STATISTICS: AN INTRODUCTION USING R

By M.J. Crawley

Exercises

5. ANALYSIS OF VARIANCE

Instead of fitting continuous, measured variables to data (as in regression), many experiments involve exposing experimental material to a range of discrete *levels* of one or more categorical variables known as *factors*. Thus, a factor might be drug treatment for a particular cancer, with 5 levels corresponding to a placebo plus 4 new pharmaceuticals. Alternatively, a factor might be mineral fertilizer, where the 4 levels represented 4 different mixtures of nitrogen, phosphorus and potassium. Factors are often used in experimental designs to represent statistical *blocks*; these are internally homogeneous units in which each of the experimental treatments is repeated. Blocks may be different fields in an agricultural trial, different genotypes in a plant physiology experiment, or different growth chambers in a study of insect photoperiodism.

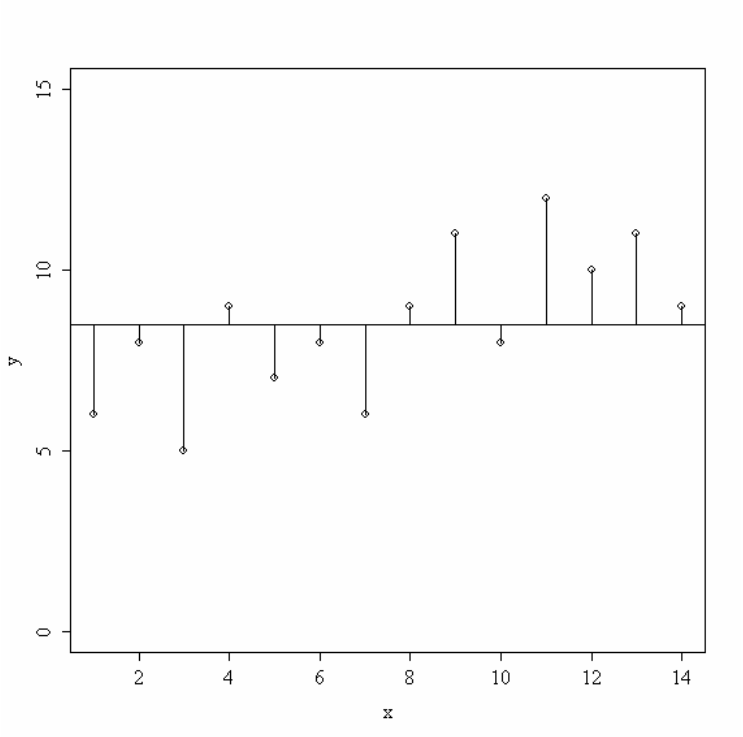
It is important to understand that regression and anova are identical approaches except for the nature of the explanatory variables. For example, it is a small step from having three levels of a shade factor (say light, medium and heavy shade cloths) then carrying out a one-way analysis of variance, to measuring the light intensity in the three treatments and carrying out a regression with light intensity as the explanatory variable. As we shall see later on, some experiments combine regression and analysis of variance by fitting a series of regression lines, one in each of several levels of a given factor (this is called analysis of covariance; see Exercises 6).

Statistical Background

The emphasis in anova has traditionally been on hypothesis testing. The aim of an analysis of variance in R is to estimate means and standard errors of differences between means. Comparing two means by a t-test involved calculating the difference between the two means, dividing by the standard error of the difference, and then comparing the resulting statistic with the value of Student's t from tables (or better still, using **qt** to calculate the critical value; Exercies 3). The means are said to be significantly different when the calculated value of t is larger than the critical value. For large samples (n > 30) a useful rule of thumb is that a t-value greater than 2 is significant. In Analysis of Variance, we are concerned with cases where we want to compare 3 or more means. For the 2-sample case, the t-test and the anova are identical, and the t-test is to be preferred because it is simpler.

It is not at all obvious how you can analyse differences between mean by looking at variances. But this is what analysis of variance is all about. An example should make clear how this works. To keep things simple, suppose we have just two levels of a

single factor. We plot the data in the order in which they were measured: first for the first level of the factor and then for the second level.



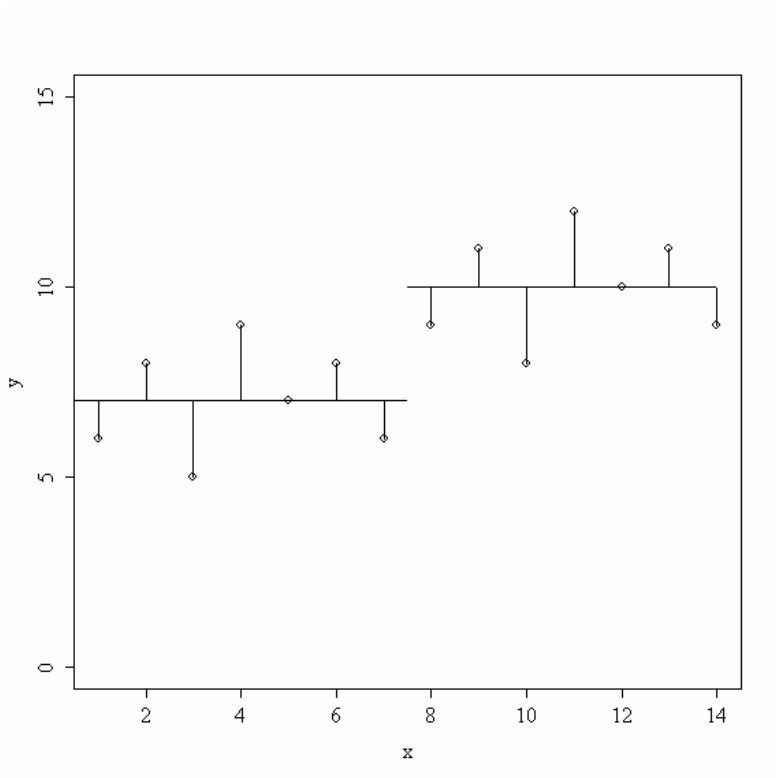
This shows the variation in the data set as a whole. SST, the total variation about the overall mean value of y, is the sum of the squares of these differences:

$$SST = \sum (y - \bar{\bar{y}})^2$$

Next we can fit each of the separate means, $\bar{y}_A$ and $\bar{y}_B$, and consider the sum of squares of the differences between each y value and its own treatment mean. We call this SSE, the error sum of squares, and calculate it like this:

$$SSE = \sum (y_A - \bar{y}_A)^2 + \sum (y_B - \bar{y}_B)^2$$

On the graph, the differences from which SSE is calculated look like this:



Now ask yourself this question. If the two means were the same, what would be the relationship between SSE and SST ? After a moments though you should have been able to convince yourself that if the means are the same, then SSE is the same as SST, because the two horizontal lines in the last plot would be in the same position as the single line in the earlier plot. Now what if the means were significantly different from one another? What would be the relationship between SSE and SST in this case? Which would be the larger? Again, it should not take long for you to see that if the means *are* different, then SSE will be less than SST. Indeed, in the limit, SSE could be zero if the replicates from each treatment fell exactly on their respective means. This is how analysis of variance works. It's amazing but true. *You can make inferences about differences between means by looking at variances* (well, at sums of squares actually, but more of that later).

