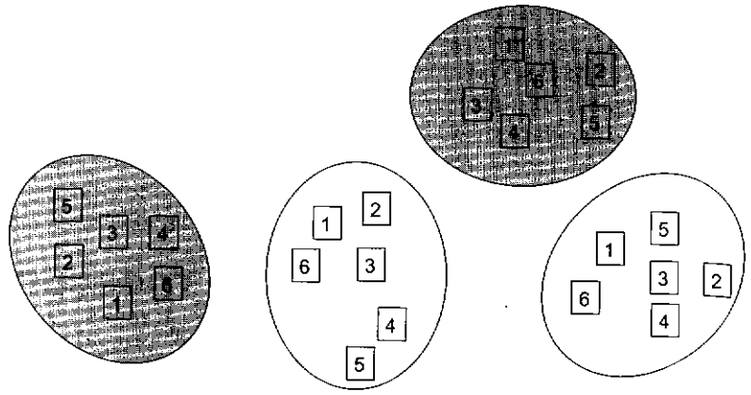

Figure 11.1 Diagrammatic representation of the split-plot experiment from Wissinger *et al.* (1996). There are four ponds (only two shown here) in each of two hydroperiods (permanent and autumnal, represented by different shading), the between plots factor. Within each pond, there were six cages, each containing one level of the within plots factor, competition treatment.

up six wood frame cages in the littoral zone and applied one of six competition treatments (low density *Asynarchus*, low density *Limnephilus*, high density *Asynarchus*, high density *Limnephilus*, high density both species, control with no caddisflies) to each cage within each pond. The role of hydroperiod (permanent or autumnal) was investigated by having four ponds in each category. The response variables were body mass and survival of each species analyzed separately, so there were only three density treatments (those containing the same species). So there are two factors: hydroperiod was "applied" (non-experimentally) to whole ponds (plots) and is termed the between plots factor and density treatment was applied to cages within plots and is termed the within plots factor. Split-plot designs are characterized by having factors applied to experimental units at different, usually spatial, scales.

There are a number of practical design issues for this experiment.

- The experimental design that would be simplest to analyze would be to have whole ponds that are subjected to levels of both factors, hydroperiod and density treatment, forming a completely randomized (CR) factorial arrangement of two hydroperiods by six density treatments with n ponds per cell. Ponds are large units and we would expect considerable variability between them, resulting in large residual variance.
- It is often difficult to install cages, especially large ones. For example, covering whole ponds with cages to maintain experimental densities would be very expensive to set up and probably require an immense amount of labour. We may find that we cannot physically deal with the required size of cages in the time available to set the experiment up, because the research grant has dried up, or we've exhausted the supply of eager volunteers in earlier experiments. We would also need a lot more ponds. The current design uses eight ponds, whereas a completely randomized design with even only two ponds per density and hydroperiod combination would need 24. That many ponds may simply not exist.
- The split-plot design chosen allows us to group our density treatments within ponds, minimizing spatial variation in environmental characteristics, and giving us a clearer test of the effects of density. It also reduces the size of cages. We have, however, linked together groups of cages, and changed our statistical model dramatically compared to the CR design. If anything happens to a pond (e.g. it dries up at the wrong time, or gets an algal bloom), we would be forced to discard all cages in that pond. If we'd used a CR design, we would lose just a single replicate in a cell.

As another example, Leonard *et al.* (1999) tested the prediction that flow had strong effects on the abundances of mussels and barnacles in an estuary but that these effects might vary with tidal height. They had a number of general design options for testing this prediction.

- They could have sampled a range of sites in the estuary. In the simplest case, they could sample replicate sites within combinations of flow regime and tidal height (i.e. a completely randomized factorial design, with two factors, flow regime and tidal height). This approach would require a large number of sites, and it may be difficult to find enough in the estuary.
- It is likely that sites of a given flow regime vary widely, and the researchers would require many replicate sites to get adequate power. They might get less variability if they sampled a given site at different tidal heights, because they could get more similar physical habitats, and make a better test of the effect of height (although the variation between sites will still be a problem for assessing the effects of flow).

Leonard *et al.* (1999) used a split-plot design: plots were sites and they "applied" six tidal heights (from 0.0 to 3.6 m above mean low water) within each site, and each site falls into one of two flow regimes, high and low flow. In analyzing this design, we need to keep the six height observations for each site together, so that we can compare their differences.

Split-plot designs can also be used when the plots or blocks do not obviously represent spatial units of replication. For example, Westly (1993) set up a split-plot experiment to examine the effects of inflorescence bud removal on asexual investment in the Jerusalem artichoke (*Helianthus tuberosus*). There were four populations of *H. tuberosus*, five genotypes (genotypes were actually tubers from single individuals) nested within each population and two treatments (normal flowering and inflorescence removed) applied to different tubers from each genotype. Genotypes were plots, population was the between plots factor and treatment was the within plots factor.

We will illustrate the analysis of split-plot designs in this chapter with a recent example from our own work on disturbances on rocky shores.

Effects of trampling on intertidal algae populations

These data come from a long-term experiment examining the impact of humans on the fauna and flora of rocky shores in southern Australia (Keough & Quinn 1998), and the full analysis is in Box 11.1. In this experiment, we were interested in disturbances caused by pedestrians, and whether a pattern of summer disturbance and autumn–winter–spring recovery results in a series of small disturbance–recovery cycles, or whether repeated disturbances eventually cause a major impact. We manipulated disturbance by trampling on marked intertidal areas each summer, using four different disturbance levels, which were the number of pedestrian passages. To determine the variability in results, we did the experiment on three different rock platforms, separated by hundreds to thousands of meters. On a smaller scale, at each site, we had two experimental plots, separated by tens of meters, and each plot contained eight experimental strips. This arrangement corresponds to a nested design, with sites (i.e. platforms), plots within sites, and strips within each plot.

Box 11.1 Worked example of split-plot design: effects of trampling on intertidal limpet populations

Keough & Quinn (1998) examined the effect of pedestrian traffic (trampling) on the abundance of macroalgae and gastropods on rocky intertidal shores. They used three sites, representing different rock platforms separated by hundreds to thousands of meters. Within each site, there were two experimental plots separated by tens of meters and four levels of trampling intensity (0, 5, 10, 25 pedestrian passages per low tide on 6–8 days each summer) were allocated to each of two strips within each plot in each site. The response variable is the number of limpets per 0.25 m^2 quadrat per strip. With only two replicates of each plot–trampling combination, it is not worth producing boxplots, and the number of limpets did not vary widely, with numbers generally less than ten, and no extreme values. We did not transform

the response variable, a decision that seems reasonable, as the model fitted the data very well (Keough & Quinn 1998, their Table 3). Model 11.2 was fitted, and includes a term for plots within sites × trampling, because we had replicate strips for each trampling treatment in every plot.

In this design, sites and plots are random factors, so you need to be sure that you use correct *F*-ratios (Table 11.3). You might need to recalculate the sites, trampling, and sites × trampling *F*-ratios from the default statistical output if your statistical software does not allow you to specify fixed and random factors.

The specific null hypotheses of interest were as follows.

No difference between sites in the mean number of limpets per strip, pooling trampling treatments.

No difference between trampling treatments in the mean number of limpets per strip, pooling sites.

No interaction between site and trampling treatment on the mean number of limpets per strip, i.e. the effect of trampling on the mean number of limpets per strip was the same at the three sites.

Because we had replicate strips for each trampling treatment within each plot at each site, we could also test two additional null hypotheses.

No added variance in mean number of limpets per strip due to all possible plots within each site.

No interaction between trampling treatment and all possible plots within each site on the mean number of limpets per strip, i.e. the effect of trampling on the mean the number of limpets per strip was the same on all possible plots within each site.

The final ANOVA table is shown below.

Source	SS	df	MS	*F*	*P*	Denominator
Sites	8.719	2	4.359	0.521	0.639	Plots(sites)
Plots(sites)	25.094	3	8.365	5.214	0.006	Residual
Trampling	18.354	3	6.118	5.071	0.044	Site × trampling
Site × trampling	7.240	6	1.207	0.485	0.805	Plots(sites) × trampling
Plots(sites) × trampling	22.406	9	2.490	1.552	0.187	Residual
Residual	38.500	24	1.604			

We would conclude that there is a significant main effect of trampling, and that the effect of trampling on the number of limpets does not vary between sites or plots. There is also significant spatial variation at the scale of plots. The number of limpets rose with the intensity of trampling and Figure 11.3(a) shows similar increases at all three platforms (with data averaged across plots). Trampling appears to benefit limpets! This effect occurred because the species most affected by trampling is a brown alga, *Hormosira banksii*, which forms dense mats on these rock platforms. Dense mats provide a poor habitat for the herbivorous limpets, with little food, so the destruction of these mats generates new, usable habitat for limpets. At the level of plots, we found wide variation in overall abundance of limpets (averaged across trampling levels) (Figure 11.3(b)). The plots with higher numbers were on different platforms (sites), as were those with low numbers, accounting for significant variation among plots, but not sites.

Within each plot, the eight strips were allocated randomly to one of the four trampling levels, with two replicates of each trampling level. With the same disturbance levels applied to all plots and sites, the factor trampling is orthogonal to sites and plots. The data used in this example are from a census of the number of limpets in each strip after three years of trampling.

11.1.2 Repeated measures designs

A simple repeated measures design, where the responses of a number of experimental units (or subjects) are recorded for a number of trials (or times), was discussed in detail in Chapter 10 and was also termed a subject by trials design. A modification of this design is a groups by trials design where the basic repeated measures design is modified to include a treatment structure between subjects, i.e. the subjects are randomly allocated to treatment groups in addition to their responses being recorded on a number of trials or times. Just as the linear model used for a subjects by trials repeated measures designs was the same as that used for a RCB design (an unreplicated two factor ANOVA model), groups by trials repeated measures designs can be analyzed in the same way as classical split-plot designs (with a partly nested ANOVA model). The term "plot" is replaced by "subject", and we simply have "between subjects" and "within subjects" effects in the same way as we had between and within plot effects. In biology and ecology, the "subjects" are experimental or sampling units (animals, plants, quadrats, etc.) and the trials are usually sequential times (von Ende 1993).

The term "repeated measures" has actually been used in a confusing manner in the literature. It really refers to repeated observations made on individual units (e.g. subjects, plots), either sequentially through time or under some treatment structure that is applied sequentially throughout time. Repeated measurements on experimental units can occur in any type of design. For example, a randomized block or split-plot design can have repeated measurements on each experimental unit within each block or plot (Gumpertz & Brownie 1993). The linear models used for repeated measures and split-plot designs are identical. The only complications are in the way the data are coded for computer analysis and which assumptions are applicable.

As an example of a group by trials repeated measures design, Schwartz *et al.* (1995) studied the effects of four temperatures (10°, 20°, 30°, 40°C) on the dark respiration rate of five species of tree (four species of *Torrea* and one species of *Taxus*). Assume that it was desirable to have around five replicates to compare the five tree species, there are not large numbers of plants available, and individual plants were also likely to have different temperature profiles (leading to possibly reduced power). What are the design options for this experiment?

- Five replicate plants per cell, by four temperatures by five species means the experiment would require 100 plants. We would analyze this experiment with a CR two-factor design (factors: species and temperature).
- One temperature profile per plant, so each plant would be used four times for the four temperatures, and only 20 plants are required (five for each species).

The second is a sensible option, to reduce the number of plants used and cut costs (and, if an experiment required sacrificing animals, reducing the number of animals killed). If we choose this option, we don't have a set of independent measurements for each temperature, but a group of five at one temperature, then another group of five for the same set of plants at the next temperature, and so on. Our analysis therefore needs to maintain the relationships between the measurements. Schwartz *et al.* (1995) used this repeated measures design with five or six plants of each species and each plant was subjected to the four temperatures. Individual plants were subjects, species was the between subjects factor and temperature was the within subjects factor. You can see the similarity to the diagrammatic representation of the pond experiment (Wissinger *et al.* 1996): the "plots" are individual plants (in repeated measures designs, these are termed "subjects"), the hydroperiod treatment corresponds to species, and the density treatments correspond to the temperatures. This experiment has one "between subjects" factor (species) and one "within subjects" factor (temperature).

In this example, the within subjects factor is a series of treatments (temperatures) applied sequentially through time. Repeated measures designs are often used when the within subjects factor does not represent different treatments but just a time sequence of interest. For example, Gange (1995) measured aphid abundance on twenty individual trees of two species of alder on twenty consecutive dates between May and September. The response variable was aphid abundance, the between subjects factor was tree species and the within subjects factor was date.

We will illustrate the analysis of groups by trials repeated measures designs with an example from a postgraduate project on physiology of amphibians.

Responses of cane toads to hypoxia

Mullens (1993) investigated the ways that cane toads (*Bufo marinus*) respond to conditions of hypoxia. Toads, the subjects, show two different kinds of breathing patterns, lung or buccal, and this breathing pattern was the between subjects factor. The second factor was O_2 concentration, which had eight levels (0, 5, 10, 15, 20, 30, 40, 50%), and was applied within subjects (toads). Various aspects of breathing rate were measured in each trial. The full analysis of this example is in Box 11.2.

Box 11.2 Worked example of groups by trials repeated measures design: responses of cane toads to hypoxia

Mullens (1993) investigated how the breathing rates of cane toads (*Bufo marinus*) respond to conditions of hypoxia. Toads, the subjects, show two different kinds of breathing patterns, lung or buccal, and this breathing pattern was the between subjects factor. The second factor was O_2 concentration, which had eight levels (0, 5, 10, 15, 20, 30, 40, 50%), and was applied within subjects (toads). The response variable was the frequency of buccal breathing and was transformed to square roots to reduce positive skewness (based on boxplots of the data for each O_2 concentration) and improve variance homogeneity (based on residual plots).

The specific null hypotheses of interest were as follows.

- No difference between breathing types in the mean square root rate of breathing, pooling O_2 levels.
- No difference between O_2 levels in the mean square root rate of breathing, pooling breathing types.
- No interaction between breathing type and O_2 level on the mean square root rate of breathing, i.e. the effect of O_2 level on the mean square root rate of breathing was the same for both breathing types.

With no replicates within each combination of breathing type, toad and O_2 level, we could not test hypotheses about the random factor toads within breathing type or O_2 levels by toads within breathing type.

The data were initially coded in classical split-plot form, where toads were plots, and the model in Equation 11.3 was fitted. Because there is only one replicate observation for each toad for each O_2 concentration, this model is fully saturated, i.e. it fits the data perfectly because all sources of variation have been accounted for. The output from your statistical software usually won't include *F* tests or *P* values. You might just need to specify each effect in the model and its appropriate denominator to get these. In this example, breathing type and oxygen are clearly fixed factors, but toad is random, so breathing type is tested against toad within breathing type

and the interaction between breathing type and O_2 level is tested against the toad within breathing type by O_2 level interaction. We could also achieve these latter tests by fitting a model without the toad within breathing type by O_2 level interaction, which would then become the residual term. Note that many statistical programs assume all factors are fixed and default to using this as the denominator for all tests, which is incorrect if B(A) is random.

Source	SS	df	MS	*F*	*P*
Breathing type	39.921	1	39.921	5.762	0.027
Toad(breathing type)	131.634	19	6.928		
O_2 level	25.748	7	3.678	4.884	<0.001
Breathing type × O_2 level	56.372	7	8.053	10.693	<0.001
Toad(Breathing type) × O_2 level	100.166	133	0.753		

We would conclude that there is a significant difference between toads with the different breathing types, but this depends on O_2 level (significant breathing type × O_2 level interaction).

We then re-analyzed the data after recoding them as a "repeated measures" design. For most software, we get even more extensive output.

BETWEEN SUBJECTS

Source	SS	df	MS	*F*	*P*
Breathing type	39.921	1	39.921	5.762	0.027
Residual	131.634	19	6.928		

Note:
This residual is actually toads nested within breathing type.

WITHIN SUBJECTS

Source	SS	df	MS	*F*	*P*	*GG P*	*HF P*
O_2 level	25.748	7	3.678	4.884	<0.001	0.004	0.002
Breathing type × O_2 level	56.372	7	8.053	10.693	<0.001	<0.001	<0.001
Residual	100.166	133	0.753				

Note:
This residual is actually toads within breathing type by O_2 level.

Greenhouse–Geisser epsilon: 0.428
Huynh–Feldt epsilon: 0.544

The "between subjects" and "within-subjects" parts of the ANOVA are distinguished and B(A), in this example toads within breathing type, is assumed to be random and all other factors fixed. The ANOVA output is, however, identical to the partly nested ANOVA above. Estimates of ε are also provided – the Greenhouse–Geisser is more conservative than the Huynh–Feldt estimate and neither is close to one, suggesting that the sphericity assumption is not met. Because both estimates of epsilon are less than 0.75, the Greenhouse–Geisser adjustment is preferred. Our conclusions would not be affected by these more conservative tests; there is a significant interaction between O_2 and breathing type. Both main effects are also significant, although it is more sensible to base further interpretation on the interaction. It is clear from Figure 11.4 that breathing rate decreases

with increasing O_2 level for buccal breathing toads but increases with O_2 level for lung breathing toads.

Because of the interaction, simple main effects tests for O_2 level at each breathing type separately might be of interest. We adjust the df for these tests based on the Greenhouse–Geisser estimate of epsilon.

BUCCAL:

Source	SS	df	MS	*F*	*P*	GG df	GG *P*
O_2 level	75.433	7	10.776	14.311	<0.001	2.997	<0.001
Residual	100.166	133	0.753			56.951	

LUNG:

Source	SS	df	MS	*F*	*P*	GG df	GG *P*
O_2 level	19.907	7	2.844	3.777	0.001	2.997	0.015
Residual	100.166	133	0.753			56.951	

There is a significant effect of O_2 level for both breathing types, although the effect seems stronger for buccal breathing toads than lung breathing toads.

For most statistical software, "repeated measures" output will include polynomial trend analyses. With eight O_2 levels, up to seventh order polynomials could be examined, although we will just look at the first three. The interaction test of these polynomials is testing whether the trend (linear, quadratic, etc.) through O_2 level differs between breathing types; the main effect test is examining whether there is a trend through O_2 level pooling breathing types.

Polynomial Test of Order 1 (Linear)

Source	SS	df	MS	*F*	*P*
O_2 level	17.010	1	17.010	8.255	0.010
Breathing type × O_2 level	40.065	1	40.065	19.444	<0.001
Residual	39.149	19	2.060		

Polynomial Test of Order 2 (Quadratic)

Source	SS	df	MS	*F*	*P*
O_2 level	5.007	1	5.007	6.967	0.016
Breathing type × O_2 level	12.326	1	12.326	17.150	0.001
Residual	13.655	19	0.719		

Polynomial Test of Order 3 (Cubic)

Source	SS	df	MS	*F*	*P*
O_2 level	1.747	1	1.747	3.263	0.087
Breathing type × O_2 level	1.784	1	1.784	3.331	0.084
Residual	10.174	19	0.535		

Both linear and quadratic trends are different between the two breathing types; there is no evidence of a cubic trend. It is clear from Figure 11.4 that the linear trends are in different directions for the two breathing types. Note that separate

error terms are used for each trend test, a requirement if there is a chance that sphericity of variances and covariances does not hold.

Finally, we get the multivariate tests of the within-subjects hypotheses.

O_2 level		Hypoth. df	Error df	*F*	*P*
Wilks' Lambda	0.115	7	13	14.277	<0.001
Pillai Trace	0.885	7	13	14.277	<0.001
Hotelling–Lawley Trace	7.688	7	13	14.277	<0.001

Breathing type × O_2 level		Hypoth. df	Error df	*F*	*P*
Wilks' Lambda	0.325	7	13	3.853	0.017
Pillai Trace	0.675	7	13	3.853	0.017
Hotelling–Lawley Trace	2.075	7	13	3.853	0.017

The conclusions from the Pillai statistic agree with the univariate analysis – a significant interaction between O_2 level and breathing type.

11.1.3 Reasons for using these designs

The examples above demonstrate the two major reasons for using split-plot or repeated measures designs. First, if our experimental or sampling units (organisms, ponds, sites) are expensive or otherwise difficult to obtain, we might consider applying a number of treatments to each (or to subunits within each) or recording each through time. Second, if we expect lots of variation between these units, and are worried that this variation might obscure effects of our treatments, we can attempt to remove this variation by taking a biological "unit" and applying different treatments to it – sampling different parts of the same pond, applying a range of oxygen concentrations, etc. The basic difference between a split-plot and a group by trials repeated measures design is that the former allocates the within plots treatments to subunits within each plot whereas the latter allocates the within subjects treatments sequentially to each subject.

11.2 Analyzing partly nested designs

We will first describe the analyses for a standard partly nested design that has three factors. One (the plots or subjects) is nested within the second, but both of these factors are crossed (orthogonal, factorial) with the third (Figure 11.2). In the split-plot example from Wissinger *et al.* (1996), we have hydroperiod (A), ponds within hydroperiods ((B(A)), and density treatment (C). In the repeated measures example from Schwartz *et al.* (1995), we have species (A), plants within each species ((B(A)), and temperature (C). In both examples, we have every combination of A and C (hydroperiod and density or species and temperature), so A and C form a factorial design. B and C are also factorial because every pond gets all density treatments and every plant gets all temperature treatments.

We could also have replicate observations from each combination of B and C (run each plant twice at each temperature, have replicate cages for each density treatment within each pond, etc.). As we will see below, in most cases, it makes little difference to how we test the effects of A and C.

