Estimators of the Squared Cross-Validity Coefficient: A Monte Carlo Investigation

Fritz Drasgow

University of Illinois at Urbana-Champaign

Neil J. Dorans

Personnel Research and Development Center

U.S. Office of Personnel Management

Ledyard R. Tucker

University of Illinois at Urbana-Champaign

A monte carlo experiment was used to evaluate four procedures for estimating the population squared cross-validity of a sample least squares regression equation. Four levels of population squared multiple correlation (R_p^2) and three levels of number of predictors (n) were factorially crossed to produce 12 population covariance matrices. Random samples at four levels of sample size (N) were drawn from each population. The levels of N , n , and R_p^2 were carefully selected to ensure relevance of simulation results for much applied research. The least squares regression equation from each sample was applied in its respective population to obtain the actual population squared cross-validity (R_{cv}^2). Estimates of R_{cv}^2 were computed using three

formula estimators and the double cross-validation procedure. The results of the experiment demonstrate that two estimators which have previously been advocated in the literature were negatively biased and exhibited poor accuracy. The negative bias for these two estimators increased as R_p^2 decreased and as the ratio of N to n decreased. As a consequence, their biases were most evident in small samples where cross-validation is imperative. In contrast, the third estimator was quite accurate and virtually unbiased within the scope of this simulation. This third estimator is recommended for applied settings which are adequately approximated by the correlation model.

In many situations, it is very important to cross-validate a sample least squares regression equation. The need for cross-validation is particularly acute when using a regression equation obtained in one sample to predict criterion scores for individuals not in that sample. Measures of the prediction equation's precision in the derivation sample can substantially exaggerate the effectiveness of that prediction equation in other samples or in the population.

One measure of the predictive accuracy of a least squares regression equation in the derivation sample is the sample multiple correlation coefficient, R_s . Mosier (1951) pointed out that a sample multiple correlation coefficient overestimates the effectiveness of the sample least squares regression equation in predicting criterion scores for future observations. To obtain a better measure of predictive accuracy, Mosier suggested the procedure known as *double cross-validation*. Here the derivation sample is split in half and a regression equation is computed in each half sample. Then regression weights from each half sample are applied to predictor scores from the other half sample to generate predicted criterion scores. An estimate of the long-run efficiency of the sample regression equation,

R_{ps} , is obtained by correlating predicted criterion scores with the actual criterion scores in a half sample. Each half sample serves as both a derivation sample and a cross-validation sample. Hence, the double cross-validation procedure produces two estimates of long-run predictive accuracy.

One problem with double cross-validation is that each regression equation is based on only half of the data. In many instances, the original derivation sample is too small to justify its division into halves. Regression weights derived in the smaller half samples have larger standard errors of estimation and tend to yield less accurate prediction in the population than weights derived in the total sample. The tendency of least squares estimation to capitalize on chance in the derivation sample is exacerbated when the derivation sample is divided in half.

Claudy (1978), on the basis of empirical results obtained via a monte carlo experiment, recommended a method of estimating the long-run efficiency of a sample prediction equation that uses all the correlation coefficients calculated in the double cross-validation procedure. This method, termed "double shrinkage of the double cross-validity estimate" by Claudy (p. 602), combines the multiple correlation in the total sample, R_s ; the two half sample multiple correlations, $R_{s(1)}$ and $R_{s(2)}$; and the two half sample R_{ps} statistics, $R_{ps(1)}$ and $R_{ps(2)}$, via

$$\text{Est}(R_{cv}) = (R_{ps(1)} + R_{ps(2)} + R_{s(1)} + R_{s(2)})/2 - R_s \quad [1]$$

to estimate the population cross-validity coefficient, R_{cv} . The statistic in Equation 1 estimates the correlation in the *population* between criterion scores and predictions based on the sample regression coefficient. (For justification of Equation 1, see Claudy, 1978.)

As an alternative to the empirical cross-validation procedures, a single regression equation based on all available data can be computed. Then the distribution theory for least squares regression weights can be used to derive a formula estimate of R_{cv}^2 .