We can calculate the difference between SST and SSE, and use this as a measure of the difference between the treatment means; this is traditionally called the *treatment sum of squares*, and is denoted by SSA:

$$SSA = SST - SSE$$

The technique we are interested in, however, is analysis of variance, not analysis of sums of squares. We convert the sums of squares into variances by dividing by their degrees of freedom. In our example, there are two levels of the factor and so there is 2-1=1 degree of freedom for SSA. In general, we might have k levels of any factor and hence k-1 d.f. for treatments. If each factor level were replicated n times, then there would be n-1 d.f. for error within each level (we lose one degree of freedom for each individual treatment mean estimated from the data). Since there are k levels, there would be $k(n-1)$ d.f. for error in the whole experiment. The total number of

numbers in the whole experiment is *kn*, so total d.f. is *kn*-1 (the single degree is lost for our estimating the overall mean, $\bar{\bar{y}}$). As a check in more complicated designs, it is useful to make sure that the individual component degrees of freedom add up to the correct total.

$$kn-1 = k-1+k(n-1) = k-1+kn-k = kn-1$$

The divisions for turning the sums of squares into variances are conveniently carried out in an anova table:

Source	SS	d.f.	MS	F	Critical F
Treatment	*SSA*	*k-1*	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{s^2}$	qf(0.95, k-1,k(n-1))
Error	*SSE*	*k(n-1)*	$s^2 \frac{SSE}{k(n-1)}$		
Total	*SST*	*kn-1*			

Each element in the sums of squares column is divided by the number in the adjacent degrees of freedom column to give the variances in the mean square column (headed MS for mean square). The significance of the difference between the means is then assessed using an F test (a variance ratio test). The treatment variance MSA is divided by the error variance, s^2, and the value of this test statistic is compared with the critical value of F using **qf** (the quantiles of the F distribution, with p = 0.95, *k*-1 degrees of freedom in the numerator, and *k*(*n*-1) degrees of freedom in the denominator). If you need to look up the critical value of F in tables, remember that you look up the numerator degrees of freedom (on top of the division) across the *top* of the table, and the denominator degrees of freedom down the rows. If the test statistic is larger than the critical value we *reject* the null hypothesis

$$H_0 : \textit{all the means are the same}$$

and accept the alternative

$$H_1 : \textit{at least one of the means is significantly different from the others}$$

If the test statistic is less than the critical value, then it could have arisen due to chance alone, and so we accept the null hypothesis.

Another way of visualising the process of anova is to think of the relative amounts of sampling variation between replicates receiving the same treatment (i.e. between individual samples in the same level), and between different treatments (i.e. between-level variation). When the variation between replicates within a treatment is large compared to the variation between treatments, we are likely to conclude that the difference between the treatment means is not significant. Only if the variation between replicates within treatments is relatively small compared to the differences between treatments, will we be justified in concluding that the treatment means are significantly different.

Calculations in anova

The definitions of the various sums of squares can now be formalised, and ways found of calculating their values from samples. The total sum of squares, SST is defined as:

$$SST = \sum y^2 - \frac{(\sum y)^2}{kn}$$

just as in regression (see Exercises 4). Note that we divide by the total number of numbers we added together to get $\sum y$ (the grand total of all the y's) which is *kn*. It turns out that the formula that we used to define SSE is rather difficult to calculate (see above), so we calculate the treatment sums of squares SSA, and obtain SSE by difference.

The treatment sum of squares, SSA, is defined as:

$$SSA = \frac{\sum C^2}{n} - \frac{(\sum y)^2}{kn}$$

where the new term is C, the *treatment total*. This is the sum of all the *n* replicates within a given level. Each of the *k* different treatment totals is squared, added up, and then divided by *n* (the number of numbers added together to get the treatment total). The formula is slightly different if there is unequal replication in different treatments, as we shall see below. The meaning of C will become clear when we work through the example later on. Notice the symmetry of the equation. The second term on the right hand side is also divided by the number of numbers that were added together (*kn*) to get the total ($\sum y$) which is squared in the numerator. Finally,

$$SSE = SST - SSA$$

to give all the elements required for completion of the anova table.

Assumptions of anova

You should be aware of the assumptions underlying the analysis of variance. They are all important, but some are more important than others:

- random sampling
- equal variances
- independence of errors
- normal distribution of errors
- additivity of treatment effects

These assumptions are well and good, but what happens if they do not apply to the data you propose to analyse? We consider each case in turn.

Random sampling

If samples are not collected at random, then the experiment is seriously flawed from the outset. The process of randomisation is sometimes immensely tedious, but none the less important for that. Failure to randomise properly often spoils what would otherwise be well-designed ecological experiments (see Hairston 1989 for examples). Unlike the other assumptions of anova, there is nothing we can do to rectify non-random sampling after the event. If the data are not collected randomly they will probably be biased. If they are biased, their interpretation is bound to be equivocal.

Equal variances

It seems odd, at first, that a technique which works by comparing variances should be based on the assumption that variances are equal. What anova actually assumes, however, is that the *sampling errors* do not differ significantly from one treatment to another. The comparative part of anova works by comparing the variances that are due to differences between the treatment means with the variation between samples within treatments. The contributions towards SSA are allowed to vary between treatments, but the contributions towards SSE should not be significantly different.

