Cross-Validation Sample Sizes

James Algina, University of Florida

H. J. Keselman, University of Manitoba

The squared cross-validity coefficient is a measure of the predictive validity of a sample linear prediction equation. It provides a more realistic assessment of the usefulness of the equation than the squared multiple-correlation coefficient. The squared cross-validity coefficient cannot be larger than the squared multiple-correlation coefficient; its size is affected by the number of predictor variables and the size of the sample. Sample-size tables are presented that should result in very small discrepancies between the squared multiple correlation and the squared cross-validity correlation, thus facilitating the selection of sample size for predictive studies.
Index terms: cross-validity coefficient, least-squares regression, multiple correlation, prediction, sample size.

When regression analysis is used in a prediction context, it is important to distinguish between the *population* linear regression function,

$$\widetilde{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_J X_J \ , \tag{1}$$

and the *sample* linear prediction function,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_J X_J \ . \tag{2}$$

The squared multiple correlation is

$$\rho^2 = \rho^2_{Y\widetilde{Y}} = \frac{\left(\boldsymbol{\beta}'\boldsymbol{\sigma}_{xy}\right)^2}{\sigma^2_y \boldsymbol{\beta}' \boldsymbol{\Sigma}_{xx} \boldsymbol{\beta}} \ , \tag{3}$$

where $\boldsymbol{\beta}$ denotes the $J \times 1$ vector of population regression coefficients, $\boldsymbol{\sigma}_{xy}$ denotes the $J \times 1$ vector of covariances between the criterion variable (Y) and the predictors ($X_1, X_2, \ldots, X_J$), and $\boldsymbol{\Sigma}_{xx}$ denotes the $J \times J$ covariance matrix for $X_1, X_2, \ldots, X_J$. The coefficient ρ^2 measures the accuracy with which the population linear regression function predicts Y.

According to Browne (1975), the accuracy of predictions based on the sample linear prediction function can be measured by the squared cross-validity coefficient

$$\omega^2 = \rho^2_{Y\hat{Y}} = \frac{\left(\hat{\boldsymbol{\beta}}'\boldsymbol{\sigma}_{xy}\right)^2}{\sigma^2_y \hat{\boldsymbol{\beta}}' \boldsymbol{\Sigma}_{xx} \hat{\boldsymbol{\beta}}} \ , \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ denotes the $J \times 1$ vector of sample regression coefficients. As pointed out by Raju, Bilgic, Edwards, & Fleer (1997),

$$\omega^2 \leq \rho^2 \ , \tag{5}$$

Applied Psychological Measurement, Vol. 24 No. 2, June 2000, 173–179

with equality if and only if $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. As numerous authors indicate (e.g., Darlington, 1978), ω^2 is more important than ρ^2 in assessing the accuracy of prediction because the sample linear prediction function is used in practice to calculate predicted values of Y.

Due to the importance of ω^2, a substantial number of procedures have been developed for estimating ω^2 (e.g., Browne, 1975; Cattin, 1980a, 1980b; Darlington, 1978). A review of these procedures and their estimation accuracy was presented by Raju et al. (1997).

Although it is important to estimate ω^2, it is equally important to plan a prediction study so that ω^2 is sufficiently close to ρ^2. Sample size (N) affects the difference between ω^2 and ρ^2. That is, the larger the value of N, the smaller the expected shrinkage or disparity between ρ^2 and ω^2 (e.g., Raju et al., 1997). Of course, the required N is a function of the number of predictor variables. However, if the N required could be determined in advance so that the difference (c) between ρ^2 and ω^2 was at a desired small value (e.g., .025, .05, .075, or .10), the selection of N for a prediction study could be facilitated. A table to accomplish this is provided here.