Herzberg (1969) presented two formula estimators of the squared cross-validity coefficient.¹ One estimator can be used to estimate the squared cross-validity coefficient when the *regression model* (predictor variables are assumed to be fixed) is appropriate. This estimator is

$$\text{Est}_1(R_{cv}^2) = 1 - \frac{N-1}{N} \frac{N+n+1}{N-n-1} (1 - R_s^2) \quad [2]$$

where R_s^2 is the squared sample multiple correlation obtained in a sample of N individuals using n predictor variables. For the *correlation model* (observations on the predictors and criterion are assumed to follow a multivariate normal distribution), the estimator presented by Herzberg is

$$\text{Est}_2(R_{cv}^2) = 1 - \frac{N-1}{N} \frac{N-2}{N-n-1} \frac{N+1}{N-n-2} (1 - R_s^2) \quad [3]$$

It has been shown that these two estimators are derived from the erroneous assumption that predictions based on a *sample* least squares regression equation are uncorrelated with their associated errors of prediction *in the population* (Drasgow, Dorans, & Tucker, 1979). Drasgow et al. demon-

¹The origin of the estimators in Equations 2 and 3 is not clear. Schmitt, Coyle, and Rauschenberger (1977) attribute Equation 2 to Lord (1950) and Equation 3 to Stein (1960). However, these estimators do *not* appear in the papers cited. To the best of the present authors' knowledge, Equations 2 and 3 were first presented by Herzberg (1969).

strate that this mistaken assumption leads to negative biases for Equations 2 and 3, which increase as the population multiple correlation decreases. It should be noted that Herzberg (1969) did not endorse these estimators, stating "It is not the purpose of this study to derive the best estimate of $[R_{cv}^2] \dots$ "

Recently, Schmitt, Coyle, and Rauschenberger (1977) investigated the estimators in Equations 2 and 3 via monte carlo methods. They found that the estimators worked quite well. The multiple correlations in the populations investigated by Schmitt et al. ranged from .634 to .965. Because the biases of Equations 2 and 3 decrease as the population multiple correlation increases, extrapolation of the results obtained by Schmitt et al. to populations characterized by low to moderate multiple correlations is not warranted.

Claudy (1978) investigated Equations 1, 2, and 3 and several other procedures, as estimators of the cross-validity coefficient R_{cv} . (This entailed taking the square roots of the estimates produced by Equations 2 and 3.) Aggregating results across 16 populations with 2 to 5 predictors and population multiple correlations ranging from .29 to .82 (Claudy, 1979), Claudy found that Equation 1 was slightly more accurate than the other estimators he studied at all but very small sample sizes ($N = 20$), where it was slightly less accurate than Equation 2. Aggregation of results across populations with widely disparate multiple correlations apparently prevented Claudy from observing the negative bias of Equations 2 and 3 in populations with small multiple correlations.

Browne (1975) and Drasgow et al. (1979) independently derived an estimator of the squared cross-validity coefficient for the correlation model,

$$\text{Est}_3(R_{cv}^2) = \frac{(N-n-3)\hat{R}_p^4 + \hat{R}_p^2}{(N-2n-2)\hat{R}_p^2 + n} \quad [4]$$

where $\hat{R}_p^2$ is an estimate of the squared multiple correlation in the population, R_p^2 , and $\hat{R}_p^4$ is an estimate of the population multiple correlation raised to the fourth power. The familiar Wherry (1931) formula

$$\hat{R}_p^2 = 1 - \frac{N-1}{N-n-1} (1 - R_s^2) \quad [5]$$

may be used to estimate R_p^2 . Browne suggested estimating R_p^4 by

$$\hat{R}_p^4 = (\hat{R}_p^2)^2 - \frac{2n(1 - \hat{R}_p^2)^2}{(N-1)(N-n+1)} \quad [6]$$

and replacing negative estimates of R_p^2 and R_p^4 by zero.

