# FORMULAE IN R: ANOVA AND OTHER MODELS, MIXED AND FIXED

JANUARY 10, 2013 | WILL

R's formula interface is sweet but sometimes confusing. ANOVA is seldom sweet and almost always confusing. And random (a.k.a. mixed) versus fixed effects decisions seem to hurt peoples' heads too. So, let's dive into the intersection of these three.

I'm aware that there are lots of packages for running ANOVA models that make things nicer for particular fields. I'm just going to ignore them all here and focus on the builtin function **aov** and the standard mixed model package **lme4**. I'm not even going to talk about the analysis you might do with such models, still less delve into the horrors of Type 1/2/3 sums of squares. This is just the model specification part.

In the following, assume that Y is a dependent variable and A, B, C, etc. are predictors, all contained in data frame d.

## FORMULA RECAP

If you use R then you probably already know this, but let's recap anyway.
Start with an additive model of Y using the linear model function **lm**

```
lm(Y ~ A + B, data=d)
```

Interactions are expressed succinctly with the asterisk

```
lm(Y ~ A * B, data=d)
```

or equivalently but more explicitly by specifying component parts using the colon notation, like

```
lm(Y ~ A + B + A:B, data=d)
```

This is useful for more complex interaction structures, e.g.

```
lm(Y ~ A * B * C, data=d)
```

which contains all main effects, all two way interactions, and a three way interaction. On the other hand

```
lm(Y ~ A + B + C + A:B + A:C + B:C, data=d)
```

is the same except for having no three way interaction.

If you're feeling fancy you can get the same effect as the model above by raising the variables to a power

```
lm(Y ~ (A + B + C)**2, data=d)
lm(Y ~ (A + B + C)^2, data=d)
```

which, if you remember your algebra, amounts to the same thing. Frankly I find this a bit too clever, not least because

```
lm(Y ~ A + B + B**2, data=d)
```

does *not* specify a model with a linear effect of A and a quadratic effect of B as every beginning R user and everyone who takes the algebra analogy too seriously feels that it

should.

Both of these specifications obey the *principle of marginality*, which requires, roughly, that all higher order interaction have their lower order siblings in the model unless you have a good reason. (Good reasons are shown below and tend to have to do with nesting). The asterisk and the power notation make sure your models obey this, whereas the colon invites you to forget something.

Squeezing the algebra analogy a bit further, another way to get the all two way interaction model is to make a three way model and then subtract the highest interaction term, like

```
lm(Y ~ A*B*C - A:B:C, data=d)
```

which is cute, but arguably not very useful.

Finally, remember that an intercept is almost always a good idea due to the principle of marginality, so R adds one by default and represents it with a 1. Consequently, these formulae specify the same model

```
lm(Y ~ A + B, data=d)
lm(Y ~ 1 + A + B, data=d)
```

In the model matrix the intercept really is a column of ones, but R uses it rather more analogically as we will see when specifying mixed models.

In the unlikely event we want to remove the intercept, it can be replaced by a zero, or simply subtracted. Consequently these formulae specify the same, not very sensible, model:

```
lm(Y ~ 0 + A + B, data=d)
```

```
lm(Y ~ A + B - 1, data=d)
lm(Y ~ -1 + A + B, data=d)
```

OK, enough warm up. On to the ANOVAs

## CLASSICAL ANOVA

We start with simple additive fixed effects model using the built in function **aov**

```
aov(Y ~ A + B, data=d)
```

To cross these factors, or more generally to interact two variables we use either of

```
aov(Y ~ A * B, data=d)
aov(Y ~ A + B + A:B, data=d)
```

So far so familiar. Now assume that B is nested within A

```
aov(Y ~ A/B, data=d)
aov(Y ~ A + B %in% A, data=d)
aov(Y ~ A + A:B, data=d)
```

so, nesting amounts to adding one main effect and one interaction.

## RANDOM EFFECTS IN CLASSICAL ANOVA

**aov** can deal with random effects too, provided everything is nicely balanced. Assume A is a lone random effect, e.g. a subject indicator

```
aov(Y ~ Error(A), data=d)
```

Now assume A is random, but B is fixed and B is nested within A.

```
aov(Y ~ B + Error(A/B), data=d)
```

or maybe B and X are crossed (interacted) within levels of random A.

```
aov(Y ~ (B*X) + Error(A/(B*X)), data=d)
```

Or perhaps B and X within random A are categorized by (non-nested) G and H:

```
aov(Y ~ (B*X*G*H) + Error(A/(B*X)) + (G*H), data=d)
```

Yuck. This **Error** business can get confusing and the balance requirements tiresome, so for random effects models its usually easier to move to **lme4**.

## MIXED AND MULTILEVEL MODELS

Let's start again with the lone random effects model

```
lmer(Y ~ 1 + (1 | A), data=d)
```

Random effects, like **(1 | A)**, are parenthetical terms containing a conditioning bar and wedged into the body of the formula. As the notation suggests, this is a conditional distribution of possible case level intercepts for each level or quantity of A. Still, the semantics should be familiar: **(B | A)** is equivalent to **(1 + B | A)** and the way to *not* automatically get an intercept added is to specify **(0 + B | A)** or perhaps more confusingly **(B – 1 | A)**.

