The following discussion took place on the CRAN website in 2005. It is very interesting from a history of statistics point of view, but is also important to know if you use R for statistics:

-----Original Message-----
From: Douglas Bates [mailto:bates_at_stat.wisc.edu]
Sent: 20 April 2005 15:07
To: michael watson (IAH-C)
Cc: r-help_at_stat.math.ethz.ch
Subject: Re: [R] Anova - adjusted or sequential sums of squares?

michael watson (IAH-C) wrote:
> Hi
>
> I am performing an analysis of variance with two factors, each with
> two levels. I have differing numbers of observations in each of the
> four combinations, but all four combinations *are* present (2 of the
> factor combinations have 3 observations, 1 has 4 and 1 has 5)
>
> I have used both anova(aov(...)) and anova(lm(...)) in R and it gave
> the same result - as expected. I then plugged this into minitab,
> performed what minitab called a General Linear Model (I have to use
> this in minitab as I have an unbalanced data set) and got a different
> result. After a little mining this is because minitab, by default,
> uses the type III adjusted SS. Sure enough, if I changed minitab to
> use the type I sequential SS, I get exactly the same results as aov()
and lm() in R.
>
> So which should I use? Type I adjusted SS or Type III sequential SS?
> Minitab help tells me that I would "usually" want to use type III
> adjusted SS, as type I sequential "sums of squares can differ when
> your design is unbalanced" - which mine is. The R functions I am
> using are clearly using the type I sequential SS.

****

I guess the real problem is this:

As I have a different number of observations in each of the groups, the results *change* depending on which order I specify the factors in the model. This unnerves me. With a completely balanced design, this doesn't happen - the results are the same no matter which order I specify the factors.

It's this reason that I have been given for using the so-called type III adjusted sums of squares...

Mick

****

Install the fortunes package and try
> fortune("Venables")

I'm really curious to know why the "two types" of sum of squares are
called "Type I" and "Type III"! This is a very common misconception,
particularly among SAS users who have been fed this nonsense quite often
for all their professional lives. Fortunately the reality is much simpler. There is,
by any sensible reckoning, only ONE type of sum of squares, and it always
represents an improvement sum of squares of the outer (or alternative) model over
the inner (or null hypothesis) model. What the SAS highly dubious
classification of sums of squares does is to encourage users to concentrate on the null
hypothesis model and to forget about the alternative. This is always a
very bad idea and not surprisingly it can lead to nonsensical tests, as in the
test it provides for main effects "even in the presence of interactions",
something which beggars definition, let alone belief.
-- Bill Venables
R-help (November 2000)

In the words of the master, "there is ... only one type of sum of
squares", which is the one that R reports. The others are awkward
fictions created for times when one could only afford to fit one or two
linear models per week and therefore wanted the output to give results
for all possible tests one could conceive, even if the models being
tested didn't make sense.

michael watson

****

Dear Mick,

The Anova() function in the car package will compute what are often called
"type-II" and "-III" sums of squares.

Without wishing to reinvigorate the sums-of-squares controversy, I'll just
add that the various "types" of sums of squares correspond to different
hypotheses. The hypotheses tested by "type-I" sums of squares are rarely
sensible; that the results vary by the order of the factors is a symptom of
this, but sums of squares that are invariant with respect to ordering of the
factors don't necessarily correspond to sensible hypotheses.

If you do decide to use "type-III" sums of squares, be careful to use a
contrast type (such as contr.sum) that produces an orthogonal row basis for
the terms in the model.

I hope this helps,
John

****

Yes, but how do you know that the test represented by the type III sums of squares makes sense in the context of your data?

The point that Bill is trying to make is that hypothesis tests always involve comparing the fits of two nested models to some data. If you can't describe what the models being compared are, how can you interpret the results of the test?

There are many concepts introduced in statistics that apply to certain, specific cases but have managed to outgrow the original context so that people think they have general application. The area of linear models and the analysis of variance is overrun with such concepts. The list includes "significance of main effects", "overall mean", "R^2", "expected mean square", ... These concepts do not stand on their own - they apply to specific models.

You are looking for *the answer* to the question "Is this main effect significant?" What Bill is saying is that the question doesn't make sense. You can ask "Does the model that incorporates this term and all other first-order terms provide a significantly better fit than same model without this one term?" That's a well-phrased question. You could even ask the same question about a model with first-order and higher-order terms versus the same model without this one term and you can get an answer to that question. Whether or not that answer makes sense depends on whether or not the model with all the terms except the one being considered makes sense. In most cases it doesn't so why say that you must get a p-value for a nonsensical test. That number does *not* characterize a test of the "significance of the main effect".