Employing monte carlo methods, Browne found that Equation 4 had a slight positive bias when R_p^2 was .10 and a slight negative bias when R_p^2 was .90. The positive bias was probably due to replacing negative estimates of R_p^2 and R_p^4 with zeros. As Browne noted, this procedure increases the bias in estimation but should reduce mean squared error of estimation.

In the present study, the characteristics of the estimators in Equations 2, 3, and 4 were investigated in populations with squared multiple correlations that ranged from .00 to .45. This research serves several purposes. First, it is important to examine the biases that result from applying Equations 2 and 3 under circumstances that resemble data sets seen in practice. Researchers in some applied areas rarely see a cross-validity greater than .40. The theoretical work by Drasgow et al. indi-

cates that the negative biases of Equations 2 and 3 may be substantial in populations with small multiple correlations. Second, it is interesting to compare Equations 2 and 3 to Equation 4. Bias and mean squared error of estimation provided two criteria for evaluating the relative merits of the estimators. A third purpose of the present research was to provide more information about the estimator in Equation 4. Browne investigated bias and mean squared error of estimation at five levels of R_p^2 . However, Browne's research only studied three combinations of N and n : (1) $N = 20$ and $n = 10$; (2) $N = 60$ and $n = 10$; and (3) $N = 60$ and $n = 50$. In the present study, three values of n were factorially crossed with four levels of N . The values of n and N were carefully chosen so that the results would have relevance for much applied research. Finally, negative estimates of R_p^2 and R_p^4 were *not* replaced with zeros in order to ensure a fairer comparison of Equations 2 and 3 to Equation 4 with respect to both bias and mean squared error. In practice, negative estimates of R_p^2 and R_p^4 should be regarded as indicants of a negligible relationship in the population between the criterion and the predictor variables.

Method

Three facets were manipulated in the present study. The first facet, squared multiple correlation in the population, had four levels: R_p^2 was set equal to .00, .15, .30, and .45. The number of predictors was the second facet: Predictor batteries of $n = 4, 8$, and 12 were simulated. Finally, four sample sizes were examined: Samples of 30, 60, 120, and 240 observations were generated. These three facets were factorially crossed to yield 48 cells in the design. One hundred replications were obtained for each cell.

The present research simulated a situation where the correlation model is appropriate. Here the criterion and predictor variables follow a $(n + 1)$ dimensional multivariate normal distribution.

The sample multiple correlation and the least squares weights depend only upon the sample covariance matrix for the predictors and the criterion. Consequently, it was not necessary to generate raw data. Instead, sample covariance matrices following a Wishart distribution were generated directly. This was possible because variables following a multivariate normal distribution yield sample covariance matrices that have a Wishart distribution (Anderson, 1958). Herzberg (1969) and Schmidt (1971) provide a complete description of a procedure that can be used to obtain sample covariance matrices that follow a Wishart distribution.

The focus of this research was upon the population squared cross-validity coefficient, R_{cv}^2 . Interestingly, for given values of R_p^2 , N , and n , the distribution of the squared cross-validity coefficient does *not* depend upon the particular form of the population covariance matrix (Browne, 1975; Drasgow et al., 1979).² As a result, without loss of generality, all population covariance matrices were constructed by the following procedure. First, each covariance matrix for the n predictor variables was set equal to an identity matrix. Then, the covariance between the first predictor variable and the criterion was set equal to R_p . All other predictors had zero covariance with the criterion. Finally, the criterion was specified to have unit variance.

In summary, four levels of R_p^2 and three levels of n were factorially crossed to produce 12 population covariance matrices. One hundred sample covariance matrices were generated for each of four sample sizes (30, 60, 120, and 240) by the Bartlett decomposition of the Wishart distribution. Sample least squares regression weights were computed and then applied in the population to obtain R_{cv}^2 . Equations 2, 3, and 4 were then utilized to estimate R_{cv}^2 . Finally, pairs of cross-validity estimates were

²See Appendix 1 for a discussion of the transformational invariances that justify this statement.

obtained by applying Mosier's (1951) double cross-validation procedure to pairs of samples, such that each sample regression equation was cross-validated in a second sample of equal size.

