When correlations go bad …

The noted statistician C. P. Winsor once established a *Society for the Suppression of the Correlation Coefficient*. According to John Tukey it had as its "guiding principle […] that most correlation coefficients should never be calculated" (Tukey, 1954, p.38). Nor were Tukey and Winsor alone:

> "The idea of regression is usually introduced in connection with the theory of correlation, but it is in reality a more general, and, in some respects, a simpler idea, and the regression co-efficients are of interest and scientific importance in many classes of data where the correlation coefficient, if used at all, is an artificial concept of no real utility."
>
> [Fisher, 1925, p.129]

Some might attribute this stance to Fisher's rivalry with Karl Pearson (of the eponymous product-moment correlation coefficient *r*). Yet this would be to miss the point.

To appreciate the advantages of regression over correlation the first step is to understand how they are related. The relationship is easiest to explain in terms of simple linear regression (a bivariate regression between two variables). In regression *Y* is predicted from *X*, and a linear regression finds the straight line that predicts most accurately (by minimizing the sums of the squared vertical distances of observations from the line). The output of the regression is an equation of the form

$$Y = b_0 + b_1X_1$$

The intercept of the line is the constant b_0 and represents the value of Y when $X = 0$ (where the regression line crosses the Y axis when plotted). The slope of the line is b_1 and represents the expected increase in Y when X increases by 1. What if we wanted to predict people's earnings (measured in US dollars) from their height (measured in inches). For a US sample[1] a simple linear regression gives the line:

$$earnings = -60515 + 1256 \times height$$

Each extra inch in height is associated with an increase in average earnings of $1,256. This equation can readily be used to predict the earnings of based on a person's height (e.g., $14,845 for a person who is 60 inches tall).

There are many different ways to view a correlation coefficient (e.g., see Rodgers & Nicewander, 1988), but the similarities and differences with regression are clear if you consider that r is itself a regression slope. However, r is the slope in the bivariate regression of the 'standardized' scores of Y on X. (i.e. the regression of z_Y on z_X). Standardizing X or Y involves centering (subtracting the variable's mean) and dividing its by its standard deviation (*SD)*. This preserves the distribution of X and Y but rescales them so that both have a mean of 0 and an *SD* of 1. The resulting regression therefore has an intercept of zero and the slope r between -1 and +1.

For the earnings data the regression of z_Y on z_X gives the best-fitting line as:

$$earnings = .24 \times height$$

[1] This example is adapted from Gelman and Hill (2007, p.53), who also consider more plausible linear models.

The slope .24 is identical to that value I'd get from calculating *r*. The difference lies in the interpretation of the *r* and b_0. The latter uses the original units of analysis and if the units are meaningful is widely regarded as easier to interpret and understand (e.g., Wilkinson & APA Task Force on Statistical Inference, 1999). It is immediately obvious – even to someone with no statistical training – that a $1,256 increase in earnings per inch of height is potentially important. On the other hand, a correlation of .24 between height and earnings is trickier to interpret. Many students are taught to interpret such a correlation as a 'small' to 'moderate' effect and regard it as relatively uninteresting (because height explains only $.24^2 \approx .058$ or 5.8% of the variance). A better interpretation is that a 1 *SD* increase in height is associated with a .24 *SD* increase in earnings. Some psychologists will realize this represents a surprisingly big effect, but it is hard to put in context unless you really know the *SD* of each variable.

A common argument in favour of standardized coefficients such as *r* is they make interpreting or comparing variables on arbitrary easier. This position is questionable. I have recently argued the opposite: psychologists will generally be better off using simple, unstandardized effect size metrics (Baguley, 2009). This applies to correlation-based measures such as *r* or R^2 or standardized mean differences such as Cohen's *d*. Even with arbitrary scales, psychologists will typically be better off using the original units. For a start, even ad hoc scales (e.g., Likert-style ratings) convey some useful information about what is going on. Knowing that two groups differ on average by 2 points on a five-point scale of agreement tells you that the difference is enough to shift someone from a neutral to an extreme response. Contrast this with a standardized effect size metric such as Cohen's *d*. If *d* = 0.5 this would represent a difference of 0.5 times the *SD* of the ratings. If the *SD* is small (because ratings are generally very consistent) this could represent a small fraction of a scale point. If the *SD* is large it might represent a shift of several points. Standardized effects are not readily interpretable unless you also know (and appreciate) how large the relevant *SD*s are. Even for, say, IQ where this information is widely known it is not clear that *d* = .20 is any easier to interpret than a difference of 3 IQ points.

