.35	53.61	54.86	64.81	73.24	80.64	18.89
Roots	6.930	6.248	5.451	6.013	5.928	6.264	7.181	7.001	4.426

Fruit	75.49	46.73
Roots	7.302	5.836

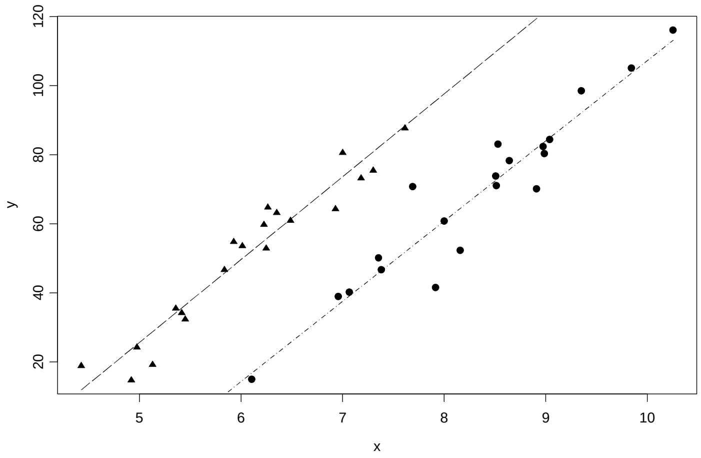
Grazed plants:

Fruit	80.31	82.35	105.1	73.79	50.08	78.28	41.48	98.47	40.15
Roots	8.988	8.975	9.844	8.508	7.354	8.643	7.916	9.351	7.066

Fruit	116.1	38.94	60.77	84.37	70.11	14.95	70.70	71.01	83.03
Roots	10.25	6.958	8.001	9.039	8.910	6.106	7.691	8.515	8.530

Fruit	52.26	46.64
Roots	8.158	7.382

The object of the exercise is to estimate the parameters of the minimal adequate model for these data: here is a plot of the final result so that you can see where we are heading: the triangles are ungrazed plants and the circles are grazed plants.



We start by working out the sums, sums of squares and sums of products for the whole data set combined (40 pairs of numbers), and then for each treatment separately (20 pairs of numbers).The data frame is called ipomopsis.txt

```
ipomopsis<-read.table("c:\\temp\\ipomopsis.txt",header=T)
attach(ipomopsis)
names(ipomopsis)
```

```
[1] "Root"    "Fruit"   "Grazing"
```

First, we’ll work out the overall totals based on all 40 data points:

```
sum(Root);sum(Root^2)
```

```
[1] 287.246
[1] 2148.172
```

Start to fill in the table of totals (it helps to be really well organised for these calculations). Check to see where (and why) the sum (287.246) and the sum of squares (2148.172) of the root diameters (the *x* values) have gone in the table:

```
sum(Fruit);sum(Fruit^2)
```

```
[1] 2376.42
[1] 164928.1
```

```
sum(Root*Fruit)
```

```
[1] 18263.16
```

That completes the overall data summary. Now we select only the “Grazed” plant data:

```
sum(Root[Grazing=="Grazed"]);sum(Root[Grazing=="Grazed"]^2)
```

```
[1] 166.188
[1] 1400.834
```

Make sure you see where (and why) these totals go where they go. Now the “Ungrazed” subtotals

```
sum(Root[Grazing=="Ungrazed"]);sum(Root[Grazing=="Ungrazed"]^2)
```

```
[1] 121.058
[1] 747.3387
```

Now for the data on Fruits for the “Grazed” plants:

```
sum(Fruit[Grazing=="Grazed"]);sum(Fruit[Grazing=="Grazed"]^2)
```

```
[1] 1358.81
[1] 104156.0
```

And the Fruit data from the “Ungrazed” plants:

```
sum(Fruit[Grazing=="Ungrazed"]);sum(Fruit[Grazing=="Ungrazed"]^2)
```

```
[1] 1017.61
[1] 60772.11
```

Finally we want the data for the sums of products: first for the “Grazed” plants:

```
sum(Root[Grazing=="Grazed"]*Fruit[Grazing=="Grazed"])
```

```
[1] 11753.64
```

and for the “Ungrazed” plants:

```
sum(Root[Grazing=="Ungrazed"]*Fruit[Grazing=="Ungrazed"])
```

```
[1] 6509.522
```