Recall the folly of attempting to compare treatment means when the variances of the samples are different in the example of three commercial gardens producing lettuce (see Practical 3). There were 3 important points to be made from these calculations:

- two treatments can have different means and the same variance
- two treatments can have the same mean but different variances
- when the variances are different, the fact that the means are identical does *not* mean that the treatments have identical effects.

We often encounter data with non-constant variance, and R is ideally suited to deal with this. With Poisson distributed data, for example, the variance is equal to the mean, and with binomial data the variance (npq) increases to a maximum and then declines with the mean (np). Many data on time to death (or time to failure of a component) exhibit the property that the coefficient of variation is roughly constant, which means that a plot of log(variance) against log(mean) increases with a slope of approximately 2 (this is known as Taylor's Power Law); gamma errors can deal with this. Alternatively, the response variable can be transformed (see Practical 4) to stabilise the variance, and the analysis carried out on the new variable.

Independence of Errors and Pseudoreplication

It is very common to find that the errors are not independent from one experimental unit to another. We know, for example, that the response of an organism is likely to be influenced by its sex, age, body size, neighbours and their sizes, history of development, genotype, and many other things. Differences between our samples in these other attributes would lead to non-independence of errors, and hence to bias in our estimation of the difference that was due to our experimental factor. For instance, if more of the irrigated plots happened to be on good soil, then increases in shoot weight might be attributed to the additional water when, in fact, they were due to the

higher level of mineral nutrient availability on those plots. The ways to minimise the problems associated with non-independence of errors are:

- use block designs, with each treatment combination applied in every block (i.e. repeat the whole experiment in several different places)
- divide the experimental material up into homogeneous groups at the outset (e.g. large, medium and small individuals), then make these groups into statistical blocks, and apply each treatment within each of them
- have high replication
- insist on thorough randomisation.

No matter how careful the design, however, there is always a serious risk of non-independence of errors. A great deal of the skill in designing good experiments is in anticipating where such problems are likely to arise, and in finding ways of blocking the experimental material to maximise the power of the analysis of variance (see Hurlbert, 1984). This topic is considered in more detail later.

Normal distribution of errors

Data often have non-normal error distributions. The commonest form of non-normality is skewness, in which the distribution has a long 'tail' at one side or the other. A great many collections of data are skewed to the right, because most individuals in a population are small, but a few individuals are very large indeed (well out towards the right hand end of the size axis). The second kind of non-normality that is encountered is kurtosis. This may arise because there are long tails on both sides of the mean so that the distribution is more 'pointed' than the bell-shaped normal distribution (so called *leptokurtosis*), or conversely because the distribution is more 'flat-topped' than a normal (so called *platykurtosis*). Finally, data may be bimodal or multi-modal, and look nothing at all like the normal curve. This kind of non-normality is most serious, because skewness and kurtosis are often cured by the same kinds of transformations that can be used to improve homogeneity of variance. If a set of data are strongly bimodal, then it is clear that any estimate of central tendency (mean or median) is likely to be *unrepresentative of most of the individuals* in the population.

Additivity of treatment effects in multi-factor experiments

In 2-way analysis there are two kinds of treatments (say drugs and radiation therapy) each with 2 or more levels. Anova is based on the assumption that the effects of the different treatments are additive. This means that if one of the drug treatments produced a response of 0.4 at low radiation, then the drug will produce the same response at high radiation. Where this is *not* the case, and the response to one factor depends upon the level of another factor (i.e. where there is *statistical interaction*), we must use factorial experiments. Even here, however, the model is still assumed to be additive. When treatments effects are multiplicative (e.g. temperature and dose are multiplicative in some toxicity experiments), then it may be appropriate to transform the data by taking logarithms in order to make the treatment effects additive. In R, we could specify a log link function in order to achieve additivity in a case like this.

A worked example of One-way Analysis of Variance

To draw this background material together, we shall work through an example by hand. In so doing, it will become clear what R is doing during its analysis of the same data. The data come from a simple growth room experiment, in which the response variable is growth (mm) and the categorical explanatory variable is a factor called Photoperiod with 4 levels: Very short, Short, Long and Very long daily exposure to light. There were 6 replicates of each treatment:

V short	Short	Long	V. long
2	3	3	4
3	4	5	6
1	2	1	2
1	1	2	2
2	2	2	2
1	1	2	3

To carry out a 1-way ANOVA we aim to partition the total sum of squares SST into just two parts: variation attributable to differences between the treatment means, SSA, and the unexplained variation which we call SSE, the error sum of squares.

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

$$SSA = \frac{10^2 + 13^2 + 15^2 + 19^2}{6} - \frac{57^2}{24} = 7.125$$

$$SSE = SST - SSA = 39.625 - 7.125 = 32.5$$

Source	SS	d.f.	MS	F
Photoperiod	7.125	3	2.375	1.462
Error	32.5	20	$S^2 = 1.625$	
Total	39.625	23		

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

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

We begin by calculating the mean growth at each photoperiod, using **tapply** in the normal way

```
tapply(Growth,Photoperiod,mean)
```

```
 Long              Short      Very.long      Very.short
  2.5           2.166667       3.166667        1.666667
```

The values look fine, with mean growth increasing with the duration of illumination. But the order in which the factor levels are printed is in alphabetical order. This usually does not matter, but here it is inconvenient because the factor Photoperiod is ordered. We can fix this very simply by declaring photoperiod to be an ordered factor, and providing an ordered list of the factor levels to reflect the fact that Very short < Short < Long < Very long.

```
Photoperiod<-ordered(Photoperiod,levels=c("Very.short","Short","Long","Very.long"))
```

Now when we use **tapply**, the means are printed in the desired sequence:

```
tapply(Growth,Photoperiod,mean)
```

```
 Very.short        Short      Long           Very.long
   1.666667    2.166667       2.5            3.166667
```

The one-way analysis of variance is carried out using the **aov** directive. The model formula is written like this:

Response.variable ~ Explanatory.variable

where it is understood that the explanatory variable is a factor. Although in this case we know that Photoperiod must be a factor because it has text as the factor levels, it is useful to know how to check that a variable is a factor (especially for variables with numbers as factor levels). We use the **is.factor** directive to find this out

```
is.factor(Photoperiod)
```

```
[1] TRUE
```

so Photoperiod *is* a factor. All we need to do now is to give a name to the object that will contain the results of the analysis (one.way), and carry out the **aov**

```
one.way<-aov(Growth~Photoperiod)
```

Nothing happens until we ask for the results. All of the usual generic functions can be used for ANOVA including **summary** and **plot**.

```
summary(one.way)
```

```
                Df  Sum of Sq Mean Sq       F Value      Pr(F)
Photoperiod      3     7.125     2.375      1.461538  0.2550719
  Residuals     20    32.500    1.625
```

It is abundantly clear that there is no significant difference between the mean growth rates with different periods of illumination. An F value of 1.46 will arise due to chance when the means are all the same with a probability greater than 1 in 4 (for significance, you will recall, we want this probability to be less than 0.05). The differences we noted at the beginning are not significant because the variance is so large ($s^2 = 1.625$) and the error degrees of freedom (d.f. = 20) rather small.