Determining Cross-Validation Sample Sizes

Method

According to Browne (1975), ω^2 is invariant under nonsingular transformations of the predictors. That is, if $X_1, X_2, \ldots, X_J$ are replaced by J linear combinations of these variables and if none of these new variables is perfectly predictable from the remaining $J-1$, ω^2 will be unchanged by the transformation. Further, because $\boldsymbol{\Sigma}_{xx}$ is nonsingular, without loss of generality, it can be assumed that

$$\boldsymbol{\Sigma}_{xx} = \mathbf{I} \; , \tag{6}$$

$$\sigma_y^2 = 1 \; , \tag{7}$$

and

$$\boldsymbol{\sigma}_{xy} = \boldsymbol{\beta} \; , \tag{8}$$

where

$$\beta_1 = \rho \tag{9}$$

and

$$\beta_j = 0, \quad j = 2, \cdots, J \; . \tag{10}$$

Data. The population regression equation used to generate the data in this study was

$$Y = \rho X_1 + 0X_2 + \cdots + 0X_J + \varepsilon\sqrt{1-\rho^2} \; , \tag{11}$$

where $X_1, X_2, \ldots, X_J$ and ε are multivariate normal and mutually uncorrelated. The mean and variance of ε were 0 and 1, respectively. Samples of multivariate normal data were generated and ω^2 was calculated (see Equation 4) for 5,000 replications of each combination of (1) the number of predictor variables (J) from 2 to 20 in steps of 2, (2) squared multiple-correlation coefficients from .15 to .75 in steps of .10, and (3) sample sizes from 25 to 950 in steps of 25. The 5,000 values of ω^2 estimated the distribution of ω^2 for a particular combination of J, N, and ρ^2. Note that ω^2

is a parametric measure of cross-validity, that it varies across samples, and that it cannot be larger than ρ^2.

Analysis. The accuracy criterion was $c = \rho^2 - \omega^2$. For each combination of J, N, and ρ^2, the proportion of replications was determined in which $c \leq .025$, .05, .075, or .10. Next, the smallest sample size (N^*) was determined for each combination of J, ρ, and c, such that the probability was at least .95 that the accuracy criterion would be met. Then, multivariate normal data were generated 5,000 times for sample sizes between $N^* - 20$ and $N^* - 5$ in steps of 5, and ω^2 was computed. The smallest sample size was found such that the estimated probability that the accuracy criterion would be met was at least .95.

The distribution of ω^2 was not estimated for $N \leq 25$. Therefore, for some combinations of J, ρ^2, and c, when $N = 25$, the estimated probability, $P[\rho^2 - \omega^2 \leq c]$, was substantially larger than .95.

Results

Table 1 shows the smallest sample sizes for which $P[\rho^2 - \omega^2 \leq c] \approx .95$, for $c = .10$, .075, .05, and .025. In general, N increased as the number of predictors increased. N also increased as $\rho^2 - \omega^2$ became more stringent. The increase in N was particularly large when c was reduced from .05 to .025. For many combinations of ρ^2 and J, N nearly doubled when c changed from .05 to .025. For $c = .10$, .075, and .05, and as ρ^2 increased from .15 to .25, N was quite inconsistent, sometimes increasing, sometimes decreasing, and sometimes staying the same. N decreased with additional increases in ρ^2, except when the lower limit of $N = 25$ was reached. For $c = .025$, N decreased as ρ^2 increased. Thus, in using Table 1 to select an appropriate N, it is important to be conservative in specifying a value of ρ^2. The N indicated in Table 1 will tend to be too small to the degree that the specified value of ρ^2 is larger than the actual value of ρ^2.

Discussion

Raju, Bilgic, Edwards, & Fleer (1999) used data from the Armed Services Vocational Aptitude Battery to evaluate estimators of ω^2. It is of interest to compare the results reported by Raju et al., which were based on real (non-normal) data, and the results of the present study, which were based on multivariate normal data. Raju et. al reported that $\rho^2 = .229$ for $J = 8$ predictors. For $N = 200$, the mean and standard deviation of ω^2 were estimated to be .203 and .015, respectively. Using a normal distribution to approximate the distribution of ω^2, the .05 percentile point of the distribution is .178. Thus, with $N = 200$, the approximate $P[\rho^2 - \omega^2 \leq .05]$ was .95. In the present study, Table 1 shows that, when $J = 8$ and $\rho^2 = .25$, $N = 200$ for $P[\rho^2 - \omega^2 \leq .05] \approx .95$.

For $N = 100$, the mean and standard deviation of ω^2 were estimated to be .180 and .026, respectively. Based on the normal approximation, the .05 percentile point of the distribution is .137. Thus, with $N = 100$, the approximate $P[\rho^2 - \omega^2 \leq .092]$ was .95. In Table 1, when $J = 8$ and $\rho^2 = .25$, $N = 100$ for $P[\rho^2 - \omega^2 \leq .10] \approx .95$. Although the calculations based on the results in Raju et al. were approximate, they supported the accuracy of N for data that are non-normal to some degree.