	$\sum$ sums	$\sum^{2}$ products	$\sum\sum$ totals
x Ungrazed	121.058	747.3387	
y Ungrazed	1017.61	60772.11	
xy Ungrazed		6509.522	
$\sum x1 \sum y1$			123189.8
x Grazed	166.188	1400.834	
y Grazed	1358.81	104156.0	
xy Grazed		11753.64	
$\sum x2 \sum y2$			225817.9
x overall	287.246	2148.172	
y overall	2376.42	164928.1	
xy overall		18263.16	
$\sum x \sum y$			682617.14

Now we have all of the information necessary to carry out the calculations of the corrected sums of squares and products, SSY, SSX and SSXY for the whole data set (*n* = 40) and for the two separate treatments (20 reps in each).

To get the right answer you will need to be extremely methodical, but there is nothing mysterious or difficult about the process. First, calculate the regression statistics for the whole experiment, ignoring the grazing treatment.

The famous 5 are

```
sum(Root)   sum(Root^2)   sum(Fruit)   sum(Fruit^2)   sum(Root*Fruit)
 287.246     2148.172     2376.42       164928.1          18263.16
```

so

$$SST = 164928.1 - \frac{2376.42^{2}}{40} = 23743.84$$

$$SSX = 2148.172 - \frac{287.246^{2}}{40} = 85.4158$$

$$SSXY = 18263.16 - \frac{287.246 \times 2376.42}{40} = 1197.731$$

$$SSR = \frac{(1197.731)^{2}}{85.4158} = 16795$$

$$SSE = 23743.84 - 16795 = 6948.835$$

The effect of differences between the two grazing treatments, SSA, is

$$SSA = \frac{1358.81^2 + 1017.61^2}{20} - \frac{2376.42^2}{40} = 2910.436$$

Next calculate the regression statistics for each of the grazing treatments separately. First, for the grazed plants:

$$SST_g = 104156 - \frac{1358.81^2}{20} = 11837.79$$

$$SSX_g = 1400.834 - \frac{166.188^2}{20} = 19.9111$$

$$SSXY_g = 11753.64 - \frac{1358.81 \times 166.188}{20} = 462.7415$$

$$SSR_g = \frac{(462.7415)^2}{19.9111} = 10754.29$$

$$SSE_g = 11837.79 - 10754.29 = 1083.509$$

so the slope of the graph of Fruit against Root for the grazed plants is given by

$$b_g = \frac{SSXY_g}{SSX_g} = \frac{462.7415}{19.9111} = 23.240$$

Now for the ungrazed plants:

$$SST_u = 60772.11 - \frac{1017.61^2}{20} = 8995.606$$

$$SSX_u = 747.3387 - \frac{121.058^2}{20} = 14.58677$$

$$SSXY_u = 6509.522 - \frac{121.058 \times 1017.61}{20} = 350.0302$$

$$SSR_u = \frac{(350.0302^2}{14.58677} = 8399.466$$

$$SSE_u = 8995.606 - 8399.466 = 596.1403$$

so the slope of the graph of Fruit against Root for the ungrazed plants is given by

$$b_u = \frac{SSXY_u}{SSX_u} = \frac{350.0302}{14.58677} = 23.996$$

Now add up the regression statistics across the factor levels (grazed and ungrazed):

$$SST_{g+u} = 11837.79 + 8995.606 = 20833.4$$

$$SSX_{g+u} = 19.9111 + 14.58677 = 34.49788$$

$$SSXY_{g+u} = 462.7415 + 350.0302 = 812.7717$$

$$SSR_{g+u} = 10754.29 + 8399.436 = 19153.75$$

$$SSE_{g+u} = 1083.509 + 596.1403 = 1684.461$$

The SSR for a model with a single common slope is given by

$$SSR = \frac{(SSXY_{g+u})^2}{SSX_{g+u}} = \frac{812.7717^2}{34.49788} = 19148.94$$

and the value of the single common slope is

$$b = \frac{SSXY_{g+u}}{SSX_{g+u}} = \frac{812.7717}{34.49788} = 23.560$$

The difference between the two estimates of SSR (SSR_{diff} = 19153.75 – 19148.94 = 4.81) is a measure of the significance of the difference between the two slopes estimated separately for each factor level. Finally, SSE is calculated by difference:

$$SSE = SST - SSA - SSR - SSR_{diff}$$

$$SSE = 23743.84 - 2910.44 - 19148.94 - 4.81 = 1679.65$$

Now we can complete the Anova table for the full model:

Source	SS	d.f.	MS	F
Grazing	2910.44	1		
Root	19148.94	1		
Different slopes	4.81	1	4.81	n.s.
Error	1679.65	36	46.66	
Total	23743.84	39		

Degrees of freedom for error are 40 – 4 = 36 because we have estimated 4 parameters from the data: 2 slopes and 2 intercepts. So the error variance is 46.66 (= SSE/36). The difference between the slopes is clearly not significant (F = 4.81/46.66 = 0.10) so we can fit a simpler model with a common slope of 23.56. The sum of squares for differences between the slopes (4.81) now becomes part of the error sum of squares:

Source	SS	d.f.	MS	F
Grazing	2910.44	1	2910.44	63.9291
Root	19148.94	1	19148.94	420.6156
Error	1684.46	37	45.526	
Total	23743.84	39		

This is the minimal adequate model. Both of the terms are highly significant and there are no redundant factor levels. The next step is to calculate the intercepts for the two parallel regression lines. This is done exactly as before, by rearranging the equation of the straight line to obtain *a* = *y* - *bx*. For each line we can use the mean values of *x* and *y*, with the common slope in each case. Thus:

$$a_1 = \overline{Y_1} - b\overline{X_1} = 50.88 - 23.56 \times 6.0529 = -91.7261$$
$$a_2 = \overline{Y_2} - b\overline{X_2} = 67.94 - 23.56 \times 8.309 = -127.8294$$

This demonstrates that the grazed plants produce, on average, 36.1 fruits *fewer* than the ungrazed plants (127.83 - 91.73). Finally, we need to calculate the standard errors for the common regression slope and for the difference in mean fecundity between the treatments, based on the error variance in the minimal adequate model:

$$s^2 = \frac{1684.46}{37} = 45.526$$

The standard errors are obtained as follows. The standard error of the common slope is found like this:

$$SE_b = \sqrt{\frac{s^2}{SSX}} = \sqrt{\frac{45.526}{19.9111 + 14.45667}} = 1.149$$

The standard error of the intercept of the regression for the grazed treatment (mean root size = 8.3094) is found as follows:

$$SE_a = \sqrt{s^2\left[\frac{1}{n}+\frac{(0-\bar{x})^2}{SSX}\right]} = \sqrt{45.526\left[\frac{1}{20}+\frac{8.3094^2}{34.498}\right]} = 9.664$$

It is clear that the intercept of –127.829 is very significantly less than zero (t = 127.829/9.664 = 13.2), suggesting that there is a threshold rootstock size before reproduction can begin. Finally, the standard error of the difference between the elevations of the two lines (the grazing effect) is given by

$$SE_{\hat{y}_1-\hat{y}_2} = \sqrt{s^2\left[\frac{2}{n}+\frac{(\overline{x_1}-\overline{x_2})^2}{SSX}\right]}$$

which, substituting the values for the error variance and the mean rootstock sizes of the plants in the two treatments, becomes:

$$SE_{\hat{y}_1-\hat{y}_2} = \sqrt{45.526\left[\frac{2}{20}+\frac{(6.0529-8.3094)^2}{34.498}\right]} = 3.357$$

This suggests that any lines differing in elevation by more than about 2 × 3.357 = 6.66 mg dry weight would be regarded as significantly different. Thus, the present difference of 36.09 clearly represents a highly significant reduction in fecundity caused by grazing (t = 10.83).