We will describe the linear model and the various forms of analysis using split-plot terminology; keep in mind, however, that the plots are simply replaced by subjects in repeated measures designs. Components for fixed and random factors in expected mean squares are represented as "variances"; remember the different interpretations of variation between means of fixed treatment

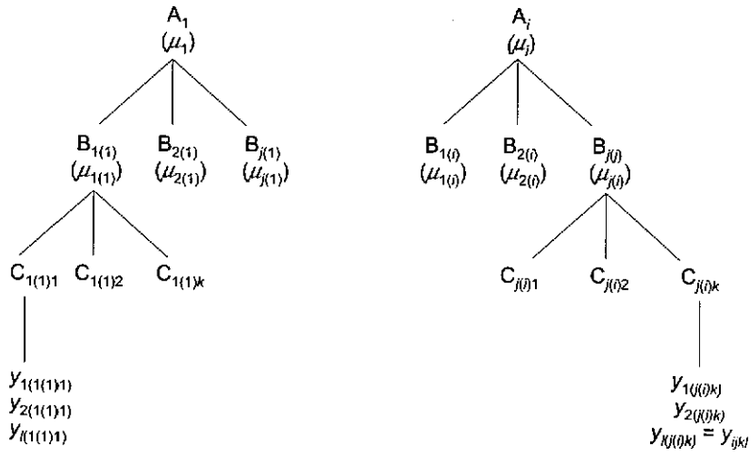

Figure 11.2 Part of data set for partly nested design, with p levels of factor A ($i = 1$ to p), q levels of factor B ($j = 1$ to q) nested within each level of A, r levels of factor C ($k = 1$ to r) crossed with factors A and B(A), and n replicate observations ($l = 1$ to n) within each combination (cell) of A, B(A) and C.

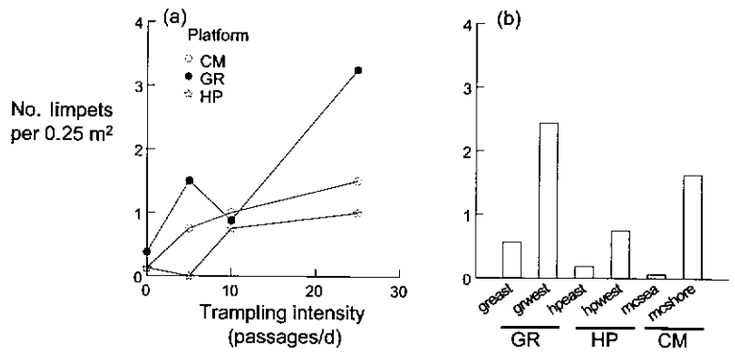

Figure 11.3 Variation in numbers of limpets under different trampling regimes and at different places, from Keough & Quinn (1998). Panel (a) shows number of limpets vs intensity of trampling for three rock platforms, and panel (b) shows variation among plots within platforms in overall abundance of limpets.

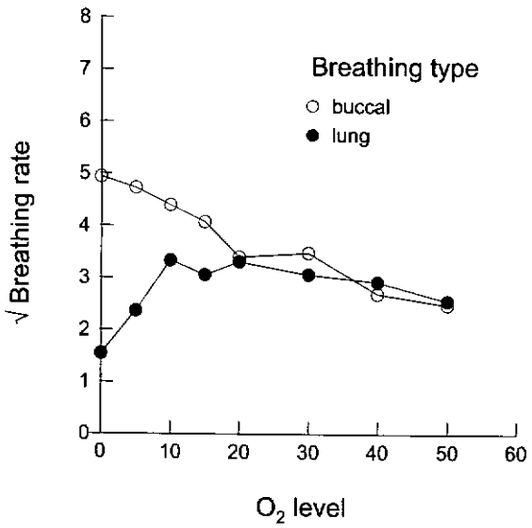

Figure 11.4 Mean square root transformed rate of buccal breathing for lung and buccal breathing toads for eight levels of O_2 concentration from Mullens (1993).

groups versus variance across all possible groups from which we have selected a subset at random – see Box 9.8.

11.2.1 Linear models for partly nested analyses

Linear effects model

Consider a design with p levels of factor A ($i = 1$ to p), q levels of factor B (plots or subjects) nested within each level of A ($j = 1$ to q) and r levels of factor C ($k = 1$ to r), crossed with both A and B (Table 11.1). From Keough & Quinn (1998), p equals three (the number of sites), q equals two (the number of plots) and r equals four (trampling treatments). From Mullens (1993), p equals two (the different breathing types), q equals eight or 13 (the number of toads of each breathing type) and r equals eight (O_2 levels). For completeness, we will describe the model with replicate observations

Table 11.1 Marginal means for a partly nested design with i levels of factor A, j levels of factor B within each level of factor A and k levels of factor C crossed with both A and B

		C_1	C_2	C_k	B marginal means	A marginal means
A_1	$B_{1(1)}$	y_{111}	y_{112}	y_{11k}	$\bar{y}_{j=1(1)}$	
	$B_{2(1)}$	y_{121}	y_{122}	y_{12k}	$\bar{y}_{j=2(1)}$	$\bar{y}_{i=1}$
	$B_{j(1)}$	y_{1j1}	y_{1j2}	y_{1jk}	$\bar{y}_{j(1)}$	
A_2	$B_{j(2)}$	y_{2j1}	y_{2j2}	y_{2jk}	$\bar{y}_{j(2)}$	$\bar{y}_{i=2}$
A_i	$B_{j(i)}$	y_{ij1}	y_{ij2}	y_{ijk}	$\bar{y}_{j(i)}$	$\bar{y}_i$
C marginal means		$\bar{y}_{k=1}$	$\bar{y}_{k=2}$	$\bar{y}_k$		

($l = 1$ to n) within each combination of A, B and C, e.g. Keough & Quinn (1998) had replicate observations (strips) within each site, plot and treatment combination. Usually, however, there is only a single observation (e.g. a single toad) of each level of C in each plot/subject.

The formal linear model for a split-plot design is:

$$y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{j(i)k} + \varepsilon_{ijkl} \quad (11.1)$$

Details of this linear model, including estimation of its parameters, are provided in Box 11.3.

From Keough & Quinn (1998):

$$\begin{aligned}(\text{number of limpets})_{ijkl} &= \mu + (\text{site})_i + \\ &(\text{plot within site})_{j(i)} + (\text{trampling})_k + \\ &(\text{interaction between site and trampling})_{ik} + \\ &(\text{interaction between plot within site and trampling})_{j(i)k} + \varepsilon_{ijkl}\end{aligned} \quad (11.2)$$

From Mullens (1993):

$$\begin{aligned}(\text{breathing rate})_{ijkl} &= \mu + (\text{breathing type})_i + \\ &(\text{toad within breathing type})_{j(i)} + (O_2 \text{ level})_k + \\ &(\text{interaction between breathing type and } O_2 \\ &\text{level})_{ik} + (\text{interaction between toad within} \\ &\text{breathing type and } O_2 \text{ level})_{j(i)k} + \varepsilon_{ijkl}\end{aligned} \quad (11.3)$$

In models 11.1 and 11.2 we have the following.

y_{ijkl} is the number of limpets in the lth strip in the kth level of trampling treatment for the jth plot at the ith site. Commonly in these designs, n equals one, although Keough & Quinn (1998) had n equals two.

μ is the overall (constant) population mean number of limpets per strip for all levels of trampling in all plots across all sites.

Factor A is fixed, so α_i is the effect of ith site on the number of limpets per strip, pooling over

Box 11.3 The partly nested linear model and its parameters

Consider a design with p levels of factor A ($i = 1$ to p), q levels of factor B (plots or subjects) nested within each level of A ($j = 1$ to q) and r levels of factor C ($k = 1$ to r), crossed with both A and B (Table 11.1) and replicate observations ($l = 1$ to n) within each combination of A, B and C.

The formal linear model for a split-plot design is:

$$y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{j(i)k} + \varepsilon_{ijkl}$$

In this model we have the following.

y_{ijkl} is the value of the response variable for the lth observation in the kth level of factor C for the jth plot/subject in the ith level of factor A. Commonly in

these designs, $I = 1$, but our two worked examples in this chapter include one in which $I = 2$ (Box 11.1) and one with $I = 1$ (Box 11.2).

μ is the overall (constant) population mean of the response variable.

If factor A is fixed, α_i is the effect of ith level of factor A ($\mu_i - \mu$), pooling over factor C. If factor A is random, α_i is a random variable with a mean of zero and a variance of σ_α^2, measuring the variance in mean values of the response variable across all possible levels of factor A that could have been used.

Plots or subjects are nearly always random so $\beta_{j(i)}$ is a random variable with a mean of zero and a variance of σ_β^2, measuring the variance in mean values of the response variable across all possible plots or subjects that could have been used within any level of A.

If factor C is fixed, γ_k is the effect of the kth level of factor C ($\mu_k - \mu$), pooling over factor A. If factor C is random, γ_k is a random variable with a mean of zero and a variance of σ_γ^2, measuring the variance in mean values of the response variable across all possible levels of factor C that could have been used.

If factors A and C are both fixed, $\alpha\gamma_{ik}$ is the effect of the interaction between the ith level of A and kth level of C. If either factor is random, then $\alpha\gamma_{ik}$ is a random variable with a mean of zero and a variance of $\sigma_{\alpha\gamma}^2$, measuring the variance across all the possible interaction terms between the fixed levels (if A is fixed) or all possible levels (if A is random) of factor A and the fixed levels (if C is fixed) or all possible levels (if C is random) of factor C.

Because plots/subjects are nearly always random, the interaction between factor C and plots/subjects $\beta\gamma_{j(i)k}$ is a random variable with a mean of zero and a variance of $\sigma_{\beta\gamma}^2$, measuring the variance across all the possible interaction terms between all the possible plots/subjects within any level of A and the fixed levels (if C is fixed) or all possible levels (if C is random) of factor C.

ε_{ijkl} is random or unexplained error associated with the lth observation in the kth level of factor C for the jth plot/subject in the ith level of factor A. These error terms are assumed to be normally distributed in each combination of A, B and C with a mean of zero and a variance of σ_ε^2. Note that classical split-plot and repeated measures designs usually do not have replication for each combination of plot and factor C ($n = 1$) so ε_{ijkl} usually cannot be separately estimated.

This effects model is overparameterized (Box 8.3) so the usual sum-to-zero constraints are imposed on the fixed effects. We can also use a cell means model (Kirk 1995) which may be useful when there are missing observations (Section 11.6).

Estimating the parameters of the partly nested model follows the procedures outlined in the previous three chapters for single factor, multifactor and randomized block models. The cell means (μ_{ijk}) are estimated by means of the observations in each cell, although there is often only a single observation for each A, B and C combination. The marginal means are shown in Table 11.1 and represent averages across the appropriate cell means (or single observations). Standard errors for these means must be based on the appropriate variance estimate (mean square), the one that is used as the denominator of the F-ratio for testing the H_0 that the means are equal (see Boxes 9.2 and 9.6).

plots and trampling treatment. If factor A was random, e.g. sites were chosen at random along the shore, then α_i is a random variable with a mean of zero and a variance of σ_α^2, measuring the variance in mean number of limpets across all possible sites that could have been used.

Plots are random so $\beta_{j(i)}$ is a random variable with a mean of zero and a variance of σ_β^2, measuring the variance in mean number of limpets across all possible plots that could have been used within any site, pooling trampling treatments.

Factor C is fixed, so γ_k is the effect of the kth level of trampling treatment, pooling over plots and sites. If factor C was random, e.g. trampling levels were chosen at random from the possible trampling levels that could have been used, then γ_k is a random variable with a mean of zero and a variance of σ_γ^2, measuring the variance in mean number of limpets across possible levels of trampling that could have been used.

Factors A and C are both fixed, so $\alpha\gamma_{ik}$ is the effect of the interaction between the ith site and kth trampling treatment. This interaction measures whether the effect of trampling is consistent at all sites used. If one factor is random, e.g. random sites, then $\alpha\gamma_{ik}$ is a random variable with a mean of zero and a variance of $\sigma_{\alpha\gamma}^2$, measuring the variance of all the possible interaction terms between all possible sites and the fixed trampling levels. This interaction would measure whether the effect of trampling was consistent across all possible sites.

Because plots are random, the interaction between trampling treatment and plots $[\beta\gamma_{j(i)k}]$ is a random variable with a mean of zero and a variance of $\sigma_{\beta\gamma}^2$, measuring the variance of all the possible interaction terms between all the possible plots within any site and the fixed trampling treatments. This measures the variation in effects of trampling at the spatial scale of plots.

$\varepsilon_{ij(kl)}$ is random or unexplained error associated with the lth strip in the kth trampling treatment for the jth plot at the ith site. This is the error associated with each strip that is not due to trampling treatment, plot or site. Note that classical split-plot and repeated measures designs usually do not have replication for each combination of plot and factor C (n equals one) so ε_{ijkl} usually cannot be separately estimated from $\sigma_{\beta\gamma}^2$.

Predicted values and residuals

If we replace the parameters in our model by their OLS estimates (Box 11.3), it turns out that the predicted or fitted values of the response variable from our linear model 11.1 are:

$$\hat{y}_{ijkl} = \bar{y}_{ijk} \qquad (11.4)$$

The error terms from model 11.1 can be estimated by the residuals, the difference between each observed and predicted value. In most applications of split-plot and groups by trials repeated measures designs, there is only a single observation per cell and these residuals all equal zero. In these circumstances, we cannot directly estimate σ_ε^2, the variance of the error terms, unless we assume that $\sigma_{\beta\gamma}^2$, the variance due to the B(A)×C interaction, equals zero. Not being able to estimate σ_ε^2 does not, however, compromise our tests of the main hypotheses of interest, those of A, C and A×C, the argument being similar to that used for analyses of RCB and simple repeated measures designs in Chapter 10 (see Section 11.2.3).

11.2.2 Analysis of variance

The partitioning of variance from the OLS fit of the linear model 11.1 is shown in Table 11.2. We do not provide computational details for sums-of-squares (SS) for each term in this ANOVA – see Winer *et al.* (1991) and Kirk (1995) for the classical formulae. In practice, however, we assume that you have access to statistical software with a general linear modeling routine when dealing with these complex designs. The SS for each source of variation are calculated by comparing the fit of the full model with the fit of an appropriate reduced model (the model including all terms except the one we wish to test in our H_0), as we described in Chapters 8, 9 and 10 for simpler ANOVA models. The general expected values of the mean squares are also provided in Table 11.2(a), as well as those for the usual case of factors A and C being fixed and factor B (plots or subjects) being random.

This ANOVA is more complicated than those from previous chapters but not really that difficult

Table 11.2 (a) Classical split-plot or repeated measures design with general expected mean squares and those for the specific case of A and C fixed, but B (plots or subjects) random, showing *F*-ratios used to test all hypotheses. For explanation of the conversion of the general model to particular combinations of fixed and random factors, see Box 9.8. (b) ANOVA for split-plot design from Wissinger *et al.* (1996), where hydroperiod and treatment are fixed factors, ponds is random and nested within hydroperiod, and from repeated measures design from Gange (1995), where species and date are fixed factor; trees is random and nested within species

(a)

Source	df	EMS (A, C fixed, B random)	Test (A, C fixed, B random)
Between plots/subjects			
A	$p-1$	$\sigma_\varepsilon^2 + n\sigma_\beta^2 + nqr\sigma_\alpha^2$	$MS_A/MS_{B(A)}$
B(A)	$p(q-1)$	$\sigma_\varepsilon^2 + n\sigma_\beta^2$	$MS_{B(A)}/MS_{Residual}$
Within plots/subjects			
C	$r-1$	$\sigma_\varepsilon^2 + nD_q\sigma_{\beta\gamma}^2 + npq\sigma_\gamma^2$	$MS_C/MS_{B(A)C}$
A × C	$(p-1)(r-1)$	$\sigma_\varepsilon^2 + nD_q\sigma_{\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2$	$MS_{AC}/MS_{B(A)C}$
B(A) × C	$p(q-1)(r-1)$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2$	$MS_{B(A)C}/MS_{Residual}$
Residual	$pqr(n-1)$	σ_ε^2	

(b)

Source	df
Between plots (i.e. ponds)	
Hydroperiod	1
Ponds within hydroperiod	38
Within plots (i.e ponds)	
Treatment	19
Hydroperiod × treatment	19
Ponds within hydroperiod × treatment	799
Between subjects (i.e. trees)	
Species	1
Trees within species	6
Within subjects (i.e. trees)	
Date	2
Species × date	2
Trees within species × date	12

– let's look at the different components. The between plots/subjects section is just a single factor ANOVA on the mean values for each plot/subject (i.e. averaging over the levels of factor C) and the plots/subjects within A (i.e. factor B) term is the equivalent of the residual term in this single factor ANOVA. The within plots/subjects section is just a number of RCB (or simple repeated measures) designs, one for each level of A. The effects of factor C and the A × C interaction are interpreted in the same way as for a two factor crossed ANOVA (Chapter 9). The C × plots within A [i.e. C × B(A)] term represents the pooled error terms from the p randomized block designs which comprise the within plots/subjects component of the analysis, i.e. for each level of A, we have a C by plots RCB design (Kirk 1995).

These ANOVA tables are illustrated in Table 11.2(b) for some of the examples we described in Section 11.1. The first is the split-plot design from Wissinger *et al.* (1996) where the between plots factor was hydroperiod, the within plots factor was density treatment and the plots were ponds nested within each hydroperiod. The second was the groups by trials repeated measures design from Gange (1995) where the between-subjects factor was tree type, the within-subjects factor was date, the subjects were individual trees and the response variable was aphid abundance. The ANOVA tables for our two worked examples are also provided in Box 11.1 and Box 11.2.

11.2.3 Null hypotheses

There are three null hypotheses of primary interest when we fit the partly nested model 11.1.

Factor A (fixed)

H_0: $\mu_1 = \mu_2 = \ldots = \mu_i$. This H_0 states that there is no difference between the factor A marginal means, pooling levels of factor C. For example, no difference in the mean number of limpets per strip between sites, pooling the trampling treatments.

This is equivalent to:

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_i = 0$, i.e. no effect of any level of factor A. For example, no effect of site on the mean number of limpets per strip, pooling the trampling treatments.

Factor C (fixed)

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$. This H_0 states that there is no difference between the factor C marginal means, pooling levels of factor A. For example, no difference in the mean number of limpets per strip between trampling treatments, pooling sites.

This is equivalent to:

H_0: $\gamma_1 = \gamma_2 = \ldots = \gamma_k = 0$, i.e. no effect of any level of factor C. For example, no effect of trampling treatment on the mean number of limpets per strip, pooling sites.

A × C interaction (fixed)

H_0: $\mu_{ijk} - \mu_i - \mu_k + \mu = 0$, which is the same as $(\alpha\gamma)_{ik} = 0$. This H_0 states that there are no interactions between A and C, e.g. the effect of site on the mean number of limpets per strip is the same for all trampling treatments and the effect of trampling treatment on the mean number of limpets per strip is the same for all sites.

The modifications of these H_0s for random factors are straightforward as described in Chapters 9 and 10.

Two other null hypotheses might also be tested in some circumstances.

Factor B(A) (random)

H_0: $\sigma_\beta^2 = 0$, i.e. the variance in mean values of the response variable across all possible plots or subjects that could have been used within any level of A equals zero. For example, no variance in the mean number of limpets per strip between plots in either site, pooling trampling treatments.

B(A) × C interaction (random)

H_0: $\sigma_{\beta\gamma}^2 = 0$, i.e. the variance across all the possible interaction terms between all the possible plots/subjects within any level of A and the fixed levels (if C is fixed) or all possible levels (if C is random) of factor C equals zero. For example, no variance in the mean number of limpets per strip across all the possible interaction terms between plots within each site and trampling treatment.

F-ratios

The *F*-ratios for testing these null hypotheses are based on the expected values of the relevant mean squares (Table 11.2(a)). When factors A and C are

both fixed, the F test for factor A uses a different denominator [$MS_{B(A)}$] than those for factor C and A×C [$MS_{CB(A)}$]. This is typical for designs with both fixed and random factors and is apparent in all these partly nested designs because the plots/subjects term is nearly always random. In the classical split-plot or repeated measures design with n equals one observation for each cell, the B(A) and C×B(A) terms cannot be tested. The implications of not being able to test the C×B(A) are analogous to the implications of having no test for a block by treatment interaction in a RCB design (Chapter 10). This makes sense given that the C×B(A) interaction comprises the pooled residual terms from the p RCB designs and each of these residual terms includes the plot/subject by treatment interaction. The first implication is that if B is random, then a strong C×B(A) interaction will reduce the power of the tests of C and A×C, although these tests are still valid because the expected mean squares of both C and A×C include the variance component due to C×B(A). This is not the case if B is fixed, where C×B(A) is an inappropriate error term for testing C and A×C (see below). The second implication is that the use of C×B(A) as an error term for C and A×C can be invalid if the observations within each plot/subject are correlated, which is almost certainly the case for repeated measures designs. For the C×B(A) to be used as an error term, we must make certain assumptions about the covariances of the observations (Section 11.3).

