Neuendorf

Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences
Second Edition

Jacob Cohen
New York University

Patricia Cohen
New York State Psychiatric Institute
and
Columbia University School of Public Health

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1983 Hillsdale, New Jersey London

3.6.2 Shrunken $\tilde{R}^2$ (Adjusted R^2)

Although we may determine from a sample R^2 that the population R^2 is not likely to be zero, it is nevertheless not true that the sample R^2 is a good estimate of the population R^2. To gain an intuitive understanding of part of the reason for this, imagine the case in which one or more of the IVs account for *no* Y variance in the population, that is, $r^2_{Yi} = 0$ in the population for one or more X_i. Because of random sampling fluctuations we would expect that only very rarely would its r^2 with Y in a sample be *exactly* zero; it will virtually always have some positive value. Thus, in most samples it would make some (possibly trivial) contribution to R^2. The smaller the sample size the larger these positive variations from zero will be, on the average, and thus the greater the inflation of the sample R^2. Similarly, the more IVs we have, the more opportunity for the sample R^2 to be larger than the true population R^2. It is often desirable to have an estimate of the population R^2 and we naturally prefer one that is more accurate than the positively biased sample R^2. Such a realistic estimate of the population R^2 (for the fixed model) is given by

$$\tilde{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}. \tag{3.6.4}$$

This estimate is necessarily (and appropriately) smaller than the sample R^2 and is thus often referred to as the "shrunken" $\tilde{R}^2$. The magnitude of the "shrinkage" will be larger for small values of R^2 than for larger values, other things being equal. Shrinkage will also be larger as the ratio of the number of IVs to the number of subjects increases. As an example, consider the shrinkage in R^2 when $n = 200$ and cases where $k = 5$, 10, and 20 IVs, thus yielding k/n ratios of 1/40, 1/20, and 1/10, respectively. When $R^2 = .20$, the $\tilde{R}^2$ values will equal, respectively, .1794, .1577, and .1106, representing shrinkage of approximately .02, .04, and .09, the last being a shrinkage of almost one-half. When $R^2 = .40$, the comparable $\tilde{R}^2$ values are, respectively, .3845, .3683, and .3330, smaller shrinkage either as differences from or proportions of R^2. For large ratios of k/n and small R^2, these $\tilde{R}^2$ may take on negative values; for example, for $R^2 = .10$, $k = 11$, $n = 100$, Eq. (3.6.4) gives $\tilde{R}^2 = -.0125$. Whenever the F value for R^2 of Eq. (3.6.1) is less than one, a negative $\tilde{R}^2$ will occur. In such cases, by convention the shrunken $\tilde{R}^2$ is reported as zero.

It may be of interest to examine the expected value of R^2 for a sample from a population in which all correlations between the dependent and the independent variables are zero. This expected value can be shown to be

$$R^2_E = \frac{k}{n-1}, \tag{3.6.5}$$

a value almost identical with the k/n ratio that signals the degree of shrinkage to be expected. Thus, in the example of the preceding paragraph, $k/n = .11$, $R_E{}^2 = .111$, so that it is hardly surprising that the observed sample R^2 of .10 shrinks to a negative $\tilde{R}^2$. (Note that when R^2_E is substituted for R^2 in the equation for $\tilde{R}^2$, the resulting value is, as it logically should be, zero.)

It should be clear from this discussion that whenever a subset of IVs has been selected post hoc from a larger set of potential variables on the basis of their relationships with Y, either by the computer performing stepwise regression (see Section 3.8.2), or by the experimenter selecting IVs because of their relatively large r_{Yi}s, not only R^2, but also $\tilde{R}^2$ computed by taking as k the number of IVs *selected*, will be too large because of the inherent tendency of such a procedure to capitalize on chance. A more realistic estimate of shrinkage is obtained by substituting for k in Eq. (3.6.4) the *total* number of IVs from which the selection was made (see footnote 11).