Re: Comparing models to find source of super low marginal R2 values

Helene Wagner <helene.wagner@utoronto.ca>

Wed 5/18/2022 4:34 PM

To: Marc Johnson <marc.johnson@utoronto.ca>;Sophie Breitbart <sophie.breitbart@mail.utoronto.ca>

Thanks, Sophie and Marc!

I'm back at my computer and could go over the two files. I agree with what you both wrote already. And, well done, Sophie, both are great summaries/presentations. Just a couple of comments/thoughts:

- Excel file: Celia is working with HMSC, which sounds like an alternative to mvabund. It may have additional features that could be useful or not. I'm fine with you using mvabund, but it could be worth chatting with Celia, maybe she can help you get off the ground quickly with HMSC.
- R notebook: I agree with Marc that it is reassuring that results are so consistent. I would not say that the random effects add noise, rather, there is a lot of variation within populations or families. And, 7.7% definitely is a sizeable effect, so that's nice to see. I'm a little uneasy about the multiplication by 10e9. That's a lot of zero's… Just makes me wonder about rounding errors / precision issues at that level of decimals to begin with. What are the BLUP values, do they represent logits (log(p/(1-p)), or probabilities (inverse logit transformed), or something else?

Keep up the great work,
Helene

From: Marc Johnson <marc.johnson@utoronto.ca>
Date: Wednesday, May 18, 2022 at 5:19 PM
To: Sophie Breitbart <sophie.breitbart@mail.utoronto.ca>, Helene Wagner <helene.wagner@utoronto.ca>
Subject: Re: Comparing models to find source of super low marginal R2 values

Dear Sophie,

Thanks for the reminder. I've now had time to look at these results and respond to the committee meeting advice that is attached to this email.

The R markdown file was really helpful in helping me to understand your results more fully. First, I was very happy to see that the results for when block and year are treated as random vs fixed are nearly identical, which aligns with my understanding of how fixed/random effects affect statistics in most situations. The change in the r2 isn't surprising or concerning, because the r2 portion is only summarizing the variation explained by fixed effects. It had to go up, and because year will always have a big effect, it went up a lot.

I was also very happy to see that the results were consistent between the individual plant level LMM analyses and the blups LM models of population means, and in fact, the combination of the two analyses really make the results more robust and help to illuminate the biology of the results more fully. Allow me to elaborate on this point.

First, one might be a little suspicious of whether the low p-values of P=0.04 for distance from the LMM are real effects or type I error, but this worry largely goes away for me once I see the same result is the same from the blup LM model testing the effect of distance. I agree this suggests that your experimental design was powerful enough to detect these "weak" effects, even though they explain very little variation on an individual level. However, I think the test statistics (chisq and p-value) are useful in the individual plant level data for determining whether there is an effect of urbanization, but I do not think the r2 (partial or marginal) from that model is useful in telling you how important that effect is biologically. This is because the r2 is completely depended on all the other noise at the field site, which is important to control for but not biologically that interesting when studying genetic divergence among populations. So I agree year and block add noise, but it is not interesting noise. I also don't think the low r2 of distance means the effect is biologically unimportant. In the individual level plant data, the slope is the best measure of the effect size.

When we transition to the blups LM model, we really see that the divergence among populations (after factoring out the noise) is in fact quite large, explaining on the order of 7-8% of the variation in population divergence – the slope from this model should be similar to that of the individual level data, and maybe a better estimator of the slope. The r2 from the blups model is more appropriate as a measure of effect size (along with slope), because now we are just focused on how much of the variation in population divergence is explained by urbanization, and VOILA! we see that it is non-trivial.

In summary, the individual level plant data is the perfect way to estimate the test statistics, whereas I think the blups analysis is best for determining the amount of variation explained (and maybe the slope) in population divergence by urbanization.

I hope this insight helps. Looking through your document was illuminating, and I look forward to seeing all the results together so I can see the full picture again.

Well done.

Marc

From: Sophie Breitbart <sophie.breitbart@mail.utoronto.ca>
Date: Monday, May 16, 2022 at 4:42 PM
To: Marc Johnson <marc.johnson@utoronto.ca>, Helene Wagner <helene.wagner@utoronto.ca>
Subject: Re: Comparing models to find source of super low marginal R2 values

Hi again- here's my first gentle pester:

1. I've attached an updated LM vs LMM report with some updated partial R-square values for the mixed models treating block and year as fixed effects. The partial R-square values for distance are quite low (0.001-0.004).
2. Re: using Kenward-Roger or Satterthwaite corrections in glmmtmb, Ben Bolker says: "neither Satterthwaite nor Kenward-Roger are trivial to implement" in this package. He does link to code used in the lmertest package to compute var(contrast) and ddf using KR-method via the pbkrtest package, though, that I might be able to use or adapt.
3. I drew up a quick summary of the suggestions I received at my committee meeting. Do you have thoughts on which are most important to pursue?

Thanks,
Sophie

--
Sophie Breitbart (she/her/hers)
PhD Candidate
Department of Ecology and Evolutionary Biology
University of Toronto
https://sbreitbart.github.io/

From: Sophie Breitbart <sophie.breitbart@mail.utoronto.ca>
Sent: Friday, May 13, 2022 3:47 PM

To: Marc Johnson <marc.johnson@utoronto.ca>; Helene Wagner <helene.wagner@utoronto.ca>
Subject: Re: Comparing models to find source of super low marginal R2 values

Hi Helene and Marc,

I was able to solve the NA issue wherein some linear models weren't returning p-values or standard errors. I think the issue came from the BLUP estimates' extremely tiny size (on the order of 1x10^-9) and was able to make the error disappear after uniformly multiplying the response (herbivory) by a constant to make it larger.

Thus, we have clearer results from the comparison between linear mixed models vs. linear models using BLUPs as the response estimate. I've tried to clarify the main results in the introduction. In general, I think the best way forward is to treat block and year as fixed effects and continue using linear mixed models.

Thanks, have a great weekend, and see you on Monday!
Sophie

--
Sophie Breitbart (she/her/hers)
PhD Candidate
Department of Ecology and Evolutionary Biology
University of Toronto
https://sbreitbart.github.io/

From: Sophie Breitbart
Sent: Thursday, May 5, 2022 3:48 PM
To: Marc Johnson <marc.johnson@utoronto.ca>; Helene Wagner <helene.wagner@utoronto.ca>
Subject: Comparing models to find source of super low marginal R2 values

Hi Marc and Helene,

Based on a discussion we had during my committee meeting as well as with Marc last week, I've created a few linear models using BLUPs as population-level estimates and compared them to linear mixed models fitted using a nested family-within-population random effect structure. I also investigated what would happen if I changed year and block from random to fixed effects.

If I did this correctly, what I gather is that year and block should be treated as fixed effects. Generally,

When block and year are random:

- the p-values remain similar (for latex; the BLUP models don't converge for the herbivory models)
- the marginal R-squared values are much higher for the BLUP models.

When block and year are fixed:

- the p-values remain similar (for latex; the BLUP models don't converge for the herbivory models)
- the marginal R-squared values are much higher for the mixed models. Additionally, the marginal R-squared values are significantly higher compared to the models where block and year are random.

It seems like making block and year fixed help account for the super low R2m issue. Those models also converge more easily. Thoughts?

Thanks,
Sophie

--

Sophie Breitbart (she/her/hers)
PhD Candidate
Department of Ecology and Evolutionary Biology
University of Toronto
https://sbreitbart.github.io/