With replicate observations in each cell, the B(A) and C×B(A) terms can be tested against the residual. Note, however, that using many replicate observations within each cell, e.g. multiple measurements on toads or multiple strips within plots, may not be providing a much better test of the terms that you really care about. The B(A) term is rarely of much interest, and you probably don't care much if factor C has a different effect across levels of B, i.e. a C×B(A) interaction. The effort expended in sampling at this lowest level may not be producing a more powerful statistical test of any of the biologically interesting effects, only increasing the cost of the design in terms of time and/or money. When there is no replication within plots, Underwood (1997) argued that tests of the main effects of C and the A×C interaction can only be done if we assume there is no C×B(A) interaction, i.e. the effects of C do not vary from plot to plot. However, it is clear from Table 11.2(a) that the expected mean squares for C and the A×C interaction include the variance due to C×B(A) interaction, so the F test for C and A×C is testing for these effects over and above the variation due to the C×B(A) interaction and any residual variation. Therefore, we can interpret a significant effect of C or the A×C interaction even if the effects of C do vary for different levels B(A), which is similar to the argument we made in Chapter 10 for RCB designs when the blocks factor was random. A non-significant C or A×C interaction is more difficult to interpret in the presence of a C×B(A) interaction, but interpreting non-significant tests is always problematical.

As you can see from Table 11.2(a), the effects of A, C and the A×C interaction are all tested against other terms in the analysis, all featuring effects of B within A. Because you cannot control the number of levels of A or C when they are fixed factors, the only way to increase the power of these tests is to increase the degrees of freedom. This can only be achieved by using more levels of B (more plots or subjects, e.g. more toads, more plots, etc.), i.e. increasing q.

The expected mean squares and appropriate F tests for other combinations of fixed and random factors are presented in Table 11.3. When A, C and plots/subjects (i.e. B) are all fixed (Table 11.3), you can see that all terms are tested using the $MS_{Residual}$ as the denominator. Note that you must have replicate observations in each cell if plots (B) are fixed because $MS_{CB(A)}$ is not an appropriate denominator here unless you are very sure that the C×B(A) interaction is negligible. In almost every case that we deal with, factor B will be random so this design is unlikely for biological experiments. If factors A and plots/subjects are random, but C is fixed, the tests are straightforward (Table 11.3), but note that again, they use different combinations of denominators for F tests for the various hypotheses. Problems occur when plots/subjects and factor C are random (Table 11.3). It does not matter whether A is fixed or random in this case. The difficulty is that the main effect of factor A (which will almost always be of central interest) cannot be tested directly, because there is no

Table 11.3 | ANOVA tables with expected mean squares (EMS) for partly nested models, showing F-ratios used to test all hypotheses

		A, B, C fixed		A, B random, C fixed		A fixed, B, C random		A, B, C random	
Source	df	EMS	Test	EMS	Test	EMS	Test	EMS	Test
A	$p-1$	$\sigma_\varepsilon^2 + npr\sigma_\alpha^2$	$MS_A/MS_{Residual}$	$\sigma_\varepsilon^2 + nr\sigma_\beta^2 + nq\sigma_\gamma^2$	$MS_A/MS_{B(A)}$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2 + nr\sigma_\beta^2 + nqr\sigma_\alpha^2$	No test	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2 + nr\sigma_\beta^2 + npr\sigma_\alpha^2$	No test
B(A)	$p(q-1)$	$\sigma_\varepsilon^2 + nr\sigma_\beta^2$	$MS_{B(A)}/MS_{Residual}$	$\sigma_\varepsilon^2 + nr\sigma_\beta^2$	$MS_{B(A)}/MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2 + nr\sigma_\beta^2$	$MS_{B(A)}/MS_{B(A)C}$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2 + npr\sigma_\beta^2$	$MS_{B(A)}/MS_{B(A)C}$
C	$r-1$	$\sigma_\varepsilon^2 + npq\sigma_\gamma^2$	$MS_C/MS_{Residual}$	$\sigma_\varepsilon^2 + nq\sigma_{\alpha\gamma}^2 + npq\sigma_\gamma^2$	MS_C/MS_{AC}	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2 + npq\sigma_\gamma^2$	$MS_C/MS_{B(A)C}$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2 + npq\sigma_\gamma^2$	$MS_C/MS_{B(A)C}$
A×C	$(p-1)(r-1)$	$\sigma_\varepsilon^2 + nq\sigma_{\alpha\gamma}^2$	$MS_{AC}/MS_{Residual}$	$\sigma_\varepsilon^2 + nq\sigma_{\alpha\gamma}^2$	$MS_{AC}/MS_{B(A)C}$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2$	$MS_{AC}/MS_{B(A)C}$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2 + nq\sigma_{\alpha\gamma}^2$	$MS_{AC}/MS_{B(A)C}$
B(A)×C	$p(q-1)(r-1)$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2$	$MS_{B(A)C}/MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2$	$MS_{B(A)C}/MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2$	$MS_{B(A)C}/MS_{Residual}$	$\sigma_\varepsilon^2 + n\sigma_{\beta\gamma}^2$	$MS_{B(A)C}/MS_{Residual}$
Residual	$pqr(n-1)$	σ_ε^2		σ_ε^2		σ_ε^2		σ_ε^2	

appropriate denominator. The only option is to use quasi *F*-ratios, which are combinations of mean squares that produce an approximate test of your hypothesis (see Chapter 9). Winer *et al.* (1991) discussed this option in detail, but you should be aware that the resulting *F* tests are only approximate and not necessarily robust, and you probably should avoid this situation if possible.

11.2.4 Comparing ANOVA models

The SS, df and MS for each term in the partly nested model 11.1 can be derived from comparing the fit of a full and a reduced model, where the reduced model omits the parameter specified to be zero in the H_0. This is the same principle we have described in Chapters 9 and 10 for other multifactor models.

11.3 Assumptions

Irrespective of whether it is for a split-plot design or a groups by trials repeated measures design, the partly nested ANOVA model 11.1 has a number of assumptions that need to be assessed. Additionally, we should always check for outliers from our fitted model. A useful first step is to examine the residuals from the fit of the model. If we only have n equals one within each combination of A, B and C, then we should omit the B(A)×C term, otherwise the model is saturated (a perfect fit) and all the residuals are zero. These residuals will indicate any obvious outliers and also indicate any strong skewness in the data. Generally, however, the assumptions, and their assessment, in these analyses are considered separately for the between plots/subjects and within plots/subjects components.

11.3.1 Between plots/subjects

The test of factor A assumes normality and homogeneity of variance and the comments about these assumptions in Chapters 8 and 9 apply here. Note that, for the usual case of B random and A and C fixed, these assumptions apply to the levels of A (pooled across C) with the mean of Y in each level of B(A) as a replicate observation. It is often useful to create a new variable that is the average across the levels of C and then use that variable in boxplots for each level of A or to examine residuals from the fit of a single factor ANOVA model with p groups to that variable.

11.3.2 Within plots/subjects and multisample sphericity

The tests for any terms including within-plots/subjects factors(s), i.e. tests of C and A×C, have the assumption of sphericity of variances and covariance, as did RCB and simple repeated measures designs (Chapter 9). Unless this assumption holds, then the B(A)×C term is an inappropriate denominator for the test of C and A×C. Remember that these partly nested designs can be envisaged as a series of RCB (factor C by blocks, plots or subjects) experiments, one for each level of factor A, so the assumption is now multisample sphericity. Not only must the variance–covariance matrices be the same for each level of factor A, they must each show sphericity, which means that the variances of the differences between the levels of the repeated factor must be the same.

In classical repeated measures designs, the levels of the within-subjects factor (C) can usually be applied in random order to each subject (Winer *et al.* 1991). Similarly, in classical split-plot designs, the levels of the within-plot factor (C) should be randomly allocated to experimental units (subplots) within each plot. Under these randomization conditions, there is no reason for the sphericity assumption not to hold; in fact, the sphericity assumption is often not discussed when general statistics texts describe split-plot designs. In contrast, the sphericity assumption is unlikely to hold in repeated measures designs when subjects are recorded through time because the differences between times closer together are likely to be less variable (i.e. more similar) than times further apart. If sphericity is not met, the *F*-ratio statistics for within subjects effects (C and A×C) will be inflated, increasing the risk of a Type I error above the nominal level (e.g. 0.05) – see Keselman & Keselman (1993), Keselman *et al.* (1995) and Rasmussen (1989). There is no easy test for the null hypothesis that the variance–covariance matrices conform to multisample sphericity. Kirk (1995) recommended the *W* test and provided critical values, although we suggest it is safer to assume that multisample sphericity is not met in

Table 11.4 Degrees of freedom for within-plots or -subjects components of partly nested ANOVA

Source	df	Adjusted df
Within plots/subjects		
C	$(r-1)$	$(r-1)\hat{\varepsilon}$
A×C	$(p-1)(r-1)$	$(p-1)(r-1)\hat{\varepsilon}$
B(A)×C	$p(q-1)(r-1)$	$p(q-1)(r-1)\hat{\varepsilon}$

Note:
Adjustment based on estimate of ε indicating how far variance–covariance matrix is from sphericity.

repeated measures type designs and use one or more of the following analytical strategies.

Adjusted univariate *F*-ratio tests

As described in Chapter 10 for RCB and simple repeated measures designs, we can make the *F* tests more conservative using adjusted df. An index of sphericity is the population parameter ε, which can be estimated by the epsilon statistic ($\hat{\varepsilon}$). Two methods of estimating ε were described in Chapter 10, the Greenhouse–Geisser (G–G) estimate or the Huynh–Feldt (H–F) estimate (Winer *et al.* 1991, Yandell 1997). These sample $\hat{\varepsilon}$s can be used to adjust the df for within plots/subjects tests downwards to make the tests more conservative, since non-sphericity increases the risk of Type I error. The adjustment is simple, being the original df multiplied by $\hat{\varepsilon}$, although the new df will not be integers (Table 11.4). If ε is greater than 0.75, the correction based on the Huynh–Feldt $\hat{\varepsilon}$ is better, when ε is less than 0.75, the correction based on the Greenhouse–Geisser $\hat{\varepsilon}$ is better (Keselman & Kesleman 1993). These adjusted tests are standard output from most statistical software.

Multivariate tests

An alternative solution to the sphericity assumption is to treat the levels of the within-subjects or within-plots factor (i.e. the repeated measures factor) as multiple response variables in a multivariate analysis of variance (MANOVA in Chapter 16; see also Keselman & Keselman 1993, Looney & Stanley 1989, Kirk 1995, Tabachnick & Fidell 1996). The MANOVA actually uses the difference between successive repeated measurements (i.e. times) for each subject or plot as response variables and tests the null hypothesis that the difference scores have a population centroid (multivariate mean) equal to zero. The MANOVA approach can be useful for these designs because it doesn't assume sphericity of variances and covariances, although it does have all the usual MANOVA assumptions (Chapter 16; Johnson & Field 1993) and has fewer degrees of freedom. Also, if the n is less than the number of differences between successive repeated measurements (i.e. less than the number of levels of the within-plots or -subjects factor minus one), then the MANOVA approach cannot be used. As discussed in Chapter 16, the Pillai trace statistic is recommended for these multivariate tests.

Profile analysis

Another approach is to summarize the responses for each plot/subject as a single value and then use these values in a single factor ANOVA model comparing the levels of A. The between plots/subjects part of the partly nested univariate ANOVA does this by summarizing the responses of each plot/subject as an average across the levels of C. If factor C is quantitative, e.g. time, we can also summarize the responses of each plot/subject as a trend or response curve, such as a linear, quadratic, etc., and analyze the coefficients of these trends in separate one factor ANOVAs (Meredith & Stehman 1991). This provides a test of whether such trends (linear, quadratic, etc.) vary across factor A, i.e. a test of a treatment–contrast interaction (Chapter 9). Such tests are usually default output from statistical software and will be discussed in Section 11.5.2.

Which strategy is the best?

As we pointed out in Chapter 10 for RCB and simple repeated measures designs, neither the epsilon-adjusted univariate nor the multivariate approach is always more powerful, unless sphericity is met, when the traditional partly nested univariate analysis is clearly preferred. Looney & Stanley (1989) recommended using both approaches and rejecting the within-subjects null hypotheses if either the adjusted univariate or multivariate tests are significant. Kirk (1995) recommended a

preliminary test for multisample sphericity and using the adjusted univariate tests if the sphericity test is significant; however, his preliminary test is not straightforward and not available in statistical software. We suggest that preliminary tests for sphericity are of limited value and support the Looney & Stanley (1989) approach and the use of profile analyses if factor C is quantitative.

11.4 Robust partly nested analyses

As for other linear model analyses, the RT (rank transform) procedure has been proposed as a general method for overcoming problems of non-normality and possibly other assumptions of the partly nested analyses of variance (see discussion in Thompson 1991b). We reiterate our comments from Chapters 9 and 10. The rank transformation is nonlinear in nature (Akritas 1991) and therefore cannot effectively deal with interactions; indeed, a significant main effect may be indicated when it is simply due to an undetected interaction (Thompson 1991b). As the A×C interaction is often of considerable interest in the designs discussed in this chapter, the RT procedure seems inappropriate. RT procedures are also inappropriate for nested factors (Thompson 1991b), which are important in the models used to analyze split-plot and groups by trials repeated measures designs. Also, as discussed in Chapter 10 for analyses of RCB designs, a rank transformation can also change the nature of variances and covariances, making the assumption of sphericity less tenable (Akritas 1991). Although Thompson (1991b) has developed a general rank-based multivariate test statistic that is applicable to repeated measures designs, its usefulness is restricted to situations where there are no interactions.

We could also fit the models in this chapter using generalized linear models (GLMs), that allow a range of different error distributions of which normal is just one (Chapter 13). Maximum likelihood techniques are used for fitting the models and estimating the parameters and likelihood ratios are used for hypothesis tests of these parameters. Care must be taken in the choice of full and reduced models for such complex analyses because some models won't make much biological or statistical sense, e.g. a model that includes B(A) but not A. Note that GLM analyses are still sensitive to the specification of the error distribution so model diagnostics are very important, just as they are for linear models. Chapter 13 includes a more detailed discussion of GLMs.

11.5 Specific comparisons

11.5.1 Main effects

Planned contrasts for the between-plots/subjects main effect are done in the same way as described in Chapter 8 and simply average across the within-plots/subjects factor levels for each experimental unit. Planned contrasts for the within-plots/subjects main effect assume multisample sphericity if the usual B(A)×C term is to be used as the denominator. The two alternatives are to adjust the df for these contrasts using the G-G or H-F estimates of ε (Section 11.3.2) or use separate error terms, e.g. $C_{contrast} \times B(A)$, for each contrast (Kirk 1995). These error terms are calculated similarly to those for analyses RCB (or simple repeated measures) designs described in Chapter 10, except that the contrasts are calculated across the levels of factor A; Kirk (1995) provides computational details but good statistical software will calculate these separate error terms. They basically represent a separate *F*-ratio testing for differences in the levels of C within each level of A. Keselman & Keselman (1993) suggested an approximate paired *t* test with separate error terms based on the two groups being compared, called the KKS test, similar to that described in Chapter 10, although Satterthwaite's adjusted df are used.

Unplanned comparisons for between-plots/subjects factors are done in the same way as described in Chapter 8, and simply average across the within-plots/subjects factor levels for each experimental unit or subject. The usual unplanned multiple comparison procedures may not be reliable for within-plot/subjects factors because the means are probably correlated to some extent, particularly for repeated measures designs. Keselman & Keselman (1993) described some new stepwise multiple comparison procedures for within-subjects/plots factors. The simplest approach might be to contrast the specific

levels of C applying a Bonferroni-type adjustment of significance levels for multiple testing if required (Chapter 3). Note that these contrasts between levels of C will use the B(A)×C term as the denominator and therefore assume multisample sphericity; adjusted df should be used based on G-G or H-F estimates of ε.

11.5.2 Interactions

In partly nested ANOVA models, the main interaction of interest is between A and C and represents an interaction between a between-plots/subjects factor and a within-plots/subjects factor. This interaction can be explored with "interaction" plots of means, where we might have the levels of factor C along the horizontal axis, the response variable along the vertical axis and each point represents the mean of factor A levels across plots/subjects within each A level. Deviations from parallel lines indicate some interaction between A and C.

Tests of simple main effects can also be done as described in Chapter 9, the only difficulty for the designs in this chapter is choosing the appropriate denominator for the *F* tests (Kirk 1995, Maxwell & Delaney 1990). In Chapter 9, we pointed out that for a two factor crossed (A, B, A×B) linear model, the SS for simple main effect tests for factor A represent partitioning of the SS_A and SS_{AB}, whereas the simple main effects tests for B represent partitioning of the SS_B and SS_{AB}. In contrast to the two factor completely randomized design, however, the test of the A term in a partly nested model with A and C fixed and B (plots/subjects) random uses a different denominator than the tests of the C and A×C interaction terms. So what denominators do we use for the simple main effects tests in a partly nested model?

The simple effects tests for C at each level of A separately, e.g. the effect of O_2 level for each breathing type separately in the Mullens (1993) example, are relatively straightforward because both C (O_2 level) and A×C (breathing type×O_2 level) use the same denominator – C×B(A). Note that if multisample sphericity does not hold, then these tests should be based on adjusted degrees of freedom using the G-G or H-F estimates of ε (Section 11.3.2). Alternatively, separate denominators should be used for each simple effect, the equivalent to calculating a simple repeated measures ANOVA testing C within each level of A separately (Chapter 10).

For the simple effects tests for A at each level of C separately, e.g. the effect of breathing type for each O_2 level separately, Kirk (1995) and Maxwell & Delaney (1990) recommended using a denominator that represents the average of the B(A) and B(A)×C terms. This is sometimes called the within-cells error term:

$$\frac{SS_{B(A)} + SS_{B(A)\times C}}{p(q-1) + p(q-1)(r-1)} \qquad (11.5)$$

Tests using the error term in expression 11.5 might be biased, especially if the two terms contributing to the pooled term are very different.

11.5.3 Profile (i.e. trend) analysis

A useful approach, which can be used in conjunction with any experimental design where at least one factor is quantitative, is to look for trends across levels of the quantitative factor (Chapter 8). For designs in this chapter, the common approach is to test for trends across the levels of factor C (the within plots/subjects factor) if C is quantitative (e.g. time, O_2 level). The simplest trends to examine are those of a polynomial form, such as linear, quadratic, cubic, etc. (see Chapter 8). Tabachnick & Fidell (1996) provide an excellent description of these methods for repeated measures designs, and therefore for partly nested models in general.

The number of polynomial contrasts that can be calculated is one less than the number of levels of the relevant factor. For example, if there were six levels of factor C, you could test for linear (X), quadratic (X^2), cubic (X^3), quartic (X^4) and quintic (X^5) polynomials, although it is often difficult to attach biological meaning to trends more complex than cubic. It is important to remember that these trend tests depend on the metric (spacing) of levels of the quantitative factor(s), as discussed in Chapter 8, and that most statistical software assumes equal spacing by default. The tests of these polynomials are statistically orthogonal (independent) of each other because each is tested using a separate component of df_C and MS_C (or A×C if trends are tested as part of the interaction), with separate components of the $df_{CB(A)}$ and $MS_{CB(A)}$ for the denominators of each trend *F* test.

Testing for trends as part of analyses of classical repeated measures designs is often termed profile analysis (Tabachnick & Fidell 1996). As with tests of trends in completely randomized factorial designs discussed in Chapter 9, there are two types of tests of interest in profile analysis:

- Main effect trends, usually across the within-plots/subjects factor C pooling the levels of factor A. For example, is there a linear trend in breathing rate of toads across oxygen concentrations, pooling the two breathing types? Is there a quadratic trend? A cubic trend? Trends could also be examined across factor A as part of the between-plots/subjects part of the analysis.
- Treatment–contrast interactions for examining the A × C interaction term. Here we compare trends of the same form (e.g. linear) across C between different levels of factor A. For example, is the linear trend in breathing rate of toads across oxygen concentrations the same for the two breathing types? Is the quadratic trend the same? These tests are often described as tests of parallelism (Tabachnick & Fidell 1996), since testing whether the linear trends are the same across level of A is clearly a test of whether the trends are parallel.

We do not provide computational details for calculating these trend tests because there is nothing additional to the information we included in Chapter 9 and these trend tests are usually default output from statistical software if the data are coded, and analysis run, as a classical repeated measures design. Note that some software will automatically test each trend MS against a separate error term so that multisample sphericity is not assumed. Alternatively, the B(A) × C term could be used, with adjustments to the df based on the G–G or H–F estimates of $\hat{\varepsilon}$. Growth curve analysis can also be useful for ecophysiological studies and involves comparisons of nonlinear regressions of a more complicated form than simple polynomials (Potvin *et al.* 1990).

An example of these trend analyses was provided by Sharpe & Keough (1998), who examined temporal trends in chlorophyll-*a* and in the density of herbivorous snails following the removal of dominant grazers from the intertidal zone of a rocky shore. The removal treatments were the between-subjects/plots factor, and time was the repeated factor. Individual boulders were the plots/subjects, so different boulders received different removal treatments. They recorded chlorophyll-*a* from randomly selected areas on each boulder, and censused a range of herbivores once a month. They contrasted the linear temporal trend in abundance of each species (as a measure of recolonization rate) between particular combinations of treatments. We also illustrate these trend analyses in the worked examples in Box 11.2 and Box 11.4.