Results

Bias

The bias of an estimator can be assessed by comparing the mean estimated squared cross-validity to the mean of the actual squared cross-validities of sample regression equations in the population.³ Table 1 presents the mean differences between estimated and actual squared cross-validity at each level of sample size for the various levels of n and R_p^2 . The entries in the first three columns of this table were obtained by averaging over replications and the four levels of R_p^2 . The last four columns contain entries obtained by averaging over replications and the three levels of n . Hence, each mean difference in the first three columns of Table 1 is based on 400 samples, while each mean difference in the last four columns is based on 300 samples.

In Table 1 it is clear that Equations 2 and 3 have negative biases. As the work by Drasgow et al. (1979) suggests, the biases for these two estimators increased as R_p^2 decreased and/or the ratio of N to n decreased. It is important to note that the biases of Equations 2 and 3 were largest in small samples; small samples represent the situations where it is most important to cross-validate. Consequently, Equations 2 and 3 exhibited their worst performance under the conditions where the researcher would desire best performance.

The estimator in Equation 4 shows very little bias in Table 1. Across all the conditions in this simulation, this estimator may have had a very slight negative bias. However, the magnitude of the possible bias of Equation 4 was small enough that it does not pose a serious problem for the practitioner.

The double cross-validation procedure produced relatively unbiased estimates of R_{cv}^2 . However, there did appear to be a very slight positive bias. Since no theoretical reason for this effect is evident, the bias seen here may be due to sampling fluctuations.

Root Mean Squared Error of Estimation

A second criterion for evaluating an estimator is the root mean squared error of the estimates of R_{cv}^2 over the 100 replications,

$$RMSE = \left\{ \frac{1}{100} \sum_{i=1}^{100} [R_{cv_i}^2 - \text{Est}(R_{cv_i}^2)]^2 \right\}^{1/2} \quad [7]$$

Table 2 presents average root mean squared errors. The averaging procedure used to obtain the entries in this table was identical to that used for Table 1. Again, Equations 2 and 3 were very inaccurate in small samples. As expected, this effect was particularly acute when R_p^2 was small and when the ratio of N to n was small. As sample size increased and R_p^2 increased, Equations 2 and 3 became as effective as Equation 4. Even in samples of 240 observations, however, Equation 4 was superior when R_p^2 was .15 or less.

³Because a cross-validity coefficient can be negative, a squared cross-validity was given the sign of the cross-validity coefficient. For example, if $R_{cv} = -.1$, then $R_{cv}^2 = -.01$.

Table 1
Mean Differences Between Estimated and True Squared
Cross-Validity for the Four Types of Estimators

Estimator	Number of Predictors			Squared Pop. Mult. Cor.			
	4	8	12	.00	.15	.30	.45
N = 30							
Eq. 2	-.095	-.168	-.215	-.297	-.180	-.089	-.072
Eq. 3	-.126	-.285	-.507	-.484	-.339	-.221	-.180
Eq. 4	-.024	-.015	.002	-.017	-.006	.005	-.028
R_{ps}^2	.005	.010	.014	.002	.018	.021	-.002
N = 60							
Eq. 2	-.041	-.059	-.092	-.151	-.064	-.033	-.009
Eq. 3	-.048	-.081	-.142	-.185	-.094	-.056	-.027
Eq. 4	-.007	.005	.002	-.010	.002	.001	.007
R_{ps}^2	.009	.011	.009	.000	.010	.015	.015
N = 120							
Eq. 2	-.011	-.036	-.048	-.075	-.030	-.011	-.011
Eq. 3	-.013	-.041	-.059	-.082	-.037	-.017	-.016
Eq. 4	.005	-.007	-.005	-.005	-.005	.003	-.002
R_{ps}^2	.010	-.001	.004	.000	.004	.010	.003
N = 240							
Eq. 2	-.009	-.012	-.023	-.037	-.012	-.007	-.004
Eq. 3	-.010	-.013	-.026	-.039	-.014	-.008	-.005
Eq. 4	-.001	.001	-.003	-.002	-.002	.000	.001
R_{ps}^2	.002	.003	-.001	.000	.001	.005	.001

Note. The entries in Table 1 were computed as means of the differences $\text{Est}(R_{cv}^2) - R_{cv}^2$.

