Analysis of Variance with Unbalanced Data: An Update for Ecology & Evolution

Andy Hector, Stefanie von Felten and Bernhard Schmid

Institute of Environmental Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Corresponding author:

Andy Hector

Email: ahector@uwinst.uzh.ch

Phone: +41 (0)44 635 48 04

Fax: +41 (0)44 635 57 11

Abstract

Factorial analysis of variance (ANOVA) with unbalanced (non-orthogonal) data is a commonplace but controversial and poorly understood topic in applied statistics. We explain that ANOVA calculates the sum of squares for each term in the model formula sequentially (type I sums of squares) and show how ANOVA tables of *adjusted* sums of squares are composite tables assembled from multiple sequential analyses. A different ANOVA is performed for each explanatory variable or interaction so that each term is placed last in the model formula in turn and adjusted for the others. The sum of squares for each term in the analysis can be calculated after adjusting only for the main effects of other explanatory variables (type II sums of squares) or, controversially, for both main effects and interactions (type III sums of squares). We summarize the main recent developments and emphasize the shift away from the search for the 'right' ANOVA table in favour of presenting one or more models that best suit the objectives of the analysis.

Keywords: Adjusted sums of squares; Type III sums of squares; ANOVA; Orthogonality; Linear models; R.

Introduction

Analysis of variance (ANOVA) continues to be one of the most widely used forms of statistical analysis in many areas of science (Gelman, 2005, Gelman and Hill, 2007). Nevertheless, factorial ANOVA with unbalanced (non-orthogonal – supplement S1) data is a controversial topic in applied statistics and one of the areas of ANOVA that is most poorly understood in ecology, evolution and environmental science. This is partly because bio-statistics textbooks appear to avoid the topic, perhaps because it is controversial. The last coverage of the topic in the ecology and evolution journals revealed disagreement on how to best approach ANOVA of unbalanced data (Shaw and Mitchell-Olds, 1993, Stewart-Oaten, 1995). There still appears to be no consensus within the statistical community, but there has been further discussion that has yet to make its way into the ecology and evolution literature. There has also been a move away from finding the 'right' ANOVA table towards presenting the one or more models that best match the objectives of the analysis.

In this paper, we give non-technical explanations of the issues involved in ANOVA of unbalanced data, particularly the different types of adjusted sums of squares. We also provide (as supplementary material) code for the analysis of worked examples of unbalanced ANOVA designs using the open-source R language for statistical computing and graphics that is fast becoming the *lingua franca* for analysis in ecology and evolution (R Development Core Team, 2009).

The Problem

With balanced designs one factor can be held constant while the other is varied independently. However, this desirable property of orthogonality is usually lost for unbalanced designs (supplement S1). When explanatory variables are correlated with each other due to imbalance in the number of replicates for different treatment combinations the values of the sums of squares depend on the position of the factors in the ANOVA model formula. Because ANOVA and regression are special cases of general linear models there is much overlap between this topic and multiple regression. In non-orthogonal designs, some of the explanatory variables (and, if present, their interactions) are positively or negatively correlated with each other; that is they are partially collinear or confounded. Using a Venn diagram (Figure 1), positive correlations can be illustrated as causing over-lapping and negative correlations under-lapping sums of squares. The desire to find a technological fix that provides a single outcome to the analysis of orthogonal and non-orthogonal data is clear. In response, some statistical software companies have developed several types of adjusted sums of squares.

Sequential and Adjusted Sums of Squares

The sums of squares used in ANOVA as originally proposed by Fisher (1925) are calculated sequentially for each main effect and each two-way or higher-order interaction following the sequence of terms at each level in the model formula. One desirable feature of sequential sums of squares is that they are additive; that is the total sum of squares is decomposed into a series of additive parts. The total sum of squares for a sequential ANOVA is the same for all orderings of the explanatory variables in the model formula even though the values for the individual variables change with their position in the sequence.

The alternative to sequential sums of squares is to use one of a variety of adjusted (also known as partial, unique, marginal, conditional or unweighted) sums of squares. These adjusted sums of squares are sometimes linked to early work by Yates (1933, 1934) as discussed by Nelder & Lane (1995) and summarized in the supplement (S2). Adjusted sums of squares can be divided into two categories (Herr, 1986, Macnaughton, 1998). As the name implies, adjusted sums of squares are calculated for a given explanatory variable after adjusting for the other variables in the statistical model formula. The different systems of adjusted sums of squares can then be categorized as to whether they adjust a given variable for the other variables at the same level (e.g. adjusting each main effect for the other main effects) or whether the adjustment also includes interactions at higher levels. Macnaughton (1998) has termed these 'higher-level terms omitted (HTO)' and 'higher-level terms included (HTI)', while Herr (1986) termed them 'each adjusted for other (EAD)' and 'standard parametric (STP)'. Other terminologies exist (supplement S2) but we find Macnaughton's the most transparent.