11.6 Analysis of unbalanced partly nested designs

Unequal sample sizes can arise in partly nested (split-plot or repeated measures) designs in two ways. First, the number of plots or subjects in each level of the between plots factor might vary. Since the between-plots tests average over the within-plots factors, this type of unequal sample size is no different to unequal sample sizes in the usual factorial ANOVAs described in Chapter 9, and our recommendations are the same. Remember that checking assumptions becomes much more important when sample sizes are unequal and that even tests of within-plots/subjects factors can be more sensitive to assumptions (e.g. sphericity) when the between-plots/subjects part of the design is unbalanced (Keselman & Keselman 1993). Second, when we have no replication within each cell (the classical split-plot or repeated measures design), then missing observations equate to missing cells. If you have a reasonable number of plots/subjects, then a simple approach is to delete the plot(s) or subject(s) with the missing observations; this causes problems if sample sizes (number of plots or subjects) are small because the between-subjects/plots part of the analysis may become severely unbalanced. Basically, most statistical software will use this approach by default if the data are set up for a classical repeated measures analysis. If you don't have many plots/subjects but lots of levels of the within-subjects/plots factor(s), then it might be better to omit the level of factor C (or the combination of levels if you have more than one within-subjects/plots factor) with the missing observations. This approach changes the null hypotheses being tested, of course, but if the hypotheses are general ones about trends through time and you have a long time sequence, then omitting one or two times may not have much effect.

An alternative solution is to simply fit the partly nested linear model (Berk 1987) and compare this full model with appropriate reduced models for specific hypotheses, as described in Chapter 10 for RCB designs. Unfortunately, the *F* tests are more sensitive to the sphericity assumption when observations are missing and most statistical software doesn't provide epsilon estimates, nor adjusted univariate tests, when the analysis is run this way, so be careful. As we recommended in Chapter 10 for RCB and simple repeated measures designs, a practical strategy may be to delete the subject(s)/plot(s) with the missing observation(s), running the analysis as a classical repeated measures design to check sphericity and then only fit a partly nested linear model to the data with all subjects/plots if that assumption is tenable. This is messy but there are not many practical options when dealing with missing observations in these designs.

More complicated solutions are provided by Berk (1987), who suggested ML and REML estimation procedures that weight the observations, by Kirk (1995), who described using the cell means model and testing a subset of hypotheses using contrasts, and by Rovine & Delaney (1990). All these methods will be difficult for practicing biologists, at least until they are standard components of statistical software.

11.7 Power for partly nested designs

As expected, power calculations become more complicated with these complex designs, with the possibility of separate power calculations for a series of main effects, and interactions. We can divide these tests into those involving only between-plot/subject terms, only within-plot/subject terms, and interactions between the two groups. For between-subjects factors, power calculations are similar to those described for Chapters 8 and 9. They are routine when main effects are of interest, and they can be made easier by recoding the data file as means, averaging across the repeated or within-plots factor levels. For the more complex within-subjects/plots effects, the power calculations can be done, with two important steps. First, specifying an effect size can be very difficult, as for all complex interactions. Second, in computing power for a particular effect, we must identify the denominator used to test that effect, and use that MS to generate the variance estimate needed to calculate power.

One special case in which the power calculation is relatively simple is the family of BACI (Before-After-Control-Impact) designs used in environmental monitoring. The test for an environmental impact is an interaction between Before-After and Control-Impact, tested using, for example, changes Before-After at replicate locations within Control and Impact categories. In the original formulation of this design, with two locations (C and I), two periods (B and A), and multiple sampling times within each period, we could use a partly nested analysis, with periods, times within periods, and samples at C and I at each time. An impact would be revealed as a change in the difference between C and I, from the Before to the After period. Stewart-Oaten *et al.* (1986) pointed out that this design can be analyzed as a *t* test, simply by calculating the difference, C-I, and comparing that difference between periods. As a consequence, rather than formulating an effect size based on the interaction, we can specify an effect size as the divergence or convergence of these C-I differences. More complex formulations of this design (e.g. Downes *et al.* 2002, Keough & Mapstone 1997) can also be simplified in this way, because the interaction of interest is between the main between- and within-plots factors, and each of them has only two levels.

11.8 More complex designs

So far we have considered partly nested designs involving one between-subjects/plots factor (A), one within-subjects/plots factor (C) and one factor representing subjects/plots (B). These

experimental designs can be extended to include more than one between-subjects/plots factor and/or more than one within-subjects/plots factor.

11.8.1 Additional between-plots/subjects factors

There is nothing difficult about additional between-subjects/plots factors, because this part of the analysis is just an ANOVA on the average of the response variable for each plot/subject. For example, a four factor design might have two between-subjects/plots factors (A and C), one within-subjects/plots factor (D), and factor B representing plots nested within A and C. For example, McGoldrick & Mac Nally (1998) studied the impact of eucalypt flowering on the dynamics of bird communities in forests of southeastern Australia. They had eight sites (i.e. plots) arranged in a two factor crossed design with factor A being habitat (two levels: dominated by ironbark eucalypts vs dominated by stringybark eucalypts) and factor B being region (two levels: north of Great Dividing Range and south of Great Dividing Range) with two sites within each combination. Each site was censused monthly for twelve months, so month was the within-plots/subjects factor. The response variables included flowering index, density of nectarivorous birds, species richness of nectarivorous birds etc. The analysis for this example is in Box 11.4, where we analyze the density of nectarivorous birds, transformed to logs after adding one to each observation to account for zero values.

Box 11.4 Impact of flowering on forest bird communities

As described in Section 11.8.1, McGoldrick & Mac Nally (1998) studied the impact of eucalypt flowering on the dynamics of bird communities in forests of S.E. Australia. They used a partly nested design with two between-plots/subject factors (habitat and region) with two sites within each combination. Each site was censused monthly for twelve months, so time was the within-plots/subjects factor. The response variable we will analyze is natural log transformed (density of nectarivorous birds + 1).

The specific null hypotheses of interest were as follows.

- No difference between habitats in the mean $\log_e$ (density of nectarivorous birds + 1), pooling regions and months.
- No difference between regions in the mean $\log_e$ (density of nectarivorous birds + 1), pooling habitat and months.
- No interaction between habitat and region on the mean $\log_e$ (density of nectarivorous birds + 1), pooling months. Rephrased, the effect of habitat on the mean $\log_e$ (density of nectarivorous birds + 1) was the same for both regions and vice versa, pooling months.
- No difference between months in the mean $\log_e$ (density of nectarivorous birds + 1), pooling habitats and regions.
- No interactions between habitat and month, region and month, or habitat and region and month on the mean $\log_e$ (density of nectarivorous birds + 1). Re-phrased, the effect of habitat, pooling regions, was the same in all months, the effect of region, pooling habitats, was the same in all months, and the interaction between habitat and region was the same in all months.

With no replicates within each combination of habitat, region, site and month, we could not test hypotheses about the random factor sites within habitat and region or months by sites within habitat and region.

Between plots/subjects Source	SS	df	MS	*F*	*P*
Habitat	88.313	1	88.313	48.975	0.002
Region	0.106	1	0.106	0.059	0.821
Habitat × region	1.334	1	1.334	0.740	0.438
Site(habitat, region)	7.213	4	1.803		

Within plots/subjects Source	SS	df	MS	*F*	*P*	*GG*
Month	48.676	11	4.425	5.941	<0.001	0.019
Habitat × month	75.152	11	6.559	8.806	<0.001	<0.006
Region × month	11.436	11	1.040	1.396	0.209	0.299
Habitat × region × month	3.858	11	0.351	0.471	0.911	0.665
Site(habitat, region) × month	32.774	44	0.745			
Greenhouse–Geisser epsilon:	0.2104					
Huynh–Feldt epsilon:	0.8907					

Our analysis agrees with that published by McGoldrick & Mac Nally (1998), although they did not present adjusted tests for within-plots/subjects tests. The adjusted df did not change our conclusions. The month effect varied between habitats and there were neither effects of region nor any interactions between habitat and region or region and month. Note that the epsilon estimates differ greatly and for the three factor interaction, the adjusted test is more liberal than the unadjusted tests.

Linear trends: Source	SS	df	MS	*F*	*P*
Time	16.056	1	16.056	12.231	0.025
Habitat × month	24.532	1	24.532	18.689	0.012
Region × month	3.028	1	3.028	2.307	0.203
Habitat × region × month	0.717	1	0.717	0.546	0.501
Site(habitat, region) × month	5.251	4	1.313		

Quadratic trends: Source	SS	df	MS	*F*	*P*
Time	13.099	1	13.099	8.897	0.041
Habitat × month	17.935	1	17.935	12.182	0.025
Region × month	1.574	1	1.574	1.069	0.360
Habitat × region × month	0.822	1	0.822	0.558	0.496
Site(habitat, region) × month	5.889	4	1.472		

Cubic trends: Source	SS	df	MS	*F*	*P*
Time	1.696	1	1.696	2.943	0.161
Habitat × month	22.695	1	22.695	39.375	0.003
Region × month	1.401	1	1.401	2.432	0.194
Habitat × region × month	0.167	1	0.167	0.290	0.619
Site(habitat, region) × month	2.306	4	0.576		

The trend analyses indicate that any linear, quadratic or cubic trends through time differ between the two habitats. It is clear from Figure 11.5 that there is little change through time in stringybark habitats but marked declines from the austral autumn and winter through to spring and summer for ironbark habitat.

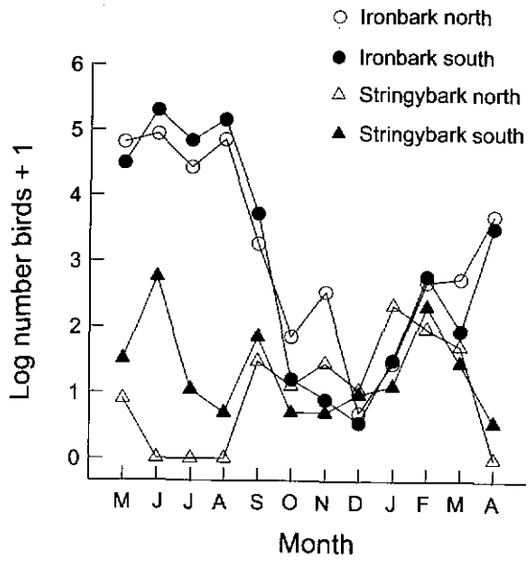

Figure 11.5 Mean $\log_e$ transformed density of birds (+1) for two habitats (ironbark and stringybark forests) and two regions (north and south) for twelve months from McGoldrick & Mac Nally (1998).

The appropriate linear model for a split-plot or repeated measures design with two crossed between-plots/subjects factors is:

$$y_{ijklm} = \mu + \alpha_i + \gamma_k + \alpha\gamma_{ik} + \beta_{j(ik)} + \delta_l + \alpha\delta_{il} + \gamma\delta_{kl} + \alpha\gamma\delta_{ikl} + \beta\delta_{j(ik)l} + \varepsilon_{ijklm} \quad (11.6)$$

From McGoldrick & Mac Nally (1998):

$$\begin{aligned}(\text{log density of nectarivorous birds plus one})_{ijklm} &= \mu + (\text{habitat})_i + (\text{region})_k + \\ &(\text{interaction between habitat and region})_{ik} + \\ &(\text{site within habitat and region})_{j(ik)} + (\text{month})_l + \\ &(\text{interaction between habitat and month})_{il} + \\ &(\text{interaction between region and month})_{kl} + \\ &(\text{interaction between habitat, region and month})_{ikl} + (\text{interaction between site within habitat and region and month})_{j(ik)l} + \varepsilon_{ijklm} \quad (11.7)\end{aligned}$$

In models 11.6 and 11.7 we find the following.

μ is the overall (constant) population mean log density of nectarivorous birds plus one.

α_i is the effect of the ith level of the first between plots factor A (effect of habitat), pooling regions and months.

γ_k is the effect of the kth level of the second between plots factor C (effect of region), pooling habitats and months.

$\alpha\gamma_{ik}$ is the effect of the interaction between the ith level of A and kth level of C (interaction between habitat and region), pooling months.

$\beta_{j(ik)}$ is the effect of the jth plot (factor B, site) within the ikth combination A and C.

δ_l is the effect of the lth level of the within-plots factor D (effect of month).

$\alpha\delta_{il}$ is the effect of the two way interaction between the ith level of A and the lth level of D (interaction between habitat and month).

$\gamma\delta_{kl}$ is the effect of the two way interaction between the kth level of C and the lth level of D (interaction between region and month).

$\alpha\gamma\delta_{ikl}$ is the effect of the three way interaction between the ith level of A, the kth level of C and the lth level of D (interaction between habitat and region and month).

$\beta\delta_{j(ik)l}$ is the effect of the interaction between the jth plot (factor B) within the ikth combination A and C and the lth level of D (interaction between site (within habitat and region) and month).

ε_{ijklm} is the error term. Note that ε_{ijklm} cannot be estimated separately from $\beta\delta_{j(ik)l}$ in this model unless there is replication within each cell, which is unusual. By recording the same sites once at each time, McGoldrick & Mac Nally (1998) did not have replicates within each combination of habitat, region and month and so could not estimate ε_{ijklm}.

The general expected mean squares are in Table 11.5, as well as those for the common case whereby A, C and D are fixed and B (plots or subjects) is random. The between-plots/subjects terms are tested against $MS_{B(AC)}$ and the within-plots/subjects terms are tested against $MS_{DB(AC)}$. Error terms for other combinations can be determined from the expected mean squares and are provided in Table 11.5 following the rules in Box 9.8 – see also Kirk (1995) and Winer *et al.* (1991).

To further illustrate this design, consider the study of Morris (1996) who examined factors affecting the density of rodents in the Rocky Mountains of the USA. He had nine locations, with two habitats (xeric and mesic) at each location (i.e. a 9×2 factorial design), with two replicate grids

Table 11.5 ANOVA table for partly nested design with two crossed between-plots factors and one within-plots factor

Source	df	General expected mean square (EMS)	EMS (A, C, D fixed; B random)	Test
Between plots/subjects				
A	$(p-1)$	$\sigma_\varepsilon^2 + D_q\sigma_{\beta\delta}^2 + qD_t\sigma_{\beta\delta}^2 + qrD_t\sigma_{\alpha\delta}^2 + qrtD_p\sigma_{\alpha\gamma\delta}^2 + tD_q\sigma_\beta^2 + qrt\sigma_\alpha^2$	$\sigma_\varepsilon^2 + t\sigma_\beta^2 + qrt\sigma_\alpha^2$	$MS_A/MS_{B(AC)}$
C	$(r-1)$	$\sigma_\varepsilon^2 + D_q\sigma_{\beta\delta}^2 + qD_p\sigma_{\gamma\delta}^2 + pqD_t\sigma_{\gamma\delta}^2 + tD_q\sigma_\beta^2 + pqt\sigma_\gamma^2$	$\sigma_\varepsilon^2 + t\sigma_\beta^2 + pqt\sigma_\gamma^2$	$MS_B/MS_{B(AC)}$
A×C	$(p-1)(r-1)$	$\sigma_\varepsilon^2 + D_q\sigma_{\beta\delta}^2 + qD_t\sigma_{\alpha\gamma\delta}^2 + tD_q\sigma_\beta^2 + qt\sigma_{\alpha\gamma}^2$	$\sigma_\varepsilon^2 + t\sigma_\beta^2 + qt\sigma_{\alpha\gamma}^2$	$MS_{AB}/MS_{B(AC)}$
B(AC)	$pr(q-1)$	$\sigma_\varepsilon^2 + D_t\sigma_{\beta\delta}^2 + t\sigma_\beta^2$	$\sigma_\varepsilon^2 + t\sigma_\beta^2$	No test
Within plots/subjects				
D	$(t-1)$	$\sigma_\varepsilon^2 + D_q\sigma_{\beta\delta}^2 + qD_p\sigma_{\alpha\gamma\delta}^2 + pqD_r\sigma_{\gamma\delta}^2 + qrD_p\sigma_{\alpha\delta}^2 + pqrD_p\sigma_\delta^2$	$\sigma_\varepsilon^2 + \sigma_{\beta\delta}^2 + pqr\sigma_\delta^2$	$MS_D/MS_{B(AC)D}$
A×D	$(p-1)(t-1)$	$\sigma_\varepsilon^2 + D_q\sigma_{\beta\delta}^2 + qD_r\sigma_{\alpha\gamma\delta}^2 + qr\sigma_{\alpha\delta}^2$	$\sigma_\varepsilon^2 + \sigma_{\beta\delta}^2 + qr\sigma_{\alpha\delta}^2$	$MS_{AD}/MS_{B(AC)D}$
C×D	$(r-1)(t-1)$	$\sigma_\varepsilon^2 + D_q\sigma_{\beta\delta}^2 + qD_p\sigma_{\gamma\delta}^2 + pq\sigma_{\gamma\delta}^2$	$\sigma_\varepsilon^2 + \sigma_{\beta\delta}^2 + pq\sigma_{\gamma\delta}^2$	$MS_{BD}/MS_{B(AC)D}$
A×C×D	$(p-1)(r-1)(t-1)$	$\sigma_\varepsilon^2 + D_q\sigma_{\beta\delta}^2 + q\sigma_{\alpha\gamma\delta}^2$	$\sigma_\varepsilon^2 + \sigma_{\beta\delta}^2 + q\sigma_{\alpha\gamma\delta}^2$	$MS_{ABD}/MS_{B(AC)D}$
B(AC)×D	$(t-1)pr(q-1)$	$\sigma_\varepsilon^2 + \sigma_{\beta\delta}^2$	$\sigma_\varepsilon^2 + \sigma_{\beta\delta}^2$	No test

Note:
Expected mean squares are provided for the general case (see Box 9.8) and for the usual case of A, C and D fixed with B (plots or subjects) random. There is only one observation within each combination of A, B, C and D.

Table 11.6 ANOVA table for partly nested design from Morris (1996) with two crossed between-plots factors (habitat and location, both fixed), one within-plots factor (time, fixed) and grids as random plots. There is only one observation within each combination of A, B, C and D

Source	Source	df	*F*-ratio denominator
Between plots/subjects	*Between grids*		
A	Habitat	1	Grid (habitat, location)
C	Location	8	Grid (habitat, location)
A×C	Habitat×location	8	Grid (habitat, location)
B(AC)	Grid (habitat, location)	18	
Within plots/subjects	*Within grids*		
D	Time	2	Grid (habitat, location)×time
A×D	Habitat×time	2	Grid (habitat, location)×time
C×D	Location×time	16	Grid (habitat, location)×time
A×C×D	Habitat×location×time	16	Grid (habitat, location)×time
B(AC)×D	Grid (habitat, location)×time	36	

for each combination of location and habitat. Grids were thus the plots or subjects and location and habitat were the between plots/subjects factors. He sampled each grid at three times (early, mid, late summer), so sampling time was the within plots/subjects factor. The ANOVA for this study is in Table 11.6, illustrating the appropriate error terms for each effect in the model, based on all factors except grids (i.e. plots) being fixed.

A more complicated version of this design was used by Letourneau & Dyer (1998), who examined the effects of top predators (beetle larvae present or absent), soil type (nutrient rich or poor) and light level (high or low) on colony size of an ant species on seedlings planted in three replicate pots (i.e. plots) in a three factor crossed design. Each pot was recorded on five occasions over 18 months, with time as the within-plots/subjects factor. The ANOVA for this study is in Table 11.7 with error terms based on all factors except plots being fixed.

A further modification of the between-plots/subjects part of the design is where A (the between plots factor) and plots are arranged as a RCB design (Table 11.8). The appropriate linear model for this design is:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \varepsilon_{ijkl} \quad (11.8)$$

In model 11.8:

μ is the overall (constant) population mean,

α_i is the effect of factor A,

γ_k is the effect of factor C,

$\alpha\gamma_{ik}$ is the interaction between factors A and C,

β_j is the effect of plots/subjects (i.e. blocks),

$\alpha\beta_{ij}$, $\beta\gamma_{jk}$, and $\alpha\beta\gamma_{ijk}$ are the interactions between A, C, A×C and plots/subjects, and

ε_{ijkl} is the error effect. Note that ε_{ijkl} cannot be estimated separately from $\alpha\beta\gamma_{ijk}$ in this model unless there is replication within each cell, which is unusual.

This is basically a three factor unreplicated ANOVA, identical to a factorial RCB design (Chapter 10). If A and C are fixed and B (plots/subjects or blocks) is random, then A is tested against A×B (plot), as in all RCB designs, C is tested against C×B and A×C is tested against A×B×C (Table 11.8). There are no tests for plot/subject (i.e. block) or its interactions with A and C, unless quasi *F*-ratios are used.

Aguiar & Sala (1997) used such an analysis in their investigation of seed movement in the Patagonia steppe. They had three sites recorded on three dates and they measured seed availability in four microsites (bare ground, grass, shrub,

Table 11.7 ANOVA table for partly nested design from Letourneau & Dyer (1998) with three crossed between-plots factors (predators, light level, soil; all fixed), one within-plots factor (time, fixed) and plots as random plots

Source	Source	df	*F*-ratio denominator
Between plots/subjects	*Between plots*		
A	Predators	1	Plot (predators, soil, light)
C	Soil	1	Plot (predators, soil, light)
D	Light	1	Plot (predators, soil, light)
A×C	Predators×soil	1	Plot (predators, soil, light)
A×D	Predators×light	1	Plot (predators, soil, light)
C×D	Soil×light	1	Plot (predators, soil, light)
A×C×D	Predators×soil×light	1	Plot (predators, soil, light)
B(ACD)	Plot (predators, soil, light)	16	
Within plots/subjects	*Within plots*		
E	Time	4	Plot (predators, soil, light)×time
A×E	Predators×time	4	Plot (predators, soil, light)×time
C×E	Soil×time	4	Plot (predators, soil, light)×time
D×E	Light×time	4	Plot (predators, soil, light)×time
A×C×E	Predators×soil×time	4	Plot (predators, soil, light)×time
A×D×E	Predators×light×time	4	Plot (predators, soil, light)×time
C×D×E	Soil×light×time	4	Plot (predators, soil, light)×time
A×C×D×E	Predators×soil×light×time	4	Plot (predators, soil, light)×time
B(ACD)×E	Plot (predators, soil, light)×time	64	

litter) in each site on each date. Site and date were between plots factors (although there was only one "plot" for each combination of site and date) and microsite was a within plot factor (Table 11.9). Although not stated in their paper, they treated site as a random block effect and assumed there was no site by date interaction since they tested the random site effect against this interaction term. A second example comes from Evans & England (1996) who looked at the effect of artificial honeydew on the numbers of adult weevil parasitoids on alfalfa plants. They had three treatments (early application of artificial honeydew followed by water, early application of water followed by artificial honeydew, two applications of water only), each allocated to one of three "subplots" in each of ten rows (plots or blocks). The numbers of parasitoids were recorded from each subplot on two separate dates about ten days apart. The ANOVA for this design is also in Table 11.9 and Evans & England (1996) fitted an additive model with no treatment×row interactions, allowing tests for row (i.e. plots) and row×date.