Comparison of Estimators

Although the average RMSE for R_{ps}^2 was usually smaller than the corresponding average for Equation 4, this does *not* mean double cross-validation is more effective. There are two problems that preclude a simple interpretation of the low bias and RMSE seen for R_{ps}^2 . Both arise because samples of $2N$ observations are actually required to obtain the results listed for N observations when double cross-validation is utilized. First, note that an average RMSE of .056 was obtained for R_{ps}^2 when $N = 30$ and $R_p^2 = .00$ and that a total of 60 observations was required to form two subsamples of 30. Only then could the squared cross-validity estimates be obtained by applying regression weights computed in one subsample to the other subsample. Hence, it may be more appropriate to compare results obtained for Equations 2, 3, and 4 given $2N$ observations to results for R_{ps}^2 given N observations. Thus, to compare Equation 4 to R_{ps}^2 for the example just mentioned, an average RMSE of .043 for Equation 4 would be compared to .056 for R_{ps}^2 .

Table 2
Average Root Mean Squared Error of
Estimate for the Four Estimators

Estimator	Number of Predictors			Squared Pop. Mult. Cor.			
	4	8	12	.00	.15	.30	.45
N = 30							
Eq. 2	.186	.282	.365	.359	.293	.237	.223
Eq. 3	.208	.378	.624	.538	.435	.341	.301
Eq. 4	.130	.130	.131	.073	.123	.153	.173
R_{ps}^2	.105	.100	.089	.056	.084	.122	.130
N = 60							
Eq. 2	.105	.133	.167	.173	.135	.121	.111
Eq. 3	.109	.147	.202	.205	.156	.134	.117
Eq. 4	.083	.086	.084	.043	.062	.102	.103
R_{ps}^2	.076	.074	.065	.028	.070	.092	.097
N = 120							
Eq. 2	.065	.078	.092	.083	.073	.080	.077
Eq. 3	.065	.081	.099	.090	.077	.082	.078
Eq. 4	.058	.057	.058	.022	.058	.076	.074
R_{ps}^2	.053	.050	.048	.013	.055	.066	.068
N = 240							
Eq. 2	.045	.047	.052	.041	.047	.056	.048
Eq. 3	.045	.048	.053	.043	.048	.056	.049
Eq. 4	.041	.040	.038	.011	.044	.055	.048
R_{ps}^2	.039	.037	.034	.007	.040	.052	.048

The second difficulty in interpreting the accuracy of R_{ps}^2 results because the cross-validity coefficients for regression equations based on N observations tended to be smaller than cross-validities for equations derived in samples of $2N$. Hence, even if the bias and average RMSE of R_{ps}^2 for N observations were smaller than the bias and average RMSE for Equation 4 given $2N$ observations, Equation 4 may provide a superior estimate of the squared cross-validity of the regression equation based on $2N$ observations. All this means that R_{ps}^2 is *not* estimating the squared cross-validity of primary interest, i.e., the squared cross-validity coefficient of the prediction equation derived from *all* available data. The quantity R_{ps}^2 estimates the squared cross-validity of prediction equations based on *half* of the data available.

To provide a clearer understanding of the estimators under consideration, detailed results for four populations are shown in Figure 1. Each point in this figure is a mean based on 100 replications.

Figure 1a presents means for a population with $R_p^2 = .15$ and four predictor variables. Across all sample sizes, both Equation 4 and R_{ps}^2 showed little bias. In contrast, both Equations 2 and 3 exhibited substantial negative bias at $N = 30$ and $N = 60$. Even in samples of 120 and 240 observations, these estimators had small negative biases.

Figure 1

Mean squared cross-validities and mean estimates of the squared cross-validity coefficient for 100 replications at each sample size.