In the following section we express these two general classes more formally and illustrate them using a simple worked example of a two-way factorial ANOVA (this is the design used in most discussion of this topic in the statistical literature). To build on earlier literature on this topic we use the hypothetical dataset from Shaw & Mitchell-Olds (1993). The dataset (Table 1) comprises height of experimental target organisms as the response variable, the experimental removal (or not) of neighbours as a first explanatory factor and the initial size of the target organisms as a second factor. Both factors have two levels since initial sizes are recorded only as two classes (small or large). The design is therefore a fully-factorial 2^2 design: that is two factors - each with two levels - crossed so that all four possible combinations (or 'cells' in a tabular representation of the design) are present. The design is unbalanced because the different combinations have different numbers of replicates but no cells are empty (a more extreme form of imbalance). Because the proportional number of replicates are not the same across treatments the design is non-orthogonal, that is the two explanatory variables are not independent of each other.

Sequential Sums of Squares

The design can be analysed with a two-way factorial ANOVA that considers the main effects of the neighbour removal treatment, the initial size class, and their interaction. Due to the imbalance, the sums of squares for the main effects of the two variables change with the two alternative sequential model formulas, which can be written using the effects notation as:

$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ink}$ (1)

$y_{ijk} = \mu + \beta_j + \alpha_i + \gamma_{ij} + \varepsilon_{ijk}$ (2)

where y_{ijk} is the response (final height) of the kth organism (k = 1,2,…, n_{ij}), in the ith level of factor α (the neighbour removal treatment), and the jth level of factor β (initial size), γ is the interaction of the two treatments, μ indicates the intercept (here the grand mean; supplement S3), and ε the within-group error. These two models can be written in the widely used statistical model formula notation of Wilkinson & Rogers (1973) as follows:

T + S + T.S (3)

S + T + T.S (4)

where T is neighbour removal treatment, S is initial plant size and T.S the interaction (which could be equivalently written as S.T). The intercept is taken as implicit in this notation. The model with treatment fitted first produces the sequential ANOVA shown in Table 2a and the model with initial size fitted first produces Table 2b.

Note that in the two sequential models the values for the interaction, residual error and total sum of squares are the same, despite differences for the main effects. These differences in the main effect sums of squares arise because treatment and initial size are not orthogonal. When treatment is fitted before size, treatment is not significant and initial size is highly significant. But when the order is reversed and initial size is put first its sum of squares is reduced (although it remains highly significant) and the sum of squares for treatment is increased so that it borders on being significant too (Table 2). The change of the treatment effect from convincingly non-significant to marginal makes clear the dangers of sequential sums of squares: fitting only one of these models could give an incomplete and potentially misleading impression. The complexity of sequential sums of squares is also clear: we have had to fit two models instead of one (for more complex models the numbers of alternatives increases dramatically)? Is one correct and the other wrong? Or, are both correct but one preferred over the other?

Adjusted Sums of Squares With Higher-Level Terms Omitted

The higher-level terms omitted adjusted sum of squares for the interaction can be written in either of the two following ways:

$SS(T.S \mid \mu + T + S)$ (5)

$SS(T.S \mid \mu + S + T)$ (6)

that is, the sum of squares for the interaction conditional on (or adjusted for) all the lower-order terms: the grand mean, the main effects of both neighbour removal treatment and initial size. The order of the main effects does not matter since their combined value is the same and therefore the sums of squares for the interaction is also the same with either formulation. Similarly, the higher-level terms omitted adjusted sum of squares for treatment (T) and for initial size (S) can be written, respectively, as:

$SS(T \mid \mu + S)$ (7)

$SS(S \mid \mu + T)$ (8)

The different models considered above (we require only 5 or 6, not both) can be written in the Wilkinson & Rogers notation, respectively, as:

T + S + T.S (9)

S + T (10)

T + S (11)

Model 9, for example, can be said to fit the effect of T *ignoring* S and then the effect of S *eliminating* T (McCullagh and Nelder, 1989). That is, for every variable in a sequential model formula preceding variables are said to be eliminated and subsequent variables ignored. The ANOVA tables for these three sequential analyses are shown in Table 3a-c. A composite ANOVA table summarising these adjusted sums of squares can be assembled from these three separate sequential models as follows. Equations 6 - 8 each specify adjusted sum of squares for a single term (T.S, T and S respectively). To get these adjusted sums of squares we fit models 9 – 11 (Table 3a - c). In each case we take only the sum of squares for the final term (excluding the residual error, which is the same in all cases) and use these to build the composite ANOVA table of adjusted sums of squares (3d). Note that the residual sum of squares are the same in both cases and that if we add up the adjusted sums of squares in the composite table the value is different from the total of the sums of squares given by the equivalent sequential ANOVA shown in Table 3a. For this example, the total sum of squares of the adjusted analysis is larger than that of the sequential analysis (some double counting has occurred). The opposite also frequently occurs when sums of squares are missing due to the correlation between variables. In the terminology of the SAS software package (SAS Institute Inc. 1985), this composite ANOVA table uses type II sums of squares (supplement S4). That is, SAS type II sums of squares are adjusted sums of squares that omit higher-level terms when making the adjustments.

