Testing the need for a random intercept

E-mail distributed on 17-06-2024

Dear all,

The most basic multilevel model—for data from a typical psychology experiment—contains a random intercept for subjects. This intercept allows that individuals deviate from the population mean response by a constant value. For example, in a reaction time experiment, some subjects will be slow in every condition, while others will be fast in every condition (regardless of differences between conditions). A random intercept takes into account such "baseline" deviations, by assuming a constant correlation between repeated measures within subjects, an assumption it shares with the sphericity model in repeated measures ANOVA.

The model parameter that captures the random intercept is a variance, which may be somewhat unexpected, since ordinarily we associate intercepts with mean parameters. This is a consequence, however, of the assumption that random effects contribute random and not fixed error to the outcome. The parameters of random variables are variances, not means, for example as in the case of residual error, where the residual values are not the parameters but their variance, MSE.

In R, users of lme4 will be familiar with the random intercept syntax in the formula of the lmer function, e.g.:

```
lmer(Y ~ X + (1|ID))

> Random effects:
>  Groups   Name        Variance Std.Dev.
>  ID       (Intercept) 0.2068   0.4547
>  Residual             1.1325   1.0642
> Number of obs: 300, groups:  ID, 100
```

Where ID is a subject identifier variable. The summary output will print the value of the random intercept variance, and its standard error, at the top. This value will almost always be non-zero, but what if we observed it to be near zero or exactly zero? Could we evaluate whether we need the random intercept at all? In some experiments, we may even have an explicit hypothesis that there should be no individual deviations from the population mean response (e.g., a task where chance-level performance is expected). To evaluate the need for a random intercept, there are two possible approaches **(a)** a significance test, or **(b)** an AIC/BIC comparison.

Significance test

The package lmerTest allows a significance test on random effects with the ranova function, and this includes a test on the random intercept, e.g.:

```
> ANOVA-like table for random-effects: Single term deletions
> Model:
> Y ~ X + (1 | ID)
>          npar  logLik    AIC    LRT Df Pr(>Chisq)
> <none>      4 -468.60 945.20
> (1 | ID)    3 -471.83 949.65 6.4547  1    0.01107 *
> ---
> Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This is a likelihood ratio test (LRT), where the test statistic has a chi-square distribution with degrees of freedom equal to the difference in random effects parameters between the full model and reduced model. Since a random intercept requires just one parameter, this difference would be equal to 1 for the random intercept test.

Unfortunately, **this chi-square distribution is not correct**! A variance parameter is bounded to the values [0,+∞[. Testing whether a variance is non-zero is therefore a test on the **boundary** of the parameter space. The true null distribution of the test statistic in this case is a 50:50 mixture of $\chi(0)$ and $\chi(1)$, with a critical threshold lower than under $\chi(1)$ (Fitzmaurice, Laird & Ware, 2004). The ranova function therefore underestimates the true *p*-value of the random intercept. How can we calculate the correct *p*-value? Fortunately, the R package emdbook has functions to calculate *p*-values under such mixture distributions. For the above LRT statistic, we thus obtain:

```
pchibarsq(6.454045,1,mix=0.5,lower.tail=FALSE)
> 0.005535001
```

When reporting such a test, we could write, e.g.: *“a likelihood ratio test on the random intercept variance indicated that the variance was significantly different from 0, $\chi(0:1) = 6.45$, $p = 0.0055$, suggesting that there were significant differences between subjects in the baseline mean response.”*

While the significance test approach would seem to have the benefit of allowing a formal decision on random intercept importance, I do not recommend it in general. This is partly because of the technical complication involving mixture chi-square distributions, but also because, in practice, random effects are rarely the subject of research hypotheses. Therefore testing their significance is redundant and inflating the possibility of Type I errors. I would advise to only use this test when the research hypothesis explicitly requires it.

Information criteria

A less complicated alternative to the significance test may be information criteria. This requires us to fit a reduced and full model without and with random intercept, calculate AIC or BIC, and determine which model has the lower AIC/BIC, exceeding an importance range of 2 points. Interestingly, the lmer

function does not allow the random intercept to be removed. The model must contain at least one random intercept, with formulas like (0|ID) invalid. For the model without random intercept, we should therefore switch to the base lm function. However, the following comparison is **not valid**:

```
AIC( lm(Y~X) )
AIC( lmer(Y~X+(1|ID)) )
> 942.59
> 945.19
```

The reason is that lmer models are fitted by default with Restricted Maximum Likelihood (REML), whereas the base lm is fitted by ordinary Maximum Likelihood (ML). In order to make a valid comparison for AIC, REML needs to be switched off:

```
AIC( lm(Y~X) )
AIC( lmer(Y~X+(1|ID), REML=FALSE) )
> 942.59
> 938.35
```

In this particular case, failing to be careful with the estimation method would even produce the wrong conclusion. Under ML, the random intercept model has the lower AIC by more than 2 points, and therefore should be retained. BIC also favors the random intercept model in this case, but by a difference less than 2 points. While one normally prefers the model with the least parameters (Occam's Razor), Fitzmaurice et al. (2004) caution against overly restrictive random effects structures, and recommend the AIC comparison. In fact, for the same reason, they recommend testing at a significance level of $\alpha = 0.1$ for the significance test discussed earlier.

Miscellaneous notes

Zero, or near-zero variance parameters in multilevel models are usually the source of singularity warnings printed by lme4 (e.g., "boundary (singular) fit"). For some researchers this warning can motivate the removal of redundant random effects, although this may not be possible in complicated random effects structures, if only some but not all parameters of a random effect are on the boundary.[1] In this case, one should simply proceed with the singular model, as the impact on the fixed effects (and inference) is likely negligible.

A final point of interest is that a 0 random intercept will occur automatically when the outcome variable is centered or standardized *within subjects*. By definition, this procedure forces each subject's mean to be 0, and hence individual baseline differences are removed. This procedure is standard in the pre-processing of some data, such as physiological signals (e.g., phasic skin conductance, pupil dilation), or any other measure where individual baselines are removed. It should be noted, however, that such baseline removal is strictly speaking unnecessary, since repeated measures models will automatically account for baseline differences.

[1] 0 for variance parameters, -1/1 for correlation parameters

References

Fitzmaurice, G.M., Laird, N.M., & Ware, J.H. (2004). Applied Longitudinal Analysis. Hoboken: Wiley-Interscience

--
Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79