Figure 1b presents the results for a population with four predictors and $R_p^2 = .45$. Although both Equations 2 and 3 still exhibited bias in small samples, they appeared to be virtually unbiased for samples of 120 and 240. These results are congruent with the findings of Schmitt et al. (1977). Note that Equation 4 showed less bias than Equations 2 and 3 in small samples.

The results for a population with 12 predictors and a population squared multiple correlation of .15 are shown in Figure 1c. Clearly, Equations 2 and 3 had very large negative biases in small samples. These biases occurred because the ratio of N to n was small as was R_p^2 . In Figure 1a, both Equations 2 and 3 exhibit less bias because the ratio of N to n was larger.

In contrast to Equations 2 and 3, Equation 4 did not appear to have a bias that depends upon R_p^2 and the ratio of N to n . In Figure 1c, Equation 4 shows little bias. For samples of size 30, the mean bias of Equation 4 was .023. Since the standard error of the mean for Equation 4 was .013, this bias should not be given undue attention.

A population with $R_p^2 = .45$ and 12 predictors is presented in Figure 1d. Again, both Equations 2 and 3 exhibited substantial negative bias in small samples. However, both these estimators show less bias in Figure 1d than in Figure 1c because R_p^2 was higher. The means for Equation 4 were close to mean values of R_{cv}^2 for all sample sizes.

The fundamental weakness of the double cross-validation procedure can be seen in Figure 1d. Note that the mean of R_{ps}^2 was fairly close to the mean of R_{cv}^2 for samples of both 30 and 60 observations. At $N = 30$, the mean of R_{ps}^2 was .260 and the mean of R_{cv}^2 was .246. However, R_{ps}^2 estimates of the squared cross-validity at $N = 30$ actually required 60 observations (30 in the derivation sample and 30 in the cross-validation sample). Thus this estimator of squared cross-validity utilized 60 observations and yielded a mean estimate of .260. Note that the mean squared cross-validity of regression equations based on *all* 60 observations is .353. Consequently, although R_{ps}^2 may be an unbiased estimator of one quantity (the squared cross-validity coefficient for a regression equation based on half of the available data), R_{ps}^2 is a *biased* estimator of the quantity of primary interest: the squared cross-validity coefficient for regression equations based on *all* available data.

Discussion

The general conclusions about Equations 2 and 3 drawn in this study differ from the conclusions drawn by Schmitt et al. (1977). Whereas Schmitt et al. found Equations 2 and 3 to be satisfactory, in the present study these estimators were found to be unacceptable. There is a simple explanation for these results: Equations 2 and 3 have biases that decrease as R_p^2 increases and/or the ratio of N to n increases. Schmitt et al. examined Equations 2 and 3 in populations with moderate to large values of R_p^2 . Further, Schmitt et al. did *not* factorially cross R_p^2 with n . Instead, they employed a design in which n was correlated with R_p^2 . Consequently, the smallest ratios of N to n occurred in populations with the highest values of R_p^2 . Thus, dependence of the biases of Equations 2 and 3 upon the ratio of N to n was not obvious in the research performed by Schmitt et al. due to their experimental design.

The performance of Equations 2, 3, and 4 was also examined in three populations with squared multiple correlations of .90. Test batteries with 4, 8, and 12 predictors were factorially crossed with sample sizes of 30, 60, 120, and 240 observations. The results obtained in these populations confirm the findings of Schmitt et al. In populations with high multiple correlations, Equations 2 and 3 were quite effective. In addition, Equation 4 was very accurate. In fact, the differences between Equations 2, 3, and 4 in terms of bias and RMSE were so small that it probably does not matter which estimator is used.

As noted above, Claudy's (1978) decision to analyze results aggregated over populations with widely disparate multiple correlations prevented him from observing the negative biases of the estimators in Equations 2 and 3. This aggregation complicates comparison of the present findings to the results obtained by Claudy, because the accuracies of Equations 2 and 3 depend on the magnitude of the population multiple correlation. Meaningful comparisons are further impeded by the fact that Claudy examined estimators of the cross-validity coefficient, whereas estimators of the *squared* cross-validity coefficient were examined in the present study. As Claudy notes, the square root of the best estimator of a parameter (the squared cross-validity coefficient) may not yield the best estimate of the square root of the parameter (the cross-validity coefficient). Similarly, squaring the best estimate of R_{cv} might not produce the best estimate of R_{cv}^2 .