Estimation of ω^2, whether by single or double cross-validation studies or by formula, is intended to provide a more realistic appraisal of the usefulness of a prediction equation. Many authors (e.g., Drasgow, Dorans, & Tucker, 1979; Raju et al., 1997, 1999) have found that formula-based procedures are as, if not more, effective for estimating the cross-validity coefficient. These same authors have indicated that the shrinkage (i.e., $\rho^2 - \omega^2$) expected is related to the number of predictor variables and the sample size. The results presented here provide sample sizes that should ensure that the difference between ρ^2 and ω^2 will be at some small value (i.e., .025, .05,

Table 1
Sample Size (N) and Estimated Probability ($\hat{P}$) Required for $P[(\rho^2 - \omega^2) \leq .10] \approx .95$ When $c = .10, .75, .05$, and .025, for $\rho^2 = .15$ to .75 and $J = 2$ to 20 Predictors

	ρ^2						
c and J	.15	.25	.35	.45	.55	.65	.75
$c = .10$							
$J = 2$							
N	35	35	30	25	25	25	25
$\hat{P}$	.950	.954	.950	.952	.965	.982	.996
$J = 4$							
N	60	60	55	50	45	35	30
$\hat{P}$	.957	.954	.956	.961	.968	.962	.976
$J = 6$							
N	75	80	75	65	55	50	40
$\hat{P}$	.953	.956	.961	.951	.951	.968	.971
$J = 8$							
N	90	100	95	85	70	60	55
$\hat{P}$	.959	.959	.960	.962	.954	.957	.955
$J = 10$							
N	100	115	110	100	85	70	55
$\hat{P}$	.957	.954	.957	.961	.957	.958	.959
$J = 12$							
N	115	130	125	115	100	80	65
$\hat{P}$	.957	.951	.958	.959	.952	.954	.964
$J = 14$							
N	125	145	140	125	110	90	70
$\hat{P}$	.954	.952	.954	.952	.951	.951	.957
$J = 16$							
N	135	160	160	140	125	105	80
$\hat{P}$	.954	.951	.954	.951	.956	.951	.956
$J = 18$							
N	150	180	175	155	135	115	90
$\hat{P}$	.957	.956	.959	.950	.951	.967	.963
$J = 20$							
N	160	190	185	170	150	120	95
$\hat{P}$	.956	.951	.957	.957	.963	.953	.957
$c = .075$							
$J = 2$							
N	50	45	40	35	25	25	25
$\hat{P}$	.958	.959	.956	.963	.952	.960	.984
$J = 4$							
N	80	80	70	65	55	45	35
$\hat{P}$	.953	.955	.952	.959	.956	.962	.961
$J = 6$							
N	110	110	100	85	75	60	45
$\hat{P}$	.950	.959	.954	.951	.955	.950	.954

continued on next page

Table 1, continued
Sample Size (N) and Estimated Probability ($\hat{P}$) Required for $P[(\rho^2 - \omega^2) \leq .10] \approx .95$ When $c = .10, .75, .05$, and $.025$, for $\rho^2 = .15$ to $.75$ and $J = 2$ to 20 Predictors

	ρ^2						
c and J	.15	.25	.35	.45	.55	.65	.75
$c = .075$ (continued)							
$J = 8$							
N	130	135	125	110	95	75	60
$\hat{P}$	.956	.953	.961	.961	.961	.953	.968
$J = 10$							
N	155	155	145	130	115	90	70
$\hat{P}$	.959	.951	.950	.956	.963	.957	.961
$J = 12$							
N	175	185	165	150	130	105	80
$\hat{P}$	.954	.956	.951	.954	.955	.955	.955
$J = 14$							
N	190	205	190	170	150	120	90
$\hat{P}$	.951	.953	.951	.952	.966	.955	.956
$J = 16$							
N	215	225	210	185	165	135	100
$\hat{P}$	.957	.952	.951	.952	.957	.954	.953
$J = 18$							
N	235	250	230	205	180	150	115
$\hat{P}$	.953	.954	.950	.952	.954	.961	.966
$J = 20$							
N	250	275	250	220	190	160	120
$\hat{P}$	.954	.961	.952	.951	.953	.954	.953
$c = .05$							
$J = 2$							
N	70	65	60	50	40	30	25
$\hat{P}$	.956	.956	.962	.952	.956	.950	.961
$J = 4$							
N	130	115	105	90	75	65	45
$\hat{P}$	.958	.950	.951	.951	.951	.960	.954
$J = 6$							
N	170	165	145	125	105	85	65
$\hat{P}$	.954	.956	.961	.955	.952	.952	.953
$J = 8$							
N	215	200	190	160	130	110	85
$\hat{P}$	.956	.953	.958	.954	.951	.961	.970
$J = 10$							
N	245	245	215	190	160	130	100
$\hat{P}$	.950	.954	.950	.951	.958	.957	.962
$J = 12$							
N	285	285	255	225	185	145	115
$\hat{P}$	.955	.961	.955	.957	.953	.950	.965
$J = 14$							
N	325	310	290	250	210	170	130
$\hat{P}$	.954	.951	.955	.951	.955	.952	.961

