Mid-Western Educational Researcher

Volume 9
Issue 4 *Regression Analyses*

Article 8

1996

Estimating the True Accuracy of Regression Predictions

Richard B. Darlington
Cornell University

Estimating the True Accuracy of Regression Predictions

Richard B. Darlington, Cornell University

Abstract

Given the lack of mathematical proof to decide upon the best estimation technique, the author presents his comparison of four closed-formula estimators (Burkett, Claudy, Rozeboom, Browne) and the omit-one method for estimating TRS, the true shrunken correlation (not to be confused with TR, the true multiple correlation). The recommendations are based on artificial populations with known TRS.

We must distinguish between two concepts that are usually confused: the true multiple correlation that I shall denote *TR*, and the true shrunken correlation that I shall denote *TRS*. Both of these differ from the observed multiple correlation *R*, which is simply the correlation in the present sample between the true criterion or dependent variable *Y*, and the estimates of *Y* made from the regression.

TR answers this question: If we derived the same regression in the total infinite population, thereby finding the true regression slopes, what value would we observe for *R*?

TRS answers a different question: Given that we have derived the regression in the sample, and have presumably not found the true population regression weights, what correlation would we find between *Y* and $\hat{Y}$ if we were to apply this set of regression weights to the entire population?

The first question asks in effect how good these *variables* are at predicting *Y*, while the second asks how good these *weights* are. When we ask about the variables, we pretend we could find the true population weights. But when we ask about the weights, we are asking about the weights we have already found.

The question about variables (involving *TR*) is usually of most interest in questions involving cause and effect, e.g., How important is education in determining income or attitude toward abortion? The question about weights (involving *TRS*) is usually of most interest in practical prediction problems. If we derive a regression in one sample, and use that regression to estimate the future performance of students or workers not in the original sample, what will be the correlation between our estimates and their actual performance?

TRS is always below *TR*, because in estimating *TR* we are asking what *R* would be if we found the true population weights. But by definition those weights are the best weights for the population, and are thus almost certainly better than the weights we have found in one particular sample. *TRS* is asking how well those sample weights would work, and they almost certainly would not work as well as the true weights. Therefore our estimate of *TRS* is always somewhat below our estimate of *TR*.

The standard formula for estimating *TR* is

$$\textit{Adjusted } R^2 = ARS = R^2 - \frac{P(1-R^2)}{N-P-1}$$

This formula is used in nearly every standard regression program. Note that N is sample size, and P is number of predictors.

The irrelevance of collinearity

At first it seems obvious that *TR* and *TRS* would fall further below *R* when collinearity is high than when it is low. After all, collinearity increases the errors with which individual regression slopes are estimated, and these errors are what cause *TR* and *TRS* to be below the observed multiple correlation *R*. Therefore it at first seems obvious that since collinearity increases those errors, it must increase the amount by which *TR* and *TRS* fall below *R*.

Surprisingly, however, collinearity can be ignored in estimating *TR* and *TRS*. The point that is ignored in the last paragraph is that under collinearity, errors in individual regression slopes tend to cancel each other out. This is one of the most remarkable features of regression. To explain it we will consider the simple case in which the regression has only two predictor variables X_1 and X_2, which we will assume are highly correlated positively. Let b_1 and b_2 denote the regression slopes of these two variables.

It turns out that the errors in b_1 and b_2 are not independent. In samples in which b_1 overestimates its true value, on the average b_2 underestimates its true value, and vice versa. But since X_1 and X_2 are highly correlated the overestimation in one slope tends to cancel out the underestimation in the other slope, with the result that on the average *Y* is estimated as accurately as if there had been no collinearity. The greater the collinearity, the more errors will cancel each other out.

The result is that even though individual slopes tend to be estimated less accurately under collinearity, *Y* and *R*

are estimated no less accurately. Thus the observed multiple correlation R tends to be no higher, relative to *TR* and *TRS*, under collinearity than when the predictor variables are mutually independent. There are rigorous mathematical proofs of this claim, in books like Draper and Smith (1981) and Graybill (1961).