11.8.2 Additional within-plots/subjects factors

Extra within-plots/subjects factors can also be included in these designs, although this complicates the analysis because multiple denominators now must be used for the F tests for the within-subjects/plots terms. With one between-plots factor (A), two within-plots factors (C and D) and plots (factor B) nested within A, the appropriate linear model is:

$$y_{ijklm} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \alpha\gamma_{ik} + \beta\gamma_{j(i)k} + \delta_l + \alpha\delta_{il} + \beta\delta_{j(i)l} + \gamma\delta_{kl} + \alpha\gamma\delta_{ikl} + \beta\gamma\delta_{j(i)kl} + \varepsilon_{ijklm} \quad (11.9)$$

In model 11.9:

μ is the overall (constant) population mean,

α_i is the effect of ith level of factor A (the between-plots factor),

$\beta_{j(i)}$ is the effect of the jth plot (factor B) within the ith level of factor A,

γ_k is the effect of the kth level of factor C (the first within-plots factor),

Table 11.8 | ANOVA table for partly nested design with a RCB between-plots component (A fixed and B = blocks random) and within-plots factor C fixed. The fitted model is non-additive, including A × Blocks interactions. An additive model means that all the ($\alpha\beta$) terms would disappear from the EMS, allowing tests for B (blocks) and B (blocks) × C

Source	df	General expected mean square (EMS)	EMS (A, C fixed, B random)	Test
Between plots/subjects				
A	$(p-1)$	$\sigma_\varepsilon^2 + D_q D_r \sigma^2_{\alpha\beta\gamma} + D_r q \sigma^2_{\alpha\gamma} + D_q r \sigma^2_{\alpha\beta} + qr\sigma_\alpha^2$	$\sigma_\varepsilon^2 + r\sigma^2_{\alpha\beta} + qr\sigma_\alpha^2$	MS_A / MS_{AB}
B (block)	$(q-1)$	$\sigma_\varepsilon^2 + D_p D_r \sigma^2_{\alpha\beta\gamma} + D_r p \sigma^2_{\beta\gamma} + D_p r \sigma^2_{\alpha\beta} + pr\sigma_\beta^2$	$\sigma_\varepsilon^2 + r\sigma^2_{\alpha\beta} + pr\sigma_\beta^2$	No test
A × B	$(p-1)(q-1)$	$\sigma_\varepsilon^2 + D_r \sigma^2_{\alpha\beta\gamma} + r\sigma^2_{\alpha\beta}$	$\sigma_\varepsilon^2 + r\sigma^2_{\alpha\beta}$	No test
Within plots/subjects				
C	$(r-1)$	$\sigma_\varepsilon^2 + D_q D_p \sigma^2_{\alpha\beta\gamma} + D_p q \sigma^2_{\alpha\gamma} + D_q p \sigma^2_{\beta\gamma} + pqr\sigma_\gamma^2$	$\sigma_\varepsilon^2 + p\sigma^2_{\beta\gamma} + pqr\sigma_\gamma^2$	MS_C / MS_{BC}
A × C	$(p-1)(r-1)$	$\sigma_\varepsilon^2 + D_q \sigma^2_{\alpha\beta\gamma} + q\sigma^2_{\alpha\gamma}$	$\sigma_\varepsilon^2 + \sigma^2_{\alpha\beta\gamma} + q\sigma^2_{\alpha\gamma}$	MS_{AC} / MS_{ABC}
B × C	$(q-1)(r-1)$	$\sigma_\varepsilon^2 + D_p \sigma^2_{\alpha\beta\gamma} + p\sigma^2_{\beta\gamma}$	$\sigma_\varepsilon^2 + p\sigma^2_{\beta\gamma}$	No test
A × B × C	$(p-1)\,(q-1)\,(r-1)$	$\sigma_\varepsilon^2 + \sigma^2_{\alpha\beta\gamma}$	$\sigma_\varepsilon^2 + \sigma^2_{\alpha\beta\gamma}$	No test

Table 11.9 | ANOVA table for partly nested designs with a RCB between-plots component

Aguiar & Sala (1997)			Evans & England (1996)		
Source	df	*F*-ratio denominator	Source	df	*F*-ratio denominator
Between plots/subjects			*Between plots/subjects*		
Date	2	Residual (site × date)	Treatment	2	Residual
Site (= block)	2		Row (= block)	9	Residual
Residual (date × site)	4		Residual	18	
Within plots/subjects			*Within plots/subjects*		
Microsite	3	Date × microsite	Date	1	Treatment × row × date
Date × microsite	6	Date × site × microsite	Treatment × date	2	Treatment × row × date
Site × microsite	6		Row × date	9	Treatment × row × date
Date × site × microsite	12		Treatment × row × date	18	

Note:
The example from Aguiar & Sala (1997) has date (fixed) and site (a random blocking factor) as between-plots (blocks) factors and microsite (fixed) as a within-plots (blocks) factor. There is only one observation for each site and date combination and the four microsites were located within each plot (block). The example from Evans & England (1996) has treatment (fixed) and row (a random blocking factor) as between-plots (blocks) factors and date as a within-plots (blocks) factor. Evans & England (1996) fitted an additive model assuming no treatment × row and no treatment × row × date interactions, allowing tests for row and row × date. There is only one observation for each treatment and row combination and each combination was recorded on two dates.

$\alpha\gamma_{ik}$ is the effect of the two way interaction between the ith level of A and kth level of C (i.e. A×C interaction),

$\beta\gamma_{j(i)k}$ is the interaction between the kth level of C and the jth plot (B) within the ith level of A (B within A×C interaction),

δ_l is the effect of the lth level of factor D (the second within-plots factor),

$\alpha\delta_{il}$ is the effect of the two way interaction between the ith level of A and the lth level of D (A×D interaction),

$\beta\delta_{j(i)l}$ is the interaction between the lth level of D and the jth plot (B) within the ith level of A (B within A×D interaction),

$\gamma\delta_{kl}$ is the effect of the two way interaction between the kth level of C and the lth level of D (C×D interaction),

$\alpha\gamma\delta_{ikl}$ is the effect of the three way interaction between the ith level of A, the kth level of C and the lth level of D (A×C×D interaction),

$\beta\gamma\delta_{j(i)kl}$ is the effect of the interaction between the kth level of C and the lth level of D and jth plot (B) within the ith level of A (B within A×C×D interaction), and

ε_{ijklm} is the error effect. Note that ε_{ijklm} cannot be estimated separately from $\beta\gamma\delta_{j(i)kl}$ in this model unless there is replication within each cell, which is unusual.

The general expected mean squares, and those when factors A, C and D are fixed and plots/subjects random, are provided in Table 11.10. Note that when A, C and D are fixed, C and A×C are tested against C×plots within A, D and A×D against D×plots within A and C×D and A×C×D against C×D×plots within A.

These designs are sometimes termed split-split-plot designs because we can have a main between-plots factor and two within-plots factors, one applied to sub-plots within each plot and one applied to sub-sub-plots within each sub-plot. More commonly, however, these experimental designs include a single within-plots factor with repeated measurements through time or two within-subjects time factors. For example, we mention the following.

- Vasquez (1996) looked at the effect of illumination (two fixed levels: bright and dark) and seed distribution (two fixed levels: dispersed and clumped) on seed consumption in experimental arenas for three species of rodents. Species was the between-subjects factor and there were approximately 17 individuals/subjects for each species. Each individual was tested under each illumination level and each seed distribution in a crossed arrangement (four combinations), so illumination and seed distribution were separate within subjects factors (Table 11.11).
- Green (1997) studied the effects of land crabs on recruitment of rainforest seedlings on Christmas Island. He used two habitats (understory and gap) in the rainforest as the between-plots factor with seven paired plots in understory habitat and three paired plots in gap habitat. These pairs were the "plots" or "subjects". One plot (i.e. "sub-plot") in each pair allowed access to crabs and one (sub)plot excluded crabs, so exclusion was one within-plots factor. Additionally, each plot and (sub)plot was recorded monthly for 23 months (although only 22 months were analyzed) so time was a second within-plots factor. This example includes a factor whose levels are allocated to (sub)plots within plots (pairs) plus a factor representing the whole plots recorded through time (Table 11.11).

Other designs can be termed doubly repeated measures designs because the within-plots factors both represent repeated measurements through time. Meserve *et al.* (1996) set up an experiment to examine the effect of predation on the survivorship of degus, a species of rodent. One factor was predation (two fixed levels: predators excluded using fencing and netting and control), with four plots within each level. The number of rodents alive was recorded on each plot at six monthly censuses over four years – year (four fixed levels) and month (six fixed levels) were within-plots factors and were crossed (Table 11.11). In all these examples, there were four different denominators used for testing hypotheses in the ANOVA.

11.8.3 Additional between-plots/subjects and within-plots/subjects factors

These partly nested analyses of variance can be applied to a variety of complex split-plot (repeated

Table 11.10 ANOVA with expected mean squares for partly nested design with one between-plots factor (A), plots/subjects (B) and two crossed within-plots factors (C and D). There is only a single observation within each combination of A, B, C and D

Source	df	EMS general	EMS (A, C, D fixed, B random)	Test
Between plots/subjects				
A	$(p-1)$	$\sigma^2_\varepsilon + D_q D_t \sigma^2_{\beta\delta} + qD_t\sigma^2_{\alpha\gamma\delta} + rD_q\sigma^2_{\beta\delta} + qrD_t\sigma^2_{\alpha\delta}$	$\sigma^2_\varepsilon + rt\sigma^2_\beta + qrt\sigma^2_\alpha$	$MS_A / MS_{B(A)}$
B(A)	$p(q-1)$	$\sigma^2_\varepsilon + D_t\sigma^2_{\beta\gamma\delta} + tD_q\sigma^2_{\beta\gamma} + rD_t\sigma^2_{\beta\delta} + rtD_q\sigma^2_\beta$	$\sigma^2_\varepsilon + rt\sigma^2_\beta$	No test
Within plots/subjects				
C	$(r-1)$	$\sigma^2_\varepsilon + D_q D_t \sigma^2_{\beta\gamma\delta} + qD_p D_t \sigma^2_{\alpha\gamma\delta} + pqD_t\sigma^2_{\gamma\delta} + tD_q\sigma^2_{\beta\gamma} + pqtD_t\sigma^2_\gamma + qtD_p\sigma^2_{\alpha\gamma} + qt_q\sigma^2_{\alpha\gamma}$	$\sigma^2_\varepsilon + t\sigma^2_{\beta\gamma} + pqt\sigma^2_\gamma$	$MS_C / MS_{B(A)C}$
A×C	$(p-1)(r-1)$	$\sigma^2_\varepsilon + D_q D_t \sigma^2_{\beta\gamma\delta} + qD_t\sigma^2_{\alpha\gamma\delta} + tD_q\sigma^2_{\beta\gamma} + qt_q\sigma^2_{\alpha\gamma}$	$\sigma^2_\varepsilon + t\sigma^2_{\beta\gamma} + qt\sigma^2_{\alpha\gamma}$	$MS_{AC} / MS_{B(A)C}$
B(A)×C	$p(q-1)(r-1)$	$\sigma^2_\varepsilon + D_t\sigma^2_{\beta\gamma\delta} + t\sigma^2_{\beta\gamma}$	$\sigma^2_\varepsilon + t\sigma^2_{\beta\gamma}$	No test
D	$(t-1)$	$\sigma^2_\varepsilon + D_q D_r \sigma^2_{\beta\gamma\delta} + qD_p D_r \sigma^2_{\alpha\gamma\delta} + pqD_r\sigma^2_{\gamma\delta} + rD_q\sigma^2_{\beta\delta} + pqrD_t\sigma^2_\delta + qrD_p\sigma^2_{\alpha\delta} + qr\sigma^2_{\alpha\delta}$	$\sigma^2_\varepsilon + r\sigma^2_{\beta\delta} + pqr\sigma^2_\delta$	$MS_D / MS_{B(A)D}$
A×D	$(p-1)(t-1)$	$\sigma^2_\varepsilon + D_q D_r \sigma^2_{\beta\gamma\delta} + qD_r\sigma^2_{\alpha\gamma\delta} + rD_q\sigma^2_{\beta\delta} + qr\sigma^2_{\alpha\delta}$	$\sigma^2_\varepsilon + r\sigma^2_{\beta\delta} + qr\sigma^2_{\alpha\delta}$	$MS_{AD} / MS_{B(A)D}$
B(A)×D	$p(q-1)(t-1)$	$\sigma^2_\varepsilon + D_r\sigma^2_{\beta\gamma\delta} + r\sigma^2_{\beta\delta}$	$\sigma^2_\varepsilon + r\sigma^2_{\beta\delta}$	No test
C×D	$(r-1)(t-1)$	$\sigma^2_\varepsilon + D_q\sigma^2_{\beta\gamma\delta} + qD_p\sigma^2_{\alpha\gamma\delta} + pq\sigma^2_{\gamma\delta}$	$\sigma^2_\varepsilon + \sigma^2_{\beta\gamma\delta} + pq\sigma^2_{\gamma\delta}$	$MS_{CD} / MS_{B(A)CD}$
A×C×D	$(p-1)(r-1)(t-1)$	$\sigma^2_\varepsilon + D_q\sigma^2_{\beta\gamma\delta} + q\sigma^2_{\alpha\gamma\delta}$	$\sigma^2_\varepsilon + \sigma^2_{\beta\gamma\delta} + q\sigma^2_{\alpha\gamma\delta}$	$MS_{ACD} / MS_{B(A)CD}$
B(A)×C×D	$p(q-1)(r-1)(t-1)$	$\sigma^2_\varepsilon + \sigma^2_{\beta\gamma\delta}$	$\sigma^2_\varepsilon + \sigma^2_{\beta\gamma\delta}$	No test

Table 11.11 Examples of partly nested designs from the literature with one between-plots factor and two crossed within-plots (subjects) factors. There is only a single observation within each combination of A, B, C and D. See Section 11.8.2 for more details of specific examples.

General source	*F*-ratio denominator	Meserve *et al.* (1996) Source	df	Vasquez (1996) Source	df	Green *et al.* (1997) Source	df
Between plots/subjects							
A	B(A)	Predation	1	Species		Habitat	1
B(A)		Plots(predation)	6	Subject(species)		Pairs(habitat)	8
Within plots/subjects							
C	B(A) × C	Year	3	Illumination		Exclusion	1
A × C	B(A) × C	Year × predation	3	Illumination × species		Exclusion × habitat	1
B(A) × C		Plots(predation) × year	18	Subject(species) × illumination		Pairs(habitat) × exclusion	8
D	B(A) × D	Month	5	Distribution		Time	21
A × D	B(A) × D	Month × predation	5	Distribution × species		Time × habitat	21
B(A) × D		Plots(predation) × month	30	Subject(species) × distribution		Pairs(habitat) × time	168
C × D	B(A) × C × D	Year × month	15	Illumination × distribution		Exclusion × time	21
A × C × D	B(A) × C × D	Year × month × predation	15	Illumination × distribution × species		Exclusion × time × habitat	21
B(A) × C × D		Plots(predation) × year × month	90	Subject(species) × illumination × distribution		Pairs(habitat) × exclusion × time	168

measures) experimental designs that include multiple between-plots/subjects factors and multiple within-subjects/plots factors. We will use the study of Gough & Grace (1998) on the effects of herbivores and productivity levels on plant species densities to illustrate such a complex design. They chose two freshwater marshes on a river near the Louisiana/Mississippi border in eastern USA. In each marsh, they established eight fenced areas (plots), to exclude herbivores like rabbit, muskrat, etc., and eight unfenced areas. So the between plots component of the design had two fixed factors (marsh and fence) in a crossed arrangement with replicate areas (i.e. plots). There were three sub-plots within each fenced or unfenced area and each sub-plot received one of three nutrient enrichment treatments (no addition, nutrient addition, and natural soil addition). So enrichment was the first within-plots factor. Additionally, each sub-plot was also censused seven times over two years, so time was a second within-plots factor. All factors were considered fixed except for area (i.e. plot). The resulting ANOVA model (Table 11.12) had 19 terms and four different denominators for testing hypotheses.

11.8.4 General comments about complex designs

Gumpertz & Brownie (1993) discussed split-plot designs that include repeated measures (usually multiple times) in some detail. They recommended using trend analyses to examine patterns in the repeated factor and its interactions with the other factors in the design (Section 11.5.3). They also recommended against analyzing such designs as univariate split-split-plot designs, i.e. using the partly nested models we have described, because of the assumption of sphericity of variance–covariance matrices across times, and preferred a multivariate approach. We agree that the sphericity assumption may be important but instead recommend epsilon-adjusted univariate tests in addition to the multivariate tests. Winer *et al.* (1991) and Kirk (1995) provide details of these complex designs and approaches to analyses; they also provide general formula for determining EMS for any combination of fixed and random factors. Kirk's (1995) unique terminology adapts well to these designs.

11.9 Partly nested designs and statistical software

Data files for these partly nested analyses can be set up in two ways. First, we could create a file for a classical "split-plot analysis" with each factor in a separate column (Table 11.13). A partly nested linear model is then fitted and most software requires that all terms are specified in the model and each term specifically tested against the appropriate denominator. Only unadjusted univariate tests are usually provided but this approach provides great flexibility in structuring the model and choosing denominators for *F* tests. Second, the data can be coded for a "repeated measures analysis", with between-subjects factors coded as usual but the different levels of the within-subjects factors are in individual columns (Table 11.13). If you have replicate observations within each cell, it can be difficult to code the data file for "repeated measures" analysis and you must either just use cell means or switch to the "split-plot" set up. Software using the "repeated measures" approach nearly always assumes B(A) is random and all other factors are fixed but provides additional output, including estimates of ε (for multisample sphericity), adjusted and unadjusted univariate tests, multivariate tests, and polynomial trend analyses; it also explicitly distinguishes "between-subjects" and "within-subjects" components of the ANOVA. Note that the unadjusted univariate tests will be identical to those provided by the first analysis. The important point is that although the two univariate analyses are functionally identical, the alternative analyses (adjusted univariate, MANOVA) and automatic extras (profile analyses) will often only be provided when the data are coded for a classical repeated measures design, not for a split-plot. The profile analyses can usually also be obtained by including contrasts as part of a split-plot analysis.

Table 11.12 ANOVA table from complex partly design from Gough & Grace (1998) with two crossed and fixed between-plots/subjects factors and two crossed and fixed within-plots/subjects factors – see Section 11.8.3

Source	Source	df	*F*-ratio denominator
Between plots	*Between plots*		
A	Marsh	1	Plots (marsh, fence)
B	Fence	1	Plots (marsh, fence)
A × B	Marsh × fence	1	Plots (marsh, fence)
C(AB)	Plots (marsh, fence)	28	
Within plots	*Within plots*		
D	Enrichment	2	Plots (marsh, fence) × enrichment
A × D	Marsh × enrichment	2	Plots (marsh, fence) × enrichment
B × D	Fence × enrichment	2	Plots (marsh, fence) × enrichment
A × B × D	Marsh × fence × enrichment	2	Plots (marsh, fence) × enrichment
C(AB) × D	Plots (marsh, fence) × enrichment	56	
E	Time	6	Plots (marsh, fence) × time
A × E	Marsh × time	6	Plots (marsh, fence) × time
B × E	Fence × time	6	Plots (marsh, fence) × time
A × B × E	Marsh × fence × time	6	Plots (marsh, fence) × time
C(AB) × E	Plots (marsh, fence) × time	168	
D × E	Enrichment × time	12	Plots (marsh, fence) × enrichment × time
A × D × E	Marsh × enrichment × time	12	Plots (marsh, fence) × enrichment × time
B × D × E	Fence × enrichment × time	12	Plots (marsh, fence) × enrichment × time
A × B × D × E	Marsh × fence × enrichment × time	12	Plots (marsh, fence) × enrichment × time
C(AB) × D × E	Plots (marsh, fence) × enrichment × time	336	

Table 11.13 Data coding for unreplicated "split-plot" analysis and for unreplicated "repeated measures" analysis

Data file for "split-plot" analysis

Factor A	Plots/subjects (B)	Factor C	Y
1	1	1	y_{111}
1	1	2	y_{112}
1	2	1	y_{121}
1	2	2	y_{122}
2	3	1	y_{231}
i	j	k	y_{ijk}

Data file for "repeated measures" analysis

Factor A	Plots/subjects (B)	C_1	C_k
1	1	y_{111}	y_{11k}
1	2	y_{121}	y_{12k}
1	3	y_{231}	y_{23k}
i	j	y_{ij1}	y_{ijk}

11.10 General issues and hints for analysis

11.10.1 General issues

- Partly nested designs are very commonly used in biology, as ways to use resources more economically – save money, kill fewer organisms, etc. There is a cost to this rationalization, as the statistical models have more assumptions than completely randomized factorial designs.
- Although they are treated differently in many textbooks, unreplicated partly nested, split-plot and repeated measures ("groups × trials") designs are analyzed with an identical linear model. For repeated measures designs, this model is usually described with a larger set of assumptions, which imposes more restrictions on the analysis. We recommend that, because the two designs (split-plot and repeated measures) require identical models, you should examine the larger set of assumptions for all partly nested designs.
- Unreplicated partly nested designs, i.e. those with only a single observation of each level of C for each plot/subject within each level of A, prevent you from testing one higher-order interaction or require that you assume that interaction to be zero, depending on the exact design. In the usual situation of all factors being fixed except B (i.e. plots or subjects), this does not preclude tests of the fixed factors or their interactions.
- These designs can include additional factors, both between-plots/subject and within-plots/subjects. Once the model is decided, the analysis is straightforward except that care must be taken to determine the correct *F*-ratios depending on which factors are fixed and which are random.