continued on next page

Table 1, continued

Sample Size (N) and Estimated Probability ($\hat{P}$) Required for $P[(\rho^2 - \omega^2) \leq .10] \approx .95$ When $c = .10, .75, .05,$ and $.025$, for $\rho^2 = .15$ to $.75$ and $J = 2$ to 20 Predictors

	ρ^2						
c and J	.15	.25	.35	.45	.55	.65	.75
$c = .05$ (continued)							
$J = 16$							
N	360	350	325	280	235	190	150
$\hat{P}$	.951	.954	.958	.954	.957	.951	.951
$J = 18$							
N	390	390	350	305	255	205	155
$\hat{P}$	.952	.956	.952	.952	.950	.951	.953
$J = 20$							
N	400	425	390	335	285	225	170
$\hat{P}$	.956	.958	.960	.954	.956	.951	.951
$c = .025$							
$J = 2$							
N	135	120	100	90	75	60	45
$\hat{P}$	.951	.954	.950	.950	.953	.961	.956
$J = 4$							
N	255	235	200	180	145	115	85
$\hat{P}$	.951	.951	.951	.957	.952	.950	.952
$J = 6$							
N	360	325	290	245	205	165	120
$\hat{P}$	.952	.953	.950	.950	.955	.959	.956
$J = 8$							
N	450	415	360	310	265	205	150
$\hat{P}$	.952	.952	.954	.952	.961	.956	.954
$J = 10$							
N	530	490	430	375	305	245	180
$\hat{P}$	.950	.951	.950	.951	.954	.955	.953
$J = 12$							
N	620	565	510	430	360	290	210
$\hat{P}$	.951	.955	.953	.957	.952	.960	.957
$J = 14$							
N	705	645	570	485	410	320	240
$\hat{P}$	.954	.951	.952	.952	.954	.950	.952
$J = 16$							
N	775	725	630	535	450	360	265
$\hat{P}$	.952	.955	.951	.951	.963	.953	.955
$J = 18$							
N	845	785	700	605	495	400	290
$\hat{P}$	.950	.952	.952	.954	.953	.952	.953
$J = 20$							
N	920	865	760	655	550	435	315
$\hat{P}$	.954	.957	.956	.952	.950	.952	.951

.075, or .10) approximately 95% of the time. These data facilitate planning predictor studies and provide confidence that the sample linear prediction function will give valid values according to the accuracy criterion used.

References

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

Cattin, P. (1980a). Estimation of a regression model. *Journal of Applied Psychology, 65,* 407–414.

Cattin, P. (1980b). Note on the estimation of the squared cross-validated multiple correlation of a regression model. *Psychological Bulletin, 87,* 63–65.

Darlington, R. B. (1978). Reduced-variance regression. *Psychological Bulletin, 85,* 1238–1255.

Drasgow, F., Dorans, N. J., & Tucker, L. R. (1979). Estimators of the squared cross-validity coefficient: A monte carlo investigation. *Applied Psychological Measurement, 3,* 171–192.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1997). Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction. *Applied Psychological Measurement, 21,* 291–305.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement, 23,* 99–115.

Author's Address

Send requests for reprints or further information to James Algina, 1403 Norman Hall, P.O. Box 117047, Gainesville FL 32611-7047. U.S.A. Email: algina@nersp.nerdc.ufl.edu.