Adjusted Sums of Squares With Higher-Level Terms Included

For sums of squares that adjust for higher-level terms, the equations given above can be amended by including the interaction:

$SS(T.S \mid \mu + T + S)$ (14)

$SS(T \mid \mu + S + T.S)$ (15)

$SS(S \mid \mu + T + T.S)$ (16)

Because the highest-level term is not affected, model 14 is the same as the earlier model 5. These models can be written in the Wilkinson & Rogers notation, respectively, as:

T + S + T.S (17)

T.S + S + T (18)

T.S + T + S (19)

Note also that model 17 is the same as the earlier model 5. The last two models, where a main effect is adjusted for the other main effect *and the interaction*, may look strange to the users of software that only use sequential sums of squares. In such packages (e.g. GenStat, GLIM and the base distribution of R used here), attempts to fit models like 18 and 19 will not produce adjusted sums of squares and we must mimic the adjustments that are made behind the scenes by other packages (supplement S5). Once models 17 - 19 have be fitted to produce the sequential ANOVAs shown in Table 4a-c, the final term (again excluding the residual error, which is the same in all cases) from each sequential model is taken to form the composite table of adjusted sums of squares (Table 4d). Note that the higher-terms included adjusted sums of squares for the main effects differ from the higher-terms omitted adjusted sums of squares because each main effect is now adjusted for the other *and the interaction*. Adjusting for the interaction changes the pattern of correlations. In the SAS terminology, these higher-terms included sums of squares are type III sums of squares. That is, SAS type III sums of squares are adjusted sums of squares that include higher-level terms when making adjustments.

Having seen how the four alternatives are obtained we next look at their advantages and disadvantages. First, the good news: in all four cases (Tables 2a, 2b, 3d, 4d) the sum of squares for the residual error and for the interaction term are the same. This means that when the result of an analysis is an interaction that is clearly significant (both statistically and biologically) the type of sum of squares used becomes of little relevance because the interaction is the central result and it is unaffected by the type of sum of squares. Once an interaction is significant, the main effects of the variables involved are usually of little interest (unless the sums of squares for the main effects are much greater than the interaction sum of squares). This is because a clear interaction tells you that both variables are important but that the effect of each depends on the other. To look at the main effect of a factor is to look at its effect averaged over the levels of the other factor, something that would normally be misleading when there is an appreciable interactive effect.

The bad news is that the values for the main effects differ for the two alternative sequential analyses and for the two different types of adjusted sums of squares. The sums of squares for the two main effects in the pair of sequential ANOVAs differ because of their non-orthogonality. The sums of squares for the main effects for the two types of adjusted sums of squares differ because in one case they are adjusted for the other main effect only and in the other case they are adjusted for the other main effect and the interaction term. The next section reviews the heated debate over sequential and adjusted sums of squares and the arguments for and against the different types.

The Case for Higher-Level Terms Included Adjusted Sums of Squares

What led so many software packages to adopt higher-terms included adjusted sums of squares as the default option? Part of the reason is probably a hang over from the early days of computing when analyses had to be programmed using punch cards and were usually done in batch mode because interactive analyses that compare multiple sequential models were too laborious (Nelder, 1994, Nelder and Lane, 1995). When computer power was limiting, the desire for software that produced the (single) answer is understandable (see the quote from Herr given in supplement S2). However, the arguments in favour of adjusted sums of squares go beyond this. Based on some of the statistical literature, Shaw & Mitchell-Olds (1993) recommended them because, *"The Type III sum of squares for each main effect is the sum of the squared differences of unweighted marginal means…[that] do not, therefore, depend on the details of the sampling structure in the data at hand…[and] Type III tests of the various factors in the model do not depend on the particular order in the model"*. Quinn & Keough (2002) recommend them for similar reasons because, *"most biologists would probably prefer their hypotheses to be independent of the cell sample sizes"*. In a sense, higher-terms included adjusted sums of squares can be thought of as testing variables in unbalanced datasets as if those datasets were actually balanced and orthogonal (see supplement S6). The recommendations from bio-statistics sources given above are based on similar recommendations in some of the statistical literature (albeit with important caveats). For example, Searle (1995) comments that, *"for all-cells-filled data, when wanting to use hypothesis testing with models that include interactions, the careful use of Type III sums of squares is the best we can do. True, hypothesis testing may not be the best thing to do, and true, also, is the fact that hypotheses…[may]…have interactions secreted within them."* The question then becomes whether or not it makes sense to test hypotheses about main effects in the presence of interactions.

Another potential argument in favour of ANOVA using type III sums of squares is that, for single degree of freedom tests (i.e. continuous variables and factors with two levels), the results of the (adjusted) F tests are consistent with the results of the t tests of the estimates given in the table of coefficients (because parameter estimates are always adjusted for all other terms in the model too). Again, the question is whether it makes sense to test main effects adjusted for interactions.