Model checking involves **plot**(one.way). You will see that while there is slight evidence of heteroscedasticity (i.e. there is some tendency for the variance to increase as the fitted values increase), but there are strong signs of non-normality in the residuals (J-shape not linear). This is perhaps not surprising given that the data are small integers. We return to these issues later. For the time being, we continue with the analysis of the same data set, but taking further information into account.

Two-way Analysis of Variance

Continuing the analysis from where we left off, we had fitted a 4-level factor for Photoperiod, but the unexplained variation, SSE, was very large (32.5). So large in fact, that the differences between mean growth in the different photoperiods was not significant. Fortunately, the experiment was well designed. Material from 6 plant Genotypes was cloned and Photoperiod treatments were allocated at random to each of 4 clones of the same Genotype. Each Photoperiod was allocated once to each Genotype. There was no replication of Genotypes within Photoperiods, so we can not carry out a Factorial ANOVA. But we can carry out a 2-way ANOVA. The sum of squares attributable to Photoperiod is unchanged, and we can add this directly to our new, expanded ANOVA table. The total sum of squares will also be unchanged:

Source	SS	d.f.	MS	F
Photoperiod	7.125	3		
Genotype		5		
Error		15		
Total	39.625	23		

We know that the sums of squares for Genotype and Error must add up to 32.5, but we need to work out one of them, before we can obtain the other by subtraction. Before we do that, it is worth noting that we can fill in the degrees of freedom column without doing any calculations. There a 6 levels of Genotype so there must be 6-1 = 5 d.f. for Genotype. We know that d.f. must always add up to the total, so

$$\text{error d.f.} = 23 - 5 - 3 = 15$$

There is an analytical way to calculate error d.f.. Technically, in an analysis like this without any replication, the error term is the interaction between the constituent factors. You will recall that interaction degrees of freedom, are the product of the component degrees of freedom. So, in this case with (c-1) = 3 d.f. for Photoperiod and (r-1) = 5 d.f. for Genotype, we have

$$\text{error d.f.} = (r-1)(c-1) = 5x3 = 15$$

The only extra calculation we need to do is for the Genotype sum of squares. The procedure is exactly analogous to computing the Photoperiod sum of squares. Photperiod differences were reflected in the column totals. Likewise Genotype differences are reflected in the row totals:

Genotype	V short	Short	Long	V. long	Row totals
A	2	3	3	4	**12**
B	3	4	5	6	**18**
C	1	2	1	2	**6**
D	1	1	2	2	**6**
E	2	2	2	2	**8**
F	1	1	2	3	**7**

So we use the squares of the row totals in computing SSB. Note that we divide by the number of numbers in a row (4 in this case; 1 for each photoperiod). For SSA we divided by the number of numbers in a column (6 in that case). In general, in ANOVA, you add up the squares of sub totals and *divide by the number of numbers that were added together to get that sub total* . The Genotype sum of squares is therefore calculated as

$$SSB = \frac{12^2 + 18^2 + 6^2 + 6^2 + 8^2 + 7^2}{4} - \frac{57^2}{24} = 27.875$$

and the new, much smaller SSE is obtained by difference

$$SSE = SST - SSA - SSB = 39.625 - 7.125 - 27.875 = 4.625$$

The 2-way ANOVA table can now be completed

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 interpretation of the experiment is completely different. Now we conclude that Photoperiod has a highly significant effect on mean growth ($F = 7.703$, d.f. =3,15, $p<0.01$). The moral is that good experimental design, aimed at reducing variation between individuals (blocking by Genotype in this case), can pay huge dividends in terms of the likelihood of detecting biologically important differences.

We tidy up by removing (**rm**) the variable names used in the previous one-way analysis, and detaching the old dataframe called oneway

```
rm(Growth, Photoperiod)
detach(oneway)
```

Two-way ANOVA in R

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

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

We begin by comparing the mean growth rates of the 6 genotypes, using **tapply**:

```
tapply(Growth,Genotype,mean)
```

```
 A      B        C     D    E        F
 3    4.5      1.5   1.5    2     1.75
```

Evidently, there are substantial differences between the mean growth rates of the different genotypes. This variation was formerly included in the error variance, so by removing the variation attributable to genotype, we will make the error variance smaller, and hence the significance of the difference between the photoperiod means greater. The 2-way ANOVA is carried out simply by specifying the names of two explanatory variables (both must be factors) in the model formula, linked by a plus sign (+). We call the object containing the results of the analysis two.way, and proceed as follows:

```
two.way<-aov(Growth~Genotype+Photoperiod)
```

Nothing happens until we specify the form of output we would like.

```
summary(two.way)
```

```
             Df    Sum of Sq   Mean Sq     F Value          Pr(F)
   Genotype 5     27.875      5.575000    18.08108    0.000007092
Photoperiod 3      7.125      2.375000     7.70270    0.002404262
  Residuals 15     4.625      0.308333
```

The 2-way ANOVA table shows the dramatic effect of including Genotype in the model. Far from being insignificant (as it was in the 1-way ANOVA) the difference between the Photoperiod means is now highly significant ($p < 0.0025$). This example

highlights the enormous potential benefits of introducing blocking into the experimental design. If the photoperiod treatments had been allocated to different genotypes at random, we would not have obtained a significant result. By ensuring that every photoperiod was applied to every genotype, the experimental design was greatly improved. Randomisation was carried out, but it involved deciding at random which of the 4 clonal plants from each genotype was allocated to a given photoperiod treatment.

Model criticism involves using plot(two.way). You will see that the non-normality problems that looked so severe after the 1-way analysis have been completely cured; evidently it was the differences between the genotype means that were the principal cause of the difficulty. If we had not known the identity of the genotypes, and hence been unable to carry out the 2-way analysis, we would have been left with this problem. We return to the issue of unexplained variation under the topic of **overdispersion**, later on.