Four Closed-Formula Estimators of TRS

There is no clear agreement about the best way to estimate TRS. The remainder of this article describes and evaluates five estimates of TRS. Three of the formulas are quite simple. One by Burket (1964) is

$$\textit{Burket estimate} = \sqrt{\frac{NR^2 - P}{R(N-P)}}$$

An estimator by Claudy (1978) is

$$\textit{Claudy estimate} = 2\sqrt{ARS} - R$$

An estimator by Rozeboom (1978) is

$$\textit{Rozeboom estimate} = \sqrt{1 - \frac{N+P}{N-P}(1-R^2)}$$

A slightly more complicated procedure by Browne (1975) requires the user to first compute an estimate of the fourth power of TR, by the formula

$$Rho4 = ARS^2 - \frac{2P(1-ARS)^2}{(N-1)(N-P+1)}$$

If this formula yields a negative value for Rho^4, then set $Rho^4 = 0$. Then TRS is estimated by

$$\textit{Browne estimate} = \sqrt{\frac{(N-P-3)Rho4 + ARS}{(N-2P-2)ARS + P}}$$

When these four estimates are plotted against the observed multiple correlation *R* for fixed values of *N* and *P*, they all approach *R* as *R* approaches 1. Thus the right end of each curve approaches a straight line with a slope of 1. As *R* declines, all four curves gradually get steeper until they hit the horizontal axis. It is unreasonable to estimate a negative value of *TRS*, so all four estimates are taken to be 0 if the above formulas yield negative estimates of *TRS* or TRS^2. When the estimators are ranked from most liberal to most conservative, they generally fall in the order: Burket (most liberal), Browne, Claudy, Rozeboom.

The Omit-one Method for Estimating TRS

All the previous formulas assume multivariate normality. We now describe an alternative approach that dispenses with this requirement. I call it the *omit-one* approach. Its computation is considerably more complex than for any of the previous approaches.

Imagine omitting one case from a sample of *N* cases, fitting the regression in the remaining sample of N-1 cases, and then using that regression to estimate *Y* for the one case that was omitted. Let the difference between the actual and estimated Y-scores for that one case be denoted *DCR*, for "deleted-case residual". Imagine computing *DCR* for every case in the sample, by running the regression *N* times, each time with one case omitted. If you then use the *N* values of *DCR* to estimate the accuracy of the regression predictions, you are using the "omit-one" approach.

Surprisingly, it turns out that one can compute the *N* values of *DCR* without actually computing the *N* omit-one regressions. Nearly every standard regression program computes residuals, and a great many programs will compute for each case a value that is called either LEVERAGE or *H*. This value measures the "atypicalness" of a case's scores on the predictor variables; a case whose predictor scores all fall exactly at the means has the lowest possible value of LEVERAGE. But for our purpose here, the important fact about LEVERAGE is that it can be used to compute *DCR* via the formula

DCR = RESIDUAL/(1 - LEVERAGE)

Then *Y* - *DCR* gives the estimates of *Y* computed by the omit-one regressions, but without the work of actually repeating the regression *N* times.

At first it would seem that simply correlating these values of *Y* - *DCR* with the actual *Y* values would give a good estimate of *TRS*. However, it turns out that this approach actually gives an overly conservative estimate of *TRS*. The reason is that if one case has an exceptionally high value of *Y*, then omitting it will lower the *Y*-mean of the remaining sample, and will thus lower the estimate of *Y* for that one case. This will tend to lower the correlation just mentioned. But errors in estimating means do not lower the true value of *TRS* at all, so we want to somehow correct for this lowering. You can do this with the help of the formula:

Mean of remaining sample = *M*N/(N-1) - Y/(N-1)*

where the "remaining sample" is the sample after the deletion of the one case, and *M* is the mean of the total sample. This formula shows that by adding *Y/(N-1)* to each of the (*N*-1) other scores, we would change their mean to *M*N/(N-1)*. That would fix the problem, since that value is independent of *Y*. But adding *Y*/(*N*-1) to the (*N*-1) other scores changes the residual of the deleted case *Y* by that same amount. Thus we can instead adjust the residual by that amount. But we use the residual simply to compute the predicted value of *Y*, so we can instead adjust that value by the same amount. The "bottom line" of this reasoning is that we can compute an "adjusted deleted-case prediction" *ADCP* for each case from the formula

ADCP = *Y* - RESIDUAL/(1-LEVERAGE) + *Y*/(*N*-1)

Correlating these *ADCP* values with the original *Y* values then gives an estimate of *TRS*.

Comparison of the five estimators

Sometimes an estimator can be proven mathematically to be the best possible estimator. No such proof is available for any of the estimators of TRS, so I have compared them numerically. I used a 7 x 6 x 16 array of values of the sample size *N*, the number of predictors *P*, and *TR* respectively. I let *N* range from 40 to 100 in increments of 10, let *P* range from 5 to 30 in increments of 5, and let *TR* range from .05 to .80 in increments of .05. For each of these 7 x 6 x 16 or 672 combinations of *N*, *P*, and *TR*, I drew 1000 samples and fitted a multiple regression in the sample. Because the true population regression slopes were known for the artificial populations I used, I could compute the exact value of *TRS* for each sample regression. I then used each of the five estimators in turn to estimate *TRS*. For each estimator I computed two statistics, *LIBCOUNT* and *RMSE*, for each of the 672 combinations. *LIBCOUNT* was defined as the number of times the estimate of *TRS* exceeded the true *TRS* for that sample, and *RMSE* was the root mean squared difference between true and estimated values of *TRS*.

