14.6 Effect size in factorial ANOVA, ANCOVA and multiple regression

Adding multiple predictors to a regression model, whether categorical or continuous, complicates the interpretation of effects. This applies also to the problem of estimating the magnitude of effects and selecting an appropriate effect size statistic. One approach, advocated earlier, is to emphasize effect size metrics such as simple differences in means and unstandardized slopes – particularly where the original units of analysis are not arbitrary (Baguley, 2004; 2009). Some of the advantages of simple effect size metrics are reduced in non-experimental designs if some of the predictors are unreliable (Ree & Caretta, 2006; Hunter & Schmidt, 2004). However, selecting an appropriate standardized effect size metric presents additional difficulties.

Several difficulties with R^2 type measures emerge from the large number of parameters in the model. First, proportion of variance explained needs to be corrected to account for over-fitting. The most common correction is adjusted R^2. This adjusts R^2 for the number of parameters in the model, but not from over-fitting arising from other sources (e.g., cherry-picking the best model from a large set of models). Second, when more than one predictor is included in the model, the choice of partial effect size statistic will matter (e.g., η_p^2 or η^2). Third, effect size statistics for multiple *df* effects don't distinguish the direction or pattern of the effect. This makes them dangerous when interpreted unthinkingly. Two effects of a similar size could represent completely different effects (e.g., in opposite directions).

These problems are particularly acute for ANOVA and ANCOVA designs. In multiple regression analyses it is more common to focus on individual predictors and there is more emphasis on the sign and magnitude of their slopes. Of particular importance in ANOVA and ANCOVA is whether to use a partial effect size statistic. These statistics are used to compare the contribution of categorical or continuous predictors within and between models. For these comparisons to be sensible it is important to consider the possible presence of factors such as range restriction or differences in reliability. Even if these factors do not distort the result, any comparison needs to consider the design of the study.

To understand the basic problem, consider two studies. One in which the factor *A* is investigated on its own (i.e., in a one-way ANOVA) and another in which both factor *A* and factor *B* are investigated (in a two-way factorial design). What effect size metric will make it possible to compare factor *A* across the two designs? The most common choice in this situation is η_p^2 (partial eta-squared). Using η_p^2, the argument goes, strips out the extra variance in the two-way design (associated with *B* and the *A* × *B* interaction). It does this by using $SS_A + SS_{error}$ in place of SS_{total} as the denominator for the proportion. A hidden assumption is that including factor *B* increases the total sums of squares in the study. This assumption is probably only reasonable if *B* is a manipulated factor – as found in a completely randomized experimental design (Gillett, 2003; Olejnik & Algina, 2003). A manipulated factor (if it has an effect) adds to the total variance in the sample by causing differences in the means that would not have otherwise have been present. In contrast, measured (individual difference) factors (Olejnik & Algina, 2003) are not expected to increase the total variance. If factor *B* were the sex of the participants, there would be no expectation that the total sums of squares would increase in the two-factor design relative to the one-factor design. This counter-argument assumes that the one-factor design sampled all levels of *B*. However, on the rare occasions that the one-factor design sampled only males or only females, *B* would behave like a manipulated factor rather than a measured one. Thus the crucial distinction is whether the second factor adds variation relative to the one-factor design used as a comparator.