Interpretation of the parameters of 2-way Anova is described later.

Two-way Factorial Analysis of Variance

The new concept here is the calculation of the *interaction sum of squares*. Like all components of ANOVA it is based on a sum of squared totals, divided by the number of numbers that were added together to get each total. For interactions, the sub totals we need are often not directly available (like the row totals and column totals of the standard 2-way analysis). To make calculation as straightforward as possible, it is a good idea to draw up a table of interaction totals; the levels of Factor A make up the rows and the levels of Factor B make up the columns. Each cell of the table contains the sum of the replicates in each factor combination. We call these totals Q (don't ask me why).

Now the interaction sum of squares, SSAB, is calculated like this:

$$SSAB = \frac{\sum Q^2}{n} - SSA - SSB - CF$$

because the interaction sub totals contain the main effects of Factor A and Factor B as well as the interaction effect we are after. That is why we subtract SSA and SSB in calculating SSAB.

Source	SS	d.f.	MS	F
Factor A	SSA	a-1	$\frac{SSA}{(a-1)}$	$\frac{MSA}{s^2}$
Factor B	SSB	b-1	$\frac{SSB}{(b-1)}$	$\frac{MSB}{s^2}$
Interaction A:B	SSAB	(a-1)(b-1)	$\frac{SSAB}{(a-1)(b-1)}$	$\frac{MSAB}{s^2}$
Error	SSE	ab(n-1)	$s^2 = \frac{SSE}{ab(n-1)}$	
Total	SST	abn-1		

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

```
   growth diet  coat
1     6.6    A light
2     7.2    A light
3     6.9    B light
4     8.3    B light
5     7.9    C light
6     9.2    C light
7     8.3    A  dark
8     8.7    A  dark
9     8.1    B  dark
10    8.5    B  dark
11    9.1    C  dark
12    9.0    C  dark
```

There are 3 levels of diet (A, B and C), 2 levels of coat colour (light and dark) and 2 replicates. The analysis proceeds very simply by using the * operator. This means fit all the main effects and their interactions. Thus the **model formula**

growth ~ diet * coat

is shorthand for

growth ~ diet + coat + diet : coat

where the colon operator means “the interaction between”.

```
attach(factorial)

model<-aov(growth~diet*coat)

summary(model)
```

```
            Df  Sum Sq Mean Sq F value  Pr(>F)
diet         2 2.66000 1.33000  3.6774 0.09069 .
coat         1 2.61333 2.61333  7.2258 0.03614 *
diet:coat    2 0.68667 0.34333  0.9493 0.43833
Residuals    6 2.17000 0.36167
```

The interaction sum of squares is 0.68667, and we should look at how it was obtained. The **interaction totals** are the sums of the 2 replicates in each of the 6 combinations of factor levels:

tapply(growth,list(coat,diet),sum)

```
          A    B    C
dark  17.0 16.6 18.1
light 13.8 15.2 17.1
```

We need the sum of the squares of these subtotals, then we divide by 2 (because each of the subtotals is the sum of 2 numbers).

SSAB<-sum(as.vector(tapply(growth,list(coat,diet),sum))^2)/2

Note the use of the **as.vector** directive to turn the table produced by **tapply** into a set of numbers that we can use in calculations. Now we need to compute the correction factor

CF<-sum(growth)^2/length(growth)

Now, as we saw earlier, the interaction sum of squares is SSAB–SSA-SSB-CF. We can steal the values of SSA and SSB from the anova table above (2.66 and 2.6133)

SSAB-CF-2.66-2.61333

[1] 0.68667

which, to our considerable relief, is the interaction sum of squares as computed by R.

The output of the anova table is interpreted like this. We always start by looking at the interaction rather than the main effects. The F value of 0.9493 falls well short of significance, so there is no evidence at all for an interaction between diet and coat colour. Now for the main effects. There is a significant effect of coat colour ($p < 0.04$) but not of diet ($p > 0.05$). Now we use **plot(model)** to carry out model criticism. The variance is constant and the errors are normal, so that is good.

It is a good idea at this stage to do model simplification to remove any non-significant parameters. We use a new directive **update** to remove the interaction term. It works like this:

model2<-update(model , ~ . - diet:coat)

which is read like this. The new model (called “model2” here) gets an **update** of the earlier model (called “model” in the first argument in update). The syntax is now very

important: it is “comma tilde dot minus”. This says take the whole of “model” (that is the “tilde dot” bit, where dot means “all of”), and remove from it (hence the minus sign) the interaction diet : coat (which is read “diet by coat”). Now we have 2 models: a complicated one (called model) and a simpler one (called model2). R is **brilliant** at comparing models. We use the **anova** directive to compare 2 or more models, like this (make sure you understand the difference between **aov** and **anova**)

```
anova(model,model2)
```

```
Analysis of Variance Table

Model 1: growth ~ diet * coat
Model 2: growth ~ diet + coat

  Res.Df Res.Sum Sq Df   Sum Sq F value Pr(>F)
1      6    2.17000
2      8    2.85667 -2 -0.68667  0.9493 0.4383
```

This says that the simpler *model2* is not significantly different in its explanatory power than the more complicated *model*. In other words, the model simplification is justified. Let’s look at the output of *model2*:

```
summary(model2)
```

```
             Df  Sum Sq Mean Sq F value  Pr(>F)
diet          2 2.66000 1.33000  3.7246 0.07190 .
coat          1 2.61333 2.61333  7.3186 0.02685 *
Residuals     8 2.85667 0.35708
```

There is a hint of an effect of diet (p = 0.0719) but we are ruthless in the elimination of non-significant terms, so we shall try leaving that out as well:

```
model3<-update(model2, ~. -diet)
```

```
anova(model2,model3)
```

```
Analysis of Variance Table

Model 1: growth ~ diet + coat
Model 2: growth ~ coat

  Res.Df Res.Sum Sq Df  Sum Sq F value Pr(>F)
1      8     2.8567
2     10     5.5167 -2 -2.6600  3.7246 0.0719 .
```

This gives us exactly the same p value as in model 2, but when we come onto more complicated statistical models this will not always be the case, and we prefer to compare models by deletion rather than by tests on their parameter values. The simplified model (known as the **minimal adequate model**, because it only contains parameters that are significantly different from zero) looks like this:

```
summary(model3)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
coat         1 2.6133  2.6133  4.7372 0.05457 .
Residuals   10 5.5167  0.5517
```

This is an interesting example of what can happen in model simplification. The effect of coat was highly significant in model 2 but just fails to reach significance when diet is left out of the model. It looks as if either both factors are important, or neither factor is unequivocally important. Only more detailed analysis will tell.

