STATISTICAL SUPPORT
SWISS CENTER FOR AFFECTIVE SCIENCES

Partial eta-squared for multilevel models

E-mail distributed on 28-09-2023

Dear all,

Effect sizes are complicated in multilevel/mixed regressions, due to the presence of the random effects. For example, Cohen's *d* requires a difference of means to be standardized by an *appropriate error term*, but this becomes ambiguous when the model's error variance consists of multiple sources (random intercepts) and possibly depends on values of predictor variables (random slopes). In the latter scenario, if the error variance changes with predictor values, then so should Cohen's *d* (see Fig. 1), and hence there will not be one consistent effect size. Although one could select "interesting" predictor values (e.g., the mean) to condition on, and calculate only the effect size for this value, the precise calculation of the required error term will become complex very quickly in models that contain multiple random slopes.

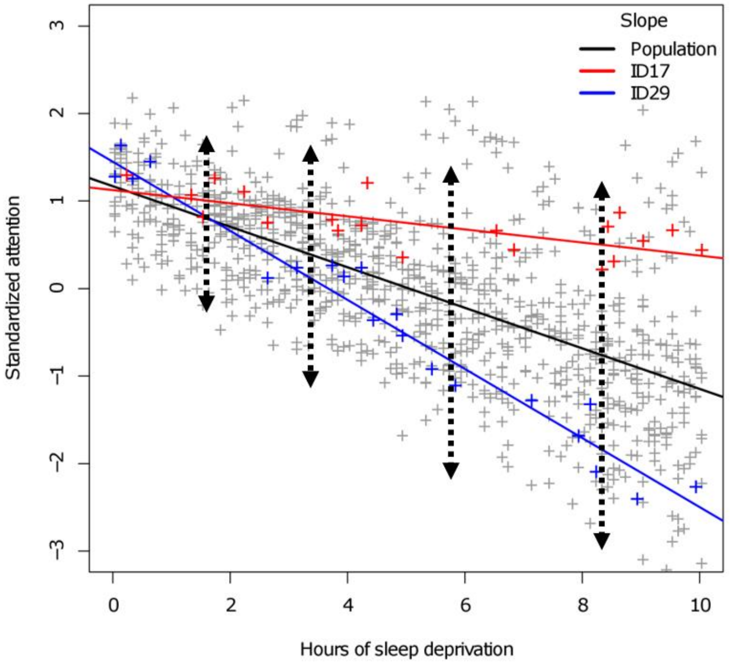

Figure 1. Heteroscedasticity due to random slopes in a sleep deprivation study.

Alternative effect sizes include **(a)** standardized regression parameters (e.g., with R packages `parameters` and `effectsize`), **(b)** Bayes factors (with R package `brms`; but complicated!), or **(c)** approximate Bayes factors with AIC or BIC ratios. However, some of these have various drawbacks, such as not allowing omnibus effects, not having confidence intervals, or being complicated technically to compute. What about **partial eta-squared**?

The "squareds" family

Recall that R-squared quantifies the proportion of variance in the outcome data that is explained by the model's predictors. Likewise, partial R-squared is the proportion of variance explained for a specific predictor/effect, when controlling for the others. It can be calculated by differencing the R-squared of a full model with the R-squared from a reduced model that omits this effect (leave-one-out approach). In ordinary regression models, especially factorial ANOVA models, partial R-squared is known as partial eta-squared.

R-squared has an important drawback, it is sensitive to overfitting, in that its value can only *increase* when adding predictors, no matter how unimportant or random they are. **Adjusted R-squared** therefore corrects R-squared by a factor related to the number of model parameters. Its partial versions are known as **partial epsilon**- and **partial omega-squared**. However, conceptually, all these measures retain more or less the same interpretation (proportion of variance explained), and their size can be interpreted qualitatively according to rules. In R, the package effectsize offers the functions `eta_squared`, `epsilon_squared`, and `omega_squared` for their computation, with various options for confidence intervals and generalized versions.