A partial statistic such as η_p^2 is appropriate if the extra predictors in a model add variance to the model and the objective is to compare effect sizes of a common effect. This most often (but not always) happens with manipulated factors. If the extra predictors don't add variance (as tends to be the case with measured factors) η^2 is more appropriate. This argument extends to ANCOVA. More often than not, covariates and continuous predictors will be measured rather than manipulated variables – particularly where a covariate is added to increase power in a randomized design or to control for a confound. While η_p^2 is a tempting choice for researchers, because it is always at least as large as η^2, it is nearly always a bad choice for ANCOVA designs. Even in ANOVA with apparently manipulated factors it can be misleading. Consider an experiment looking at memory for faces. One factor might be the symmetry of the face (symmetrical or asymmetrical). This is manipulated by the experimenter (using image processing software), but the choice of the effect size statistic is not trivial. Faces used as experimental stimuli and those encountered in everyday life vary in symmetry. To generalize to another study involves knowing whether the variability in the experiment is similar or dissimilar to that in the experimental materials used by others (which may be unknown). Generalizing to everyday life requires knowing the degree of symmetry typically encountered in the population. Baguley (2009) also points out that variance explained measures traditionally treat factors as fixed effects (see Key Concept 7.1). Where levels of a factor are samples from a larger population, and that population is the target of generalization, standard measures may greatly overestimate standardized effect size.

Where it is known that a predictor adds to the variability in one study (relative to another) it is possible to compute Olejnik and Algina's η_g^2 (generalized eta-squared). Olejnik and Algina (2003) provide detailed examples, but the general principle is to compute proportion of variance explained including only factors and covariates that add to the total variance (that for convenience will be referred to as manipulated variables):

$$\eta_g^2 = \frac{SS_{effect}}{I \times SS_{effect} + \sum_{meas} SS_{meas} + \sum_{error} SS_{error}} \qquad \text{Equation [14.21]}$$

In this formula I is a dummy indicator variable that equals 1 if the effect is manipulated and 0 otherwise. The $\sum_{meas} SS_{meas}$ term is the sum of *SS* for measured factors or covariates and $\sum_{error} SS_{error}$ is the sum of all sources of error variance (e.g., the pooled error term for a typical independent measures design). The $\sum_{meas} SS_{meas}$ term also includes any interactions terms between measured and manipulated factors.

As eta-squared statistics are based on sums of squares they are descriptive measures of sample variance and inevitably overestimate the proportion of population variance accounted for by a factor or covariate. One alternative is ω^2 (omega-squared). This is an estimate of the proportion of explained population variance in ANOVA. Olejnik and Algina describe how to calculate a generalized version of the statistic. For an independent measures design it can be represented as:

$$\omega_g^2 = \frac{SS_{effect} - df_{effect} \times MS_{error}}{I \times \left(SS_{effect} - df_{effect} \times MS_{error}\right) + \sum_{meas} \left(SS_{meas} - df_{meas} \times MS_{error}\right) + N \times MS_{error}}$$

Equation [14.22]

The $\sum_{meas} \left(SS_{meas} - df_{meas} \times MS_{error}\right)$ term sums over all measured variables whether factors or covariates (and including their interactions with manipulated variables). As before, manipulated and measured variables are shorthand for variables that do or don't add variance to what is being measured. Olejnik and Algina (2003) provide tables of formulas for specific designs with different combinations of manipulated and measured variables. These tables can simplify computation of ω_g^2 or η_g^2.

Of these statistics only η_p^2 is routinely reported by statistics packages. It has a direct connection to R^2, being the change in R^2 when a factor or covariate is added to a regression. It is therefore convenient for sample size estimation (at least if an identical or near-identical replication is assumed). It overestimates the proportion of variance explained, either because of its bias as a population estimate or because it excludes measured variables. It should not be relied on for comparing effects within or between studies. Its generalized counterpart η_g^2 should be preferred and is almost as easy to calculate (provided you can evaluate whether a factor or covariate adds variance to the analysis relative to some comparator study). A safer option may be to focus on simple, unstandardized effect size metrics. For generalizing to a population, ω_g^2 is preferable to η_g^2, but is hardly ever reported (which limits is usefulness for comparing between studies). Although both statistics are biased, the bias is smaller for omega-squared based measures (Howell, 2002; Olejnik & Algina, 2003).

A final concern is that any measure that decomposes the total sums of squares is only really meaningful in ANOVA or ANCOVA with orthogonal factors (or where imbalance is negligible). If the choice of method for partitioning *SS* changes the outcome of the calculation (as it does in unbalanced designs) then ω^2 and η^2 metrics may not be interpretable.