The case against higher-level terms included adjusted sums of squares

Missing and double-counted sums of squares

One of the main arguments against adjusted sums of squares is that they result in missing or double-counted variation. Recall (see above) that ANOVA tables of adjusted sums of squares do not sum to the total model sum of squares (as sequential sums of squares do). Depending on the nature of the correlations between explanatory variables, the sum of the adjusted sums of squares can be less than the total model sum of squares or more than it: the greater the imbalance the greater the discrepancy. It is easiest to think about the case where the total of the adjusted sums of squares is less than the total sum of squares for the sequential model.

Consider the simplest example with two main effects, A and B, and *no interaction*. If explanatory variables A and B are positively correlated then they can be thought of as 'sharing' sums of squares. In a Venn diagram (Figure 1) the sums of squares for A and B would be partially overlapping circles (for a similar graphical approach see Schmid et al., 2002). In this case, adjusting both main effects (each for the other) results in the shared or overlapping sums of squares not being counted. It is these missing sums of squares which account for the difference between the sum of the adjusted sums of squares and the total sum of squares for the whole model (e.g. the total of the adjusted sums of squares in tables 3 and 4 vs the total of the sequential squares in table 2). The alternative situation is where the correlation leads to 'underlapping' sums of squares. These are much harder to illustrate graphically but the situation is the reverse of what we have just described: instead of the total of the adjusted sums of squares being less than the total model sum of squares it is greater because of the 'double-counted' variation. Our example here omits the interaction purely because it was beyond our abilities to graphically illustrate it, but the basic principles concerning overlapping and underlapping sums of squares extend to examples involving interactions (as demonstrated in supplement S7 using an example from Aitkin (1977)).

Marginality of main effects and interactions

One of the key criticisms of sums of squares that adjust for higher terms is that they do not respect marginality (Nelder, 1977, Nelder and Lane, 1995). In the context of unbalanced ANOVA, marginality refers to the relationship between higher- and lower-order (or level) terms. Respecting the marginality relations of variables in a model formula means taking account of their position in the hierarchy of main effects and interactions. The principle can be simply illustrated using the two-way factorial analysis example. To respect marginality, models including the interaction term should also include both main effects. More generally, when a higher-level interaction is included in a model, all lower-level interactions and main effects should be included too. For our example, this means a model that includes the interaction should also include the main effects of size and removal treatment. The main effects are said to be marginal to the interaction. Furthermore, marginality implies that when interpreting an ANOVA with interactions we should start at the bottom of the table, looking at the highest-order terms first. If an interaction is significant, then the null hypothesis of additive main effects can be rejected, and we know that the effect of one variable depends on the other. The significant interaction already tells us that the main effects are also important, but that they do not have simple independent effects that can be expressed by averaging over the levels of the other factors. Therefore, it normally makes little sense to interpret a main effect in the presence of a significant interaction (supplement S8). Venables (2000) and Venables & Ripley (2002) make essentially the same argument against adjusting for higher-level terms, as do Aitkin (1978, 1995) and colleagues (Aitkin et al., 2009) and Stewart-Oaten (1995), who says in this context that higher-terms included adjusted sums of squares are, *"best for a test of main effects only when it makes little sense to test main effects at all."*

The null hypothesis of no main effect in the presence of an interaction

The null hypothesis tested for the main effects when using higher-terms omitted sums of squares is unlikely to be true (although to be fair this is a criticism of null hypothesis testing generally). McCullagh (2005) reviews the situation as follows: *"Nelder (1977) and Cox (1984) argue that statistical models having a zero average main effect in the presence of interaction are seldom of scientific interest. McCullagh (2000) reaches a similar conclusion…By definition, non-zero interaction implies a non-constant treatment effect, so a zero treatment effect in the presence of non-zero interaction is a logical contradiction."*

For the null hypothesis of no main effect (for either factor) to be true in the presence of a significant interaction the effect of one factor would have to differ depending on the level of the other (the non-additivity that defines an interaction) but in such a way that the differences cancel exactly such that the effect of one factor averaged across the levels of the other factor is zero. Many statisticians (above) see this as extremely unlikely, although Stewart-Oaten (1995) considers some hypothetical situations where this might occur and we provide some further possibilities (supplement S8).

Marginality - special cases

Most statisticians seem to consider respecting marginality to be the sensible thing to do in general, even those who support the use of higher-terms included sums of squares in some situations (Fox, 2002, Quinn and Keough, 2002, Searle, 1995). What are these special situations? An obvious one is when the degree of imbalance is minor and sequential and higher-terms included adjusted sums of squares produce qualitatively similar answers and the adjusted sums of squares avoid the complexity of presenting the alternative (but similar) sequential analyses. Another situation may be in the case of large complex datasets where there is a desire to test main effects despite interactions. Searle (1995) gives an example of a large and complex dataset, *"involving 9 factors having a total of 56 levels, more than 5 million cells and 8577 data points. Assessing interactions from the whole data set was out of the question."* As discussed below, other statisticians do not agree with this approach to complex unbalanced datasets.