11.10.2 Hints for individual analyses

- These designs are complex, and generally have mixtures of fixed and random factors. As a first step, before doing the experiment, write out the linear model, the ANOVA table, and include details of the *F*-ratios.
- The different designs will change the df, and hence the power, of many of your tests of hypotheses. Before doing the experiment, look at all of the relevant degrees of freedom, and decide whether this arrangement of your experimental units and resources will give you the best compromise between power and cost.
- The assumption of normality is less a problem for the between-plots/subjects factors, as those analyses effectively use means of other data, allowing the Central Limit Theorem to be invoked.
- Tests of the between-plots/subjects factors assume homogeneity of between-group variances.
- The assumption of sphericity is important for the tests of within-plots/subjects factors and incorporates the homogeneity of variance assumption for this component of the analysis. Examine the various measures of the validity of this assumption (particularly the Greenhouse–Geiser and Huynh–Feldt estimates of ε), and, if $\hat{\varepsilon}$ values are low, use the conservative corrections to the *F* tests or the MANOVA approach. There is no agreed-upon test for the assumption of multisample sphericity.

- If the design is unreplicated, consider coding the data file up as repeated measures, allowing you to routinely get the $\hat{\varepsilon}$ values, corrected univariate *F*-ratios, and the multivariate equivalents. We follow Looney & Stanley (1989), and suggest you look for a significant result in either the univariate or multivariate analyses.

Chapter 12

Analyses of covariance

In Chapter 10, we described a technique for reducing the residual or unexplained variation in an experiment by grouping experimental units into spatial or temporal blocks. Another approach to reducing the residual variation is to measure one or more concomitant continuous variables for each experimental unit along with the response variable. These concomitant variables, or covariates, are usually considered as continuous predictor variables, with the one or more factors being categorical predictors. A linear models analysis of this design is sometimes called an analysis of covariance (ANCOVA), where the effect of the covariate on the response variable is removed from the unexplained variability by regression analysis. The final ANCOVA tests the difference between factor level means, adjusted for the effect of the covariate.

Another use of ANCOVA is to compare the slopes and/or intercepts of two or more regression lines, although this use is less common. We will cover basic methods for ANCOVA in this chapter, but also pay particular attention to complex designs and situations with regression slopes that are heterogeneous between the factor levels (see also Figure 12.1).

12.1 Single factor analysis of covariance (ANCOVA)

The simplest ANCOVA design is one analogous to a single factor ANOVA where we have a single categorical predictor variable (factor). In addition to a single continuous response variable, we also record the value of a continuous covariate from each experimental or sampling unit. Some examples from the biological literature include the following.

- Tollrian (1995) studied the effect of a chemical cue (kairomone) released by predators (midge larva *Chaoborus*) on morphology of the aquatic cladoceran *Daphnia*. The response variable was body mass of *Daphnia*, the factor was kairomone treatment (two levels: presence, resulting in neckteeth-induced morphs, and absent, resulting in typical morphs) and the covariate was body length. If body length explains some of the variation in body mass, a more powerful test of kairomone treatment will be obtained.

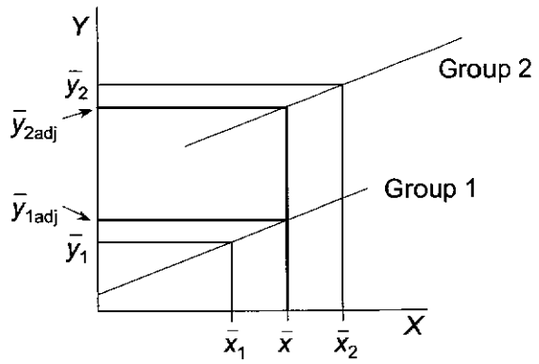

Figure 12.1 Diagrammatic representation of adjusted means in ANCOVA. The adjusted *Y* means are based on the overall *X* mean, not the *X* means for each group. Note that the difference between the adjusted *Y* means is smaller than the difference between the unadjusted *Y* means, although this does not always occur in ANCOVA adjustment.

- Mothershead & Marquis (2000) looked at the effects of increased leaf damage (two levels: natural herbivore damage, artificially increased damage mimicking increased herbivory) on floral traits of flowers of the perennial herb *Oenothera macrocarpa* in Missouri, USA. The response variables were corolla diameter and floral tube length, changes in which would result in changes in pollinator preference and efficiency. They used flower order (successive seasonal flowering) as a covariate to help explain some of the variation in floral traits and provide a more powerful test of damage effects.

We illustrate ANCOVA with two examples from the biological literature.

Sex and fruitfly longevity

Partridge & Farquhar (1981) examined the effect of number and type of mating partners on longevity (response variable) of fruitflies. There was a single factor (partner type) with five treatments: one virgin female per day, eight virgin females per day, a control group with one newly inseminated female per day, a control group with eight newly inseminated females per day, a control group with no females. Also, the thorax length of each individual fly was recorded as a covariate. If thorax length explains some of the variation in longevity, then the test of the effect of partner type on longevity adjusted for thorax length will be more powerful. The analysis of these data is presented in Box 12.1.

Shrinking in sea urchins

Constable (1993) studied the role of sutures (joins between plates in the test) in the shrinking of the test of the sea urchin *Heliocidaris erythrogramma*. He compared widths of inter-radial sutures (mm), the response variable, between urchins kept under high and low food regimes and an initial sample, the factor with three groups, with body volume (ml, cube root transformed) as the covariate. The analysis of these data is presented in Box 12.2.

Box 12.1 Worked example of ANCOVA: sex and fruitfly longevity

Partridge & Farquhar (1981) studied the effect of number of mating partners on longevity of fruitflies. There were five treatments: one virgin female per day, eight virgin females per day, a control group with one newly inseminated female per day, a control group with eight newly inseminated females per day, a control group with no females. Also, the thorax length of each individual fly was recorded as a covariate. If thorax length explains some of the variation in longevity, then the test of the effect of treatments on longevity adjusted for thorax length will be more powerful. The raw data were extracted by reading from Figure 2 in the original paper (see also description and discussion in Hanley & Shapiro 1994). Our general H_0 was that there was no effect of partner treatment on longevity of male fruitflies, adjusting for thorax length.

An ANCOVA model relating longevity to treatment group with thorax length as a covariate (model 12.2) was fitted and the model residuals examined. The plot of residuals against predicted longevity showed evidence of heterogeneous variances (Figure 12.2(a)). The model was refitted with $\log_{10}$ transformation of longevity. The residual plot was much improved with consistent variances for different levels of the covariate (Figure 12.2(b)). There was no indication that the treatments affected thorax length (ANOVA on thorax length: $F_{4,120} = 1.26$, $P = 0.289$).

The specific H_0 was that there was no effect of partner treatment on $\log_{10}$ longevity of male fruitflies, adjusting for thorax length.

The ANCOVA from the fit of the model based on $\log_{10}$ longevity against treatment group with thorax length as a covariate is as follows.

Source	df	MS	F	P
Treatment	4	0.196	27.97	<0.001
Thorax	1	1.017	145.44	<0.001
Residual	119	0.007		

There was a significant difference between adjusted treatment means. The pooled within-groups regression coefficient was 1.194. The $MS_{Residual}$ for an ANOVA on $\log_{10}$ longevity (without thorax as covariate) was 0.015 with 120 df, so including a covariate has reduced the unexplained variation by around 50%.

Adjusted and unadjusted OLS treatment means were as follows.

Treatment	Adjusted mean	Unadjusted mean
1	1.808	1.789
2	1.771	1.789
3	1.794	1.799
4	1.717	1.737
5	1.589	1.564

The standard errors were 0.017 for adjusted means and 0.025 for unadjusted means. Note that the covariance adjustment reduced the mean $\log_{10}$ longevity of treatment one relative to treatments two and three.

The test for homogeneity of within-groups regression slopes was done by fitting a model that related $\log_{10}$ longevity to treatment group, thorax length and the interaction between treatment group and thorax length, the latter term testing the H_0 of equal slopes.

Source	df	MS	F	P
Treatment × thorax length	4	0.011	1.56	0.189
Residual	115	0.007		

The null hypothesis of equal within-group regression slopes was not rejected and it is clear from Figure 12.3 that there was little evidence for non-parallel slopes.

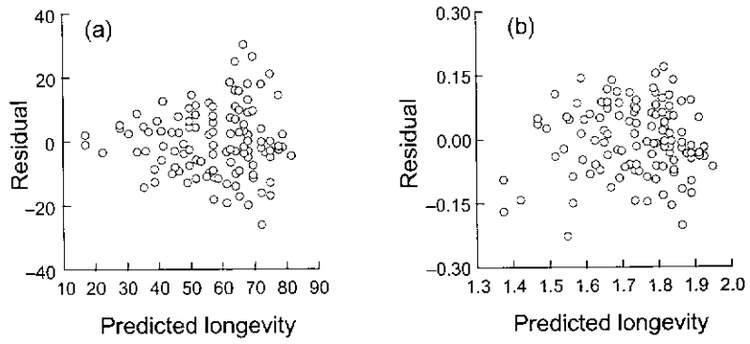

Figure 12.2 Plots of residuals versus predicted values of the response variable from ANCOVA models fitted to data from Partridge & Farquar (1981). (a) Untransformed longevity and (b) $\log_{10}$-transformed longevity.

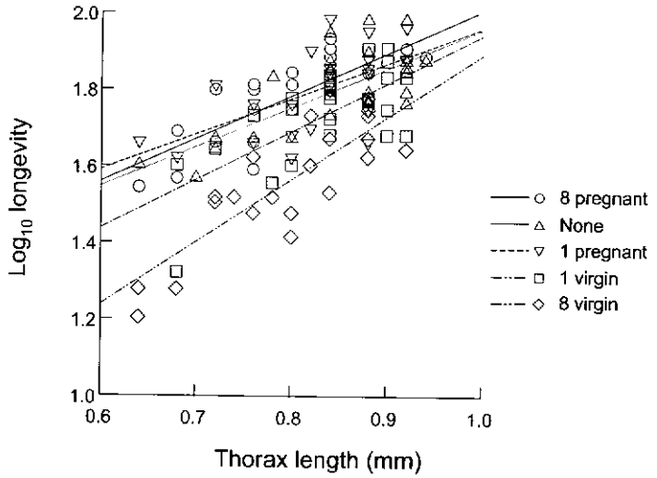

Figure 12.3 Scatterplots with linear regression lines of $\log_{10}$ longevity (days) against thorax length (mm) for male fruitflies under each of the five partner treatment groups. The number and type of female partners for each treatment group are indicated in the legend.

Box 12.2 Worked example of ANCOVA: shrinking in sea urchins

Constable (1993) studied the role of sutures in the shrinking of the test of the sea urchin *Heliocidaris erythrogramma*. He compared widths of inter-radial sutures (mm) between urchins kept under high and low food regimes and an initial sample (one factor with three groups) with body volume (ml, cube root transformed) as the covariate and $n = 24$ urchins in each group. There was a significant interaction between the factor and the covariate ($F_{2,66} = 4.701$, $P = 0.012$), indicating heterogeneous slopes (Figure 12.4). Constable (1993) used the Wilcox modification of the Johnson–Neyman procedure (Box 12.4) to determine over which values of body volume the groups were significantly different.

Initial > Low food for cube root volume > 2.95
High food > Initial for cube root volume > 1.81
High food > Low food for cube root volume > 2.07

So initial suture width was greater than low food suture width for body volumes greater than 2.95, high food suture width was greater than initial for volumes greater than 1.81 and high food suture width was greater than low food suture width for volumes greater than 2.07.

12.1.1 Linear models for analysis of covariance

Linear effects model

So far, we have focused on linear models where all the predictors are continuous (classical regression analyses in Chapters 5 and 6) or categorical (classical analyses of variance in Chapters 8–11). In Chapter 6, we explained how a linear model could include both categorical (factors) and continuous (covariates) predictors. Now consider a data set where factor A is a fixed categorical predictor variable with p groups ($i = 1$ to p), X is a continuous predictor variable and we have a continuous response variable Y, with both Y and X recorded for each experimental or sampling unit within each group. In the example from Partridge & Farquhar (1981), factor A is partner

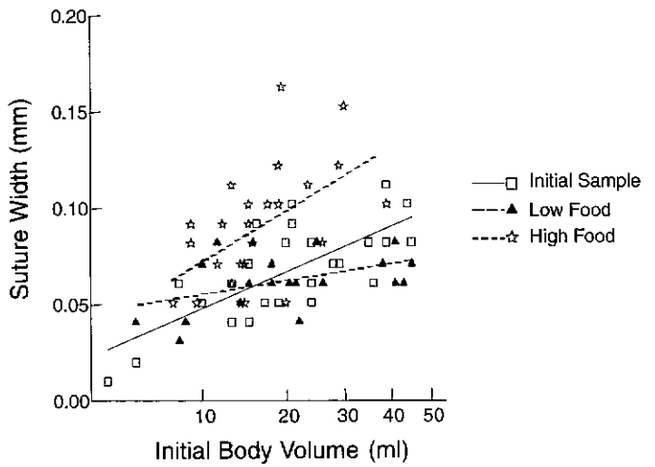

Figure 12.4 Scatterplots with linear regression lines of suture width against cube root transformed body volume for sea urchins under each of the three food level groups. Treatment groups: low food, high food and initial sample.

type ($p = 5$), X is thorax length, the response variable Y is longevity and each experimental unit is a fruitfly. In Constable's (1993) study, factor A is food regime ($p = 3$), X is body volume, Y is suture width with individual urchins as the experimental units.

The ANCOVA model is a linear model with one continuous predictor (covariate) and one categorical predictor (factor) but where we focus on the effects of the factor levels, adjusted for the covariate. The ANCOVA model is best considered as an "ANOVA" model with a covariate included, rather than a "regression" model with a categorical predictor.

The usual form of the ANCOVA model is:

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij} \qquad (12.1)$$

The details of the linear ANCOVA model, including estimation of its parameters and means, are provided in Box 12.3. Note that if there is no relationship between the response variable and the covariate, i.e. $\beta = 0$, then model 12.1 simply becomes the single factor ANOVA model described in Chapter 8. If there are no effects of the treatments, i.e. all $\alpha_i = 0$, then model 12.1 becomes a simple linear regression model described in Chapter 5. These reduced models will be discussed further in Section 12.1.4.

Box 12.3 The linear ANCOVA model and its parameters

Consider a data set with n observations ($j = 1$ to n) where factor A is a categorical predictor variable with p groups ($i = 1$ to p), X is a continuous predictor variable and we have a continuous response variable Y, with both Y and X recorded for each experimental or sampling unit within each group. Based on the Constable (1993) example, we could code factor A as two dummy variables (Chapter 6), so that X_1 equals 1 for high food and 0 for otherwise and X_2 equals 1 for low food and 0 for otherwise, and call the covariate X_3. The (multiple) linear (regression) model we could fit to these data, explicitly ignoring the group structure is:

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3} + \varepsilon_j$$

From Constable (1993):

$$(\text{suture width})_j = \beta_0 + \beta_1(\text{high food vs initial})_j + \beta_2(\text{low food vs initial})_j + \beta_3(\text{body volume})_j + \varepsilon_j$$

In these two models we have the following.

y_j is the jth replicate observation of the response variable, e.g. the suture width for the jth urchin.

β_0 is the intercept of the linear model, the mean of Y when X_1, X_2, and $X_3 = 0$, e.g. the suture width for an urchin with zero body volume in the initial sample.

β_1 is the partial regression slope for X_1, e.g. the regression slope relating suture width to the difference between high food and initial sample groups, holding the difference between low food and initial sample groups, and body volume, constant.

β_2 is the partial regression slope for X_2, e.g. the regression slope relating suture width to the difference between low food and initial sample groups, holding the difference between high food and initial sample groups, and body volume, constant.

β_3 is the partial regression slope for X_3, e.g. the regression slope relating suture width to body volume, holding the difference between high food and initial sample groups, and low food and initial sample groups, constant.

ε_j is random or unexplained error associated with the jth replicate observation.

The interpretations here are those of a standard multiple regression with one continuous and one categorical predictor (Chapter 6). No specific adjustment is made to the values of the response variable or the means of the response variable for each group, although the interpretation of each regression coefficient is based on holding the other predictors constant.

The usual form of the ANCOVA model is:

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij}$$

In this model we have the following.

y_{ij} is the value of the response variable for jth observation in the ith level of factor A.

μ is the overall (constant) mean value of the response variable.

α_i is effect of ith level factor A, defined as the difference between each A mean and the overall mean $(\mu_i - \mu)$.

β is a combined regression coefficient representing the pooling of the regression slopes of Y on X within each group. A basic assumption is that the regression slopes within each group are the same, otherwise pooling them to produce β can result in interpretation of factor effects that are misleading.

x_{ij} is the covariate value for the jth replicate observation from the ith level of factor A.

$\bar{x}$ is the mean value of the covariate.

ε_{ij} is random or unexplained error associated with the jth replicate observation from the ith level of factor A, representing the component of the response variable not explained by the effects of the factor or the relationship with the covariate. These error terms are assumed to be normally distributed at each level of factor A, with a mean of zero $[E(\varepsilon_{ij}) = 0]$ and a variance of σ_ε^2.

This model is overparameterized so, when factor A is fixed, then the usual constraint $\sum_{i=1}^{p} \alpha_i = 0$ applies, so that parameters in the effects model can be estimated.

Note that we have centered the X-values, by subtracting the mean. If we don't, the model is:

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$$

and μ is now a population intercept (for $X = 0$) rather than an overall population mean of Y. It doesn't matter for the partitioning of variance and testing hypotheses which version we use, although the first is most common in the literature.

The focus in the usual ANOVA models is on estimating group or cell means. In ANCOVA models, we wish to estimate group means adjusted for the effects of the covariate, i.e. adjusted means. These are the means of the adjusted values of the response variable defined in expression 12.7. For group i, the adjusted mean represents the mean value of the response variable if the mean value of the covariate for that group equals the overall mean value for the covariate:

$$\mu_{i(\text{adjusted})} = \mu_i - \beta(\bar{x}_i - \bar{x})$$

This is estimated by:

$$\mu_{i(\text{adj})} = \bar{y}_i - b(\bar{x}_i - \bar{x})$$

The standard error of the adjusted mean is:

$$s_{\bar{y}_{i(\text{adjusted})}} = \sqrt{MS_{\text{Residual}}\left(\frac{1}{n_i} + \frac{(\bar{x}_i - \bar{x})^2}{SS_{\text{Residual}(X)}}\right)}$$

where MS_{Residual} is from the ANCOVA partitioning of variation (Table 12.1) and $SS_{\text{Residual}(X)}$ is from an ANOVA on the covariate.

We may also wish to estimate β, the pooled within-groups regression coefficient relating Y to X. Unfortunately in terms of computation, this is neither the estimate of the regression slope of Y on X pooling all observations, as pointed out above, nor is it a simple average of the within-group regression slope estimates. Fortunately, the general linear model routine in most statistical software will provide this estimate (b) and its standard error (s_b), although the former can be calculated from:

$$b = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)}{SS_{\text{Residual}(X)}}$$

where the numerator is the sum, across groups, of the covariance between Y and X within each group.

From Partridge & Farquhar (1981):

$$(\text{longevity})_{ij} = \text{overall mean} + (\text{partner treatment})_i + \beta[(\text{thorax length})_{ij} - (\text{mean thorax length})] + \varepsilon_{ij} \quad (12.2)$$

From Constable (1993):

$$(\text{suture width})_{ij} = \text{overall mean} + (\text{food treatment})_i + \beta[(\text{body volume})_{ij} - (\text{mean body volume})] + \varepsilon_{ij} \quad (12.3)$$

In models 12.1 and 12.3 we have the following:

y_{ij} is the value of suture width for the jth urchin in the ith food treatment.

μ is the overall (constant) mean value of suture width.

α_i is effect of ith food treatment on suture width. This effect is defined as the difference between each food treatment mean and the overall mean $(\mu_i - \mu)$.