If that were the only problem with standardization I'd be fairly relaxed about their ubiquity in psychology. There is a deeper issue. Standardized effect sizes are calculated using the sample *SD*, but researchers nearly always (implicitly or explicitly) assume that they can be interpreted in terms of the population *SD*. This is one of the main objections to correlation coefficients. Anything that influences a sample *SD* but not the population *SD* has the potential to distort standardized effect size as measure of the population effect. As it happens, there is quite a long list of factors (including sampling error) that do exactly that. Worse still, many of these factors systematically distort the sample *SD* relative to the population *SD*. In Baguley (2009) I discuss these factors under three main headings: *reliability*, *range restriction* and *study design*. Most outcomes that psychologists are interested are unreliable. For example, Schmidt and Hunter (1999) suggest that measurement error in psychological research is frequently of the order of 50% of the total variance. In the simplest case, measurement error in an outcome measure inflates the sample *SD*, and so reduces the estimated value of *d* or *r*.

Range restriction occurs whenever the range of values in a sample differs from those in the population of interest. Consider the selection of the *X* variable in regression. If the range of *X* in the sample is restricted relative to the population, this reduces the *SD* of *X*. If *X* and *Y* are correlated (provided the correlation is not perfect) the *SD* of *Y* will also decrease, but to a lesser degree. This differential impact on the *SD* of *X* and *Y* in turn depresses *r* in the restricted sample. Range restriction also works in reverse; sampling the extremes of a population will inflate *r* in a sample (Preacher et al., 2005). The effects of range restriction can be quite extreme. Figure 1 shows the effect of selecting the middle 100 *X* values on the *X*-*Y* correlation. (Here, X and Y are sampled from normally distributed variables with a population correlation of .80). In the full sample of 500 simulated participants the correlation is .82, while the correlation in the restricted sample is only .19.[2]

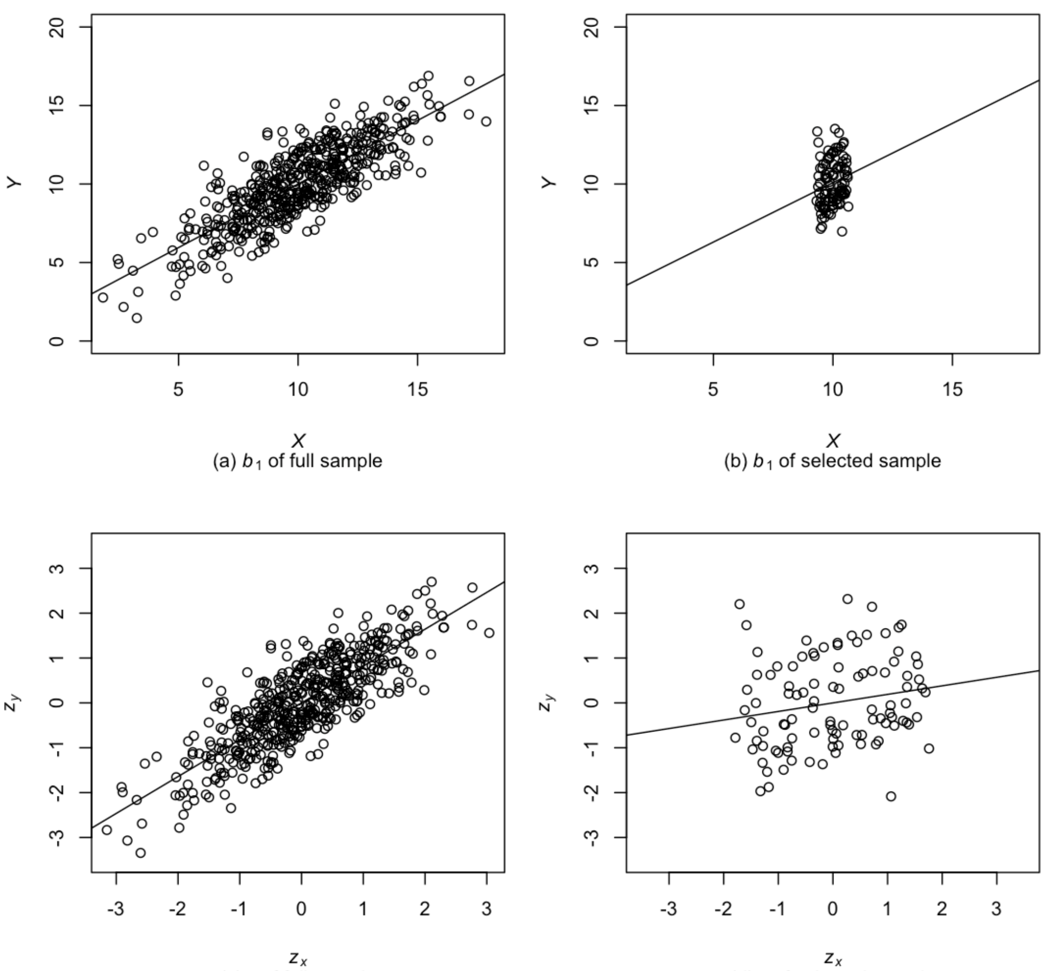
In contrast, for simple situations such as these neither range restriction nor reliability will bias simple effect size metrics such as the unstandardized regression slope b_1.[3] Although increased sampling error or reduced sample size makes estimation more 'noisy' a statistic such as b_1 is not directly influenced by the *SD* of *X* or *Y*.