There are also some special cases where the usual marginality relations do not apply. Nelder (1994) gives an example of a special case of analysis of covariance (ANCOVA) where it might make sense to remove the intercept even in the presence of an interaction (differences in slopes) on theoretical grounds (supplement S9). Nelder's (1977) criticisms of higher-terms included adjusted sums of squares also prompted other suggestions where it might make sense to look at main effects in the presence of an interaction, including one from Tukey (1977) which is summarized in supplement S10.

Summary of the sequential versus adjusted sums of squares debate

We can summarize the debate over unbalanced ANOVA as follows, based on our reading of the literature and earlier reviews (Herr 1986; Macnaughton 1998). The main motivation for higher-terms included sums of squares appears to have been the desire for a single outcome to unbalanced ANOVA where the values for the sums of squares are not dependent on the order of the variables in the model formula and where hypothesis tests are not affected by differences in sample sizes for the treatment combinations. This desire seems to have led many statistical software packages to use higher-terms included adjusted sums of squares as the default type.

On the other hand, many statisticians are critical of the use of higher-terms included adjusted sums of squares. The arguments against these type III sums of squares centre on a group of criticisms that relate to their disregard for marginality. While there may be special cases where the usual marginality relations do not apply, most statisticians seem to recommend respecting marginality as a good general principle. Statistical software packages remain divided in their approaches, with some using higher-terms adjusted sums of squares as the default type and others providing only sequential sums of squares. Some recent papers have recommended that higher-terms omitted (SAS type II) sums of squares would be a better choice for software that wants to use a type of adjusted sums of squares as the default setting (Macnaughton 1998; Langsrud 2003) while others recommend comparing a nested series of sequential (type I) models in an approach similar to backwards-deletion multiple regression (e.g. Nelder & Lane 1995; Venables & Ripley 2002; Aitkin et al. 2005).

Recent developments

The last decade has seen a continued shift in emphasis away from hypothesis tests and probability values in favour of parameter estimation. In this context, it is worth pointing out that tests performed on the parameter estimates from unbalanced ANOVA (using t-tests or confidence intervals based on the relevant standard errors) will not always match the results of the F tests from the sequential ANOVA. For single degree of freedom tests of variables in balanced datasets the results of F and t tests do match: $F = t^2$ (Venables & Ripley 2002). However, for unbalanced datasets, there will be a mismatch between some of the F and t tests. This is because, as explained above, the sums of squares used to perform the F tests are calculated sequentially whereas the point estimates and standard errors of each variable are assessed after controlling for all others (supplement). This causes a problem in assessing variables in non-orthogonal analyses with positively correlated explanatory variables that are significant when placed first in the sequential model but non-significant when placed later. The results of these analyses are ambiguous because, as we have explained, the parameter estimates and intervals from the different sequential models will be the same and will support the adjusted (non-significant) F tests.

Another important development is the increase in the popularity of multi-model inference. Model selection approaches like backward-deletion multiple regression using P values tend to result in selection of a single model, despite recommendations to consider more than one model when appropriate (McCullagh and Nelder, 1989). Inferences based on a set of models are now becoming more popular due to the wider recognition of the problem of model selection uncertainty and the increasing use of information criteria (Anderson, 2008, Burnham and Anderson, 2002).

The Example Dataset Revisited: Objective-Led Modelling

To illustrate the shift from searching for the 'right' ANOVA table towards presenting one or more models that best match the objectives of the analysis we revisit the two-way factorial ANOVA of the hypothetical data in Shaw & Mitchell-Olds (1993) on the effects of neighbour removal treatment (T), initial size (S) and the interaction (T.S). They presented three alternative analyses summarised in ANOVA tables: the sequential (type I) model T + S + T.S, the higher-terms omitted (SAS type II) and the higher-term included (SAS type III) adjusted sums of squares. They recommend the SAS type III sum of squares analysis as it uses unweighted marginal means rather than taking into account the differing sample sizes per treatment combinations. However, we argue that consideration of the objectives of the analysis leads to a different solution. If the goal of the ANOVA is to test for significant differences between treatments after accounting for differences in initial size, then we propose an analysis of covariance (ANCOVA) type approach where we want to control for differences in initial size before assessing the effects of the neighbour removal treatment (in a typical ANCOVA initial size would be a continuous covariate). This consideration of the objectives suggests, *a priori,* a sequential model with initial size fitted before neighbour removal treatment: S + T + T.S. The null hypothesis tested is of no effect of neighbour removal after controlling for differences in initial target organism size. This model was discussed in the Shaw & Mitchell-Olds paper but not presented in their table 2. In this analysis, adjusting for initial size causes treatment to become marginally significant. This is a simple example, but it illustrates the shift away from the search for the single ‘right’ ANOVA table, to fitting the model (or models) that best match the objectives of the analysis.

