

3. Statistical methodology

In this chapter I describe the methods, statistics, and models used and reported. The same methodology is used across the whole technical report, unless otherwise stated.

Correlation measures

Correlations between two variables

The Pearson's product-moment correlation coefficient is used to measure the correlation between pairs of variables. This measure, or its square, can be used to measure the size of the effect.

When using the correlation coefficient alone, values close to -1 or 1 are taken to indicate a strong relationship; those about -0.5 or 0.5 to indicate a moderate relationship, those about -0.3 or 0.3 to indicate a weak relationship, and those between -0.2 and 0.2 to indicate very little or no relationship.

The square of the correlation coefficient gives the proportion of the variability in the outcome variable that is explained by the variability in the explanatory variable (or of the one variable that is explained by the other if neither can be considered to be an "outcome" variable). This means that a correlation of 0.2 explains 4 percent of the variability, one of 0.5 explains 25 percent of the variability, and one of 0.7 explains 49 percent.

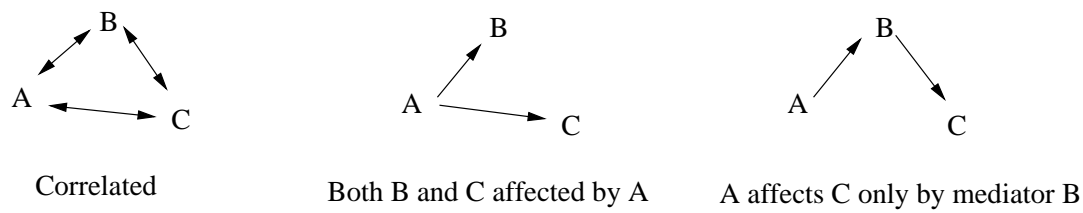
Correlations between more than two variables

Where we are interested in "clumps" of interrelated variables, simple correlations tell only part of the story. If we know, for example, that engagement, being absorbed in learning, not being disengaged, and being focused and responsible are all correlated, what pattern do we have as a basis for the correlatedness? Maybe all four variables do affect the values of the others. Or maybe, some only appear to do so.

A classical example of this is the apparent correlation between the number of books published in a year and the number of traffic accidents in the same year. In actual fact, there is no direct relationship between them, but both are reflecting economic and demographic changes that result in both the number of cars on the road (and so accidents) and the number of books bought (and so published) increasing over time. This could be illustrated by the middle diagram in 0, where the year is A , and B and C are the number of accidents and books, respectively.

A third possibility is that there is a mediating variable. A mediating variable is one where variable A is correlated with variable B which is in turn correlated with variable C . This means that A is apparently correlated with C , but in fact is so only through the mediating variable B .

Figure 1: Possible correlation mechanisms



In the first diagram on the left, if the model was causal, and C was the outcome variable, then the arrows from A and B to C would have a single arrow, at the C-end. Depending on the relationship between A and B that, too, could be a single-headed arrow (if one of A or B “caused” changes in the other), or it could be double-headed (if no causal relationship exists, they are correlated).

In this first diagram, as all three variables are correlated, the correlation between any two of them will be less strong once the effect of the third variable is taken into account. In particular, in a linear model in which the explanatory variables are intercorrelated, the simple correlation between any one of the explanatory variables and the outcome variable is stronger than the *partial correlation*¹ which is corrected for or accounting for the other explanatory variables. In a sense, this is the “left-over” correlation between one explanatory variable and the outcome variable, after accounting for the other explanatory variables in the model. The more strongly the explanatory variables are correlated, the greater the reduction in size between the simple correlation and the partial correlation. In this report, where the models contain many approximately continuous variables, we report the partial correlations.

Models fitted

In all cases where we model an outcome variable on several dependent variables, we used R (R Development Core Team, 2007) to fit a linear model (using the `lm` function).

Because there were so many possible variables to include in the model, we report each time the minimum model, containing only the variables that were statistically significant. An added reason for reporting the minimum model is that where there were two or more strongly correlated variables that could have been included in the model, fitting reduced models allowed us to select the variable that accounted for more variability in the model. Two strongly correlated variables cannot be included in the same model because of multicollinearity. This term reflects that if two or more variables are measuring almost the same thing, at least one is redundant; a model containing both is very difficult to estimate accurately, and the model estimates obtained can vary widely if small changes are made to the model, or be of the opposite sign to that expected. The variance inflation factor (the extent to which the variance of a regression estimate is inflated by multicollinearity) gives a measure of the extent to which multicollinearity is present in a model. If the

¹ The partial correlation between *X* and *Y* holding a set of variables fixed (list) will have the same sign as the multiple regression coefficient of *X* when *Y* is regressed on *X* and the set of variables being held fixed. The partial regression can be calculated from the *t*-statistic for the coefficient of *X* in the multiple regression of *Y* on *X* and the variables in the list, and the residual number of degrees of freedom:

$$r_{XY.\text{list}} = \frac{t}{\sqrt{t^2 + \text{Resid. df}}}$$

explanatory variables are independent, the variance inflation factors (VIFs) are all 1. VIFs of 4 or more (certainly of 10 or more) can indicate possible problems.

All our models were checked for multicollinearity, and where there was an indication of a problem, at least one of the collinear variables was dropped from the model.