Let's look at the means for the three diets:

```
tapply(growth,diet,mean)
```

```
   A    B    C
7.70 7.95 8.80
```

It looks as if diets A and B produce a rather similar response, but perhaps diet C produces faster growth than the other 2 ? We calculate a new factor, diet2, which is 1 for diets A and B and 2 for diet C:

```
diet2<-factor(1+(diet=="C"))
diet2
```

```
 [1] 1 1 1 1 2 2 1 1 1 1 2 2
Levels:  1 2
```

Make sure you understand what we've done here. All of the diets are represented by a number 1, except for diet C which is represented by a 2 (1+0 = 1, 1+1 = 2 only when it is TRUE that diet equals "C"). Now we try adding this revised dietary factor to the minimal model:

```
model4<-update(model3, ~. +diet2)
```

and see whether it makes a significant difference using **anova**:

```
anova(model3,model4)
```

```
Analysis of Variance Table

Model 1: growth ~ coat
Model 2: growth ~ coat + diet2
  Res.Df Res.Sum Sq Df Sum Sq F value  Pr(>F)
1     10     5.5167
2      9     2.9817  1 2.5350  7.6518 0.02189 *
```

Yes, it does. Saving that extra degree of freedom, by reducing the levels of diet from 3 to 2 has turned a non-significant effect into a significant one. Let's see if there is an interaction with this new, simpler factor:

```
model5<-update(model4, ~. +diet2:coat)
anova(model4,model5)
```

```
Analysis of Variance Table

Model 1: growth ~ coat + diet2
Model 2: growth ~ coat + diet2 + coat:diet2

  Res.Df Res.Sum Sq Df  Sum Sq F value Pr(>F)
1      9    2.98167
2      8    2.70000  1 0.28167  0.8346 0.3877
```

No, there isn't. This means that model4 is the minimal adequate model.

summary(model4)

```
             Df Sum Sq Mean Sq F value  Pr(>F)
coat          1 2.6133  2.6133  7.8882 0.02042 *
diet2         1 2.5350  2.5350  7.6518 0.02189 *
Residuals     9 2.9817  0.3313
```

This example shows the benefits of model simplification. Our interpretation is now much simpler and much clearer than before. In brief, there is a significant effect of coat colour phenotype on mean growth, and there is a significant difference between diet C and the other two diets in mean growth rate. There is no evidence, however, that different coat colours respond to diet C in different ways (i.e. there is no interaction effect). Diagnostic plots on model4 show that the assumption of the anova in regard to normality of errors seems reasonable, but there is a hint of variance declining with the fitted values. Given the low replication, this is not serious.

Three-way Factorial Analysis of Variance

The only new element in a 3-way analysis is the extra weight of calculation involved. The principles are exactly the same as in 2-way factorial ANOVA. We have *a* levels of factor A, *b* levels of factor B and *c* levels of factor C. There are *n* replicates of each treatment combination (unequal replication can be handled, but it makes the formulas unnecessarily complicated, so we consider only the equal replication case in the hand calculations). The main effect totals are denoted by A, B and C respectively. The main effect sums of squares are calculated as in 1-way ANOVA, subtracting CF (the correction factor) from the sums of squares of the treatments totals divided by the relevant replication:

$$CF = \frac{\left[\sum y\right]^2}{abcn}$$

$$SSA = \frac{\sum A^2}{bcn} - CF$$

$$SSB = \frac{\sum B^2}{acn} - CF$$

$$SSC = \frac{\sum C^2}{abn} - CF$$

The three 2-way interactions are calculated in exactly the same way as they were in 2-way ANOVA, based on 2-way interaction tables of subtotals (see above). Only the 3-way formula is given here: this is based on a new set of tables of subtotals: there is a separate AB table for every level of C. The sums of squares of each of the elements, T, of these tables are added together and divided by *n*, the number of numbers added together to obtain each of these subtotals. The 3-way interaction sum of squares is then calculated like this:

$$SSABC = \frac{\sum T^2}{n} - SSA - SSB - SSC - SSAB - SSAC - SSBC - CF$$

The term $\sum T^2 / n$ contains all three 2-way interaction sums of squares as well as all 3 main effects sums of squares, so we need to subtract these in order to obtain the 3-way interaction sum of squares that we are after. Now the error sum of squares is obtained by subtracting all the other component sums of squares from the total sum of squares in the usual way:

$$SSE = SST - SSA - SSB - SSC - SSAB - SAC - SSBC - SSABC$$

and the ANOVA table can be completed.

Source	SS	d.f.	MS	F
Factor A	SSA	a-1	$\frac{SSA}{(a-1)}$	$\frac{MSA}{s^2}$
Factor B	SSB	b-1	$\frac{SSB}{(b-1)}$	$\frac{MSB}{s^2}$
Factor C	SSC	c-1	$\frac{SSC}{(c-1)}$	$\frac{MSC}{s^2}$
Interaction A:B	SSAB	(a-1)(b-1)	$\frac{SSAB}{(a-1)(b-1)}$	$\frac{MSAB}{s^2}$
Interaction A:C	SSAC	(a-1)(c-1)	$\frac{SSAC}{(a-1)(c-1)}$	$\frac{MSAC}{s^2}$
Interaction B:C	SSBC	(b-1)(c-1)	$\frac{SSBC}{(b-1)(c-1)}$	$\frac{MSBC}{s^2}$
Interaction A:B:C	SSABC	(a-1)(b-1)(c-1)	$\frac{SSABC}{(a-1)(b-1)(c-1)}$	$\frac{MSABC}{s^2}$
Error	SSE	abc(n-1)	$s^2 = \frac{SSE}{abc(n-1)}$	
Total	SST	$abcn$-1		

Now, bracing ourselves, we start the calculations by hand. The response variable is the population growth rate of *Daphnia*. There are 3 categorical response variables: Water (with 2 levels from the rivers Tyne and Wear), Detergent (with 4 levels, imaginatively named BrandA, BrandB, BrandC and BrandD) and *Daphnia* clone (with 3 levels, Clone1, Clone2 and Clone3). Each treatment combination was replicated 3 times. So our factor level codings from the equations above are $a = 2$, $b = 4$, $c = 3$ and $n = 3$.