Example 14.9. Consider the diagram data from Example 13.7. The model for these data had a single factor and a single covariate. For any model, η^2 is the proportion of effect *SS* relative to total *SS*. SS_{total} is 3023.9 so the η^2 statistics are:

$$\eta^2_{(factor)} = \frac{879.26}{3023.90} = .291 \quad \eta^2_{(covariate)} = \frac{339.65}{3023.90} = .112$$

The η_p^2 statistic replaces the denominator with the sum of SS_{effect} and $SS_{residual}$. As $SS_{residual}$ = 2100.54, the η_p^2 statistics are:

$$\eta^2_{p(group)} = \frac{879.26}{583.70+2100.54} = .328 \quad \eta^2_{p(covariate)} = \frac{339.65}{339.65+2100.54} = .139$$

This statistic partials out the effect of any other predictors in the model.

Calculating generalized versions of the statistic involves determining whether factors and covariates are measured or manipulated. The factor would usually be considered a manipulated variable because it represents an experimental manipulation (the type of instructional text participants are exposed to). This manipulation adds to the variability of *Y* (the description quality). A covariate would normally be considered a measured variable that varies naturally and is not manipulated by the experimenter. This is probably true for the diagram study. If so, η_g^2 for the group effect is:

$$\eta_g^2 = \frac{SS_{factor}}{I \times SS_{factor} + \sum_{meas} SS_{meas} + \sum_{error} SS_{error}} = \frac{879.26}{1 \times 879.26 + 339.65 + 2100.54} = \eta^2 = .291$$

For the covariate it is:

$$\eta_g^2 = \frac{SS_{covariate}}{I \times SS_{covariate} + \sum_{meas} SS_{meas} + \sum_{error} SS_{error}} = \frac{339.65}{0 \times 339.65 + 339.65 + 2100.54} = \eta_p^2 = .139$$

With only two effects it is inevitable that η_g^2 reduces to either η^2 or η_p^2, but with additional effects it could fall somewhere between these two extremes. The calculations clarify the role of the indicator *I*. If the indicator were not present, $SS_{\text{covariate}}$ would be counted twice. The indicator is there to stop an effect that is itself a measured variable contributing more than once.

What is the interpretation of the generalized statistic? For the factor, η_g^2 is the proportion of sample variance accounted for by group. For the covariate it is the proportion of sample variance that would be accounted for by the covariate had there been no experimental manipulation. In principle, both statistics are now comparable to similar studies with different combinations of predictors. The logic is that the manipulated factor needs to account for the extra variation it introduces to the study, but the measured factor does not. The variation due to a measured factor should be present in the error term of other similar studies (regardless of whether it is included in the model).

There is one further thing to reflect on. Olejnik and Algina (2003) treat covariates as measured variables. This is reasonable most of the time, but sometimes not. Reading time presumably varies from text to text and person to person. Treating it as a measured variable makes sense if the goal is to determine the practical impact of the covariate on reading an instructional text. If so, ω_g^2 (which is a population estimate) would be better. But if the goal is to compare proportion of variance explained between experiments (a reasonable aim for η^2 measures) then perhaps reading time should be treated as a manipulated factor. Many similar experiments control exposure to the text by limiting reading time to a fixed period (e.g., 15 minutes). In such studies reading time is not free to vary. To compare between experiments it would be necessary to treat this particular covariate as a manipulated factor (i.e., additional variation attributable to the design). The decision of whether a variable is manipulated or measured therefore depends on the type of variable, the context of the study and the design of the study you wish to compare the statistic with.

14.7 Statistical power to detect interactions

The statistical power to detect an interaction effect is often very low. McClelland and Judd (1993) reviewed a number of factors thought to be responsible for this lack of power. The problem is most acute in observational as opposed to experimental research. There are several reasons for the difference, but the most important have to do with the increased power that comes from being able to manipulate variables or minimize error in an