Ancova in R using lm

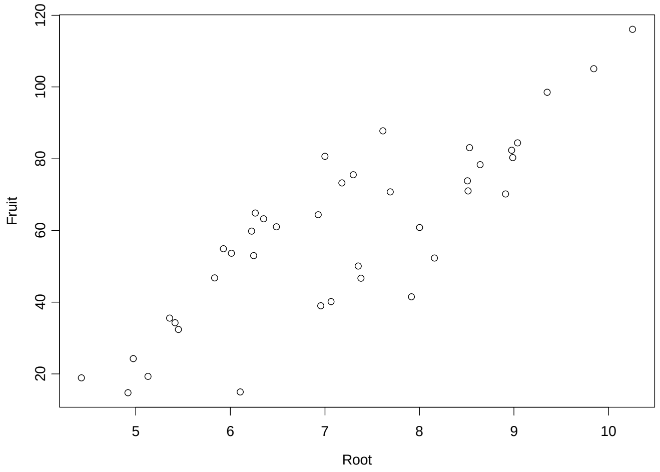
We now repeat the analysis using **lm**. The response variable is fecundity, and there is one experimental factor (grazing) with two levels (ungrazed and grazed) and one covariate (initial root stock diameter). There are 40 values for each of these variables. The odd thing about these data is that grazing seems to *increase* fruit production, a highly counter intuitive result:

```
tapply(Fruit,Grazing, mean)
```

```
 Grazed Ungrazed
 67.9405  50.8805
```

How could this have come about ? We begin by inspecting the data:

```
plot(Root,Fruit)
```



This demonstrates clearly that size matters: the plants that had larger rootstocks at the beginning produced more fruit when they matured. This plot does not help with the real question, however, which is “how did grazing affect fruit production” ? What we need to do is to plot the data separately for the grazed and ungrazed plants. First we plot blank axes:

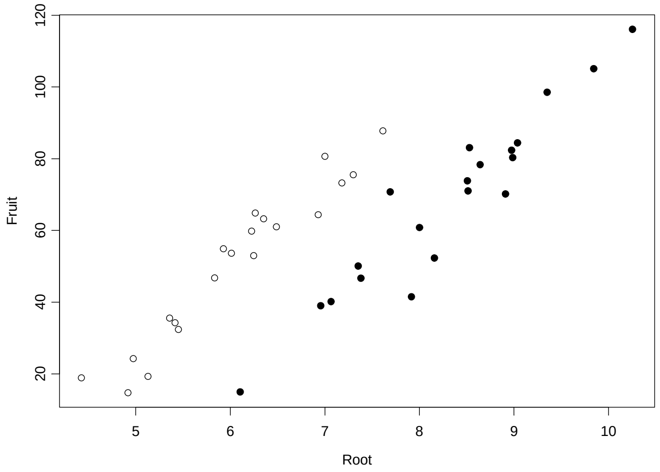
```
plot(Root,Fruit,type="n")
```

Next we select only those points that refer to Ungrazed plants, and add them using **points**

```
points(Root[Grazing=="Ungrazed"],Fruit[Grazing=="Ungrazed"])
```

Now, using a different plotting symbol (the filled circle, pch=16), we add the points that relate to the Grazed plants (use the Up Arrow to edit the last line):

```
points(Root[Grazing=="Grazed"],Fruit[Grazing=="Grazed"],pch=16)
```



This shows a striking pattern which suggests that the largest plants were allocated to the grazed treatments (the filled symbols). But the plot also indicates that for a *given* rootstock diameter (say 7 mm) the grazed plants produced *fewer* fruits than the ungrazed plants (not more, as a simple comparison of the means suggested). This is an excellent example of where analysis of covariance comes into its own. Here, the correct analysis using Ancova completely *reverses* our interpretation of the data.

The analysis proceeds in the following way. We fit the most complicated model first, then simplify it by removing non-significant terms until we are left with a minimal adequate model, in which all the parameters are significantly different from zero. For Ancova, the most complicated model has different slopes and intercepts for each level of the factor. Here we have a 2-level factor (Grazed and Ungrazed) and we are fitting a linear model with 2 parameters ($y = a + bx$) so the most complicated mode has 4 parameters (2 slopes and 2 intercepts). To fit different slopes and intercepts we use the asterisk * notation:

```
ancova<-lm(Fruit~Grazing*Root)
```

You should realise that *order matters*: we would get a different output if the model had been written `Fruit ~ Root * Grazing` (more of this later).

```
summary(ancova)
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -125.173     12.811  -9.771 1.15e-11 ***
GrazingUngrazed        30.806     16.842   1.829   0.0757 .
Root                   23.240      1.531  15.182  < 2e-16 ***
GrazingUngrazed:Root    0.756      2.354   0.321   0.7500

Residual standard error: 6.831 on 36 degrees of freedom
Multiple R-Squared: 0.9293,     Adjusted R-squared: 0.9234
F-statistic: 157.6 on 3 and 36 DF,  p-value: < 2.2e-16
```