In a typical problem the investigator of course knows *N* and *P*, but does not know *TR*. It seems unreasonable to average *RMSE* or other measures of accuracy across the various values of *TR*, because this assumes that these values occur with equal frequency in real problems. Since the Burket and Rozeboom formulas are respectively the most liberal and most conservative of the four closed-formula estimators, Burket tends to be best when *TR* is high, while Rozeboom tends to be best when *TR* is low. To avoid this problem I always used the *worst* value of *LIBCOUNT* or *RMSE* across the 16 values of *TR* studied for a given combination of *N* and *P*, calling these worst values *LIBMAX* and *RMSEMAX* respectively. By this means, the 672 values of *LIBCOUNT* and *RMSE* for each method are reduced to 42 values of *LIBMAX* and *RMSEMAX*.

Clearly the worst value of *RMSE* is the largest. For each estimator these largest values always occured for *TR* values between .30 and .70; they were never at the highest or lowest values of *TR* studied. I also defined *LIBMAX* as the largest of the 16 values of *LIBCOUNT*, rather than the value farthest from 500. (500 is half the number of samples used for each combination of *N* and *P*.) This seems reasonable to me because the whole purpose of estimating *TRS* is to avoid an overly liberal estimate of a regression's predictive power, and *RMSE* provides an alternate statistic that treats overestimates and underestimates equally.

It seems reasonable to consider LIBMAX values of 550 and below as acceptable. This allows for a little random error caused by the fact that only 1000 samples were used, and also allows a modest amount of nonrandom error. By this criterion all 42 values of *LIBMAX* were acceptable for Browne and for Omit-one; their highest values of *LIBMAX* were 549 and 546 respectively. For all five methods the highest *LIBMAX* value came at the highest *P* and lowest *N* studied, with $P = 30$ and $N = 40$. For Burket, *LIBMAX* was acceptable only if $P = 5$, or if $P = 10$ and $N \geq 80$; its highest *LIBMAX* value otherwise was 877. For Claudy, *LIBMAX* was acceptable only if $P = 5$, or if $P = 10$ and $N \geq 70$, or if $P=15$ and $N \geq 80$; its highest *LIBMAX* value otherwise was 759. For Rozeboom, *LIBMAX* was acceptable only if $P = 5$, or if $P = 10$ and $N \geq 60$, or if $P = 15$ and $N \geq 80$; its highest *LIBMAX* value otherwise was 730.

For each of the 42 combinations of *N* and *P*, I also identified the method with the lowest value of *RMSEMAX*. With one apparently random exception, the Burket method was always best by this criterion when $P = 5$; and with two apparently random exceptions, the Omit-one method was always best by this criterion when $P \geq 10$. When the Omit-one method is ignored as too complex, it turns out that with one exception at the margin, the Burket method is best when $N/P > 3.5$, while the Browne method is best when $N/P < 3.5$.

In summary, though Claudy and Rozeboom do give moderately good estimates, there seems to be no good reason to use those estimators, since others consistently do better by both the *LIBMAX* and *RMSEMAX* criterion. The remaining three methods can be ranked in terms of simplicity, with Burket simplest, Browne next, and Omit-one most complex. When multivariate normality can be assumed and $N/P > 8$, the simple Burket rule is quite satisfactory since its *LIBMAX* value stays below 550. When multivariate normality can be assumed and $N/P < 8$, Browne seems superior. All 42 of its *RMSEMAX* values exceeded those of Omit-one, but never by as much as 10%. But if computing power is no major obstacle then Omit-one seems the clear choice. None of its *RMSEMAX* values exceed those of Burket by more than 1.5%, and none at all exceed those of Browne. And Omit-one has the further major advantage of requiring no assumption of multivariate normality—an assumption that is quite important for the competing methods.

References

Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology*, 28, 79-87.

Burket, G. R. (1964). A study of reduced rank models for multiple prediction. *Psychometric Monographs* (No. 12).

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, 32, 311-322.

Draper, Norman & Smith, Harry. (1981) *Applied regression analysis, 2nd edition.* New York: Wiley.

Graybill, Franklin. (1961) *An introduction to linear statistical models, volume 1.* New York: McGraw-Hill.

Rozeboom, W. W. (1978). The estimation of cross-validated multiple correlation: a clarification. *Psychological Bulletin*, 85, 1349-1351.