It should be noted that the present research examined a situation where the correlation model is appropriate: The predictors and criterion form a multivariate normal distribution. For prediction problems in psychology, the assumption that observations are randomly sampled is probably more appropriate than the assumption that the values of the predictor variables are predetermined or fixed by the researcher. However, the distribution theory for the correlation model is more complex than for the regression model. As a result, the regression model is generally employed with the hope that the differences between the two models are small in practice (Burket, 1964). This assumption is not usually examined carefully. The findings in the present study bypass this problem because the more appropriate correlation model was investigated.

Two major restrictions on the generality of the results obtained here are evident. First, violations of the normality assumption were not examined. No empirical evidence can be provided for the effectiveness of Equation 4 when the criterion or predictors follow a non-normal distribution.⁴ Further, the robustness of Equation 4 to contamination caused by occasionally sampling an observation from another population is unknown.

The second major restriction on the present research is that Equation 4 may be used only for least squares regression weights. At present, there is no formula estimator that can be used to estimate the squared cross-validity coefficient for alternative prediction schemes such as principal components regression, ridge regression, or Bayesian regression. Here the only viable procedure for cross-validation appears to be double cross-validation.

In summary, Equation 4 provides virtually unbiased estimation of the squared cross-validity coefficient within the scope of this simulation. For data sets that adequately conform to the assumptions of the correlation model, Equation 4 provides the most satisfactory estimate of the squared cross-validity coefficient for least squares regression equations.

References

- Anderson, T. W. *Multivariate statistical analysis*. New York: Wiley, 1958.
- Browne, M. W. Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology*, 1975, 28, 79–87.
- Burket, G. R. A study of reduced rank models for multiple prediction. *Psychometric Monographs*, 1964 (No. 12).
- Claudy, J. G. Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, 1978, 2, 595–607.
- Claudy, J. G. Personal communication, 1979.
- Dragow, F., Dorans, N. J., & Tucker, L. R. *Estimators of the squared cross-validity coefficient: Some theoretical considerations*. Manuscript submitted for publication, 1979.

⁴The double cross-validation procedure may be preferable when sampling from non-normal populations, since it is a distribution-free statistic.

- Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950.
- Herzberg, P. A. The parameters of cross-validation. *Psychometrika Monograph Supplement*, 1969, 34 (2, Pt. 2, No. 16).
- Lord, F. M. *Efficiency of prediction when a regression equation from one sample is used in a new sample* (RB-50-40). Princeton, NJ: Educational Testing Service, 1950.
- Mosier, C. I. Problems and designs of cross-validation. *Educational and Psychological Measurement*, 1951, 11, 5-11.
- Schmidt, F. L. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, 1971, 31, 699-714.
- Schmitt, N., Coyle, B. W., & Rauschenberger, J. A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. *Psychological Bulletin*, 1977, 84, 751-758.
- Stein, C. Multiple regression. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press, 1960.
- Wherry, R. J. A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, 1931, 2, 440-457.

Acknowledgments

Much of this research was done while the second author was studying at the University of Illinois, partially supported by a NIMH Training Grant in Measurement Psychology (MH14257-03) to the University of Illinois.

Author's Address

Send requests for reprints or further information to Fritz Drasgow, Yale School of Organization and Management, Box 1A, New Haven, CT 06520.

Appendix 1

In this appendix, the squared cross-validity coefficient is shown to be invariant with respect to nonsingular transformations in the predictor space. Then it is argued that the distribution of squared cross-validity coefficients is unaffected by nonsingular transformations when the predictors and criterion follow a multivariate normal distribution.