---

[2] If the simulation were repeated these numbers would change but the general pattern would be similar.

[3] In more complicated situations (e.g., involving unreliability in *X* as well as *Y* or correlated predictors) even unstandardized effect size estimates can be distorted (e.g., Ree & Caretta, 2006; Hunter & Schmidt, 2004).

The effect of range restriction of X on b_1 and r



(a) b_1 of full sample

(b) b_1 of selected sample

(c) r of full sample

(d) r of selected sample

Aspects of the design of the study can also make it very difficult to compare standardized effects between otherwise very similar studies. These factors include whether independent or repeated measures are used, the choice of stimuli and the characteristics of the samples (Baguley, 2009). In some cases it is possible to work round these problems by computing a standardized effect size statistic in a particular way or by employing corrections for reliability and range restriction. Hunter and Schmidt (2004) consider many of these corrections in the context of meta-analysis. But in a single study with low *n* the information needed to make these corrections may be unavailable, of insufficient quality or the corrections themselves may be too difficult to implement. Furthermore, these corrections or fixes will often be unnecessary in uncomplicated studies (particularly experiments).

There is sometimes a case to be made for standardizing variables – perhaps as an initial default in an overall modeling strategy (Gelman, 2008) or in psychometrics where correlation coefficients are convenient ways to measure reliability and validity (Baguley, 2009). However, the widely held belief that standardization necessarily places variables or effects on a common scale is false. Careless and routine application of standardization in psychology (without any awareness of the potential pitfalls) is dangerous. In relation to the *Society for the Suppression of the Correlation Coefficient*, Brillinger may be correct to conclude:

> "It is probably more needed now than it was back in the 1940's. Perhaps someone will start a website."
>
> [Brillinger, 2001, p.216]

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology. 100*, 603-617.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods, 10*, 178-192.

Brillinger, D. R. (2001). John Tukey and the correlation coefficient, *Computing Science and Statistics, 33*. 204-218.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: correcting error and bias in research findings*. (2nd Ed.). Thousand Oaks, CA: sage.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*, 2865-2873.

Ree, M. J., & Carretta, T. R. (2006). The role of measurement error in familiar statistics. *Organizational Research Methods, 9*, 99-112.

Rodgers, J. L., & Nicewander, W. L. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*, 59–66.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*, 183-198.

Tukey, J. W. (1954). Causation, regression and path analysis. In O. Kempthorne, T. A. Bancroft, J. W. Gowen & J. L. Lush (Eds.), *Statistics and mathematics in biology* (pp. 35-66). Ames, IA: Iowa State College Press.

Wilkinson, L., & APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.