Conclusions

Our aim is not to assert that we have solved the debate over the best approach to unbalanced ANOVA. Far from it, there is still much debate amongst statisticians and, as we have shown above, authoritative backing can be marshalled for all of the approaches reviewed here. This ongoing debate amongst statisticians argues for open-mindedness. By this we do not mean that anything goes! Rather we mean that we (as teachers, analysts, reviewers, editors *etc.*) ought to be open to sensible arguments for a given approach. However, this still calls for good arguments in support of a chosen analysis rather than falling back on a 'cook-book' approach using whatever recipe is known or close to hand. We finish by making some recommendations that we hope will be of general use:

1. Consider whether the objectives and design imply one (or a few) sequential models.
2. Perform tests where you can specify the corresponding biological hypotheses.
3. Investigate imbalance: why has it occurred (was it accident or is it a property the biology of the situation: 'biological colinearity'?). What correlations has it caused and what patterns in the sums of squares for the different sequential analyses (cf. Figure 1)?
4. Test the interactions that are of interest first. If an interaction is significant (biologically and statistically) you have your main answer and one which is independent of the choice of sums of squares (sequential and adjusted sums of squares give the same value for the highest-order interaction). An interaction tells you that all factors involved are important but that their effects depend on each other. Appropriate graphs are a useful way of investigating the nature and strength of interactions.
5. When the imbalance is small, the difference between sequential and adjusted sums of squares may be minor with no difference in the qualitative outcome of the analysis (but remember the examples cited here that show cases where the differences are larger and do matter).
6. Comparing the results of different sequential analyses (including the adjusted sums of squares values contained within them) often leads to a deeper understanding than a single analysis. Focus on the model, or models, that best match the objectives of the analysis rather than searching for the single 'right' ANOVA table.

Acknowledgements We thank: John Nelder and Donald Macnaughton for helpful discussions of their published work on this topic and William Venables and Douglas Bates and for advice on analysis and R; Christa Mulder, Lindsay Turnbull, Andy Wilby and the "Brown Bag Lunch" discussion group for their comments on the article, and Maja Weilenmann for help in preparing the manuscript.

References

Aitkin, M. (1977) A reformulation of linear models - Discussion. *Journal of Royal Statistical Society. Series A,* **140,** 66.

Aitkin, M. (1978) Analysis Of Unbalanced Cross-Classifications. *Journal Of The Royal Statistical Society Series A-Statistics In Society,* **141,** 195-223.

Aitkin, M. (1995) Comments on: J.A. Nelder 'The statistics of linear models: back to basics'. *Statistics And Computing,* **5,** 85-86.

Aitkin, M., Francis, B., Hinde, J. & Darnell, R. (2009) *Statistical Modelling in R,* Oxford University Press, Oxford.

Anderson, D. R. (2008) *Model Based Inference in the Life Sciences,* Springer, New York.

Burnham, K. P. & Anderson, D. R. (2002) *Model selection and multimodel inference: A practical information-theoretic approach,* Springer, New York.

Cox, D. R., Atkinson, A. C., Box, G. E. P., Darroch, J. N., Spjotvoll, E. & Wahrendorf, J. (1984) Interaction. *International Statistical Review,* **52,** 1-31.

Fisher, R. A. (1925) *Statistical Methods for Research Workers,* Oliver & Boyd, Edinburgh.

Fox, J. (2002) *An R and S-Plus companion to applied regression,* Sage Publications, Thousand Oaks.

Gelman, A. (2005) Analysis of variance - Why it is more important than ever. *Annals Of Statistics,* **33,** 1-31.

Gelman, A. & Hill, J. (2007) *Data Analysis Using Multiple Regression and Multilevel/Heirarchical Models,* Cambridge University Press, Cambridge.

Herr, D. G. (1986) On the history of ANOVA in unbalanced, factorial designs. *American Statistician,* **40,** 265-270.

Macnaughton, D. B. (1998) Which sums of squares are best in unbalanced analysis of variance? MatStat Research Consulting Inc.

McCullagh, P. (2000) Invariance and factorial models. *Journal of the Royal Statistical Society, series B,* **62,** 209-256.

McCullagh, P. (2005) Exchangeability and regression models. *Celebrating Statistics: Papers in honour of Sir David Cox on the occasion of his 80th birthday* (eds A. C. Davison, Y. Dodge & N. Wermuth), pp. 89-110. Chapman & Hall, London.

McCullagh, P. & Nelder, J. A. (1989) *Generalized Linear Models,* Chapman and Hall, London.

Nelder, J. (1994) The Statistics Of Linear Models: back to basics. *Statistics And Computing.,* **4,** 221-234.

Nelder, J. & Lane, P. (1995) The Computer Analysis Of Factorial Experiments: In Memoriam - Frank Yates. *The American Statistician,* **49,** 382-385.

Nelder, J. A. (1977) A reformulation of linear models. *Journal of Royal Statistical Society. Series A,* **140,** 48-77.

Quinn, G., P. & Keough, M., J. (2002) *Experimental Design and Data Analysis for Biologists,* Cambridge University Press, Cambridge.