This shows that initial root size has a massive effect on fruit production (t = 15.182), but there is no indication of any difference in the slope of this relationship between the two grazing treatments (this is the Grazing by Root interaction with t = 0.321, *p* >> 0.05). The Anova table for the maximal model looks like this (the same as **summary.aov(ancova)**):

```
anova(ancova)
```

```
             Df Sum of Sq  Mean Sq  F Value     Pr(F)
     Grazing  1   2910.44  2910.44  62.3795 0.0000000
        Root  1  19148.94 19148.94 410.4201 0.0000000
Grazing:Root  1      4.81     4.81   0.1031 0.7499503
   Residuals 36   1679.65    46.66
```

The next step is to delete the non-significant interaction term from the model. We can do this manually or automatically: here we'll do both for the purposes of demonstration. The directive for manual model simplification is **update**. We update the current model (here called ancova) by deleting terms from it. The syntax is important: the punctuation reads "comma tilde dot minus" . We define a new name for the simplified model like ancova2

```
ancova2<-update(ancova, ~ . - Grazing:Root)
```

Now we compare the simplified model with just 3 parameters (1 slope and 2 intercepts) with the maximal model using **anova** like this:

```
anova(ancova,ancova2)
```

```
Analysis of Variance Table

Model 1: Fruit ~ Grazing * Root
Model 2: Fruit ~ Grazing + Root

  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     36 1679.65
2     37 1684.46 -1     -4.81 0.1031   0.75
```

This says that model simplification was justified because it caused a negligible reduction in the explanatory power of the model (p = 0.75; to retain the interaction term in the model we would need $p < 0.05$). The next step in model simplification involves testing whether or not grazing had a significant effect on fruit production once we control for initial root size. The procedure is similar: we make a new model name, say ancova3, and use **update** to remove Grazing from ancova2 like this:

```
ancova3<-update(ancova2, ~ . - Grazing)
```

Now we compare the two models using **anova**:

```
anova(ancova2,ancova3)
```

```
Analysis of Variance Table

Model 1: Fruit ~ Grazing + Root
Model 2: Fruit ~ Root

  Res.Df Res.Sum Sq Df  Sum Sq F value     Pr(>F)
1     37     1684.5
2     38     6948.8 -1 -5264.4  115.63 6.107e-013 ***
```

This model simplification is a step too far. Removing the Grazing term causes a massive reduction in the explanatory power of the model, with an F value of 115.63 and a vanishingly small *p* value. The effect of grazing in reducing fruit production is highly significant and needs to be retained in the model. Thus ancova2 is our minimal adequate model, and we should look at its summary table to compare with our earlier calculations carried out by hand:

```
summary(ancova2)
```

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

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
```

You know when you have got the minimal adequate model, because every row of the coefficients table has one or more significance stars (there are 3 stars in this case, because the effects are all so strong). Unfortunately, the interpretation is not crystal clear from this table because *the variable name, not the level name* appears in the first column, and you might misread this as saying that “Grazing is associated with + 36.103 mg of Fruit production”. This is wrong. It is *the second level of Grazing* (which is “Ungrazed” in the present case) that is associated with this positive difference in intercepts. For a given root size, the Grazed plants (factor level 1) produce 36.103 mg of fruit *less* than the Ungrazed plants (factor level 2). R arranges the factor levels in alphabetical order unless you tell it to do otherwise.

```
anova(ancova2)
```

```
          Df Sum of Sq  Mean Sq  F Value         Pr(F)
  Grazing  1   2910.44  2910.44  63.9291 1.397196e-009
     Root  1  19148.94 19148.94 420.6156 0.000000e+000
Residuals 37   1684.46    45.53
```

These are the values we obtained longhand on p. 152. Now we repeat the model simplification using the automatic model-simplification directive called **step**. It couldn’t be easier to use. The full model is called ancova:

```
step(ancova)
```

This directive causes all the terms to be tested to see whether they are needed in the minimal adequate model. The criterion used is AIC, the Akaike Information Criterion (more of which later). In the jargon, this is a “penalised log likelihood”. What this means in simple terms is that it weighs up the inevitable trade off between degrees of freedom and fit of the model. You can have a perfect fit if you have a parameter for every data point, but this model has zero explanatory power. Thus ***deviance goes down as degrees of freedom in the model goes up***. The AIC adds 2 times the number of parameters in the model to the deviance (to penalise it). Deviance, you will recall, is twice the log likelihood of the current model.

Anyway, AIC is a measure of lack of fit; big AIC is bad, small AIC is good. The full model (4 parameters, 2 slopes and 2 intercepts) is fitted first, and AIC calculated as 157.5

```
Start:  AIC= 157.5
 Fruit ~ Grazing + Root + Grazing:Root
```

Then **step** tries removing the most complicated term (the Grazing by Root interaction)

```
               Df Sum of Sq     RSS     AIC
- Grazing:Root  1      4.81 1684.46  155.61
<none>                      1679.65  157.50
```

```
Step:  AIC= 155.61
 Fruit ~ Grazing + Root
```

This has reduced AIC to 155.61 (an improvement, so the simplification is justified)

```
          Df Sum of Sq     RSS     AIC
<none>                  1684.5   155.6
```

```
- Grazing  1    5264.4  6948.8   210.3
- Root     1   19148.9 20833.4   254.2

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

Coefficients:
(Intercept)      Grazing         Root
    -127.83        36.10        23.56
```

No further simplification is possible (as we saw when we used **update** to remove the Grazing term from the model) because AIC goes up to 210.3 when Grazing is removed and up to 254.2 if Root size is removed. Thus, **step** has found the minimal adequate model (it doesn’t always, as we shall see later; it is “good but not perfect”). Again, with this form of output it is possible to misinterpret the result of Grazing; it would be more informative to have the intercepts labelled by the factor levels Grazed and Ungrazed, as when we used **tapply** at the beginning of the exercise.

Drawing the regression lines through the scatterplot following Ancova

You should already have the scatterplot on the graphics window. All you need to do now is to use abline to draw the two lines (one for the grazed plants and one for the ungrazed). The first argument to abline is the intercept (-127.829 for the grazed plants) and the second is the slope (23.56). The parameter values are in the model called ancova2:

```
summary(ancova2)
```

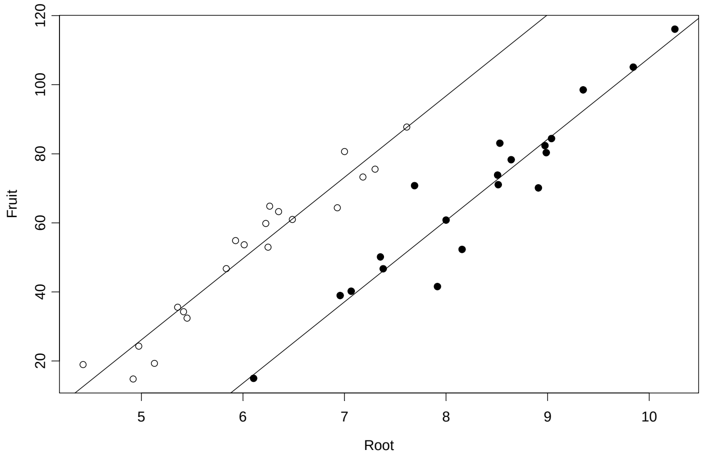
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -127.829      9.664  -13.23 1.35e-15 ***
GrazingUngrazed   36.103      3.357   10.75 6.11e-13 ***
Root              23.560      1.149   20.51  < 2e-16 ***
```

So, to get the first line, we write

```
abline(-127.829 , 23.56)
```

The second line is obtained simply by adding the difference between the intercepts (the ungrazed plants have 36.1032 mg more fruit than the grazed plants): use the Up arrow to edit the last command

```
abline(-127.829+36.1032 , 23.56)
```



Ancova and experimental design

There is an extremely important general message in this example for experimental design. No matter how carefully we randomise at the outset, our experimental groups are likely to be heterogeneous. Sometimes, as in this case, we may have made initial measurements that we can use as covariates later on, but this will not always be the case. There are bound to be important factors that we did not measure. If we had not measured initial root size in this example, we would have come to entirely the wrong conclusion about the impact of grazing on plant performance.