R-squared in multilevel models

In multilevel regression, the calculation of R-squared is complicated, again due to the presence of random effects. For the full model, users may already be familiar with the distinction between **marginal R-squared** and **conditional R-squared**. The former exclusively considers variance explained by fixed effects (averaged over random effects), and resembles most closely traditional R-squared in ordinary regression. The latter conditions explained variance on random effects. For many data sets, conditional R-squared is substantially higher than marginal R-squared, especially when strong individual differences are present. In fact, the difference between marginal and conditional R-squared could be taken as a rough estimate of the proportion of variance explained by individual differences in the data, not accounted for by fixed effects. Since one cannot explain *why* such individual differences remain unaccounted for, however, the general utility of conditional R-squared is limited, and its high values not necessarily impressive. In R, the function `MuMIn::r.squaredGLMM` calculates these statistics for a multilevel regression fitted with `lmer` from `lme4`.

R-squared for multilevel models is not without controversies. Rights and Sterba (2019) have suggested that there as many as 12 different ways of partitioning variance in multilevel regression, and even when limited to just 2 variants, different R implementations sometimes give different numbers. Moreover, a multilevel equivalent of **adjusted R-squared** is not available, as yet. For practical purposes, however, marginal R-squared for a full multilevel model can be interpreted more or less the same as R-squared for ordinary regression, and should definitely be reported in your results.

Partial eta-squared in multilevel models

The picture complicates even further when we wish to consider partial R-squared for a multilevel model. In general, it is not recommended to calculate such values manually, i.e., by comparing a full and reduced model, because the two models may differ in estimation method (ML versus REML), and because of the potential ambiguity that is created when a fixed effect is removed without removing its corresponding random effect (if it is present).

In R, the function `r2glmm::r2beta`, gives a "semi-partial" marginal R-squared breakdown for a fitted multilevel regression, including confidence intervals. However, there are some problems with this breakdown:

- No omnibus R-squareds for multi-parameter effects (e.g., factors with more than 2 levels)
- The values for the individual parameters do not add up to the model's total marginal R-squared
- The values tend to be small, even for effects that appear to be important
- Three different calculation methods are available, with no clear choice among them

Especially the lack of omnibus values makes this breakdown impractical to use for factorial ANOVAs in multilevel models.

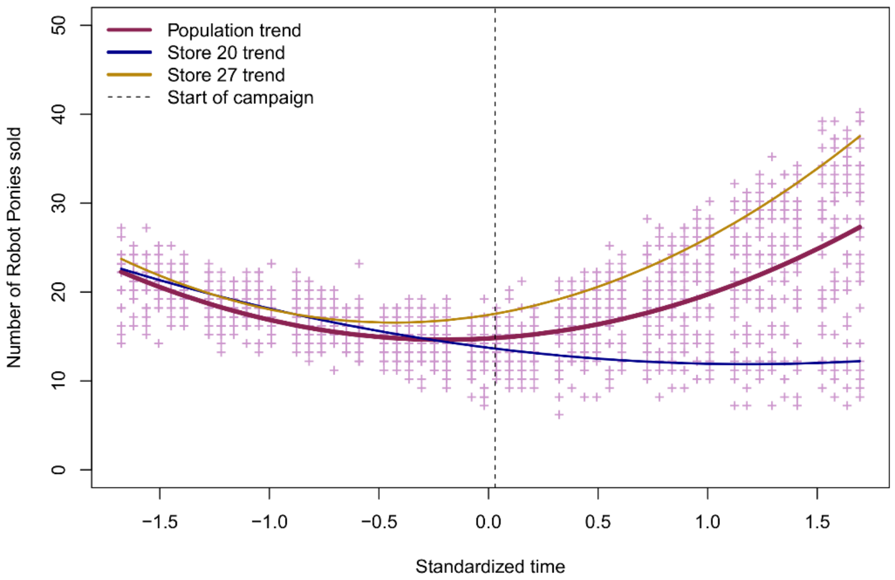

Figure 2. Sales data for toy stores, with advertising intervention.