In models that were not collinear, variable selection was done by a combination of examining the *F*-test (Type III sums of squares) for the variable, or whether the variable added significantly to the model if it were fitted last, change in the AIC (Akaike's Information Criterion) if the variable were excluded from the model, and the LMG estimates of the percentage of the variability accounted for by the model (see below).

Effect sizes

In all the linear models fitted, the size of the effect (coefficient) is given. In addition, several measures of "effect size", indices that measure the strength of association between one variable and another (or "signal to noise ratio"), are given.

The most commonly used effect size measure, Cohen's *d*, is seldom used here, as it is most appropriately used to compare two groups, or an experimental and control group. The measures that are used are those appropriate for measures of correlation, or for linear models where several explanatory (independent) variables are used to explain the variability in a single outcome (dependent) variable.

Correlation measures

Both the simple (Pearson's product-moment) correlation coefficients and partial correlation coefficients can be considered measures of effect size. The coefficients measure the strength of the relationship, and the squared coefficients measure the proportion of variability explained.

In the linear models fitted, the partial correlation coefficients are given as one of the measures of effect size.

Percentage of variability accounted for

In more complex models there are several ways in which to measure this.

Multiple R²

R^2 is the percentage of variability in the outcome variable that is accounted for or explained by all the explanatory variables.

$$R^2 = \frac{SSM}{SST},$$

Where *SSM* is the model sum of squares and *SST* is the total sum of squares. R^2 is the proportion of the variability in the outcome variable (as measured by the total sum of squares) that is accounted for by the model (as measured by the model sum of squares). This value tends to slightly overpredict the true value (it is biased). In models including only continuous variables (one parameter per variable) we quote R^2 .

A less biased estimate is given by the adjusted R^2 , which uses mean squares rather than sums of squares, and so is in a sense adjusted for the number of parameters in the models (related to the number of explanatory variables). The difference between the two versions is slight if there is a single explanatory variable, and is more marked in larger models, particularly those with discrete variables (ANOVA or ANCOVA) fitted using dummy variables. In models including several discrete variables (often more than one parameter per variable) we quote adjusted R^2 .

The overall (adjusted) R^2 gives an indication of how well the model as a whole fits the data. However, interest is usually greater in how useful each explanatory variable is in explaining the variation in the outcome variable. There are two approaches here, the one calculating the change in the (adjusted) R^2 when one of the explanatory variables is included in the model, and the other, the "eta squareds", gives an estimate of the proportion of the variability in the outcome variable that is accounted for by one of the explanatory variables. η^2 or η_p^2 are similar measures, but use a different estimate of variability in the denominator.

R² change

To calculate these quantities, a series of models needs to be fitted: one including all explanatory variables, and then a number of models leaving out each explanatory variable in turn. These changes in R^2 can be calculated whether the explanatory variables are categorical or approximately continuous. The traditional version of these measures has the disadvantage that the percentages calculated for each explanatory variable will typically not add up to exactly the value of the total R^2 for the whole model, as in social science research the data are usually not balanced (the numbers of observations in each category of an explanatory variable are not equal). But the relative sizes of the changes in R^2 will give an idea of the relative importance of each of the explanatory variables.

Recently, alternative measures that do add to R^2 have been proposed. These measures are more computing intensive, as they involve fitting the model in all possible orders (so for three explanatory variables, A, B, and C, fitting the models in the orders ABC ACB BAC BCA CAB CBA) and taking the average amount of variability explained for each variable if it was fitted last (Grömping, 2006, 2007).

When all the variables in a model are continuous or binary (taking only two values, like gender), we present the LMG (Lindeman, Merenda, & Gold, 1980) measure of relative importance, calculated using the relaimpo package (Grömping, 2006) in R (R Development Core Team, 2007) which is the unweighted relative importance across all possible specification orderings of the model. At present, no equivalent measure using the same software is available for discrete variables with more than two levels. Bootstrapped (see below) confidence intervals for the LMG estimates are available (although they may be somewhat liberal, or too wide).

The eta squareds

There are three of these effect size measures: the first two are estimates of the degree of association in the sample, and the third is a measure of the degree of association in the population. All give, in a sense, the proportion of variability in the outcome variable that is explained by an explanatory variable.

$\eta^2 = SS_{\text{effect}}/SS_{\text{total}}$ gives the proportion of the total variability (in the outcome variable) that is explained by the particular effect (explanatory variable). This measure is not that commonly used, as it tends to be positively biased (it gives estimates that are too high), and is sensitive to the number of observations and other explanatory variables in the model.

If there is only one explanatory variable, then $\eta^2 = R^2$.

A preferred measure (this is the one produced by, for example, SPSS), is partial η^2 or

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}}.$$

η_p^2 gives the proportion of the variability explained by a variable relative to the total amount of variability explained by that variable plus unexplained variability (or error variability). Each η_p^2 calculated for a model takes account of the effect of all the other variables, as the error SS used in the calculations is the residual error left over after all the explanatory variables have been added to the model. This does, however, mean that the η_p^2 s are not additive, and in fact they can sum to a number greater than 1 (or 100 percent).

When the model includes one or more discrete variables with more than two levels, we give η_p^2 .

Bootstrapping

Bootstrapping is a computer-intensive method of estimating sampling error. Using the existing data as a "population", repeated samples with replacement (i.e., an observation can be selected more than once in any one sample) are selected from the data, and the variability between these samples is used to obtain estimates of confidence intervals (or other parameters) that would otherwise be difficult or impossible to estimate, or where the assumptions required for theoretical estimates are unlikely to be met.