Now assume that B is fixed but A is random

```
lmer(Y ~ B + (1 | A), data=d)
```

```
lmer(Y ~ 1 + B + (1 | A), data=d)
```

Now let's reconsider nested variables. If A is random, B is fixed, and B is nested within A then

```
lmer(Y ~ B + (1 | A:B), data=d)
```

Now the advantage of using **lmer** is that it is easy to state the relationship between two random effects. For example, if A and B are both random and *crossed* i.e. marginally independent, then

```
lmer(Y ~ 1 + (1 | A) + (1 | B), data=d)
```

Interestingly, if levels of (random) B are nested within levels of (random) A then the formula *looks* very much the same. However, this leads to an ambiguity.

Assume each level of A nests six levels of B, for example if we took six samples (B) from each of five subjects (A). If we label each subject's samples 1, 2, 3, 4, 5, and 6 then although there are five subjects with B=1 these samples are completely unrelated. Unfortunately, this is just the way the data would look if A and B were actually crossed factors.

Bates suggests avoiding this design ambiguity by generating a new variable to represent the nesting. In general, this is just A:B, just as it was above.

```
lmer(Y ~ 1 + (1 | A) + (1 | A:B), data=d)
lmer(Y ~ 1 + (1 | A/B), data=d)
```

This expresses the nesting and ensures that we don't accidentally do a crossed factor analysis.

Moving further into multilevel regression territory, let's assume that A is random and both

the intercept and the slope of B depend on A. With no constraint on the slope-intercept relationship we have

```
lmer(Y ~ 1 + B + (1 + B | A), data=d)
```

but if we want to force B's intercept and slope to be independent conditional on A then

```
lmer(Y ~ 1 + B + (1 | A) + (0 + B | A), data=d)
```

Here is a rare situation where it is sensible to remove an intercept, but only because the other random effect looks after it.
The second is a more parsimonious model but of course we'd want to check that the we weren't missing anything important by making slope and intercept independent.

## SOURCES AND FURTHER READING

Much of this information was gleaned from the personality-project's pages on doing ANOVA in R, from various Doug Bates course handouts, e.g. this one, and an R News article (pp.27-30), and from experimentation. A more ANOVA-focused piece is at statmethods.

R

< CONSTANTS IN LOGIT SCALES

ON THE USE AND ABUSE OF WEASELS IN SCIENCE JOURNALISM >

## 23 THOUGHTS ON "FORMULAE IN R: ANOVA AND OTHER

MODELS, MIXED AND FIXED”

**PHOSPHORELATED**

MAY 15, 2013 AT 19:42

You say:
“if levels of (random) B are nested within levels of (random) A then the formula is exactly the same. However, if both A and B are random and B is nested in A then the simple random intercept model is”

What is the difference between “levels of (random) B are nested within levels of (random) A” and “both A and B are random and B is nested in A”?

Thanks!

Reply

**WILL**

MAY 15, 2013 AT 21:31

You’re right, that was rather vague. I’ve altered the text to try to make things clearer. Also I’ve added a link to some slides at the end that give a more thorough presentation, in case that’s useful.

Reply

**PHILIPPE**

MAY 17, 2013 AT 14:41

Hi, thanks for your post which is very clear and relevant to my data.

A question, though, how would you code a two-way anova with interaction to which a random effect is applied? I.e. I got x1 and x2 explaining my y, and a z variable that have an influence on the whole experience.

```
lmer(y ~ x1*x2 + (1|z), data = my.data)
```

Would it be correct? If so, do you have recommendations regarding post-hoc tests that could apply to this formulation? I know the glht() function from the multcomp package can cope with `lmer(y ~ x1 + x2 + (1|z), data = my.data)` but as I tried to add the interaction, glht() won't work anymore...

Any help is welcome and thanks again for your post.
Best,
Philippe

Reply

**WILL**

OCTOBER 21, 2013 AT 11:59

I can't really suggest anything about the multcomp package since I've never used it. My general feeling about multiple comparisons issues is summarised in Gelman et al. 2012, though I appreciate that may not be helpful to you.

Reply

**DAZ KAMBO**

OCTOBER 20, 2013 AT 23:07

Hello,

Thanks for your post. I'm fairly new to R and if you could help me out with a permutation

of something that you've already posted, I'd appreciate it.

If I wanted: A to be fixed, B to be random and B nested within A, how would I do that using lmer?

Thanks so much!

Reply

**WILL**

OCTOBER 21, 2013 AT 11:55

I think this discussion might be relevant to your question, particularly the last answer.

Reply

**SANIYA**

JANUARY 15, 2014 AT 17:36

Thank you for a clear explanation with code examples. Of all the lme4 tutorials I've seen, you break it down the best. Thanks for saving my sanity!