Table 12.1 | ANOVA table for single factor ANCOVA based on factor A being fixed and response variable adjusted for the effects of the covariate

Source of variation	df	Mean square	Expected mean square	*F*-ratio
Factor A (adjusted)	$(p-1)$	$\frac{SS_{A(adjusted)}}{(p-1)}$	$\sigma_\varepsilon^2+\frac{n\sum_{i=1}^{n}\alpha_i^2}{p-1}$	$\frac{MS_{A(adjusted)}}{MS_{Residual(adjusted)}}$
Residual (adjusted)	$p(n-1)-1$	$\frac{SS_{Residual(adjusted)}}{p(n-1)-1}$	σ_ε^2	
Total (adjusted)	$pn-2$			

β is a combined regression coefficient representing the pooling of the regression slopes relating suture width to body volume within each food treatment group. A basic assumption is that the regression slopes within each group are the same, otherwise pooling them to produce β can result in interpretations of factor effects that are misleading.

x_{ij} is the value of body volume for the jth urchin from the ith food level group.

ε_{ij} is random or unexplained error associated with the jth urchin in the ith food level group not explained by the food treatment or the body volume.

Although our model includes both effects of a categorical predictor (α_i) on the response variable and the slope (β) of a regression line relating a continuous predictor to the response variable, the interpretation of the parameters is familiar. We measure the effects of treatments (factor A) adjusting for the covariate, i.e. holding it constant. The ANCOVA can therefore be considered as an ANOVA on data adjusted by the regression slope of Y on the covariate X. Each adjusted observation (the value of an observation "corrected" for the effects of the covariate) can be expressed as:

$$y_{ij(adj)} = y_{ij} - \beta(x_{ij} - \bar{x}) = \mu + \alpha_i + \varepsilon_{ij} \quad (12.4)$$

These adjusted observations are also the residuals from the fit of a regression model of Y on X (Winer *et al.* 1991). In model 12.4, α_i is effect of ith level factor A, adjusted for the effects of the covariate ($\mu_{i(adj)} - \mu_{(adj)}$). Substituting the OLS estimate of the pooled within-groups regression slope, we obtain:

$$y_{ij(adj)} = y_{ij} - b(x_{ij} - \bar{x}) \quad (12.5)$$

Each adjusted value is the value of Y for an observation in any group adjusted (centered) to the mean value of the covariate. For example, the suture width of an urchin is adjusted for the effects of the covariate X by subtracting a term that represents a shift, using the regression of suture width on body volume, of the body volume for that urchin to the mean body volume of all urchins in the experiment:

$$(\text{suture width})_{ij(adj)} = (\text{suture width})_{ij} - b[(\text{body volume})_{ij} - (\text{mean body volume})] \quad (12.6)$$

The focus in the usual ANOVA models is on estimating group or cell means. In ANCOVA models, we wish to estimate group means adjusted for the effects of the covariate, i.e. adjusted group means. These are the means of the adjusted values of the response variable defined in Equation 12.4 and, for group i, represent the mean value of the response variable if the mean value of the covariate for that group equals the overall mean value for the covariate:

$$\mu_{i(adjusted)} = \mu_i - \beta(\bar{x}_i - \bar{x}) \quad (12.7)$$

This is estimated by:

$$\bar{y}_{i(adj)} = \bar{y}_i - b(\bar{x}_i - \bar{x}) \quad (12.8)$$

From Constable (1993):

$$(\text{mean suture width})_{i(adj)} = (\text{mean suture width})_i - b[(\text{mean body volume})_i - (\text{overall mean body volume})] \quad (12.9)$$

Details on estimating adjusted means and their standard errors are provided in Box 12.3.

Table 12.2 | Analyses of variance from Partridge & Farquhar (1981) for $\log_{10}$ longevity of fruitflies for different partner treatments, showing ANOVA on Y ($\log_{10}$ longevity), regression of Y ($\log_{10}$ longevity) on covariate X (thorax length) and ANCOVA on Y ($\log_{10}$ longevity) adjusting for covariate X (thorax length). The SS_{Total} and df_{Total} for ANCOVA sum $SS_{Treatment}$ and $SS_{Residual}$ for data adjusted for effects of covariate

	ANOVA Y		Regression Y on X		ANCOVA Y	
Source	SS	df	SS	df	SS	df
Treatment	0.977	4			0.783	4
Regression on thorax length			1.212	1	1.017	1
Residual	1.850	120	1.615	123	0.833	119
Total	2.827	124	2.827	124	1.615	123

Predicted values and residuals

In practice, the ANCOVA model fitted is that in Equation 12.4 where a single factor ANOVA model is fitted to observations adjusted for the effects of the covariate. The predicted values from this model are based on the regression adjustment and the treatment group:

$$\hat{y}_{ij} = \bar{y}_i - b(\bar{x}_i - \bar{x}_{ij}) \quad (12.10)$$

These predicted values are different for each observation within each group, in contrast to the ANOVA model where the predicted values within each group were the same, i.e. the group mean. The residuals from the fitted ANCOVA model are the differences between each observed Y-value and the predicted Y-value:

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i + b(\bar{x}_i - x_{ij}) \quad (12.11)$$

These residuals in Equation 12.11 incorporate the effects of both the continuous covariate and the categorical factor. As for all linear models, residuals provide the basis of the OLS estimate of σ_ε^2 and they are valuable diagnostic tools for checking assumptions and fit of our model.

12.1.2 Analysis of (co)variance

The $SS_{Total(adj)}$ from the ANCOVA is simply the SS_{Total} from an ANOVA on Y less the $SS_{Regression}$ from a linear regression of Y on X, the latter representing the adjustment to the Y-values based on the relationship between Y and X. This $SS_{Total(adj)}$ can be partitioned into that due to the difference between adjusted A group means ($SS_{A(adj)}$) and that not explained by factor A ($SS_{Residual(adj)}$). The $df_{A(adj)}$ is the number of groups minus one and the $df_{Residual(adj)}$ is the total number of observations minus the number of groups minus one for the regression of Y on the covariate. These sum to the $df_{Total(adj)}$, the total number of observations minus two (one for the regression of Y on the covariate). The mean squares are the SS divided by the df as usual and the expected values of these mean squares are identical to those from a single factor ANOVA (Chapter 8), except that the analysis is based on Y-values adjusted for the covariate.

The relationship between the analyses of variance from fitting a single factor ANOVA model to unadjusted Y-values, a simple regression model fitted to unadjusted Y-values against the covariate, and the ANCOVA model fitted to adjusted Y-values is illustrated for the data from Partridge & Farquhar (1981) in Table 12.2. The $SS_{Total(adj)}$ represents the total variation in unadjusted Y (SS_{Total} from the ANOVA on Y) less that explained by the regression of unadjusted Y on X across the whole data set ($SS_{Regression}$ from regression analysis on complete data set). The unexplained variation in unadjusted Y ($SS_{Residual}$ from ANOVA on Y) is split into the variation due to the pooled within-groups regression of Y on X (SS for the covariate from ANCOVA) and the variation in adjusted Y not explained by the treatment groups ($SS_{Residual(adj)}$ from ANCOVA). Note that the $SS_{Regression}$ from the whole data set is not the same as the SS for the covariate from the ANCOVA because the latter is the variation explained by the pooled within-groups regression.

12.1.3 Null hypotheses

The H_0 for a single factor ANCOVA with a single covariate is based on adjusted means and adjusted

treatment effects, i.e. means and effects of A adjusted for the covariate:

H_0: $\mu_{1(adj)} = \mu_{2(adj)} = \ldots = \mu_{i(adj)} = \ldots = \mu_{(adj)}$
H_0: $\alpha_{1(adj)} = \alpha_{2(adj)} = \ldots = \alpha_{i(adj)} = \ldots = 0$

The adjusted means are simply group (treatment) means of the adjusted observations. They are also the mean values of Y in each group when the covariate is adjusted to equal $\bar{x}$, using the estimate of pooled within-groups regression slope of Y on X (β). Because of the assumption that the slopes of the individual within-group regression lines are the same (see Section 12.3), the differences between adjusted means are the same as the differences between adjusted Y-values for any value of X. When $X = 0$, we are dealing with Y intercepts for regression models with the common pooled within-groups regression slope fitted to the population of observations in each group. Any test of equality of adjusted population group means is also a test of equality of population group intercepts.

The expected values of the mean squares for the ANCOVA in Table 12.1 indicate that the test of the H_0 of no difference between adjusted group means uses an F-ratio of $MS_{A(adjusted)}$ to $MS_{Residual(adjusted)}$. This F-ratio is compared to an F distribution with $(p-1)$ and $p(n-1)-1$ df in the usual manner.

12.1.4 Comparing ANCOVA models

We can also test the H_0 of no effects of factor A using full and reduced models. The full model 12.1 is:

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij} \quad (12.11)$$

The reduced model is a simple linear regression model based on no group effects (H_0: all α_is equal zero):

$$y_{ij} = \mu + \beta(x_{ij} - \bar{x})\varepsilon_{ij} \quad (12.12)$$

Here, β is the regression slope of Y on X for all groups combined. The SS_{Total} from the ANCOVA (i.e. $SS_{Total(adjusted)}$) is simply the $SS_{Residual}$ from the full model and $SS_{Residual(adjusted)}$ is simply the $SS_{Residual}$ from the reduced model, analogous to the model fitting procedure described in previous chapters.

We could also compare the full model 12.1 with a reduced ANOVA model ignoring the covariate:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (12.13)$$

This tests the null hypothesis that pooled within-groups regression slope between Y and X equals zero. If this H_0 is true, then we would expect the covariate not to contribute to explaining the variation in Y and $SS_{Residual}$ from the ANOVA model would be the same as that from the ANCOVA model. Note that model 12.13 is fitted to unadjusted observations so is not the same as model 12.4.

12.2 Assumptions of ANCOVA

The assumptions for ANCOVA include those for regression models (Chapter 5) and ANOVA models (Chapter 8). The error terms from our fitted ANCOVA model should be normally distributed, they should have similar variances between groups and they should be independent. Note that these error terms are the errors from the linear regression of Y on X (model 12.1) and from the ANOVA model fitted to the adjusted observations (model 12.4). We use the residuals in 12.11 to check these assumptions. Because our ANCOVA model has a regression component, these residuals will be different for observations within each group as well as between groups. Plots of residuals against adjusted group means are the best check of the assumption of homogeneous variances. Transformations of Y will often help if the heterogeneous variances are due to skewed distributions of Y within each group and generalized linear models (Chapter 13) are also applicable.

Because the ANCOVA is a linear model with both categorical and continuous predictors, some other assumptions are discussed below. A fundamental assumption underlying the application of the ANCOVA model and calculation of adjusted group means, that the within-group regression slopes relating Y to X are equal, will be examined in Section 12.3.

12.2.1 Linearity

The relationship between Y and X in each group should be linear. As always, scatterplots are a good way of checking this assumption and transformations should be used where appropriate. Specific forms of nonlinearity between Y and the covariate may be dealt with by including a polynomial term as an extra covariate (Maxwell *et al.* 1993; see also Section 12.7.1). These analyses will not be straightforward because there will probably be collinearity between the covariate and its polynomial term (see Chapter 6). Also, the test of homogeneity of within-groups regression slopes is more complex because there are at least two slopes for each group, one for X and one for each polynomial term.

12.2.2 Covariate values similar across groups

ANCOVA also assumes that the covariate has the same distribution, especially the range of covariate values, for all groups. This assumption is basically one of no collinearity between the continuous and categorical predictors in the model. We are assuming that the covariate is independent of the treatment groups, i.e. the covariate values do not depend on the groups. This assumption means in practice that you should avoid situations in which there is a range of covariate values that is present in one group, but absent from others. The problem is that the adjustment procedure would involve extrapolation of the regression between Y and X beyond the range of X values in some groups. Note that a correlation between the covariate and the factor is not the same as an interaction between the covariate and the factor on the response variable. The latter is about homogeneity of within-group regression slopes and will be considered in Section 12.3.

There is no hard and fast rule about what constitutes too little overlap of covariate values between groups, but if the covariate means are not significantly different between groups (from a single factor ANOVA on the covariate), then the ANCOVA is probably reliable. If you have problems with this assumption, the only solution is to omit observations within groups that have unusually high or low covariate values.

The other important implication of this assumption is that ANCOVA models should not be used as a correction for different values of the covariate in each group of an experiment. For example, if the initial body sizes of animals are different between treatments at the start of a growth experiment, then using initial size as a covariate to "adjust" for this difference is inappropriate.

12.2.3 Fixed covariate (X)

The covariate X is assumed to be a fixed variable with no error associated with it. This is the standard fixed X assumption of linear regression (Chapter 5). This assumption is almost never valid for ANCOVA in biological settings because the covariate is usually a random variable, just like the response variable. As we pointed out in Chapter 5, X being a random variable in regression analysis usually results in underestimation of the true regression slope. If the assumptions about homogeneity of variance, range of covariate values and parallel slopes hold, there is no reason to suspect that the underestimation of the true pooled within-groups regression coefficient between Y and X will vary between treatments. Therefore, tests of significance should still be reliable. We know of no extension of the Model II regression approach (Chapter 5) to ANCOVA.

12.3 Homogeneous slopes

The comparison of adjusted means relies on the slopes of the regressions of Y on X being the same between groups, i.e. homogeneity (equality) of slopes. The adjustment of the Y-values to produce adjusted group means and effects is based on a pooled within-groups regression coefficient. This pooled slope must be a reasonable representation of the individual slopes, which will only be true if the individual slopes are similar, i.e. the individual regression lines are parallel.

12.3.1 Testing for homogeneous within-group regression slopes

The H_0 of equal within group regression slopes ($\beta_1 = \beta_2 = \beta_i = \beta$) is tested by examining whether the interaction between the categorical predictor (factor A) and the continuous predictor (covariate) equals zero, i.e. no interaction. We have already

examined interactions between continuous predictors (Chapter 6) and between categorical predictors (Chapters 9 to 11). An interaction between a categorical and a continuous predictor is interpreted as a change in the slope of the regression line of Y on X for different levels of the factor (i.e. different groups). No interaction indicates that the regression coefficients (slopes) are the same in the different groups. This assumption is tested by comparing the fit of a full model with a factor by covariate interaction term to a reduced model with no interaction term. The formal model terminology for interactions between covariate and factors is tedious so we will illustrate the models for the data from Constable (1993). The full model is:

$$(\text{suture width})_{ij} = \text{overall mean} + (\text{food level})_i + (\text{body volume})_{ij} + (\text{food level} \times \text{body volume})_{ij} + \varepsilon_{ij} \quad (12.14)$$

The reduced model, assuming homogeneous regression slopes between groups, is:

$$(\text{suture width})_{ij} = \text{overall mean} + (\text{food level})_i + (\text{body volume})_{ij} + \varepsilon_{ij} \quad (12.15)$$

In practical terms, heterogeneous slopes cause problems for interpreting our data. If the regression lines in the different groups are not parallel, and you are trying to decide if their adjusted means or intercepts differ, your answer depends on where along the range of X-values you do the comparison. For some X-values, the adjusted means or intercepts will be closer together than for others.

Maxwell *et al.* (1993) suggested that homogeneity of slopes should not be thought of as merely an assumption. While main effects are difficult to interpret in the presence of interactions, interactions between factors and covariates usually represent effects of considerable biological interest. Differences between the slopes of the regression lines indicate that the treatments affect the relationship between Y and the covariate and explaining this might be as important as interpreting differences between adjusted means.

There is one important problem that often occurs, especially with data like morphometrics. If your factor has many levels, or you have large numbers of observations in each group, or the linear regression model of Y on X fits the data in each group very well (i.e. r^2 is very high), you may have a very sensitive test of the H_0 of no interaction. You may find yourself rejecting H_0, even though a scatterplot suggests the regression lines are almost parallel. This is always a difficulty when using formal significance tests for checking assumptions before a linear model analysis. We suggest plotting the lines to see how different they look and to examine the individual regression slopes. If they seem parallel, consider doing the ANCOVA anyway or else simply use the Wilcox procedure described below for heterogeneous slopes.

12.3.2 Dealing with heterogeneous within-group regression slopes

When slopes are clearly heterogeneous, there are a number of possible approaches, which depend on the questions of interest. First, if the slopes themselves are of primary interest, you can contrast slopes across treatment combinations. This is like using treatment–contrast interactions to examine the Y by covariate interaction (Chapter 9). Second, if the treatment (group) effects are the main interest, Huitema (1980) recommended a test called the Wilcox procedure (Wilcox 1987b), which is a modification of the original Johnson–Neyman technique (Box 12.4; see also Maxwell *et al.* 1993). This test compares groups in a pairwise fashion, and identifies ranges of the covariate for which the group means are significantly different, and ranges for which there are no differences. It is analogous to a test for simple main effects in a factorial ANOVA (Maxwell *et al.* 1993; Chapter 9), asking for what values of the covariate are the treatments significantly different. It is essentially an unplanned comparison technique, which, with the Wilcox modification, adjusts probability levels to take account of the number of tests. We recommend comparing only a few pairs of treatments, treating them as essentially planned contrasts where possible. Constable (1993) described the application of the Wilcox procedure to compare treatments in sea urchins (see Box 12.2).

A related approach is to choose certain values of covariate and compare groups at those specific values, e.g. using the mean of X or the value of X for which the distance between regression lines has the most precision (Maxwell *et al.* 1993, Rogosa 1980).

Box 12.4 Computations for the Wilcox modification of the Johnson–Neyman procedure for testing over which ranges of the covariate are the group means different

Significantly different slopes in an analysis of covariance indicates that the relationship between the response variable and the covariate differs between treatments. The differences between the regressions may be examined by plotting the 95% confidence bands around each line, and seeing whether these bands overlap. However, the interpretation of these differences is difficult, because the relative effects of the treatments become obscure when the distributions of values about each line begin to overlap. Generally, it is interesting to know the range of the covariate over which the treatments differ. This is analogous to tests of simple main effects (i.e. means) in a multi-factorial analysis of variance when there is a significant interaction, e.g. identifiying the levels of factor B for which there is an effect of treatment A (see Huitema 1980 for a discussion).

One procedure for making such a comparison determines the lower and upper limits of the covariate (X_{lower} and X_{upper}) between which we are 95% certain that two treatments under consideration are *not* significantly different, i.e. region over which the lines cross. Johnson & Neyman (1936; J–N) originally designed a procedure for comparing two treatments at single values of the covariate. Huitema (1980) and Wilcox (1987b) have suggested ways of controlling experiment-wise Type I error rates for simultaneous comparisons of two treatments at more than one region of the covariate (i.e. defining regions of significant differences), as well as for controlling error rates for simultaneous comparisons of more than two treatments. Huitema's (1980) method simply exchanges the F-ratio in the original J–N formulae with a modified Bonferroni F-ratio, which acccounts for the total number of comparisons to be made between treatments (see Huitema 1980, pp. 292–293). Wilcox (1987b) developed formulae similar to the J–N technique, but based on the Tukey–Kramer simultaneous multiple comparisons procedure and Studentized range distribution, rather than on the F distribution. In these formulae, he accounts for unequal variances, allows simultaneous determination of the lower and upper limits of the regions of non-significance between all pairs of treatments and the subsequent generalizations, as well as controlling the potential effects of differences between the range of the covariate in the treatments. Wilcox (1987b) also developed a statistic, 'h' (table included in his paper) to help control the error rates due to repeated comparisons of both intercepts and slopes of all treatment regressions. For comparisons of treatments with very large sample sizes or large differences in sample sizes, h should be substituted by $\sqrt{}(2SMM)$, where SMM, the Studentized Maximum Modulus, is read from the table in Rohlf & Sokal (1969).

The Wilcox (J–N) procedure is computationally tedious but there is a computer program (WILCOX.EXE), written by Andrew Constable from the Antarctic Division (Australia), to do the analysis and it is available from our website. It requires some of the standard ANCOVA output from statistical software.

The procedure adopted here comprises the comparisons for unequal variances of Wilcox (1987b, p. 91), which we have called the "Wilcox comparisons". To compare two groups, j and k:

$$X_{upper} = -B - \sqrt{}(B^2 - 4AC)/\,2A$$

$$X_{lower} = -B + \sqrt{}(B^2 - 4AC)\,/\,2A$$

where

$$A = (b_{1j} - b_{1k})^2 - \left(\frac{h^2}{2}\right)\left(\frac{MS_{Residual(j)}}{SS_{\bar{x}_j}} + \frac{MS_{Residual(k)}}{SS_{\bar{x}_k}}\right)$$

$$B = 2(b_{1j} - b_{1k})(b_{0j} - b_{0k}) + h^2\left(MS_{Residual(j)}\frac{\bar{x}_j}{SS_{\bar{x}_j}} - MS_{Residual(k)}\frac{\bar{x}_j}{SS_{\bar{x}_k}}\right)$$

$$C = (b_{0j} - b_{0k})^2 - E - \left(\frac{h^2}{2}\right)\left(MS_{Residual(j)}\frac{\bar{x}_j^2}{SS_{\bar{x}_j}} + MS_{Residual(k)}\frac{\bar{x}_k^2}{SS_{\bar{x}_k}}\right)$$

and

$$E = \left(\frac{h^2}{2}\right)\left(\frac{MS_{Residual(j)}}{n_j} + \frac{MS_{Residual(k)}}{n_k}\right)$$

with

b_{0j}, b_{0k} the intercepts of the regressions for groups j and k
b_{1j}, b_{1k} the slopes of the regressions for groups j and k
$MS_{Residual(j)}, MS_{Residual(k)}$ the residual mean squares from the regression within each group
$\bar{x}_j, \bar{x}_k$ the mean values of the covariate in each group
$SS_{\bar{x}j}, SS_{\bar{x}k}$ the sums of squares for the means of each covariate in each group
n_j, n_k sample sizes in each group
$h, h_{\alpha,j,df}$ read from Table 1 in Wilcox (1987b)
and
α significance level (usually 0.05)
J number of groups for factor A
df degrees of freedom for the comparison between j and k based on the Satterthwaite and Welch adjustment (Chapter 3).