	Growth.rate	Water	Detergent	*Daphnia*
1	2.919086	Tyne	BrandA	Clone1
2	2.492904	Tyne	BrandA	Clone1
3	3.021804	Tyne	BrandA	Clone1
4	2.350874	Tyne	BrandA	Clone2
5	3.148174	Tyne	BrandA	Clone2
6	4.423853	Tyne	BrandA	Clone2
7	4.870959	Tyne	BrandA	Clone3
8	3.897731	Tyne	BrandA	Clone3
9	5.830882	Tyne	BrandA	Clone3
10	2.302717	Tyne	BrandB	Clone1
11	3.195970	Tyne	BrandB	Clone1
12	2.829021	Tyne	BrandB	Clone1
13	3.027500	Tyne	BrandB	Clone2
14	5.285108	Tyne	BrandB	Clone2
15	4.260955	Tyne	BrandB	Clone2
16	4.049528	Tyne	BrandB	Clone3
17	4.623400	Tyne	BrandB	Clone3
18	5.625845	Tyne	BrandB	Clone3
19	2.634182	Tyne	BrandC	Clone1
20	3.513746	Tyne	BrandC	Clone1
21	3.714657	Tyne	BrandC	Clone1
22	3.862106	Tyne	BrandC	Clone2
23	3.955181	Tyne	BrandC	Clone2
24	3.044308	Tyne	BrandC	Clone2
25	3.358051	Tyne	BrandC	Clone3
26	5.633479	Tyne	BrandC	Clone3
27	4.613175	Tyne	BrandC	Clone3
28	2.200593	Tyne	BrandD	Clone1
29	2.233800	Tyne	BrandD	Clone1
30	3.357184	Tyne	BrandD	Clone1
31	3.111764	Tyne	BrandD	Clone2
32	4.842598	Tyne	BrandD	Clone2
33	4.362592	Tyne	BrandD	Clone2
34	2.230210	Tyne	BrandD	Clone3
35	3.104323	Tyne	BrandD	Clone3
36	4.762764	Tyne	BrandD	Clone3

	Growth.rate	Water	Detergent	*Daphnia*
37	2.319406	Wear	BrandA	Clone1
38	2.098191	Wear	BrandA	Clone1
39	3.541969	Wear	BrandA	Clone1
40	3.888784	Wear	BrandA	Clone2
41	3.960038	Wear	BrandA	Clone2
42	5.742290	Wear	BrandA	Clone2
43	5.244230	Wear	BrandA	Clone3
44	4.795048	Wear	BrandA	Clone3
45	5.380755	Wear	BrandA	Clone3
46	3.610532	Wear	BrandB	Clone1
47	2.247651	Wear	BrandB	Clone1
48	3.388947	Wear	BrandB	Clone1
49	4.061630	Wear	BrandB	Clone2
50	5.478983	Wear	BrandB	Clone2
51	4.303407	Wear	BrandB	Clone2
52	3.973060	Wear	BrandB	Clone3
53	4.632224	Wear	BrandB	Clone3
54	5.284316	Wear	BrandB	Clone3
55	2.480984	Wear	BrandC	Clone1
56	3.950710	Wear	BrandC	Clone1
57	2.133731	Wear	BrandC	Clone1
58	5.586468	Wear	BrandC	Clone2
59	6.918344	Wear	BrandC	Clone2
60	5.270421	Wear	BrandC	Clone2
61	2.015707	Wear	BrandC	Clone3
62	3.467042	Wear	BrandC	Clone3
63	5.028927	Wear	BrandC	Clone3
64	2.307174	Wear	BrandD	Clone1
65	2.962274	Wear	BrandD	Clone1
66	2.699762	Wear	BrandD	Clone1
67	6.569412	Wear	BrandD	Clone2
68	6.405130	Wear	BrandD	Clone2
69	5.990973	Wear	BrandD	Clone2
70	2.553241	Wear	BrandD	Clone3
71	2.592766	Wear	BrandD	Clone3
72	1.761603	Wear	BrandD	Clone3

We begin by calculating SST because it is easy and this will make us feel better.

$$CF = \frac{277.3372^2}{2 \times 4 \times 3 \times 3} = \frac{76915.9}{72} = 1068.276$$

$$SST = \sum y^2 - CF = 1185.434 - 1068.276 = 117.1572$$

Now we can work out the 3 main effect sums of squares, using the equations above:

$$SSA = \frac{132.691^2 + 144.6461^2}{4 \times 3 \times 3} - CF = 1.98506$$

$$SSB = \frac{69.927^2 + 72.1808^2 + 71.1812^2 + 64.0482^2}{2 \times 3 \times 3} - CF = 2.21157$$

$$SSC = \frac{68.157^2 + 109.8509^2 + 99.3293^2}{2 \times 4 \times 3} - CF = 39.1777$$

Calculation of the three 2-way interactions requires that we draw up **3 tables of subtotals**: an AB table summing over *Daphnia* Clones and replicates, an AC table summing over Detergents and replicates, and a BC table summing over Water and replicates. The sub totals in each of the 3 tables are squared, added up and divided by the number of numbers that were added together to get each sub total (3x3 = 9, 4x3 = 12 and 2x3 = 6 respectively). Here are the three tables, and the sums of their squares:

	BrandA	BrandB	BrandC	BrandD
Tyne	32.95627	35.20004	34.32889	30.20583
Wear	36.97071	36.98075	36.85233	33.84233

$$\sum T^2 = 9653.83$$

	Clone1	Clone2	Clone3
Tyne	34.41566	45.67501	52.60035
Wear	33.74133	64.17588	46.72892

$$\sum T^2 = 13478.05$$

	Clone1	Clone2	Clone3
BrandA	16.39336	23.51401	30.01961
BrandB	17.57484	26.41758	28.18837
BrandC	18.42801	28.63683	24.11638
BrandD	15.76078	31.28247	17.00491

$$\sum T^2 = 6781.597$$

The 2-way interaction sums of squares now look like this:

$$SSAB = \frac{9653.83}{3 \times 3} - SSA - SSB - CF = 0.17486$$

$$SSAC = \frac{13478.05}{4 \times 3} - SSA - SSC - CF = 13.732$$

$$SSBC = \frac{6781.597}{2 \times 3} - SSB - SSC - CF = 20.6006$$

Calculation of the 3-way interaction sum of squares involves working out all of the individual subtotals for the 2x4x3 = 24 treatment combinations, each of which is the sum of $n = 3$ replicates.

```
8.433794  7.959566  8.327708  9.247130  9.862586
8.565425  7.791576  7.969209  9.922901  13.591112
12.573563 13.844020 10.861595 17.775234 12.316954
18.965514 14.599572 15.420034 14.298773 13.889600
13.604706 10.511676 10.097297 6.907610
```

Now we need to square each of these, add them up and divide by 3

$$SSABC = \frac{34.56.017}{3} - SSA - SSB - SSC - SSAB - SSAC - SSBC - CF = 5.8476$$

The error sum of squares is the last thing we need to calculate:

$$SSE = 117.152 - 1.98506 - 2.21157 - 39.1777 - 0.17486 - 13.732 - 20.6006 - 5.8476 = 33.42776$$

Here is the complete ANOVA table.

Source	SS	d.f.	MS	F
Water	1.985	1	1.985	2.852
Detergent	2.116	3	0.705	1.013
Daphnia	39.178	2	19.589	28.145
Water:Detergent	0.1749	3	0.058	0.083
Water:*Daphnia*	13.732	2	6.866	9.865
Detergent:*Daphnia*	20.6006	6	3.433	4.932
3-way Interaction	5.8476	6	0.975	1.401
Error	33.4278	48	$s^2 = 0.696$	
Total	117.1572	71		

The 3-way interaction is not significant, nor is the Water:Detergent interaction, but both other 2-way interactions Water:*Daphnia* and Detergent :*Daphnia* are highly significant. All three factors therefore appear in at least one significant interactions, so model simplification ends here. The apparently insignificant main effects of Water and Detergent should *not* be interpreted as meaning that these factors have no significant effect on population growth rate. Always check the highest order

interactions first, and stop model simplification as soon as all of the terms have appeared in at least one significant interaction.

We can compare our answers with those produced by the computer.

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

```
[1] "Growth.rate"   "Water"        "Detergent"   "Daphnia"
```

The model formula uses the * notation to specify the fit of all interaction terms. In this case, the model calls for the estimation of 23 parameters, using up 6 degrees of freedom for the 3-way interaction, 11 for the three 2-way interactions and 6 for the 3 main effects.

```
factorial<-aov(Growth.rate~Water*Detergent*Daphnia)
```

```
summary(factorial)
```

```
                          Df     Sum of Sq   Mean Sq    F Value    Pr(F)
                Water     1      1.98506      1.98506  2.85042   0.0978380
             Detergent    3      2.21157      0.73719  1.05855   0.3754783
              Daphnia     2      39.17773    19.58887 28.12828   0.0000000
        Water:Detergent   3      0.17486      0.05829  0.08370   0.9686075
          Water:Daphnia   2      13.73204     6.86602  9.85914   0.0002587
      Detergent:Daphnia   6      20.60056     3.43343  4.93017   0.0005323
Water:Detergent:Daphnia   6      5.84762      0.97460  1.39946   0.2343235
             Residuals    48     33.42776     0.69641
```

As always, the interpretation begins with the highest order interactions. The present data provide no support for the hypothesis that the interaction between Water and *Daphnia* depends upon the kind of Detergent (p=0.23). There is no significant interaction between Water and Detergent, but highly significant two way interactions between Water and *Daphnia* and Detergent and *Daphnia*. All 3 factors appear in one or more significant interactions, so all factors are important. The non-significant main effect for Detergent should ***not***, therefore, be interpreted as meaning that Detergent has no significant effect on *Daphnia* growth rate. The correct interpretation is that the effect of Detergent on *Daphnia* growth rate varies from one *Daphnia* clone to another. Model criticism is carried out using **plot**(factorial).

In order to see how the interaction works, we can use **tapply** to calculate a 2-dimensional table of mean growth rates for each combination of Detergent and *Daphnia* clone (note the use of **list** to get the 2-dimensions we want):

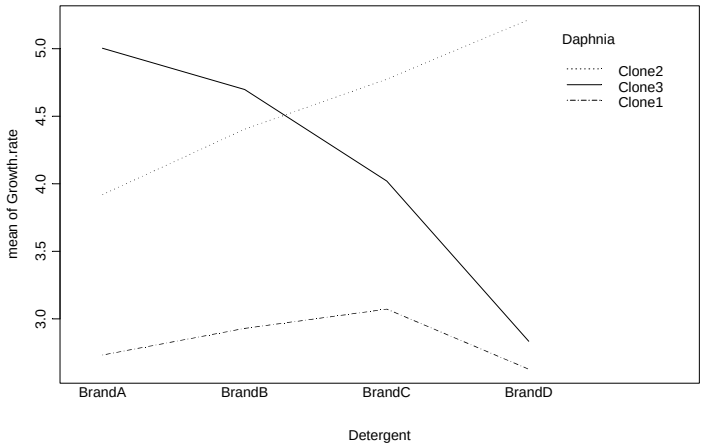
```
tapply(Growth.rate,list(Detergent,Daphnia),mean)
```

	Clone1	Clone2	Clone3
BrandA	2.732227	3.919002	5.003268
BrandB	2.929140	4.402931	4.698062
BrandC	3.071335	4.772805	4.019397
BrandD	2.626797	5.213745	2.834151

The interaction between Detergent and Clone is interpreted as follows. Consider Brand A. Mean growth rate rises from Clone 1 to Clone 2 To Clone 3. Brands B and C show a different pattern, with growth rate not increasing from Clone 2 to Clone 3. Brand D is different again, with growth rate declining steeply from Clone 2 to Clone 3. In general, it is a good idea to produce plots to show these interaction effects.

This is achieved by using the **interaction.plot** directive, like this (note that the response variable is **last** in the list, and that the factor to make the x axis is first in the list):

```
interaction.plot(Detergent,Daphnia,Growth.rate)
```



In general, interactions show up as non-parallelness in the lines of the interaction plot.

- How would you simplify this model ?
- What is the minimal adequate model for these data