Reply

**JENNIFER**

NOVEMBER 15, 2014 AT 20:18

Hi, great post!
Is there a way to do all this on nonparametric data? (strongly heteroskedastic in my case)

Reply

**LIZE**

DECEMBER 13, 2014 AT 06:45

I like your post, thank you.
Please advise me.
Consider fixed A and B, with A nested in B.
B has levels “red” and “green”.
I need separate p-values for the effect of A in red and the effect of A in green.

Reply

**WILL**

APRIL 28, 2015 AT 12:11

This sounds like an ideal question for http://stats.stackexchange.com

Reply

**EWA**

APRIL 28, 2015 AT 10:42

Hi,
Do you know why do I get different p values and F-scores when I use: aov ( y ~ x + z ) and different when aov ( y ~ z + x ) ? (sequence of x and z is changed).
Thanks!

Reply

**WILL**

APRIL 28, 2015 AT 12:09

That's because R by default reports 'incremental' a.k.a. 'Type I' tests with its ANOVA functions. There are two other alternatives – called 'Type II' and 'Type III', naturally. Falk Scholer gives a clear discussion http://goanna.cs.rmit.edu.au/~fscholer/anova.php

Reply

**SANTIAGO**

MARCH 8, 2016 AT 10:25

Hi Will,

in nested random effects, the notations "(1|A/B)" or "re(random=~1|A/B)" or "random(A)+random(A:B)", are the same? are equivalent? Thank you,

Santiago.

Reply

**BEATRICE**

JULY 14, 2016 AT 05:51

Hey, nice post!

If you have a situation like this:
lmer(Y ~ 0 + A:B + (1 | C:D:E), data=d) with A=continuous variable, (B,C,D,E)=factors
How can you remove the intercept given by the random effects as well? That is, how can

you shift the whole model so that it's origin is in (0;0)?

Thanks,
Bea

Reply

**SHAKIRAT**
NOVEMBER 28, 2016 AT 07:51

the code to use in r to Solve two way anova that is random

Reply

**SAMER AMER**
JANUARY 31, 2017 AT 05:51

Hi,
Thanks for your valuable post,
I 've used lmer to make a mixed model, it it's pretty well, but neither the ANOVA nor the summary commands are retrieving me the p values

Reply

**WILL**
FEBRUARY 3, 2017 AT 22:15

There are good reasons for that. The regular "t-test on a coefficient" thing is a bit hard to define for mixed models. Some discussion is here and a more detailed explanation

from lme4's author is here.

Reply

**HAORAN**
FEBRUARY 26, 2017 AT 18:47

Hi,

aov( y ~ A + B %in% A, data= d )

is generally not the same with
aov(y ~ A + A:B, data= d)

The degree of freedom differs.

Reply

**HAORAN**
FEBRUARY 26, 2017 AT 18:52

Oh, sorry no way! This is the same.

Reply

**JULI**
FEBRUARY 27, 2017 AT 16:50

my design is

p crossed in I, but nested in S. Any help for code?

Reply

### RICH

AUGUST 13, 2017 AT 20:31

Hi, thanks for the excellent summary of how to code the formula for different ANOVA designs. This page is my go-to resource for this now.

However one point I'm not clear on is your use of the term nested. You say " if we took six samples (B) from each of five subjects (A)" then "each level of A nests six levels of B".

But this isn't correct is it? If each level of A contains every level of B, then A and B are crossed. We could see this if we wrote out the cross-tabulation of A and B (http://www.theanalysisfactor.com/the-difference-between-crossed-and-nested-factors/)

Reply

### ADEKUNLE TOHEEB OPEYEMI

OCTOBER 12, 2017 AT 00:43

I think the correct definition of B nested in A is that:if different levels of B occur in each level of A.Take for instance,consider an investigation concerning the effect of a number of schools(A) and the effectiveness of three mathematics teachers(B) selected in each of the schools.Here,the three mathematics teachers are not concerned in any way with any of the other schools.Therefore the three mathematics teachers are nested within each schools since the teachers in one school will be different from that of the other schools.

Reply

**ÉLISE**

DECEMBER 4, 2017 AT 18:48

Thanks for this useful post.

I have a question about the way to code nested factor in the data. Should each level be identified with a unique code (I think this is required for lmer) or each factor should be repeated. Example:
3 treatments (enrichment), repeated 3 times (different places, uniques):
enrichment place place_unique
A 1 1
A 2 2
A 3 3
B 1 4
B 2 5
B 3 6

aov(y~enrichment+Error(place_unique),data=data)
#same result as:
lm(y~enrichment,data=data)

So I guess I should use repeated ID, even if each place is unique?

Reply

## LEAVE A REPLY

Your email address will not be published. Required fields are marked *

COMMENT

NAME *

EMAIL *

WEBSITE

POST COMMENT

Search … SEARCH

## PAGES

Blog

Contact

Biography

Publications

Research

Software

Austin

Content Analysis in Python

Events

Java Content Analysis tools

JFreq

Re-encoder

YKConverter

Yoshikoder

Will Lowe