Let Σ_{xx} denote the n -by- n covariance matrix for the n predictor variables, σ_{xy} denote the n -by-1 vector of predictor-criterion covariances, and σ_y^2 denote the variance of the criterion. Furthermore, let S_{xx} and s_{xy} be the sample counterparts to Σ_{xx} and σ_{xy} , respectively. Sample least squares weights are

$$\hat{\beta} = S_{xx}^{-1} s_{xy} \quad [A.1]$$

and the squared cross-validity coefficient for these weights is

$$R_{cv}^2 = \frac{(\hat{\beta}' \sigma_{xy})^2}{\sigma_y^2 (\hat{\beta}' \Sigma_{xx} \hat{\beta})} \quad [A.2]$$

Now let an arbitrary, nonsingular transformation, T , be defined from the original predictors, X , to a new set of predictors, Z :

$$Z = TX \quad [A.3]$$

From the theory of linear transformations (Gulliksen, 1950), are obtained

$$\Sigma_{zz} = T \Sigma_{xx} T' \quad [A.4]$$

and

$$\sigma_{zy} = T\sigma_{xy} \quad [\text{A.5}]$$

Similarly, the sample counterparts for the transformed predictors are

$$S_{zz} = TS_{xx}T' \quad [\text{A.6}]$$

and

$$s_{zy} = Ts_{xy} \quad [\text{A.7}]$$

The least squares regression weights for the transformed predictors are

$$\begin{aligned} \hat{\beta} &= S_{zz}^{-1}s_{zy} = (T')^{-1}S_{xx}^{-1}T^{-1}Ts_{xy} \\ &= (T')^{-1}\hat{\beta} \end{aligned} \quad [\text{A.8}]$$

and the squared cross-validity of the sample regression weights in the transformed space is

$$\begin{aligned} \tilde{R}_{cv}^2 &= \frac{(\hat{\beta}'\sigma_{zy})^2}{\sigma_y^2(\hat{\beta}'\Sigma_{zz}\hat{\beta})} \\ &= \frac{(\hat{\beta}'T^{-1}T\sigma_{xy})^2}{\sigma_y^2[\hat{\beta}'T^{-1}T\Sigma_{xx}T'(T')^{-1}\hat{\beta}]} \\ &= R_{cv}^2 \end{aligned} \quad [\text{A.9}]$$

Equation A.9 demonstrates that the squared cross-validity coefficient of sample least squares weights for the transformed predictors is identical to the squared cross-validity of regression weights for the original predictors. Thus, a nonsingular transformation of the predictors does *not* affect the squared cross-validity coefficient.⁵

Given Σ_{xx} and σ_{xy} , the distribution of the squared cross-validity coefficient depends solely upon the distributions of sample covariance matrices S_{xx} and s_{xy} because (1) sample least squares regression weights are entirely a function of sample covariances and (2) the squared cross-validity is a function of sample least squares weights. When the predictors and criterion follow an $(n + 1)$ dimensional multivariate normal distribution, it is known that the distribution of the $(n + 1)$ -by- $(n + 1)$ sample covari-

⁵With slight modification of the derivation above, it can be shown that the population multiple correlation is also unaffected by nonsingular transformations, i.e., $\tilde{R}_p^2 = R_p^2$.

ance matrix, S , based on N observations is Wishart (Anderson, 1958) and can be denoted $W(\Sigma, N)$, where Σ is the population counterpart to S .

Since linear combinations of normally distributed variables are also normally distributed, the transformed variables in Equations A.3 and the criterion follow a multivariate normal distribution. Consequently, covariance matrices sampled from the transformed variables are distributed as $W(T^*\Sigma T^{*'}, N)$, where T^* is T augmented by a row and column vector of zeros and a one as the diagonal element. In sum, covariance matrices sampled from the original predictor space are distributed as $W(\Sigma, N)$ and sample covariance matrices for the transformed predictors are distributed $W(T^*\Sigma T^{*'}, N)$. Finally, because of the relationship in Equation A.9, the distribution of the squared cross-validity coefficient is unaffected by the difference between sampling from $W(\Sigma, N)$ and $W(T^*\Sigma T^{*'}, N)$ distributions: Sampling from either the original or the transformed predictor space yields squared cross-validity coefficients with identical distributions.