How does R come in to this? Well, with R you can fit models of great complexity to reasonably large data sets quickly and easily so, if you can formulate the hypothesis of interest to you by comparing the fit of an inner model to an outer model, then you simply fit them and compare them, usually with anova(fm1, fm2). That's it. That's all that the analysis of variance is about. The complicated formulas and convoluted reasoning that we were all taught are there solely for the purpose of trying to "simplify" the calculations for this comparison. They're unnecessary. With a tool like R you simple fit model 1 then fit model 2 and compare the fits. The only kicker is that you have to be able to describe your hypothesis in terms of the difference between two models.

With tools like R we have the potential to change statistics is viewed as a discipline and especially the way that it is taught. Statistics is not about formulas - statistics is about models. R allows you to think about the models and not grubby details of the calculations.

Douglas Bates

****

> *I guess the real problem is this:*
>
> *As I have a different number of observations in each of the groups, the*
> *results *change* depending on which order I specify the factors in the*
> *model. This unnerves me. With a completely balanced design, this*
> *doesn't happen - the results are the same no matter which order I*
> *specify the factors.*
>
> *It's this reason that I have been given for using the so-called type III*
> *adjusted sums of squares...*

...and that is completely wrong!

If there ever is a reason for using Type III SSDs, it should be that the results do not really depend "very much" on the order. This is conceivably the case in "nearly balanced" designs. (I.e. it can be viewed as an attempt to regain the nice property of balanced designs where you can read everything off of the ANOVA table.)

If the unbalance is severe, then results simply *are* dependent on which other factors are in the model - the effect of weight diminishes when height is added to a model, etc. It's a fact of life and you just have to deal with it.

```
--
   O__  ---- Peter Dalgaard
```

****

This is one of many examples of an attempt to provide a mathematical answer to something that isn't a mathematical question.

As people have already pointed out, in any practical testing situation you have two models you want to compare. If you are working in an interactive statistical environment, or even in a modern batch-mode system, you can fit the two models and compare them. If you want to compare two other models, you can fit them and compare them.

However, in the Bad Old Days this was inconvenient (or so I'm told). If you had half a dozen tests, and one of the models was the same in each test, it was a substantial saving of time and effort to fit this model just once.

This led to a system where you specify a model and a set of tests: eg I'm going to fit y~a+b+c+d and I want to test (some of) y~a vs y~a+b, y~a+b vs y~a+b+c and so on. Or, I want to test (some of) y~a+b+c vs y~a+b+c+d, y~a+b+d vs y~a+b+c+d and so on. This gives the "Types" of sums of squares, which are ways of specifying sets of tests. You could pick the "Type" so that the total number of linear models you had to fit was minimized. As these are merely a computational optimization, they don't have to make any real sense. Unfortunately, as with many optimizations, they have gained a life of their own.

The "Type III" sums of squares are the same regardless of order, but this
is a bad property, not a good one. The question you are asking when
you test "for" a term X really does depend on what other terms are in the
model, so order really does matter. However, since you can do anything
just by specifying two models and comparing them, you don't actually need
to worry about any of this.

-thomas lumley

****

Folks:

At the great risk of having my ignorance publicly displayed, let me say:
1) I enjoyed and learned from the discussion;
2) I especially enjoyed WNV's paper linked by BDR -- I enjoyed both the
wisdom of the content and the elegance and humor of style. Good writing is a
rare gift.

Anyway, I would like to add what I hope is just a bit of useful perspective on WNV's comment in his paper that:(p.14) "... there are some very special occasions where some clearly defined estimable function of the parameters that would qualify as a definition of a main effect to be tested, even when there is an interaction in place, but like the regression through the origin case, such instances are extremely rare and special."

Well, maybe not so rare and special: Consider a two factor model with one factor representing, say process type and the other, say, type of raw materials. The situation is that we have several different process types each of which can use one of the several sources of raw materials. We are interested in seeing whether the sources of raw materials can be used interchangeably for the different processes. We are interested both in the issue of whether the sources of raw materials are assoicated with some consistent effect over **all ** processes and also in the more likely issue of whether only some processes might be sensitive and others not. This latter issue can be explored -- with the caveats expressed in this discussion -- by testing for interactions in a simple 2-way model. However, it seems to me to be both reasonable and of interest to test for the main effect term given the interactions. This expresses the view that the interactions are in fact more likely than main effects; i.e. one expects perhaps a few of the processes to be sensitive in different ways, but not most of them and not in a consistent direction. I think that this is, in fact, not so uncommon a situation in many different contexts.

Of course, whether under imbalance one can actually test a hypothesis that meaningfully expresses this notion is another story ...

As always, I would appreciate other perspectives and corrections, either on list or privately.

-- Bert Gunter