A far better design for this experiment would have been to measure the root stock diameters of all the plants at the beginning of the experiment (as was done here), but then to place the plants in matched pairs with similar sized rootstocks. Then, one of the plants is picked at random and allocated to one of the two grazing treatments (e.g. by tossing a coin); the other plant of the pair then receives the unallocated gazing treatment. Under this scheme, the size ranges of the two treatments would overlap, and the analysis of covariance would be unnecessary.

A more complex Ancova: 2 factors and 1 continuous covariate

This experiment with Weight as the response variable involved Genotype and Sex as two categorical explanatory variables and Age as a continuous covariate. There are 6 levels of Genotype and 2 levels of Sex.

```
gain<-read.table("c:\\temp\\gain.txt",header=T)
attach(gain)
names(gain)
```

```
[1] "Weight"   "Sex"      "Age"      "Genotype" "Score"
```

We begin by fitting the maximal model with its 24 parameters; different slopes and intercepts for every one of the 12 combinations of Sex (2) and Genotype (6).

```
m1<-lm(Weight~Sex*Age*Genotype)
```

```
summary(m1)
```

```
Coefficients:
                       Estimate Std. Error t value  Pr(>|t|)
(Intercept)             7.80053    0.24941  31.276  < 2e-016 ***
Sex                    -0.51966    0.35272  -1.473   0.14936
Age                     0.34950    0.07520   4.648 4.39e-005 ***
GenotypeCloneB          1.19870    0.35272   3.398   0.00167 **
GenotypeCloneC         -0.41751    0.35272  -1.184   0.24429
GenotypeCloneD          0.95600    0.35272   2.710   0.01023 *
GenotypeCloneE         -0.81604    0.35272  -2.314   0.02651 *
GenotypeCloneF          1.66851    0.35272   4.730 3.41e-005 ***
Sex.Age                -0.11283    0.10635  -1.061   0.29579
Sex.GenotypeCloneB     -0.31716    0.49882  -0.636   0.52891
Sex.GenotypeCloneC     -1.06234    0.49882  -2.130   0.04010 *
Sex.GenotypeCloneD     -0.73547    0.49882  -1.474   0.14906
Sex.GenotypeCloneE     -0.28533    0.49882  -0.572   0.57087
Sex.GenotypeCloneF     -0.19839    0.49882  -0.398   0.69319
Age.GenotypeCloneB     -0.10146    0.10635  -0.954   0.34643
Age.GenotypeCloneC     -0.20825    0.10635  -1.958   0.05799 .
Age.GenotypeCloneD     -0.01757    0.10635  -0.165   0.86970
Age.GenotypeCloneE     -0.03825    0.10635  -0.360   0.72123
Age.GenotypeCloneF     -0.05512    0.10635  -0.518   0.60743
Sex.Age.GenotypeCloneB  0.15469    0.15040   1.029   0.31055
Sex.Age.GenotypeCloneC  0.35322    0.15040   2.349   0.02446 *
Sex.Age.GenotypeCloneD  0.19227    0.15040   1.278   0.20929
Sex.Age.GenotypeCloneE  0.13203    0.15040   0.878   0.38585
Sex.Age.GenotypeCloneF  0.08709    0.15040   0.579   0.56616

Residual standard error: 0.2378 on 36 degrees of freedom
Multiple R-Squared: 0.9742,     Adjusted R-squared: 0.9577
F-statistic: 59.06 on 23 and 36 degrees of freedom,
p-value:     0
```

Model simplification

There are one or two significant parameters, but it is not at all clear that the 3-way or 2-way interactions need to be retained in the model. As a first pass, let's use **step** to see how far it gets with model simplification:

```
step(m1)
```

```
Start:  AIC= -155.01
 Weight ~ Sex + Age + Genotype + Sex:Age + Sex:Genotype +
Age:Genotype +  Sex:Age:Genotype
```

```
                   Df Sum of Sq      RSS      AIC
- Sex:Age:Genotype  5     0.349    2.385 -155.511
<none>                             2.036 -155.007
```

It definitely doesn’t need the 3-way interaction, despite the effect of Sex.Age.GenotypeCloneC which gave a significant t-test on its own. How about the 3 2-way interactions ?

```
Step:  AIC= -155.51
 Weight ~ Sex + Age + Genotype + Sex:Age + Sex:Genotype + 
Age:Genotype


               Df Sum of Sq      RSS      AIC
- Sex:Genotype  5     0.147    2.532 -161.924
- Age:Genotype  5     0.168    2.553 -161.423
- Sex:Age       1     0.049    2.434 -156.292
<none>                         2.385 -155.511
```