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schmid, B., Hector, A., Huston, M. A., Inchausti, P., Nijs, I., Leadley, P. W. & Tilman, D. (2002) The design and analysis of biodiversity experiments. *Biodiversity and Ecosystem Functioning* (eds M. Loreau, S. Naeem & P. Inchausti), pp. 61-78. Oxford University Press, Oxford.

Searle, S. R. (1995) Comments on: J.A. Nelder 'The statistics of linear models: back to basics'. *Statistics And Computing.,* **5,** 103-107.

Shaw, R. G. & Mitchell-Olds, T. (1993) ANOVA for unbalanced data: an overview. *Ecology,* **74,** 1638-1645.

Stewart-Oaten, A. (1995) Rules and Judgements in Statistics: Three Examples. *Ecology,* **76,** 2001-2009.

Tukey, J. W. (1977) A reformulation of linear models - Discussion. *Journal of Royal Statistical Society. Series A,* **140,** 72.

Venables, W. N. (2000) Exegeses on Linear Models (paper presented to the S-Plus User's Conference, Washington D.C., 8-9th October 1998). Washington DC.

Venables, W. N. & Ripley, B. D. (2002) *Modern applied statistics with S,* Springer-Verlag, Berlin.

Wilkinson, G. N. & Rogers, C. E. (1973) Symbolic description of factorial models for analysis of variance. *Applied Statistics,* **22,** 392-329.

Yates, F. (1933) The principles of orthogonality and confounding in replicated experiments. *The Journal of Agricultural Sciences,* **23,** 108-145.

Yates, F. (1934) The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Society,* **29,** 51-66.

Treatment:	Control (no removal)	Removal (of neighbours)	
Initial size class			Marginal means
Small	50	57	
Small	57	71	[62.25]
Small	-	85	
Small	-	-	
Cell means:	**[53.5]**	**[71.0]**	
Large	91	105	
Large	94	120	[108.87]
Large	102	-	
Large	110	-	
Cell means:	**[99.25]**	**[112.5]**	
Marginal means:	[76.37]	[91.75]	

Table 1: Hypothetical example data (n = 11) reproduced from Shaw & Mitchell-Olds (1993). The response variable, study organism height, is cross-classified by experimental treatment (experimental removal or not of neighbours) and initial target organism size (small or large). Marginal means and cell means are given in square brackets. Note that to make the degree of imbalance clearer we have indicated missing values (-) for all treatment combinations with less than four values (the maximum observed for any combination in the original dataset). In the supplement we discuss the analysis of an artificial balanced (4 x 4) dataset that could be formed by replacing the missing values in each treatment combination with the relevant cell mean.

	Source	DF	SS	MS	F	P
a)	Treatment	1	35.3	35.3	0.33	0.58315
	Size	1	4846.0	4846.0	45.37	0.00027
	Interaction	1	11.4	11.4	0.11	0.75338
	Residual	7	747.8	106.8		
	Total	10	5640.5	564.1		
b)	Size	1	4291.2	4291.2	40.17	0.00039
	Treatment	1	590.2	590.2	5.52	0.05105
	Interaction	1	11.4	11.4	0.11	0.75338
	Residual	7	747.8	106.8		
	Total	10	5640.5	564.1		

Table 2: The two alternative sequential ANOVAs for the example data.

	Source	DF	SS	MS	F	P
a)	Treatment	1	35.3	35.3	0.33	0.5831
	Size	1	4846.0	4846.0	45.37	0.0003
	Interaction	1	11.4	11.4	0.11	0.7534
	Residual	7	747.8	106.8		
	Total	10	5640.5	564.1		
b)	Size	1	4291.2	4291.2	45.22	0.0001
	Treatment	1	590.2	590.2	6.22	0.0373
	Residual	8	759.2	94.9		
	Total	10	5640.5	564.1		
c)	Treatment	1	35.3	35.3	0.37	0.5586
	Size	1	4846.0	4846.0	51.07	0.0001
	Residual	8	759.2	94.9		
	Total	10	5640.5	564.1		
d)	Treatment	1	590.2	590.2	6.2	0.0373
	Size	1	4846.0	4846.0	51.1	0.0001
	Interaction	1	11.4	11.4	0.1	0.7534
	Residual	7	747.8	94.9		
	Adjusted total	10	6195.4			

Table 3: Higher-terms omitted adjusted sums of squares (SAS type II). Sequential models that produce adjusted sums of squares for the (a) interaction, (b) main effect of treatment, and (c) main effect of initial size are shown with (d) the composite table of adjusted sums of squares.