As an alternative, the R package `effectsize` now offers a partial eta-squared breakdown for multilevel models, either by plugging a `lmer` model or its ANOVA breakdown (e.g., with `car::Anova`[1])

[1] Note that `car::Anova` uses the Kenward-Roger approximation to the DDF, which can be computationally heavy. For the faster Satterthwaite approach, the `lmerTest::anova` function is recommended.

into the `eta_squared` function. Likewise, similar breakdowns can be obtained with the `epsilon_squared` and `omega_squared` functions. For example, in the robot pony sales data from my multilevel workshop, we want to fit a quadratic curve to the longitudinal sales, while checking for moderation by store location, and individual differences in sales curve for specific stores (see Fig. 2):

```
library(lme4) ; library(lmerTest)
library(car) ; library(effectsize)
ponies <-
read.table("https://drive.switch.ch/index.php/s/EhA63ePqbJrqUod/download",
header=TRUE,sep=",",as.is=FALSE)

model <- lmer(Sales~poly(Time,2)*Location+(1+poly(Time,2)|Store),
data=ponies, REML=TRUE)
summary(model)
aov <- Anova(model,type=2,test="F")
aov

eta_squared(aov,alternative="two.sided")
```

Which produces:

```
# Effect Size for ANOVA (Type II)
Parameter              | Eta2 (partial) |       95% CI
------------------------------------------------------
poly(Time, 2)          |           0.98 | [0.97, 0.99]
Location               |       3.93e-03 | [0.00, 0.15]
poly(Time, 2):Location |           0.02 | [0.00, 0.16]
```

Note the use of Type II ANOVA, in order to produce proper marginal main effects in the presence of interactions! So we get a neat breakdown of effects, including omnibus effect sizes for multi-parameter effects like the quadratic Time effect, and we get 95% confidence intervals. Where are these values coming from? The package documentation explains that these partial eta-squareds are derived from the ANOVA's test statistics (*F*-values). This approach comes with a **massive caution**.

Test statistics from effect size

Test statistics reflect both effect size and sample size (e.g., the larger the sample, the larger the *F*-value). Therefore, one can in principle obtain an effect size when removing the sample size bias, for example with a formula that includes the denominator degrees-of-freedom (DDF) of the test distribution. This is the idea behind converting test statistics to effect sizes, and would work alright as long as the DDFs are well-defined, which is unfortunately not the case for multilevel regressions.

In multilevel models, the DDFs of test statistics depend substantially on **(a)** the approximation method (e.g., none, Satterthwate, Kenward-Roger[2]) and **(b)** the complexity of random effects (e.g., multiple random intercepts, random slopes). In a typical repeated measures dataset, these choices can

[2] Some have argued that the null distribution of these test statistics is unknown, and therefore approximating any DDF is pointless.

produce DDFs as high as the number of observations, and as low as the number of subjects. For example, in a multiple-trial experiment with 30 participants and 60 trials, the DDF could be somewhere between (approximately) 30 and 1800. Therefore, for the same *F*-values, the resulting partial eta-squareds can differ substantially, when the DDF-correction is 30 or 1800. For the example above, removing the random time curve has a dramatic impact:

```
# Effect Size for ANOVA (Type II)
Parameter              | Eta2 (partial) |       95% CI
------------------------------------------------------
poly(Time, 2)          |           0.45 | [0.35, 0.49]
Location               |           0.02 | [0.00, 0.21]
poly(Time, 2):Location |           0.01 | [0.00, 0.02]
```

This is because the DDF increased from 27 to 1526, for a roughly similar *F*-value! The change is especially paradoxical for these data, because while the more complex model should be preferred (the random time curve is important), it lowers the DDF so much that almost no correction is applied when `eta_squared` is run on the *F*-values, resulting in an unrealistically high partial eta-squared of nearly 1! In this case, the model without the random time curve would produce the more realistic effect sizes.

Note finally that, In this example, there is no appreciable difference between eta- epsilon-, and omega-squared, but this may be different in larger models containing many predictors. The latter two values produce negative estimates for the Location effect and the interaction but this is a known possibility for adjusted R-squared measures.

Conclusions

As a general conclusion, I continue to urge caution with effect sizes for multilevel models, including partial eta-squared. The breakdown offered by package `effectsize` has many appealing advantages, such as Type II breakdowns, omnibus effects, and confidence intervals. However, one must keep in mind that the values were converted from each effect's *F*-values and DDFs, and hence depend strongly on the model's random effects complexity.

Best,
Ben

--
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