It has left out Sex by Genotype and now assesses the other two:

```
Step:  AIC= -161.92
 Weight ~ Sex + Age + Genotype + Sex:Age + Age:Genotype

               Df Sum of Sq      RSS      AIC
- Age:Genotype  5     0.168    2.700 -168.066
- Sex:Age       1     0.049    2.581 -162.776
<none>                         2.532 -161.924
```

No need for Age by Genotype. Try removing Sex by Age:

```
Step:  AIC= -168.07
 Weight ~ Sex + Age + Genotype + Sex:Age

           Df Sum of Sq      RSS      AIC
- Sex:Age   1     0.049    2.749 -168.989
<none>                     2.700 -168.066
- Genotype  5    54.958   57.658    5.612
```

Nothing. What about the main effects?

```
Step:  AIC= -168.99
 Weight ~ Sex + Age + Genotype

           Df Sum of Sq      RSS      AIC
<none>                     2.749 -168.989
```

```
- Sex       1    10.374   13.122  -77.201
- Age       1    10.770   13.519  -75.415
- Genotype  5    54.958   57.707    3.662
```

They are all highly significant. This is R's idea of the minimal adequate model. Three main effects but no interactions. That is to say, that the slope of the graph of weight gain against age does not vary with sex or genotype, but the intercepts *do* vary.

It would be a good idea to look at the Anova table for this model:

```
m2<-aov(Weight~Sex+Age+Genotype)

summary(m2)
```

```
Coefficients:

            Df Sum Sq Mean Sq F value     Pr(>F)
Sex          1 10.374  10.374  196.23 < 2.2e-016 ***
Age          1 10.770  10.770  203.73 < 2.2e-016 ***
Genotype     5 54.958  10.992  207.93 < 2.2e-016 ***
Residuals   52  2.749   0.053
```

That certainly looks pretty convincing. What does **lm** produce ?

```
summary.lm(m2)
```

```
Coefficients:

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)     7.93701    0.10066  78.851  < 2e-016 ***
Sex            -0.83161    0.05937 -14.008  < 2e-016 ***
Age             0.29958    0.02099  14.273  < 2e-016 ***
GenotypeCloneB  0.96778    0.10282   9.412 8.07e-013 ***
GenotypeCloneC -1.04361    0.10282 -10.149 6.22e-014 ***
GenotypeCloneD  0.82396    0.10282   8.013 1.21e-010 ***
GenotypeCloneE -0.87540    0.10282  -8.514 1.98e-011 ***
GenotypeCloneF  1.53460    0.10282  14.925  < 2e-016 ***

Residual standard error: 0.2299 on 52 degrees of freedom
Multiple R-Squared: 0.9651,     Adjusted R-squared: 0.9604
F-statistic: 205.7 on 7 and 52 degrees of freedom,
p-value:     0
```

This is where Helmert contrasts would come in handy. Everything is 3-star significantly different from Genotype[1] Sex[1], but it is not obvious that the intercepts for Genotypes B and D need different values (+0.96 and +0.82 above Genotype A with s.e. difference = 0.1028), nor is it obvious that C and E have different intercepts (-1.043 and –0.875).

Perhaps we could reduce the number of factor levels of Genotype from the present 6 to 4 without any loss of explanatory power ?

Factor-level reduction

We create a new categorical variable called newgen with separate levels for clones A and F, and for B & D combined and C & E combined.

```
newgen<-factor(1+(Genotype=="CloneB")+(Genotype=="CloneD")+
   2*(Genotype=="CloneC")+2*(Genotype=="CloneE")+3*(Genotype=="CloneF"))
```

Then we re-do the modelling with newgen (4 levels) instead of Genotype (6 levels)

```
m3<-lm(Weight~Sex+Age+newgen)
```

and check that the simplification was justified

```
anova(m2,m3)
```

Remember:

In ancova,

The intercept belongs to the factor level that comes first in the alphabet

The slope belongs to the first factor level in the alphabet

Categorical effects are differences between intercepts

Interaction terms are differences between slopes