12.3.3 Comparing regression lines

We mentioned at the start of this chapter that ANCOVA can also be used as a way of comparing regression lines between groups. The comparison of regression slopes across groups, a test for parallelism, uses the methods described in the previous section, testing the factor group by covariate interaction term. If the within-group regression slopes are different, then there is usually no interest in comparing intercepts because the difference between intercepts is not maintained for other values of X. If the regression lines are found to be not significantly different from parallel, then a pooled within-group regression slope is used to "force" the lines to be parallel and the differences between intercepts represent differences for any value of X. So the test comparing intercepts is simply the test comparing adjusted means (Section 12.1.3) once a pooled within-groups regression slope has been fitted.

12.4 Robust ANCOVA

There has been a surprising amount of theoretical work on robust alternatives to ANCOVA, just about all based on ranks (see review in Maxwell *et al.* 1993). Such robust methods may be required if there is non-normality in the response variable (Y) or nonlinearity in the relationship between Y and the covariate. Puri & Sen's (1969) test, one of the first, ranks Y and X separately then calculates a special test statistic. Alternatively, a simple rank transform (RT) approach could be used whereby the usual ANCOVA is done on rank transformed data (both Y and X). Olejnik & Algina (1987) indicated that the different rank transform tests in ANCOVA generally perform similarly but their results showed that only when the parametric assumptions were seriously compromised did the parametric ANCOVA do badly. The rank transform approaches are probably most useful when inexplicable outliers are present or when the relationship between Y and X is nonlinear, effectively requiring a non-parametric regression. Given the concerns expressed in Chapter 9 about the ability of rank transform tests to detect interactions in ANOVA designs, their ability to pick up heterogeneity of slopes in ANCOVA designs must also be in doubt.

Randomization tests could also be used if we consider the ANCOVA model as a multiple regression and do multiple randomizations of experimental or sampling units to groups (as in a single factor ANOVA design – see Chapter 8), keeping the pairing between Y and the covariate (Manly 1997).

12.5 Unequal sample sizes (unbalanced designs)

There are no specific difficulties associated with ANCOVAs with unequal sample sizes between groups beyond what we have already discussed in Chapter 8 for single factor ANOVAs. We have to be more careful about checking assumptions with unequal sample sizes and if our design has two or more factors as well as a covariate, we recommend using Type III SS (see Chapter 9).

12.6 Specific comparisons of adjusted means

12.6.1 Planned contrasts

Contrasts among adjusted means can be done with a t test:

$$t = \frac{c_1\bar{y}_{1(adjusted)} + c_2\bar{y}_{2(adjusted)} + \cdots}{\sqrt{MS_{Residual}\left[c_1^2/n_1 + c_2^2/n_2 + \cdots + \frac{(c_1\bar{x}_1 + c_2\bar{x}_2 + \cdots)^2}{SS_{Residual(X)}}\right]}} \quad (12.16)$$

This daunting equation is simply the usual t test for a contrast in a standard ANOVA except it takes into account the covariate means and the covariate residual variation. The c_is are the contrast coefficients, $MS_{Residual}$ is from the ANCOVA and $SS_{Residual(X)}$ is from an ANOVA on the covariate. There will be an equivalent F test ($F = t^2$) that can be partitioned from the $SS_{A(adjusted)}$. Note that most statistical software will provide adjusted means as output from fitting an ANCOVA model and also allow contrasts on adjusted means as part of the general linear models routines.

12.6.2 Unplanned comparisons

To do unplanned multiple comparisons of adjusted means, use either the Bryant–Paulson-Tukey (B–P–T) test or the Conditional Tukey–Kramer test (Day & Quinn 1989, p. 461). The latter test is simpler, because it uses the usual q distribution. The B–P–T test uses special tables (Kirk 1995). Both can be used as stepwise Ryan's tests. As a general rule, however, we recommend that you avoid unplanned multiple comparisons and try and plan a small number of sensible contrasts wherever possible. Most statistical software won't do either multiple comparison test, so an alternative is to do all pairwise contrasts based on the t tests in Equation 12.16 with a Bonferroni-style adjustment of significance levels to correct for multiple testing (Chapter 3).

12.7 More complex designs

Single factor ANCOVA models are relatively straightforward, but things get more complicated with multiple factors and/or multiple covariates. The adjustment procedure is just an extension of the single factor design, but the test of homogeneity of slopes is much trickier. Each broad type of design will be considered separately in this section.

12.7.1 Designs with two or more covariates

In designs with multiple covariates, the regression component of the ANCOVA becomes a multiple regression. The adjustment for a design with one factor and two covariates (X and Z) is:

$$\bar{y}_{i(adjusted)} = y_{ij} - b_{YX}(x_{ij} - \bar{x}) - b_{YZ}(z_{ij} - \bar{z}) \quad (12.17)$$

Table 12.3 Factorial ANCOVA with factor A (p levels), factor B (q levels) and covariate

Source	Morse & Bazzaz (1994)	df	Morse & Bazzaz (1994)
A	Temperature	$(p-1)$	2
B	CO_2	$(q-1)$	1
A × B	Temperature × CO_2	$(p-1)(q-1)$	2
Covariate	*Biomass*	*1*	*1*
Residual	Residual	$pq(n-1)-1$	231

Note:
Example is from Morse & Bazzaz (1994), who had unequal numbers of plants within each cell. The covariate term does not contribute to the $SS_{Total(adjusted)}$.

where b_{YX} is the estimate of the pooled within-groups regression slope relating Y to X (β_{YX}) and b_{YZ} is the estimate of the pooled within-groups regression slope relating Y to Z (β_{YZ}). The analysis is then done on these adjusted values in a similar manner to when there is a single covariate. We must be very careful about collinearity problems, particularly correlations between the two covariates. Two highly correlated covariates provide redundant information so won't help in reducing the residual variation much anyway. Additionally, if either of the covariates are different between the groups, the adjustment requires extrapolation of either the Y on X or Y on Z regression lines.

For most statistical software, we simply include the multiple covariates when we fit our linear ANCOVA model. Checking homogeneity of within-group regression slopes is more difficult. Essentially the assumption is about parallelism of a series of planes (or higher-dimensional spaces!), rather than simple lines. We suggest that you check the homogeneity of slopes for each covariate separately, by testing the interactions between the factor and each covariate.

12.7.2 Factorial designs

Factorial designs that include one or more covariates measured on each experimental or sampling unit are common. For example, Morse & Bazzaz (1994) did an experiment to test the effects of three temperature regimes and two levels of CO_2 on the number of nodes (an estimate of developmental age) of individuals of two species of annual plants (*Abutilon theophrasti*, a C_3 plant, and *Amaranthus retroflexus*, a C_4 plant). Each species was analyzed separately with a factorial linear model with replicate plants in each cell. Because the number of nodes might be affected by size independently of age, the aboveground biomass (i.e. size) was also used as a covariate for these analyses.

The ANCOVA model for this design is based on adjusting the Y-values using a within-cells regression slope pooled across all the combinations of factors A and B (the pq cells):

$$y_{ijk(adj)} = y_{ijk} - b(x_{ijk} - \bar{x}) \qquad (12.18)$$

This adjustment is based on the estimate (b) of the pooled within-cells regression slope (β). The analysis then uses these adjusted values in a two factor crossed ANOVA (Table 12.3). For most software, the model fitted is the usual two factor crossed ANOVA model including a covariate term.

Maxwell *et al.* (1993) point out that the effects of the two factors in crossed ANCOVAs are not orthogonal, i.e. we have the same difficulty partitioning the $SS_{Total(adj)}$ as we do trying to partition the SS_{Total} in a crossed ANOVA design with unequal sample sizes (Chapter 9). Our recommendation for Type III SS in unbalanced factorial models also applies to factorial ANCOVAs, even when the sample sizes are equal. When random factors are included in these models, the denominators of the F tests for the fixed factors will change, as described in Chapter 9.

Since the adjustment in Equation 12.18 is based on the pooled within-cells regression slope, the test for homogeneity of slopes in these factorial designs should compare the regression slopes across all pq cells. For a two factor (A and B) with one covariate (X) design, the following model would be fitted:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + X + \alpha_i X + \beta_j X + (\alpha\beta)_{ij} X + \varepsilon_{ijk} \qquad (12.19)$$

Note that β_j here refers to the effect of factor B, not the regression slope for the covariate. The regression slopes are implied by the covariate term X and its interactions in the model. Model 12.19 will result in three heterogeneity of slopes terms: A × X, B × X and A × B × X. Huitema (1980) recommended that these terms be combined and tested against the $MS_{Residual}$ from this model. This tests for any variation between slopes across all cells. This is the same test that we would get if we fitted a model that considered the factor combinations as levels of a single factor (a cell means model) and tested the factor by covariate interaction term.

Tests of main effects in factorial designs pool across the levels of the other factor(s), so it might be more appropriate to do separate tests for homogeneity of slopes for each effect based on adjusted means. So we would test homogeneity of slopes for the A × B interaction (test A × B × X against the Residual), test for homogeneity of slopes for the A main effect (test A × X against Residual) and again for the B main effect (test B × X against Residual). We have not seen this approach discussed in the literature, although we suggest a version of it for nested designs in Section 12.7.3.

If the H_0 of equal slopes is rejected, you can then test simple main effects with separate ANCOVAs or examine the interaction between the factors and the covariates in more detail. Either way, the Wilcox (J–N) procedure will again play an important role. As before, we recommend doing only a small number of possible comparisons, as planned contrasts. If the homogeneity of slopes test is not significant, then those interaction terms involving the covariate can be omitted and the model refitted – this then is a standard ANCOVA model.

Table 12.4 Nested ANCOVA with fixed factor A (p levels), random factor B (q levels) nested within A and covariate

Source	Leonard *et al.* (1999)	df	Leonard *et al.* (1999)	Denominator
A	Predation	$p-1$	1	B(A)
B(A)	Site(Predation)	$p(q-1)$	4	Residual
Covariate	*Length*	*1*	*1*	*Residual*
Residual	Residual	$pq(n-1)-1$	204	

Note:
Example is from Leonard *et al.* (1999) who had unequal numbers of mussels within each site within each predation level. Denominator for F test of H_0 for each term provided. The covariate term does not contribute to the $SS_{Total(adjusted)}$.

12.7.3 Nested designs with one covariate

Nested designs can also include covariates. For example, Leonard *et al.* (1999) examined attachment strength of intertidal mussels at sites with either high levels of crab predation or low levels of crab predation. The prediction was that attachment strength would be greater at sites where predation was important. This was a nested design, with factor A being high vs low predation, there were three sites (factor B) nested within each predation level and attachment strength (Y) was measured on randomly chosen mussels. Because attachment strength might also be related to mussel size (larger mussels have stronger attachments), shell length was recorded for each mussel as a covariate (X).

The ANCOVA model for this design is based on adjusting the Y-values using a pooled within-cells regression slope. This adjustment is the same as used for a factorial design, based on the estimate (b) of the pooled within-cells regression slope (β) – see Equation 12.18 in previous section.

The nested ANCOVA then uses these adjusted values (Table 12.4). The model fitted is the usual nested ANOVA model including the covariate term.

Testing for homogeneity of regression slopes tests can be done in two ways. First, we can test for any differences in slopes of the regression models for Y on X across all cells. This test is done by fitting the following model:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + X + \alpha_i X + \beta_{j(i)} X + \varepsilon_{ijk} \quad (12.20)$$

This model includes the $A \times X$ and the $B(A) \times X$ interactions and we would combine these into a single test of homogeneity of within-cells regression slopes.

The second approach acknowledges that factor B is usually random in these designs and A is then tested against B(A). This suggests that we might do a separate test of homogeneity of slopes among the levels of A, using the $B(A) \times X$ interaction terms as the error ($A \times X$ against $B(A) \times X$). The question that we are now asking is, "Is there significant variation in slopes between levels of A, relative to variation in slopes between the levels of B within levels of A?" This seems to be the approach taken by Leonard *et al.* (1999), who tested the Predation × Length interaction against either the Site(Predation) or the Site(Predation) × Length term. We suggest the latter denominator is more appropriate, especially when the question focuses on adjusted A level means. There will be some cases where you explicitly want to compare slopes across all cells in a nested design, and then the pooled test of homogeneity of regression slopes is applicable.

12.7.4 Partly nested models with one covariate

In split-plot and groups by trials repeated measures designs (Chapter 11), there are two ways a covariate can be included. First, separate measures for the covariate are taken for each sub-plot within each plot or for each subject at each time or within-subjects group. In the example from Mullens (1993) described in Chapter 11, blood pressure might be measured as a covariate for each toad each time breathing rate is recorded. Second, a single covariate measure is associated with each plot or with each subject, irrespective of sub-plot or level of within subjects factor. Again from Mullens (1993), body size or basal breathing rate might be used as a covariate and there would be only a single value for each toad, as this would not vary with O_2 level. Krupnick & Weis (1999) used this second type of partly nested ANCOVA to analyze their experiments on the effect of florivory on plant success. They had a repeated measures design with individual plants of the perennial shrub *Isomeris arborea* as the subjects. The between-subjects factor was three insecticide treatments (protected from herbivory by insecticide spraying, exposed to herbivory but sprayed with water control, exposed to herbivory without spray). The within-subjects factor was date as each plant was recorded on numerous occassions in each of three years – separate analyses were done for each year (Table 12.5). The response variable was fruit production but because this might also be affected by plant size, the number of branches on each plant was recorded as a covariate. This covariate did not vary for each plant during the experiment.

In the more general first scenario, there are regressions of Y on X at two levels – between plots or subjects, and within plots or subjects. The second situation is just a special case of the first where the covariate measure is the same for every observation on each subject or plot. In this situation, there is only a between-subject or plot regression, so only the between-subjects means are adjusted in practice. In the first case, the adjustment is done for between-subjects effects (A) and within-subjects effects (C and $A \times C$).

The ANCOVA adjustment for the first case is (Kirk 1995):

$$y_{ijk(adj)} = y_{ijk} - b_{Between}(\bar{x}_i - \bar{x}) - b_{Within}(x_{ijk} - \bar{x}_i) \quad (12.21)$$

where $b_{Between}$ is the estimate of the pooled within A groups (i.e. between plots/subjects) regression slope ($\beta_{Between}$) and b_{Within} is the estimate of the pooled within C groups (i.e. within plots/subjects) regression slope (β_{Within}). When the covariate has a single value for each plot or subject, then the second component of the adjustment in 12.21 simply becomes zero. The analysis then uses these adjusted values in a partly nested ANOVA (Table 12.5). The model fitted is the usual partly nested model including the covariate term.

Testing homogeneity of slopes in these designs is tricky and rarely discussed in textbooks. Even a recent review of ANCOVAs for split-plot designs (Federer & Meredith 1992) did not describe testing

Table 12.5 | Partly nested ANCOVA with factor A (p levels) and factor B (plots/subjects with q levels) nested within A, factor C (r levels) as within-plots/subjects factor and a covariate measured on each plot/subject

Source	Krupnick & Weis (1999)	df	Krupnick & Weis (1999)
Between plots/subjects			
A	Treatment	$(p-1)$	2
Covariate (X)	No. branches	1	1
B(A)	Plants within treatment	$p(q-1)-1$	26
Within plots/subjects			
C	Date	$(r-1)$	25
$A \times C$	Treatment × date	$(p-1)(r-1)$	50
$C \times X$	Date × no. branches	$(r-1)$	25
$B(A) \times C$	Plants within treatment × date	$p(q-1)(r-1)-1$	650

Note:
Example is for 1992 fruit production in *Isomeris arborea* from Krupnick & Weis (1999) – their Table 2. Factor A was insecticide treatment, plots/subjects were plants, factor C was date and the covariate was number of branches on each plant.

for homogeneity of slopes, although their paper emphasized estimation, not hypothesis testing. For the general case with separate covariate measures for each sub-plot or each subject at each level of the within-subjects factor, a model that includes all factor by covariate interactions is fitted:

$$y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + (\alpha\gamma)_{ik} + \beta\gamma_{j(i)k} + X + \alpha_i X + \beta_{j(i)} X + \gamma_k X + (\alpha\gamma)_{ik} X + \varepsilon_{ijkl} \quad (12.22)$$

Note that l usually equals one in these designs so each observation is actually y_{ijk}. In such a design, we cannot separately estimate $\beta\gamma_{j(i)k}$ and ε_{ijkl} nor can we separately estimate the covariate by factor interaction terms $\beta_{j(i)}$ by X and $(\alpha\gamma)_{ik}$ by X. We suggest testing homogeneity of slopes for A, C and $A \times C$ separately using the appropriate error terms, i.e. $A \times X$ against B(A), $C \times X$ and $A \times C \times X$ against $B(A) \times C$. For the case where we have only single covariate measure for each plot or subject, homogeneity of slopes is only relevant across levels of A, so only the interaction of $A \times X$ would be included in the model and tested against the B(A) term. This approach was used by Krupnick & Weis (1999) who tested for an interaction between insecticide treatment (the between-subjects factor) and number of branches (covariate).

Note that if the covariate measures are different for each sub-plot or level of the within-subjects factor, your data file will need to be coded for a classical split-plot analysis (Chapter 11), even if you have a repeated measures design. Note also that the number of terms in these models, including all the interactions with the covariate, can get large and this can cause computational problems when the number of observations (especially plots/subjects within A) is relatively small.

12.8 General issues and hints for analysis

12.8.1 General issues

- Including one or more covariates can reduce the unexplained variation in ANOVA designs and increase precision of estimates of group means and power of tests.
- The basic ANCOVA tests null hypotheses about adjusted means and factor effects, where the linear relationship between the covariate and the response variable (Y) is taken into account. These means are adjusted to the overall mean value for the covariate by the relationship between Y and the covariate.
- Since a pooled within-groups regression slope is used for the adjustment, the assumption of homogeneous slopes across groups is very important for interpreting ANCOVA models

and should always be checked. The Johnson–Neyman (J–N) procedure is applicable for simple designs if this assumption is not met.
- Contrasts and unplanned multiple comparisons between adjusted means require different methods than for unadjusted means, taking into account the linear relationship with the covariate.
- Covariates can be included in more complex ANOVA models (nested, factorial, and partly nested), the major difficulty being deriving tests for homogeneity of slopes.

12.8.2 Hints for analysis

- Most common statistical software offers ANCOVA as a menu option, but in most of them, you will be fitting an ANCOVA model that assumes homogeneity of slopes. To fit a model testing for heterogeneous slopes, you will generally need to specify the model fully through the general linear models option.
- Homogeneity of within-group regression slopes is tested by including factor by covariate interaction terms in a preliminary model. In complex models, homogeneity of slopes can be checked by combining all factor by covariate terms into a single interaction term that is tested or by treating the design as a single factor means model and testing the single factor by covariate term. Alternatively, homogeneity of slopes may be better tested separately for each component of the analysis, e.g. homogeneity of slopes for main effects separately.
- If slopes are heterogeneous, the comparison of adjusted means using the Johnson–Neyman (J–N) procedure is not available as part of most statistical software, and must be computed manually (Box 12.4) or with the program WILCOX.
- Assumptions such as normality, homogeneity of variances and linearity are best examined with graphical techniques such as residual plots and scatterplots.

Chapter 13

Generalized linear models and logistic regression

So far, most of the analyses we have described have been based around linear models that assume normally distributed populations of the response variable and of the error terms from the fitted models. Most linear models are robust to this assumption, although the extent of this robustness is hard to gauge, and transformations can be used to overcome problems with non-normal error terms. There are situations where transformations are not effective in making errors normal (e.g. when the response variable is categorical) and, in any case, it might be better to model the actual data rather than data that are transformed to meet assumptions. What we need is a technique for modeling that allows other types of distributions besides normal. Such a technique was introduced by Nelder & Wedderburn (1972) and further developed by McCullough & Nelder (1989) and is called generalized linear modeling (GLM). In this chapter, we will examine two common applications of GLMs: logistic regression, used when the response variable is binary, and Poisson regression, when the response variable represents counts. In the next chapter, we will describe log-linear models when both response and predictor variables are categorical and usually arranged in the form of a contingency table.

13.1 Generalized linear models

Generalized linear models (GLMs) have a number of characteristics that make them more generally applicable than the general linear models we have considered so far. One of the most important is that least squares estimation no longer applies and maximum likelihood methods must be used (Chapter 2).

A GLM consists of three components. First is the random component, which is the response variable and its probability distribution (Chapter 1). The probability distribution must be from the exponential family of distributions, which includes normal, binomial, Poisson, gamma and negative binomial. If Y is a continuous variable, its probability distribution might be normal; if Y is binary (e.g. alive or dead), the probability distribution might be binomial; if Y represents counts, then the probability distribution might be Poisson. Probability distributions from the exponential family can be defined by the natural parameter, a function of the mean, and the dispersion parameter, a function of the variance that is required to produce standard errors for estimates of the mean (Hilbe 1993). For distributions like binomial and Poisson, the variance is related to the mean and the dispersion parameter is set to one. For distributions like normal and gamma, the dispersion parameter is estimated separately from the mean and is sometimes called a nuisance parameter.

Second is the systematic component, which represents the predictors (X variables) in the model. These predictors might be continuous and/or categorical and interactions between predictors, and polynomial functions of predictors, can also be included.

Third is the link function, which links the random and the systematic component. It