December 2003: Don't Summarize Regression Sampling Schemes with Correlation (Rule 3.2)

Rules of the month are numbered in accordance with the numbering in the book. Thus, Rule 1.1 refers to the first rule in Chapter 1. And so on. These comments do not repeat the material in the book but highlights and amplifies it. A rule is stated as found in the book and then discussed.

"Do not summarize regression sampling schemes with correlations" (Rule 3.2)

Further Comments on the Rule

Correlation is a symmetric measure of covariation. To use it in asymmetric situations is precarious. Equation 3.7 in the text developed the relationship between a correlation estimated from random sampling from a population and a correlation obtained by regression. For example, in the first situation, randomly sample subjects from a population, measure their height and weight and correlate. In the second situation, specify a set of heights, randomly sample from the subpopulations with those heights, measure their weights and correlate. The relationship between the correlations calculated by the two sampling schemes is given by Equation 3.7 in the text—modified slightly here to improve the notation. Let X be the predictor variable and

$\sigma^2_{x,true}$ = the actual variance of the predictor variable, X, in the population,

$s^2_{x,regression}$ = the variance of the predictor variable, X, in the regression situation,

ρ^2_{true} = the true correlation between predictor and dependent variable,

and

$r^2_{regression}$ = the regression coefficient estimated from the data.

For purposes of this discussion assume that the sample sizes are large so that good precision can be obtained. Then the following relationship is derived in the text,

$$\frac{r^2_{\text{regression}}}{1-r^2_{\text{regression}}} = \frac{\rho^2_{\text{true}}}{1-\rho^2_{\text{true}}} \frac{s^2_{\text{x,regression}}}{\sigma^2_{\text{x,true}}}. \quad (1)$$

Only when the variance in the predictor variable is equal to the variance of the predictor in the population is there a valid estimate of the population correlation. If the range of the predictor is oversampled (as if often deliberately done in regression situations in order to improve precision of the estimate of the slope), the sample variance will be larger than the true variability and the sample correlation will tend to be larger than the true correlation. Table 1 gives some idea of the effect.

Table 1. Effect of sampling on the correlation estimated from regression, $r^2_{\text{regression}}$.

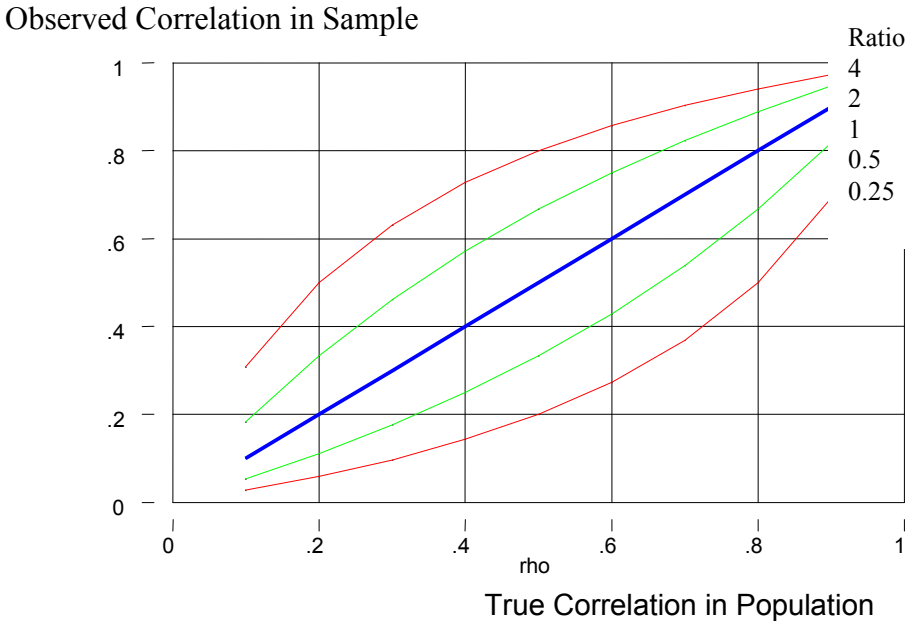
$\frac{s^2_{\text{x,regression}}}{\sigma^2_{\text{x,true}}}$	True correlation^2, ρ^2_{true}					
	0	0.10	0.20	0.30	0.40	0.50
0.25	0	0.03	0.06	0.10	0.15	0.20
0.50	0	0.05	0.11	0.18	0.25	0.33
1.00	0	0.10	0.20	0.30	0.40	0.50
2.00	0	0.18	0.33	0.46	0.57	0.67
4.00	0	0.31	0.50	0.63	0.73	0.80

Table 1 shows that only under two conditions is the population correlation estimated correctly in a regression sampling scheme. First, when there is no correlation. Second, when the variance of the predictor variable in the sample mimics the variance of the predictor in the population. In all other cases there is bias. There can be substantial misjudgment. For example, suppose that regression sampling was done in such a way that the variance of the predictor variable is four times that of the true variability. When the true correlation squared is 0.20, the estimate from the sampling is 0.50. Thus the claim that the predictor variable "explains 50% of the variability in the dependent variable" is off by 250%! An investigator basing sample size calculations for a future study assuming random sampling will be sorely disappointed if the value of 0.50 is used.

If the population variance of the predictor variable is known, then the correlation estimated from the regression situation can be adjusted to give an unbiased estimate of the population correlation.

Figure 1 displays the effect graphically. For example If the correlation in the population is 0.8 the correlation in the sample is estimated to be about 0.5 when the sample variance of the predictor variable is one fourth that of the population variance.

Figure 1. Observed correlation in sample as function of ratio of sample variance in predictor variable to true variance of the predictor variable in the population. Ratio = $s^2_{x,\text{regression}} / \sigma^2_{x,true}$. Ratios of 4, 2 ,1, 0.50 and 0.25.



Equation (1) can be written in the logit scale,

$$\text{logit}(r^2_{\text{regression}}) = \text{logit}(\rho^2_{\text{true}}) + \ln\left(\frac{s^2_{\text{regression}}}{\sigma^2_{\text{true}}}\right). \qquad (2)$$

This equation indicates that on the logit scale the 45° line is shifted up or down by a quantity that depends on the logarithm of the ratio of the variances (or the ratio of the standard deviations since the exponents cancel out). Equation (2) also shows that the adjustment is symmetrical about the ratio of the sample and population variances.

There are two bottom lines. First, you must know how the sampling was done in a situation of covariation. Perhaps the default rule is that there was selection of values of one of the variables (regression sampling). Second, either disregard the claim of "percent variability explained" or try to determine how the variability of the predictor variable in the sample corresponds to that in the population from which the sample came.