	Source	DF	SS	MS	F	P
a)	Treatment	1	35.3	35.3	0.33	0.58318
	Size	1	4846.0	4846.0	45.37	0.00027
	Interaction (=TS)	1	11.4	11.4	0.11	0.75338
	Residual	7	747.8	106.8		
	Total	10	5640.5			
b)	TS	1	43.7	43.7	0.41	0.54284
	Size	1	4251.9	4251.9	39.80	0.00040
	Treatment	1	597.2	597.2	5.59	0.05001
	Residual	7	747.8	106.8		
	Total	10	5640.5			
c)	TS	1	43.7	43.7	0.41	0.54284
	Treatment	1	41.2	41.2	0.39	0.55438
	Size	1	4807.9	4807.9	45.01	0.00028
	Residual	7	747.8	106.8		
	Total	10	5640.5			
d)	Treatment	1	597.2	597.2	5.59	0.05001
	Size	1	4807.9	4807.9	45.01	0.00027
	Interaction	1	11.4	11.4	0.11	0.75338
	Residual	7	747.8	94.9		
	Adjusted total	10	6164.3			

Table 4: Higher-terms included adjusted sums of squares (SAS type III). Sequential models that produce adjusted sums of squares for the (a) interaction, (b) main effect of treatment, and (c) main effect of initial size are shown with (d) the composite ANOVA table of these adjusted sums of squares.

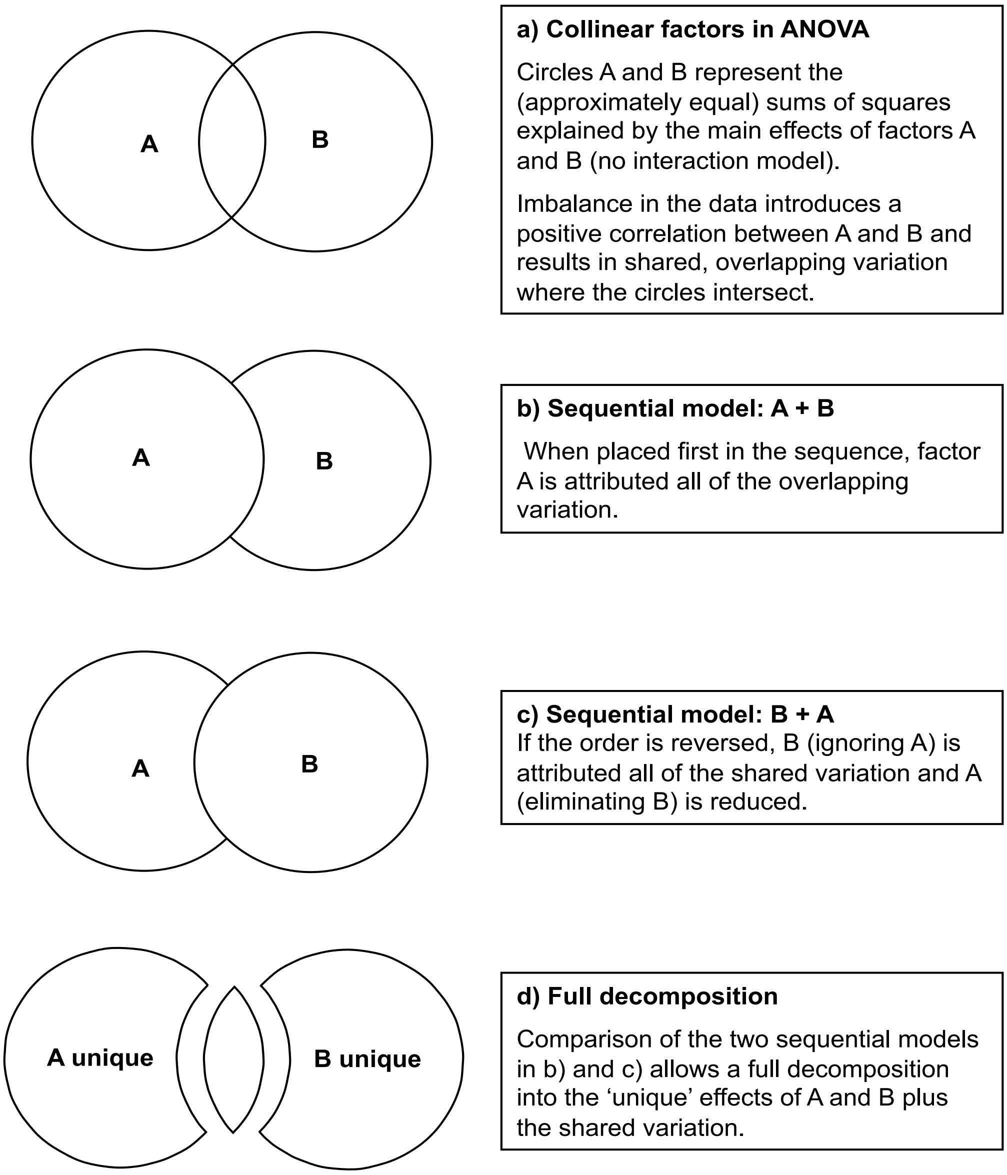

Figure 1 Venn diagram illustration of sums of squares partitioning for non-orthogonal factors A and B (without interaction) using different sequential ANOVA models (a – d). Only the sums of squares for the main effects of A and B are illustrated (the total and error sums of squares are not shown).